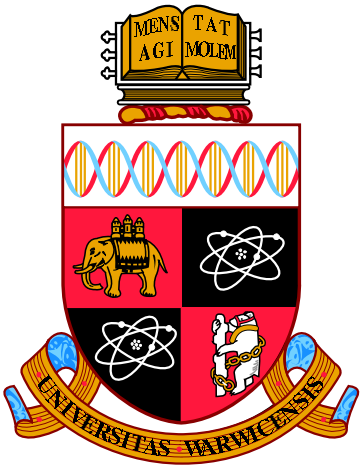


Optimization-centric Generalizations of Bayesian Inference

by

Jeremias Knoblauch

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

December 2021

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I am extremely thankful to my supervisor Theo Damoulas for consistently enabling and encouraging me to pursue my own research interests, letting me collaborate freely with other researchers, and supporting me with every endeavour I undertook. Over the five years we spent together, Theo was always there for me, and never shied away from advising me on even the most insignificant details that I enquired about. Throughout, I felt very much taken care of on a personal, academic, and intellectual level: Theo taught me how to think about and conduct research, how to write papers, and how to present one's work and oneself. The debt of gratitude I owe him certainly is beyond the brief space I can allocate to my written acknowledgements—and so I will simply say that without him, I would not be the researcher—and more importantly: person—that I am today.

I also want to thank Lara Vomfell for going through most of this adventure with me as my closest friend and partner, and being supportive and understanding throughout—none of this would have been possible without you. Special thanks also go to Jack Jewson, without whom I would never have started thinking about generalized Bayesian methods in the first place. Beyond becoming a close friend, Jack also became one of my most important collaborators; and I will forever remain indebted to him for parts of the intellectual journey I took over the course of my PhD. Similarly, I want to thank my collaborator and friend Hisham Husain for the endless whiteboard sessions, the fantastic discussions, and the lion's roar. Hisham's intellectual curiosity and creativity remains inspiring to this day, and something I will forever seek to emulate. I also owe an enormous debt of gratitude to Francois-Xavier Briol, who has become a close friend and collaborator; and has been instrumental

in helping me both academically and career-wise.

In no particular order, I also want to thank Takuo Matsubara, Chris Oates, Charita Dellaportas, Ollie Hamelijnck, Juan Maronas, Sebastian Schmon, Patrick Cannon, Claire Vernade, Omar Rivasplata, Tom Diethe, David Dunson, and Pierre Alquier for allowing me to collaborate with them on various projects that have been instrumental for the contents of this thesis. Particular thanks goes to David Dunson, who hosted me at Duke University for 4 months and introduced me to many researchers and ideas during my stay.

A big thanks also goes to my friends; and particularly Fabian Prietzel and Sander Aarts, who were always there when I needed them. I am also thankful to the 'Botley Crew', and my OxWaSP cohort more generally for making my PhD as a whole and the first few years in particular such a socially pleasant affair.

Last but not least, I am also very grateful to the Engineering and Physical Sciences Research Council (EPSRC) as well as Facebook Research for supporting me and my research financially. the Alan Turing Institute in general as well as both the Data-Centric Engineering Group and Mark Girolami in particular for my time there as visiting researcher.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented was carried out by the author. Parts of this thesis have been previously published by the author as the sole lead contributor as listed below.

1. J. Knoblauch and T. Damoulas. Spatio-temporal Bayesian on-line change-point detection with model selection. *Proceedings of the 35th International Conference on Machine Learning*, pages 2718-2727, 2018.

2. J. Knoblauch, J. Jewson, and T. Damoulas. Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β-divergences. *Advances in Neural Information Processing Systems 31*, 2018.

3. J. Knoblauch, J. Jewson, and T. Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *to appear in the Journal of Machine Learning*, 2022.

4. J. Knoblauch. Robust Deep Gaussian Processes. *technical report; arXiv preprint 1904.02303*, 2019.

5. J. Knoblauch. Frequentist consistency of Generalized Variational Inference. *arXiv preprint 1912.04946*, 2019.

Additional research that forms part of this thesis was performed in collaboration as second author.

6. H. Husain, J. Knoblauch. Adversarial Interpretation of Bayesian Inference, *submitted to ALT 2022: International Conference on Algorithmic Learning Theory*

   ▷ The main ideas for this work originated from conversations and discussions with Hisham Husain, and were worked out relatively evenly between the two of us. I took the lead on interpreting our findings, on writing, and on the interpretation of the results. Hisham took the lead on the more involved functional analysis, and on the connection with Wasserstein Autoencoders.

7. T. Matsubara, J. Knoblauch, F.X. Briol, and C. Oates. Robust generalised Bayesian inference for intractable likelihoods. *submitted to the Journal of the Royal Statistical Society, Series B*, 2021.

   ▷ Most of the main ideas of this article were developed by Takuo Matsubara and myself. For the research, Takuo executed the majority of the work being supported by the other co-authors: I focused on questions of asymptotic analysis; Francois-Xavier Briol focused on extensions of conjugacy as they pertain to generalised Bayesian methods; and Chris Oates improved the work through his extensive knowledge of kernel-based methods and contributed the majority of the code.

Additional published research that was undertaken and does **not** form part of this thesis is listed below:

8. J. Knoblauch and L. Vomfell (equal contribution). Robust Bayesian Inference for Discrete Outcomes with the Total Variation Distance. *arXiv preprint 1912.04946*, 2019.

9. J. Knoblauch, H. Husain, and Tom Diethe. Optimal Continual Learning has perfect memory and is NP-hard. *Proceedings of the 37th International Conference on Machine Learning*, pages 5327-5337, 2020.

10 S. Schmon, P. Cannon, and J. Knoblauch. Generalized Posteriors in Approximate Bayesian Computation. *3rd Symposium on Advances in Approximate*

Bayesian Computation, 2020.

11. J. Maronas, O. Hameljinck, J. Knoblauch, and T. Damoulas. Transforming Gaussian processes with normalizing flows. *Artificial Intelligence and Statistics*, 2021.

12. C. Dellaportas, J. Knoblauch, T. Damoulas, and F. X. Briol. Robust Bayesian Inference for Simulator-based Models via the MMD Posterior Bootstrap. *submitted to Artificial Intelligence and Statistics*, 2022.

Abstract

The mathematical machinery underlying Bayesian inference is Bayes' Rule—an important and elegant result dating back to the 18th century (Bayes, 1763). But in statistical practice, mathematical elegance alone is not enough: for Bayes' Rule to be a useful practical device, we need to impose a number of stringent assumptions that often do not reflect the realities of modern statistical and Machine Learning applications. This thesis sets out to propose and apply formalisms that are useful in situations where the assumptions underlying Bayes' Rule are dramatically violated. These assumptions include the presumption of a correctly specified statistical model, prior information of sufficient quality to improve the posterior belief, and adequate computational power. The violations of these assumptions and the proposed remedies will be explored theoretically and methodologically, but also empirically on a number of Machine Learning applications.

Acronyms

AM	Adams & MacKay (2007).
AR	Autoregression.
ARGPCP	Autoregressive Gaussian Process changepoint (model).
BAR	Bayesian autoregression.
BBGVI	Black Box generalized variational inference.
BBVI	Black Box variational inference.
BF	Bayes factor.
BF	Stochastic variance-reduced gradient.
BLR	Bayesian linear regression.
BMM	Bayesian mixture models.
BNN	Bayesian neural network.
BOCPD	Bayesian On-line Changepoint Detection.
BOCPDMS	Bayesian On-line Changepoint Detection with Model Selection.
BVAR	Bayesian vector autoregression.

CP	Changepoint.
DGM	Data-generating mechanism.
DGP	Deep Gaussian Process.
DVI	Discrepancy variational inference.
ELBO	Evidence Lower Bound.
EP	Expectation Propagation.
FDR	False discovery rate.
FL	Fearnhead & Liu (2007).
GARCH	Generalized autoregressive conditionally heteroscedastic.
GBI	Generalized Bayesian Inference.
GP	Gaussian Process.
GPTSCP	Gaussian Process time series changepoint (model).
GVI	Generalized Variational Inference.
HMM	Hidden Markov model.
IMQ	Inverse multi-quadratic (kernel).
IPM	Integral Probability Metric.
KLD	Kullback-Leibler Divergence.
KSD	Kernel-Stein Discrepancy.

LLN	Law of Large Numbers.
MAP	Maximum a posteriori.
MCMC	Markov Chain Monte Carlo.
MD	Mahalanobis distance.
ML	Maximum likelihood.
MMD	Maximum Mean Discrepancy.
MSE	Mean square error.
MVN	Multivariate normal (distribution).
NIG	Normal inverse-gamma.
NLL	Negative log likelihood.
NOX	Nitrogen Oxide.
NSGP	Non-stationary Gaussian Process (model).
OLS	Ordinary least squares.
PAC	Probably Approximately Correct.
PIB	Predictive Information Bottleneck.
PPM	Product partition model.
RLD	Run-length distribution.
RMSE	Root mean square error.

ROT	Rule of Three.
SGD	Stochastic gradient descent.
SGV	Standardized generalized variance.
SIC	Standard industrial classification.
SPD	Semi-positive definite.
SSBVAR	Spatially structured Bayesian vector autoregression.
SVI	Stochastic variational inference.
VAE	Variational Auto-encoder.
VAR	Vector autoregression.
VARMA	Vector autoregressive moving average.
VI	Variational inference.
YWE	Yule-Walker estimator.

Symbols

$p(\cdot\|\boldsymbol{\theta})$	Likelihood model on the data space, indexed by parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
$\boldsymbol{\Theta}$	Parameter space; throughout assumed to admit Polish topology.
$\mathcal{X}$	Data space; throughout assumed to admit Polish topology.
$x_{1:n}$	Data set of n observations so that $x_{1:n} \in \mathcal{X}^n$.
π	Prior belief distribution $\pi \in \mathcal{P}(\boldsymbol{\Theta})$ on the parameter space $\boldsymbol{\Theta}$.
$q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$	Standard Bayes posterior (for a given prior π and likelihood function $p(\cdot\|\boldsymbol{\theta})$; $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta}) = P(-\log p(\cdot\|\boldsymbol{\theta}), \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$.
L	Loss function $L : \boldsymbol{\Theta} \times \mathcal{X}^n \to \mathbb{R}$.
ℓ	Loss function $\ell : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}$ that is assumed to be additive so that $L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$.
D	Statistical divergence $D : \mathcal{P}(\boldsymbol{\Theta})^2 \to \mathbb{R}_{\geq 0}$.
Π	Some subset of $\mathcal{P}(\boldsymbol{\Theta})$.
$\mathcal{Q}$	A subset of $\mathcal{P}(\boldsymbol{\Theta})$ parametrized by a set of parameters $\boldsymbol{K}$; $\mathcal{Q}$ is also sometimes called a variational family.
$q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$	Generalized Bayes posterior (for a given prior π and loss function $L : \boldsymbol{\Theta} \times \mathcal{X}^n \to \mathbb{R}$; $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta}) = P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$.
$P(L, D, \Pi)$	Shorthand notation for posteiors induced by the Rule of Three (RoT) specified through the loss L, the divergence D, and space Π.

$q^*_{\mathrm{A}}(\boldsymbol{\theta})$	A posterior in some parametrized subset $\mathcal{Q}$ approximating $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$ or $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ by some means.
$q^*_{\mathrm{VI}}(\boldsymbol{\theta})$	A posterior approximating $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$ or $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ via Variational Inference (VI); so that $q^*_{\mathrm{VI}}(\boldsymbol{\theta}) = P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$.
$q^*_{\mathrm{DVI}}(\boldsymbol{\theta})$	A posterior approximating $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$ or $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ via Discrepancy Variational Inference (DVI); so that $q^*_{\mathrm{DVI}}(\boldsymbol{\theta}) = \arg\min_{q\in\mathcal{Q}} \left\{ D(q(\boldsymbol{\theta}) \| q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})) \right\}$.
$\mathcal{F}_b(\boldsymbol{\Theta})$	The set of bounded and measurable functions on $\boldsymbol{\Theta}$.
$\mathcal{B}(\boldsymbol{\Theta})$	The set of finitely-additive measures on $\boldsymbol{\Theta}$.
$\mathcal{G}_{L,w^{-1}D,\Pi}$	The objective value associated with $P(L, w^{-1}D, \Pi)$.
$\mathrm{co}\,(A)$	The convex hull of a set A.
$\overline{\mathrm{co}}\,(A)$	The closure of the convex hull of a set A.
$D^\star_\pi(L')$	The Legendre-Fenchel conjugate of a statistical divergence $D(\cdot\|\pi) : \mathcal{P}(\boldsymbol{\Theta}) \to \mathbb{R}_{\geq 0}$ relative to some loss $L' \in \mathcal{F}_b(\boldsymbol{\Theta})$.
$\mathsf{E}_\Pi(L)$	The minimum of L over elements of Π, ie $\mathsf{E}_\Pi(L) = \inf_{q\in\Pi} \mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta}, x_{1:n})\right]$.
ρ	A perturbation $\rho \in \mathcal{F}_b(\boldsymbol{\Theta})$ by an adversary (relative to L).
$\mathcal{H}$	A subset of $\boldsymbol{\Theta}^{\mathbb{R}}$ such as a Reproducing Kernel Hilbert Space or $\mathcal{F}_b(\boldsymbol{\Theta})$.
$d_{\mathcal{H}}(q, \pi)$	The IPM between q and π relative to the function space $\mathcal{H}$ so that $d_{\mathcal{H}}(q,\pi) = \sup_{h\in\mathcal{H}} \{\mathbb{E}_q[h] - \mathbb{E}_\pi[h]\}$.
$\boldsymbol{\theta}^*$	The population-optimal value of $\boldsymbol{\theta}$ for a given loss $\ell : \boldsymbol{\Theta}\times\mathcal{X} \to \mathbb{R}$ so that for some $\mu \in \mathcal{P}(\mathcal{X})$, we have $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]$ If we are in the well-specified case so that $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ and μ admits a density $p(\cdot|\boldsymbol{\theta}^*)$, then $\boldsymbol{\theta}^*$ is also the value of $\boldsymbol{\theta}$ for which $p(\cdot|\boldsymbol{\theta})$ recovers the true data-generating process.

O_{VI}	The objective of a VI posterior whose variational family is parametrized by $\boldsymbol{\kappa} \in \boldsymbol{K}$, so that $O_{\text{VI}} : \boldsymbol{K} \to \mathbb{R}$.
O_{GVI}	The objective of a GVI posterior whose variational family is parametrized by $\boldsymbol{\kappa} \in \boldsymbol{K}$, so that $O_{\text{GVI}} : \boldsymbol{K} \to \mathbb{R}$.
$\widehat{O_{\text{VI}}}$	Estimate of O_{VI}.
$\widehat{O_{\text{GVI}}}$	Estimate of O_{GVI}.
$D_{AR}^{(\alpha)}$	Rényi's α-divergence parametrized with parameter α.
$D_{A}^{(\alpha)}$	α-divergence parametrized with parameter α.
$D_{B}^{(\beta)}$	β-divergence parametrized with parameter β.
$D_{G}^{(\gamma)}$	γ-divergence parametrized with parameter γ.
$\mathcal{L}_p^{\beta}$	Loss based on β-divergence; $\mathcal{L}_p^{\beta} : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}$; we also write $L^{\beta}(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^{n} \mathcal{L}_p^{\beta}(\boldsymbol{\theta}, x_i)$.
$\mathcal{L}_p^{\gamma}$	Loss based on γ-divergence; $\mathcal{L}_p^{\gamma} : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}$; we also write $L^{\gamma}(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^{n} \mathcal{L}_p^{\gamma}(\boldsymbol{\theta}, x_i)$.
S_h	Slack term based on a hyperparameter h.
$\mathcal{IG}(a, b)$	Inverse-Gamma distribution with shape a and scale b.
$\mathcal{N}(\boldsymbol{\mu}, \Sigma)$	(Multivariate) Normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
$\mathcal{O}(g(n))$	Big-O notation; a function $f(n)$ is of order $O(g(n))$ if it scales as $c \cdot g(n)$ for large enough.
$L^q(\mathcal{X}, \mathbb{Q})$	For $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$, $L^q(\mathcal{X}, \mathbb{Q})$ is both the set of functions $f : \mathcal{X} \to \mathbb{R}$ for which $\|f\|_{L^q(\mathcal{X},\mathbb{Q})} := (\int_{\mathcal{X}} \lvert f \rvert^q \mathrm{d}\mathbb{Q})^{1/q} < \infty$ and the normed space in which two elements $f, g \in L^q(\mathcal{X}, \mathbb{Q})$ are identified if they are $\mathbb{Q}$-almost everywhere equal.
$L^q(\mathcal{X})$	$L^q(\mathcal{X}, \mathbb{Q})$, if $\mathbb{Q}$ is a Lebesgue measure.
$\mathcal{P}_{\text{S}}(\mathbb{R}^d)$	The subset of Borel measures $\mathcal{P}(\mathbb{R}^d)$ on $\mathbb{R}^d$ that admit an everywhere positive probability density function (pdf) $p : \mathcal{X} \to \mathbb{R}_{>0}$ with continuous partial derivatives.
$\mathcal{S}_{\mathbb{Q}}$	A Stein operator defined relative to a Stein set $\mathcal{H}$ with $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$ so that $\mathbb{E}_{X \sim \mathbb{Q}}[\mathcal{S}_{\mathbb{Q}}[h](X)] = 0 \quad \forall h \in \mathcal{H}$.

$\mathcal{H}$	A Stein set $\mathcal{H}$ defined relative to a Stein operator $\mathcal{S}_{\mathbb{Q}}$ with $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$ so that $\mathbb{E}_{X\sim\mathbb{Q}}[\mathcal{S}_{\mathbb{Q}}[h](X)] = 0 \quad \forall h \in \mathcal{H}$.
$\mathrm{SD}(\mathbb{Q}\|\mathbb{P})$	A Stein discrepancy between distributions $\mathbb{Q}, \mathbb{P} \in \mathcal{P}(\mathcal{X})$, which for a Stein operator $\mathcal{S}_{\mathbb{Q}}$ and a Stein set $\mathcal{H}$ is given as $\mathrm{SD}(\mathbb{Q}\|\mathbb{P}) = \sup_{\|h\|_{\mathcal{H}}\leq 1} \left\lvert \mathbb{E}_{X\sim\mathbb{P}} \left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] \right\rvert$.
$\mathbb{P}_n$	Empirical measure $\mathbb{P}_n = \sum_{i=1}^n \delta_{x_i}$ based on $x_{1:n}$.
$\mathbb{P}_{\boldsymbol{\theta}}$	Measure $\mathbb{P}_{\boldsymbol{\theta}}$ induced by the model density $p(\cdot\|\boldsymbol{\theta})$.
K / k	Matrix-valued / vector-valued kernel function associated with a reproducing kernel Hilbert space $\mathcal{H}$.
$\mathrm{KSD}(\mathbb{Q}\|\mathbb{P})$	Kernel-Stein Discrepancy between measures $\mathbb{Q}, \mathbb{P} \in \mathcal{P}(\mathcal{X})$ given by $\mathrm{KSD}^2(\mathbb{Q}\|\mathbb{P}) := \mathbb{E}_{X,X'\sim\mathbb{P}} \left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(X, X')\right]$.
$\|\cdot\|_2$	Euclidean norm.
$C(\mathcal{X})$	The set of continuous functions $f : \mathcal{X} \to \mathbb{R}$.
$C_b^1(\mathbb{R}^d)$	The set of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that both f and the partial derivatives $x \mapsto (\partial/\partial x_{(i)})f(x)$ are bounded and continuous on $\mathbb{R}^d$.
$C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d)$	The set of bivariate functions $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that both f and the partial derivatives $(x, x') \mapsto (\partial/\partial x_{(i)})(\partial/\partial x'_{(j)})f(x, x')$ are bounded and continuous on $\mathbb{R}^d \times \mathbb{R}^d$.
$\mathcal{S}(\mathcal{X}; \mathbb{R}^k)$	For an arbitrary set $\mathcal{S}(\mathcal{X})$ of functions $f : \mathcal{X} \to \mathbb{R}$, denote by $\mathcal{S}(\mathcal{X}; \mathbb{R}^k)$ the set of $\mathbb{R}^k$-valued functions whose components belong to $\mathcal{S}(\mathcal{X})$.
∇	The gradient operator on $\mathbb{R}^d$; often we write ∇_x to indicate the argument x to which the operator is applied (eg $\nabla_x f(x, y)$).
$\nabla\cdot$	The divergence operator on $\mathbb{R}^d$; often we write $\nabla_x\cdot$ to indicate the argument x to which the operator is applied (eg, $\nabla_x \cdot f(x, y)$).

$L_{\mathrm{KSD}}(x_{1:n}, \boldsymbol{\theta})$	The Loss based on the squared KSD^2 given by $L_{\mathrm{KSD}}(x_{1:n}, \boldsymbol{\theta}) = n \cdot \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n)$.
$\pi_n^{\mathrm{KSD}}(\boldsymbol{\theta})$	The KSD-Bayes posterior given by $\pi_n^{\mathrm{KSD}}(\boldsymbol{\theta}) = P(L_{\mathrm{KSD}}, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, so that $\pi_n^{\mathrm{KSD}}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\{-wL_{\mathrm{KSD}}(x_{1:n}, \boldsymbol{\theta})\}$.
∂^k	The partial derivative operator (∂^k/∂) for $k \in \mathbb{N}$; so that $[\nabla_x f(x)]_{(h)} = (\partial/\partial x_{(h)}) f(x)$ and $[\nabla_x^2 f(x)]_{(h,k)} = (\partial^2/\partial x_{(h)} x_{(k)}) f(x)$.
$\mathrm{PIF}(y, \boldsymbol{\theta}, \mathbb{P}_n)$	The posterior influence function given by $\mathrm{PIF}(y, \boldsymbol{\theta}, \mathbb{P}_n) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \pi_n^{L(\boldsymbol{\theta}; \mathbb{P}_{n,\varepsilon,y})}|_{\varepsilon=0}$, where $\mathbb{P}_{n,\varepsilon,y} = (1-\varepsilon)\mathbb{P}_n + \varepsilon\delta_y$ is the ε-contamination model.

Foreword & About

This thesis collects my thoughts on and contributions to the field of generalized Bayesian methodology. In my roughly five years within the Oxford-Warwick Statistics Programme, I have come to witness that there is a large gap between the foundation and application of standard Bayesian methodology—particularly in computationally demanding and emerging fields such as Machine Learning or simulator-based inference. This thesis and my continued research is part of the effort to close this gap, and I believe that over the coming years we will succeed in doing this—by extending and adjusting what it means to conduct Bayesian analysis. It is my personal conviction that we will achieve this by thoroughly departing from the idea of using Bayes' Rule as the de-facto default method for deriving belief distributions; and it is my hope that this thesis will convince the esteemed reader of this vision.

The structure of the thesis is fourfold: The introduction in Chapter 1 presents the main object of study throughout the thesis: the Rule of Three (RoT). The RoT is an optimization-centric generalization of Bayesian inference that recovers previous extensions of Bayes-like procedures, and can be motivated both intuitively and axiomatically. In the thesis' first part (Chapters 2 and 3), we study various theoretical properties of this optimization-centric generalization of Bayesian inference. Some of the highlights relate to a new interpretation of Bayesian inference as an adversarially robust procedure that has a game-theoretic interpretation; as well as a broadly applicable result on frequentist consistency. The second part of the thesis (Chapters 4–6) introduces the methodology of Generalized Variational Inference (GVI) as one of the main practical fallouts from the RoT for Machine Learning. This is complemented by the third part of the thesis (Chapters 7 and 8), which illustrates how the theory and methods discussed in the thesis can aid in two applications that are often adversely affected by the severe misalignment between the theoretical basis of standard Bayesian inference and the real world: on-line changepoint detection models and intractable likelihoods. Lastly, we discuss our contributions in Chapter 9 and provide further details in Appendices A, B, and C.

Chapter 1

Introduction & Motivation

Summary: In this first chapter, we take a closer look at the Bayesian paradigm and introduce the core theme of this thesis. Specifically, we aim to answer the following questions: what is the logic underlying Bayesian inference? What makes this logic inappropriate for modern large-scale statistical problems such as Machine Learning applications? And most importantly: which device are we proposing to remedy these issues, and how does the approach put forward in this thesis differ from what has been suggested in related previous work?

Though first discovered by the Reverend Thomas Bayes (1763), the version of Bayes' Theorem that a modern audience would be familiar with is much closer to the one in De Laplace (1774). Bayes' rule is one of the most fundamental results in probability theory and states that for two events A, B, it holds that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)}.$$

where as usual, $\mathbb{P}(A|B)$ denotes the conditional probability of event A given that event B occurred. It would take nearly two more centuries for this mathematical result to be used as the basis for an entire school of statistical inference (for a full account of this history, see Fienberg, 2006). More precisely, Fisher (1950) provides the first mention of the term *Bayesian* in accordance with our modern understanding (David, 1998).

Bayesian statistics uses Bayes' Theorem to conduct inference on an unknown and unobservable event A. Specifically, suppose that one can compute for an observable event B the probability $\mathbb{P}(B|A)$ and has a prior belief $\mathbb{P}(A)$ about the event A before observing B. Now, Bayes' rule tells us that we should be able to draw

probabilistic inferences on $A|B$ by computing the probability $\mathbb{P}(A|B)$. In practice, the event A quantifies the uncertainty about a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ indexing a statistical model, and so is of the form $A \subset \boldsymbol{\Theta}$. The prior beliefs about events A are usually specified by some probability density $\pi : \boldsymbol{\Theta} \to \mathbb{R}_+$ inducing the probability measure $\mathbb{P}(A) = \int_A d\pi(\boldsymbol{\theta})$. This leaves us with the need to specify a probability distribution $\mathbb{P}(B|A)$ that relates the (unobserved) parameter $\boldsymbol{\theta}$ to the (observable) event B. In practice, B will correspond to the event that n random variables $\boldsymbol{x}_{1:n}$ take certain observable values $x_{1:n} \in \mathcal{X}^n$. The next step is to hypothesise a distribution of $B|A$, which amounts to positing a likelihood function $p(x_{1:n}|\boldsymbol{\theta})$ and setting $\mathbb{P}(B|A) = p(x_{1:n}|\boldsymbol{\theta})$.[1] If both the prior π and the likelihood function p admit densities with respect to the Lebesgue measure, we can put all this together and obtain the *standard Bayesian posterior* which we denote as $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ throughout the thesis and which is given by

$$q^*_{n,\mathrm{SB}}(\boldsymbol{\theta}) = \frac{p(x_{1:n}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z}.$$

Here, $Z = \int_{\boldsymbol{\Theta}} p(x_{1:n}|\boldsymbol{\theta})d\pi(\boldsymbol{\theta})$ is the normalizing constant—also known as the partition function—whose intractability is what makes computing the Bayesian posterior an often rather involved problem.

Bayesian inference is appealing both conceptually and practically: First and foremost, through Bayes' Rule, it is based on a clear mathematical principle. Further and unlike alternative frameworks of analysis such as Frequentist statistics, Bayesian methods allow inferences to be informed by domain expertise in the form of a carefully specified prior belief $\pi(\boldsymbol{\theta})$. Furthermore, Bayesian inference produces belief distributions (rather than point estimates) over the parameter of interest $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. As a consequence, Bayesian inferences automatically quantify uncertainty about $\boldsymbol{\theta}$. This is practically useful in many situations, but especially if one uses $\boldsymbol{\theta}$ predictively: Integrating over $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ avoids being over-confident about the best value of $\boldsymbol{\theta}$, substantially improving predictive performance (see e.g. Aitchison, 1975). Amongst other benefits, it is this enhanced predictive performance that has cast Bayesian inference as one of the predominant paradigms in contemporary large-scale statistical inference and Machine Learning.

While Bayesian methods automatically quantify the uncertainty about their inferences, this comes at a cost: In the translation of Bayes' rule into the Bayesian

[1] For pedagogical reasons, we have treated both $x_{1:n}$ and $\boldsymbol{\theta}$ as if they were discrete so that $A = \{\boldsymbol{\theta} = \boldsymbol{\theta}'\}$ and $B = \{\boldsymbol{x}_{1:n} = x_{1:n}\}$ are well-defined events for any $\boldsymbol{\theta}' \in \boldsymbol{\Theta}$ and $x_{1:n} \in \mathcal{X}^n$. While this does not describe most situations of interest, the underlying logic of setting $\mathbb{P}(B|A) = p(x_{1:n}|\boldsymbol{\theta})$ is applicable more broadly with careful caveats, and in particular with continuously-valued variables.

posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$, we have made three implicit but crucial assumptions. First, we assumed that the modeller has a prior belief which is worth being taken into account and which the modeller is capable of writing out mathematically as $\pi(\boldsymbol{\theta})$. Second, we specified the likelihood function $p(x_{1:n}|\boldsymbol{\theta})$ as a conditional probability. In other words, we assumed that the model is correctly specified, which is to say that $p(x_{1:n}|\boldsymbol{\theta}^*) = \mathbb{P}(x_{1:n})$ for some unknown value of $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$. Third, we assumed the availability of enough computational power to make use of the often intractable posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$. In many situations, these three assumptions built into $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ are harmless. For a range of modern large-scale statistical problems including high-dimensional inference, simulator-based models, or Machine Learning however, they are frequently violated.

To address this, we take a step back from the traditional interpretation of the Bayesian posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ as an updating rule—Instead, we adopt an optimization-centric view point. Throughout, we motivate this with the tensions and contradictions between the three main assumptions underlying standard Bayesian inference on the one hand, and the real world characteristics of many contemporary statistical applications on the other hand. Aimed at resolving these tensions and contradictions, we define an optimization-centric generalization of Bayesian inference that we call the Rule of Three (RoT). The RoT is specified by an optimization problem over the space of Borel probability measures $\mathcal{P}(\boldsymbol{\Theta})$ on $\boldsymbol{\Theta}$ with three arguments. These arguments are a loss function $L : \boldsymbol{\Theta} \times \mathcal{X}^n \to \mathbb{R}$ which will often be additive so that $L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$, a divergence D measuring the deviation of the posterior from the prior and a space $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$ of feasible solutions. Together, these three ingredients define posterior beliefs of the form

$$q^*(\boldsymbol{\theta}) = \operatorname*{arg\,min}_{q \in \Pi} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right] + D(q \| \pi) \right\} \stackrel{\text{def}}{=} P(L, D, \Pi). \tag{1.1}$$

While this objective clearly also depends on two additional arguments—data $x_{1:n}$ and a prior π—we consider these fixed throughout and thus notationally suppress this dependence. Note that whenever the loss is additive so that $L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$, we also define

$$P(L, D, \Pi) \stackrel{\text{def}}{=} P(\ell, D, \Pi).$$

Though this may not be obvious, (1.1) in fact has an intimate relationship with the standard Bayesian posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$, as the following result shows.

Theorem 1.1. If $L(\boldsymbol{\theta}, x_{1:n}) = -\log p(x_{1:n}|\boldsymbol{\theta})$, $D = \mathrm{KLD}$, $\Pi = \mathcal{P}(\boldsymbol{\Theta})$; and if $Z =$

$\int_{\boldsymbol{\Theta}} \exp\{-L(\boldsymbol{\theta}, x_{1:n})\}\,\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is so that $0 < Z < \infty$, then $P(L, D, \Pi) = q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$.

Proof. One may rewrite the objective of (1.1) as

$$\begin{aligned} q^*(\boldsymbol{\theta}) &= \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \int_{\boldsymbol{\Theta}} \left[\log\left(\exp\{L(\boldsymbol{\theta}, x_{1:n})\}\right) + \log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)\right] q(\boldsymbol{\theta})d\boldsymbol{\theta} \right\} \\ &= \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \int_{\boldsymbol{\Theta}} \log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})\exp\{-L(\boldsymbol{\theta}, x_{1:n})\}}\right) q(\boldsymbol{\theta})d\boldsymbol{\theta} \right\}. \end{aligned}$$

As one only cares about the minimizer $q^*(\boldsymbol{\theta})$ (and not the objective value), it also holds that for any constant $Z > 0$, the above is equal to

$$\begin{aligned} q^*(\boldsymbol{\theta}) &= \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \int_{\boldsymbol{\Theta}} \log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})\exp\{-L(\boldsymbol{\theta}, x_{1:n})\}\,Z^{-1}}\right) q(\boldsymbol{\theta})d\boldsymbol{\theta} - \log Z \right\} \\ &= \underset{q\in\mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \mathrm{KLD}\left(q(\boldsymbol{\theta}) \middle\| \pi(\boldsymbol{\theta})\exp\{-L(\boldsymbol{\theta}, x_{1:n})\}\,Z^{-1}\right)\right\}. \end{aligned}$$

Lastly, one sets $Z = \int_{\boldsymbol{\theta}} \exp\{-L(\boldsymbol{\theta}, x_{1:n})\}\,\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ and notes that as the KLD is a statistical divergence, it is minimized uniquely if its two arguments are the same, so $q^*(\boldsymbol{\theta}) = q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$. □

The proof of Theorem 1.1 is essentially a restatement from the supplement of Bissiri et al. (2016), and is just a simple application of well-known results (see for instance Csiszár (1975) or Donsker and Varadhan (1975)). It is important in the context of this thesis because it implies that that the standard Bayesian posterior can be thought of as the solution of an infinite-dimensional optimisation problem that is structurally the same as the RoT. From this, it is clear that the RoT recovers $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$. Beyond that, it also recovers previous generalizations of Bayesian inference, including those inspired by Gibbs posteriors (e.g. Ghosh and Basu, 2016; Bissiri et al., 2016; Jewson et al., 2018; Nakagawa and Hashimoto, 2019; Chérief Abdellatif and Alquier, 2020), tempered posteriors (e.g. Grünwald, 2011, 2012; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019), as well as PAC-Bayesian approaches (for a recent overview, see Guedj, 2019). Unlike any of these previous generalizations however, posteriors taking the form $P(L, D, \Pi)$ need **not** be exponentially additive. We point out this relationship between the RoT and other Bayes-like procedures in Figure 1.1.

Before we move on to studying this generalization, its applications, and its associated methodologies in more detail, we first need to answer one crucial question: why is it needed?

1.1 A reality check: Re-examining the traditional Bayesian paradigm

This thesis argues for a generalized view on Bayesian methodology. The remainder of this introduction explains why. In particular, we illuminate the misalignment between the assumptions underlying the traditional Bayesian paradigm and the way in which modern statistical Machine Learning uses (approximate) Bayesian posteriors to conduct inference. We do so in three consecutive steps:

First, **Section 1.1.1** elaborates on the three crucial assumptions underlying the standard Bayesian posterior: Appropriate specification of prior **(P)** and likelihood **(L)** as well as an infinite computational budget **(C)**.

Next, **Section 1.1.2** exposes the misalignment of these three assumptions with inferential practices in contemporary statistical analysis for Machine Learning, and complex large-scale inference problems.

Lastly, **Sections 1.1.3–1.1.5** points to the adverse real-world consequences arising from violating these assumptions.

1.1.1 The traditional Bayesian paradigm

Due to their direct correspondence with the fundamental rules of probability, Bayesian posteriors $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ are desirable objects to be basing inference on. To see why, suppose the following three conditions hold true.

(P) The **P**rior $\pi(\boldsymbol{\theta})$ is correctly specified: It encodes the best available judgement about $\boldsymbol{\theta}$ based on *all* information available to the modeller. Crucially, the distribution $\boldsymbol{\pi}(\boldsymbol{\theta})$ is assumed to reflect this prior belief *exactly*. This implies that $\pi(\boldsymbol{\theta})$ should *completely* reflect all information available to the modeller

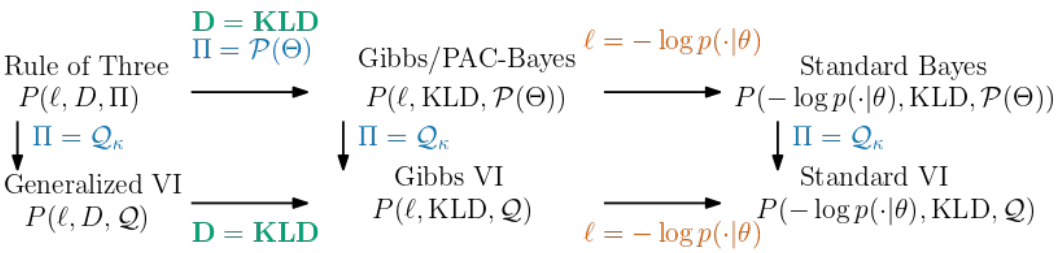


Figure 1.1: A taxonomy of some important belief distributions as special cases of the RoT.

such as previously observed observations $x_{-m:0}$ of the same phenomenon or domain expertise relating to the problem domain and the statistical model.

(L) There exists an (unknown but fixed) $\boldsymbol{\theta}^*$ making the **L**ikelihood model equivalent to the data generating mechanism of x_i. This is to say that $x_i \sim p(x_i|\boldsymbol{\theta}^*)$.[2]

(C) The budget for **C**omputation is infinite, so the complexity of computing the belief $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ can be ignored; and both the prior and the likelihood can be chosen without having to consider the implications their choices have on computational complexity.

If **(L)**, **(P)** and **(C)** are satisfied, Bayes' Rule immediately implies that the best belief about the best parameter value given the data $\{\boldsymbol{\theta}^* = \boldsymbol{\theta}\}|\{\boldsymbol{x}_{1:n} = x_{1:n}\}$ is given by

$$d\mathbb{P}\left(\boldsymbol{\theta}|x_{1:n}\right) \propto d\mathbb{P}\left(\boldsymbol{\theta}\right) \prod_{i=1}^{n} d\mathbb{P}\left(x_i|\boldsymbol{\theta}\right) = \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \propto q^*_{n,\mathrm{SB}}(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{1.2}$$

The crucial insight is that **(P)** and **(L)** lend a practically meaningful interpretation to Bayes' rule in form of conditional probability updates. Complementing this, **(C)** ensures that it is feasible to compute the often intractable resulting posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$. Accordingly, **(C)** generally is interpreted to mean that a Markov Chain Monte Carlo algorithm can be run for long enough to accurately represent $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$. In summary, if **(P)**, **(L)** and **(C)** hold, $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ is the only desirable posterior belief distribution. But how well does reality align with **(P)**, **(L)** and **(C)**? Turning attention to **(C)** first, most traditional scientific disciplines have little need to worry about computational complexity and will resort to sampling schemes for two reasons: First, the models are often relatively simple and thus straightforward to infer. Second—and even for more complicated models—the experimental setup as well as the cost of data collection typically far outweigh those of computation $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$. As

[2] We note here that to keep the presentation simpler, we are giving conditions that are stricter than what is required for Bayesian analysis. In particular, **(L)** corresponds to an objectivist treatment of the likelihood and can be weakened under the subjectivist paradigm for Bayesian analysis. In this paradigm, the treatment of the likelihood mirrors that of the prior: It now simply corresponds to the modeller's belief about the process that generated the data. While this first sounds like a weaker requirement, it ends up producing the same misspecification problems as **(L)**. Specifically, a subjectivist treatment of the likelihood requires the modeller to express her beliefs about the likelihood function *exactly*. This forces her to make more probability statements than she realistically has time or introspection for (see e.g. Goldstein, 1990; O'Hagan and Oakley, 2004; Goldstein, 2006). The result is that the likelihood function supplied by the modeller is *at best* going to be an approximate description of the modeller's beliefs. This provides the subjectivist interpretation of misspecification. Notice that it directly mirrors the objectivist interpretation of misspecification in **(L)**: The likelihood function supplied is *at best* going to be an approximate description of the true data generating mechanism.

for (P) and (L), neither prior nor likelihood are ever perfect reflections of one's full prior beliefs (see e.g. Goldstein, 1990; O'Hagan and Oakley, 2004; Goldstein, 2006) or the data generating mechanism (see e.g. Bernardo and Smith, 2009). In other words, (P) and (L) are invariably violated when interpreted literally. However and as enshrined in Box's aphorism that *all models are wrong, but some are useful*, this is not a problem so long as these violations are sufficiently small. In traditional statistics, ensuring that these violations are small has typically been enforced through a simple recursion (e.g. Box, 1980; Berger et al., 1994). Specifically, until you are confident that both (P) and (L) are close enough to the truth, repeat the following: Check if (L) or (P) are violated severely for the data you wish to analyse. If they are, choose a more appropriate likelihood and prior. In order to operationalize this iterative logic, batteries of descriptive statistics, tests and model selection criteria have been developed.

In summary then, ignoring the computational overhead and iteratively refining likelihoods and priors is rightfully the predominant inferential strategy for traditional scientific endeavours. Not only is domain expertise relevant for designing priors and likelihoods, but the process of finding an appropriate model often provides valuable insights in itself. Further, the expensive part of the analysis is typically data *collection*. Consequently, performing inference even with the most computationally expensive of sampling schemes is often not a major practical concern. In line with this, most methodological contributions in Bayesian statistical sciences rely to a substantial degree on (P), (L) and (C).

1.1.2 Machine Learning: A case study in the shortcomings of traditional Bayesian inference

Contemporary large-scale Machine Learning applications have frequently turned the traditional schematic of statistical model design upside down: Rather than carefully designing an appropriate likelihood model $p(\cdot|\boldsymbol{\theta})$ for a specific data domain, statistical Machine Learning research is typically characterized by the search of a flexible algorithm that can fit *any* data set $x_{1:n}$ well enough to produce useful inferences. The resulting likelihood models are typically not attempting to describe any data generating processes in the sense of (L). Rather, they are highly over-parameterized functions of $\boldsymbol{\theta}$ and typically un-identifiable, meaning that the parameter $\boldsymbol{\theta}^*$ that recovers the true data-generating mechanism is neither interpretable nor unique—if it exists at all. Such models have three major issues under the traditional paradigm of Bayesian inference that are readily identified:

(↯P) Invariably, the **P**rior is misspecified. Two factors compound this issue: Firstly, the large number of parameters over-parameterizing the likelihoods of many statistical Machine Learning models are no longer interpretable. This often prohibits domain experts from carrying out carefully guided prior elicitation. Secondly, priors are typically selected at least in part for their computational feasibility. This fundamentally alters the interpretation of the prior: Rather than the result of an attempt to capture the modeller's knowledge before observing the data, the prior takes the role of a more or less arbitrary reference measure whose primary function is ensure a form of smoothness or regularization. To make matters worse, the number of parameters is often large relative to n, which means that the priors have a disproportionate effect on inference—a problem we discuss in Example 1.1 in the context of Bayesian Neural Networks.

(↯L) Clearly, the **L**ikelihood is misspecified. This often has adverse side effects: While using an over-parameterized or off-the-shelf likelihood function can provide a good fit for the typical behaviour of the data, it will struggle with heterogeneous or untypical data points. We demonstrate this phenomenon on a black box model in Chapter 6.

(↯C) With increasingly complex statistical models, efficient computation has become only more important. Often, likelihood and prior choice are explicitly taken to facilitate computation—and so **(C)** has proven an increasingly infeasible description of reality. Accordingly, this problem has inspired numerous directions of research, including variational methods and Laplace approximations. To illustrate this, Example 1.2 goes over some of the research seeking to reduce computation time for the case of Gaussian Processes—a class of models that would be impossible to apply to large scale problems without the significant advances that have been made in reducing its computational complexity.

Under the challenges outlined in **(↯P)**, **(↯L)** and **(↯C)**, standard Bayesian posteriors often do not provide posterior belief distributions that are appropriate for downstream analysis and decision making. In the remainder, we will explain how and why this is the case for many parts of modern large-scale inference.

1.1.3 Prior misspecification

For most finite-dimensional parameters, even severely misspecified priors can often be harmless. For example, prior misspecification is typically no problem in the

asymptotic sense. Specifically, so long as (L) holds, it suffices that $\pi(\boldsymbol{\theta}^*) > 0$ for standard Bayesian posteriors to contract around $\boldsymbol{\theta}^*$ at rate $O(n^{-1/2})$ (see e.g. Ghosal, 1998; Ghosal et al., 2000; Shen and Wasserman, 2001; Walker, 2004, and references therein).

Often, these results are used as an apology to neglect the role of prior specification. While it is reassuring that the sequence of standard Bayesian posteriors shrinks to the population-optimum as $n \to \infty$, this does not describe the real world: n is usually fixed and only a single posterior is computed. Further, it is possible to specify arbitrarily bad priors for any n fixed observations. This means that once one departs from assuming that (P) is at least approximately correct, the standard Bayesian posterior belief about $\boldsymbol{\theta}^*$ can be made arbitrarily inappropriate. In summary, prior specification is particularly precarious whenever (i) the parameter space is large relative to n or (ii) it is impossible to specify priors in a principled way. As we discuss in the next example, a model invariably affected by both problems is the Bayesian Neural Network (BNN).

Example 1.1 (Deep Bayesian models as violations of (P))**.** Bayesian Neural Networks (BNNs) (MacKay, 1996; Neal, 2012) combine Deep Learning with Bayesian uncertainty quantification. For the parameter vector $\boldsymbol{\theta}$ of network weights, let $F(\boldsymbol{\theta})$ be the function specified by a Neural Network. One way of thinking about BNNs is as an over-parameterized likelihood function with a large number of parameters $d = |\boldsymbol{\Theta}|$. This is to say that one believes that (at least approximately), $x_i \sim p(x_i|F(\boldsymbol{\theta}^*))$ for at least one $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$. For a prior $\pi(\boldsymbol{\theta})$ about $\boldsymbol{\theta}$, this means that BNNs seek to do inference on the posterior given by

$$q^*_{n,\mathrm{SB}}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} p(x_i|F(\boldsymbol{\theta})).$$

This approach is conceptually appealing: One circumvents most issues with (L) by making the likelihood function almost arbitrarily flexible, and also quantifies uncertainty in the usual Bayesian manner. While both observations are correct, they mask a severe practical issue with this approach: Specifying $\pi(\boldsymbol{\theta})$ in a principled way and in accordance with (P) is generally impossible.

There are two main reasons for this: Firstly, the vector $\boldsymbol{\theta}$ indexes a black box model and is not interpretable, making domain expertise useless for prior elicitation. Secondly, computational aspects are a major concern for BNNs, so that one typically is constrained to choose priors that factorize over $\boldsymbol{\theta}$. As a consequence, practitioners often resort to choosing "default priors" that do not even attempt to approximately satisfy (P). Specifically, one typically just picks $\pi(\boldsymbol{\theta}) = \prod_{j=1}^{d} \pi_j(\boldsymbol{\theta}_j)$, where $\pi_j(\boldsymbol{\theta}_j)$

is a standard normal distribution for all j. Choosing priors in this ad-hoc fashion violates the principles underlying classical Bayesian modelling and is especially problematic when n is small relative to d (so that the prior has relatively strong influence). At the same time, reliable uncertainty quantification is most important whenever n is small relative to d. In fact, this is a well-known issue and addressed in various contributions by up-weighting the likelihood (down-weighting the KLD term in the ELBO), see Zhang et al. (2018); Rossi et al. (2020, 2019); Sønderby et al. (2016).

We do not mean to suggest that it is impossible to specify meaningful or useful priors for BNNs. For example, Toussaint et al. (2006) uses the principles of transformation invariance and maximum entropy, Nalisnick et al. (2021) calibrates priors via their predictive distribution and a 'reference' model, and Matsubara et al. (2021b) focuses on the prior's impact on the prediction space (see also Gelman et al., 2017) and in particular its covariance structure to specify more principled priors. While these approaches are all conceptually elegant, they also are computationally cumbersome—thus compounding the issues outlined in (↯C). As a result, the fully factorized priors discussed above are the de-facto default choices for most applications of BNNs.

For completeness, we note that this thesis does not discuss uninformative and so-called objective priors (see, e.g. Jeffreys, 1961; Zellner, 1977; Bernardo, 1979; Berger and Bernardo, 1992; Jaynes, 2003; Berger, 2006). Such priors are constructed to be as uninformative as possible. In some ways, they would be a natural, principled alternative to ill-informed priors. Generally however, their construction results in improper prior densities that do not correspond to a finite measure and thus do not integrate to one. While this is not generally prohibitive, it would severely complicate further developments because most divergences are not well-defined for improper priors[3].

1.1.4 Likelihood Misspecification

While prior misspecification affects inference adversely, the issue for inferential practice is even more serious if (L) is violated: Whenever the likelihood model for x_i is severely misspecified, inference outcomes suffer dramatically. Moreover, not even the asymptotic regime offers a remedy: The adverse effects of misspecification persist as $n \to \infty$. The traditional approach to addressing this issue is straightforward: If the likelihood model $p(x_i|\boldsymbol{\theta})$ is misspecified, simply investigate why exactly it

[3] The KLD is the exception to this rule: As it depends on the log normalizer of $\pi(\boldsymbol{\theta})$ in an additive fashion, improper priors can still be admissible.

fits the data poorly. After residual analysis, intense study of descriptive statistics and consultation with domain experts, redesign it to arrive at a likelihood model $p'(x_i|\boldsymbol{\theta}')$, which provides a better fit to the data and (approximately) satisfies (L). In other words, the traditional view is that any problem with misspecification is really a problem with careless modelling.

As outlined in Section 1.1.2, this strategy is neither practiced nor feasible with contemporary large-scale models. The naive interpretation of likelihoods as corresponding to an appropriately good description of the true data generating process in the sense of (L) is thus wholly inappropriate. This is especially important as many large-scale models are mainly interested in capturing the *typical* behaviour of the data—rather than *fully* modelling every aspect of a population. While this may appear to be a minor point at first glance, it has serious consequences for inferential practice. To see why, suppose a population contains a small number of outlying observations, local heterogeneities or spiky noise. The naive interpretation of the likelihood as in (L) *assumes* that these untypical aspects are encoded in the likelihood function. Hence, if x_i is an outlier, the inference machinery of traditional statistics interprets this as a strong signal: Since the likelihood model is an approximately correct description of the data, the most informative observations are those that do *not* fit the model fitted to the rest of the data. This is why aberrant parts of the data will have a disproportional impact on inference outcomes—leading standard inference methods to break down (see also Jewson et al., 2018).

Moreover, the often-invoked intuition that a sufficiently flexible likelihood family (such as likelihoods parameterized by Neural Networks) will not suffer these problems is dangerously incorrect in at least two ways: firstly, increasing the dimension of the model space for a fixed number of observations amounts to placing more weight on the prior—and so amounts to merely shifting the problem from (L) into (P). Secondly, the symmetries and degeneracies of such likelihoods can be shown to induce generalization errors that increase with the number of observations n (see e.g. Watanabe, 2018, Example 19 and Remark 20).

1.1.5 Mismatch between theoretically required and available computational resources

As Theorem 1.1 shows, the Bayesian posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ is the result of optimizing over the infinite-dimensional space $\mathcal{P}(\boldsymbol{\Theta})$. Generally, this implies that the posterior itself also does not live in a finite-dimensional space. In fact, the only case in which $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ can be represented through a finite-dimensional parameter is when prior and likelihood are conjugate to one another—a fact independently established by

Koopman (1936), Pitman (1936), and Darmois (1935) and thus commonly referred to as Koopman-Pitman-Darmois Theorem. This means that inference with $q^*_{n,\text{SB}}(\boldsymbol{\theta})$ is generally a hard problem, which manifests itself through the need to deal with the posterior's intractable normalizing constant. To address this problem, Markov Chain Monte Carlo algorithms are typically used. Such algorithms produce an exact representation of $q^*_{n,\text{SB}}(\boldsymbol{\theta})$ if the chain runs indefinitely and collects infinitely many samples. In practice, collecting a finite number of samples from the chain yields can represent $q^*_{n,\text{SB}}(\boldsymbol{\theta})$ almost exactly whenever $d = |\boldsymbol{\Theta}|$ is not too large. For large enough d however, the number of samples required to make the approximation useful is often too large to make samplers computationally viable: For example, in the *best* case scenario, Random Walk Metropolis Hastings scales like $\mathcal{O}(d^2)$ (Roberts et al., 1997), the Metropolis-adjusted Langevin algorithm like $\mathcal{O}(d^{4/3})$ (Roberts and Rosenthal, 1998) and Hamiltonian Monte Carlo like $\mathcal{O}(d^{5/4})$ (Beskos et al., 2013). Note that these results assume independence and Gaussianity—so on more complex models, scaling rates are even worse.

Approximation strategies constitute an alternative way to avoid explicit computation of normalizing constants. These methods project $q^*_{n,\text{SB}}(\boldsymbol{\theta})$ into some parameterized subset $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$. Clearly then, they produce approximations $q^*_{\text{A}}(\boldsymbol{\theta})$ of high quality only if the set $\mathcal{Q}$ is chosen to be sufficiently large. In practice however, most posterior belief distributions $q^*_{\text{A}}(\boldsymbol{\theta})$ computed this way barely deserve to be called approximations of $q^*_{n,\text{SB}}(\boldsymbol{\theta})$. For example, consider the mean field normal variational family given by

$$\mathcal{Q}_{\text{MFN}} = \left\{ \prod_{j=1}^{d} \mathcal{N}(\boldsymbol{\theta}_j | \mu_j, \sigma_j^2) : \mu_j \in \mathbb{R}, \sigma_j^2 \in \mathbb{R}_{>0} \text{ for all } j \right\}. \tag{1.3}$$

For most interesting non-trivial posterior distributions $q^*_{n,\text{SB}}(\boldsymbol{\theta})$, there will not exist an element $q \in \mathcal{Q}_{\text{MFN}}$ that could be considered a meaningful approximation to $q^*_{n,\text{SB}}(\boldsymbol{\theta})$. This is perhaps unsurprising: After all, $\mathcal{Q}_{\text{MFN}}$ assumes $\mathcal{O}(d^2)$ independence relationships in the approximate posterior belief for $\boldsymbol{\theta}$. Worse still: As approximations are particularly attractive when $|\boldsymbol{\Theta}| = d$ is large, in practice we will resort to such insufficiently expressive "approximations" to $q^*_{n,\text{SB}}(\boldsymbol{\theta})$ *precisely when* the elements in $\mathcal{Q}_{\text{MFN}}$ are structurally most dissimilar from $q^*_{n,\text{SB}}(\boldsymbol{\theta})$. To improve the quality of these approximations, numerous directions of research have proposed ever more flexible variational families in order to make $\mathcal{Q}$ more expressive. Examples include implicit distributions (e.g. Tran et al., 2017; Tiao et al., 2018; Shi et al., 2018; Ma et al., 2019), normalizing flows (e.g. Rezende and Mohamed, 2015), or the

variational Gaussian Process (Tran et al., 2016). Once again, there is no free lunch: more expressive families $\mathcal{Q}$ will incur higher computational cost and compound the issues with (C).

In this thesis, we advocate an optimization-centric view of posterior belief computation. As a side-product of this view, we believe that it is often unhelpful to think of $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ as an approximation to $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$. Rather, we prefer to think of $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ as defining a new and distinct posterior belief distribution in its own right—which happens to also be an approximation to $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ if $\mathcal{Q}$ is sufficiently expressive.

To highlight the importance that computational considerations have played in research on Bayesian Machine Learning, we end their discussion by pointing to some of the recent research on Bayesian computation for Gaussian Processes.

Example 1.2 (large-scale Gaussian processes as violations of (C))**.** Many Bayesian Machine Learning models prohibit exact computation. One particularly interesting case are Gaussian Process (GP) models: Even in the special cases where they admit closed form posteriors, it may well be impossible to compute them exactly for sufficiently large inference problems. The reason is that for n observations, direct computation of the associated GP posterior takes $\mathcal{O}(n^3)$ time. As a consequence, an entire literature is dedicated to bringing down this computational complexity (see for instance Williams and Seeger, 2001; Quiñonero Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006; Titsias) and developing software or computer-architecture specific methods geared towards inference with GPs (e.g. Matthews et al., 2017; Gardner et al., 2018; Balandat et al., 2020; Wang et al., 2019b). Furthermore, with deep (i.e., hierarchical) approaches to GPs introduced in Damianou and Lawrence (2013) and extended in various directions (e.g. Dai et al., 2016; Hegde et al., 2019), this challenge has only become more important (see e.g. Bui et al., 2016; Cutajar et al., 2017b; Salimbeni and Deisenroth, 2017).

1.2 Existing modifications of Bayesian inference

The last section explained why standard Bayesian posteriors may be an inappropriate tool for performing inference in the context of Machine Learning. Before further discussing how this can be alleviated by the RoT, we first briefly review and discuss some of the most important previous extensions and generalisations of Bayes' Rule depicted in Figure 1.1. Most of these generalizations are motivated by similar observations as those made in the previous sections, and aimed at fixing specific shortcomings of a standard Bayesian approach to statistical inference. This will be important to motivate further developments, but also to contrast the added benefit

that the RoT can bring relative to existing methodologies. Lastly, it will also help the reader to better grasp the most important ideas, techniques, and applications of generalized Bayesian methods.

1.2.1 Probably Approximately Correct (PAC) Bayesian methods

While PAC-Bayesian results often have intimate links with traditional Bayesian inference (see e.g. Germain et al., 2016; Grünwald and Van Ommen, 2017), their motivations and origins are rather distinct (see e.g. Shawe Taylor and Williamson, 1997; Guedj, 2019): Unlike Bayesian inference, PAC-Bayesian results are not constructed based on assuming a correct likelihood or prior to be available and in this sense circumnavigate both (L) and (P). In fact, they do not rely on likelihoods at all and—much like the remainder of this thesis—treat the negative log likelihood as just one choice of loss (amongst many possible). The aim of such PAC-Bayesian bounds is in their name: they seek to derive generalization bounds for belief distributions $q(\boldsymbol{\theta}) \in \mathcal{P}(\boldsymbol{\Theta})$ defined over some hypothesis space $\boldsymbol{\Theta}$ relative to a loss function ℓ. For example, under a prior belief $\pi(\boldsymbol{\theta})$, a loss ℓ and a data generating mechanism for $x_{1:n}$ satisfying appropriate regularity conditions and for any $q(\boldsymbol{\theta}) \in \mathcal{P}(\boldsymbol{\Theta})$ as well as for any fixed value of $\varepsilon > 0$, McAllester's seminal bound (McAllester, 1999a,b) says that with probability at least $1 - \varepsilon$,

$$\mathbb{E}_{q(\boldsymbol{\theta})}\left[\mathbb{E}_{\boldsymbol{x}_{1:n}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, \boldsymbol{x}_i)\right]\right] \leq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i)\right] + \sqrt{\frac{\mathrm{KLD}(q,\pi) + \log\frac{2\sqrt{n}}{\varepsilon}}{2n}}. \quad (1.4)$$

Minimizing the right hand side of this bound with respect to $q(\boldsymbol{\theta})$ over some set $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$ immediately recovers a special case for the RoT given by $P(\ell, D_{\mathrm{McAllester}}, \Pi)$. Here, $D_{\mathrm{McAllester}}$ is just the last term of the above bound, with a subtracted constant and rescaled by n:

$$D_{\mathrm{McAllester}}(q\|\pi) = \sqrt{n} \cdot \left(\sqrt{\frac{\mathrm{KLD}(q,\pi) + \log\frac{2\sqrt{n}}{\varepsilon}}{2}} - \sqrt{\frac{\log\frac{2\sqrt{n}}{\varepsilon}}{2}}\right).$$

Subtraction of the constant ensures that $D_{\mathrm{McAllester}}(q\|\pi) = 0$ if and only if $\pi = q$. The rescaling is necessary as we have to multiply both sides of eq. (1.4) by n to bring them into the RoT form. Note that neither the addition of the constant nor the rescaling affects the minimizer.

A similar logic can be applied to virtually all PAC-Bayesian bounds, crucially also for bounds based on divergences other than the KLD such as those of Bégin

et al. (2016), Alquier and Guedj (2018), or Ohnishi and Honorio (2021).[4] In light of this, PAC-Bayesian analysis may prove crucial in deciding which divergence should be used for inference in a given problem: The bounds of Bégin et al. (2016), Alquier and Guedj (2018), and Ohnishi and Honorio (2021) all depend on divergences other than the KLD, and provide generalization guarantees for less restrictive settings than the KLD. For example, the bounds of Alquier and Guedj (2018) depend on f-divergences, and provide generalization guarantees even if the observation sequence exhibits a substantial degree of heterogeneity or temporal dependence. Similarly, unlike bounds based on the KLD, the bounds of Ohnishi and Honorio (2021) provide generalization guarantees even if ℓ is an unbounded loss function.

While PAC-Bayesian bounds appear quite similar to the posterior belief distributions computed via the generalized posteriors proposed in this thesis, there are a number of important differences. Firstly, PAC-Bayes bounds are mainly a theoretical device. Unlike much of the methodology developed as part of this thesis, their main interest is typically in establishing learning rates and theoretical guarantees—often even for algorithms that themselves do not use a prior distribution as input. Secondly and as a consequence of this, the choice of priors in PAC-Bayesian learning is often geared towards optimising an error bound. This is completely different from how priors would be approached in traditional Bayesian inference: rather than viewing the prior as a source of information, PAC-Bayesian bounds typically treat them as nuisance parameters that ought to be minimized over (see for instance the distribution-dependent priors derived by Lever et al. (2013)). Thirdly and on a related note, the literature on PAC-Bayes often takes no interest in actually computing the PAC-Bayesian posterior. Rather, its role is purely conceptual: by plugging it into a PAC-Bayesian bound, one obtains a generalization guarantee or error rate with desirable properties. Lastly, there is a strong trade-off in PAC-Bayesian bounds between the strength of the theoretical guarantee on the one hand, and the class of loss functions for which this guarantee holds on the other hand. In fact, most PAC-Bayes bounds of practical interest assume that the loss ℓ is bounded both from above and below, which immediately precludes us from using $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ for any probability density $p(x_i|\boldsymbol{\theta})$ for which we can find a sequence of parameters $\{\boldsymbol{\theta}_n\}_{n\in\mathbb{N}}$ so that $p(x_i|\boldsymbol{\theta}_n) \to \infty$ as $n \to \infty$. While such sequences typically do not exist if the data are discretely-valued, they can be found for most probability densities on Euclidean spaces, such as normal distributions. As a consequence, the practicality of PAC-Bayesian bounds is limited: in particular, they are far more practically relevant for classification problems than they are for regression problems. For a more

[4] For more classical bounds based on $D = \mathrm{KLD} \neq D_{\mathrm{McAllester}}$, see Catoni (2007); Zhang (2006).

thorough review into the principles and limitations of PAC-Bayesian inference as well as their relationship with traditional Bayesian methodology, see Guedj (2019).

1.2.2 Gibbs posteriors & general Bayesian updating

For a loss function $L : \boldsymbol{\Theta} \times \mathcal{X}^n \to \mathbb{R}$ and a prior π on $\boldsymbol{\Theta}$ for which it can be shown that $\int_{\boldsymbol{\Theta}} \exp\{-L(\boldsymbol{\theta}, x_{1:n})\} d\pi(\boldsymbol{\theta}) < \infty$, the corresponding Gibbs posterior (often also called **g**eneralised **B**ayes posterior or pseudo posterior) is defined as

$$q^*_{n,\mathrm{GB}}(\boldsymbol{\theta}) = \frac{\exp\{-L(\boldsymbol{\theta}, x_{1:n})\}\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \exp\{-L(\boldsymbol{\theta}, x_{1:n})\} d\pi(\boldsymbol{\theta})}. \tag{1.5}$$

Gibbs posteriors can be recovered as a special case of the RoT, as the following extension of Theorem 1.1 shows.

Proposition 1.1. If $\int_{\boldsymbol{\Theta}} \exp\{-L(\boldsymbol{\theta}, x_{1:n})\} d\pi(\boldsymbol{\theta}) < \infty$, then it holds that $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta}) = P(\ell, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$.

Proof. Note that the proof of Theorem 1.1 did not depend on the choice of L, so the same arguments as before yield the result. □

Belief distributions like $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$ and their relationship to the **s**tandard **B**ayes posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ have been widely studied for particular special cases of L. This includes so-called power posteriors which raise the likelihood to a power w so that $L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^n -\log p(x_i|\boldsymbol{\theta})^w$ (see e.g. Grünwald, 2011, 2012; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019), so-called pseudo likelihood or composite likelihood approaches (e.g. Varin et al., 2011; Pauli et al., 2011; Ribatet et al., 2012), as well as divergence- and disparity-based Bayesian methods that take $L(\boldsymbol{\theta}, x_{1:n}) \approx D(p(\cdot|\boldsymbol{\theta}), p_{\mathrm{empirical}}(\cdot))$ as some (approximate) statistical discrepancy measure D between a likelihood model $p(\cdot|\boldsymbol{\theta})$ and the empirical data distribution $p_{\mathrm{empirical}}(\cdot)$ (e.g., Hooker and Vidyashankar, 2014; Ghosh and Basu, 2016; Futoshi Futami et al., 2018; Jewson et al., 2018; Chérief Abdellatif and Alquier, 2020; Matsubara et al., 2021a).

Perhaps the most important contribution in this space in recent years has been the idea of generalized Bayesian updates in Bissiri et al. (2016), which advocates for Gibbs posteriors with arbitrary additive losses of the form $L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$. Much as what this thesis outlined in (↯L), a key motivation for generalized Bayes updates is the unavailability of correctly specified models in many situations. To depart from the assumption of correct model specification, Bissiri et al. (2016) poses that all one needs for belief updates is a way to link data x_i to

a parameter of interest $\boldsymbol{\theta}$ via some loss function ℓ, and a prior belief about 'good' values of $\boldsymbol{\theta}$. In particular, the paper shows that the only belief update rule satisfying coherence will be the generalized Bayesian belief update that leads to $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$. In a nutshell, coherence as defined in Bissiri et al. (2016) says that posteriors have to be generated according to some function $\psi : \mathbb{R}^2 \to \mathbb{R}$ which for the prior $\pi(\boldsymbol{\theta})$ and loss terms $\ell(\boldsymbol{\theta}, x_1), \ell(\boldsymbol{\theta}, x_2)$ behaves as

$$\psi\left(\ell(\boldsymbol{\theta}, x_2), \psi\left(\ell(\boldsymbol{\theta}, x_1), \pi(\boldsymbol{\theta})\right)\right) = \psi\left(\ell(\boldsymbol{\theta}, x_1) + \ell(\boldsymbol{\theta}, x_2), \pi(\boldsymbol{\theta})\right).$$

This requirement effectively enforces a multiplicative update via exponential additivity as in (1.5). Put differently and in terms of (1.1), this requirement enforces that $D = \mathrm{KLD}$ and $\Pi = \mathcal{P}(\boldsymbol{\Theta})$.

Imposing the requirement of coherence upon generalized posterior beliefs has two main disadvantages: firstly, coherence is only well-defined and achievable for additive losses. For this reason, it excludes many successful methodologies for generalized posteriors derived from non-additive losses such as those based on V-statistics or U-statistics (see for instance Hooker and Vidyashankar (2014), Chérief Abdellatif and Alquier (2020), or Matsubara et al. (2021a)). Further, the idea of coherence crucially relies on trusting the prior: it treats the prior belief as perfectly and exactly encapsulating our full knowledge before we see any data; and therefore being a quantity that is worth updating. Note that the premise for generalized Bayesian updating is model misspecification, so the modeller is already in a setting with limited information. This makes it likely that the prior is in fact *not* a perfect encapsulation of our full knowledge. This is crucial, since if the prior is poorly specified as outlined in **(↯P)**, coherence will in fact be an undesirable design choice for a posterior belief. For these reasons, the RoT does not impose coherence.

1.2.3 Variational approximations to Bayes posteriors

While the logic of multiplicative updates inherent in Bayes' rule and (1.5) is conceptually elegant, the intractable normalization constants of $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ and $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$ often make these approaches infeasible in practice. Specifically, exact computation of these posteriors is often only possible through sampling methods; and thus using a posterior of this form typically incurs a large computational burden. To alleviate this problem, many approximate Bayesian inference schemes have been proposed. Their principal idea is to force the posterior belief into some parametric form. Specifically, one seeks to approximate $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta}) \approx q^*_{\mathrm{A}}(\boldsymbol{\theta})$ (or $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta}) \approx q^*_{\mathrm{A}}(\boldsymbol{\theta})$), where $q^*_{\mathrm{A}}(\boldsymbol{\theta}) \in \mathcal{Q}$

and

$$\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in \boldsymbol{K}\} \tag{1.6}$$

is a family of distributions on $\boldsymbol{\Theta}$ parametrized by $\boldsymbol{\kappa}$. This significantly reduces the computational burden, because it transforms the problem of inference into a finite-dimensional optimization problem.

The literature on such approximations is extensive and has diverse origins. Their development arguably started with Laplace Approximations (see e.g. the seminal papers of Tierney and Kadane, 1986; Shun and McCullagh, 1995; MacKay, 1998), which have recently been refined into Integrated Nested Laplace Approximations (Rue et al., 2009). A second family of approximation methods known as Expectation Propagation (Opper and Winther, 2000; Minka, 2001) was motivated through factor graphs and message passing (Minka, 2005). The third and arguably most successful approach originated by connecting the Expectation-Maximization algorithm (Dempster et al., 1977) and the variational free energy from statistical physics (Neal and Hinton, 1998), culminating in Variational Inference (VI) (Jordan et al., 1999; Beal, 2003). For these methods, $\mathcal{Q}$ is called the *variational family*.

Two main interpretations of VI prevail. Firstly, one may derive its objective function as an Evidence Lower Bound (ELBO). Secondly, one can show that VI minimizes the KLD between $\mathcal{Q}$ and $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ (or $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$).

VI from an Evidence Lower Bound (ELBO)

One context in which VI was originally derived is the task of model selection. In Bayesian model selection, the integral $p(x_{1:n}) = \int_{\boldsymbol{\Theta}} \exp\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$—called *evidence* or *marginal likelihood* whenever $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ for some likelihood model $p(\cdot|\boldsymbol{\theta})$—takes centre stage. Roughly speaking, one selects the model for which this integral takes the largest value. But since $p(x_{1:n})$ is generally intractable, one finds an approximation to it. In particular, one notes that for any $q(\boldsymbol{\theta}) \in \mathcal{Q}$,

$$\begin{aligned}
\log p(x_{1:n}) &= \log\left(\int_{\boldsymbol{\Theta}} \exp\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\right) \\
&= \log\left(\int_{\boldsymbol{\Theta}} \exp\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\}\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}q(\boldsymbol{\theta})d\boldsymbol{\theta}\right) \\
&\overset{\text{Jensen's IE}}{\geq} \int_{\boldsymbol{\Theta}} \log\left(\exp\left\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right\}\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)q(\boldsymbol{\theta})d\boldsymbol{\theta}.
\end{aligned} \tag{1.7}$$

If the loss function is $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ for some likelihood model $p(\cdot|\boldsymbol{\theta})$, then the right hand side of eq. (1.7) is called the Evidence Lower Bound (ELBO). Rewriting the integral, one now obtains the VI posterior as

$$q^*_{\text{VI}}(\boldsymbol{\theta}) = P(\ell, \text{KLD}, \mathcal{Q}) \stackrel{\text{def}}{=} \arg\min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right] + \text{KLD}\left(q||\pi\right) \right\}, \quad (1.8)$$

where $q^*_{\text{VI}}(\boldsymbol{\theta}) = q(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)$ for some optimal parameter $\boldsymbol{\kappa}^* \in \boldsymbol{K}$. Note that here, we have directly defined $q^*_{\text{VI}}(\boldsymbol{\theta})$ relative to an arbitrary additive loss ℓ rather than relative to the negative log likelihood. Consequently, $q^*_{\text{VI}}(\boldsymbol{\theta})$ is defined as an approximation to arbitrary generalized/Gibbs/pseudo Bayes posteriors $q^*_{n,\text{GB}}(\boldsymbol{\theta})$, and we will use this definition in the remainder of the thesis.

Taking inspiration from the interpretation of $q^*_{\text{VI}}(\boldsymbol{\theta})$ as minimizing a lower bound on the evidence, alternative approximations target generalized Evidence Lower Bounds (e.g. Chen et al., 2018; Domke and Sheldon, 2018; Burda et al., 2016). For a given bound $\log p(x_{1:n}) \geq \text{G-ELBO}(q)$, such methods produce posteriors as

$$q^*_{\text{G-ELBO}}(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \left\{ -\text{G-ELBO}(q) \right\}.$$

Multi-sample bounds (see e.g. Burda et al., 2016) are a particularly prominent example. As the name implies, these bounds interpret the ELBO term given in eq. (1.8) by

$$\text{ELBO}(q) = \mathbb{E}_{\boldsymbol{\theta} \sim q(\boldsymbol{\theta})} \left[\log \left(\frac{\exp\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)\pi(\boldsymbol{\theta})\}}{q(\boldsymbol{\theta})} \right) \right]$$

as a bound constructed from a single sample of $\boldsymbol{\theta}$ and replace the objective with its K-sample version obtained by

$$\text{MS-ELBO}(q, K) = \mathbb{E}_{\boldsymbol{\theta}_{1:K} \sim \prod_{j=1}^{K} q(\boldsymbol{\theta}_j)} \left[\log \frac{1}{K} \sum_{j=1}^{K} \left(\frac{\exp\{-\sum_{i=1}^{n} \ell(\boldsymbol{\theta}_j, x_i)\pi(\boldsymbol{\theta})\}}{q(\boldsymbol{\theta}_j)} \right) \right].$$

The rationale for doing so is that $\text{MS-ELBO}(q, 1) = \text{ELBO}(q)$, and that the resulting bound on the (generalized) evidence is tighter than the standard ELBO. More precisely, for any $K \in \mathbb{N}$, $\log p(x_{1:n}) \geq \text{MS-ELBO}(q, K+1) \geq \text{MS-ELBO}(q, K) \geq \text{MS-ELBO}(q, 1) = \text{ELBO}(q)$.

VI as KLD-minimization and as Discrepancy-based VI (DVI)

A second well-known perspective on standard VI posteriors is motivated by rewriting the objective in eq. (1.8) in terms of the Kullback-Leibler Divergence (KLD) as follows:

$$q^*_{\text{VI}}(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \left\{ \text{KLD}\left(q(\boldsymbol{\theta}) \middle\| q^*_{n,\text{GB}}(\boldsymbol{\theta}) \right) \right\}$$

In words, the above shows that standard VI finds $q^*_{\text{VI}}(\boldsymbol{\theta}) \in \mathcal{Q}$ closest to $q^*_{n,\text{GB}}(\boldsymbol{\theta})$ in the KLD-sense. The relevant algebraic arguments are simple, but worth stating formally to clarify the crucial role of the logarithm in this equivalence.

Proposition 1.2. If $\int_{\boldsymbol{\Theta}} \exp\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\}\, d\pi(\boldsymbol{\theta}) < \infty$, then it holds that $P(\ell, \text{KLD}, \mathcal{Q}) = \arg\min_{q \in \mathcal{Q}} \left\{ \text{KLD}\left(q(\boldsymbol{\theta}) \middle\| q^*_{n,\text{GB}}(\boldsymbol{\theta}) \right) \right\}$.

Proof. The arguments are once again the same as for the proof of Theorem 1.1: taking $Z = \int_{\boldsymbol{\Theta}} \exp\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\}\, d\pi(\boldsymbol{\theta})$, it holds that

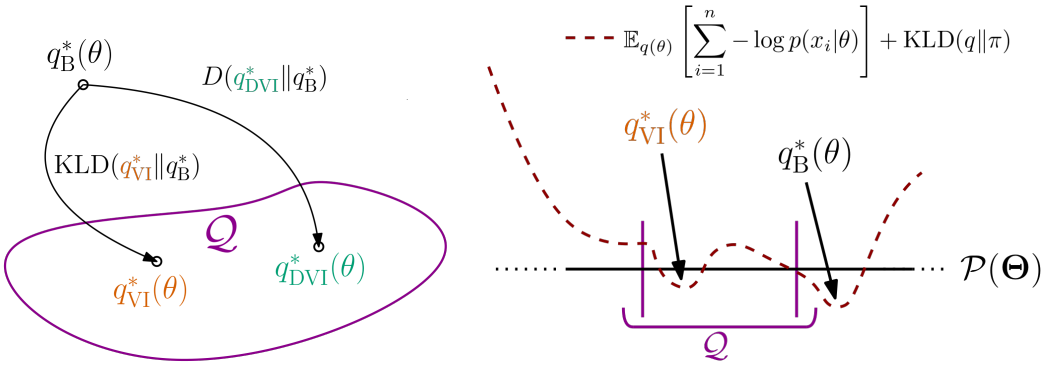
$$\begin{aligned}
P(\ell, \text{KLD}, \mathcal{Q}) &= \arg\min_{q \in \mathcal{Q}} \left\{ \int_{\boldsymbol{\Theta}} \left[\log\left(\exp\left\{ \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right\} \right) + \log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right) \right] q(\boldsymbol{\theta}) d\boldsymbol{\theta} \right\} \\
&= \arg\min_{q \in \mathcal{Q}} \left\{ \int_{\boldsymbol{\Theta}} \log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) \exp\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\}} \right) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \right\}. \\
&= \arg\min_{q \in \mathcal{Q}} \left\{ \int_{\boldsymbol{\Theta}} \log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}) \exp\{-\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\} Z^{-1}} \right) q(\boldsymbol{\theta}) d\boldsymbol{\theta} - \log Z \right\} \\
&= \arg\min_{q \in \mathcal{Q}} \left\{ \text{KLD}\left(q(\boldsymbol{\theta}) \middle\| \pi(\boldsymbol{\theta}) \exp\left\{ -\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right\} Z^{-1} \right) \right\},
\end{aligned}$$

so that the result follows. □

This insight has produced a growing literature seeking to minimize (local or global) discrepancies D between $\mathcal{Q}$ and $q^*_{n,\text{GB}}(\boldsymbol{\theta})$ different from the KLD (e.g. Minka, 2001; Opper and Winther, 2000; Li and Turner, 2016; Dieng et al., 2017; Hernández Lobato et al., 2016; Yang et al., 2019; Cichocki and Amari, 2010; Ranganath et al., 2016; Wang et al., 2018; Saha et al., 2019). For a disrepancy measure $D : \mathcal{P}(\boldsymbol{\Theta}) \times \mathcal{P}(\boldsymbol{\Theta}) \to \mathbb{R}$, these methods compute

$$q^*_{\text{DVI}}(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \left\{ D\left(q(\boldsymbol{\theta}) \middle\| q^*_{n,\text{GB}}(\boldsymbol{\theta}) \right) \right\}.$$

In the remainder, we will call such procedures **Discrepancy Variational Infer-**



(a) **DVI** interpretation of **VI** (b) Interpretation of **VI** as in Theorem 1.3

Figure 1.2: Best viewed in color. Depicted is a schematic to clarify the conceptual distinction between two interpretations of **VI**. **DVI** methods interpret **VI** as the KLD-projection of $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$ into the variational family $\mathcal{Q}$. New methods are then derived by replacing the KLD with alternative projection operators. Alternatively, **VI** posteriors can also be seen as the best solution to a constrained optimization problem: specifically, rather than finding the global optimum $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$ of the optimization problem associated to $P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, **VI** finds $P(L, \mathrm{KLD}, \mathcal{Q})$, which is simply the $\mathcal{Q}$-constrained solution in the subset $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$.

ence (DVI) methods whenever $D \neq \mathrm{KLD}$.[5] We graphically summarize the interpretation of DVI in Figure 1.2a. Note that DVI methods do not fall into our RoT framework: they are generally not recoverable as $\mathcal{P}(L, D, \mathcal{Q})$ for any choice of L, D, $\mathcal{Q}$.

VI as constrained optimization

Because it will prove convenient for the remainder of the thesis, we introduce another interpretation of VI that is obvious once seen. Surprisingly, this interpretation had not been formally presented before work conducted as part of this thesis. To set the stage for this interpretation, we first extend Theorem 1.1 in an obvious way. This is done simply to emphasize that the proceeding discussion applies to both $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ and $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$. For convenience, we will state everything in terms of $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$, since it recovers $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ as a special case.

Corollary 1.1. Take $L(\boldsymbol{\theta}, x_{1:n}) = -\log p(x_{1:n}|\boldsymbol{\theta})$, $D = \mathrm{KLD}$, $\Pi = \mathcal{P}(\boldsymbol{\Theta})$. If $Z = \int_{\boldsymbol{\Theta}} \exp\{-L(\boldsymbol{\theta}, x_{1:n})\}\, \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ and $0 < Z < \infty$, then $P(L, D, \Pi) = q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$.

Proof. The proof is the same as for Theorem 1.1. □

[5] Whenever $D = \mathrm{KLD}$, we will refer to these methods just as standard Variational Inference (VI).

The implication of this Corollary is that

$$q^*_{n,\mathrm{GB}}(\boldsymbol{\theta}) \overset{\text{def}}{=} \underset{q \in \mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right] + \mathrm{KLD} \left(q || \pi \right) \right\}. \tag{1.9}$$

Comparing this to (1.8), it immediately becomes clear that $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ is the $\mathcal{Q}$-constrained counterpart to $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$. This immediately implies that the following trivial result:

Proposition 1.3 (Optimality of standard VI)**.** Relative to the objective associated with $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta}) = P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, the variational posterior $q^*_{\mathrm{VI}}(\boldsymbol{\theta}) = P(L, \mathrm{KLD}, \mathcal{Q})$ is the optimal solution in $\mathcal{Q}$.

Proof. This essentially follows by definition. First, notice that the VI posterior belief distribution $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ and the Bayesian posterior belief distribution $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$ both seek to minimize

$$\mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right] + \mathrm{KLD}(q \| \pi)$$

over $q(\boldsymbol{\theta})$. Second, notice that $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ is the minimizer of this objective relative to $\mathcal{Q}$ while $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$ is the minimizer relative to $\mathcal{P}(\boldsymbol{\Theta})$. Third, note that $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$. □

This provides another meaningful interpretation of $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ depicted in Figure 1.2b. Specifically, the result endows standard VI with a special property: in an optimization-centric view on Bayesian inference, we should prefer $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ to all other possible approximations within $\mathcal{Q}$ *provided* we believe that the optimization objective defining the Bayesian posterior is appropriate for the problem at hand.

In this sense, the result also implies the sub-optimality of alternative approximation methods within the same variational family $\mathcal{Q}$.

Corollary 1.2 (Suboptimality of alternative methods)**.** Relative to the objective associated with $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta}) = P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, any approximation method that produces some $q^* \in \mathcal{Q}$ so that $q^* \neq P(L, \mathrm{KLD}, \mathcal{Q})$ produces sub-optimal posterior beliefs.

Proof. This follows immediately from Proposition 1.3, but for pedagogical reasons we give another proof by contradiction. Suppose we are given a posterior belief $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ that could not have alternatively been produced by standard VI. First, by definition of standard VI, it holds that that for *any* sequence of observations $x_{1:n}$

and for all n,

$$\mathbb{E}_{q^*_{\text{VI}}(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},x_i)\right]+\text{KLD}\left(q^*_{\text{VI}}||\pi\right)\leq\mathbb{E}_{q^*_{\text{A}}(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},x_i)\right]+\text{KLD}\left(q^*_{\text{A}}||\pi\right).$$

Since we also assumed that $q^*_{\text{A}}(\boldsymbol{\theta})$ could not have alternatively been produced by standard VI, it also holds that the inequality is strict, i.e.

$$\mathbb{E}_{q^*_{\text{VI}}(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},x_i)\right]+\text{KLD}\left(q^*_{\text{VI}}||\pi\right)<\mathbb{E}_{q^*_{\text{A}}(\boldsymbol{\theta})}\left[\sum_{i=1}^{n}\ell(\boldsymbol{\theta},x_i)\right]+\text{KLD}\left(q^*_{\text{A}}||\pi\right).$$

This yields the desired result. □

Corollary 1.2 says that for a fixed variational family $\mathcal{Q}$, any alternative approximation $q^*_{\text{A}}(\boldsymbol{\theta})\in\mathcal{Q}$ that is not equal to $q^*_{\text{VI}}(\boldsymbol{\theta})$ will be sub-optimal under an optimization-centric view on Bayesian inference. This concerns a host of methods, including generalized evidence lower bound formulations, alternative Discrepancy Variational Inference (DVI) methods or Expectation Propagation (EP) approaches.

Importantly, the result does *not* imply that these alternative posterior approximations will perform worse than VI in practice. In fact, from an optimization-centric standpoint it is quite clear why such alternative approximations can deliver empirical success: if $q^*_{\text{A}}(\boldsymbol{\theta})$ performs better than the standard variational approximation $q^*_{\text{VI}}(\boldsymbol{\theta})$, the objective underlying $q^*_{\text{A}}(\boldsymbol{\theta})$ must implicitly be targeting a more appropriate posterior belief for the problem at hand—an observation that partially motivates some later developments of this thesis.

1.2.4 Links with Information Theory

One can also draw a close connection between the RoT and another latent variable model: the Predictive Information Bottleneck (PIB) (see Tishby et al., 2000; Bialek et al., 2001). Given a data generating process ϕ so that $\boldsymbol{x}_{1:\infty}\sim\phi$ and a compressed representation $\boldsymbol{\theta}$ of the random variables $\boldsymbol{x}_{1:n}$, the PIB poses the following optimization problem:

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})=\underset{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})\in\Pi_{\text{PIB}}}{\arg\min}\;\{-I(\boldsymbol{\theta},\boldsymbol{x}_{n+1:\infty})\}\quad\text{s.t. }I(\boldsymbol{\theta},\boldsymbol{x}_{1:n})\leq I_0,\tag{1.10}$$

where all random variables admit densities p with respect to the Lebesgue measure,

$$I(\boldsymbol{Z},\boldsymbol{Y})=\text{KLD}\left(p(\boldsymbol{Z},\boldsymbol{Y})\|p(\boldsymbol{Z})p(\boldsymbol{Y})\right)$$

denotes the mutual information between random variables $\boldsymbol{Z}$ and $\boldsymbol{Y}$, and

$$\Pi_{\text{PIB}} = \left\{ q \in \mathcal{P}(\boldsymbol{\Theta}|\mathcal{X}^n) : \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) p(\boldsymbol{x}_{1:n}) d\boldsymbol{\theta} = p(\boldsymbol{x}_{1:n}) \right\}$$

is the set of admissible conditional distributions. This shows that the PIB maximizes the mutual information $I(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty})$ between the future $\boldsymbol{x}_{n+1:\infty}$ and the compression (i.e. model) $\boldsymbol{\theta}$ subject to an upper bound I_0 on the mutual information $I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n})$ between said model and the distribution of the training data $\boldsymbol{x}_{1:n}$. The PIB owes its name to the requirement that $I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) \leq I_0$: in words, this bound prevents the compression from being arbitrarily expressive and forces us to squeeze the information contained in $\boldsymbol{x}_{1:n}$ through a bottleneck.

This PIB form is generally hard to solve, but can be rewritten as a RoT-like objective

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \arg\min_{q \in \Pi_{\text{PIB}}} \left\{ \mathbb{E}_q \left[L_{n,\text{PIB}}(q) \right] + D_{\text{PIB}}(q \| \pi_{\text{PIB}}) \right\}.$$

The nature of $L_{n,\text{PIB}}$ and D_{PIB} as well as mathematical details for arriving at this form are deferred to Appendix B.1. While the structure of the problem closely resembles that of the RoT, there are some important differences. Most important among them, the PIB relates to the full distributional characterizations of the random variables $\boldsymbol{x}_{1:n}$ via $p(\boldsymbol{x}_{1:n})$—rather than to any actual observations $x_{1:n}$. As a consequence, the space of feasible solutions Π_{PIB} contains *all* possible conditional distributions $\{q^*(\boldsymbol{\theta}|x_{1:n})\}_{x_{1:n} \in \mathcal{X}^n}$—rather than a single conditional distribution $q^*(\boldsymbol{\theta}|x_{1:n})$ depending on a single realization $x_{1:n}$ of $\boldsymbol{x}_{1:n}$ only.

As shown in Alemi (2019) however, the PIB can also be variationally lower-bounded and approximated with observations $x_{1:n}$ to arrive at the usual data-dependent form of the RoT. Specifically, if we are willing to assume that $\boldsymbol{x}_{1:n}$ are independent, then we may rewrite $p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}) = \prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})$. This allows the coarse approximation $\mathbb{E}_{p(\boldsymbol{x}_{1:n})} \left[\log p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}) \right] = \mathbb{E}_{p(\boldsymbol{x}_{1:n})} \left[\sum_{i=1}^n \log p(\boldsymbol{x}_i|\boldsymbol{\theta}) \right] \approx \sum_{i=1}^n \log p(x_i|\boldsymbol{\theta})$, which replaces dependence on $p(\boldsymbol{x}_{1:n})$ by dependence on a finite sample $x_{1:n}$. This yields the approximate lower bound

$$\begin{aligned} I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) &\geq H(\boldsymbol{x}_{1:n}) + \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})} \left[\mathbb{E}_{p(\boldsymbol{x}_{1:n})} \left[\log p(\boldsymbol{x}_{1:n}|\boldsymbol{\theta}) \right] \right] \\ &\approx H(\boldsymbol{x}_{1:n}) + \mathbb{E}_{p(\boldsymbol{\theta}|x_{1:n})} \left[\sum_{i=1}^n \log p(x_i|\boldsymbol{\theta}) \right]. \end{aligned}$$

For any $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, we can also derive another approximate upper bound via

$$I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n} | \boldsymbol{x}_{n+1:\infty}) \leq \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})p(\boldsymbol{x}_{1:n})} \left[\log \left(\frac{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})}{\pi(\boldsymbol{\theta})} \right) \right]$$
$$\approx \mathbb{E}_{p(\boldsymbol{\theta}|x_{1:n})} \left[\log \left(\frac{p(\boldsymbol{\theta}|x_{1:n})}{\pi(\boldsymbol{\theta})} \right) \right]$$

Writing out the resulting bound and minimizing over $p(\boldsymbol{\theta}|x_{1:n}) \in \mathcal{P}(\boldsymbol{\Theta})$, we find that its minimizer is $P(-\log p(\cdot|\boldsymbol{\theta}), \beta\text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$.

Method	$\ell(\boldsymbol{\theta}, x_i)$	D	Π
Standard Bayes	$-\log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Power Likelihood Bayes[1]	$-\log p(x_i\|\boldsymbol{\theta})$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{P}(\boldsymbol{\Theta})$
Composite Likelihood Bayes[2]	$-w_i \log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Divergence-based Bayes[3]	divergence-based ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Gibbs/PAC-Bayes [4]	any ℓ	various D	$\mathcal{P}(\boldsymbol{\Theta})$
VAE[5,†]	$-\log p_{\boldsymbol{\zeta}}(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{Q}$
β-VAE[6,†]	$-\log p_{\boldsymbol{\zeta}}(x_i\|\boldsymbol{\theta})$	$\beta \cdot$ KLD, $\beta > 1$	$\mathcal{Q}$
Bernoulli-VAE[7,†]	continuous Bernoulli	KLD	$\mathcal{Q}$
Standard VI	$-\log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{Q}$
Power VI[8]	$-\log p(x_i\|\boldsymbol{\theta})$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{Q}$
Utility VI[9]	$-\log p(x_i\|\boldsymbol{\theta}) + \log u(h, x_i)$	KLD	$\mathcal{Q}$
Regularized Bayes[10]	$-\log p(x_i\|\boldsymbol{\theta}) + \phi(\boldsymbol{\theta}, x_i)$	KLD	$\mathcal{Q}$
Gibbs VI[11]	any ℓ	KLD	$\mathcal{Q}$
Posteriors in Online Learning [12]	any ℓ	f-divergences	$\mathcal{P}(\boldsymbol{\Theta})$ / $\mathcal{Q}$

Table 1.1: Relationship of $P(\ell, D, Q)$ to a selection of existing methods. [1](e.g. Grünwald, 2011, 2012; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019), [2](e.g. Varin et al., 2011; Pauli et al., 2011; Ribatet et al., 2012; Hamelijnck et al., 2019), [3](e.g. Hooker and Vidyashankar, 2014; Ghosh and Basu, 2016; Futoshi Futami et al., 2018; Jewson et al., 2018; Chérief Abdellatif and Alquier, 2020), [4](Bissiri et al., 2016; Germain et al., 2016; Guedj, 2019; Syring and Martin, 2019), [5](Kingma and Welling, 2013), [6](Higgins et al., 2017), [7](Loaiza Ganem and Cunningham, 2019) [8](e.g. Yang et al., 2020; Huang et al., 2018) [9](e.g. Kuśmierczyk et al., 2019; Lacoste Julien et al., 2011) [10](Ganchev et al. (2010), but only if the regularizer can be written as $\mathbb{E}_{q(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}, \boldsymbol{x})]$ as in Zhu et al. (2014)), [11](e.g. Alquier et al., 2016) [12](e.g. Alquier, 2021) [†]For notational clarification for the VAE entries in the table, see Appendix B.2.

1.3 Other directions of related research

It should go without saying that many other directions of research have sought to extend or modify Bayesian posteriors in various ways. This includes utility-based inference, so-called regularized Bayesian inference, divergence-based Bayesian methods, belief distributions motivated via Online Learning, and even Variational Autoencoders (VAEs). Notably, virtually all of these approaches can be seen as special cases of the RoT. While it is beyond the scope or purpose of this introduction to discuss each of these approaches and do them justice, Table 1.1 gives an overview of some of the most important ones.

1.4 Axiomatic Derivation

So far, we have focused on relating the RoT to existing generalizations aimed at fixing various shortcomings of Bayesian inference, and showing that it recovers them as a special case. Since statistics is foremost an epistemological science, another question well worth finding an answer to is under which conditions it is advisable for a statistician to construct posteriors via the RoT. To this end, the following section provides a simple axiomatic foundation. In essence, we argue that a useful generalization of standard Bayesian posteriors should have three main properties: First—and like standard Bayesian inference—it should be able to trade off prior information against information in the data by means of an optimization problem over probability measures. Second, the structure of this optimization problem should be the same regardless of the prior, loss, and data used to compute the belief. Third, the generalization must able to recover standard Bayesian inference.

Axiom I (Representation) *The posterior $q^* \in \mathcal{P}(\boldsymbol{\Theta})$ solves an minimisation problem over some space $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$. For any finite sample $\{x_i\}_{i=1}^n$, the minimisation problem's objective is increasing in two arguments:*

(i) An expected in-sample loss $L(\boldsymbol{\theta}, x_{1:n})$ taken with respect to $q^*(\boldsymbol{\theta})$.

(ii) Deviation from the prior $\pi(\boldsymbol{\theta})$ as measured by some statistical divergence D.

Simply put, Axiom I formalizes the optimization-centric view on Bayesian inference. More precisely, it tells us that for a fixed prior π, posteriors are specified through an optimization problem with three parts: The loss ℓ, the divergence $D(\cdot\|\pi)$ and the space Π over which the objective is optimized. Making this insight more precise, we can derive the following representation—which really is just a more convenient restatement of the Axiom.

Theorem 1.2 (Form 1)**.** Under Axiom I, posterior belief distributions can be written as

$$q^*(\boldsymbol{\theta}) = \arg\min_{q \in \Pi} \left\{ f\left(\mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right], D(q||\pi) \right) \right\},$$

where $f : \mathbb{R}^2 \to \mathbb{R}$ is some function that increases in both arguments, and may depend on $\pi, \Pi, \ell, \{x_i\}_{i=1}^n$, or D.

Proof. This follows directly from Axiom I: Firstly, any posterior belief distribution $q^*(\boldsymbol{\theta})$ is the solution to an optimization problem over Π. Thus, for an appropriately structured objective Obj, one can write

$$q^*(\boldsymbol{\theta}) = \arg\min_{q \in \Pi} \{ \mathrm{Obj}(q) \}.$$

By (i) and (ii) of Axiom I, we also know that the optimization's objective depends only on $D(q||\pi)$ and $\mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})]$. Clearly then, for some function $f : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathrm{Obj}(q) = f\left(\mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right], D(q||\pi) \right),$$

which completes the proof. □

This result is a first and helpful step, but in itself does not suffice to yield objectives that are useful in practice. Specifically, we need to get a handle on the function f. It is clear that under Axiom I alone, very little can be said about f. In fact, the mathematical mechanism f by which we compute posteriors may depend on (random) data and the loss, which is clearly undesirable and notably not a feature of standard Bayesian inference.[6] Further, since our explicit target is a generalization of the Bayesian inference problem, we will have to restrict the form of f so that Theorem 1.2 admits only the Bayesian posterior whenever $D = \mathrm{KLD}$ and $\Pi = \mathcal{P}(\boldsymbol{\Theta})$. Both these issues are addressed in the following Axiom.

Axiom II (Recovers Bayesian Posteriors) *Function f in Theorem 1.2 does not depend on $\pi, \Pi, \ell, \{x_i\}_{i=1}^n$, or D. Further, q^* is the posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ if $D = \mathrm{KLD}$, $\Pi = \mathcal{P}(\boldsymbol{\Theta})$.*

The intution of Axiom II is clear: in the case of $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$, $D = \mathrm{KLD}$, $\Pi = \mathcal{P}(\boldsymbol{\Theta})$; and f does not depend on the data $\{x_i\}_{i=1}^n$, the prior π, the loss ℓ, etc. As we want to recover $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$, we thus impose the same conditions for other posteriors. Fortunately, this also drastically simplifies the representation of Theorem 1.2.

[6] In particular, Bayes' rule does not depend on the choice of likelihood, prior, or the data observed.

Theorem 1.3. Suppose the posterior belief $q^* \in \mathcal{P}(\boldsymbol{\Theta})$ satisfies Axioms I and II. Then the objective of Theorem 1.2 can be written as $f(x, y) = x + y$ so that

$$q^*(\boldsymbol{\theta}) = \underset{q \in \Pi}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right] + D(q \| \pi) \right\} = P(L, D, \Pi). \tag{1.11}$$

Proof. This follows by combining Theorem 1.2 with Axiom II and the identity

$$q^*_{n,\mathrm{SB}}(\boldsymbol{\theta}) = \underset{q \in \mathcal{P}(\boldsymbol{\Theta})}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[-\log p(x_i|\boldsymbol{\theta}) \right] + \mathrm{KLD}(q \| \pi) \right\},$$

so that the result follows. □

The last result suggests that the RoT is not only an intuitively appealing flexible recipe for the optimization-centric design of new posterior distributions, but also a principled one. To conclude this section, we shall point out in two further remarks why the RoT is an attractive proposition. The first of these explains some connections the RoT has with the fundamental properties of standard Bayesian procedures. The second notes an attractive modularity property that ties in directly with the shortcomings of standard Bayesian methods as belaboured in Chapter 1.

Remark 1.1. Theorem 1.3 demonstrates that in combination with Axiom I, Axiom II enforces an additive relationship between the expected loss and prior regularizer. This additive relationship is desirable for a number of reasons, some of which include

- **Invariance to additive, but not multiplicative constants**: adding constants to L will not change the posterior. In other words for, any $C \in \mathbb{R}$, we have $P(L, D, \Pi) = P(\ell + C, D, \Pi)$. However, multiplying L by w (or equivalently, D by $\frac{1}{w}$) for some $w \in \mathbb{R}$ changes the posterior that is computed. This means that we recover a well-known feature of other Bayes-like procedures. In fact, exponentiating likelihoods $p(\cdot|\boldsymbol{\theta})^w$—which is equivalent to multiplying $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ with some $w \in (0, 1)$—is a popular tool in the existing literature on generalized Bayesian methods with $D = \mathrm{KLD}$ (e.g. Grünwald, 2011, 2012; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019) and serves to up-weight (or down-weight) the information of the data relative to the prior.

- **Recovery of (D-approximated) prior without additional Information**: Given no information from the observations (i.e. if $L = 0$), the solution of the optimization problem in Theorem 1.3 is the member of Π that is

closest to the prior $\pi(\boldsymbol{\theta})$ as measured by D. Put differently, $P(0, D, \Pi) = \arg\min_{q \in \Pi} D(q\|\pi)$. Clearly then, if $\pi \in \Pi$ we have that $P(0, D, \Pi) = \pi$.

- **Generalized (weak) likelihood principle**: Data $x_{1:n}$ favours $\boldsymbol{\theta}_1$ over $\boldsymbol{\theta}_2$ if and only if $L(\boldsymbol{\theta}_1, x_{1:n}) < L(\boldsymbol{\theta}_2, x_{1:n})$. This is the natural generalization of the 'weak likelihood principle' outlined by Sober (2008) from $\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta})$ to general loss functions , see also Mayo Wilson and Saraf (2020).

In summary, Theorem 1.3 results in a number of natural properties that one would want to hold for any belief distribution trading off prior against data-driven information.

Remark 1.2. Beyond the axiomatic development presented, there is another advantage of the RoT: each component of the optimization problem defined by the posterior $P(L, D, \Pi)$ serves a specific and separate purpose.

(💡L) A **loss** $L : \boldsymbol{\Theta} \times \mathcal{X}^n \to \mathbb{R}$. The loss defines the parameter of interest $\boldsymbol{\theta}$ by linking it to the observations $x_{1:n}$. To simplify presentation, we will assume that all losses depend only on a parameter $\boldsymbol{\theta}$ rather than on a latent variable.[7]

(💡P) A divergence $D : \mathcal{P}(\boldsymbol{\Theta}) \times \mathcal{P}(\boldsymbol{\Theta}) \to \mathbb{R}_+$ that **regularizes** the posterior by penalizing deviations from the prior $\pi(\boldsymbol{\theta})$. Beyond regularization, D also determines the nature of the uncertainty induced by π. To see this, consider $D = 0$ and the (non-RoT) problem

$$\widehat{q}(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{P}(\boldsymbol{\Theta})} \left\{ \mathbb{E}_q \left[\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \right] \right\}. \tag{1.12}$$

Denoting $\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \{L(\boldsymbol{\theta}, x_{1:n})\}$ and $\delta_y(x)$ as the Dirac measure at y, it is clear that $\widehat{q}(\boldsymbol{\theta}) = \delta_{\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{\theta})$, which holds as $\delta_{\widehat{\boldsymbol{\theta}}_n} \in \mathcal{P}(\boldsymbol{\Theta})$. Clearly, the absence of D corresponds to the absence of any uncertainty in the posterior. Similarly, the nature of D determines the nature in which uncertainty about $\boldsymbol{\theta}$ is quantified.

(💡C) A set of **feasible posteriors** $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$: By definition, any $q \in \Pi$ is a feasible solution for the optimization problem associated to the posterior $P(L, D, \Pi)$. The larger we choose the set Π, the more complicated we allow the resulting optimization problem to be. Consequently, the choice of Π amounts to regulating our computational budget.

[7] This requirement is not at all essential and easily relaxed: for instance, in the experiments on Deep Gaussian Processes in Chapter 6, all losses are defined relative to latent variables.

In summary, each of these three arguments directly addresses one of the problems (↯P), (↯L) and (↯C) in Section 1.1: Firstly, the loss L determines the parameter and thus can be used to tackle model misspecification and other violations of (L). Secondly—assuming one has specified the best possible prior—the divergence D can tackle (P) by shaping the nature in which priors affect the way in which the posterior quantifies uncertainty. Thirdly, the choice of Π can directly address (C): The more computational power is available, the more complex Π is allowed to become.

1.5 Structure of this thesis

The core contribution of this thesis will be the exploration and study of the Rule of Three (RoT). This proceeds in three parts: In the first part (Chapters 2 and 3), we will study its theoretical properties. This includes conditions under which RoT posteriors exist and are unique (Chapter 2). Beyond that, we study the dual of the minimization problem defining the RoT; and show that it allows us to re-interpret generalized Bayesian procedures as adversarially robust games (Chapter 2). Lastly, we also study frequentist consistency for RoT posteriors when the space over which the optimization is performed is parameterized (Chapter 3).

In the second part (Chapters 4–6), we study one of the main applications for the RoT of relevance in Machine Learning: Generalized Variational Inference (GVI). First, we discuss how to compute GVI posteriors using modern computing methods that have emerged with the increasing popularity of ordinary Variational Inference (VI) in Chapter 4. Next, we illustrate how changing the prior regularizer D in GVI posteriors can provide robustness if the prior belief is misspecified (Chapter 5); and how this leads to performance improvements when compared to a standard Bayesian approach. This is demonstrated on a range of Bayesian Neural Network (BNN) examples. Lastly, Chapter 6 discusses how GVI can make probabilistic inferences in black box Machine Learning methods robust to model misspecification. This implications are illustrated on Deep Gaussian Processes (DGPs)—a canonical black box Bayesian model of practical importance in many applications.

In the third and last part of the thesis (Chapters 7 and 8), we will study the methodological ramifications of the RoT on two classes of models: Changepoint (CP) models, and intractable likelihood models. For different reasons, both these statistical problems are often severely and adversely affected by the misalignment between the mathematical foundations of standard Bayesian inference on the one hand, and the real world on the other hand.

Part I

Theoretical Advances

Chapter 2

Existence, Uniqueness, and Duality

Summary: We give a full overview of the theoretical findings that relate to the optimization problem posed by the Rule of Three (RoT). In particular, we first study conditions under which RoT posteriors exist and are unique. While uniqueness is generally harder to show if D is not strictly convex in its first argument, existence is easier to prove. The second question we answer is what the dual problem of the minimization underlying the RoT looks like. Doing so, we form a direct and hitherto unknown connection between both standard and generalized versions of Bayesian inference on the one hand, and adversarial robustness on the other hand.

For simplicity, we will avoid introducing measure-theoretic notation in the remainder. To this end, we will typically assume that all probability measures of interest have densities with respect to the Lebesgue measure. We also slightly abuse notation in two ways: We often write $q \in \mathcal{P}(\boldsymbol{\Theta})$ for probability densities $q(\boldsymbol{\theta})$ of Borel measures on $\boldsymbol{\Theta}$, even though probability densities themselves are not in $\mathcal{P}(\boldsymbol{\Theta})$. However, $q(\boldsymbol{\theta})$ induces a Borel measure $\mu_q \in \mathcal{P}(\boldsymbol{\Theta})$ as $\mu_q(A) = \int_A q(\boldsymbol{\theta}) d\boldsymbol{\theta}$ for any measurable set $A \subset \boldsymbol{\Theta}$. Thus, whenever we write $q \in \mathcal{P}(\boldsymbol{\Theta})$ and it is clear that q is a density from context, what we mean is that $\mu_q \in \mathcal{P}(\boldsymbol{\Theta})$. Similarly, when we write $q_1 \neq q_2$, we mean that there exists a measurable set $A \in \boldsymbol{\Theta}$ such that $\mu_{q_1}(A) \neq \mu_{q_2}(A)$.

While we informally introduced the RoT in Chapter 1, we begin our analysis with a number of more formal definitions that will be applicable throughout this thesis.

Definition 2.1 (Loss Function)**.** Losses are functions $L : \boldsymbol{\Theta} \times \mathcal{X}^n \to \mathbb{R}$ which are lower bounded. For observations $x_{1:n} \in \mathcal{X}^n$, their empirical risk minimizers are given by

$$\widehat{\boldsymbol{\theta}}_n \in \operatorname*{arg\,inf}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ L(\boldsymbol{\theta}, x_{1:n}) \right\},$$

and all elements of $\operatorname{arg\,inf}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ L(\boldsymbol{\theta}, x_{1:n}) \right\}$ are finite-valued.

Definition 2.2 (Statistical Divergence)**.** Statistical divergences are functions $D : \mathcal{P}(\boldsymbol{\Theta}) \times \mathcal{P}(\boldsymbol{\Theta}) \to \mathbb{R}$ so that $D(q\|\pi) \geq 0$ and $D(q\|\pi) = 0 \iff q = \pi$.

With these definitions out of the way, we can now proceed to formally stating the RoT—the conceptual corner stone of this thesis.

Definition 2.3 (Rule of Three (RoT))**.** For observations $x_{1:n} \in \mathcal{X}^n$, a prior $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, a space $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$, a loss function $L : \boldsymbol{\Theta} \times \mathcal{X}^n \to \mathbb{R}$ and a divergence $D(\cdot\|\pi) : \Pi \to \mathbb{R}_{\geq 0}$, we say that a posterior belief distribution q^* has been constructed via the Rule of Three (RoT) if it can be written as

$$q^* \in P(L, D, \Pi) = \operatorname*{arg\,inf}_{q \in \Pi} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right] + D(q\|\pi) \right\}.$$

Here, $P(L, D, \Pi)$ is a short-hand notation for the RoT suppressing dependence on $x_{1:n}$ and π. If L decomposes additively as $L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$, then we additionally define $P(L, D, \Pi) = P(\ell, D, \Pi)$.

Contrary to the introductory exposition, Definition 2.3 has used the arg inf (rather than the arg min) operator to define the RoT—meaning that the RoT is defined even if the minimizer is not attained inside Π. However, while we cannot generally assume that the minimizer $q^*(\boldsymbol{\theta})$ is unique, in the more methodologically oriented chapters of this thesis we will often ignore this problem and treat posteriors derived via the RoT as if they had unique minimizers; and write $P(L, D, \Pi) = \operatorname{arg\,min}_{q \in \Pi} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_{i=1}^n L(\boldsymbol{\theta}, x_{1:n}) \right] + D(q\|\pi) \right\}$ for convenience. As our first result will reveal, this assumption is unproblematic so long as $D(\cdot\|\pi)$ is strictly convex on Π.

2.1 Existence and Uniqueness

Since we will often treat $P(L, D, \Pi)$ as if a minimizer in the interior of Π could be guaranteed to exist, an important question is under which conditions this is

justified. Using basic analysis, the next result gives some sufficient conditions for an even stronger result: uniqueness.

Proposition 2.1. Suppose that Π is convex, that $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})]$ is linear, and that $q \mapsto D(q\|\pi)$ is a strictly convex function on Π. Then, $P(L, D, \Pi)$ is a singleton.

Proof. This follows by basic analysis: if, $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})]$ is linear, it is a convex function. Thus, if $q \mapsto D(q\|\pi)$ is strictly convex, the function $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] + D(q\|\pi)$ also is. Since L and D are both lower bounded (see Definitions 2.1 and 2.2), the function has a unique minimizer and the result follows. □

This result is simple, but quite useful since most divergences of practical interest will be convex if we choose $\Pi = \mathcal{P}(\boldsymbol{\Theta})$ to be the set of Borel measures on $\boldsymbol{\Theta}$. For instance, it holds for all strictly convex f-divergences.

Corollary 2.1. Let L be any loss function, and take $\Pi = \mathcal{P}(\boldsymbol{\Theta})$. If D is an f-divergence so that f is a strictly convex function on $\mathbb{R}_+$ with $f(1) = 0$ and

$$D(q\|\pi) = \mathbb{E}_{\pi(\boldsymbol{\theta})}[f(q(\boldsymbol{\theta})/\pi(\boldsymbol{\theta}))].$$

Then, $P(L, D, \Pi)$ is a singleton.

Proof. This follows by applying Proposition 2.1, since it is well-known that $q \mapsto D(q\|\pi)$ is strictly convex whenever f is. □

The same result could be derived for numerous divergences outside the class of f-divergences by following the same steps—including Rényi-divergences, a number of Integral Probability Metrics, and certain members of the family of β- and γ-divergences.

In some cases, we will not be able to show that $P(L, D, \Pi)$ is a singleton. In these cases, we may still want to show that a minimizer exists, which is to say that $P(L, D, \Pi)$ is not the empty set. There are two ways of guaranteeing this: via a weaker convexity requirement on D, or via an assumption of coerciveness on L. The former is a simple modification of Proposition 2.1, while the latter is technically quite involved and relies on tightness and Prokhorov's Theorem.

Proposition 2.2. Suppose that $D(\cdot\|\pi) \mapsto \mathbb{R}$ is a convex function on Π, and that Π is a convex, closed and bounded set. Then, $P(L, D, \Pi)$ is non-empty.

Proof. This follows by similar arguments as Proposition 2.1: $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] + D(q\|\pi)$ is convex, L and D are lower bounded, and so a minimizer exists. □

We next prove a more involved result that also guarantees that $P(L, D, \mathcal{P}(\boldsymbol{\Theta}))$ is non-empty, but under much milder assumptions on D. In particular, all that is required is that D is lower semi-continuous on the set of Borel measures $\mathcal{P}(\boldsymbol{\Theta})$ on $\boldsymbol{\Theta}$. This is an extremely weak requirement, and will be satisfied by virtually any divergence of practical interest.

Definition 2.4 (lower-semicontinuity)**.** A function $F : X \to \overline{\mathbb{R}}$ is lower-semicontinuous at $x \in X$ if for every sequence $\{x_n\}_{n\in\mathbb{N}}$ so that $x_n \to x$,

$$f(x) \leq \liminf_{n\to\infty} f(x_n).$$

We say that a function F is lower-semicontinuous on X if F is lower-semicontinuous for all $x \in X$.

In order to state the proof, we need a technical Lemma that relates coerciveness and the existence of minimizers.

Lemma 2.1. Suppose the functional $q \mapsto F(q)$ defined over $\mathcal{P}(\mathbb{R}^d)$ for some $d \in \mathbb{N}$ is lower-bounded, coercive, and lower semi-continuous. Then, there exists q^* so that $\inf_{q\in\mathcal{P}(\boldsymbol{\Theta})} F(q) = F(q^*)$.

Proof. Let $\{q_n\}_{n\in\mathbb{N}}$ be a minimizing sequence so that $\inf_{q\in\mathcal{P}(\boldsymbol{\Theta})} F(q) = \lim_{n\to\infty} F(q_n)$. Because F is coercive, its sub-level sets are closed and compact. Because of this and thanks to Prokhorov's Theorem, it must hold that there exists $q^* \in \mathcal{P}(\boldsymbol{\Theta})$ so that weakly, $q_n \to q^*$ as $n \to \infty$. By definition of the lower-semicontinuity of F, it also holds that

$$F(q^*) \leq \lim_{n\to\infty} F(q_n) = \inf_{q\in\mathcal{P}(\boldsymbol{\Theta})} F(q).$$

Since $q^* \in \mathcal{P}(\boldsymbol{\Theta})$, we also have that $\inf_{q\in\mathcal{P}(\boldsymbol{\Theta})} F(q) \leq F(q^*)$, so that $F(q^*) = \inf_{q\in\mathcal{P}(\boldsymbol{\Theta})} F(q)$. □

Theorem 2.1. Let $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ for $d \in \mathbb{N}$ be a (subset of a) Euclidean space. Suppose that $\boldsymbol{\theta} \to L(\boldsymbol{\theta}, x_{1:n})$ is norm-coercive on $\boldsymbol{\Theta}$ or that $\boldsymbol{\Theta}$ is a compact space. Further, assume that $q \mapsto D(\cdot\|\pi)$ is lower semi-continuous on $\mathcal{P}(\boldsymbol{\Theta})$. Then $P(L, D, \mathcal{P}(\boldsymbol{\Theta}))$ is non-empty.

Proof. We prove this by showing that the function given by

$$q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] + D(q\|\pi) \tag{2.1}$$

is coercive (which is to say that its sub-level sets are both closed and compact). Once this is established, we apply Lemma 2.1 to conclude the proof.

First, note that $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] + D(q\|\pi)$ has closed sub-level sets, since the functional is lower semi-continuous. Next, observe that it suffices to show that $\{q \in \mathcal{P}(\boldsymbol{\Theta}) : \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] \leq t\}$ is compact for any $t \in \mathbb{R}$, since it clearly holds that $\{q \in \mathcal{P}(\boldsymbol{\Theta}) : \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] + D(q\|\pi) \leq t\} \subseteq \{q \in \mathcal{P}(\boldsymbol{\Theta}) : \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] \leq t\}$ by virtue of the fact that $D(q\|\pi) \geq 0$ for any $q \in \mathcal{P}(\boldsymbol{\Theta})$. More specifically, being a subset here implies (countable) compactness because $\mathcal{P}(\boldsymbol{\Theta})$ is a metric space (via Prokhorov's metric), and because $\{q \in \mathcal{P}(\boldsymbol{\Theta}) : \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] + D(q\|\pi) \leq t\}$ is a closed set by virtue of $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] + D(q\|\pi)$ being a lower semi-continuous function.[1] Thus, we proceed by showing that $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})]$ has compact lower sub-level sets.

First, suppose that $\boldsymbol{\Theta}$ is compact in the Euclidean metric. It is well-known that this immediately implies that $\mathcal{P}(\boldsymbol{\Theta})$ is tight. By Prokhorov's Theorem, this implies that $\mathcal{P}(\boldsymbol{\Theta})$ is compact, which means that any subset of $\mathcal{P}(\boldsymbol{\Theta})$ is compact. Thus, $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})]$ is coercive whenever $\boldsymbol{\Theta}$ is compact.

Now instead of compactness for $\boldsymbol{\Theta}$, suppose that $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta}, x_{1:n})$ is norm-coercive on $\boldsymbol{\Theta}$. In other words, $L(\boldsymbol{\theta}, x_{1:n}) \to \infty$ as $\|\boldsymbol{\theta}\| \to \infty$, where $\|\cdot\|$ denotes the usual Euclidean norm. First, define the sub-level sets as $\mathcal{S}_t = \{q \in \mathcal{P}(\boldsymbol{\Theta}) : \mathbb{E}_{q(\boldsymbol{\theta})}[L(\boldsymbol{\theta}, x_{1:n})] \leq t\}$. Since $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta}, x_{1:n})$ is coercive in $\boldsymbol{\Theta}$, for any constant $C \in \mathbb{R}$, there exists $\boldsymbol{\theta}_C$ so that for a sufficiently large ball $B_{\boldsymbol{\theta}_C}(r_C) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|^{1/2} \leq r_C\}$ of radius r_C around $\boldsymbol{\theta}_C$, $L(\boldsymbol{\theta}, x_{1:n}) \geq C$ for all $\boldsymbol{\theta} \notin B_{\boldsymbol{\theta}_C}(r_C)$. Thus, for any $q \in \mathcal{S}_t$ and for any arbitrarily small and fixed $\Delta > 0$,

$$C \int_{\boldsymbol{\Theta} \setminus B_{\boldsymbol{\theta}_C}(r_C)} q(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq \int_{\boldsymbol{\Theta} \setminus B_{\boldsymbol{\theta}_C}(r_C)} L(\boldsymbol{\theta}, x_{1:n}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} < t + \Delta < \infty.$$

Rearranging terms, this immediately implies that

$$\int_{\boldsymbol{\Theta} \setminus B_{\boldsymbol{\theta}_C}(r_C)} q(\boldsymbol{\theta}) d\boldsymbol{\theta} < \frac{t + \Delta}{C}$$

Since C was chosen arbitrarily and can be picked arbitrarily large thanks to the coerciveness of $L(\boldsymbol{\theta}, x_{1:n})$, this immediately implies that $\mathcal{S}_t$ is tight. Again, by Prokhorov's Theorem this implies that $\mathcal{S}_t$ is compact, which completes the proof. □

While all existence and uniqueness results derived in this section consider problems for convex sets Π, in practice one is often forced to compute $P(L, D, \mathcal{Q})$ for

[1] In metric spaces, if K is compact and $K' \subset K$, it suffices to show that K' is closed to conclude that K' must be compact.

some set of parametric distributions $\mathcal{Q} \subset \mathcal{P}(\boldsymbol{\Theta})$ as in variational methods. Crucially, $\mathcal{Q}$ will typically not be convex: for example, if we choose $\mathcal{Q}$ to be the set of normal distributions, $\mathcal{Q}$ will not contain mixtures of normals. Though the current thesis does not provide theoretical guarantees for the uniqueness of these posteriors, this is not treated as a problem throughout. There are two main reasons for this: firstly, this problem is all but new. In fact, in practice most standard variational inference (VI) posterior approximations of the form $P(-\log p(\cdot|\boldsymbol{\theta}), \mathrm{KLD}, \mathcal{Q})$ proceed without guarantees on their uniqueness (or in fact even existence). Secondly—and closely mirroring the standard VI case—our empirical findings suggest that it is reasonable to expect objects of the form $P(L, D, \mathcal{Q})$ to be unique in most cases of practical interest.

2.2 Duality & Adversarial Robustness

So far, the only property of the RoT we have established relates to its fundamental properties as an optimization problem. We shall stick with this theme for now, and ask a second question: when thinking about the minimization problem $P(L, D, \Pi)$, how can we derive its dual form? And even more importantly, what can we learn from the dual?

2.2.1 Preliminaries

To this end, the remainder of this section uses the optimization-centric formulation of Bayesian procedures to tap into a very different branch of optimization: Duality Theory. Doing so allows us to leverage the structural and constraint properties of the optimization problem to explain the robustness of generalized Bayesian procedures specified through the RoT via *Fenchel* duality. While Fenchel duality has been extremely useful for the theoretical study of other Machine Learning methods such as Generative Adversarial Networks (see Farnia and Tse, 2018; Liu and Chaudhuri, 2018; Husain et al., 2019) and regularization (Husain, 2020), the remainder is the first analysis of this kind for (generalised and standard) Bayesian methods, and lead to the discovery of a fundamental connection between risk robustness and the variational optimization problem underlying (generalised and standard) Bayesian inference. Understanding this connection advances our insight into Bayesian methods: Specifically, it provides a new, concise, and rigorous explanation why a large class of Bayes-like methods typically outperform point estimation methods.

Before we can state it formally, we will need to define additional notation. We take $\mathcal{F}_b(\boldsymbol{\Theta})$ as the set of bounded and measurable functions on $\boldsymbol{\Theta}$, and $\mathcal{B}(\boldsymbol{\Theta})$

as its dual space—the set of finitely-additive measures on $\boldsymbol{\Theta}$. Throughout, we will denote $\mathcal{P}(\boldsymbol{\Theta})$ as the set of Borel measures on $\boldsymbol{\Theta}$. (Note that $\mathcal{P}(\boldsymbol{\Theta}) \subset \mathcal{B}(\boldsymbol{\Theta})$.) Further, we take $w \in \mathbb{R}_+$, and we define the objective value associated with $P(wL, D, \Pi) = P(L, w^{-1}D, \Pi)$ as

$$\mathcal{G}_{L,w^{-1}D,\Pi} := \inf_{q \in \Pi} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right] + w^{-1} D(q \| \pi) \right\}. \tag{2.2}$$

Note that the role of w^{-1} here is to up- or down-weight the regularizer relative to the loss. Equivalently, the role of w is to up- or down-weight the loss relative to the regularizer. For the special case of Gibbs posteriors, the role of w has been subject of frequent discussion (e.g. Grünwald, 2011, 2012; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019) and can be interpreted directly, since

$$P(L, w^{-1}\mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta})) = P(w \cdot L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta})) = \frac{\exp\{-wL(\boldsymbol{\theta}, x_{1:n})\}\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \exp\{-wL(\boldsymbol{\theta}, x_{1:n})\} d\pi(\boldsymbol{\theta})}.$$

For any set $A \subseteq \mathcal{B}(\boldsymbol{\Theta})$ and $h \in \mathcal{F}_b(\Theta)$, we use $\sigma_A(h) = \sup_{\nu \in A} \langle h, \nu \rangle$ and $\iota_A(\nu) = \infty \cdot [\![\nu \notin A]\!]$ to denote the *support* and *indicator* functions such as in Rockafellar (1970).

Throughout the development of our duality theory, we will make use of the following regularity condition.

Assumption 2.1. $\boldsymbol{\Theta}$ and $\mathcal{X}$ admit Polish topology. D is a divergence that is both convex and lower semi-continuous in its first argument, the loss $L(\cdot, x_{1:n})$ is an element of $\mathcal{F}_b(\boldsymbol{\Theta})$, and the prior π is an element of $\mathcal{P}(\boldsymbol{\Theta})$.

With this in hand, we can now define one of the main tools of our analysis: the Legendre-Fenchel dual.

Definition 2.5. For $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, the Legendre-Fenchel conjugate of a regularizer $D(\cdot\|\pi) : \mathcal{P}(\boldsymbol{\Theta}) \to \mathbb{R}$ is

$$D^\star_\pi(L') = \sup_{\mu \in \mathcal{B}(\boldsymbol{\Theta})} \left\{ \int_{\boldsymbol{\Theta}} L'(\boldsymbol{\theta}) d\mu(\boldsymbol{\theta}) - D(\mu\|\pi) \right\},$$

for any $L' \in \mathcal{F}_b(\boldsymbol{\Theta})$.

For convenience, we also define an auxiliary minimization problem which appears as part of the Legendre-Fenchel dual.

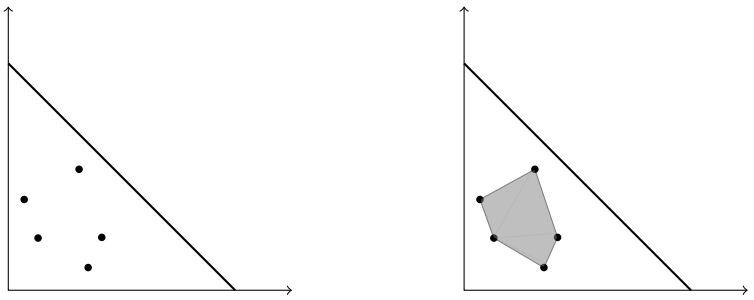


Figure 2.1: The left image illustrates a choice for Π which consists of five probability vectors over $\boldsymbol{\Theta} = \{a, b, c\}$. The right illustrates $\overline{\text{co}}\,(\Pi)$ over this choice where one can see that the selection of probabilities increases vastly.

Definition 2.6. For any set of probability distributions $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$, we define for any $L(\cdot, x_{1:n}) \in \mathcal{F}_b(\boldsymbol{\Theta})$

$$\mathsf{E}_\Pi(L) = \inf_{q \in \Pi} \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right]$$

In words, $\mathsf{E}_\Pi(L)$ denotes the smallest possible value achievable by integrating the loss $L(\cdot, x_{1:n})$ with an element from the class of probability distributions Π. For example, if there is a unique minimizer $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}, x_{1:n})$ and a sequence $q_n \in \Pi$ so that $q_n \to \delta_{\hat{\boldsymbol{\theta}}}$ as $n \to \infty$, then we will have $\mathsf{E}_\Pi(L) = \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}, x_{1:n})$.

Lastly, we introduce the *closed convex hull* of a set Π of admissible solutions to the optimization problem in Definition 2.3. For a set of potential posteriors Π, $\overline{\text{co}}\,(\Pi)$ denotes the smallest *closed* and *convex* set containing Π. In particular, we will have

$$\lambda \cdot q + (1 - \lambda) \cdot q' \in \overline{\text{co}}\,(\Pi)\,, \tag{2.3}$$

for all $q, q' \in \Pi$ and $\lambda \in [0, 1]$. We illustrate this in Figure 2.1 for a discrete parameter space $\boldsymbol{\Theta} = \{a, b, c\}$ with only three elements. In this setting, $\mathcal{P}(\boldsymbol{\Theta})$ is simply the set of vectors in $\mathbb{R}^3_{\geq 0}$ whose co-ordinates sum to 1; though they can be viewed as elements in $\mathbb{R}^2$ enclosed in a triangle with vertices $(0, 0)$, $(0, 1)$ and $(1, 0)$. While the definition of the convex hull is best understood in the geometric sense, it also has a clear probabilistic counterpart in mixture models: For example, if Π is the set of normal distributions, then $\overline{\text{co}}\,(\Pi)$ is the set of all (finite and infinite) mixtures of normal distributions on $\boldsymbol{\Theta}$.

For pedagogical reasons, we now proceed by illustrating all relevant concepts and definitions for this section on a simple example based on a least squares

loss. This example is of little practical interest, and mostly meant for the reader's convenience.

Example 2.1. Given a dataset $\{(z_i, y_i)\}_{i=1}^n$ and $x_i = (z_i, y_i)$ so that $z_i, y_i \in \mathbb{R}$ and $\mathcal{X} = \mathbb{R}^2$, and the parameter space $\boldsymbol{\Theta} = [0, 1]$, define the corresponding least squares loss

$$L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^{n} (z_i^\top \boldsymbol{\theta} - y_i)^2.$$

Further, consider for some $m \in \mathbb{N}$ the discretely supported and uniform prior

$$\pi(\boldsymbol{\theta}) = \frac{1}{m+1} \sum_{j=0}^{m} \delta_{(j/m)}(\boldsymbol{\theta})$$

and the variational family supported on the same atoms as

$$\Pi = \left\{ \sum_{j=0}^{m} w_j \delta_{(j/m)}(\boldsymbol{\theta}) : w_j \geq 0, \sum_{j=0}^{m} w_j = 1 \right\}.$$

Clearly, both $\pi \in \Pi$ and $\overline{\text{co}}\,(\Pi) = \Pi$. Considering as regularizer the χ^2-divergence given for any $q \in \Pi$ by

$$\chi^2(q \| \pi) = \mathbb{E}_{\pi(\boldsymbol{\theta})} \left[\left(\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} - 1 \right)^2 \right],$$

its Legendre-Fenchel conjugate is defined as

$$\chi^{2,\star}_\pi(L') = \sup_{\mu \in \mathcal{B}(\boldsymbol{\Theta})} \left\{ \int_{\boldsymbol{\Theta}} L'(\boldsymbol{\theta}) d\mu(\boldsymbol{\theta}) - \chi^2(\mu \| \pi) \right\},$$

for any $L' \in \mathcal{F}_b(\boldsymbol{\Theta})$. Further, taking $w^{-1} = 1$, the corresponding RoT is

$$P(L, w^{-1}\chi^2, \Pi) = \arg\inf_{q \in \Pi} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right] + \chi^2(q, \pi) \right\}.$$

Lastly, we have that

$$\mathsf{E}_\Pi(L) = \inf_{w_{1:n} : \sum_{i=1}^m w_i = 1, w_i \geq 0 \forall i} \left\{ \sum_{j=0}^{m} \sum_{i=1}^{n} w_j (z_i \cdot (j/m) - y_i)^2 \right\}.$$

Before we can state the two main Theorems of this section, we will need to introduce a number of key technical Lemmas. For completeness, we state these Lemmas here but defer their proofs to Appendix C.1.

Lemma 2.2.1. For any $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$ and $L \in \mathcal{F}_b(\boldsymbol{\Theta})$, we have

$$\mathsf{E}_{\overline{\mathrm{co}}(\Pi)}[L] = \mathsf{E}_{\Pi}[L].$$

Lemma 2.2.2. For any prior $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, we have

$$D(q\|\pi) = \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \{\mathbb{E}_{q(\boldsymbol{\theta})}[\rho(\boldsymbol{\theta})] - D^{\star}_{\pi}(\rho)\}.$$

Lemma 2.2.3. For any prior $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, regularizer D and set $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$, define a function $F : \mathcal{P}(\boldsymbol{\Theta}) \times \mathcal{F}_b(\boldsymbol{\Theta}) \to \mathbb{R}$ as

$$F(q, \rho) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})}\left[\rho(\boldsymbol{\theta})\right] - D^{\star}_{\pi}(\rho) + \iota_{\overline{\mathrm{co}}(\Pi)}(q).$$

It holds that

$$\inf_{q \in \mathcal{P}(\boldsymbol{\Theta})} \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} F(q, \rho) = \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \inf_{q \in \mathcal{P}(\boldsymbol{\Theta})} F(q, \rho).$$

2.2.2 Main Results regarding Duality

We can now finally prove the main results of the current section. The first of these is a strong duality result that reveals an interesting structure of the dual problem associated with the RoT.

Theorem 2.2 (Strong Duality)**.** For any $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$, $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, and loss $L \in \mathcal{F}_b(\boldsymbol{\Theta})$, it holds for any $w^{-1} > 0$ that

$$\mathcal{G}_{L,D,\overline{\mathrm{co}}(\Pi)} = \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \left\{ \mathsf{E}_{\Pi}(L + \rho) - w^{-1} D^{\star}_{\pi}\left(\frac{\rho}{w^{-1}}\right) \right\}. \tag{2.4}$$

Proof.

$$
\begin{aligned}
&\inf_{q\in\overline{\mathrm{co}}(\Pi)} \left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta})\right] + D(q\|\pi)\right\} \\
&\overset{(1)}{=} \inf_{q\in\overline{\mathrm{co}}(\Pi)} \left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta})\right] + \sup_{\rho\in\mathcal{F}_b(\boldsymbol{\Theta})} \left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[\rho(\boldsymbol{\theta})\right] - D_\pi^\star(\rho)\right\}\right\} \\
&= \inf_{q\in\mathcal{P}(\boldsymbol{\Theta})} \sup_{\rho\in\mathcal{F}_b(\boldsymbol{\Theta})} \left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})}\left[\rho(\boldsymbol{\theta})\right] - D_\pi^\star(\rho) + \iota_{\overline{\mathrm{co}}(\Pi)}(q)\right\} \\
&\overset{(2)}{=} \sup_{\rho\in\mathcal{F}_b(\boldsymbol{\Theta})} \inf_{q\in\mathcal{P}(\boldsymbol{\Theta})} \left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})}\left[\rho(\boldsymbol{\theta})\right] - D_\pi^\star(\rho) + \iota_{\overline{\mathrm{co}}(\Pi)}(q)\right\} \\
&= \sup_{\rho\in\mathcal{F}_b(\boldsymbol{\Theta})} \left\{\inf_{q\in\mathcal{P}(\boldsymbol{\Theta})} \left\{\mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta}) + \rho(\boldsymbol{\theta})\right] + \iota_{\overline{\mathrm{co}}(\Pi)}(q)\right\} - D_\pi^\star(\rho)\right\} \\
&\overset{(3)}{=} \sup_{\rho\in\mathcal{F}_b(\boldsymbol{\Theta})} \left\{\mathsf{E}_{\overline{\mathrm{co}}(\Pi)}(L+\rho) - D_\pi^\star(\rho)\right\} \\
&\overset{(4)}{=} \sup_{\rho\in\mathcal{F}_b(\boldsymbol{\Theta})} \left\{\mathsf{E}_{\Pi}(L+\rho) - D_\pi^\star(\rho)\right\},
\end{aligned}
$$

where (1) is due to Lemma 2.2.2, (2) is due to Lemma 2.2.3, (3) is by definition of E_Π and (4) holds due to Lemma 2.2.1. The proof concludes by noting that the dual of $w^{-1}D(\cdot,\pi)$ is $w^{-1}D_\pi^\star(\cdot/w^{-1})$. □

The above result is interesting because it shows that the RoT has a close correspondence to adversarial games. Specifically, the adversary in the game of Theorem 2.2 changes the original loss L via some adversarial perturbation ρ so that the minimum achievable integrated and perturbed loss $\mathbb{E}_{q(\boldsymbol{\theta})}[L+\rho]$—in other words, $\mathsf{E}_\Pi[L+\rho]$—is as large as possible. For perturbing the loss in this way however, the adversary pays a price $w^{-1}D_\pi^\star(\rho/w^{-1})$. Note in particular that the nature of this price for worsening the loss depends on the choice of prior π and the choice of divergence D. As we shall see in the examples presented later, the role of π becomes that of a cost function: in essence, it is more expensive for the adversary to perturb L in regions of $\boldsymbol{\Theta}$ with high probability under π. This prevents the perturbation ρ from being flexible enough to make $\mathsf{E}_\Pi[L+\rho]$ infinitely large.

Regarding the result, note also that $\Pi \subseteq \overline{\mathrm{co}}\,(\Pi)$, which implies that in general, $\mathcal{G}_{L,D,\Pi} \geq \mathcal{G}_{L,D,\overline{\mathrm{co}}(\Pi)}$. In practice, this means that unless our RoT is built on a set Π for which it holds that $\Pi = \overline{\mathrm{co}}\,(\Pi)$, we cannot use the above strong duality result. However, even if $\Pi \neq \overline{\mathrm{co}}\,(\Pi)$ we would still be able to conclude that

$$
\mathcal{G}_{L,D,\Pi} \geq \sup_{\rho\in\mathcal{F}_b(\boldsymbol{\Theta})} \left\{\mathsf{E}_\Pi(L+\rho) - w^{-1}D_\pi^\star\left(\frac{\rho}{w^{-1}}\right)\right\}. \tag{2.5}
$$

For the standard choice $\Pi = \mathcal{P}(\boldsymbol{\Theta})$, it clearly holds that $\Pi = \overline{\mathrm{co}}\,(\Pi)$. Hence,

we obtain the following strong duality result that holds for standard Bayesian inference problems as well as for posteriors derived via the RoT with $\Pi = \mathcal{P}(\boldsymbol{\Theta})$. This includes the generalised posteriors studied in the previous chapter, as well as those considered by Reid et al. (2015); Knoblauch (2019); Alquier (2021).

Corollary 2.2. If $\Pi = \mathcal{P}(\boldsymbol{\Theta})$ then for any $\pi, \in \mathcal{P}(\boldsymbol{\Theta})$, $L \in \mathcal{F}_b(\boldsymbol{\Theta})$ and $w > 0$, it holds that

$$\mathcal{G}_{L,D,\Pi} = \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \left\{ \mathsf{E}_\Pi(L + \rho) - w^{-1} D_\pi^\star \left(\frac{\rho}{w^{-1}} \right) \right\}.$$

While Corollary 2.2 connects the two *values* of the objectives at the optimum, we can make this dual connection much firmer. In fact, the RoT primal and its adversarial dual share another—perhaps even more important—connection: The RoT posterior minimizes the loss that results from the adversary's perturbation.

Theorem 2.3 (Adversarial Robustness of GVI)**.** Let ρ^* denote a maximizer of the dual in (2.5). If Π is convex and closed, it holds that

$$P(L, D, \Pi) = \arg\inf_{q \in \Pi} \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}) + \rho^*(\boldsymbol{\theta}) \right].$$

Proof. Take $q_{D,L,\Pi} \in P(L, D, \Pi)$. Then, note that

$$\mathbb{E}_{q_{L,D,\Pi}(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}) + \rho^*(\boldsymbol{\theta}) \right] \geq \inf_{q \in \Pi} \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}) + \rho^*(\boldsymbol{\theta}) \right]$$

by definition. For the other direction, we have

$$\begin{aligned}
&\inf_{q \in \Pi} \mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}) + \rho^*(\boldsymbol{\theta}) \right] - \mathbb{E}_{q_{L,D,\Pi}(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}) + \rho^*(\boldsymbol{\theta}) \right] \\
&= \left(\mathbb{E}_\Pi[L + \rho^*] - w^{-1} D_\pi^\star \left(\frac{\rho^*}{w^{-1}} \right) \right) + w^{-1} D_\pi^\star \left(\frac{\rho^*}{w^{-1}} \right) - \mathbb{E}_{q_{L,D,\Pi}(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}) + \rho^*(\boldsymbol{\theta}) \right] \\
&\overset{(1)}{=} \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \left\{ \mathbb{E}_\Pi[L + \rho] - w^{-1} D_\pi^\star \left(\frac{\rho}{w^{-1}} \right) \right\} + w^{-1} D_\pi^\star \left(\frac{\rho^*}{w^{-1}} \right) - \mathbb{E}_{q_{L,D,\Pi}(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}) + \rho^*(\boldsymbol{\theta}) \right] \\
&\overset{(2)}{=} \inf_{q \in \Pi} \left\{ \mathbb{E} q[L] + w^{-1} D(q, \pi) \right\} + w^{-1} D_\pi^\star \left(\frac{\rho^*}{w^{-1}} \right) - \mathbb{E}_{q_{L,D,\Pi}(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}) + \rho^*(\boldsymbol{\theta}) \right] \\
&\overset{(3)}{=} \mathbb{E}_{q_{L,D,\Pi}(\boldsymbol{\theta})}[L] + w^{-1} D \left(q_{L,D,\Pi} \| \pi \right) + w^{-1} D_\pi^\star \left(\frac{\rho^*}{w^{-1}} \right) - \mathbb{E}_{q_{L,D,\Pi}(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}) + \rho^*(\boldsymbol{\theta}) \right] \\
&= w^{-1} D \left(q_{L,D,\Pi} \| \pi \right) + w^{-1} D_\pi^\star \left(\frac{\rho^*}{w^{-1}} \right) - \mathbb{E}_{q_{L,D,\Pi}(\boldsymbol{\theta})} \left[\rho^*(\boldsymbol{\theta}) \right] \\
&\overset{(4)}{\geq} 0,
\end{aligned}$$

where (1) is due to the optimality of ρ^*, (2) is via Theorem 2.2 noting that Π is closed and convex by assumption, (3) is due to optimality of $q_{L,D,\Pi}$ and (4) holds by applying the Fenchel-Young inequality on D. □

This last result is striking: It tells us that posteriors derived via the RoT are adversarially robust. More specifically, they produce optimal beliefs in the presence of an adversary whose cost for perturbing the original loss L by ρ is given by $w^{-1}D_\pi^\star\left(\rho/w^{-1}\right)$. Though this idea is not pursued in the applications of this thesis, the finding is particularly appealing for designing RoT posteriors: if we work out the divergence D from a desired penalty $w^{-1}D_\pi^\star\left(\rho/w^{-1}\right)$ imposed upon the adversary, then we can use this result to construct posteriors motivated directly by adversarial robustness considerations.

2.2.3 Examples

To illustrate the meaning of these results in more detail in a less abstract context, we present some examples using two popular families of divergences: f-divergences and Integral Probability Metrics (IPMs). For both f-divergences and IPMs, the derivations used for the examples can be found in Appendix C.1

f-divergences

For a convex lower semicontinuous function $f : \mathbb{R} \to (-\infty, \infty]$, the corresponding f-divergence is $D(q\|\pi) = \int_{\boldsymbol{\Theta}} f(q(\boldsymbol{\theta})/\pi(\boldsymbol{\theta}))d\pi(\boldsymbol{\theta})$ if π is absolutely continuous with respect to q and $D(q\|\pi) = \infty$ otherwise. This includes the popular Kullback-Leibler divergence (KLD) when $f(t) = t\log t$ and the χ^2-divergence if $f(t) = (t-1)^2$.

Hence, our theory applies to regularizers that are members of the f-divergences. To gain some intuition about the penalization function these divergences entail for the adversary, we will study the general case, as well as the special cases of the KLD and the χ^2-divergence more closely.

Example 2.2 (KLD)**.** If $D = \text{KLD}$ then the dual problem is

$$\mathcal{G}_{L,\text{KLD},\mathcal{P}(\boldsymbol{\Theta})} = \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \left\{ \mathsf{E}_\Pi(L+\rho) - w^{-1}\log \int_{\boldsymbol{\Theta}} \exp\left(\frac{\rho(\boldsymbol{\Theta})}{w^{-1}}\right) d\pi(\boldsymbol{\Theta}) \right\}.$$

As this example reveals, the KLD penalizes the adversary for deviations ρ from L proportionally to the prior belief π. As pointed out above, this means that π plays the role of the adversary's cost function in the dual form. Jensen's inequality

also gives a lower bound on the penalty associated with the KLD. While coarse, this bound is perhaps more interpretable than the exact form:

$$w^{-1} \log \int_{\boldsymbol{\Theta}} \exp\left\{\frac{\rho(\boldsymbol{\theta})}{w^{-1}}\right\} d\pi(\boldsymbol{\theta}) \geq \int_{\boldsymbol{\Theta}} \rho(\boldsymbol{\theta}) d\pi(\boldsymbol{\theta}).$$

Specifically, it reveals that regularization via the KLD implies a perturbation penalty that is *at least as costly* as a linear penalization. The linear penalty (i.e., $\int_{\boldsymbol{\Theta}} \rho(\boldsymbol{\theta}) d\pi(\boldsymbol{\Theta})$) is a useful benchmark to compare against: Linear penalties punish the adversary simply by weighting its perturbation ρ with the prior. Note that this conforms with our interpretation of a prior: at least a priori, we believe that the value of L matters most in regions of high prior mass. Naturally then, it should be in these regions that it is most expensive for an adversarial agent to increase our loss and harm our inference procedure by means of a perturbation.

Having derived the perhaps most canonical form of the dual with $D = \text{KLD}$, a natural next question is what we can say for the more general case where D is chosen to be *any* f-divergence.

Example 2.3 (f-divergence)**.** For any f-divergence D based upon a lower semicontinuous convex function $f : \mathbb{R} \to (-\infty, \infty]$ with $f(1) = 0$, we have

$$\mathcal{G}_{L, D_f, \mathcal{P}(\boldsymbol{\Theta})} = \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \left\{ \mathsf{E}_\Pi(L + \rho) - \inf_{b \in \mathbb{R}} \left[\int_{\boldsymbol{\Theta}} f^\star(\rho(\boldsymbol{\theta}) - b) d\pi(\boldsymbol{\theta}) + b \right] \right\},$$

where $f^\star(t) = \sup_{t' \in \text{dom}(f)} \{t \cdot t' - f(t')\}$

Note for any f as in the above example, it also holds that $f^\star(t) \geq t$ and so immediately we get

$$\inf_{b \in \mathbb{R}} \left[\int_{\boldsymbol{\Theta}} f^\star(\rho(\boldsymbol{\theta}) - b) d\pi(\boldsymbol{\theta}) + b \right] \geq \int_{\boldsymbol{\Theta}} \rho(\boldsymbol{\theta}) d\pi(\boldsymbol{\theta}), \tag{2.6}$$

showing that what held true for the KLD also holds true for general f-divergences: the linear penalization is a lower bound on the cost function implied by any f-divergence. The linear penalty term also sometimes re-surfaces directly and rather elegantly. As the next example shows, this happens the case of the χ^2-divergence, for which we can easily solve $\inf_{b \in \mathbb{R}} \left[\int_{\boldsymbol{\Theta}} f^\star(\rho(\boldsymbol{\theta}) - b) d\pi(\boldsymbol{\theta}) + b \right]$.

Example 2.4 (χ^2-divergence)**.** If $D = w^{-1} \cdot \chi^2$ for some $\lambda \geq 0$ then the dual problem is

$$\mathcal{G}_{L, \chi^2, \mathcal{P}(\boldsymbol{\Theta})} = \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \left\{ \mathsf{E}_\Pi(L + \rho) - \int_{\boldsymbol{\Theta}} \rho(\boldsymbol{\theta}) d\pi(\boldsymbol{\theta}) - \frac{1}{4w^{-1}} \text{Var}_\pi(\rho) \right\}. \tag{2.7}$$

Here, the role of w^{-1} is particularly interesting: in particular, the example shows that for larger values of w^{-1}, the variation of the perturbation function ρ (as measured by $\mathrm{Var}_\pi(\rho)$) is penalized less and less. The example also shows that for the case of the χ^2-divergence, we can exactly quantify the slack term arising from (and therefore the looseness of) the bound in (2.6).

Integral Probability Metrics (IPMs)

Choosing f-divergences constitute an appealing way of penalizing how far a posterior deviates from the prior. This is due to the fact that f-divergences requires absolute continuity—so that the posterior is forced to be supported wherever the prior is. This is not necessarily true for other divergences. For example, the family of Integral Probability Metrics (IPMs) that we study next do not have this property, and thus can be considered to be weaker regularizers.

Definition 2.7 (Integral Probability Metric)**.** For a set of functions $\mathcal{H} \subseteq \mathcal{F}_b(\mathbf{\Theta})$, the IPM between $q, \pi \in \mathcal{P}(\mathbf{\Theta})$ is

$$d_{\mathcal{H}}(q, \pi) = \sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_q[h] - \mathbb{E}_\pi[h] \right\}. \tag{2.8}$$

IPMs have often been studied for theoretical interest in Machine Learning as they define metrics over probability spaces (Müller, 1997), with one famous example being the 1-Wasserstein distance (Villani, 2008). Another example of an IPM is the kernel-based Maximum Mean Discrepancy, which is comparatively easy to compute in practice.

The downside of IPMs is that for a general class $\mathcal{H}$, they cannot be easily computed. In spite of this, IPMs have recently been popularized by work on Generative Adversarial Networks, where deep neural networks have played the role of $\mathcal{H}$ with various kinds of parametrizations (Arbel et al., 2018; Li et al., 2017; Arjovsky et al., 2017; Mroueh et al., 2018; Mroueh and Sercu, 2017). Beyond that, they have also been used in the Wasserstein Autoencoder (Tolstikhin et al., 2018). This case is of special interest to us because the Wasserstein Autoencoder can be shown to be a special case of the RoT, as we will demonstrate later. Before we can do this, we will first apply our general result to the case where D is an IPM.

Example 2.5 (Integral Probability Metric)**.** For a set of functions $\mathcal{H} \subseteq \mathcal{F}_b(\mathbf{\Theta})$, and

the IPM given by $d_{\mathcal{H}}$, if $D = d_{\mathcal{H}}$, then

$$\mathcal{G}_{L,\text{IPM},\mathcal{P}(\boldsymbol{\theta})} = \sup_{\rho \in w^{-1} \cdot \mathcal{H}} \left\{ \mathsf{E}_{\Pi}(L + \rho) - \int_{\boldsymbol{\theta}} \rho(\boldsymbol{\theta}) d\pi(\boldsymbol{\theta}) \right\},$$

where we have used the notation $w^{-1} \cdot \mathcal{H} = \{w^{-1} \cdot h : h \in \mathcal{H}\}$.

Notice that once again, the linear penalization makes an appearance. However, the dual of the RoT with an IPM-regularizer imposes a particularly *strong* penalty on the adversary: Any perturbation ρ that is not in the set $w^{-1} \cdot \mathcal{H} = \{w^{-1} \cdot h : h \in \mathcal{H}\}$ incurs an infinitely large penalty. Translating this to the primal form, we can conclude that an IPM in general will be a relatively *weak* regularizer. That being said, the choice of w allows us to regulate this somewhat, and indeed $w \to 0$ (or equivalently, $w^{-1} \to \infty$) reduces the constraints on the penalty to an arbitrary degree.

Having stated the result for general IPMs, we can now use it to demonstrate that the well-known Wasserstein Autoencoder (Tolstikhin et al., 2018) can be written as the dual of a particular RoT problem with an IPM-regularizer. Before we can do so, we first need to define the Wasserstein Autoencoder.

Definition 2.8 (Wasserstein Autoencoder)**.** Consider $\boldsymbol{\theta} = \mathcal{Z} \times \mathcal{X}$, where $\mathcal{Z}$ is typically referred to as a *latent* space. Take $G : \mathcal{Z} \to \mathcal{X}$ to be a fixed mapping, and suppose we have a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For $\boldsymbol{\theta} = (z, x)$, and a fixed measure $P_X \in \mathcal{P}(\mathcal{X})$, define

$$\begin{aligned} L(\boldsymbol{\theta}) &= c(G(z), x) \\ \Pi &= \{q \in \mathcal{P}(\boldsymbol{\theta}) : q(\mathcal{Z} \times A) = P_X(A), A \text{ measurable}\}. \end{aligned} \tag{2.9}$$

The Wasserstein Autoencoder objective is given by

$$\inf_{G, q \in \Pi} \left\{ \int_{\boldsymbol{\Theta}} c(G(z), x) dq(z, x) + w^{-1} D(F_{\#} q, \pi) \right\},$$

where D is a divergence, $F : \boldsymbol{\Theta} \to \mathcal{Z}$ is a projection mapping defined as $F(z, x) = z$ so that $F_{\#} q$ is the marginal of q on $\mathcal{Z}$; and $\pi \in \mathcal{P}(\mathcal{Z})$ is a prior distribution over $\mathcal{Z}$.

Without going into too much detail about variational autoencoders, the Wasserstein Autoencoder (WAE) problem consists in the usual optimization over two different objects: the so-called decoder G, and the so-called encoder q. In the literature on autoencoders, the term $\int_{\boldsymbol{\Theta}} c(G(z), x) dq(z, x)$ will often be referred to

as reconstruction cost, while the divergence term $D(F_{\#}q, \pi)$ usually is motivated by ensuring smoothness of the encoder on the latent space $\mathcal{Z}$. Note that in practice, the choice of D as the Maximum Mean Discrepancy for the WAE as introduced in Tolstikhin et al. (2018) was somewhat adhoc, and based mostly on computational considerations. Since then, other work has considered the Wasserstein-1 distance for D instead (Patrini et al., 2020; Zhang et al., 2019). Note that both choices for D are IPMs, so that the duality result we are about to present for the WAE holds regardless of the particular choice of D.

While the objective in Definition 2.8 looks similar to those of the RoT, there are two minor complications: the additional optimization over G, as well as the fact that D is defined only over $\mathcal{Z}$—rather than $\mathbf{\Theta}$. Both are easily addressed: we simply fix any arbitrary G, take $\tilde{\mathcal{H}} = \{f(x, z) = h(z) : h \in \mathcal{H}\}$, and define $\tilde{D} = d_{\tilde{\mathcal{H}}}$ as well as $\tilde{\pi} = \pi \times \nu$ where $\nu \in \mathcal{P}(\mathcal{X})$ is an arbitrary probability measure. It then follows that the WAE objective is precisely $\mathcal{G}_{L,\tilde{D},\Pi}$ with prior $\tilde{\pi}$. We now invoke our main result, noting that Π is closed and convex to derive the dual:

$$\mathcal{G}_{L_G,\tilde{D},\Pi} = \sup_{\rho \in \mathcal{H}} \left\{ \mathsf{E}_\Pi \left[L + \rho \right] - \int_{\mathcal{Z}} \rho(z) d\pi(z) \right\},$$

where both L and Π are as defined in Definition 2.8. So far, we have kept the decoder G fixed. What if we allow it to vary? Interestingly, the minimization problem over G can now be interpreted as a min-max problem:

$$\inf_G \mathcal{G}_{L_G,\tilde{D},\Pi} = \inf_G \sup_{\rho \in \mathcal{H}} \left\{ \mathsf{E}_\Pi \left[L + \rho \right] - \int_{\mathcal{Z}} \rho(z) d\pi(z) \right\}.$$

In other words, the function G is minimizing the worst case reconstruction loss as altered by an adversary who can choose perturbations in $\mathcal{H}$. The magnitude of these perturbations is controlled via a penalty based on the prior π. This finding allows us to reinterpret the Wasserstein Autoencoder (WAE) of Tolstikhin et al. (2018) not only as a member of the RoT—and therefore a Bayes-like method—but also as an adversarially robust procedure.

Chapter 3

Frequentist consistency

Summary: In this chapter, we study the frequentist properties of variational forms for the RoT. It is generally quite difficult to prove frequentist properties for RoT posteriors. The reason for this is relatively simple: unlike the Gibbs posterior or standard Bayesian setting where $D = \text{KLD}$ and $\Pi = \mathcal{P}(\boldsymbol{\Theta})$, RoT posteriors generally will not have an analytically available solution. While this problem has been addressed for the standard variational approximation setting where $D = \text{KLD}$, $\Pi = \mathcal{Q}$, these proofs rely on the fact that one is forming an approximation to an analytically available object. In contrast, one loses this approximating interpretation if one chooses $D \neq \text{KLD}$. For these reasons, the techniques developed here to prove consistency for the RoT are very much unlike other work on the large sample behaviour of generalized Bayesian objects. Specifically, we will need to introduce the notions of Γ-convergence and associated concepts from functional analysis. And even with these tools, it will not be possible to derive the speed of convergence—all we will be able to find is the limit.

Before we can state and prove the results in this section formally, we need to cover a number of concepts and definitions relating to Γ-convergence and functional analysis that will not be used in the remainder of the thesis. Most of these will be taken from Dal Maso (2012), which provides an excellent introductory reference into Γ-convergence.

3.1 Γ-convergence

For a topological space X, let $N(x)$ the be set of all open neighbourhoods of $x \in X$. Further, take F_n to be a sequence of functions so that $F_n : X \to \overline{\mathbb{R}}$, where

$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Because it is not instructive for a reader without vested interest in advanced functional analysis, we defer the general Definition of Γ-convergence to Appendix A.1. For the purposes of our proof, it is more important that once a sequence of function $\{F_n\}_{n\in\mathbb{N}}$ can be shown to be equi-coercive and Γ-convergent, we can conclude that its minimizers converge. This raises two questions: what does it mean for a function to be equi-coercive, and what conditions do we need to establish in order to conclude that a function Γ-converges?

Before we can answer these questions in detail, it is worth recalling an equivalence between lower-semicontinuity of a function on the one hand, and closedness of a function's sub-level sets on the other.

Proposition 3.1. F is lower-semicontinuous if and only if for all $t \in \mathbb{R}$, $\{x \in X : F(x) \leq t\}$ is closed.

Next, we can state a result that we shall use to prove the Γ-convergence of functions.

Proposition 3.2 (Remark 5.5, (Dal Maso, 2012))**.** If $\{F_n\}_{n\in\mathbb{N}}$ is an increasing sequence of lower-semicontinuous functions which converges pointwise to a function F, then F is lower-semicontinuous and F_n Γ-converges to F.

As Proposition 3.1 shows, the sub-level sets $\{x \in X : F(x) \leq t\}$ are closed whenever F is lower-semicontinuous. Another property we will need these sub-level sets to have for some of our proofs is compactness, a property that is also known as coerciveness of F.

Definition 3.1 (coercive function)**.** A function $F : X \to \overline{\mathbb{R}}$ is coercive if the closure of $\{x \in X : F(x) \leq t\}$ is (countably) compact in X for every $t \in \mathbb{R}$.

In contrast to coerciveness, equi-coerciveness is a property of a sequence of functions $\{F_n\}_{n\in\mathbb{N}}$ rather than a property of a single function.

Definition 3.2 (equi-coerciveness)**.** A function $\{F_n\}_{n\in\mathbb{N}}$ is equi-coercive on X if for every $t \in \mathbb{R}$, there is a closed (countably) compact set $K_t \subset X$ so that $\{x \in X : F_n(x) \leq t\} \subset K_t$ for every $n \in \mathbb{N}$

Because it is difficult to prove this property directly from its definition, we shall use the following equivalent condition.

Proposition 3.3 (Proposition 7.7, (Dal Maso, 2012))**.** $\{F_n\}_{n\in\mathbb{N}}$ is equi-coercive if and only if there exists a coercive, lower-semicontinuous function Ψ so that $F_n \geq \Psi$ on X for all $n \in \mathbb{N}$.

3.2 Preliminaries and Notation

Throughout, $(\boldsymbol{\Theta}, \|\cdot\|)$ will be a normed space of finite dimension. For virtually all cases of practical interest, we will be in the situation where $\boldsymbol{\Theta} \subset \mathbb{R}^d$. Unlike in previous chapters, the current chapter only considers the case where $\Pi = \mathcal{Q}$ for some parameterized set of Lebesgue densities $\mathcal{Q}$ on $\mathcal{P}(\boldsymbol{\Theta})$. For its intimate relationship with Variational Inference, we will often refer to this as the the *variational setting*. Further to that, our study of consistency relies on additive losses for L, so that throughout, we will only study RoT posteriors with losses of the form

$$L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i).$$

Accordingly and as introduced in Definition 2.3, we will use the notation $P(\ell, D, \mathcal{Q})$ instead of $P(L, D, \mathcal{Q})$.

As in the remainder of the thesis, certain notational liberties are taken. In particular, if the measure $\nu \in \mathcal{P}(\boldsymbol{\Theta})$ admits a density q_ν on $\boldsymbol{\Theta}$, we often write $q_\nu \in \mathcal{P}(\boldsymbol{\Theta})$ to mean that $\nu \in \mathcal{P}(\boldsymbol{\Theta})$. Importantly for the current chapter, we extend this to statements about convergence. For example, whenever we write $q_n \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}$, this means that the sequence of measures $\nu_n \in \mathcal{P}(\boldsymbol{\Theta})$ with densities q_n converges weakly to the measure $\delta_{\boldsymbol{\theta}^*} \in \mathcal{P}(\boldsymbol{\Theta})$.

Unlike in previous chapters, we are interested in frequentist properties, which necessitates a stochastic treatment of the observation sequence. In particular, it is assumed that the fixed numbers $x_i \in \mathcal{X}$ are realizations of random variables $\boldsymbol{x}_i : \Omega \to \mathcal{X}$. In principle, our treatment allows $\boldsymbol{x}_{1:n}$ to be non-identically distributed, and even to exhibit certain forms of dependency. In other words, the general form of our result is not limited to the case where $\boldsymbol{x}_{1:n}$ are independent and identically distributed (i.i.d.) copies of the random variable $\boldsymbol{x}_1$. Rather, we only require that a strong law of large numbers holds: Specifically, we need that for some probability measure μ, $\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)$ converges μ-almost surely (μ-a.s.) to $\mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]$ as $n \to \infty$. Throughout this chapter, we will denote this type of convergence by writing $\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) \overset{\mu-a.s.}{\longrightarrow} \mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]$ as $n \to \infty$; and the mode of convergence will be assumed to be pointwise (in $\boldsymbol{\Theta}$) unless stated otherwise. The interpretation of μ will depend on the specific application or problem, but it will be helpful to think of it as the stationary distribution for $\boldsymbol{x}_{1:n}$ as $n \to \infty$. For the special case where $\boldsymbol{x}_i$ are i.i.d., this means that μ is simply the law of $\boldsymbol{x}_1$.

Throughout, we will hope that our posteriors collapse to a particular parameter value defined as $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]$, sometimes called the *population-*

optimal value of $\boldsymbol{\theta}$.

3.3 Proof strategy and high-level summary

The goal of this chapter is to provide a generic proof strategy applicable to most forms of the RoT using minimal assumptions. To this end, the proofs cannot assume that the solution set $P(\ell, D, \mathcal{Q})$ consists of an analytically available singleton. In plain English: we have no idea what our posterior looks like, and the best we can do in practice is numerically approximate it. Due to these extraordinarily challenging conditions, the results are weaker than what one would expect for the standard Bayesian setting or standard variational approximations[1], and we have to rely on a somewhat exotic theoretical tool kit: we deploy the machinery of Γ-convergence, which was also used in the context of standard variational approximations by Lu et al. (2017) and Wang and Blei (2018). Roughly speaking, the role of Γ-convergence in the present work is as follows: If a sequence of functions F_n Γ-converges to a function F, then the sequence $q_n = \arg\inf_{q\in\mathcal{Q}} F_n(q)$ of its minimizers converges to the minimizer of F under mild regularity conditions. Provided that we can prove the minimizer of F to be a point mass at $\boldsymbol{\theta}^*$, this would prove frequentist consistency.

To this end, the current chapter studies the (stochastic) sequence of objectives associated with $P(\ell, D, \mathcal{Q})$ as well as their minimizers. Thus, we define for our convenience a number of objects that will be used in the remainder of the chapter:

$$F_n(q) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i)\right] + \frac{1}{n}D(q||\pi),$$
$$q_n = \arg\inf_{q\in\mathcal{Q}} F_n(q).$$

Inspecting F_n, the intuition behind our proof becomes obvious: mild regularity conditions should ensure that $\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i) \overset{\mu-a.s.}{\longrightarrow} \mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]$ as $n \to \infty$ for an appropriate probability measure μ on $\mathcal{X}$. Similarly, it is usually reasonable to expect that for well-behaved losses, $\widehat{\boldsymbol{\theta}}_n = \arg\min\{\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i)\} \to \boldsymbol{\theta}^* = \arg\min\{\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\}$ as $n \to \infty$, μ-almost surely. Intuitively then, one expects the sequence q_n to converge in distribution to $\delta_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})$ under mild regularity conditions, μ-almost surely. In other words, the remainder of the chapter will aim to show that for $\mathcal{F}_b(\boldsymbol{\Theta})$ the set

[1] in particular, the results of this chapter say nothing about the speed of convergence

of continuous, bounded functions from $\boldsymbol{\Theta}$ to $\mathbb{R}$,

$$\begin{aligned}&\mathbb{P}_\mu\left(q_n(\boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})\right)\\ &= \mathbb{P}_\mu\left(\forall f \in \mathcal{F}_b(\boldsymbol{\Theta}) : \lim_{n\to\infty}\left\{\mathbb{E}_{q_n(\boldsymbol{\theta})}\left[f(\boldsymbol{\theta})\right] - f(\boldsymbol{\theta}^*)\right\} = 0\right) = 1. \qquad (3.1)\end{aligned}$$

However, any direct way of showing that this intuition holds would require establishing Γ-convergence of $q \mapsto F_n(q)$ to $q \mapsto \mathbb{E}_{q(\boldsymbol{\theta})}\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$. Unfortunately, the stochasticity of F_n (introduced via $x_{1:n}$) makes it hard to prove this directly. The key insight of the proof technique presented here is a way to circumvent these technical complications: We simplify the problem, and then analyze the *deterministic* sequence of functions $\overline{F}_n$ and its minimizers $\overline{q}_n$ given by

$$\begin{aligned}\overline{F}_n(q) &= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right] + \frac{1}{n} D(q||\pi)\\ \overline{q}_n &= \arg\inf_{q\in\mathcal{Q}} \overline{F}_n(q).\end{aligned}$$

For this new objective $\overline{F}_n(q)$, establishing Γ-convergence to $\mathbb{E}_{q(\boldsymbol{\theta})}\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$ is much simpler. The last and most important part of the proofs will then be to show that the sequences q_n and $\overline{q}_n$ become arbitrarily close as $n \to \infty$ (μ-almost surely). More precisely, we will show that the sequence $\{q_n\}_{n=1}^\infty$ constitutes a sequence of ε_n-minimizers of $\overline{F}_n$, i.e.

$$\overline{F}_n(q_n) \leq \inf_{q\in\mathcal{Q}} \overline{F}_n(q) + \varepsilon_n,$$

where ε_n is a stochastic sequence converging to zero (μ-almost surely). This—together with Γ-convergence and equi-coerciveness of $\overline{F}_n$—suffices to show that as desired, eq. (3.1) holds. To summarize, we will show consistency of the RoT for additive losses and $\Pi = \mathcal{Q}$ in three steps:

(S1) Establishing that $\overline{F}_n$ is equi-coercive and Γ-converges to $\mathbb{E}_{q(\boldsymbol{\theta})}\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$, from which it follows that $\overline{q}_n \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}$ as $n \to \infty$;

(S2) Showing that the minimizers q_n of the stochastic sequence F_n are ε_n-minimizers of $\overline{F}_n$, so that $\overline{F}_n(q_n) \leq \inf_{q\in\mathcal{Q}} \overline{F}_n(q) + \varepsilon_n$ for sufficiently large n and μ-almost surely;

(S3) Proving that ε_n goes to zero μ-almost surely as $n \to \infty$. This together with the first two findings finally implies that as desired, $q_n \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}$ holds μ-almost surely.

3.4 Standing assumptions

First, we present a collection of harmless assumptions that will generally be satisfied for most problems of practical interest. We start by providing mild regularity conditions on the loss and its interaction with the data-generating mechanism.

Assumption 3.1. For the loss $\ell : \boldsymbol{\Theta} \to \mathbb{R}$, the law $\mu \in \mathcal{P}(\mathcal{X})$, and the prior $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, it holds that

(a) The minimizers $\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right\} \in \boldsymbol{\Theta}$ exist for all sufficiently large n so that $\left| \min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right\} \right| < \infty$ holds μ-almost surely;

(b) The loss satisfies a strong law of large numbers, i.e. $\frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \overset{\mu-a.s.}{\longrightarrow} \mathbb{E}_\mu \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) \right]$;

(c) The minimizer $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_\mu \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) \right]$ exists and is unique;

(d) The expected loss and the expected prior loss are finite, i.e. $\mathbb{E}_\mu \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) \right] < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbb{E}_\pi \left[\mathbb{E}_\mu \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) \right] \right] < \infty$.

Remark 3.1. The interpretation of μ for the case where $x_i \overset{iid}{\sim} \boldsymbol{x}_1$ is clearly that of the probability measure corresponding to $\boldsymbol{x}_1$. Things are perhaps less obvious in the dependent case. For example, suppose that $\ell(\boldsymbol{\theta}, x_i) = \ell(\boldsymbol{\theta}, x_i; x_{i-1})$ is the negative log likelihood of a sequentially dependent model (like a first order autoregressive process) and that this model accurately describes how $x_i | x_{i-1}$ was generated. Then—provided that a strong law of large numbers holds—μ is essentially the stationary distribution of the process. Dependencies like these are notationally suppressed for readability, but would not affect any of the results derived in the current chapter unless in those cases where independence of observations is specifically required.

Remark 3.2. One may also be interested in the convergence properties of posteriors built with a sequence of heterogeneous losses $\{\ell_i(\boldsymbol{\theta}, x_i)\}_{i=1}^\infty$ where $\ell_i \neq \ell_j$ for some i, j. In this case, all derived convergence results follow after an easy adaptation of the above assumption. Specifically, one requires that the minimizers exist for $\frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{\theta}, x_i)$ instead. Further, one requires that there exists some function $\widetilde{\ell} : \boldsymbol{\Theta} \times \mathcal{X} \to \mathbb{R}$ such that $\frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{\theta}, x_i) \overset{\mu-a.s.}{\longrightarrow} \mathbb{E}_\mu \left[\widetilde{\ell}(\boldsymbol{\theta}, \boldsymbol{x}) \right]$. Replacing the old convergence requirement in the above Assumption by the new one and $\mathbb{E}_\mu \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) \right]$ by $\mathbb{E}_\mu \left[\widetilde{\ell}(\boldsymbol{\theta}, \boldsymbol{x}) \right]$ completes the adaptation to heterogeneous losses.

While the previous standing assumption relates to the loss function ℓ and the stochasticity in the data $x_{1:n}$, the next one relates to the interplay between the prior π, the divergence D, and the variational family $\mathcal{Q}$.

Assumption 3.2. For the prior $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, the statistical divergence $D : \mathcal{P}(\boldsymbol{\Theta})^2 \to \mathbb{R}_{\geq 0}$, and the variational family $\mathcal{Q} \subseteq \mathcal{P}(\boldsymbol{\Theta})$, it holds that

(a) $\mathcal{Q}$ is a collection of Lebesgue densities on $\boldsymbol{\Theta}$ parameterized by $\boldsymbol{\kappa} \in \boldsymbol{K}$. Further, for any $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, there exist sequences $\{\boldsymbol{\kappa}_k\}_{k=1}^{\infty}$ of variational parameters so that $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_k) \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}(\boldsymbol{\theta})$ as $k \to \infty$. Since $q(\cdot|\boldsymbol{\kappa}_k)$ are densities, this means that $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_k) \to \infty$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, and $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_k) \to 0$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$;

(b) It holds that $\pi \in \mathcal{Q}$;

(c) The prior π, regularizer D, and variational family $\mathcal{Q}$ are chosen so that for all $q \in \mathcal{Q}$, $D(q||\pi) < \infty$.

(d) Both $\boldsymbol{\kappa} \mapsto D(q(\boldsymbol{\theta}|\boldsymbol{\kappa})\|\pi)$ and $\boldsymbol{\kappa} \mapsto \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$ are lower semi-continuous with respect to the Euclidean metric. Further, $\left\{\boldsymbol{\kappa} : \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right] \leq t\right\}$ is bounded for all $t \in \mathbb{R}$.

Remark 3.3. The role of Assumption 3.2 (a) is to ensure that an analysis of frequentist consistency is possible in the first place: frequentist consistency necessarily implies that the variational family $\mathcal{Q}$ be such that it can get arbitrarily close to dirac measures on $\boldsymbol{\Theta}$.

Remark 3.4. Assumption 3.2 (b) is purely technical, and its only function is to guarantee that the RoT posteriors do not become *worse* than the prior belief after observing data. Since this is a situation that is very unlikely to occur in practice for even small sample sizes, in practice one will obtain frequentist consistency even if one removes π from $\mathcal{Q}$.

Remark 3.5. Assumption 3.2 (c) is the variational equivalent to the canonical requirement that $\pi(\boldsymbol{\theta}) > 0$ in a neighbourhood of $\boldsymbol{\theta}^*$ that is imposed for traditional Bayesian consistency proofs. Its role is to ensure that concentration on the dirac measure at $\boldsymbol{\theta}^*$ is possible. In fact, the requirement is slightly stronger than necessary for consistency: It would suffice to require that there exists a sequence $p_n \in \mathcal{Q}^{\boldsymbol{\theta}}$ so that (i) $p_n \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}$ and (ii) $D(p_n||\pi) < \infty$ for all finite n. If $D = \mathrm{KLD}$, it is clear that this latter requirement is satisfied if and only if $\mathcal{Q}$ satisfies Assumption 3.2 (a) and

$\pi(\boldsymbol{\theta}) > 0$ in a neighbourhood of $\boldsymbol{\theta}^*$. Indeed, this equivalence holds over the wider class of choices for divergences D in the set

$$\{D : D \text{ satisfies Assumption 3.2 (c)}\} \cap$$
$$\{D(q||\pi) = \infty \text{ for all } q \text{ that are not absolutely continuous w.r.t. } \pi\}.$$

Examples of divergences in this set are the KLD, the α-divergence, Rényi's α-divergence as well as the family of f-divergences.

Remark 3.6. In technical terms, Assumption 3.2 (d) will ensure that the functional F_n is equi-coercive (in $\boldsymbol{K}$), because the lower semi-continuity requirement together with boundedness of the sub-level sets implies that $\left\{\boldsymbol{\kappa} : \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right] \leq t\right\}$ is a compact set for all t. Note that continuity implies lower semi-continuity, so that a sufficient condition for part (d) would be that $\boldsymbol{\kappa} \mapsto D(q(\boldsymbol{\theta}|\boldsymbol{\kappa})\|\pi)$ and $\boldsymbol{\kappa} \mapsto \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$ are continuous, and that $\boldsymbol{K}$ is a compact set. For most problems of interest, continuity in $\boldsymbol{\kappa}$ of this form is a reasonable assumption. Similarly, compactness is not a prohibitive assumption: any non-compact $\boldsymbol{K}$ can be reparameterized component-wise with some continuous one-to-one function $r : \boldsymbol{K} \to [0, 1]$ with continuous inverse to ensure compactness of the new variational parameter space. For example, if $\boldsymbol{K} = \mathbb{R}^d$, then one can reparameterize component-wise via

$$r(\boldsymbol{\kappa}) = \left(\frac{1}{1 + e^{-\boldsymbol{\kappa}_1}}, \frac{1}{1 + e^{-\boldsymbol{\kappa}_2}}, \dots \frac{1}{1 + e^{-\boldsymbol{\kappa}_d}}\right)^T.$$

Since the inverse function r^{-1} exists and is continuous, and defining the new compactified space $\boldsymbol{K}^r = \{\boldsymbol{\kappa}^r = r(\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in \boldsymbol{K}\}$, it also holds that $\boldsymbol{\kappa}^r \mapsto D(q(\boldsymbol{\theta}|\boldsymbol{\kappa}^r)\|\pi)$ and $\boldsymbol{\kappa}^r \mapsto \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa}^r)}\left[\mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$ are continuous.

3.5 Main Results

Having outlined the general thrust of the theoretical argument in steps (S1)—(S3), we now state the main results without any formal derivations. Instead, detailed derivations are given and elaborated upon in the next section. Broadly speaking, there are two classes of results we derive: the first type of result relies only on conditions on the loss function, and therefore imposes conditions that can typically be verified very easily (Theorem 3.1). While the advantage of the first result is that it is easy to verify and does not require data to be independently and identically distributed (i.i.d.), its disadvantage is that the conditions deployed therein require that either the loss itself or its derivatives are bounded in certain ways.

This is why we also derive a second type of result, whose strength is that the boundedness conditions are replaced by mild regularity conditions on moments and local continuity (Theorem 3.2). While these requirements are more general, their drawback is that they also tend to be harder to verify. Additionally, we require the data to be i.i.d. in order to deploy these types of results.

Theorem 3.1. Suppose that Assumptions 3.1 and 3.2 hold. Further, suppose one of the following three conditions holds:

(i) $\sup_{x_i \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \leq M$ for some constant M, μ-almost surely;

(ii) $\ell(\boldsymbol{\theta}, \boldsymbol{x})$ and $\mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]$ are both jointly continuous in $\boldsymbol{\theta}$ and $\boldsymbol{x}$, μ-almost surely; and both $\mathcal{X}$ and $\boldsymbol{\Theta}$ are compact;

(iii) The function $\boldsymbol{\theta} \mapsto \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$ is continuously differentiable on $\boldsymbol{\Theta}$ μ-almost surely and for all n large enough. Additionally, it holds that $\sup_{x_i \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} |\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, x_i)| < \infty$, or that $\sup_{x_{1:n} \in \mathcal{X}^n, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \left|\nabla_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right| < \infty$, or that $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left|\nabla_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right| < \infty$ μ-almost surely and for all n large enough;

Then $q_n \xrightarrow{D} \delta_{\boldsymbol{\theta}^*}$ μ-almost surely.

Before we can introduce a result that replaces the boundedness conditions on the loss function with moment and continuity requirements, we need a very mild additional requirement for the variational family.

Assumption 3.3. The variational family $\mathcal{Q}$ is such that

(i) for each $q \in \mathcal{Q}$, there exists M so that for all $\|\boldsymbol{\theta}\|_2 > M$, q is decreasing as we move in any direction pointing outside of the set $\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta}\|_2 \leq M\}$;

(ii) for any $q \in \mathcal{Q}$, and on any compact set S of $\boldsymbol{\Theta}$, we can lower bound q on S so that $\inf_{\boldsymbol{\theta} \in S} q(\boldsymbol{\theta}) > 0$.

It is not difficult to see that this requirement will be met by virtually all variational families of practical interest, including fully factorized normals (see Lemma 3.8), multivariate normals, student's t-distributions, uniform, and Beta distributions to name but a few (see Lemma 3.9 and Corollary 3.3).

Theorem 3.2. Suppose that Assumptions 3.1, 3.2, and 3.3 hold, that $x_{1:n} \overset{i.i.d.}{\sim} \mu$, and $\mathbb{E}_\mu \left[\mathbb{E}_\pi \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|\right]^2\right] < \infty$. Further, for $A \subset \boldsymbol{\Theta}$ a compact set containing $\boldsymbol{\theta}^*$, suppose that $\mathbb{E}_\mu \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x}) \cdot 1_A(\boldsymbol{\theta})|^{2+\delta}\right] < \infty$ for all $\boldsymbol{\theta} \in A$ for some $\delta > 0$, that $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{\theta}, x_i)$ is μ-almost surely continuous on A, and that $\boldsymbol{\theta} \mapsto \mathbb{E}_\mu \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|\right]$ is continuous on A. Then $q_n \xrightarrow{D} \delta_{\boldsymbol{\theta}^*}$ μ-almost surely.

3.6 General Derivations for Technical results

Having set out the Assumptions to be used in the most general form of our derivation, we now follow the steps (S1)—(S3) in the remainder of this section. For each subsection, we provide a short summary of the proof strategy. Whenever we write a mapping in terms of q, this is a shorthand: all analysis happens on the level of the parametric space into which q is embedded. In other words, whatever statement we make or prove about the behaviour of a function $F(q)$ will be a statement about the function $\boldsymbol{\kappa} \to F(q(\cdot|\boldsymbol{\kappa}))$.

3.6.1 Establishing convergence for the auxiliary objective (S1)

Role of (S1) for the overall proof: by virtue of Assumptions 3.1 and 3.2 (a), (c), and (d) one can show that the sequence of functions $\overline{F}_n$ is equi-coercive (Lemma 3.1). Secondly, one can show that $\overline{F}_n$ Γ-converges to $\mathbb{E}_q\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$ (Lemma 3.2). Together, this implies that one of the main workhorses for proving consistency can be deployed. Specifically, Corollary 7.24 in (Dal Maso, 2012) holds, proving that $\overline{q}_n \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}$ (Corollay 3.1).

Lemma 3.1 (Equi-coerciveness)**.** If Assumption 3.2 (d) holds, $\{\overline{F}_n\}_{n=1}^{\infty}$ is equi-coercive on the space $\boldsymbol{K}$ associated with $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in \boldsymbol{K}\}$.

Proof. $\{\overline{F}_n\}_{n=1}^{\infty}$ is equi-coercive if and only if there exists a coercive function Ψ for which $\Psi \leq \overline{F}_n$ for all n by Proposition 3.3. Clearly, since $D(\cdot\|\pi) \geq 0$, it holds that $\Psi(q) = \mathbb{E}_q\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$ yields a lower bound on $\overline{F}_n(q)$ for all n and all $q \in \mathcal{Q}$. All that remains is to prove that $\boldsymbol{\kappa} \mapsto \Psi(q(\cdot|\boldsymbol{\kappa}))$ is coercive, which holds by virtue of Assumption 3.2 (d) . □

Lemma 3.2 (Γ-convergence)**.** If Assumptions 3.2 (c) and (d) hold, $\overline{F}_n(q)$ Γ-converges to $\mathbb{E}_q\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$.

Proof. Clearly, it holds that $\overline{F}_n(q) \leq \overline{F}_{n-1}(q)$ so that $\overline{F}_n$ is a decreasing sequence of functions. Moreover, it is clear that pointwise (i.e. for fixed q), $\overline{F}_n(q) \to \mathbb{E}_q\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$. This holds trivially if $\overline{F}_n = \infty$ for all n and for the finite-valued case provided that $D(q||\pi) < \infty$, which holds by Assumption 3.2 (c). Taken together, this implies that $\overline{F}_n$ Γ-converges to the lower-semicontinuous envelope of its pointwise limit by Proposition 5.7 in Dal Maso (2012). Now since $\mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right]$ is itself lower semi-continuous on $\boldsymbol{K}$ by Assumption 3.2 (d) , it is its own lower-semicontinuous envelope. This completes the proof. □

Corollary 3.1 (Consistency)**.** If Assumptions 3.1 and 3.2 hold, then $\overline{q}_n \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}$, i.e. the minimizers of $\overline{F}_n$ weakly converge to a point mass at $\boldsymbol{\theta}^*$ as $n \to \infty$. Moreover, $\overline{F}_n(\overline{q}_n) \to \mathbb{E}_\mu [\ell(\boldsymbol{\theta}^*, \boldsymbol{x})]$ as $n \to \infty$.

Proof. This is a simple application of Corollary 7.24 in Dal Maso (2012): by Lemmas 3.1 and 3.2, $\overline{F}_n$ is both equi-coercive and Γ-convergent to $\mathbb{E}_q [\mathbb{E}_\mu [\ell(\boldsymbol{\theta}, \boldsymbol{x})]]$ (on $\boldsymbol{K}$). To complete the proof, we only need to show that the limiting functional has a unique minimizer. Thanks to Assumptions 3.1 and 3.2 (a), the infimum of $\mathbb{E}_q [\mathbb{E}_\mu [\ell(\boldsymbol{\theta}, \boldsymbol{x})]]$ over $\boldsymbol{K}$ is unique and given by the value of $\tilde{\boldsymbol{\kappa}} \in \mathrm{cl}(\boldsymbol{K})$ for which it holds that $\lim_{\boldsymbol{\kappa}_n \to \tilde{\boldsymbol{\kappa}}} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) = \delta_{\boldsymbol{\theta}^*}$. □

3.6.2 Showing that $\{q_n\}_{n\in\mathbb{N}}$ are ε_n-minimizers of $\overline{F}_n$ (S2)

Role of (S2) for overall proof: While step (S1) showed that the auxiliary objective's minimizer $\overline{q}_n$ behaves as we would hope, this is not useful unless we can establish that q_n and $\overline{q}_n$ are not too different as $n \to \infty$. To this end, in step (S2) we prove that the minimizers q_n are ε_n-minimizers of $\overline{F}_n$. Specifically, Lemma 3.3 guarantees that q_n corresponds to a μ-almost surely finite objective value for all n. This can directly be used to show that the sequence ε_n consists only of μ-almost surely finite-valued terms, which we use in Lemma 3.4 to derive an explicit form for a finite ε_n. Crucially, this form does *not* depend on q_n, but on $\overline{q}_n$. This will turn out to substantially ease remaining proofs: Unlike q_n which is a function of $x_{1:n}$ and thus is random, $\overline{q}_n$ is a fixed quantity.

Lemma 3.3. If Assumptions 3.1 and 3.2 (b) hold, then it also holds that

(i) $\mathbb{E}_\pi [\mathbb{E}_\mu [|\ell(\boldsymbol{\theta}, \boldsymbol{x})|]] < \infty$ and $\mathbb{E}_{\overline{q}_n} [\mathbb{E}_\mu [|\ell(\boldsymbol{\theta}, \boldsymbol{x})|]] < \infty$;

(ii) $\mathbb{E}_\pi [\mathbb{E}_\mu [\ell(\boldsymbol{\theta}, \boldsymbol{x})]] = \mathbb{E}_\mu [\mathbb{E}_\pi [\ell(\boldsymbol{\theta}, \boldsymbol{x})]]$ and $\mathbb{E}_{\overline{q}_n} [\mathbb{E}_\mu [\ell(\boldsymbol{\theta}, \boldsymbol{x})]] = \mathbb{E}_\mu \left[\mathbb{E}_{\overline{q}_n} [\ell(\boldsymbol{\theta}, \boldsymbol{x})]\right]$;

(iii) $\mathbb{E}_\pi [\ell(\boldsymbol{\theta}, x_i)] < \infty$ and $\mathbb{E}_{\overline{q}_n} [\ell(\boldsymbol{\theta}, x_i)] < \infty$ μ-almost surely.

for any $n \in \mathbb{N}$.

Proof. For simplicity, define $\overline{q}_0 = \pi$.

(i) First, observe that by Assumption 3.2 (b) and by the definition of the objective $\overline{F}_n$, it holds that

$$\infty > \mathbb{E}_{\overline{q}_0} [\mathbb{E}_\mu [\ell(\boldsymbol{\theta}, \boldsymbol{x})]] \geq \cdots \geq \mathbb{E}_{\overline{q}_n} [\mathbb{E}_\mu [\ell(\boldsymbol{\theta}, \boldsymbol{x})]] \geq \mathbb{E}_{\overline{q}_{n+1}} [\mathbb{E}_\mu [\ell(\boldsymbol{\theta}, \boldsymbol{x})]] \geq \ldots \quad (3.2)$$

It remains to show that this also holds if one takes the absolute value of the loss. Denote by $1_A(x)$ the indicator function that equals 1 if $x \in A$ and zero otherwise;

and by μ the probability measure as defined via Assumption 3.1. With this in mind, it holds that

$$\begin{aligned}
&\mathbb{E}_{\overline{q}_n}\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right] \\
&= \int_{\boldsymbol{\Theta}} \int_{\mathcal{X}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \mu(d\boldsymbol{x}) \overline{q}_n(\boldsymbol{\theta}) \\
&= \int_{\boldsymbol{\Theta}} \int_{\mathcal{X}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \cdot 1_{\{\ell(\boldsymbol{\theta}, \boldsymbol{x}) \leq 0\}}(\boldsymbol{x}) \mu(d\boldsymbol{x}) \overline{q}_n(\boldsymbol{\theta}) + \int_{\boldsymbol{\Theta}} \int_{\mathcal{X}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \cdot 1_{\{\ell(\boldsymbol{\theta}, \boldsymbol{x}) > 0\}}(\boldsymbol{x}) \mu(d\boldsymbol{x}) \overline{q}_n(\boldsymbol{\theta})
\end{aligned}$$

Because $\overline{q}_n \geq 0$, this also immediately implies that one can compute the absolute expectation via

$$\begin{aligned}
&\mathbb{E}_{\overline{q}_n}\left[\mathbb{E}_\mu\left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|\right]\right] \\
&= -\int_{\boldsymbol{\Theta}} \int_{\mathcal{X}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \cdot 1_{\{\ell(\boldsymbol{\theta}, \boldsymbol{x}) \leq 0\}}(\boldsymbol{x}) \mu(d\boldsymbol{x}) \overline{q}_n(\boldsymbol{\theta}) \\
&\quad + \int_{\boldsymbol{\Theta}} \int_{\mathcal{X}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \cdot 1_{\{\ell(\boldsymbol{\theta}, \boldsymbol{x}) > 0\}}(\boldsymbol{x}) \mu(d\boldsymbol{x}) \overline{q}_n(\boldsymbol{\theta}),
\end{aligned} \tag{3.3}$$

which will be finite if both integrals by themselves are finite. As it turns out, this is indeed the case: by virtue of Assumption Assumption 3.2 (b) and eq. (3.2), $\mathbb{E}_{\overline{q}_n}\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right] < \infty$. Moreover, by Assumption 3.1, $\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]$ is bounded below by $\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})\right] = C < \infty$ so that it also holds that

$$\int_{\boldsymbol{\Theta}} \int_{\mathcal{X}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \cdot 1_{\{\ell(\boldsymbol{\theta}, \boldsymbol{x}) \leq 0\}}(\boldsymbol{x}) \mu(d\boldsymbol{x}) \overline{q}_n(\boldsymbol{\theta}) \leq \min\{0, C\} < \infty.$$

Thus, the only remaining term in eq. (3.3) must also be finite and (i) follows.

(ii) By virtue of **(i)**, one may apply the Fubini-Tonelli Theorem to conclude that $\mathbb{E}_{\overline{q}_n}\left[\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right] = \mathbb{E}_\mu\left[\mathbb{E}_{\overline{q}_n}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right] < \infty$.

(iii) By definition of the expectation, it is clear that $\mathbb{E}_\mu\left[\mathbb{E}_{\overline{q}_n}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right] < \infty$ if and only if $\mathbb{P}_\mu\left(\mathbb{E}_{\overline{q}_n}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right] = \infty\right) = 0$, or equivalently if $\mathbb{P}_\mu\left(\mathbb{E}_{\overline{q}_n}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right] < \infty\right) = 1$. In other words, $\mathbb{E}_{\overline{q}_n}\left[\ell(\boldsymbol{\theta}, x_i)\right] < \infty$ holds μ-almost surely. □

Lemma 3.4 (ε_n-minimizers)**.** If Assumptions 3.1, and 3.2 (a), (b) hold, then the sequence $\{q_n\}_{n=1}^\infty$ produces finite valued objectives, i.e. $F_n(q_n) < \infty$. Moreover, q_n is a ε_n-solution of $\overline{F}_n$, i.e.

$$\overline{F}_n(q_n) \leq \inf_{q \in \mathcal{Q}} \overline{F}_n(q) + \varepsilon_n$$

for a sequence $\{\varepsilon_n\}_{n=1}^\infty$ with $\varepsilon_n < \infty$ μ-almost surely for all sufficiently large n and

given by

$$\varepsilon_n = 2 \left| \mathbb{E}_{\overline{q}_n} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) \right] \right] \right| \tag{3.4}$$

Proof. **μ-almost surely finite-valued objectives:** This immediately follows by Assumption 3.1 and Lemma 3.3. Recall that the Lemma implies that $\mathbb{E}_{\pi} \left[\ell(\boldsymbol{\theta}, x_i) \right] < \infty$ μ-almost surely, which means that $F_n(q_n) \leq F_n(\pi) < \infty$, μ-almost surely for all n. To complete the argument, note that for all sufficiently large n, Assumption 3.1 (a) implies a lower bound on $F_n(q_n)$.

Finite-valued ε_n: First, define the difference between $F_n(q)$ and $\overline{F}_n(q)$ as

$$e_n(q) = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) \right] \right] d\boldsymbol{\theta}.$$

It is clear that ε_n is finite-valued if and only if $e(\overline{q}_n)$ is. Now, notice that

$$e_n(\overline{q}_n) \leq \frac{1}{n} \sum_{i=1}^{n} \underbrace{\int_{\boldsymbol{\Theta}} \mathbb{E}_{\overline{q}_n} \left[\ell(\boldsymbol{\theta}, x_i) \right]}_{<\infty, \text{ Lemma 3.3}} - \underbrace{\mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right]}_{<\infty, \text{ Assumption 3.1}} < \infty.$$

ε_n-solution: Note that

$$\overline{F}_n(q_n) + e_n(q_n) = F_n(q_n) = \inf_{q \in \mathcal{Q}^{\boldsymbol{\theta}}} \left[\overline{F}_n(q) + e_n(q) \right] \leq \overline{F}_n(\overline{q}_n) + e_n(\overline{q}_n), \tag{3.5}$$

μ-almost surely. Further, by definition of $\overline{q}_n$ as the minimizer of $\overline{F}_n$, it also holds that $\overline{F}_n(q_n) \geq \overline{F}_n(\overline{q}_n)$ μ-almost surely, so that one may conclude that

$$0 \leq \overline{F}_n(q_n) - \overline{F}_n(\overline{q}_n) \leq e_n(\overline{q}_n) - e_n(q_n), \tag{3.6}$$

μ-almost surely, from which it clearly follows that

$$e_n(q_n) \leq e_n(\overline{q}_n), \tag{3.7}$$

μ-almost surely. This allows to conclude that

$$0 \leq e_n(\overline{q}_n) - e_n(q_n) \leq |e_n(\overline{q}_n)| + |e_n(q_n)| \leq 2|e_n(\overline{q}_n)|,$$

μ-almost surely. With this last result in hand, one can now define the sequence

$$\varepsilon_n = 2|e_n(\overline{q}_n)| < \infty$$

which together with eq. (3.6) yields that,

$$\overline{F}_n(q_n) \leq \inf_{q \in \mathcal{Q}^{\boldsymbol{\theta}}} \overline{F}_n(q) + \varepsilon_n.$$

so that the result follows. □

3.6.3 Proving that $\varepsilon_n \to 0$, μ-a.s. (S3)

Role of (S3) for the overall proof: While (S2) established that the minimizers q_n are ε_n-minimizers of $\overline{F}_n$, and while (S1) established that the minimizers $\overline{q}_n$ of $\overline{F}_n$ converge to the correct point, these insights do not suffice to prove consistency unless we can show that ε_n goes to zero (μ-almost surely).

To achieve this, the generic strategy is as follows: Since $\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]$ converges μ-almost surely to zero for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we would hope that this result extends to the integral that defines ε_n. In other words, we would hope that μ-almost surely,

$$\lim_{n\to\infty} \int_{\boldsymbol{\Theta}} \overline{q}_n(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})] \right] d\boldsymbol{\theta} = 0,$$

as this immediately implies that ε_n goes to zero μ-almost surely.

The key step in proving this is to find sufficient conditions under with the lim-operator can be pulled into the integral so that one may write

$$\begin{aligned} &\lim_{n\to\infty} \int_{\boldsymbol{\Theta}} \overline{q}_n(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})] \right] d\boldsymbol{\theta} \\ \overset{\text{(I)}}{=} &\int_{\boldsymbol{\Theta}} \lim_{n\to\infty} \left\{ \overline{q}_n(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})] \right] \right\} d\boldsymbol{\theta} \overset{\text{(II)}}{=} 0, \end{aligned} \tag{3.8}$$

where convergence is meant to occur μ-almost surely. Here, part (II) of this chain of equalities is much easier to establish (Lemma 3.5).

Part (I) of this chain of equalities is more difficult to show, as it amounts to finding conditions that are sufficient to prove the convergence of an indefinite integral over a generally unbounded random function $\overline{q}_n(\boldsymbol{\theta})\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$ to the integral over its deterministic pointwise limit. We provide two different strategies

for proving this. The first one proceeds via conditions on the loss function or the difference $\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu}[\ell(\boldsymbol{\theta}, \boldsymbol{x})]$ and relies on dominated convergence theorems or stochastic equicontinuity (Lemma 3.6, Corollary 3.2). The second method uses a Law of Large Numbers (LLN) on the triangular array $\{\{\mathbb{E}_{\overline{q}_n}[\mathbb{E}_{\mu}[\ell(\boldsymbol{\theta}, \boldsymbol{x})]]\}_{i=1}^{n}\}_{n=1}^{\infty}$ in conjunction with a restriction argument.

The second method relies on assumptions that are generally more difficult to establish, but remain unproblematic for a range of canonical settings in the variational case, where the variational family is Gaussian and the observations correspond to independent and identically distributed random variables.

Proving that (II) holds

Lemma 3.5. Suppose that Assumptions 3.1 and 3.2 hold. Further, suppose that (I) in (3.8) holds. Then, μ-almost surely,

$$\lim_{n\to\infty}\left\{\int_{\boldsymbol{\Theta}}\overline{q}_n(\boldsymbol{\theta})\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})\right]\right]d\boldsymbol{\theta}\right\} = 0.$$

Proof. First, note that by virtue of Assumption 3.2 (a), it holds that $\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})\right] \longrightarrow 0$, μ-a.s. for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Further, we also have $\overline{q}_n \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}$ by Corollary 3.1. Note in particular that $\overline{q}_n \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}$ together with Assumption **??** implies that the limit $\lim_{n\to\infty}\overline{q}_n(\boldsymbol{\theta}) = 0$ exists for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$. Moreover, it holds for any finite n that

$$\int_{\boldsymbol{\Theta}}\overline{q}_n(\boldsymbol{\theta})\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})\right]\right]d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}\setminus\{\boldsymbol{\theta}^*\}}\overline{q}_n(\boldsymbol{\theta})\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})\right]\right]d\boldsymbol{\theta},$$

so that this equality will hold in the limit, too. With this, we can now apply the multiplication rule of limits to conclude that

$$\begin{aligned}
&\lim_{n\to\infty}\left\{\int_{\boldsymbol{\Theta}}\overline{q}_n(\boldsymbol{\theta})\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})\right]\right]d\boldsymbol{\theta}\right\} \\
&= \int_{\boldsymbol{\Theta}\setminus\{\boldsymbol{\theta}^*\}}\lim_{n\to\infty}\{\overline{q}_n(\boldsymbol{\theta})\}\underbrace{\lim_{n\to\infty}\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu}\left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})\right]\right]\right\}}_{=0,\ \mu\text{-a.s.}}d\boldsymbol{\theta} = 0,
\end{aligned}$$

which concludes the proof. □

Proving that (I) holds via conditions on the loss

Lemma 3.6. Suppose Assumptions 3.1, 3.2 hold and that **(A)** or **(B)** is true where

(A) $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\ell(\boldsymbol{\theta}, \boldsymbol{x}) - \mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]| \leq h(\boldsymbol{x})$ μ-almost surely, where $\mathbb{E}_\mu[h(\boldsymbol{x})] < \infty$;

(B) $\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$ is asymptotically uniformly μ-almost surely bounded, i.e.

$$\mathbb{P}_\mu\left(\lim_{n\to\infty}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right| \leq h(\boldsymbol{x})\right) = 1,$$

where $\mathbb{E}_\mu[h(\boldsymbol{x})] < \infty$. Notice that for $h(\boldsymbol{x}) = 0$, this is equivalent to requiring a strong uniform law of large numbers to hold for $\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$ over $\boldsymbol{\Theta}$;

then, it also follows that μ-almost surely, both (I) and (II) in (3.8) hold.

Proof. It is clear that (II) holds by application of Lemma 3.5.
From **(A)**, (II) follows since $\mathbb{E}_{\overline{q}_n}\left[\mathbb{E}_\mu[h(\boldsymbol{x})]\right] = \mathbb{E}_\mu[h(\boldsymbol{x})]$, so the dominated convergence theorem implies that we can pull the limit operator under the integral sign in $\lim_{n\to\infty}\int_{\boldsymbol{\Theta}} \overline{q}_n(\boldsymbol{\theta})\left[\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]\right] d\boldsymbol{\theta}$. From **(B)**, (II) follows since by assumption, for all $n \geq N$ for some $N < \infty$, it holds that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})\right]\right| \leq M + h(\boldsymbol{x})$ μ-almost surely for some constant $M < \infty$. This allows us to once again use the dominated convergence theorem. □

While the conditions of Lemma 3.6 may seem somewhat obtuse, they hold for a range of mild conditions. The following Corollary elucidates this by providing some of these conditions that are easy to check.

Corollary 3.2. Suppose that Assumptions 3.1 and 3.2 hold. Suppose that additionally, any one of the following holds:

(i) $\sup_{x_i\in\mathcal{X},\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \leq M$ for some constant M, μ-almost surely;

(ii) $\ell(\boldsymbol{\theta}, \boldsymbol{x})$ and $\mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]$ are both jointly continuous in $\boldsymbol{\theta}$ and $\boldsymbol{x}$, μ-almost surely; and both $\mathcal{X}$ and $\boldsymbol{\Theta}$ are compact;

(iii) The function $\boldsymbol{\theta} \mapsto \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$ is continuously differentiable on $\boldsymbol{\Theta}$ μ-almost surely. Additionally, it holds that $\sup_{x_i\in\mathcal{X},\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}, x_i)| < \infty$, or that $\sup_{x_{1:n}\in\mathcal{X}^n,\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\nabla_{\boldsymbol{\theta}}\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right| < \infty$, or that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\nabla_{\boldsymbol{\theta}}\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right| < \infty$ μ-almost surely;

Then, both (I) and (II) hold so that $q_n \xrightarrow{D} \delta_{\boldsymbol{\theta}^*}$ μ-almost surely.

Proof. Conditions (i) and (ii) are just special cases of **(A)** in Lemma 3.6. This is immediate for (i), since $\sup_{x_i \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}, \boldsymbol{x}) \leq M$ implies that $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\mathbb{E}_\mu [\ell(\boldsymbol{\theta}, \boldsymbol{x})]| \leq M$ so that we can choose $h(\boldsymbol{x}) = 2M$. A similar argument holds for (ii): joint continuity and compactness imply that one can apply the Extreme Value Theorem for $(\boldsymbol{\theta}, x_i) \mapsto \ell(\boldsymbol{\theta}, x_i)$ and $\boldsymbol{\theta} \mapsto \mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]$, which then entails that $\sup_{x_i \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}, x_i) \leq M_1$ and $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]| \leq M_2$ holds for some constants M_1, M_2; so that we can set $h(\boldsymbol{x}) = M_1 + M_2$.

Condition (iii) is a little more nuanced and amounts to showing that **(B)** in Lemma 3.6 holds for $h(\boldsymbol{x}) = 0$. In other words, condition (iii) establishes sufficient conditions for a strong uniform law of large numbers. To this end, we invoke Theorem 21.8 of Davidson (1994). This result tells us that for a sequence of functions $\{f_n\}_{n \in \mathbb{N}}$ so that $f_n : \boldsymbol{\Theta} \to \mathbb{R}$, f_n converges uniformly μ-almost surely to f if and only if $f_n(\boldsymbol{\theta}) \overset{\mu-\text{a.s.}}{\longrightarrow} f(\boldsymbol{\theta})$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and if $\{f_n\}_{n \in \mathbb{N}}$ is strongly stochastically equi-continuous. As shown in Theorem 21.10 of Davidson (1994), a function sequence $\{f_n\}_{n \in \mathbb{N}}$ is strongly stochastically equi-continuous on $\boldsymbol{\Theta}$ if there exists a stochastic sequence C_n independent of $\boldsymbol{\theta}$ so that $|f_n(\boldsymbol{\theta}) - f_n(\boldsymbol{\theta}')\| \leq C_n \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ and $\limsup_{n \to \infty} C_n < \infty$. Setting $f_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)$, the fact that f_n is continuously differentiable by assumption implies that we can apply the Mean Value Theorem to conclude that $|f_n(\boldsymbol{\theta}) - f_n(\boldsymbol{\theta}')\| \leq C_n \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ for $C_n = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|f_n(\boldsymbol{\theta})\|_2$. It is now easy to see that $\lim_{n \to \infty} C_n < \infty$ μ-a.s. (and therefore that $\limsup_{n \to \infty} C_n < \infty$ μ-a.s.) if $\sup_{x_i \in \mathcal{X}, \boldsymbol{\theta} \in \boldsymbol{\Theta}} |\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}, x_i)| < \infty$, or $\sup_{x_{1:n} \in \mathcal{X}^n, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \left|\nabla_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right| < \infty$, or $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left|\nabla_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i)\right| < \infty$ holds μ-almost surely. Since we also have that $f_n(\boldsymbol{\theta}) \longrightarrow f(\boldsymbol{\theta})$ μ-almost surely for $f(\boldsymbol{\theta}) = \mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})]$ and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we can invoke Theorem 21.8 of Davidson (1994).

To conclude the proof, simply note that (I) and (II) guarantee that ε_n goes to zero μ-almost surely, which in turn yields the desired consistency result by application of Corollary 7.24 in (Dal Maso, 2012). □

Proving that (I) holds with a restriction argument and a triangular Laws of Large Numbers

Throughout this section, we define for convenience the triangular array $\{Z_i^{(n)}\}_{n \in \mathbb{N}}$, where $Z_i^{(n)} = \mathbb{E}_{\overline{q}_n}[1_A(\boldsymbol{\theta}) \cdot \ell(\boldsymbol{\theta}, x_i)]$ for some compact set A that contains $\boldsymbol{\theta}^*$. The two main conditions required for our alternative proof of (I) are as follows: Firstly, we require that $\{\frac{1}{n} \sum_{i=1}^n Z_i^{(n)}\}_{n \in \mathbb{N}}$ satisfies a strong law of large numbers. Secondly, it must be viable to restrict the analysis of the integral in 3.8 to the same compact

set A in the sense that

$$\lim_{n\to\infty} \int_{\boldsymbol{\Theta}} \overline{q}_n(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) \right] \right] d\boldsymbol{\theta}$$
$$= \lim_{n\to\infty} \int_{A} \overline{q}_n(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu \left[\ell(\boldsymbol{\theta}, \boldsymbol{x}) \right] \right] d\boldsymbol{\theta}. \tag{3.9}$$

Proving that (I) holds with a restriction argument and a triangular Laws of Large Numbers: the restriction argument (3.9)

As it turns out, the condition (3.9) holds if $\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\pi \left[|\ell(\boldsymbol{\theta}, x_i)| \right]\}_{n\in\mathbb{N}}$ satisfies a strong law of large numbers, and if we have that for large enough n, $\overline{q}_n \leq \pi$ everywhere except in a compact set around the parameter value $\boldsymbol{\theta}^*$ that it converges towards. The next Lemma demonstrates this.

Lemma 3.7. Suppose that Assumptions 3.1 and 3.2 (a), (d) and (b) hold, that we have $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\pi \left[|\ell(\boldsymbol{\theta}, x_i)| \right] \longrightarrow \mathbb{E}_\mu \left[\mathbb{E}_\pi [|\ell(\boldsymbol{\theta}, \boldsymbol{x})|] \right]$ μ-almost surely, and that for a compact set A containing $\boldsymbol{\theta}^*$, there exists N so that $\overline{q}_n \leq \pi$ for all $\boldsymbol{\Theta} \setminus A$ whenever $n \geq N$. Then, (3.9) holds for all n that are sufficiently large.

Proof. Note that (3.9) is equivalent to proving

$$\lim_{n\to\infty} \left\{ \int_{\boldsymbol{\Theta}} \overline{q}_n(\boldsymbol{\theta}) 1_{\boldsymbol{\Theta}\setminus A}(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_\mu \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right] \right] d\boldsymbol{\theta} \right\} = 0,$$

which we will do by justifying application of the Dominated Convergence Theorem. To this end, first note that because $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\pi \left[|\ell(\boldsymbol{\theta}, x_i)| \right] \longrightarrow \mathbb{E}_\mu \left[\mathbb{E}_\pi [|\ell(\boldsymbol{\theta}, \boldsymbol{x})|] \right]$ μ-almost surely, it also holds that we can find $\tilde{N}$ so that for a fixed and finite $\varepsilon > 0$ and $n \geq \tilde{N}$, it holds μ-almost surely that

$$\left| \mathbb{E}_\pi \left[\frac{1}{n} \sum_{i=1}^{n} |\ell(\boldsymbol{\theta}, x_i)| \right] - \mathbb{E}_\mu \left[\mathbb{E}_\pi \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})| \right] \right] \right| \leq \varepsilon.$$

Since we also have $\overline{q}_n \leq \pi$ outside some compact set A containing $\boldsymbol{\theta}^*$ and all $n \geq N$,

the following bound holds for all $n \geq \max\{N, \tilde{N}\}$:

$$
\begin{aligned}
&\int_{\boldsymbol{\Theta}} \left| \overline{q}_n(\boldsymbol{\theta}) 1_{\boldsymbol{\Theta} \setminus A}(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right] \right] \right| d\boldsymbol{\theta} \\
&\leq \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) 1_{\boldsymbol{\Theta} \setminus A}(\boldsymbol{\theta}) \left| \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right] \right| d\boldsymbol{\theta} \\
&\leq \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) \left| \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right] \right| d\boldsymbol{\theta} \\
&\leq \int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta}) \left\{ \frac{1}{n} \sum_{i=1}^{n} |\ell(\boldsymbol{\theta}, x_i)| + |\mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right]| \right\} d\boldsymbol{\theta} \\
&= \mathbb{E}_{\pi} \left[\frac{1}{n} \sum_{i=1}^{n} |\ell(\boldsymbol{\theta}, x_i)| + |\mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right]| \right] \\
&\leq \mathbb{E}_{\mu} \left[\mathbb{E}_{\pi} \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})| \right] \right] + \varepsilon + |\mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right]| \\
&= \mathbb{E}_{\pi} \left[\mathbb{E}_{\mu} \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})| \right] \right] + \varepsilon + |\mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right]| \\
&= \int_{\boldsymbol{\Theta}} \left| \mathbb{E}_{\mu} \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})| \right] + \varepsilon + |\mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right]| \right| d\boldsymbol{\theta} < \infty.
\end{aligned}
$$

The first inequality follows since n is chosen large enough so that $\overline{q}_n \leq \pi$ outside A, the second because $1 \geq 1_{\boldsymbol{\Theta} \setminus A}$, the third by repeated application of the triangle inequality, and the fourth by virtue of the fact that $n \geq \tilde{N}$. The second-to-last equality follows by application of Fubini's Theorem. The finiteness of the last integral follows because $\varepsilon < \infty$ by assumption, $\mathbb{E}_{\mu} \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})| \right] < \infty$ by virtue of Lemma 3.3, and $\mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right] < \infty$ since the minimizer $\boldsymbol{\theta}^*$ is unique by Assumption 3.1. Therefore, and defining

$$
g(\boldsymbol{\theta}) = \mathbb{E}_{\mu} \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})| \right] + \varepsilon + |\mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right]| ,
$$

we have that g is integrable and upper bounds

$$
\left| \overline{q}_n(\boldsymbol{\theta}) 1_{\boldsymbol{\Theta} \setminus A}(\boldsymbol{\theta}) \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i) - \mathbb{E}_{\mu} \left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x}) \right] \right] \right|
$$

over $\boldsymbol{\Theta}$ for all sufficiently large n. This implies that we can apply the Dominated

Convergence Theorem and conclude that μ-almost surely,

$$\lim_{n\to\infty}\left\{\int_{\boldsymbol{\Theta}}\overline{q}_n(\boldsymbol{\theta})1_{\boldsymbol{\Theta}\setminus A}(\boldsymbol{\theta})\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta},x_i)-\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}^*,\boldsymbol{x})\right]\right]d\boldsymbol{\theta}\right\}$$
$$=\int_{\boldsymbol{\Theta}\setminus A}\underbrace{\lim_{n\to\infty}\left\{\overline{q}_n(\boldsymbol{\theta})\right\}}_{=0}\cdot\underbrace{\lim_{n\to\infty}\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{\theta},x_i)-\mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}^*,\boldsymbol{x})\right]\right]\right\}}_{\text{pointwise limit exists }\mu\text{-a.s.}}d\boldsymbol{\theta}=0,$$

where the first pointwise limit follows by Assumption 3.2 (a), and the pointwise limit of the second term exists (μ-almost surely) by virtue of Assumption 3.1. □

We now provide a simple example under which the previous result can be applied.

Lemma 3.8. Suppose that $x_{1:n}\overset{i.i.d.}{\sim}\mu$, $\mathbb{E}_\mu\left[\mathbb{E}_\pi\left[|\ell(\boldsymbol{\theta},\boldsymbol{x})|\right]^2\right]<\infty$. Letting $\mathcal{Q}'$ be the collection of all fully factorized normal distributions on $\boldsymbol{\Theta}=\mathbb{R}^d$ parameterized by $\boldsymbol{\kappa}=(\boldsymbol{\sigma},\boldsymbol{m})$ so that for each $q\in\mathcal{Q}'$, there exist a positive definite vector $\boldsymbol{\sigma}\in\mathbb{R}^d$ and a vector $\boldsymbol{m}\in\mathbb{R}^d$ for which

$$q(\boldsymbol{\theta}|(\boldsymbol{\Sigma},\boldsymbol{m}))=(2\pi)^{-d/2}\left(\prod_{i=1}^{d}\boldsymbol{\sigma}_i\right)^{-1/2}e^{-0.5\cdot(\boldsymbol{\theta}-\boldsymbol{m})^T(\boldsymbol{\sigma}I)^{-1}(\boldsymbol{\theta}-\boldsymbol{m})},$$

take any one-to-one, invertible and bounded transformation $f:\mathbb{R}^d_{\geq 0}\times\mathbb{R}^d\to[0,1]^d\times[0,1]^d$ and choose the variational family $\mathcal{Q}$ that is obtained by re-parametrizing $\boldsymbol{\kappa}$ with $f(\boldsymbol{\kappa})$. Further, suppose that Assumptions 3.1 and 3.2 hold. Then, (3.9) holds for all n that are sufficiently large.

Proof. First, note that since $\mathbb{E}_\mu\left[\mathbb{E}_\pi\left[|\ell(\boldsymbol{\theta},\boldsymbol{x})|\right]^2\right]<\infty$ Kolmogorov's strong law of large numbers implies that $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\pi\left[|\ell(\boldsymbol{\theta},x_i)|\right]]\longrightarrow\mathbb{E}_\mu\left[\mathbb{E}_\pi[|\ell(\boldsymbol{\theta},\boldsymbol{x})|]\right]$ μ-almost surely.

Next, note that the transformation f applied to $\boldsymbol{\kappa}$ is invertible and one-to-one, so we can conduct our analysis in the parameter space $\boldsymbol{K}$ of $\boldsymbol{\kappa}$ (see Remark 3.6). Further, by Assumption 3.2, we have that $\pi\in\mathcal{Q}'$, which means that there is a positive vector $\boldsymbol{\sigma}_\pi$ and a mean vector $\boldsymbol{m}_\pi$ for which $\pi(\boldsymbol{\theta})=q(\boldsymbol{\theta}|(\boldsymbol{\sigma}_\pi,\boldsymbol{m}_\pi))$. Further, since we can apply Lemma 3.1, we know that $\overline{q}_n(\boldsymbol{\theta}|(\boldsymbol{\sigma}_n,\boldsymbol{m}_n))$ converges to a point mass at $\boldsymbol{\theta}^*$, which implies that $(\boldsymbol{m}_n,\boldsymbol{\sigma}_n)\longrightarrow(\boldsymbol{\theta}^*,\boldsymbol{0})$. Now by virtue of being Gaussian, π has monotonically decreasing tails in all directions. This allows us to fix a compact set A containing $\boldsymbol{\theta}^*$ as well as $\boldsymbol{m}_\pi$ in the following way: set $d_1=\|\boldsymbol{\mu}_\pi-\boldsymbol{\theta}^*\|_2$, and define $A=\{\boldsymbol{\theta}\in\boldsymbol{\Theta}:\|\boldsymbol{m}_\pi-\boldsymbol{\theta}\|_2\leq d_1+1\}$. Since $\boldsymbol{m}_\pi\in A$, we know that $\pi(\boldsymbol{\theta})$ decreases as we move away from the boundary of A. This allows us

to conclude that the smallest value of $\pi(\boldsymbol{\theta})$ in A occurs on the boundary, and we can define it as $\pi_{\min} = \min_{\boldsymbol{\theta} \in A} \pi(\boldsymbol{\theta})$. The next step consists in noting that there must exist N so that $\boldsymbol{m}_n \in B$ for all $n \geq N$ and $B = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|_2 \leq 0.5\}$. Note that $B \subset A$. More importantly, the boundary of B has a distance of at least 0.5 to the boundary of A. Now, since $\overline{q}_n$ is a normal distribution concentrating to a dirac measure at a distance of at least 0.5 from the boundary of A, we know that there must also exist $\tilde{N}$ so that for all $n \geq \tilde{N}$, the largest value attained by $\overline{q}_n$ on the boundary of A is at most as large as $\pi_{\min}$. More formally, we know that there exists $\tilde{N}$ so that $\sup_{\boldsymbol{\theta} \in \partial A} \overline{q}_n(\boldsymbol{\theta}) \leq \pi_{\min}$ for all $n \geq \tilde{N}$, where ∂A denotes the boundary of A. Figure 3.1 illustrates this visually. Consequently, we can apply Lemma 3.7 and

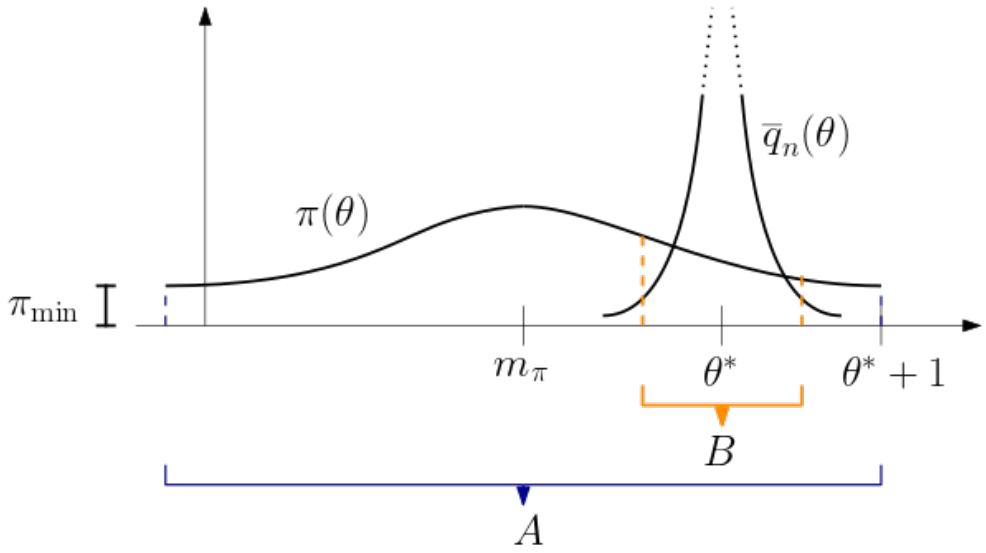


Figure 3.1: Depicted is the strategy for the proof in Lemma 3.8.

the result follows. □

Remark 3.7. In the proof of Lemma 3.8, we have used normality of the prior (which follows from Assumption 3.2 (b)) for pedagogical reasons. Clearly, this is much stricter than what we require in the proof: as a quick glance at the proof and Figure 3.1 reveals, all that is really required is the existence of a compact set A containing $\boldsymbol{\theta}^*$ so that (i) $\boldsymbol{\theta}^*$ has at least some positive distance $\varepsilon > 0$ to the boundary of A (where we chose $\varepsilon = 1$ in the proof), (ii) $\pi(\boldsymbol{\theta})$ is decreasing as we move in any direction pointing outside of A, (iii) we can lower bound π on A so that $\inf_{\boldsymbol{\theta} \in A} \pi(\boldsymbol{\theta}) = \pi_{\min} > 0$. All these properties are only sufficient (and not necessary); and are true for virtually all commonly used priors, including full (rather than factorized) multivariate Gaussians, priors distributed as Student's t-distribution,

the β-distribution, the uniform distribution, trapezoidal distributions, etc.

Remark 3.8. Similarly to the prior discussed in Remark 3.7, the choice of variational posterior family for Lemma 3.8—again, for pedagogical reasons—is far more rigid than what is necessary. In fact, all the proof uses is that the tails of normal distributions are decreasing. More generally speaking, we can use the same arguments for any variational family which satisfies that for each $q \in \mathcal{Q}$, there exists M so that for all $\|\boldsymbol{\theta}\|_2 > M$, q is decreasing as we move in any direction pointing outside of the set $\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta}\|_2 \leq M\}$.

Using the insights of Remarks 3.7 and 3.8, it becomes clear that a much more general version of Lemma 3.8 can be shown to hold.

Lemma 3.9. Suppose that $x_{1:n} \overset{i.i.d.}{\sim} \mu$, $\mathbb{E}_\mu \left[\mathbb{E}_\pi \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})| \right]^2 \right] < \infty$, and that $\mathcal{Q}$ is a variational family such that

(i) for each $q \in \mathcal{Q}$, there exists M so that for all $\|\boldsymbol{\theta}\|_2 > M$, q is decreasing as we move in any direction pointing outside of the set $\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta}\|_2 \leq M\}$;

(ii) for any $q \in \mathcal{Q}$, and on any compact set S of $\boldsymbol{\Theta}$, we can lower bound q on S so that $\inf_{\boldsymbol{\theta} \in S} q(\boldsymbol{\theta}) > 0$.

Further, suppose that Assumptions 3.1 and 3.2 hold. Then, (3.9) holds for all n that are sufficiently large.

Proof. This follows by the same logic as the proof of Lemma 3.8 and by using the insights of Remarks 3.7 and 3.8. Note that the requirements (i),(ii), and (iii) in Remark 3.7 hold because $\pi \in \mathcal{Q}$ by Assumption 3.2 (b); and that the requirement outlined in Remark 3.8 holds by assumption. □

We can now use the previous Lemma to show that the restriction condition in (3.9) will hold for a wide range of variational families. Note that the list of distributions in the Corollary is obviously not complete, and many other candidate families satisfy the conditions of the previous Lemma.

Corollary 3.3. Suppose that $x_{1:n} \overset{i.i.d.}{\sim} \mu$, $\mathbb{E}_\mu \left[\mathbb{E}_\pi \left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})| \right]^2 \right] < \infty$, and that $\mathcal{Q}$ is the collection of multivariate normals, multivariate student's t-distributions, Beta distributions, or uniform distributions on $\boldsymbol{\Theta}$ re-parameterized by an invertible, bounded, and one-to-one transformation. Further, suppose that Assumptions 3.1 and 3.2 hold. Then, (3.9) holds for all n that are sufficiently large.

Proving that (I) holds with a restriction argument and a triangular Laws of Large Numbers: The Law of Large Numbers (LLN)

Having established suitably mild conditions for our restriction condition, we will proceed with showing that $\{\frac{1}{n}\sum_{i=1}^{n} Z_i^{(n)}\}_{n\in\mathbb{N}}$ satisfies a strong law of large numbers (LLNs) for relatively mild settings. Recall that $Z_i^{(n)} = \mathbb{E}_{\bar{q}_n}[1_A(\boldsymbol{\theta}) \cdot \ell(\boldsymbol{\theta}, x_i)]$ for some compact set A that contains $\boldsymbol{\theta}^*$. It is possible to establish LLNs for the triangular array $\{Z_i^{(n)}\}_{n\in\mathbb{N}}$ with canonical i.i.d. assumption together with a mild moment and continuity condition. The moment condition is unproblematic, but somewhat difficult to verify without imposing needlessly strict conditions.

Lemma 3.10. Suppose that Assumptions 3.1 and 3.2 (a), (d) and (b) hold, that $x_{1:n} \overset{i.i.d}{\sim} \mu$, that A is a compact set containing $\boldsymbol{\theta}^*$, and that $\mathbb{E}_\mu\left[|\ell(\boldsymbol{\theta}, \boldsymbol{x}) \cdot 1_A(\boldsymbol{\theta})|^{2+\delta}\right] < \infty$ for all $\boldsymbol{\theta} \in A$ and for some $\delta > 0$. Lastly, assume that $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{\theta}, x_i)$ is μ-almost surely continuous on A, and that $\boldsymbol{\theta} \mapsto \mathbb{E}_\mu\left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|\right]$ is continuous on A. Then, μ-almost surely we have that

$$\frac{1}{n}\sum_{i=1}^{n} Z_i^{(n)} - \mathbb{E}_\mu[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})] \longrightarrow 0.$$

Proof. We first show that

$$\frac{1}{n}\sum_{i=1}^{n} Z_i^{(n)} - \mathbb{E}_\mu\left[\mathbb{E}_{\bar{q}_n}[\ell(\boldsymbol{\theta}, \boldsymbol{x}) 1_A(\boldsymbol{\theta})]\right] \longrightarrow 0. \tag{3.10}$$

using Corollary 1 of Hu et al. (1989). Because $x_{1:n} \overset{i.i.d.}{\sim} \mu$, we clearly have that

$$\mathbb{E}_\mu\left[\sum_{i=1}^{n} Z_i^{(n)} - \mathbb{E}_\mu\left[\mathbb{E}_{\bar{q}_n}[\ell(\boldsymbol{\theta}, \boldsymbol{x}) 1_A(\boldsymbol{\theta})]\right]\right] = 0 \tag{3.11}$$

The only other condition we need before we can apply said Corollary is the finiteness condition $\sup_{n\geq 1} \mathbb{E}_\mu\left[\left|\mathbb{E}_{\bar{q}_n}[\ell(\boldsymbol{\theta}, \boldsymbol{x}) 1_A(\boldsymbol{\theta})]\right|^{2+\delta}\right] < \infty$. To this end, we can bound

$$\begin{aligned}
&\sup_{n\geq 1} \mathbb{E}_\mu\left[\left|\mathbb{E}_{\bar{q}_n}[\ell(\boldsymbol{\theta}, \boldsymbol{x}) 1_A(\boldsymbol{\theta})]\right|^{2+\delta}\right] \\
\leq &\sup_{n\geq 1} \mathbb{E}_\mu\left[\mathbb{E}_{\bar{q}_n}[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|^{2+\delta} 1_A(\boldsymbol{\theta})]\right] \\
= &\sup_{n\geq 1} \mathbb{E}_{\bar{q}_n}\left[\mathbb{E}_\mu[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|^{2+\delta} 1_A(\boldsymbol{\theta})]\right] \\
\leq &\sup_{\boldsymbol{\theta}\in A} \mathbb{E}_\mu\left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|^{2+\delta}\right] < \infty.
\end{aligned}$$

Here, the first inequality follows by Jensen's inequality, the equality by application of Fubini's Theorem; the second inequality by bounding $\mathbb{E}_{\overline{q}_n}[f(\boldsymbol{\theta})1_A(\boldsymbol{\theta})] \leq \sup_{\boldsymbol{\theta}\in A} f(\boldsymbol{\theta})$; and the finiteness implied by the last inequality follows by application of the Extreme Value Theorem to the continuous function $\boldsymbol{\theta} \mapsto \mathbb{E}_\mu\left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|^{2+\delta}\right]$ on A. Note that the function is indeed continuous since $\mathbb{E}_\mu\left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|^{2+\delta}\right] < \infty$ for $\boldsymbol{\theta} \in A$, and $\mathbb{E}_\mu\left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|\right]$ is continuous.

The only step that is missing is to show that $m_n = \mathbb{E}_\mu\left[\mathbb{E}_{\overline{q}_n}\left[\ell(\boldsymbol{\theta}, \boldsymbol{x})1_A(\boldsymbol{\theta})\right]\right] \longrightarrow \mathbb{E}_\mu[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})]$ as $n \to \infty$. This follows by standard arguments: first, note that by assumption, we have $m_n < \infty$. Further, because $\mathbb{E}_\mu\left[\mathbb{E}_{\overline{q}_n}[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|1_A(\boldsymbol{\theta})]\right] < \infty$, we can apply Fubini's Theorem together with the fact that $\overline{q}_n \xrightarrow{\mathcal{D}} \delta_{\boldsymbol{\theta}^*}$ implied by Corollary 3.1 to conclude that

$$\lim_{n\to\infty} m_n = \lim_{n\to\infty} \mathbb{E}_{\overline{q}_n}\left[\mathbb{E}_\mu[\ell(\boldsymbol{\theta}, \boldsymbol{x})1_A(\boldsymbol{\theta})]\right] = \mathbb{E}_\mu\left[\ell(\boldsymbol{\theta}^*, \boldsymbol{x})\right],$$

where the last equality follows by the Portmanteau Theorem and because $\boldsymbol{\theta} \mapsto \ell(\boldsymbol{\theta}, x_i)$ is μ-almost surely continuous on A. □

Remark 3.9. The only condition in the preceeding result that could be difficult to justify is the moment condition that $\mathbb{E}_\mu\left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|^{2+\delta}\right] < \infty$ for all $\boldsymbol{\theta} \in A$. The main challenge with this assumption is that while it is easy to verify if the loss is bounded on A (i.e., if $\sup_{\boldsymbol{\theta}\in A, x\in\mathcal{X}} |\ell(\boldsymbol{\theta}, x)| < \infty$), this is a much stricter condition than the moment condition. In this sense, the moment condition often has to be taken on faith. In contrast, the continuity conditions on A are not typically challenging in practice: most losses will be continuous, especially in a small region around the minimum; and it is reasonable to expect that the same is true when one integrates out $\boldsymbol{x}$ with μ. Even if continuity were a problem however, we could get rid of this assumption so long as it holds that $\sup_{n\geq 1} \mathbb{E}_\mu\left[\left|\mathbb{E}_{\overline{q}_n}[\ell(\boldsymbol{\theta}, \boldsymbol{x})1_A(\boldsymbol{\theta})]\right|^{2+\delta}\right] < \infty$ or $\sup_{\boldsymbol{\theta}\in A} \mathbb{E}_\mu\left[|\ell(\boldsymbol{\theta}, \boldsymbol{x})|^{2+\delta}\right] < \infty$.

3.7 Proof of the Main Results

With everything now in place, we can now prove the two main results of the current chapter.

3.7.1 Proof of Theorem 3.1

Clearly, this result is a straightforward application of Corollary 3.2.

3.7.2 Proof of Theorem 3.2

This follows by showing that both the restriction condition (3.9) and the desired Law of Large Numbers (LLN) hold. Clearly, this can be achieved by combining Lemma 3.9 with Lemma 3.10.

Part II

Methodological Advances

Chapter 4

Generalized Variational Inference, Part 1: Computation

Summary: In this and the following two chapters, we study the special case for the Rule of Three (RoT) for which optimization happens over a set of parameterized distributions. For its obvious relationship to previous variational methods, we call this family of algorithms Generalized Variational Inference (GVI). In this first of three chapters devoted to GVI, we explain its computational properties. In particular, we explain how—unlike most other RoT posteriors—GVI posteriors can usually be computed (at least approximately). Particular attention is paid to the computational toolkit from variational methods that enables this computation, how it needs to be adjusted to fit GVI, and special cases for which the optimisation problem becomes particularly easy to solve.

In this chapter, we study posteriors obtained via Generalized Variational Inference (GVI). These posteriors are a particular version of RoT posteriors, and are particularly feasible for scalable inference in complex, high-dimensional problems.

Definition 4.1 (Generalized Variational Inference (GVI))**.** Computing any RoT posterior of form $P(\ell, D, \mathcal{Q})$ for $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in \boldsymbol{K}\}$ being a subset of $\mathcal{P}(\boldsymbol{\Theta})$ parameterized by $\boldsymbol{\kappa} \in \boldsymbol{K}$ (also called a variational family) constitutes a procedure we call Generalized Variational Inference (GVI). Accordingly, the resulting posterior is referred to as a GVI posterior.

As Definition 4.1 reveals, the feature that makes GVI posteriors particularly appealing is their practicability: while it is unclear how to compute posteriors via the RoT when optimization is performed over non-parameterized infinite-dimensional

spaces, this is not an issue when $\Pi = \mathcal{Q}$: In this case, optimization is instead performed over the finite-dimensional space $\boldsymbol{K}$.

In the next three chapters, we proceed as follows: In the current chapter, we explain how GVI relates to other variational methods in Bayesian statistics, but also how it is fundamentally different from them. Most importantly, we also explore how to compute GVI posteriors. Next, Chapters 5 and 6 will motivate practical reasons for studying GVI, and explain how GVI posteriors can address questions of robustness to ill-specified priors or likelihood functions. We demonstrate this on two models that are commonly used in Bayesian Machine Learning: Deep Gaussian Processes (DGPs) and Bayesian Neural Networks (BNNs).

4.1 Standard variational methods and GVI

The driving idea behind the RoT as well as GVI is that undesirable inference outcomes are synonymous with an inappropriately designed optimization objective in Definition 2.3—an observation we call the optimization-centric view on Bayesian inference. Following this line of reasoning, the most transparent way of improving posteriors is a *direct* adjustment of the optimization problem that generated them—and GVI posteriors are a tractable way of achieving this. GVI posteriors have a number of desirable properties beyond tractability. Of particular importance is that they inherit the modularity outlined in Remark 1.2. The ramifications are twofold:

(1) GVI can address prior misspecification (as well as the nature of uncertainty quantification) by changing D;

(2) GVI can address model misspecification by changing L.

In the context of potential misspecification problems, this modularity means that GVI posteriors $P(L, D, \mathcal{Q})$ are appealing alternatives to $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ or $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$. More precisely, if one can identify whether the assumptions underlying standard Bayesian inference are violated via the likelihood or the prior, GVI can be used to address this by directly modifying L or D. This means that GVI has an inherently different motivation from other variational methods such as standard Variational Inference (VI) or Discrepancy Variational Inference (DVI) introduced in Section 1.2.3: rather than seeking to approximate $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ (or $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$) with a projection operation of the form

$$q^*_A(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} D(q \| q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})),$$

GVI designs and computes an inherently different—and hopefully better-suited—posterior belief for the problem at hand.

4.1.1 Approximating $q^*_{n,\mathbf{SB}}(\theta)$ vs. specifying a new posterior

Some practitioners may argue that this feature makes GVI *less* desirable than alternative variational methods: why would we prefer these posteriors over approximations to $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$? In principle, this is a valid point: in fact, if the assumptions underlying Bayesian inference are at least approximately correct, and if $\mathcal{Q}$ contains qualitatively good approximations to $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$, one will *want* to use a method that is motivated as approximation to $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$. Unfortunately, this idealization often does not reflect the situation we encounter in practice. In fact, thinking of variational methods as approximations is often misleading in the first place—even if likelihoods and priors are correctly specified. The reason for this is that in many applications, the set $\mathcal{Q}$ does not contain any distributions suitable for approximating $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ in any meaningful way. For example, it is questionable if one obtains a meaningful approximation to $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ if the latter is a multi-modal distribution and $\mathcal{Q}$ consists of all uni-modal normal distributions on $\boldsymbol{\Theta}$. In this setting—which is rather commonplace in Machine Learning applications—variational methods motivated by approximating $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ have a clear conceptual drawback when compared to GVI posteriors: they are not interpretable as a modularly specified belief distribution; and so their only quality benchmark should be their approximating behaviour. We demonstrate this tension in Example 4.1 and Figure 4.1, and will revisit this issue with our experiments in the following two chapters.

Example 4.1 (Label switching and multi-modality)**.** A recurrent theme in the research on variational approximations $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ to $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ is the observation that if $\mathcal{Q}$ is a mean field normal family, $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ will center closely around the maximum likelihood estimate (e.g. Turner and Sahani, 2011).[1] This phenomenon is often referred to as the zero-forcing behaviour of the KLD-projection (Minka, 2005). Its effect are undesirably overconfident variational posteriors $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$. Moreover, this problem is especially pronounced when the approximated posterior beliefs $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ are multi-modal. Popular approaches to address this issue are Expectation Propagation (EP) (Minka, 2001; Opper and Winther, 2000) and Divergence Variational Inference (DVI) methods as introduced in Section 1.2.3 (e.g. Hernández Lobato et al., 2016; Li and Turner, 2016; Dieng et al., 2017). All of these approaches seek to (locally

[1] Similarly, if we approximate $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$, $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$ will center closely around the empirical risk minimizer $\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}, x_{1:n})$

or globally) minimize an alternative zero-avoiding[2] divergence D between $\mathcal{Q}$ and $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$. In contrast with GVI, changing the divergence in the DVI-sense explicitly encodes the desire to approximate $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$.

Using Bayesian mixture models (BMMs), we show that this can indeed be a problem in practice. In particular, we show that this can accidentally interfere with the negative log loss targeted by the underlying Bayes posterior. BMMs produce multi-modal posteriors as the likelihood function is invariant to switching parameter labels. In other words, BMMs have multiple parameter values that constitute equally good fits to the data. With this in mind, we simulate $n = 100$ observations from

$$p\left(x|\boldsymbol{\theta} = (\mu_1, \mu_2)\right) = 0.5 \cdot \mathcal{N}\left(x|\mu_1, 0.65^2\right) + 0.5 \cdot \mathcal{N}\left(x|\mu_2, 0.65^2\right)$$

for two different parameterizations 1) $\boldsymbol{\theta} = (0, 0.75)$ and 2) $\boldsymbol{\theta} = (0, 2)$. For inference, we use the well-specified prior belief $\mu_j \sim \mathcal{N}(0, 2^2)$, $j = 1, 2$. Using the correctly specified likelihood function $\ell(\boldsymbol{\theta}, x_i) = -\log p\left(x_i|\boldsymbol{\theta} = (\mu_1, \mu_2)\right)$, we compare the **standard Bayesian** posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$, the standard VI posterior $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$, a DVI posterior based on Rényi's α-divergence ($D^{(\alpha)}_{AR}$) as described by Li and Turner (2016), and a GVI posterior using $D = D^{(\alpha)}_{AR}$ (see eq. (5.2)). For $\mathcal{Q}$, we use the collection of fully-factorized normals on Θ.

Figure 4.1 shows the results. Because $p\left(x|\boldsymbol{\theta} = (\mu_1, \mu_2)\right) = p\left(x|\boldsymbol{\theta} = (\mu_2, \mu_1)\right)$, there are two equally good parameter values describing the data—implying that the full posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ is bi-modal. By choice of $\mathcal{Q}$ however, the posteriors are forced to be unimodal, which endows them with a straightforward interpretation: Firstly, the modes of these posteriors should be close to (one of the two) best parameter values of $\boldsymbol{\theta} = (\mu_1, \mu_2)$. Secondly, their variances quantify the uncertainty about this best value. For both settings of the true value for $\boldsymbol{\theta}$, DVI produces a posterior that reflects a highly undesirable belief: the mode of the DVI posterior is located at a (locally) worst value of $\boldsymbol{\theta}$. Unsurprisingly and as the bottom right plot shows, this adversely affects predictive performance. This behaviour is entirely attributable to the fact that unlike GVI posteriors, DVI do not inherit the modularity that stems from the form in Definitions 2.3 and 4.1. In this context, Figure 4.1 serves as a morality tale: In the GVI framework, changing $D = \mathrm{KLD}$ to another divergence only changes uncertainty quantification and does **not** affect the way the best parameter is defined: that part of the inference problem is determined by the loss L. In sharp contrast, the DVI framework comes with no such guarantee! Accordingly, posteriors produced by DVI with $D \neq \mathrm{KLD}$ conflate uncertainty quantification and the way

[2]The origin of this term is that approximations $q^*_{\mathrm{A}}(\boldsymbol{\theta})$ derived from these divergences—in contrast to $q^*_{\mathrm{VI}}(\boldsymbol{\theta})$—avoid being close to zero for regions of high probability mass under $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$.

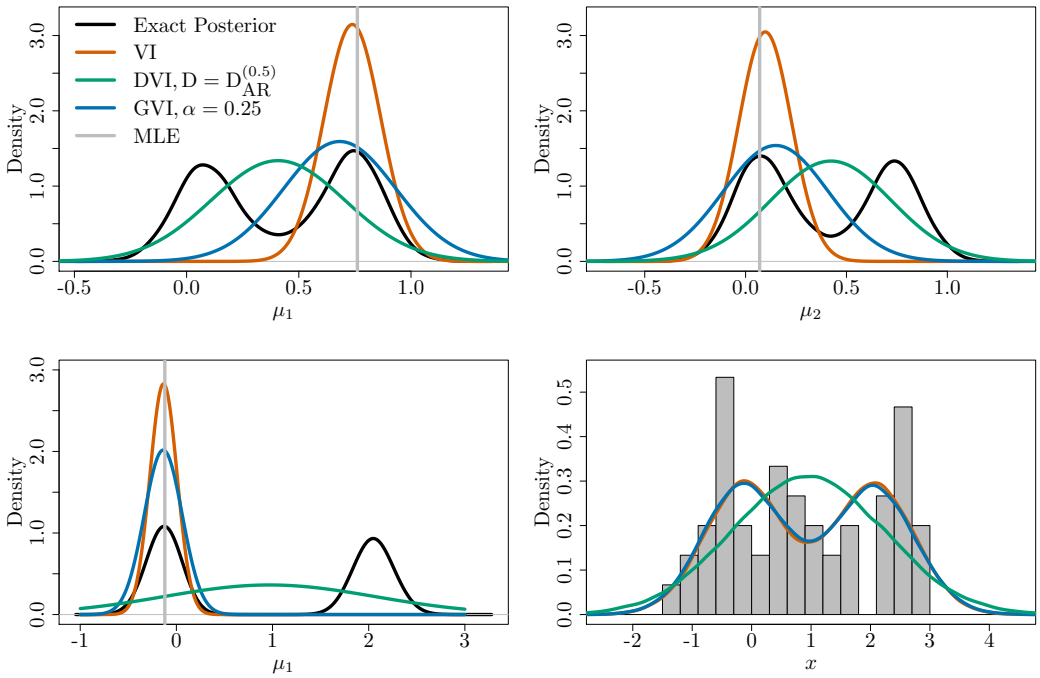


Figure 4.1: Best viewed in color. Depicted are inference outcomes for a Bayesian Mixture Model (BMM), namely the (multimodal) **standard Bayesian** posterior, **standard VI** posterior, a **DVI**-approximation based on minimizing $D_{AR}^{(\alpha)}$ between $\mathcal{Q}$ and $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ (Li and Turner, 2016), and a **GVI** posterior taking $D = D_{AR}^{(\alpha)}$. **Top:** Posterior marginals for $\mu_1 = 0, \mu_2 = 0.75$. The mode of the **DVI** posterior is a locally worst value for $\boldsymbol{\theta}$ relative to the **exact Bayesian** posterior. In contrast, **standard VI** and **GVI** respect the loss: They produce a posterior belief centered around one (of the two) values of $\boldsymbol{\theta}$ minimizing the loss. **Bottom left:** Posterior marginal for $\mu_1 = 0, \mu_2 = 2$. The effects of the top row become even stronger as the modes move further apart. **Bottom right:** Posterior predictive for $\mu_1 = 0, \mu_2 = 2$ against the histogram depicting the actual data. **VI**, **GVI** and **exact Bayesian** inference perform well and almost identically. **DVI** performs poorly, failing to capture the mixture components of the BMM.

the best parameter is found.

4.2 VI, DVI, and GVI: a common problem

While GVI, DVI, and VI posteriors have many important conceptual differences, they do have a major commonality: As their names suggest, they are all 'variational' methods. But what does this mean? The naming convention emphasizes that unlike most optimisation problems in Machine Learning and statistics, such methods

optimise over a space of functions: *variational calculus* is essentially the study of optimisation in function spaces. But what we will call variational methods in this thesis can be even more narrowly defined: in particular, the function spaces in question are parameterized sub-spaces $\mathcal{Q}$ of the space of probability distributions on $\boldsymbol{\Theta}$.

This means that fundamentally, all these methods aim at solving different variations of the same type of problem. In the remainder of the current chapter, we leverage this strong connection to adapt a range of techniques for GVI that have recently been popularized in the context of VI and DVI methods. Beyond that, we also discuss both challenges and solutions that are unique to GVI. For example, we will see that changing the loss function or divergence of a RoT problem sometimes yields closed form objectives and derivatives.

4.3 Background: How are VI posteriors computed?

Before we are ready to adapt the techniques of existing variational methods to the GVI setting, it is worth re-acquainting ourselves with the setup of standard VI. Simply put, a standard VI posterior is specified by $P(L, \mathrm{KLD}, \mathcal{Q})$ with the restrictions that (i) $L(\boldsymbol{\theta}, x_{1:n}) = -\log p(x_{1:n}|\boldsymbol{\theta})$ is a negative log likelihood for some likelihood function $p(\cdot|\boldsymbol{\theta})$, and that (ii) $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) : \boldsymbol{\kappa} \in \boldsymbol{K}\}$ is a set of distributions parametrized by a set $\boldsymbol{K}$ of parameters, where we typically have that $\boldsymbol{K} \subseteq \mathbb{R}^d$. This means that we aim at solving the problem

$$\boldsymbol{\kappa}^* \in \operatorname*{arg\,inf}_{\boldsymbol{\kappa} \in \boldsymbol{K}} \left\{ \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right] + \mathrm{KLD}(q \| \pi) \right\}, \tag{4.1}$$

since we can then conclude that $q(\boldsymbol{\theta}|\boldsymbol{\kappa}^*) = P(L, \mathrm{KLD}, \mathcal{Q})$. This is the origin for the (strictly speaking wrong but popular) phrase in Machine Learning that *Variational inference converts Bayesian inference from a sampling problem into an optimization problem.*[3]

4.3.1 Challenges in variational problems

Even though the optimization problem of (4.1) is typically *easier* to work with than $P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, it is not *easy* to work with. For example, it is typically impossible to prove that $\boldsymbol{\kappa}^*$ is unique—and in many cases even that it is attained

[3] The less sexy but more accurate version of this phrase would be that *Variational inference [over a parametric subset of $\mathcal{P}(\boldsymbol{\Theta})$] converts an infinite-dimensional into a finite-dimensional optimization problem.*

in $\boldsymbol{K}$. In particular, outside a number of niche cases the objective of (4.1) will not be convex—so even though gradient-based optimization methods are the de-facto default choice for variational posteriors, it is often unclear how good they actually are at finding the true optimum $\boldsymbol{\kappa}^*$. Further, it should be self-evident that taking gradients with respect to $\boldsymbol{\kappa}$ is itself a formidable challenge: Unless we can write down $\mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[L(\boldsymbol{\theta}, x_{1:n})\right]$ and $\mathrm{KLD}(q(\cdot|\boldsymbol{\kappa})\|\pi)$ in closed form, it is not obvious how the gradient should be computed in general.

In this thesis, we take the pragmatic approach of Machine Learning practitioners: we are comfortable taking a leap of faith, and simply hope that gradient-based methods for the variational problems we pose will generally lead to at least reasonable, and hopefully even close-to-optimal solutions. In practical terms, this means that we can focus on the actual computational devices—and in particular on gradient-based methodology.

4.3.2 Gradient-based methodology for VI

While early applications of VI relied on at least partially available closed forms for the expression in (4.1) and its gradient in $\boldsymbol{\kappa}$ (e.g. Ghahramani and Beal, 2001), the far more wide-spread approach today is stochastic optimization. The most attractive feature of stochastic optimization for VI is that modern computing equipment allows us to perform this optimization as a *black box*—an idea made explicit by Ranganath et al. (2014), but used to various degrees by a wide range of previous papers (e.g. Wingate and Weber, 2013; Kingma and Welling, 2013; Carbonetto et al., 2009; Titsias and Lázaro Gredilla, 2014) Specifically, if one can show that the gradient exists, then modern automatic differentiation software means that we do not have to derive this derivative by hand—a task that is not only tedious and time-consuming, but will have to be re-done for every new type of model we want to do inference for. The modern practitioner has many such automatic differentiation tools at their disposal, including `JAX` (James Bradbury et al., 2018), `NumPyro` (Phan et al., 2019; Eli Bingham et al., 2019), and `Tensorflow Probability` (Dillon et al., 2017).

The only remaining question is how one should compute or approximate the expectations relative to which the gradient needs to be taken. For this, we can use Monte Carlo gradient estimation—a vast field of research that has been studied for decades across different disciplines. For readers interested in details, the recent review of Mohamed et al. (2020) provides an excellent reference over this field of scientific inquiry.

The predominant Monte Carlo gradient estimator in the context of variational methods is the REINFORCE estimator (Williams, 1992)—which is also known

as the score function estimator (Kleijnen and Rubinstein, 1996; Rubinstein et al., 1996), as well as the likelihood ratio method (Glynn, 1990). To apply this estimator, rewrite the objective in (4.1) as a single expectation:

$$O_{\text{VI}}(\boldsymbol{\kappa}) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[L(\boldsymbol{\theta}, x_{1:n}) + \log q(\boldsymbol{\theta}|\boldsymbol{\kappa}) - \log \pi(\boldsymbol{\theta})\right].$$

Next, simply rewrite the gradient $\nabla_{\boldsymbol{\kappa}} O_{\text{VI}}(\boldsymbol{\kappa})$ as follows:

$$\begin{aligned}
\nabla_{\boldsymbol{\kappa}} O_{\text{VI}}(\boldsymbol{\kappa}) &= \nabla_{\boldsymbol{\kappa}} \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) \left[L(\boldsymbol{\theta}, x_{1:n}) + \log q(\boldsymbol{\theta}|\boldsymbol{\kappa}) - \log \pi(\boldsymbol{\theta})\right] d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\Theta}} \nabla_{\boldsymbol{\kappa}} \left\{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) \left[L(\boldsymbol{\theta}, x_{1:n}) + \log q(\boldsymbol{\theta}|\boldsymbol{\kappa}) - \log \pi(\boldsymbol{\theta})\right]\right\} d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\Theta}} \nabla_{\boldsymbol{\kappa}} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) \left[L(\boldsymbol{\theta}, x_{1:n}) - \log \pi(\boldsymbol{\theta})\right] d\boldsymbol{\theta} + \int_{\boldsymbol{\Theta}} \nabla_{\boldsymbol{\kappa}} \left\{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) \log q(\boldsymbol{\theta}|\boldsymbol{\kappa})\right\} d\boldsymbol{\theta} \\
&= \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\nabla_{\boldsymbol{\kappa}} \log q(\boldsymbol{\theta}|\boldsymbol{\kappa}) \left(L(\boldsymbol{\theta}, x_{1:n}) - \log \pi(\boldsymbol{\theta})\right)\right]. \qquad (4.2)
\end{aligned}$$

Here, the second equality follows due to the dominated convergence theorem[4], the third from the fact that $\nabla_{\boldsymbol{\kappa}} \log \pi(\boldsymbol{\theta}) = 0$, and the last from the derivation that $\int_{\boldsymbol{\Theta}} \nabla_{\boldsymbol{\kappa}} \left\{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) \log q(\boldsymbol{\theta}|\boldsymbol{\kappa})\right\} d\boldsymbol{\theta} = \nabla_{\boldsymbol{\kappa}} \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}[q(\boldsymbol{\theta}|\boldsymbol{\kappa})] = \nabla_{\boldsymbol{\kappa}} 1 = 0$; where we have once again used the dominated convergence theorem together with the fact that $\nabla_x \log f(x) = \frac{\nabla_x f(x)}{f(x)}$. Equation (4.2) is convenient, as it allows for a simple Monte Carlo estimator of $\nabla_{\boldsymbol{\kappa}} O_{\text{VI}}(\boldsymbol{\kappa})$ given by

$$\widehat{\nabla_{\boldsymbol{\kappa}} O_{\text{VI}}(\boldsymbol{\kappa})} = \frac{1}{S} \sum_{b=1}^{B} \nabla_{\boldsymbol{\kappa}} \log q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}) \left[L(\boldsymbol{\theta}^{(s)}, x_{1:n}) - \log \pi(\boldsymbol{\theta}^{(s)})\right],$$

where $\boldsymbol{\theta}^{(1:S)} \overset{i.i.d.}{\sim} q(\boldsymbol{\theta}|\boldsymbol{\kappa})$. Oftentimes, the variance of these gradients can be rather large; and so numerous variance reduction techniques have been proposed to curtail any negative side effects of using them.

While a host of other gradient-based estimators exist (such as pathwise gradient estimators or measure-valued gradients, see Mohamed et al., 2020), these are used only sparingly in practice. We therefore do not discuss them, and confine ourselves to adapting only the score-based gradient estimator to GVI.

4.4 Black Box GVI: stochastic computation for GVI

Standard VI is scalable using doubly stochastic, model-agnostic optimization techniques (e.g. Paisley et al., 2012; Hoffman et al., 2013; Titsias and Lázaro Gredilla,

[4]Numerous mild conditions will suffice to ensure that we can apply this theorem; and we will not dwell on them here since in practice, these will hold for all but the most pathological of cases

2014; Salimans and Knowles, 2014; Wu et al., 2019) collectively known as black box VI (Ranganath et al., 2014). We extend these methods and introduce black box GVI (BBGVI), an algorithm which is easily built into existing software: For example, adapting the Deep Gaussian Process implementation of Salimbeni and Deisenroth (2017) in Chapter 6 required <100 lines of Python code.

Suppose that for all $q \in \mathcal{Q}$, one can sample $\boldsymbol{\theta} \sim q$. Suppose also that the derivatives $\nabla_{\boldsymbol{\kappa}} \log q(\boldsymbol{\theta}|\boldsymbol{\kappa})$ and $\nabla_{\boldsymbol{\kappa}} D(q||\pi)$ exist (almost surely relative to the measure on Θ induced by any $q \in \mathcal{Q}$). Now, define the GVI objective corresponding to $P(L, D, \mathcal{Q})$ as

$$O_{\text{GVI}}(\boldsymbol{\kappa}) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[L(\boldsymbol{\theta}, x_{1:n})\right] + D(q\|\pi).$$

Lastly, assume that the conditions for the dominated convergence theorem are met so that $\nabla_{\boldsymbol{\kappa}} \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[L(\boldsymbol{\theta}, x_{1:n})\right] = \int_{\boldsymbol{\Theta}} \nabla_{\boldsymbol{\kappa}} \left\{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) L(\boldsymbol{\theta}, x_{1:n})\right\} d\boldsymbol{\theta}$. For many choices of D, $\mathcal{Q}$ and π that are of practical interest, $\nabla_{\boldsymbol{\kappa}} D(q||\pi)$ is available in closed form. In this case, GVI posteriors can be computed through gradient-based schemes built on the unbiased gradient estimate

$$\widehat{\nabla_{\boldsymbol{\kappa}} O_{\text{GVI}}} = \frac{1}{S} \sum_{s=1}^{S} \left\{ L(\boldsymbol{\theta}^{(s)}, x_{1:n}) \cdot \nabla_{\boldsymbol{\kappa}} \log q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}) \right\} + \nabla_{\boldsymbol{\kappa}} D(q||\pi), \tag{4.3}$$

where $\boldsymbol{\theta}^{(1:S)} \overset{i.i.d}{\sim} q(\boldsymbol{\theta}|\boldsymbol{\kappa})$. The derivation of this family of estimators follows the same logic as (4.2). Throughout the remainder of the thesis, all numerical examples and applications in Chapters 5 and 6 admit closed forms for $\nabla_{\boldsymbol{\kappa}} D(q||\pi)$. In other words, (4.3) provides the de-facto gradient estimator we use for all of this thesis' applications of GVI that require stochastic approximation.

If a closed form for $\nabla_{\boldsymbol{\kappa}} D(q||\pi)$ is not available but we can write $D(q||\pi) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})\right]$ as an expectation for a function $d_{\boldsymbol{\kappa},\pi} : \boldsymbol{\Theta} \to \mathbb{R}$, one can use the alternative unbiased gradient estimate

$$\widehat{\nabla_{\boldsymbol{\kappa}} O_{\text{GVI}}} = \frac{1}{S} \sum_{s=1}^{S} \Bigg\{ \left[L(\boldsymbol{\theta}^{(s)}, x_{1:n}) + d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}) \right] \cdot \nabla_{\boldsymbol{\kappa}} \log q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}) + \nabla_{\boldsymbol{\kappa}} d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}) \Bigg\}. \tag{4.4}$$

This logic applies to virtually all popular divergences, including all f-divergences. In particular, it is easy to see that this recovers the standard VI black box gradient for $d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) = \log q(\boldsymbol{\theta}|\boldsymbol{\kappa}) - \log \pi(\boldsymbol{\theta})$. In some cases however, divergences will not be linear in q so that one has $D(q\|\pi) = \tau\left(\mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})\right]\right)$ for some non-linear

function $\tau : \mathbb{R} \to \mathbb{R}$. In this case, BBGVI can be performed based on the biased gradient estimate

$$\widehat{\nabla_{\boldsymbol{\kappa}} O_{\text{GVI}}} = \frac{1}{S}\sum_{s=1}^{S}\left\{L(\boldsymbol{\theta}^{(s)}, x_{1:n}) \cdot \nabla_{\boldsymbol{\kappa}} \log(q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}))\right\} + \tau\left(\frac{1}{S}\sum_{s=1}^{S} d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)})\right) \cdot \frac{1}{S}\sum_{s=1}^{S} \nabla_{\boldsymbol{\kappa}} d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}). \tag{4.5}$$

Note that the induced bias of this estimator could be eliminated using the techniques of Grathwohl et al. (2018).

4.4.1 Closed forms for the divergence term

It should be obvious that the implied reduction in variance from having $D(q\|\pi)$ available in closed forms makes using (4.3) preferable over both (4.4) and (4.5). Naturally, this raises the question which divergences will be available in closed form.

One family of robust divergences—called the $\alpha\beta\gamma$-family by Cichocki and Amari (2010), which includes α-divergences ($D_A^{(\alpha)}$), Rényi's α-divergences, β-divergences ($D_B^{(\beta)}$), as well as γ-divergences ($D_G^{(\gamma)}$)—is of particular interest for robustness to ill-specified priors. We defer their detailed discussion and formal definitions to Chapter 5 and specifically Section 5.1.1, since discussing them more thoroughly makes more sense later on in the thesis. While these divergences are discussed later, we provide Proposition 4.1 at this stage, since this result shows that $\nabla_{\boldsymbol{\kappa}} D(q\|\pi)$ will generally be available for $\alpha\beta\gamma$-divergences if all elements of $\mathcal{Q}$ are part of the same exponential family.

Proposition 4.1 (Closed form D)**.** Let q, π with natural parameters $\boldsymbol{\eta}_q, \boldsymbol{\eta}_\pi$ be in the exponential family $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\eta}) = h(\boldsymbol{\theta}) \exp\{\boldsymbol{\eta}' T(\boldsymbol{\theta}) - A(\boldsymbol{\eta})\} : \eta \in \mathcal{N}\}$ with natural parameter space $\mathcal{N} = \{\boldsymbol{\eta} : \exp\{A(\boldsymbol{\eta})\} < \infty\}$. Then,

(1) $D_A^{(\alpha)}(q||\pi)$ and $D_{AR}^{(\alpha)}(q||\pi)$ have closed form if $\alpha \in (0,1)$, or $\alpha\boldsymbol{\eta}_q + (1-\alpha)\boldsymbol{\eta}_\pi \in \mathcal{N}$

(2) $D_B^{(\beta)}(q||\pi)$ has a closed form if $h(\boldsymbol{\theta}) = h$ does not depend on $\boldsymbol{\theta}$ and additionally, $(\beta - 1) \cdot \boldsymbol{\eta}_1 + \boldsymbol{\eta}_2 \in \mathcal{N}$ for any $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathcal{N}$

(3) $D_G^{(\gamma)}(q||\pi)$ has closed form if $D_B^{(\beta)}(q||\pi)$ does for $\beta = \gamma$.

The above Proposition is meaningful, since exponential families are typically of particular interest in Variational Inference schemes due to their computational

convenience. In the context of Chapters 5 and 6, it means that we can use (4.3) throughout all the experiments presented therein.

4.4.2 Black box variance reduction

Regardless of the exact form the gradient estimator takes, one may want to apply black box variance reduction techniques in order to make inference both faster and more reliable—even if the estimate in (4.3) can be used. In the context of standard variational methods, more-or-less black box techniques for reducing variance have previously been suggested by numerous authors, including Ranganath et al. (2014); Wu et al. (2019); Grathwohl et al. (2018). An overview over such techniques for the standard variational case is given in Mohamed et al. (2020).

Preliminaries and assumptions

In the remainder, we will find it helpful to distinguish a number of different cases for variance reduction techniques. In general, most black box variance reduction techniques for standard VI rely to varying degrees on three assumptions. These are often not stated explicitly in the relevant papers, but of crucial importance to assess which techniques we can transfer from standard VI into GVI problems. For the purposes of what follows, we will state these assumptions explicitly as (A1)–(A3) in the notation and context of GVI, and then proceed with explaining how we can obtain GVI posteriors based on different sets of these assumptions holding true.

(A1) We can factorize the variational family into k independent factors as $\mathcal{Q} = \{q(\boldsymbol{\theta}|\boldsymbol{\kappa}) = \prod_{j=1}^{k} q_j(\boldsymbol{\theta}_j|\boldsymbol{\kappa}_j) : \boldsymbol{\kappa}_j \in K_j \text{ for all } j : 1 \leq j \leq k\}$.

(A2) For the k factors $\boldsymbol{\theta}_j : 1 \leq j \leq k$, we have $\boldsymbol{\theta}_{(j)}$ so that $\boldsymbol{\theta}_j \cap \boldsymbol{\theta}_{(j)} = \emptyset$, and for which we can additively decompose $L(\boldsymbol{\theta}, x_{1:n}) = L^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_{1:n}) + L^{(-j)}(\boldsymbol{\theta}_{-j}, x_{1:n})$. Here, $L^{(j)}$ is an additive component of the loss L that only depends on the j-th factor and $\boldsymbol{\theta}_{(j)}$, while $L^{(-j)}$ is an additive component of the loss that may depend on all of $\boldsymbol{\theta}$ except for its j-th factor. In particular, $L^{(-j)}$ may depend on $\boldsymbol{\theta}_{(j)}$, but *cannot* depend on $\boldsymbol{\theta}_j$.

(A3) $D = \frac{1}{w} \cdot \text{KLD}$ (with $w = 1$ for standard VI).

Note that (A1) is always satisfied for both standard VI and GVI, because any variational family factorizes into at least a single factor. In contrast, note that (A2) does not even necessarily hold. Its meaning is that there are k ways to rewrite the loss function additively as $L^{(j)} + L^{(-j)}$, so that each of these k ways splits up

the parameters into three different blocks: $\boldsymbol{\theta}_j$, $\boldsymbol{\theta}_{(j)}$, and $\boldsymbol{\theta}_{-j}$. This means we can rewrite $L(\boldsymbol{\theta}, x_{1:n}) = \frac{1}{k}\sum_{j=1}^{k} L^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_{1:n}) + L^{(-j)}(\boldsymbol{\theta}_{-j}, x_{1:n})$. Note that even if $L(\boldsymbol{\theta}, x_{1:n}) = -\log p(x_{1:n}|\boldsymbol{\theta})$, this form of decomposability will not generally hold unless one imposes some conditional independence structure on the factors $\boldsymbol{\theta}_j$—in which case $\boldsymbol{\theta}_{(j)}$ is called the *Markov blanket* of $\boldsymbol{\theta}_j$. Since the additivity of conditionally independent components is a simple consequence of the log score associated with standard VI, for GVI the interpretation of the additivity property as conditional independence does not generally hold. Conversely, it also means that additivity could hold for parameters that are *not* conditionally independent in a probability model (and indeed for parameters that are not part of a probability model at all). Generally then, both (A2) and (A3) do not necessarily hold for GVI. If they do however, they can greatly simplify BBGVI or improve its numerical performance.

In addition to the assumptions on parameter factorization, we also need to discern two settings regarding the divergence D:

(D1) $\nabla_{\boldsymbol{\kappa}} D(q\|\pi)$ has closed form for all $q \in \mathcal{Q}$;

(D2) $D(q\|\pi) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\ell^{D}_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})\right]$ for some function $\ell^{D}_{\boldsymbol{\kappa},\pi} : \boldsymbol{\Theta} \to \mathbb{R}$.

Under each condition, we find a different solution using as much of the available information as possible to improve inference outcomes.

Standard black box VI with (A2) and (A3)

Clearly, if the regularizer used is still a rescaled version of the KLD, (D2) always holds, so that one recovers an internally rescaled version of the classical VI objective in (4.2).[5] Naturally, if (D1) holds, then this gradient can be made even more elegant using a version of the gradient estimate presented in (4.3). Next, we turn attention to the cases that are more interesting: If (A3) does not hold (so that $D \neq$ KLD) and when the losses are not necessarily negative log likelihoods, meaning that (A2) requires more careful consideration.

BBGVI under (A2): Rao-Blackwellization

First, recall that if we are *not* using the additional information in (A2), then we obtain the objective of (4.3) if (D1) holds. Similarly, we obtain (4.4) if (D2) holds.

However, one can employ Rao-Blackwellization for variance reduction if the losses satisfy (A2). The first step is to rewrite for $q_{-j}(\boldsymbol{\theta}_{-j}|\boldsymbol{\kappa}_{-j}) = \prod_{l=1,l\neq j}^{k} q_l(\boldsymbol{\theta}_l|\boldsymbol{\kappa}_l)$

[5] Simply replace $\log \pi(\boldsymbol{\theta})$ with $w^{-1}\log \pi(\boldsymbol{\theta})$

the partial derivatives as

$$\nabla_{\boldsymbol{\kappa}_j} O_{\text{GVI}}(\boldsymbol{\kappa}) = \nabla_{\boldsymbol{\kappa}_j} \mathbb{E}_{q_j(\boldsymbol{\theta}_j|\boldsymbol{\kappa}_j)} \left[\mathbb{E}_{q_{-j}(\boldsymbol{\theta}_{-j}|\boldsymbol{\kappa}_{-j})} \left[L(\boldsymbol{\theta}, x_{1:n}) + D(q\|\pi)|\boldsymbol{\theta}_j \right] \right].$$

The hope is then to get around computing as many of the inner expectations over $q_{-j}(\boldsymbol{\theta}_{-j}|\boldsymbol{\kappa}_{-j})$ as possible. Assume for the moment that at least (D2) holds. Further, denote $q_{-j}(\boldsymbol{\theta}_{-j}|\boldsymbol{\kappa}_{-j}) = q_{-j}$, $q_j(\boldsymbol{\theta}_j|\boldsymbol{\kappa}_j) = q_j$, and in similar fashion the distributions $q_{(j)}$, $q_{-(j)}$, q. Now, assuming that (A2) holds relative to the factors $\boldsymbol{\theta}_j$, one finds

$$\begin{aligned}\nabla_{\boldsymbol{\kappa}_j} O_{\text{GVI}}(\boldsymbol{\kappa}) = \mathbb{E}_{q_j} \Bigg[\nabla_{\boldsymbol{\kappa}_j} \log(q_j(\boldsymbol{\theta})) \Bigg(\mathbb{E}_{q_{-j}} \left[L^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_{1:n}) \right] + \mathbb{E}_{q_{-j}}[L^{(-j)}(\boldsymbol{\theta}_{-j}, x_{1:n})] \\ + \mathbb{E}_{q_{-j}}[d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})] \Bigg) \Bigg] + \mathbb{E}_{q_{-j}}[\nabla_{\boldsymbol{\kappa}_j} d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})].\end{aligned}$$

Observing that $\mathbb{E}_{q_j}[\nabla_{\boldsymbol{\kappa}_j} \log(q_j(\boldsymbol{\theta}_j))] = 0$ and that $\mathbb{E}_{q_{-j}}[L^{-(j)}(\boldsymbol{\theta}_{-j})]$ is constant in $\boldsymbol{\theta}_j$ by definition of $L^{-(j)}$ and $\boldsymbol{\theta}_{-j}$, this drastically simplifies to

$$\begin{aligned}\nabla_{\boldsymbol{\kappa}_j} O_{\text{GVI}}(\boldsymbol{\kappa}) = \mathbb{E}_{q_j} \Bigg[\nabla_{\boldsymbol{\kappa}_j} \log q_j(\boldsymbol{\theta}_j|\boldsymbol{\kappa}_j) \cdot \Bigg(\mathbb{E}_{q_{-j}} \left[L^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_{1:n}) \right] + \mathbb{E}_{q_{-j}} \left[d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) \right] \Bigg) \\ + \mathbb{E}_{q_{-j}} \left[\nabla_{\boldsymbol{\kappa}_j} d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) \right] \Bigg].\end{aligned}$$

Next, observe that by virtue of how $L^{(j)}$ was constructed, it holds that we can also simplify

$$\mathbb{E}_{q_j} \left[\nabla_{\boldsymbol{\kappa}_j} \log(q_j) \mathbb{E}_{q_{-j}} \left[L^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_{1:n}) \right] \right] = \mathbb{E}_{q_{(j)}} \left[L^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_{1:n}) \right].$$

Putting the above together, we finally arrive at

$$\begin{aligned}&\nabla_{\boldsymbol{\kappa}_j} O_{\text{GVI}}(\boldsymbol{\kappa}) \\ &= \mathbb{E}_{q_j} \left[\nabla_{\boldsymbol{\kappa}_j} \log q_j(\boldsymbol{\theta}_j) \left(\mathbb{E}_{q_{-j}} \left[L^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_{1:n}) \right] + \mathbb{E}_{q_{-j}}[d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})] \right) + \mathbb{E}_{q_{-j}}[\nabla_{\boldsymbol{\kappa}_j} d_{\boldsymbol{\kappa},\pi}] \right] \\ &= \mathbb{E}_{q_{(j)}} \left[\nabla_{\boldsymbol{\kappa}_j} \log q_j(\boldsymbol{\theta}_j) L^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_{1:n}) \right] + \mathbb{E}_q \left[\nabla_{\boldsymbol{\kappa}_j} \log q_j(\boldsymbol{\theta}_j) d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\kappa}_j} d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}) \right].\end{aligned}$$

which is the final form under (D1). Should (D1) hold, one can instead use the lower variance estimate

$$\nabla_{\boldsymbol{\kappa}_j} O_{\text{GVI}}(\boldsymbol{\kappa}) = \mathbb{E}_{q_{(j)}} \left[\nabla_{\boldsymbol{\kappa}_j} \log q_j(\boldsymbol{\theta}_j) L^{(j)}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{(j)}, x_{1:n}) \right] + \nabla_{\boldsymbol{\kappa}_j} D(q\|\pi).$$

The k terms $\nabla_{\boldsymbol{\kappa}_j} O_{\text{GVI}}(\boldsymbol{\kappa})$ can then be combined into a global gradient estimate

simply by setting

$$\nabla_{\boldsymbol{\kappa}} O_{\mathrm{GVI}}(\boldsymbol{\kappa}) = \left(\nabla_{\boldsymbol{\kappa}_1} O_{\mathrm{GVI}}(\boldsymbol{\kappa}), \nabla_{\boldsymbol{\kappa}_2} O_{\mathrm{GVI}}(\boldsymbol{\kappa}), \ldots \nabla_{\boldsymbol{\kappa}_k} O_{\mathrm{GVI}}(\boldsymbol{\kappa})\right)^T.$$

As before, one will in practice need to approximate the gradients with a sample $\boldsymbol{\theta}^{(1:S)}$ drawn from $q(\boldsymbol{\theta}|\boldsymbol{\kappa})$. The relevant terms are computed as

$$\nabla_{\boldsymbol{\kappa}_j} \widehat{O_{\mathrm{GVI}}}(\boldsymbol{\kappa}) = \frac{1}{S} \sum_{s=1}^{S} \left\{ \nabla_{\boldsymbol{\kappa}_j} \log q_j(\boldsymbol{\theta}_j^{(s)}|\boldsymbol{\kappa}_j) L^{(j)}(\boldsymbol{\theta}_j^{(s)}, \boldsymbol{\theta}_{(j)}^{(s)}, x_i) \right\} + \nabla_{\boldsymbol{\kappa}_j} D(q\|\pi)$$

for some closed form function $\nabla_{\boldsymbol{\kappa}_j} D(q\|\pi)$. If (D2) holds and there is no closed form for the prior regularizer, $\nabla_{\boldsymbol{\kappa}_j} D(q\|\pi)$ is replaced by the stochastic estimator

$$\nabla_{\boldsymbol{\kappa}_j} \widehat{D(q\|\pi)} = \frac{1}{S} \sum_{s=1}^{S} \left\{ \nabla_{\boldsymbol{\kappa}_j} \log q_j(\boldsymbol{\theta}_j^{(s)}|\boldsymbol{\kappa}_j) d_{\pi,\boldsymbol{\kappa}}(\boldsymbol{\theta}^{(s)}) + \nabla_{\boldsymbol{\kappa}_j} d_{\pi,\boldsymbol{\kappa}}(\boldsymbol{\theta}^{(s)}) \right\}.$$

We end this section by making the meaning of (A2) more tangible for the case of general losses through a short example in the context of multivariate regression.

Example 4.2 (Additivity as per (A2) for general losses)**.** Suppose each $x_i = (x_{i,1}, x_{i,2}, x_{i,3})'$ consists of three measurements that we wish to relate to some other observables y_i through

$$\begin{aligned} x_{i,1} &= a + y_i b + \xi_1 \\ x_{i,2} &= b + y_i c + \xi_2 \\ x_{i,3} &= d + \xi_3 \end{aligned}$$

where ξ_j are unknown slack variables (or errors), the parameters of interest are $\boldsymbol{\theta} = (a, b, c, d, e)$ and we wish to produce a belief distribution over $\boldsymbol{\theta}$ that is informative about good values of $\boldsymbol{\theta}$ relative to some prediction loss $L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)$, where the individual loss terms are

$$\begin{aligned} \ell(\boldsymbol{\theta}, x_i) = & \underbrace{\|f_1^1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{(1)}, y_i) - x_{i,1}\|_p^p + \|f_2^1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{(1)}, y_i) - x_{i,2}\|_p^p}_{\text{first factor, } j=1} \\ + & \underbrace{\|f_3^2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_{(2)}, y_i) - x_{i,3}\|_p^p}_{\text{second factor, } j=2}, \end{aligned}$$

where $\|\cdot\|_p^p$ denotes some p-norm for $p \geq 1$ and f_l^j seeks to predict only the l-th dimension of $x_i = (x_{i,1}, x_{i,2}, x_{i,3})^T$ by means of the j-th factor $\boldsymbol{\theta}_j$ as well as $\boldsymbol{\theta}_{(j)}$. Suppose that f_l^j will correspond to the l-th row written down in the above model

for x_i (excluding of course the error term), which means that

$$\begin{aligned} f_1^1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{(1)}) &= a + y_i b \\ f_2^1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{(1)}, y_i) &= b + y_i c \\ f_3^2(\boldsymbol{\theta}_2, \boldsymbol{\theta}_{(2)}, y_i) &= d \end{aligned}$$

In this case, the two factors of $\boldsymbol{\theta}$ will clearly be given by

$$\boldsymbol{\theta}_1 = (a, b, c)^T, \qquad \boldsymbol{\theta}_2 = (d),$$

and both $\boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{(2)} = \emptyset$.

BBGVI if neither (A2) nor (A3) hold

It is of course possible that neither (A2) nor (A3) hold. Even if the assumptions do hold, it may simply be convenient to build a software implementation of BBGVI that does not impose any assumptions. Naturally, one can use the gradient estimates of (4.3)–(4.5) in this case. Beyond that, one can also apply black box variance reduction techniques that work without the assumptions underlying Rao-Blackwellization. The next paragraphs present these techniques, which are adapted from the standard variational case as presented in Ranganath et al. (2014).

Generically applicable variance reduction

While the Rao-Blackwellization variance reduction will generally be more effective, some variance reduction techniques can work in circumstances where Rao-Blackwellization does not. Conversely, this means that if Rao-Blackwellization is applicable, one can actually deploy two variance reduction schemes at once to substantially speed up convergence. The control variate we use is simply

$$h(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\kappa}} \log q(\boldsymbol{\theta}|\boldsymbol{\kappa})$$

with an optimal scaling parameter that can be estimated as

$$\hat{a}^* = \frac{\sum_{s=1}^S \widehat{\mathrm{Cov}}(R(\boldsymbol{\theta}^{(s)}), h(\boldsymbol{\theta}^{(s)}))}{\sum_{s=1}^S \widehat{\mathrm{Var}}(h(\boldsymbol{\theta}^{(s)}))},$$

where we have generically written $R(\boldsymbol{\theta}^{(s)})$ as the stochastically estimated part of the gradient estimate that can be decomposed additively. For example, in the estimator of (4.3), we would have $R(\boldsymbol{\theta}^{(s)}) = L(\boldsymbol{\theta}^{(s)}, x_{1:n}) \cdot \nabla_{\boldsymbol{\kappa}} \log q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa})$, while we would

Algorithm 1 Black box GVI (BBGVI)

Input: $x_{1:n}$, π, D, ℓ, $\mathcal{Q}$, h, StoppingCriterion, $\boldsymbol{\kappa}_0$, K, S, $t = 0$, LearningRate

done $\leftarrow$ False
while not done **do**

// Step 1: Get a subsample from $x_{1:n}$ of size K
$\rho_{1:K} \leftarrow \text{SampleWithoutReplacement}(1:n, K)$
$x(t)_{1:K} \leftarrow x_{\rho_{1:K}}$

// Step 2: Sample from $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_t)$ and compute losses
$\boldsymbol{\theta}^{(1:S)} \overset{i.i.d.}{\sim} q(\boldsymbol{\theta}|\boldsymbol{\kappa}_t)$
$\ell_{i,s} \leftarrow \ell(\boldsymbol{\theta}^{(s)}, x(t)_i) \cdot \nabla_{\boldsymbol{\kappa}_t} \log q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}_t)$ for all $s = 1, 2, \dots S$ and $i = 1, 2, \dots, K$
$\ell_s \leftarrow \frac{n}{K} \sum_{i=1}^{K} \ell_{i,s}$ for all $s = 1, 2, \dots S$

// Step 3: Compute divergence term
if $D(q\|\pi)$ admits closed form **then**
$\ell_s \leftarrow \ell_s + \nabla_{\boldsymbol{\kappa}} D(q\|\pi)$ for all $s = 1, 2, \dots S$
else if $D(q\|\pi) = \mathbb{E}_q[\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})]$ **then**
$\ell_s \leftarrow \ell_s + \ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)}) \nabla_{\boldsymbol{\kappa}_t} \log q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}_t) + \nabla_{\boldsymbol{\kappa}_t} \ell^D_{\boldsymbol{\kappa}_t,\pi}(\boldsymbol{\theta}^{(s)})$ for all $s = 1, 2, \dots S$
else if $D(q\|\pi) = \tau\left(\mathbb{E}_q[\ell^D_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta})]\right)$ **then**
$\ell_s \leftarrow \ell_s + \tau\left(\frac{1}{S}\sum_{s=1}^{S} \ell^D_{\boldsymbol{\kappa}_t,\pi}(\boldsymbol{\theta}^{(s)})\right) \cdot \nabla \boldsymbol{\kappa}_t \ell^D_{\boldsymbol{\kappa}_t,\pi}(\boldsymbol{\theta}^{(s)})$ for all $s = 1, 2, \dots S$

// Step 4: Apply variance reduction via h if desired
if $h \neq$ None **then**
$h_s \leftarrow h(\boldsymbol{\theta}^{(s)}, \ell_s)$
$\ell_s \leftarrow \ell_s - h_s$ for all for all $s = 1, 2, \dots S$

// Step 5: Update $\boldsymbol{\kappa}_t$ and stopping criterion
$\rho_t \leftarrow \text{LearningRate}(t)$
$L \leftarrow \frac{1}{S}\sum_{s=1}^{S} \ell_s$
$\boldsymbol{\kappa}_{t+1} \leftarrow \boldsymbol{\kappa}_t + \rho_t \cdot L$
done $\leftarrow \text{StoppingCriterion}(\boldsymbol{\kappa}_{t+1}, \boldsymbol{\kappa}_t, t)$
$t \leftarrow t + 1$

have $R(\boldsymbol{\theta}^{(s)}) = \left[L(\boldsymbol{\theta}^{(s)}, x_{1:n}) + d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)})\right] \cdot \nabla_{\boldsymbol{\kappa}} \log q(\boldsymbol{\theta}^{(s)}|\boldsymbol{\kappa}) + \nabla_{\boldsymbol{\kappa}} d_{\boldsymbol{\kappa},\pi}(\boldsymbol{\theta}^{(s)})$ for the estimator of (4.4). Based on this, one can then compute the black box variance-

reduced term $\widehat{O_{\text{GVI,VR}}(\boldsymbol{\kappa})}$ from $\widehat{O_{\text{GVI}}(\boldsymbol{\kappa})}$ as

$$\widehat{O_{\text{GVI,VR}}(\boldsymbol{\kappa})} = \widehat{O_{\text{GVI}}(\boldsymbol{\kappa})} - \hat{a}^* \cdot \sum_{s=1}^{S} h(\boldsymbol{\theta}^{(s)}).$$

Of course, the exact same logic could be applied to already Rao-Blackwellized terms—thereby reducing the variance twice.

Algorithm 1 summarizes a generic BBGVI procedure. Because this allows an additional layer of speed-up via data sub-sampling, we will assume throughout the algorithm that $L(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i)$ for some additively decomposable loss ℓ.[6] The algorithm is adaptable to the non-additive loss case in an obvious way. Similarly, the algorithm could be modified to apply variance reduction through Rao-Blackwellization before the black box variance reduction in Step 4 is applied.

4.5 Pseudo-conjugate GVI objectives

While the majority of GVI posteriors will need to be computed using the stochastic approximations outlined in the previous section, we can sometimes obtain closed forms for $\nabla_{\boldsymbol{\kappa}} O_{\text{GVI}}(\boldsymbol{\kappa})$. One such special case arises from losses based on the β- and γ-divergences, which will be formally introduced in Chapter 6. Pseudo-conjugate objectives are an edge case, and better introduced for the settings in which they are used. For this reason, we defer the conditions for closed forms in the case of the β-divergence to Theorem 7.3 in Chapter 7, where the result is used extensively for on-line inference in changepoint models. Note that Proposition 4.2 and Theorem 7.3 are also extended in Chapter 6 (see Theorem 6.1) to both the β- and γ-divergence loss for the special case of Deep Gaussian Processes (DGPs).

Since the result for the γ-divergence loss is of independent interest for computing GVI posteriors (albeit not used in any particular application in this thesis), we state the result here. The conditions for the β- and γ-divergences are relatively similar, and essentially require that the prior $\pi(\boldsymbol{\theta})$ is conjugate to the likelihood $p(\cdot|\boldsymbol{\theta})$ together with some additional minor regularity conditions.

Proposition 4.2 (Closed form GVI objectives with the γ-divergences)**.** Let $\mathcal{L}_p^{\beta}$ be the γ-divergence based scoring rule for likelihood $p(\cdot|\boldsymbol{\theta})$ given by $L^{\gamma}(\boldsymbol{\theta}, x_{1:n}) =$

[6] The idea of data sub-sampling is that for large enough n, it will be wasteful to evaluate all individual loss terms $\ell(\boldsymbol{\theta}, x_i)$ so that we instead work with $L \approx \frac{1}{m}\sum_{j=1}^{m} \ell(x'_j|\boldsymbol{\theta})$, where $m << n$ and x'_j are i.i.d. draws from the empirical measure formed by $x_{1:n}$.

$\sum_{i=1}^{n} \mathcal{L}_p^\gamma(\boldsymbol{\theta}, x_i)$ for

$$\mathcal{L}_p^\gamma(\boldsymbol{\theta}, x_i) = -\frac{1}{\gamma - 1} p(x_i|\boldsymbol{\theta})^{\gamma-1} \cdot \frac{\gamma}{I_{p,\gamma}(\boldsymbol{\theta})^{\frac{\gamma-1}{\gamma}}}; \qquad I_{p,\gamma}(\boldsymbol{\theta}) = \int_{\mathcal{X}} p(x_i|\boldsymbol{\theta})^\gamma dx$$

Suppose $p(\cdot|\boldsymbol{\theta})$ admits conjugacy relative to the exponential distributions given by $\mathcal{Q}$ and let the conjugate prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0) \in \mathcal{Q}$. Writing

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= h(x) \exp\left\{g(x)^T T(\boldsymbol{\theta}) - B(x)\right\}, \\ q(\boldsymbol{\theta}|\boldsymbol{\kappa}) &= h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\}, \\ \mathcal{N} &= \{\boldsymbol{\kappa} : \exp\{A(\eta(\boldsymbol{\kappa}))\} < \infty\}, \end{aligned}$$

the objective of $P(\mathcal{L}^\gamma, \text{KLD}, \mathcal{Q})$ has closed form if for observations $x_{1:n}$ and all $q \in \mathcal{Q}$

$$I_{p,\gamma}(\boldsymbol{\theta}) = \int_{\mathcal{X}} p(x_i|\boldsymbol{\theta})^\gamma dx, \quad F_1(\boldsymbol{\kappa}) = \int_{\boldsymbol{\Theta}} T(\boldsymbol{\theta}) q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta}, \quad F_2(\boldsymbol{\kappa}) = \int_{\boldsymbol{\Theta}} I^{(\gamma)}(\boldsymbol{\theta})^{\frac{1-\gamma}{\gamma}} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta}$$

are closed form functions of $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$ for all x_i such that $(\eta(\boldsymbol{\kappa}) + (\gamma - 1)g(x_i)) \in \mathcal{N}$.

The proof of this result can be found in Appendix B.4

Chapter 5

Generalized Variational Inference, Part 2: Regularizer

Summary: In this, the preceding as well as the following chapter, we study the special case for the Rule of Three (RoT) for which optimization happens over a set of parameterized distributions. For its obvious relationship to previous variational methods, we call this family of algorithms Generalized Variational Inference (GVI). Broadly speaking, this second of three chapters of GVI explains how this methodology can address poorly specified priors in Bayesian methods, and specifically in the context of modern Machine Learning models. The benefits of GVI are explored by approximate bounds that quantify the difference between GVI and standard Variational inference (VI), as well as an extensive empirical evaluation. To conclude, we also study how GVI can help with poorly specified priors in Bayesian Neural Networks (BNNs)—a canonical application of significant interest in the world of Bayesian Machine Learning.

5.1 Quantifying the difference between VI and GVI

While it is clear that GVI posteriors do not aim to approximate the standard Bayes posterior $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta})$ or even necessarily a Gibbs posterior $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$, it is reasonable to expect that $P(L, D, \mathcal{Q})$ will be close to the variational approximation $q^*_{\mathrm{VI}}(\boldsymbol{\theta}) = P(L, \mathrm{KLD}, \mathcal{Q})$ of $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta}) = P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$ whenever $D(\cdot\|\pi) \approx \mathrm{KLD}(\cdot\|\pi)$. This is relevant because often, we will choose $D = D^h$ as a divergence parameterized by some hyperparameter h so that $\lim_{h\to 1} D^h(q\|\pi) = \mathrm{KLD}(q\|\pi)$ for any fixed $q, \pi \in \mathcal{P}(\boldsymbol{\Theta})$.

5.1.1 Parameterized divergences

This phenomenon is illustrated in Figure 5.1 with three different divergences: Rényi's α-divergence, the β-divergence—sometimes also known as density power divergence (see Jones et al., 2001)—as well as the γ-divergence. These are parameterized divergences that recover the KLD as their parameterization approaches unity. While one can unify all these divergences into a three-parameter family called the $\alpha\beta\gamma$-divergence (see Cichocki and Amari, 2010), we elect not to present them in this way, as it is not useful for what we set out to do here. Instead, we present them as variants of the KLD based on generalized notions of the log function. Specifically, it turns out that the parameterized robust divergences of the $\alpha\beta\gamma$-family are all based on various parameterized 'generalized' log functions $\log^h$ with the property that $\lim_{h\to 1} \log^h(x) = \log(x)$. The motivation for this is the role of the log function in the KLD, which is defined as

$$\mathrm{KLD}(q\|\pi) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\left(\frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)\right]. \tag{5.1}$$

In other words, these divergences confer robustness by using a function close (but not equal to) the log function, and then mimicking the behaviour of the KLD. Specifically, all of them use the so-called *replica trick*, which says that

$$\lim_{h\to 1} \frac{x^h - 1}{h} = \log(x).$$

For the readers interested in seeing how this works for each of the divergences we introduce in the remainder, we refer to Cichocki and Amari (2010).

Throughout, we will study several divergences constructed with this logic: The arguably most important of them is Rényi's α-divergence (Rényi, 1961). We denote it by $D_{AR}^{(\alpha)}$ and use the parameterization of Liese and Vajda (1987); Cichocki and Amari (2010) rather than its original parameterization because it links the divergence more obviously to other robust alternatives of the KLD.

Definition 5.1 (Rényi's α-divergence ($D_{AR}^{(\alpha)}$) (Rényi, 1961))**.** Rényi's α-divergence is defined as

$$D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(\alpha-1)} \log\left(\int q(\boldsymbol{\theta})^{\alpha}\pi(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta}\right), \tag{5.2}$$

where $\alpha > 0$.

Originally, Rényi's α-divergence was motivated as the *geometric mean* information to discriminate between the two hypotheses $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$ of order

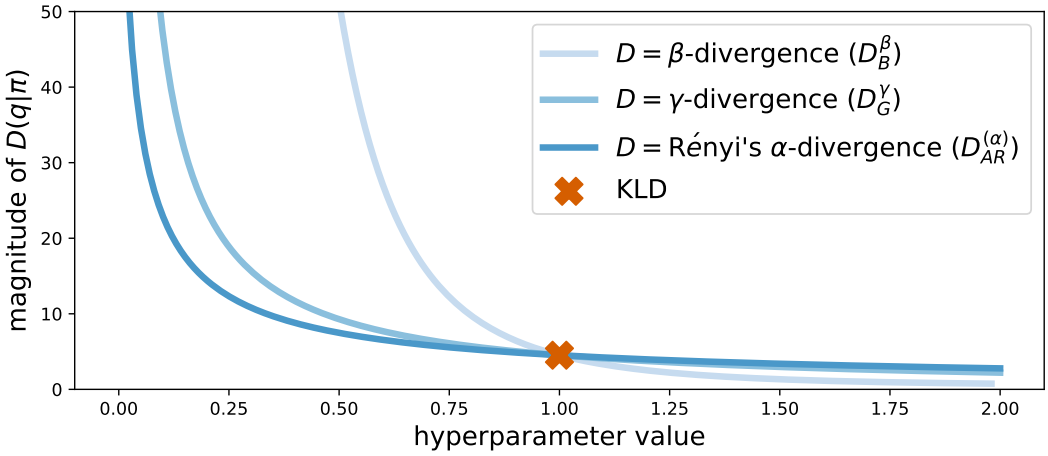


Figure 5.1: Depicted is the magnitude $D(q\|\pi)$ for different **robust divergences** ***D*** and the **KLD** for two Normal Inverse Gamma distributions given by $q(\boldsymbol{\theta}) = \mathcal{NI}^{-1}(\boldsymbol{\theta}; \boldsymbol{\mu}_q, \boldsymbol{V}_q, a_q, b_q)$ and $\pi(\boldsymbol{\theta}) = \mathcal{NI}^{-1}(\boldsymbol{\theta}; \boldsymbol{\mu}_\pi, \boldsymbol{V}_\pi, a_\pi, b_\pi)$ with $\boldsymbol{\mu}_\pi = (0,0)^T$, $\boldsymbol{V}_\pi = 25 \cdot I_2$, $a_\pi = 500$, $b_\pi = 500$ and $\boldsymbol{\mu}_q = (2.5, 2.5)^T$, $\boldsymbol{V}_q = 0.3 \cdot I_2$, $a_q = 512$, $b_q = 543$.

α. This is directly based on the original motivations for the KLD, which itself was interpreted and justified as the *arithmetic mean* information to discriminate between $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$ (Kullback and Leibler, 1951).[1] Intuitively speaking, geometric means are generally more robust measures of central tendency than arithmetic means (so long as $\alpha \in (0,1)$), and so the $D_{AR}^{(\alpha)}$ is a more robust discrepancy measure for $\alpha \in (0,1)$.

Rényi's α-divergence is arguably the most well-known divergence seeking to robustify the KLD, but it is not the oldest. This honour falls to the α-divergence, whose special case for $\alpha = 0.5$ is well-known as the Hellinger Distance. The α-divergence is also the only of the parameterized robust divergences of the $\alpha\beta\gamma$-family that is part of the family of f-divergences.

Definition 5.2 (The α-divergence ($D_A^{(\alpha)}$) (Chernoff, 1952; Amari, 2012))**.** The α-divergence is defined as

$$D_A^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(1-\alpha)} \left\{ 1 - \int q(\boldsymbol{\theta})^{\alpha} \pi(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta} \right\}, \tag{5.3}$$

where $\alpha \in \mathbb{R} \setminus \{0,1\}$.

The logic underlying the α-divergence is directly related to the replica trick:

[1] This way of measuring information stems from information theory, where information is measured in log-units because of their connection to storing information in systems based on binary logic: it takes $\mathcal{O}(\log(n))$ bits to store an integer $n \in \mathbb{Z}$.

rather than using the standard log function to measure average information, it uses a generalised log function of form $\log^{\alpha}(x) = \frac{x^{\alpha}-1}{\alpha}$; and so a trivial derivation shows that we recover the KLD as the limiting case for $\alpha \to 1$.

The next divergence we introduce is the β-divergence (also called *density power divergence*). Amongst the parameterized robust divergences discussed here, it is the only one that also belongs to the family of Bregman divergences (Cichocki and Amari, 2010).

Definition 5.3 (The β-divergence ($D_B^{(\beta)}$) (Basu et al., 1998; Mihoko and Eguchi, 2002))**.** The β-divergence is defined as

$$D_B^{(\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\beta(\beta-1)}\int q(\boldsymbol{\theta})^{\beta}d\boldsymbol{\theta} + \frac{1}{\beta}\int \pi(\boldsymbol{\theta})^{\beta}d\boldsymbol{\theta} - \frac{1}{\beta-1}\int q(\boldsymbol{\theta})\pi(\boldsymbol{\theta})^{\beta-1}d\boldsymbol{\theta}, \tag{5.4}$$

where $\beta \in \mathbb{R} \setminus \{0, 1\}$.

The last and least well-known divergence we introduce is the γ-divergence. We should note that this name is somewhat imprecise, since there have been multiple variants of strongly related discrepancy measures that have all been named γ-divergences. We will spare the reader a discussion of this literature and define γ-divergence very narrowly. The reason we choose our particular parameterization of the γ-divergence can be found in Cichocki and Amari (2010): In particular, one can show that it be generated from the β-divergence by applying the transformation

$$c_0 \int g(x)^{c_1} f(x)^{c_2} dx \to c_0 \log \int g(x)^{c_1} f(x)^{c_2} dx$$

to all three of the $D_B^{(\beta)}$ terms. In this sense, the γ-divergence is to the β-divergence what Rényi's α-divergence is to the original α-divergence—since Rényi's α-divergence can be generated from the α-divergence by applying this same transformation of its two terms.

Definition 5.4 (The γ-divergence ($D_G^{(\gamma)}$) (Fujisawa and Eguchi, 2008))**.** The γ-divergence is defined as

$$D_G^{(\gamma)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\gamma(\gamma-1)} \log \frac{\left(\int q(\boldsymbol{\theta})^{\gamma} d\boldsymbol{\theta}\right)\left(\int \pi(\boldsymbol{\theta})^{\gamma} d\boldsymbol{\theta}\right)^{\gamma-1}}{\left(\int q(\boldsymbol{\theta})\pi(\boldsymbol{\theta})^{\gamma} d\boldsymbol{\theta}\right)^{\gamma}},$$

where $\gamma \in \mathbb{R} \setminus \{0, 1\}$.

Remark 5.1. Before we move on, a word on terminology: For this chapter (and indeed this thesis), we call a divergence $D(q\|\pi)$ more robust than $\text{KLD}(q\|\pi)$ if D penalizes deviations of q from π *less* extremely than KLD whenever the data are at odds with π. In this sense, when we say that we want a RoT or GVI posterior to be more robust to prior misspecification, this implies that we want to choose a divergence D which is almost adaptive: the ideal robust D will behave similarly to KLD if the prior is well-specified, but will ignore the prior if it is contradicted too strongly by the data.

5.1.2 Closed forms of robust divergences

A big advantage of the $\alpha\beta\gamma$-divergences introduced in the last section is that we can obtain them in closed form for a wide range of variational families that are of practical interest. As it is cumbersome and not educational to state them here, we defer the precise results to Appendix B.5. In essence, the results say that for virtually all exponential families of practical interest in the context of variational methods, $D_{AR}^{(\alpha)}$, $D_{A}^{(\alpha)}$, $D_{B}^{(\beta)}$, and $D_{G}^{(\gamma)}$ have closed forms whenever $\alpha \in (0,1)$, $\beta \in (0,1)$, or $\gamma \in (0,1)$. Even when the parameters exceed 1, they are still typically available in closed forms in all but extreme cases. In practical terms, this means that for all numerical experiments presented in the remainder of this chapter, we have closed forms for the gradients $\nabla_{\boldsymbol{\kappa}} D(q\|\pi)$ discussed in Chapter 4.

5.1.3 Parameterized Divergences as Prior Regularizers D: Does GVI approximate a (generalized) posterior?

The parameterized divergences outlined in the last paragraphs all can be made arbitrarily close to the KLD in a point-wise sense. This motivates the question whether GVI posteriors $P(L, D^h, \mathcal{Q})$ with $\lim_{h\to 1} D^h(q\|\pi) = \text{KLD}(q\|\pi)$ can be thought of as approximations of $q^*_{n,\text{GB}}(\boldsymbol{\theta}) = P(L, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$ similarly as $q^*_{\text{VI}}(\boldsymbol{\theta}) = P(L, \text{KLD}, \mathcal{Q})$.

The following results study $P(L, D, \mathcal{Q})$ for $D \in \{D_{AR}^{(\alpha)}, D_{B}^{(\beta)}, D_{G}^{(\gamma)}\}$ and are geared towards answering this question. Since the derivations are not particularly educational, they are deferred to Appendix B.5.

Theorem 5.1 (GVI as approximate Evidence Lower bound with $D = D_{AR}^{(\alpha)}$)**.** The objective associated with $P(L, D_{AR}^{(\alpha)}, \mathcal{Q})$ is a lower bound on the $c(\alpha)$-scaled (generalized) evidence lower bound of $P(w(\alpha)\cdot L, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$:

$$\mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta}, x_{1:n})\right] + D_{AR}^{(\alpha)}(q||\pi) \geq -c(\alpha)\cdot \text{ELBO}^{w(\alpha)L}(q) + S_1(\alpha, q, \pi) \tag{5.5}$$

where $\text{ELBO}^{w(\alpha)L}$ is the Evidence Lower Bound associated with the generalized Bayesian posterior $q^*_{n,\text{GB}}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\left(-w(\alpha)L(\boldsymbol{\theta}, x_{1:n})\right)$ and given by

$$\text{ELBO}^{w(\alpha)L}(q) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[w(\alpha) \cdot L(\boldsymbol{\theta})\right] + \text{KLD}(q\|\pi),$$

$S_1(\alpha, q, \pi) = \mathbb{1}(0 < \alpha < 1) \cdot \left\{D^{(\alpha)}_{AR}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}) - \text{KLD}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))\right\}$ is a slack term, $c(\alpha) = \min\{1, \alpha^{-1}\}$ and $w(\alpha) = \max\{1, \alpha\}$.

Equation (5.5) shows that the slack term $S_1(\alpha, q, \pi)$ introduces the main difference between $P(L, D^{(\alpha)}_{AR}, \mathcal{Q})$ and $P(w(\alpha)\cdot L, \text{KLD}, \mathcal{Q})$. It is possible (but tedious) to make analytically more concise statements about $S_1(\alpha, q, \pi)$, and we will do so next. In short, this will reveal that the slack term makes $P(L, D^{(\alpha)}_{AR}, \mathcal{Q})$ more robust to misspecification of the prior than that of $P(w(\alpha) \cdot L, \text{KLD}, \mathcal{Q})$, and that this behaviour becomes more pronounced for smaller α. This phenomenon is summarized in Figure 5.2: since $w(\alpha) = 1$ for $\alpha \in (0, 1)$, if we ignore $S_1(\alpha, q, \pi)$ then the bound on the right of eq. (5.5) is just the ELBO of the Standard VI posterior $P(L, \alpha^{-1}\text{KLD}, \mathcal{Q}))$ for all $P(L, D^{(\alpha)}_{AR}, \mathcal{Q})$ with $\alpha \in (0, 1)$. As the Figure reveals, these two posteriors are quite different—making the slack term rather important in relating $P(L, D^{(\alpha)}_{AR}, \mathcal{Q})$ to $P(L, \text{KLD}, \mathcal{Q}))$. Since $P(L, D^{(\alpha)}_{AR}, \mathcal{Q})$ inflates variance relative to $P(L, \text{KLD}, \mathcal{Q}))$, one may expect that up-weighting the KLD term with $\frac{1}{\alpha}$ may produce similar posteriors. Thus, Figure 5.2 additionally compares $P(L, D^{(\alpha)}_{AR}, \mathcal{Q})$ with $P(L, \frac{1}{\alpha}\text{KLD}, \mathcal{Q}))$. Doing so reveals that while $P(L, D^{(\alpha)}_{AR}, \mathcal{Q}) \approx P(L, \frac{1}{\alpha}\text{KLD}, \mathcal{Q}))$ for reasonable prior specification, the distributions diverge substantially as the prior becomes more and more misspecified. This clarifies the role of the slack term $S_1(\alpha, q, \pi)$: while it ensures that $P(L, D^{(\alpha)}_{AR}, \mathcal{Q}) \approx P(L, \frac{1}{\alpha}\text{KLD}, \mathcal{Q}))$ whenever π is well-specified, it robustifies $P(L, D^{(\alpha)}_{AR}, \mathcal{Q})$ (relative to $P(L, \frac{1}{\alpha}\text{KLD}, \mathcal{Q}))$) for choices of π that are strongly at odds with the observed data.

Similar results can be derived both for the β- and γ-divergences.

Theorem 5.2 (GVI as approximate Evidence Lower bound with $D = D^{(\beta)}_B$)**.** The objective associated with $P(L, D^{(\beta)}_B, \mathcal{Q})$ is a lower bound on the $c(\beta)$-scaled (generalized) evidence lower bound of $P(w(\beta) \cdot \ell, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$:

$$\mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta}, x_{1:n})\right] + D^{(\beta)}_B(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) \geq -c(\beta)\text{ELBO}^{w(\beta)\ell}(q) + S_1(\beta, q, \pi) \quad (5.6)$$

where $\text{ELBO}^{w(\beta)L}$ is the Evidence Lower Bound associated with the generalized Bayesian posterior $q^*_{n,\text{GB}}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\left(-w(\beta)L(\boldsymbol{\theta}, x_{1:n})\right)$ and given by

$$\text{ELBO}^{w(\beta)L}(q) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[w(\beta) \cdot L(\boldsymbol{\theta}, x_{1:n})\right] + \text{KLD}(q\|\pi),$$

$c(\beta) = \min\{1, \beta^{-1}\}$, $w(\beta) = \max\{1, \beta\}$, and where $S_1(\beta, q, \pi)$ is a slack term with

$$S_1^{D_B^{(\beta)}}(\beta, q, \pi) = \begin{cases} \frac{1}{\beta(\beta-1)} \mathbb{E}_{q(\boldsymbol{\theta})} \left[q(\boldsymbol{\theta})^{\beta-1} \right] - \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log q(\boldsymbol{\theta}) \right] - \frac{1}{\beta-1} & \text{if } 0 < \beta < 1 \\ \frac{1}{\beta} \mathbb{E}_{q(\boldsymbol{\theta})} \left[\log \pi(\boldsymbol{\theta}) \right] - \frac{1}{\beta-1} \mathbb{E}_{q(\boldsymbol{\theta})} \left[\pi(\boldsymbol{\theta})^{\beta-1} \right] - \frac{1}{\beta(\beta-1)} & \text{if } \beta > 1. \end{cases} \tag{5.7}$$

Theorem 5.3 (GVI as approximate Evidence Lower bound with $D = D_G^{(\gamma)}$)**.** The objective associated with $P(L, D_G^{(\gamma)}, \mathcal{Q})$ is a lower bound on the $c(\gamma)$-scaled (generalized) evidence lower bound of $P(w(\gamma) \cdot L, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$:

$$\mathbb{E}_{q(\boldsymbol{\theta})} \left[L(\boldsymbol{\theta}, x_{1:n}) \right] + D_G^{(\gamma)}(q(\boldsymbol{\theta}) || \pi(\boldsymbol{\theta})) = -c(\gamma) \text{ELBO}^{w(\gamma)L}(q) + S(\gamma, q, \pi) \tag{5.8}$$

where $\text{ELBO}^{w(\gamma)L}$ is the Evidence Lower Bound associated with the generalized

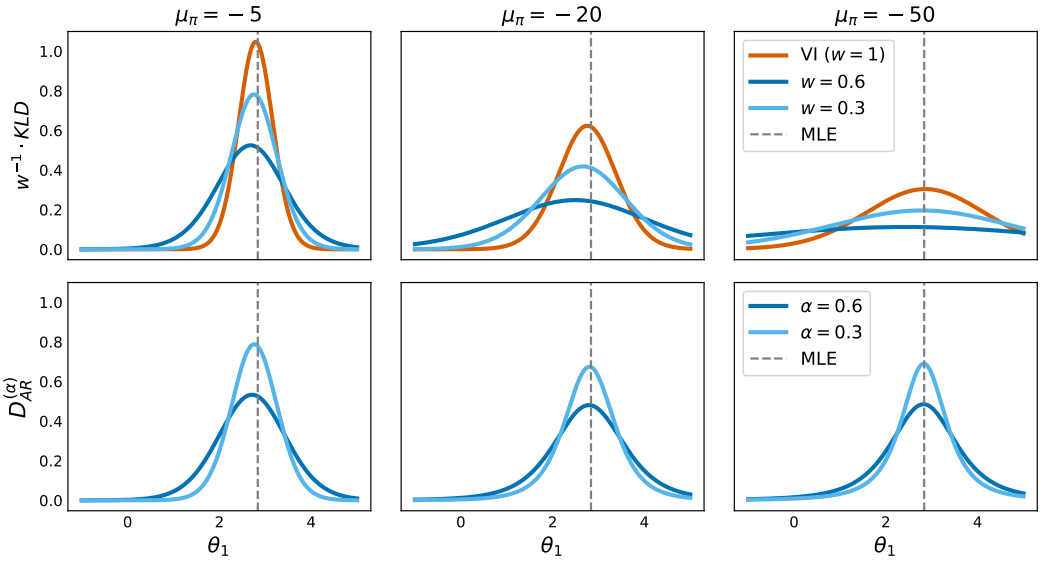


Figure 5.2: Best viewed in color. Marginal **VI** compared to different **GVI** posteriors for the coefficient θ_1 of data simulated from a d-dimensional Bayesian linear model with different priors (see Section 5.2.1). The prior for the coefficients is a Normal Inverse Gamma distribution given by $\boldsymbol{\mu} \sim \mathcal{NI}^{-1}(\mu_\pi \cdot 1_d, v_\pi \cdot I_d, a_\pi, b_\pi)$ with $v_\pi = 4 \cdot I_d$, $a_\pi = 3$, $b_\pi = 5$ and various values for μ_π. For all posteriors, the loss ℓ is the correctly specified negative log likelihood of the true data generating mechanism. Further, all variational posteriors are constrained to lie inside a mean field normal family $\mathcal{Q}$. Notice that the **standard VI** posterior corresponds to the ELBO component on the right hand side of the bound in eq. (5.5). In contrast, the **GVI** posteriors are obtained by maximizing the left hand side of the same bound.

Bayesian posterior $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\left(-w(\gamma)L(\boldsymbol{\theta}, x_{1:n})\right)$ and given by

$$\mathrm{ELBO}^{w(\gamma)L}(q) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[w(\gamma) \cdot L(\boldsymbol{\theta}, x_{1:n})\right] + \mathrm{KLD}(q\|\pi),$$

where $c(\gamma) = \min\{1, \gamma^{-1}\}$, $w(\gamma) = \max\{1, \gamma\}$ and where $S_1(\gamma, q, \pi)$ is a slack term with

$$S_\gamma(q, \pi) = \begin{cases} \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\boldsymbol{\theta})}\left[q(\boldsymbol{\theta})^{\gamma-1}\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log q(\boldsymbol{\theta})\right] & \text{if } 0 < \gamma < 1 \\ \frac{1}{\gamma}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log \pi(\boldsymbol{\theta})\right] - \frac{1}{\gamma-1} \log \mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right] & \text{if } \gamma > 1. \end{cases} \tag{5.9}$$

Theorems 5.2 and 5.3 provide a lower bound on an objective function associated with the relevant GVI posterior. Interpreting this lower bound provides some insight into the behaviour of the GVI posterior, and in particular whether it can be seen as an approximation to some other posterior $q^*_{n,\mathrm{GB}}(\boldsymbol{\theta})$. First, we investigate the case where the hyperparameters β and γ are in $(0, 1)$. In this parameter range, we find that the GVI objectives $P(L, D, \mathcal{Q})$ for $D \in \{D_B^{(\beta)}, D_G^{(\gamma)}\}$ produce posterior variances that are larger than those of $q^*_{\mathrm{VI}}(\boldsymbol{\theta}) = P(L, \mathrm{KLD}, \mathcal{Q})$. Secondly, we investigate the case where the hyperparameters β and γ are > 1, where we find that the opposite effect takes place (see Section 5.2).

Case 1: $0 < \beta = \gamma < 1$**.** For $0 < \beta = \gamma < 1$ the terms $c(\beta) = c(\gamma)$ and $w(\beta) = w(\gamma)$ ensure that $c(\beta)\mathrm{ELBO}^{w(\beta)L} = c(\gamma)\mathrm{ELBO}^{w(\gamma)L}$ are precisely the objective corresponding to $P(L, \beta^{-1}\mathrm{KLD}, \mathcal{Q}) = P(L, \gamma^{-1}\mathrm{KLD}, \mathcal{Q})$; so that the first of the two terms in (5.7) and (5.9) amounts to the exact objective $P(L, \gamma^{-1}\mathrm{KLD}, \mathcal{Q}) = P(\gamma L, \mathrm{KLD}, \mathcal{Q})$ of standard VI. This suggests that GVI continues to do something similar to minimizing the KLD between the variational and generalized Bayesian posterior based on the loss $\gamma L = \beta L$. Unlike for standard VI however, GVI with $D = D_B^{(\beta)}$ or $D = D_G^{(\gamma)}$ additionally minimises the slack terms $S_1^{D_B^{(\beta)}}(\beta, q, \pi)$ or $S_\gamma(q, \pi)$. These adjustment terms encourage the solution to $P(L, D_B^{(\beta)}, \mathcal{Q})$ with $0 < \beta < 1$ and $P(L, D_G^{(\gamma)}, \mathcal{Q})$ with $0 < \gamma < 1$ to have greater variance than the standard VI posterior given by $P(L, \mathrm{KLD}, \mathcal{Q})$. For the $D_B^{(\beta)}$, we can see this by rewriting

$$S_\beta(q, \pi) = -\frac{1}{\beta} h_T^{(\beta)}(q(\boldsymbol{\theta})) + h_{\mathrm{KLD}}(q(\boldsymbol{\theta})) + \frac{1-\beta}{\beta}.$$

Here, $h_{\mathrm{KLD}}(q(\boldsymbol{\theta}))$ is the Shannon entropy of $q(\boldsymbol{\theta})$ and $h_T^{(\beta)}(q(\boldsymbol{\theta}))$ is the Tsallis entropy of $q(\boldsymbol{\theta})$ with parameter β. Applying Lemma B.2 (see Appendix B.6), we find that for $0 < \beta < 1$, $h_T^{(\beta)}(q(\boldsymbol{\theta})) > h_{\mathrm{KLD}}(q(\boldsymbol{\theta}))$. It immediately follows that minimising

$-\frac{1}{\beta}h_T^{(\beta)}(q(\boldsymbol{\theta})) + h_{\text{KLD}}(q(\boldsymbol{\theta}))$ for $0 < \beta < 1$ will make $h_T^{(\beta)}(q(\boldsymbol{\theta}))$ large—an effect that is achieved by increasing the variance of $q(\boldsymbol{\theta})$.

Applying the same type of logic to the $D_G^{(\gamma)}$, one can rewrite

$$S_\gamma(q, \pi) = -\frac{1}{\gamma}h_R^{(\gamma)}(q(\boldsymbol{\theta})) + h_{\text{KLD}}(q(\boldsymbol{\theta})).$$

As before, $h_{\text{KLD}}(q(\boldsymbol{\theta}))$ is the Shannon entropy of $q(\boldsymbol{\theta})$, but unlike before, $h_R^{(\gamma)}(q(\boldsymbol{\theta}))$ now is the Rényi-entropy of $q(\boldsymbol{\theta})$ with parameter γ. Note that with this, one can also extend Theorem 3 in Van Erven and Harremos (2014) to show that $h_R^{(\gamma)}(q(\boldsymbol{\theta}))$ is decreasing in γ. Since it is also well-known that $\lim_{\gamma\to 1} h_R^{(\gamma)}(q(\boldsymbol{\theta})) = h_{\text{KLD}}(q(\boldsymbol{\theta}))$, it follows that minimising $-\frac{1}{\gamma}h_R^{(\gamma)}(q(\boldsymbol{\theta})) + h_{\text{KLD}}(q(\boldsymbol{\theta}))$ for $0 < \gamma < 1$ will make $h_R^{(\gamma)}(q(\boldsymbol{\theta}))$ large—an effect that is again achieved by increasing the variance of $q(\boldsymbol{\theta})$.

Case 2: $\beta = \gamma = k > 1$**.** For $k = \gamma = \beta > 1$, $c(k) = \frac{1}{k}$ and $w(k) = k$. Minimizing $\text{KLD}(q||q_k^*)$ for $k > 1$ will encourage $P(D_B^{(\beta)}, \ell, Q)$ or $P(D_G^{(\gamma)}, \ell, Q)$ to be more concentrated around the empirical risk minimizer $\hat{\boldsymbol{\theta}}_n$ of ℓ than the standard VI posterior given by $P(\text{KLD}, \ell, Q)$. Additionally, one can show that minimising the adjustment term also favours shrinking the variance of $q(\boldsymbol{\theta})$. To see this for the case of $D_B^{(\beta)}$, rewrite

$$S_\beta(q, \pi) = \frac{1}{\beta}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right] - \frac{1}{\beta - 1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1} - 1\right] - \frac{1}{\beta}. \tag{5.10}$$

Applying Lemma B.2 (Appendix B.6) then shows that for $\beta > 1$,

$$\frac{1}{\beta - 1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1} - 1\right] \geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right] \geq \frac{1}{\beta}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right].$$

From this, it follows that minimising Eq. (5.10) will make $\frac{1}{\beta-1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right]$ large. Fixing $\pi(\boldsymbol{\theta})$, maximising $\frac{1}{\beta-1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right]$ plus $\frac{1}{\beta}\times$ the Tsallis entropy of $q(\boldsymbol{\theta})$ is equivalent to minimising $D_B^{(\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))$. Because $D_B^{(\beta)}$ is a divergence, this maximization would naturally seek to choose $q(\boldsymbol{\theta})$ close to $\pi(\boldsymbol{\theta})$. The Tsallis entropy term in this formulation would have acted to increase the variance of $q(\boldsymbol{\theta})$. But since we maximize only $\frac{1}{\beta-1}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\beta-1}\right]$—i.e. without adding the Tsallis entropy of $q(\boldsymbol{\theta})$—choices of $\beta > 1$ will lead to shrinking the variance of $q(\boldsymbol{\theta})$ relative to standard VI.

For the $D_G^{(\gamma)}$, Jensen's inequality shows that for $\gamma > 1$,

$$\frac{1}{\gamma - 1}\log\mathbb{E}_{q(\boldsymbol{\theta})}\left[\pi(\boldsymbol{\theta})^{\gamma-1}\right] \geq \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right] \geq \frac{1}{\gamma}\mathbb{E}_{q(\boldsymbol{\theta})}\left[\log(\pi(\boldsymbol{\theta}))\right].$$

As a result, minimising $S_\gamma(q, \pi)$ will seek to make $\frac{1}{\gamma-1} \log \mathbb{E}_{q(\boldsymbol{\theta})} \left[\pi(\boldsymbol{\theta})^{\gamma-1}\right]$ large. Fixing again $\pi(\boldsymbol{\theta})$, maximising $\frac{1}{\gamma-1} \log \mathbb{E}_{q(\boldsymbol{\theta})} \left[\pi(\boldsymbol{\theta})^{\beta-1}\right]$ plus $\frac{1}{\gamma}\times$ the Rényi entropy of $q(\boldsymbol{\theta})$ is equivalent to minimising $D_G^{(\gamma)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))$, and thus seeks $q(\boldsymbol{\theta})$ close to $\pi(\boldsymbol{\theta})$. The Rényi entropy term would have acted to increase the variance of $q(\boldsymbol{\theta})$. Therefore and similarly to the case of $D_B^{(\beta)}$, maximising $\frac{1}{\gamma-1} \log \mathbb{E}_{q(\boldsymbol{\theta})} \left[\pi(\boldsymbol{\theta})^{\gamma-1}\right]$ without adding the Rényi entropy will lead to shrinkage of the variance of $q(\boldsymbol{\theta})$.

In conclusion, while GVI posteriors with parameterized divergences D^h so that $\lim_{h\to 1} D^h(q\|\pi) = \mathrm{KLD}(q\|\pi)$ can indeed be interpreted as approximately targeting a generalized Bayes posterior, the devil is in the detail: the validity of thinking of GVI posteriors as approximations depends strongly on the form of the slack terms S_h introduced in Theorems 5.1, 5.2, and 5.3. To emphasize that this slack term is indeed of crucial importance, we provide a short empirical comparison between different robust regularizers and weighted versions of the KLD in the next section.

5.2 An Empirical Comparison

While Theorems 5.1, 5.2, and 5.3 and their slack terms give us some theoretical guidance as to the behaviour we expect from posteriors of the form $P(L, D, \mathcal{Q})$ for $D \in \{D_{AR}^{(\alpha)}, D_B^{(\beta)}, D_G^{(\gamma)}\}$, we can at best form vague interpretations and base our expectations on these. Moreover, these interpretations are relatively uniform for all parameterized robust divergences, which does not answer the question which of them one should prefer in practice. To bridge this gap in what our theory can achieve, we now present a small empirical comparison.

5.2.1 Experimental setup

Throughout, we use the log likelihood loss of a correctly specified model. For this, we use a simple Bayesian linear regression (BLR) with two highly correlated predictors. Formally, we study the Bayesian model given by

$$\begin{aligned} \sigma^2 &\sim \mathcal{IG}(a_0, b_0) \\ \boldsymbol{\theta}|\sigma^2 &\sim \mathcal{N}\left(\boldsymbol{\mu}_0, \sigma^2 V_0\right) && (5.11) \\ y_i|\boldsymbol{\theta}, \sigma^2 &\sim \mathcal{N}\left(X_i\boldsymbol{\theta}, \sigma^2\right). && (5.12) \end{aligned}$$

We choose this example because it provides a closed form exact Bayesian posteriors and closed form objectives for the variational objectives of both standard **VI** and **GVI**. Consequently, no stochastic optimization or sampling is required—neither for calculating the exact posterior nor for the optimization of the **GVI** and

VI posteriors—so that numerical errors and uncertainties are kept to a minimum.

Studying the exact closed form Bayesian (normal) posterior for $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$, one observes that if the two predictors are correlated, then the posterior covariance of $\boldsymbol{\theta}$ will inherit this correlation. This is convenient: it is well-known that for posteriors with highly correlated dimensions, standard VI will strongly underestimate the marginal variances (Turner and Sahani, 2011); so that we will also be able to study if GVI can address this shortcoming of standard VI. In particular, we simulate the highly correlated predictors

$$(x_1, x_2)^T \sim \mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}\right)$$

and compare the performance of the different GVI and VI posteriors on the resulting BLR. All posteriors are based on the the mean field normal variational family given by

$$\begin{aligned}
\mathcal{Q} &= \{q(\theta_1|\sigma^2, \boldsymbol{\kappa}_n)q(\theta_2|\sigma^2, \boldsymbol{\kappa}_n)q(\sigma^2|\boldsymbol{\kappa}_n)\} \text{ so that} \\
\boldsymbol{\kappa}_n &= (a_n, b_n, \mu_{1,n}, \mu_{2,n}, v_{1,n}, v_{2,n})^T \\
&\text{with } a_n, b_n, v_{1,n}, v_{2,n} > 0 \text{ and } \mu_{1,n}, \mu_{2,n} \in \mathbb{R} \\
q(\sigma^2|\boldsymbol{\kappa}_n) &= \mathcal{IG}(\sigma^2|a_n, b_n) \\
q(\theta_1|\sigma^2, \boldsymbol{\kappa}_n) &= \mathcal{N}\left(\theta_1|\mu_{1,n}, \sigma^2 v_{1,n}\right) \\
q(\theta_2|\sigma^2, \boldsymbol{\kappa}_n) &= \mathcal{N}\left(\theta_2|\mu_{2,n}, \sigma^2 v_{2,n}\right).
\end{aligned}$$

For all experiments, $n = 25$ observations are simulated from eq. (5.12) with $\boldsymbol{\theta} = (2, 3)$ and $\sigma^2 = 4$. We use the negative log-likelihood ℓ of the correctly specified model as given in eq. (5.12) as loss function. The results are depicted in Figs. 5.3 and 5.5-5.8. We summarize the most interesting results in the following subsections.

5.2.2 A cautionary tale about boundedness

The attentive reader will note that while we have introduced the α-divergence in Definition 5.2, we did not derive a corresponding robustness result for it in the previous section. The reason for this is a practical one: As our results in Figure 5.3 show, the $D_A^{(\alpha)}$ is not a reliable prior regularizer within the GVI framework—at least not for $\alpha \in (0, 1)$. In particular, the plot shows that the solutions to $P(\ell, D_A^{(\alpha)}, \mathcal{Q})$ can produce essentially degenerate posteriors if $\alpha \in (0, 1)$. Note also that this happens in spite of the relatively small sample size of $n = 25$. For example, when $\alpha = 0.5$, $P(\ell, D_A^{(\alpha)}, \mathcal{Q})$ is visually indistinguishable from a point mass at the

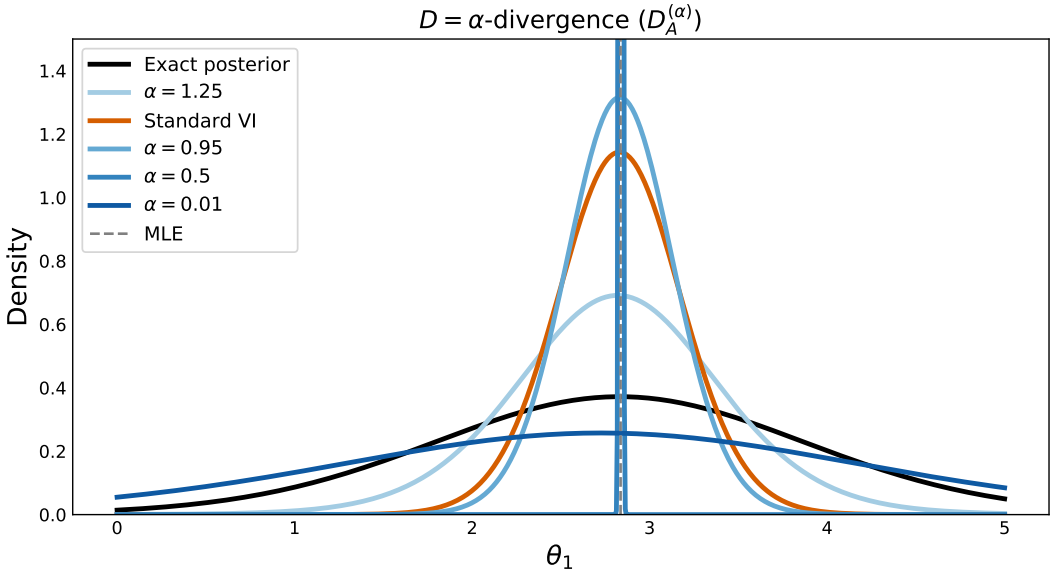


Figure 5.3: Best viewed in color. Marginal **VI** and **GVI** posterior for the $\boldsymbol{\theta}_1$ coefficient of a Bayesian linear model under the $D_A^{(\alpha)}$ prior regularizer for different values of α. The boundedness of the $D_A^{(\alpha)}$ causes **GVI** posteriors to severely over-concentrate if α is not carefully specified. Prior Specification: $\sigma^2 \sim \mathcal{IG}(20, 50)$, $\theta_1|\sigma^2 \sim \mathcal{N}(0, 25\sigma^2)$ and $\theta_2|\sigma^2 \sim \mathcal{N}(0, 25\sigma^2)$.

maximum likelihood estimate. This is a simple consequence of the boundedness of $D_A^{(\alpha)}$ for $\alpha \in (0, 1)$: Specifically, it holds that $D_A^{(\alpha)} \leq (\alpha(1-\alpha))^{-1}$ for $\alpha \in (0, 1)$. As α decreases from 1, this upper-bound initially also decreases until reaching its minimum for $\alpha = 0.5$. As a result, decreasing α from unity to 0.5 significantly decreases the maximal penalty for posterior beliefs far from the prior. In turn, this forces the posterior to focus mostly on minimising the in-sample loss.

This phenomenon is illustrated in Figure 5.4, which also shows that the divergence magnitude increases again as α approaches zero or if $\alpha > 1$. Comparing the plot with that in Figure 5.1, it is clear why hyperparameter selection for the other members of the $D_G^{(\alpha,\beta,r)}$ family of divergences is a less complicated endeavour than for the α-divergence. This does not mean that the $D_A^{(\alpha)}$ cannot be used for producing **GVI** posteriors: For example, some GVI posteriors in Figure 5.3 based on the $D_A^{(\alpha)}$ are able to achieve marginal variances that more closely correspond to the exact posterior than **VI**—notably for $\alpha = 1.25$ and $\alpha = 0.01$. Generally speaking, for values of α close to zero or above unity, it is possible to achieve more conservative uncertainty quantification. Yet, the $D_A^{(\alpha)}$ also functions primarily as a cautionary tale: Without understanding the properties of the prior regularizer D sufficiently well, **GVI** may well yield unsatisfactory posteriors.

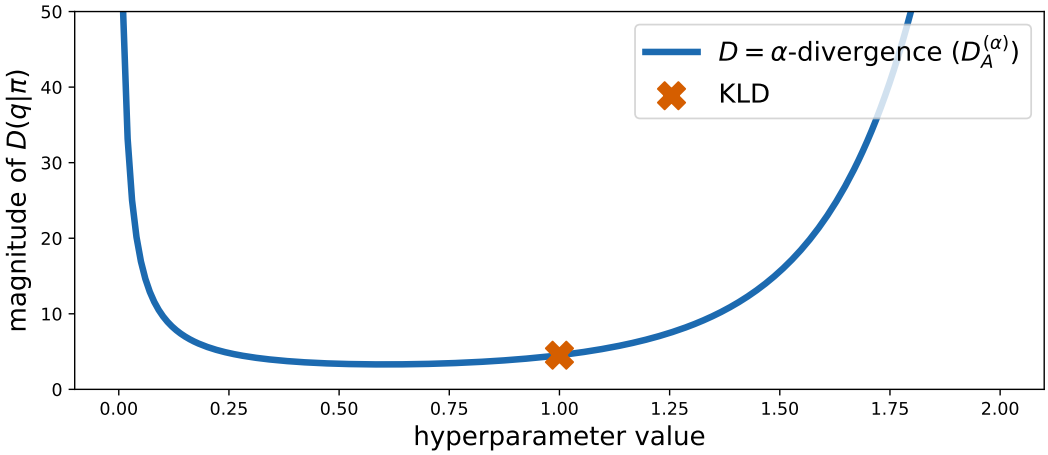


Figure 5.4: A comparison of the size of $D_A^{(\alpha)}$ for various values of α between two bivariate Normal Inverse Gamma distributions with $a_n = 512$, $b_n = 543$, $\boldsymbol{\mu}_n = (2.5, 2.5)$, $\mathbf{V}_n = \text{diag}(0.3, 2)$ and $a_0 = 500$, $b_0 = 500$, $\mu_0 = (0, 0)$, $V_0 = \text{diag}(25, 2)$.

5.2.3 Robustness to the prior

Next, we compare the impact of changing the prior regularizer on the posterior's sensitivity to appropriate specification of the prior for divergences that are of more practical interest than the $D_A^{(\alpha)}$—in particular, $D_{AR}^{(\alpha)}$, $D_B^{(\beta)}$, $D_G^{(\gamma)}$ and a weighted version of the KLD, $\frac{1}{w}$KLD. The aim of this comparison is to identify the differences in inference outcomes when compared to standard variational approaches: essentially, we wish to find out the practical differences implied by the different slack terms in Theorems 5.1, 5.2, and 5.3.

To this end, when we compare $\frac{1}{w}$KLD with $D_{AR}^{(\alpha)}$ and $D_G^{(\gamma)}$, we fixed $\alpha = \gamma = w$. Setting the values of these various hyperparameters to be the same makes sense: the slack terms of **GVI** in Theorems 5.1, 5.2, and 5.3 are such that the first term always corresponds to an ELBO whose KLD-regularizer is weighted by $\frac{1}{\alpha}$, $\frac{1}{\beta}$, and $\frac{1}{\gamma}$. As a consequence, setting $\alpha = \gamma = w$ allows us to study the effect of the slack terms in isolation. For the $D_B^{(\beta)}$, different values of β had to be selected to ensure that the objective to be optimized is available in closed form.

Weighted KLD ($\frac{1}{w}$KLD)

To set a baseline, Figure 5.5 examines how changing the weight w affects the posteriors $P(\ell, \frac{1}{w}\text{KLD}, \mathcal{Q}) = P(w\ell, \text{KLD}, \mathcal{Q})$. It should be clear that choosing $w < 1$ leads to posteriors that encourage larger variances, leading to more conservative uncertainty

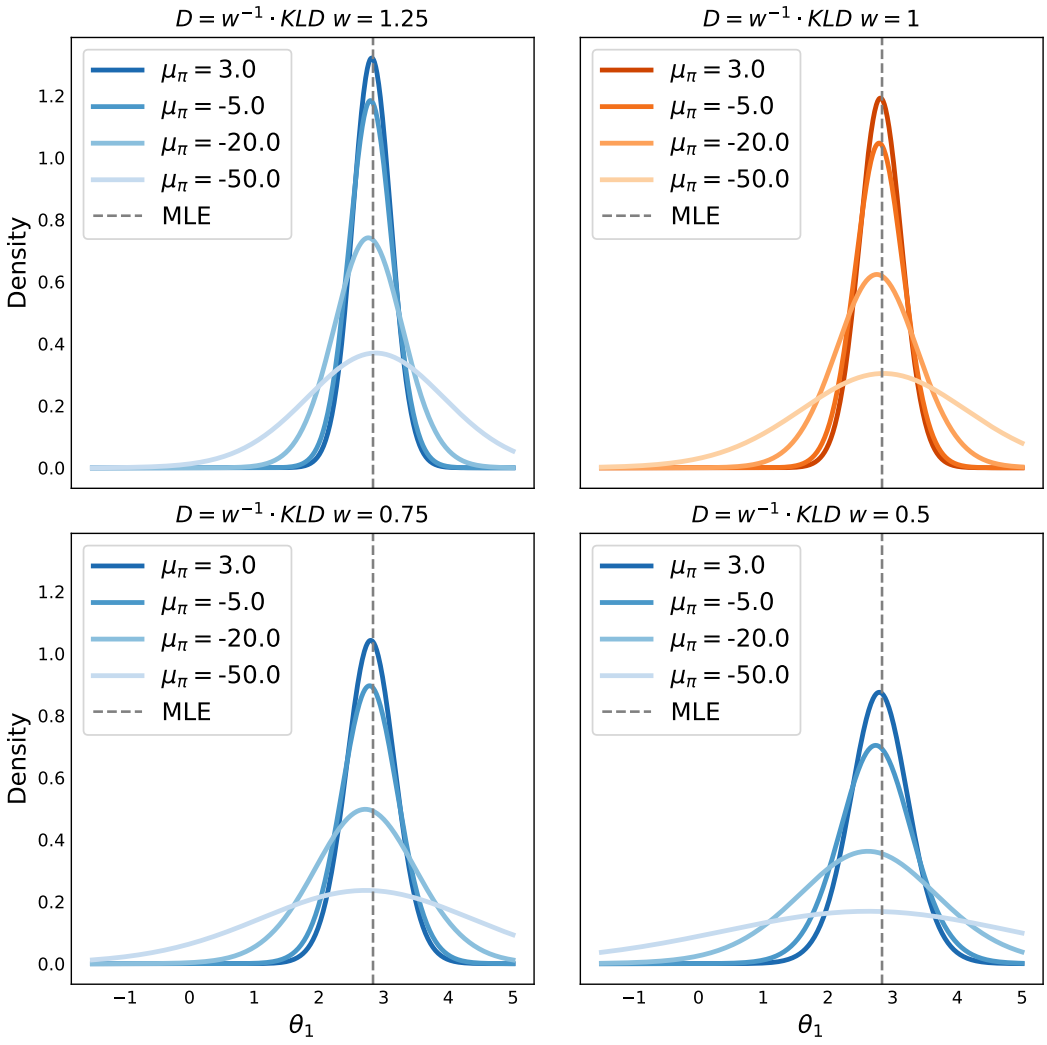


Figure 5.5: Best viewed in color. Marginal **VI** and **GVI** posterior for the coefficient of a Bayesian linear model under different priors using $D = \frac{1}{w}\text{KLD}$ as prior regularizer ($\frac{1}{w}\text{KLD}$ recovers KLD for $w = 1$). The prior specification is given by $\theta_1|\sigma^2 \sim \mathcal{N}(\mu_\pi, \sigma^2)$ with $\sigma^2 \sim \mathcal{IG}(3, 5)$.

quantification. Unfortunately and unsurprisingly, this comes at the cost of making posteriors *more* sensitive to the prior: After all, one up-weights the term penalizing deviations from the prior. Conversely, $w > 1$ will result in posteriors that are less sensitive to the prior than standard VI. At the same time, they will also be more concentrated around the Maximum Likelihood Estimator. This leads to a serious problem when using $\frac{1}{w}$KLD to achieve robustness to poorly specified priors: If we want to be robust to poorly specified priors, we should set $w > 1$. But this implies that we are obtaining a more concentrated posterior belief! in other words, even though we are *less certain* about one of our key ingredients to obtaining a belief distribution, the posterior belief ends up being *more certain* about which regions of the parameter space represent 'good' values. As we shall see, this undesirable trade-off is *not* shared by the other (robust) divergences considered in this section. Unlike the $\frac{1}{w}$KLD or Brexit, they often provide a way to have your cake and eat it, too.

Rényi's α-divergence ($D_{AR}^{(\alpha)}$)

Figure 5.6 demonstrates the sensitivity of $P(\ell, D_{AR}^{(\alpha)}, \mathcal{Q})$ to prior specification and also shows $P(\ell, \text{KLD}, \mathcal{Q}) = P(\ell, D_{AR}^{(1)}, \mathcal{Q})$ in the top right panel for comparison. For $0 < \alpha < 1$, the posterior exhibits the kind of behaviour that is desirable in the presence of prior misspecification, but which is difficult to attain with **VI**: It both produces larger marginal variances *and* is robust to badly specified priors. This is no longer true if $\alpha > 1$, since it holds that $D_{AR}^{(\alpha)} \leq \text{KLD}$ for this parameter range—which is why **GVI** no longer produces larger marginal variances than **VI** based on the KLD for $\alpha > 1$. This flip in robustness as α crosses from $(0, 1)$ into values larger than unity may seem strange, but can be understood by investigating the form of the $D_{AR}^{(\alpha)}$:

$$D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(\alpha-1)} \log \int q(\boldsymbol{\theta})^{\alpha} \pi(\boldsymbol{\theta})^{1-\alpha} d\boldsymbol{\theta} = \frac{1}{\alpha(\alpha-1)} \log \int \frac{q(\boldsymbol{\theta})^{\alpha}}{\pi(\boldsymbol{\theta})^{\alpha-1}} d\boldsymbol{\theta}.$$

It is clear that the magnitude of the divergence is determined by a ratio of two densities. Glancing closer, for $\alpha > 1$ this means that if $q(\boldsymbol{\theta})$ is large in an area where $\pi(\boldsymbol{\theta})$ is not, then a severe penalty is incurred. This limits how far $q(\boldsymbol{\theta})$ can move from the prior and thus results in lack of prior robustness. Conversely, if $\alpha \in (0, 1)$, then $\pi(\boldsymbol{\theta})^{\alpha-1} > \pi(\boldsymbol{\theta})$ for regions where $\pi(\boldsymbol{\theta}) < 1$, which allows the posterior to spread its mass in a less concentrated way than for $\alpha > 1$. In fact, this very finding is also implicitly stated in Theorem 5.1.

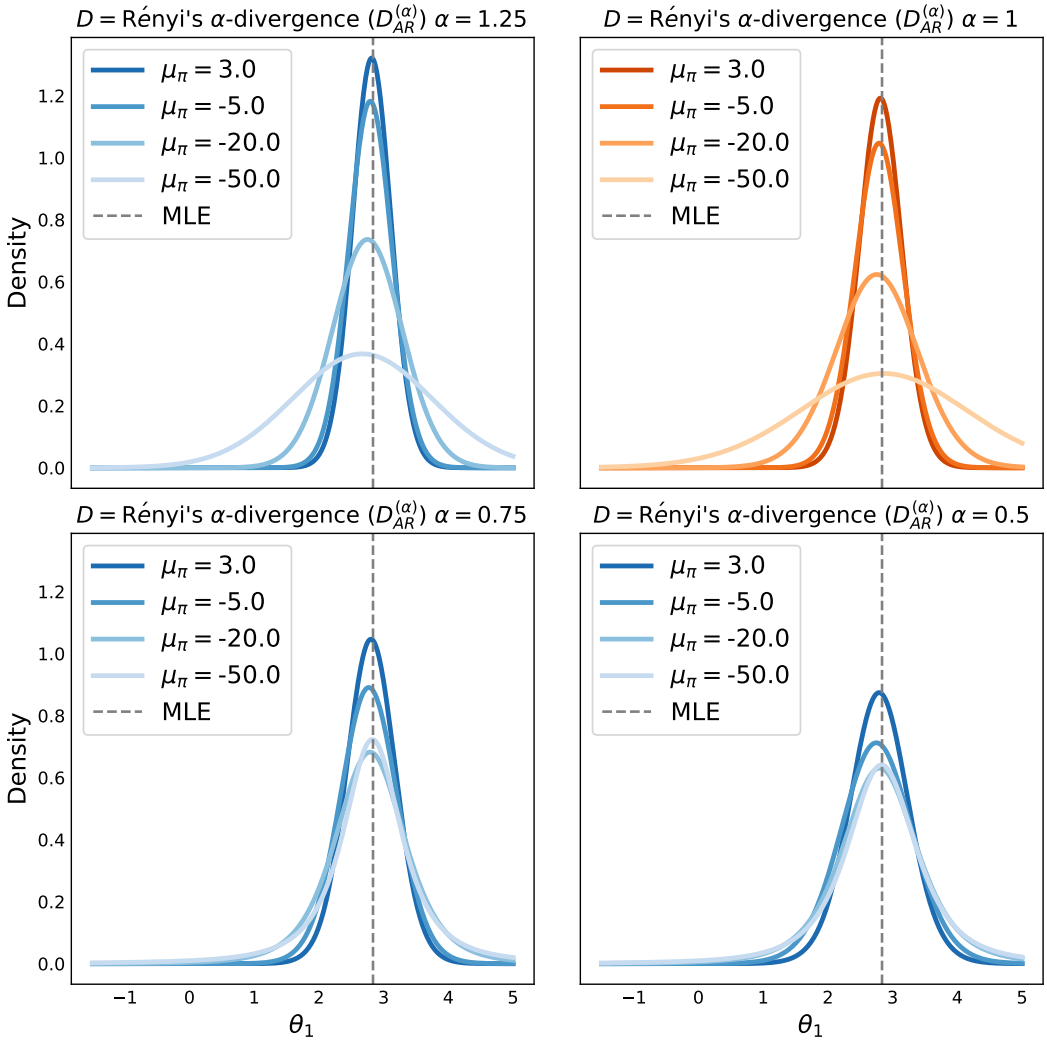


Figure 5.6: Best viewed in color. Marginal **VI** and **GVI** posterior for the coefficient of a Bayesian linear model under different priors using $D = D_{AR}^{(\alpha)}$ as prior regularizer ($D_{AR}^{(\alpha)}$ recovers KLD as $\alpha \to 1$). The prior specification is given by $\theta_1|\sigma^2 \sim \mathcal{N}(\mu_\pi, \sigma^2)$ with $\sigma^2 \sim \mathcal{IG}(3, 5)$.

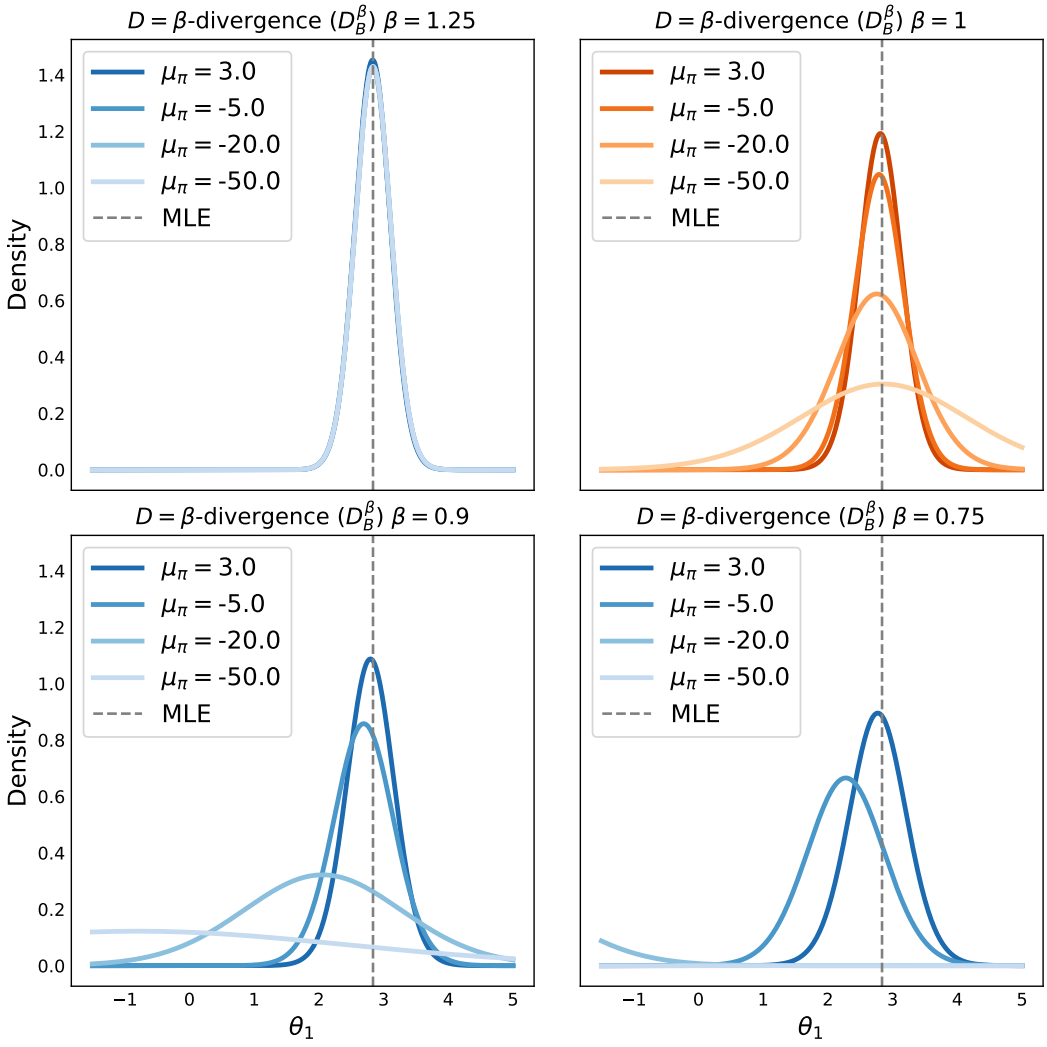


Figure 5.7: Best viewed in color. Marginal **VI** and **GVI** posterior for the coefficient of a Bayesian linear model under different priors using $D = D_B^{(\beta)}$ as prior regularizer ($D_B^{(\beta)}$ recovers KLD as $\beta \to 1$). The prior specification is given by $\theta_1|\sigma^2 \sim \mathcal{N}(\mu_\pi, \sigma^2)$ with $\sigma^2 \sim \mathcal{IG}(3, 5)$.

β-divergence ($D_B^{(\beta)}$)

Figure 5.7 demonstrates the sensitivity of $P(\ell, D_B^{(\beta)}, \mathcal{Q})$ to prior specification. The plot shows that $\beta > 1$ is able to achieve extreme robustness to the prior, while $\beta < 1$ causes extreme sensitivity to the prior. This phenomenon is a result of the fact that the $D_B^{(\beta)}$ decomposes into three integrals, one containing just the prior, one containing just $q(\boldsymbol{\theta})$ and one containing an interaction between them:

$$D_B^{(\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\beta}\int \pi(\boldsymbol{\theta})^{\beta} d\boldsymbol{\theta} - \frac{1}{\beta-1}\int \pi(\boldsymbol{\theta})^{\beta-1} q(\boldsymbol{\theta}) d\boldsymbol{\theta} + \frac{1}{\beta(\beta-1)}\int q(\boldsymbol{\theta})^{\beta} d\boldsymbol{\theta}.$$

The integral depending only on the prior does not on depend $q(\boldsymbol{\theta})$, so we can ignore it (since the prior is fixed across the different values of β). For $\beta \in (0, 1)$, the signs of both of the remaining terms flip and it is instructive to rewrite the middle term as $\frac{1}{1-\beta}\int \frac{q(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})^{1-\beta}} d\boldsymbol{\theta}$ with $1 - \beta > 0$. This shows that the prior appears as a denominator. The consequences of this are similar to the behaviour of the $D_{AR}^{(\alpha)}$ for $\alpha > 1$: if $q(\boldsymbol{\theta})$ has large density in regions where $\pi(\boldsymbol{\theta})$ has small density, then we divide a not-so-small number by a very small number and a huge penalty is incurred for this. As a result, the corresponding posterior will be very close to the prior. (In fact, notice that two of the four posteriors for $\beta = 0.75$ in Figure 5.7 favour the prior so much that the density around the maximum likelihood estimate is virtually zero.) For $\beta > 1$ the opposite effect is observed. The prior no longer appears as a denominator and therefore deviations from the prior are punished in a milder manner by the middle term. This allows the third term, which depends on $q(\boldsymbol{\theta})$ independently of the prior, to have greater influence on how uncertainty is quantified. This third integral will become very large if the variance of $q(\boldsymbol{\theta})$ gets very small, which prevents it from quickly converging to a point mass at the maximum likelihood estimate. As a consequence, $D_B^{(\beta)}$ is able to provide virtually prior-invariant uncertainty quantification for $\beta > 1$.

γ-divergence ($D_G^{(\gamma)}$)

Lastly, Figure 5.8 demonstrates the sensitivity of $P(\ell, D_G^{(\gamma)}, \mathcal{Q})$ to prior specification. For $\gamma < 1$ it appears as though the $D_G^{(\gamma)}$ reacts similarly to the $\frac{1}{w}$KLD for $w < 1$. The $D_G^{(\gamma)}$ with $\gamma > 1$ produces greater robustness to the prior than the $\frac{1}{w}$KLD prior regularizer with $w > 1$, but this robustness is less pronounced than that achieved with $D = D_B^{(\beta)}$. The reason for this is that although the $D_G^{(\gamma)}$ consists of the same three integral terms as the $D_B^{(\beta)}$, these terms are now transformed into the logarithmic scale. This means that the three integrals are combined multiplicatively (for $D_G^{(\gamma)}$)

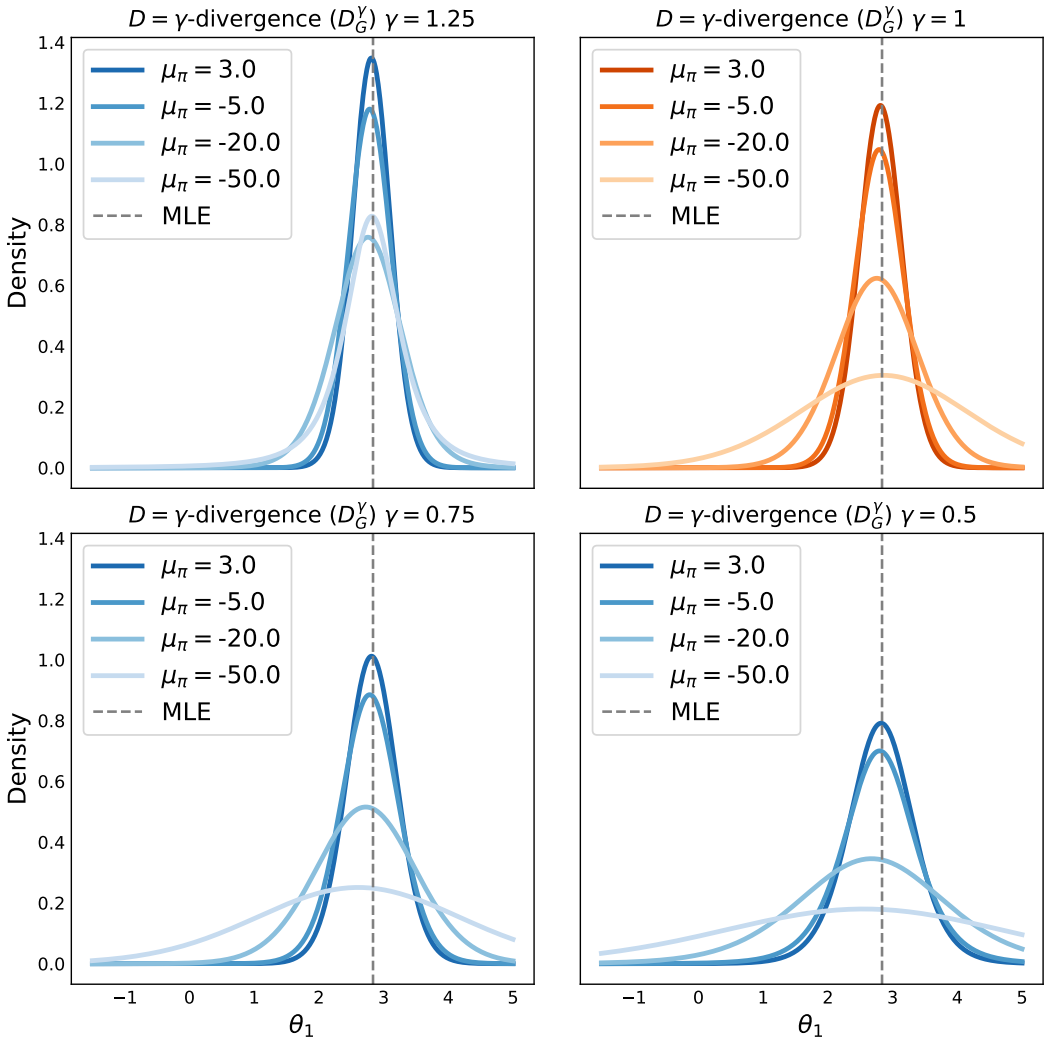


Figure 5.8: Best viewed in color. Marginal **VI** and **GVI** posterior for the coefficient of a Bayesian linear model under different priors using $D = D_G^{(\gamma)}$ as prior regularizer ($D_G^{(\gamma)}$ recovers KLD as $\gamma \to 1$). The prior specification is given by $\theta_1|\sigma^2 \sim \mathcal{N}(\mu_\pi, \sigma^2)$ with $\sigma^2 \sim \mathcal{IG}(3, 5)$.

rather than additively (for $D_B^{(\beta)}$), which makes the variation of $D_G^{(\gamma)}$ as a function of γ much smoother than that of $D_B^{(\beta)}$ as a function of β. Roughly speaking, this smoothness implies that unlike for the $D_B^{(\beta)}$, minimising the $D_G^{(\gamma)}$ no longer disregards any one term in order to minimise the others.

5.3 Applications: Bayesian Mixture Models (BMMs) & Bayesian Neural Networks (BNNs)

The final part of this chapter explores the use cases of GVI posteriors in the setting of misspecified priors on two examples: Bayesian Mixture Models (BMMs), and Bayesian Neural Networks (BNNs). Since our initial experimental evaluation suggested that Reńyi's α-divergence ($D_{AR}^{(\alpha)}$) is the most promising candidate for this kind of treatment, we focus our evaluation on GVI posteriors with $D = D_{AR}^{(\alpha)}$.

Conceptually, we also study two different forms of desiderata for robustness: Since the BMM is a classical statistical model whose number of parameters is small relative to the number of data points, our interpretation of robustness is also more classical. In particular, a procedure robust to ill-specified priors should produce posteriors that incorporate *more* parameter uncertainty about the (interpretable) parameters of our model. For the BNN, our notion of robustness is necessarily different: since this is an overparameterized black box model, individual parameters have no clear interpretable meaning. Therefore, robustness needs to be evaluated on the predictive precision of the model. In practice, this means that robustness to a misspecified prior entails that we should pay less attention to the priors, and more to the data. In other words, robustness in a BNN implies that we wish for *less* parameter uncertainty (not more!). For the $D_{AR}^{(\alpha)}$, this means that we will be expecting parameterizations in the range $\alpha \in (0, 1)$ to work best for the BMM, while parameterizations $\alpha > 1$ should be expected to be more reliable for the black box BNN model.

5.3.1 Bayesian Mixture Model (BMM)

Throughout, n observations are generated from the d-dimensional BMM with two equally likely normal mixture components $z = 0, 1$ with dimension-wise unit variance and mean given by

$$\boldsymbol{\mu}^z = (\mu_1^z, \mu_2^z, \dots \mu_d^z)^T = \begin{cases} 2 \cdot e_d & \text{if } z = 0 \\ -2 \cdot e_d & \text{if } z = 1 \end{cases},$$

where $e_d = (1, 1, \dots 1)^T$ is the d-dimensional column vector of ones. The n observations x_i are drawn with equal probability from the two mixture components, meaning that

$$\begin{aligned} \boldsymbol{z}_i &\overset{i.i.d.}{\sim} \text{Bernoulli}(0.5) \\ x_i|\{\boldsymbol{z}_i = z_i\} &\overset{i.i.d.}{\sim} \mathcal{N}(x_i|\boldsymbol{\mu}^{z_i}, \boldsymbol{I}_d). \end{aligned} \tag{5.13}$$

Notice in particular that this generates n latent variables $\boldsymbol{z}_{1:n}$ that indicate mixture memberships for $x_{1:n}$, but are unobserved. With this, inference is conducted on $\boldsymbol{\mu}^z$ for $z = 1, 2$ via the negative log likelihood loss of the correct model. For $\boldsymbol{\theta} = (\boldsymbol{\mu}^1, \boldsymbol{\mu}^2)$, this loss is given by

$$\ell(\boldsymbol{\theta}, x_i, \boldsymbol{z}_i) = -\log \mathcal{N}(x_i|\boldsymbol{\mu}^{\boldsymbol{z}_i}, \boldsymbol{I}_d).$$

The variational family $\mathcal{Q}$ used for all experiments is the collection of mean-field normal distributions given as $\mathcal{Q} = \mathcal{Q}_{\text{MFN}}$ in (1.3). In our experiments, we consider the benefits of alternative choices of D for the fixed number of observations $n = 50$. To this end, $B = 100$ artificial data sets are generated according to the above description.

If the prior is poorly specified, $D = \text{KLD}$ will produce posterior beliefs that place the same weight on the prior as they do on the data. In contrast, robust alternatives to the KLD do not suffer from this problem: They can produce posterior beliefs that take the prior into account, but are robust to prior misspecification. To illustrate the phenomenon empirically, we compare the KLD with Rényi's α-divergence ($D_{AR}^{(\alpha)}$) for $\alpha = 0.5$ under two settings: A well-specified prior $\pi_1(\boldsymbol{\theta})$ and a misspecified prior $\pi_2(\boldsymbol{\theta})$, which are given by

$$\begin{aligned} \pi_1(\boldsymbol{\theta}) &= \mathcal{N}\left(\boldsymbol{\theta}|0_d, \sqrt{10}\boldsymbol{I}_d\right) \\ \pi_2(\boldsymbol{\theta}) &= \mathcal{N}\left(\boldsymbol{\theta}| - 10 \cdot e_d, \sqrt{0.1}\boldsymbol{I}_d\right) \end{aligned}$$

Across the $B = 100$ data sets generated, Figure 5.9 reports the average posterior computed as

$$\mathcal{N}\left(\bar{m}, \bar{s}\right), \quad \text{where} \quad \bar{m} = \frac{1}{100}\sum_{j=1}^{2d}\sum_{b=1}^{B} m_{b,j}, \quad \bar{s} = \frac{1}{100}\sum_{j=1}^{2d}\sum_{b=1}^{B} s_{b,j}.$$

Here, $s_{b,j}$ corresponds to the standard deviation computed for the j-th dimension of the mean field normal posterior on the b-th artificial data sets. Similarly, $m_{b,j}$

corresponds to the mean of the same parameter posterior, albeit re-centered around the true value of the inferred parameter.

As Figure 5.9 shows, $D_{AR}^{(\alpha)}$ is an interesting alternative to the KLD in finite samples: If the prior is misspecified (top row), the KLD produces belief distributions that take the prior too strongly into account and are far from the truth. In contrast, the $D_{AR}^{(\alpha)}$ provides both prior robustness as well as better uncertainty quantification under misspecification. At the same time, $D_{AR}^{(\alpha)}$ has no tangible disadvantage relative to the KLD if the prior is well-specified (bottom row). Note that the posteriors in the bottom row plot are very concentrated because the prior is already quite informative.

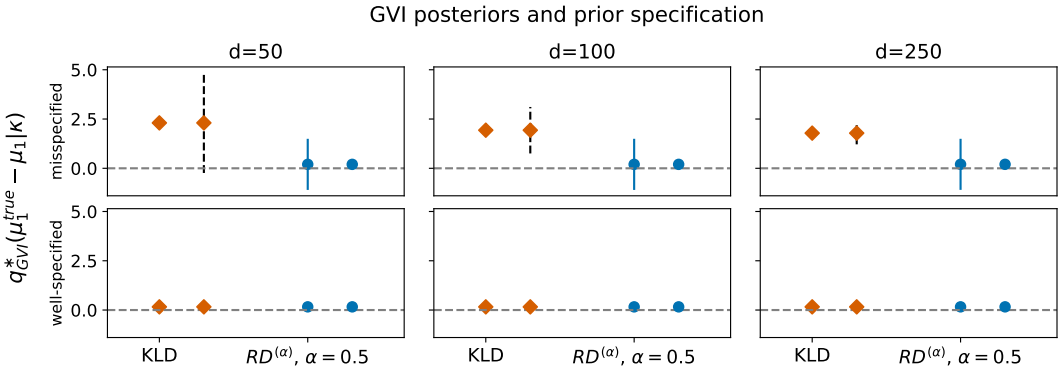


Figure 5.9: The **first column** of each setting depicts the inferred **VI** and **GVI** posteriors for $\boldsymbol{\theta}$ in the BMM of eq. (5.13). Here, the **GVI** posteriors use $D = D_{AR}^{(\alpha)}$ for $\alpha = 0.5$. All inferred posterior beliefs are normals, so dots and whiskers mark posterior means and standard deviations. The posteriors are re-centered so that the y-axis measures the magnitude by which the posterior belief deviates from the truth. The **second column** of each setting shows the inferred posterior mean and its standard error across the 100 data sets on which the experiment was run. The plots clearly show that the adverse effect of the prior stabilizes as the number d of affected parameters increases.

Can varying D fix model misspecification?

The sceptics amongst the readers may wish to see proof that addressing model misspecification should be done via the loss function—rather than via the regularizer. To this end, we now present a second variation of the above experiment. It not only illustrates that the frequentist consistency results of Chapter 3 extend far beyond the conditions outlined in the theorems presented there (in particular, they seem to extend to latent variable models!) but also demonstrates that tackling model misspecification cannot be tackled by changing the prior regularizer.

To show this, two settings are compared: In the first setting, the data is generated as before. In the second setting, additional noise is injected. Specifically, after x_i is generated according to eq. (5.13), inference is based on the polluted observation $\widetilde{x}_i$ generated as

$$\begin{aligned} \widetilde{x}_i^o &= x_i^o + u_i \cdot \eta_i \cdot e_d \\ u_i &\overset{i.i.d.}{\sim} \text{Bernoulli}(0.05) \\ \eta_i &\overset{i.i.d.}{\sim} \mathcal{N}(10, \sqrt{3}). \end{aligned} \tag{5.14}$$

To re-iterate our point—that model misspecification of this kind should be addressed by the loss rather than the regularizer—two different types of loss functions are compared: Firstly, the standard negative log likelihood given by

$$\ell(\boldsymbol{\theta}, x_i) = -\log p(x_i|\boldsymbol{\theta}).$$

Secondly, a robust scoring rule derived from the γ-divergence (Hung et al., 2018) given by

$$\mathcal{L}_p^\gamma(\boldsymbol{\theta}, x_i) = -\frac{1}{\gamma - 1} p(x_i|\boldsymbol{\theta})^{\gamma-1} \cdot \frac{\gamma}{I_{p,\gamma}(\boldsymbol{\theta})^{-\frac{\gamma-1}{\gamma}}}.$$

For the relevant background on robust minimium scoring rules like this, the reader will have to wait until Chapter 6. Alternatively, Dawid et al. (2016) and Jewson et al. (2018) provide a more thorough overview. For the moment, it is only important to note that the log score is not robust to misspecification (see e.g. Jewson et al., 2018, and references therein). In contrast, $\mathcal{L}_p^\gamma$ defines a scoring rule that is strongly robust to contamination (Fujisawa and Eguchi, 2008; Hung et al., 2018; Nakagawa and Hashimoto, 2019). The degree of robustness is regulated by γ: While $\gamma > 1$ produces more robust inferences than the log score, $\mathcal{L}_p^\gamma$ recovers the log score as $\gamma \to 1$. Consequently, one should expect $\mathcal{L}_p^\gamma$ for $\gamma = 1 + \varepsilon$ for very small values of $\varepsilon > 0$ to produce desirable inferences. For $\gamma = 1 + \varepsilon$, inferences are nearly as data-efficient as under the log score if the model is correctly specified. This small loss in efficiency buys us something priceless: the inferences remain more reliable under misspecification.

Figure 5.10 depicts this behaviour and connects it to the consistency findings in Chapter 3. The plot demonstrates three phenomena: Firstly, robustness to model misspecification cannot be achieved by adjusting D. Secondly, the exact path and speed of the convergence for GVI posteriors depends on the choice for D, especially

for small sample sizes. Thirdly, the overall patterns are the same across all choices of D and are dictated by the choice of ℓ. This should not come as a surprise: For $n \to \infty$, the GVI posteriors concentrate around $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_\mu [\ell(\boldsymbol{x}, \boldsymbol{\theta})]$, which does not depend on D. Note also that while $\mathcal{L}_p^\gamma$ recovers the true parameter values of eq. (5.13) for $n \to \infty$ in both the misspecified and well-specified setting, the log only manages to recover the true parameter values in the well-specified setting. In particular, GVI posteriors based on the log score concentrate around a sub-optimal parameter value in the misspecified setting, regardless of the choice for D.

5.3.2 Bayesian Neural Networks (BNNs)

As alluded to in Example 1.1, BNNs should be expected to suffer from prior misspecification: they are black box models, so that specifying a meaningful prior belief over their parameterizations is a daunting task. Focusing on the regression case, we wish to alleviate this problem using GVI and thus focus on varying D. Accordingly, we fix the loss function to the usual negative log likelihood $\ell(\boldsymbol{\theta}, y_i, x_i, \sigma^2) = -\log p_{\mathcal{N}}(y_i | x_i, \sigma^2, F(\boldsymbol{\theta}))$ for

$$p_{\mathcal{N}}(y_i | x_i, \sigma^2, F(\boldsymbol{\theta})) = \mathcal{N}(y_i | F(\boldsymbol{\theta}), x_i, \sigma^2),$$

where we choose $\mathcal{Q} = \mathcal{Q}_{\text{MFN}}$ as the normal mean field variational family given in eq. (1.3). With this in hand, we compare three different constructions of posterior beliefs:

(1) **Standard VI**;

(2) A **DVI** method motivated as approximations to the standard Bayesian posterior $q^*_{n,\text{SB}}(\boldsymbol{\theta})$ that find $q^*_{\text{A}}(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} D(q \| q^*_{n,\text{SB}}(\boldsymbol{\theta}))$; with D being the α-divergence (Hernández Lobato et al., 2016)[2] and Rényi's α-divergence (Li and Turner, 2016);

(3) **GVI** with $D = D^{(\alpha)}_{AR}$.

To make comparisons as fair as possible, our implementation is built on top of that used for the results of Li and Turner (2016) and only changes the objective being optimized. Similarly, all settings and data sets for which the methods are compared are unchanged and taken directly from Li and Turner (2016) and Hernández Lobato et al. (2016): We use a single-layer network with 50 ReLU nodes on all experiments.

[2] We align the parameterization of the $D^{(\alpha)}_A$ with the thesis, meaning $1 - \alpha_{\text{current}} = \alpha_{\text{H.-L. et al. (2016)}}$

Inference is performed via probabilistic back-propagation (Hernández Lobato and Adams, 2015) and the ADAM optimizer (Kingma and Ba, 2014) with its default settings, 500 epochs and a batch size of 32. Priors and variational posteriors are both fully factorized normal distributions. Further, the results are also evaluated on the same selection of UCI data sets (Lichman, 2013) and in the same way as they were in Li and Turner (2016) and Hernández Lobato et al. (2016): Using 50 random splits of the relevant data into training (90%) and test (10%) sets, the inferred models are evaluated predictively on the test sets using the average negative log likelihood (NLL) as well as the average root mean square error (RMSE). For each of the 50 splits, predictions are computed based on 100 samples from the variational posterior.

We summarize the two main results of our experiments as follows: First, Figure 5.11 depicts what appears to be the most typical relationship between **VI**, **DVI** and **GVI** on BNNs. Second, Figure 5.12 explores a surprising finding about the typical relationship further and connects it back to the modularity inherent in **GVI**, but absent from **DVI**. Appendix A.2 contains some further results that reinforce these findings.

Typical patterns (Figure 5.11)

As Figure 5.11 demonstrates, several findings form a consistent pattern across a range of data sets. Three findings are most poignant.

(A) **DVI** can often achieve a performance gain for the NLL relative to **standard VI**, but much less so for RMSE. On both metrics, there is no clear pattern of improvement.

(B) Relative to **standard VI**, **GVI** significantly improves performance for both NLL and RMSE if $\alpha > 1$. Conversely, **GVI** worsens performance if $\alpha \in (0, 1)$. In other words, *larger posterior variances adversely affect predictive quality.*

(C) **GVI** performance is a clear banana-shaped function of α across all data sets: While predictive performance benefits as α gets larger than one, the improvement flattens out and bends back in a banana shape as α grows too large. In other words, *decreasing uncertainty relative to the standard variational posterior improves predictive performance, but becoming 'overconfident' worsens it.*

Finding (B) has a straightforward interpretation: Since it holds that $D_{AR}^{(\alpha)} \leq \text{KLD}$ for

$\alpha > 1$ (see Van Erven and Harremos (2014)[3] and Figure 5.1), the GVI posteriors associated with $D_{AR}^{(\alpha)}$ for $\alpha > 1$ are *more* concentrated than the standard **VI** posteriors, a phenomenon also depicted on toy models in Figure 5.6. In other words: Ignoring more of the poorly specified prior and consequently being closer to a point mass at the empirical risk minimizer is beneficial for predictive performance. As alluded to in Example 1.1, this is to be expected: it is doubtful if a literal interpretation of the prior as in **(P)** is appropriate for BNNs. As finding (C) shows however, this does not mean that point estimates are preferable to posterior beliefs: Increasing the value of α shrinks the variances too much, eventually impeding predictive performance.

An advantage of GVI over DVI: A transparent optimization problem

While findings (B) and (C) should not come as a surprise by themselves, they do raise an interesting question: In particular, **GVI** for $D_{AR}^{(\alpha)}$ with $\alpha = 0.5$ is the *worst-performing* setting across the board. This is remarkable because this setting also constructs the only **GVI** posteriors in our experiments with *wider variances* than standard **VI**. At the same time, producing wider variances and more conservative uncertainty quantification is one of the main motivations for Expectation Propagation (EP) and the presented **DVI** methods, see for example Figure 1(a) in Li and Turner (2016) or Figure 8 in Hernández Lobato et al. (2016). This is puzzling: Are wider variances for $\boldsymbol{\theta}$ somehow beneficial for **DVI** posteriors' predictive performance while damaging that of **GVI** posteriors? As it turns out, this is not the case. Rather, while both **GVI** with $\alpha = 0.5$ and all **DVI** methods produce parameter posteriors with larger variances, in the case of **DVI** this does not translate into predictive uncertainty—as would be expected in standard Bayesian inference.

This phenomenon is depicted in Figure 5.12, which clearly shows that the additional uncertainty in the **DVI** parameter posteriors $q^*_{\text{DVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)$ is completely overshadowed by an extreme degree of variance shrinkage in the corresponding posterior predictives. In other words, the increased uncertainty in $\boldsymbol{\theta}$ is outweighed by extremely small values for σ^2. The plot demonstrates this by comparing the push-forward $F_{\#}q^*_{\text{DVI}}(\cdot|\boldsymbol{\kappa}^*)$ with the posterior predictives. Formally, the push-forward is given by

$$p(\mu|x_i) = \left(F_{\#}q^*_{\text{DVI}}(\cdot|\boldsymbol{\kappa}^*)\right)(\mu),$$

where the operation $_{\#}$ is simply a formalization of the following two operations:

[3] Note that their result holds for a different parameterization of the $D_{AR}^{(\alpha)}$, but it is easy to show that our parameterization is strictly smaller than theirs for $\alpha > 1$.

(i) sample $\boldsymbol{\theta} \sim q^*_{\text{DVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)$, (ii) compute $\mu = F(\boldsymbol{\theta})$. The posterior predictive then integrates the push-forward measure $p(\mu|x_i)$ over the likelihood function as

$$p(y_i|x_i) = \int_{\boldsymbol{\Theta}} p_{\mathcal{N}}(y_i|x_i, \sigma^2, F(\boldsymbol{\theta}))q^*_{\text{DVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)d\boldsymbol{\theta} = \int_{\mathbb{R}} p_{\mathcal{N}}(y_i|x_i, \sigma^2, \mu)p(\mu|x_i)d\mu.$$

As Figure 5.12 shows, the push-forward (i.e. the posterior predictive) behaves as expected for both **GVI** and **VI**. For **DVI**, the same cannot be said: specifically, the posterior predictive generally has much *less* variance than that of standard **VI**.

This surprising phenomenon is due to hyperparameter optimization for σ^2: Since Variational Inference on σ^2 complicates the **DVI** objectives, both Hernández Lobato et al. (2016) and Li and Turner (2016) do not infer σ^2 probabilistically. Instead, it is optimized over their objectives. This approach poses an optimization problem which for $D = D_A^{(\alpha)}$ and $D = D_{AR}^{(\alpha)}$ is given by

$$\widehat{\sigma}^2, q^*_{\text{DVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*) = \arg\min_{\sigma^2} \left\{ \arg\min_{q \in \mathcal{Q}} D(q(\boldsymbol{\theta}|\boldsymbol{\kappa})||q^*_{\text{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n})) \right\}. \quad (5.15)$$

Crucially, the inner part of this objective conditions on the exact Bayesian posterior for a *fixed* value of σ^2 and then seeks to approximate the posterior belief given by

$$q^*_{\text{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n}) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^{n} p_{\mathcal{N}}(y_i|x_i, \sigma^2, F(\boldsymbol{\theta})).$$

At the same time however, the outer part of the objective seeks to find a value for σ^2 which makes the posterior $q^*_{\text{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n})$ as easily approximable as possible. In other words, an objective which is explicitly motivated as a projection of $q^*_{\text{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n})$ into $\mathcal{Q}$ also changes the very point from which to project into $\mathcal{Q}$.

Though it would be computationally easy to perform probabilistic inference on σ^2 within **GVI**, we also optimize σ^2 as a hyperparameter for comparability. Thus, we pose the alternative optimization problem

$$\widehat{\sigma}^2, q^*_{\text{GVI}}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*) = \arg\min_{\sigma^2} \left\{ \arg\min_{q \in \mathcal{Q}} \left\{ \mathbb{E}_q \left[\sum_{i=1}^{n} -\log p_{\mathcal{N}}(y_i|x_i, \sigma^2, F(\boldsymbol{\theta})) \right] + D_{AR}^{(\alpha)}(q||\pi) \right\} \right\}. \quad (5.16)$$

As Figure 5.12 shows, the outcomes are drastically different: Unlike in the **DVI** case, the predictive uncertainty for the **GVI** posterior moves in the same direction as parameter uncertainty as α varies. The modularity of **GVI** makes it obvious what the optimization over σ^2 corresponds to in eq. (5.16): Rather than

choosing a posterior $q^*_{\mathrm{B}}(\boldsymbol{\theta}|\sigma^2, x_{1:n}, y_{1:n})$ which minimizes the cost of projecting into $\mathcal{Q}$ via $D^{(\alpha)}_{AR}$, the optimization problem simply seeks to find the best possible loss $\ell_{\sigma^2}(y_i|x_i, F(\boldsymbol{\theta})) = -\log p(y_i|x_i, \sigma^2, F(\boldsymbol{\theta}))$ over all $\sigma^2 \in \mathbb{R}_+$.

5.4 Conclusion & Summary

As outlined in Section 1.1.3, inference outcomes are adversely affected if the prior does not at least approximately reflect the best available judgement about good values of $\boldsymbol{\theta}$ before any data is seen. This is a problem whenever the prior is specified according to some (more or less arbitrary) default setting. For example, for the case of Bayesian Neural Networks (BNNs) that we have studied in the current chapter, a typical choice of prior is a multivariate standard normal distribution that factorizes over all network weights. While this may seem harmless or even uninformative, a supposedly uninformative prior specification of this kind actually encompasses a large degree of information, e.g.

(U) The prior belief is *unimodal*. In other words, we believe that there exists a *uniquely most likely* parameterization of the network before observing any data.

(I) The prior belief is that all network weights of a BNN are uncorrelated. In fact, we even believe that all network weights of a BNN are both *pairwise and mutually independent*.[4]

The above implications are in direct and strong contradiction to our best possible judgements about BNNs and thus violate **(P)**:

(↯U) Neural Networks are well-understood to have multiple parameter settings that are equally good (e.g. Choromanska et al., 2015). The unimodality assumption outlined in **(U)** is thus clearly not a reflection of the best judgement available: A prior belief in accordance with **(P)** would encode multimodality.

(↯I) By construction, Neural Networks encode a significant degree of dependence in their parameters: The best values for parameters in the l-th layer will strongly depend on the best values for parameters in the $(l-1)$-th layer (and vice versa). Hence, assuming uncorrelatedness (much less so independence!) directly contradicts our best judgement.

[4] For joint normal distributions, variables are uncorrelated if and only if they are independent.

From this, it is obvious that a fully factorized normal distribution is hardly an appropriate default prior for BNNs in the sense of **(P)** in Section 1.1.1. At the same time, it is often prohibitive or computationally infeasible to construct alternative prior beliefs that reflect our best judgements more accurately. In other words, we are stuck with a sub-optimal prior. Under the standard Bayesian paradigm, this is not an acceptable position. In contrast, the optimisation-centric view on Bayesian inference underlying the RoT and GVI do not require the prior to be flawless. We can thus use our very imperfect prior to design more appropriate posterior beliefs: Simply adapt the argument D which regularizes the posterior belief against the prior. In particular, we want to adapt D such that the resulting posteriors satisfy two criteria: Firstly, they should be more robust to priors which strongly contradict the observed data. Secondly, they should still provide reliable uncertainty quantification.

There is a host of robust alternatives to the KLD that we may hope behave in this way, most of which fall within the family of $\alpha\beta\gamma$-divergences. In this chapter, we studied the way in which these divergences affect prior robustness and uncertainty quantification in great detail. Some of the most important findings were that

- D should be unbounded over $\mathcal{Q}$ to prevent the posterior from collapsing to a point mass in finite samples. This rules out the family of α-divergences as well as the Total Variation Distance. Further and unsurprisingly, the larger the regularizers D, the larger the induced posterior variances.

- Using $D = \frac{1}{w}\text{KLD}$ for $w \in (0, 1)$ makes marginal variances larger, but is highly non-robust to misspecified priors. This should not come as a surprise, since all we do is giving more weight to the same regularizer that we were trying to fix in the first place. While $w > 1$ decreases the adversarial effects of misspecified priors, it also rapidly shrinks the posterior's marginal variances.

- The robust families of β- and γ-divergences induce fairly similar behaviour. While they are robust to misspecified priors for $\beta > 1$ (or $\gamma > 1$), this robustness comes at the price of a smaller marginal variance.

- Amongst all robust divergences that we examined, Rényi's α-divergence seems to exhibit the most desirable properties. Specifically, it guarantees prior robustness *without* tightening the marginal variances. Thus, it provides the prior robustness of β- and γ-divergences without the associated overconfident uncertainty quantification, see Appendix 5.2.3)

In conclusion, we find that Rényi's α-divergence provides prior robustness in the most practically useful way. For values of $\alpha \in (0, 1)$, it generally also provides

larger marginal variances than the KLD. Conversely, values of $\alpha > 1$ provide tighter marginal variances than the KLD. Specifically, this divergence produces similar posteriors as $D = \text{KLD}$ if the prior is *correctly* specified. However, unlike KLD, choosing Rényi's α-divergence continues to produce desirable uncertainty quantification when the prior is misspecified.

While $D_{AR}^{(\alpha)}$ behaves robustly, we should mention that it has one clear practical drawback relative to other potential regularizers such as the KLD or f-divergences. Specifically, eq. (5.2) defines it as a log expectation—meaning that standard stochastic inference techniques do not provide unbiased estimates for D. In the experiments of this thesis, we do not need to face this issue, as all experiments consider variational families $\mathcal{Q}$ that permit a closed form of $D_{AR}^{(\alpha)}$, so that the gradient estimator in (4.3) can be deployed. Note that this requirement is not particularly restrictive, as Rényi's α-divergence has closed forms for essentially all exponential family members (see Proposition 4.1).

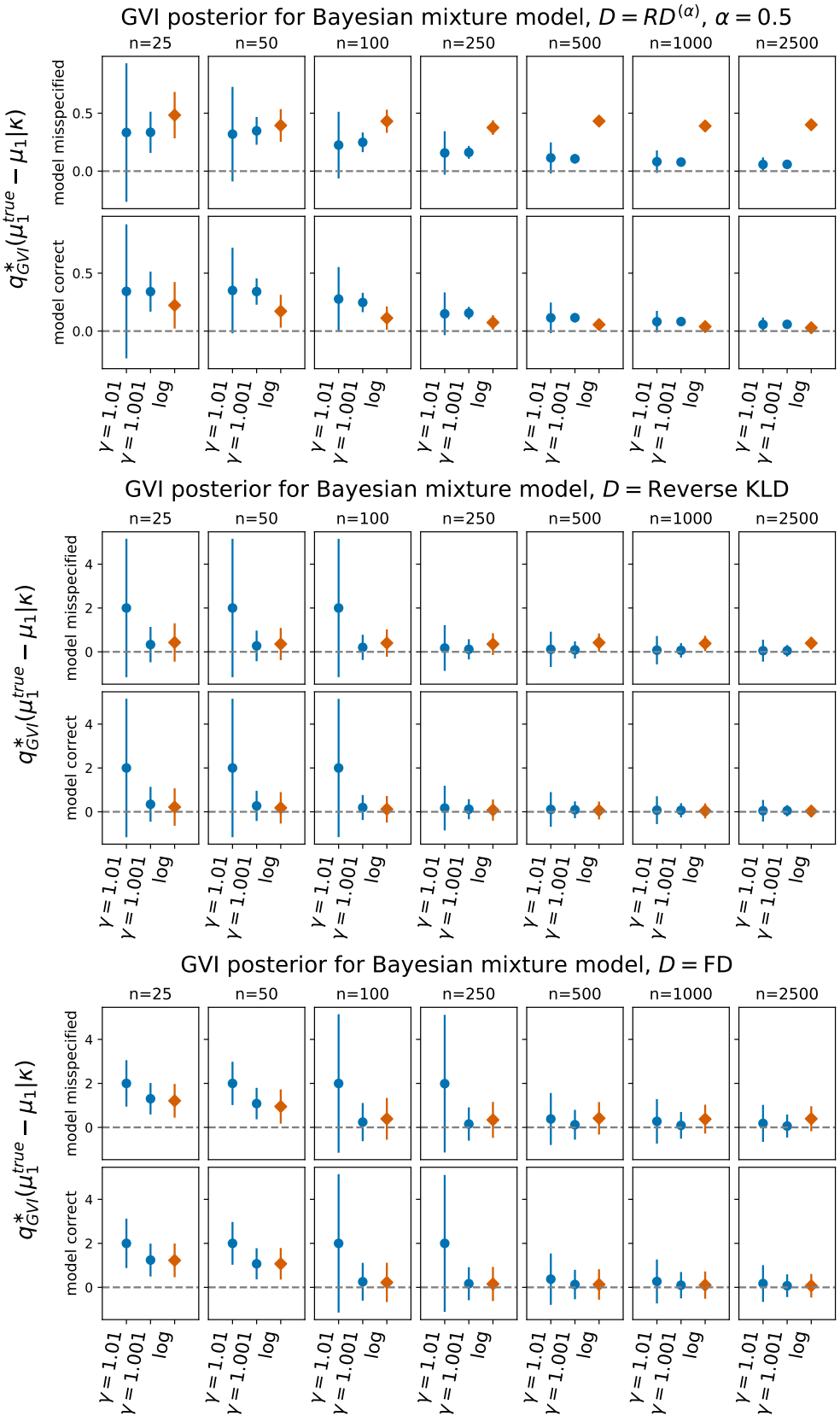


Figure 5.10: Depicted are the inferred **VI** and **GVI** posteriors for μ. Here, the **GVI** posteriors use $D = D_{AR}^{(\alpha)}$ for $\alpha = 0.5$ (top row), the reverse KLD (middle row), and the bottom row (Fisher divergence). Because all inferred posterior beliefs are normals, dots are used to mark out the posterior mean and whiskers to denote the posterior standard deviation. All posteriors are re-centered around the true value of β_1.

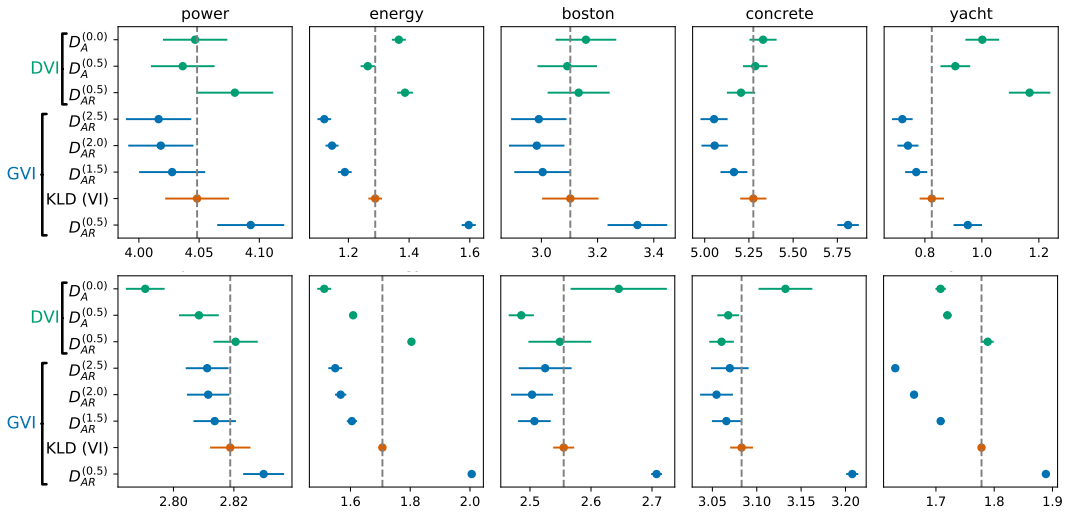


Figure 5.11: Best viewed in color. Top row depicts RMSE, bottom row the NLL across a range of data sets using BNNs. Dots correspond to means, whiskers to standard errors. The further to the left, the better the predictive performance. For the depicted selection of data sets, a clear common pattern exists for the performance differences between **standard VI**, **DVI** and **GVI**.

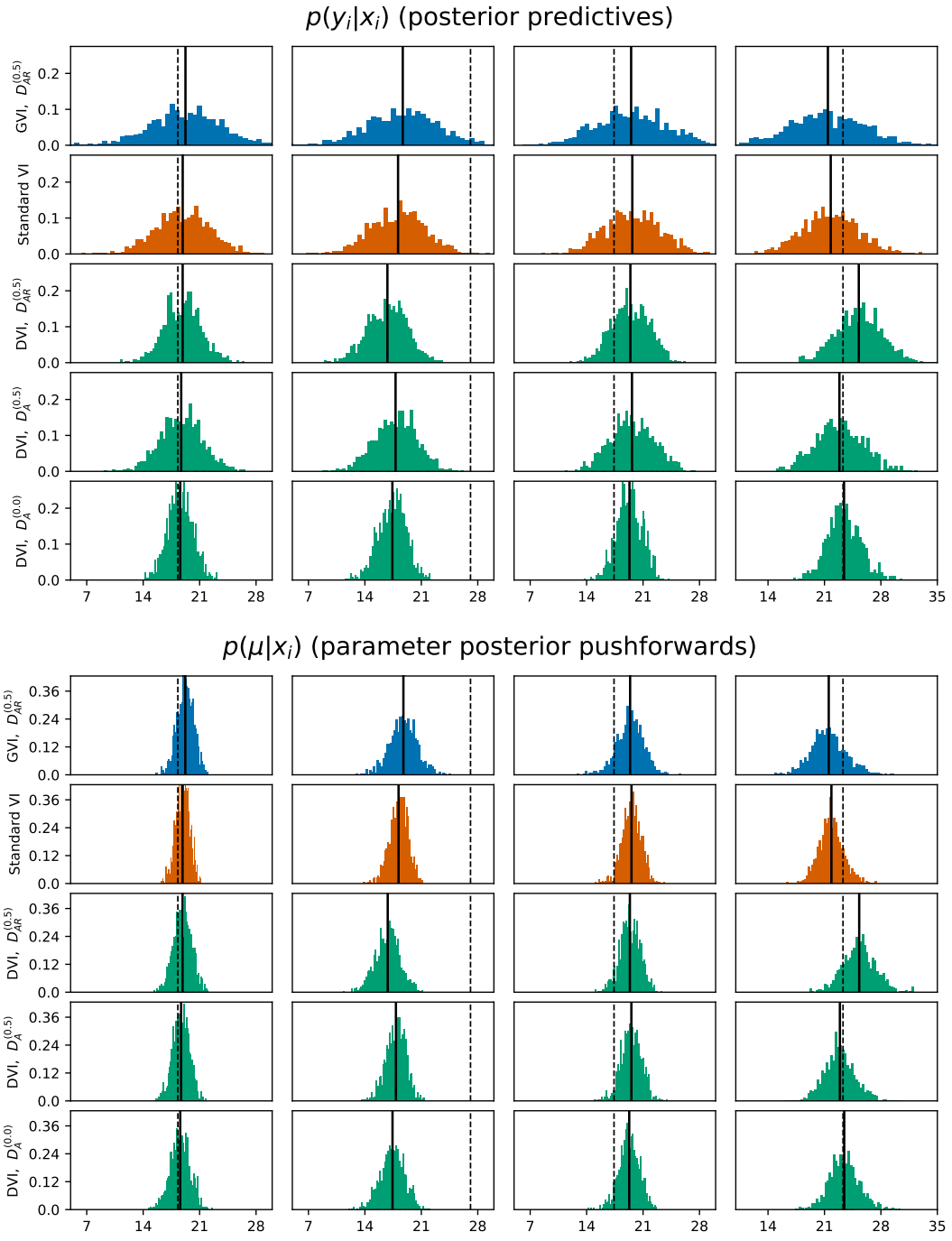


Figure 5.12: Best viewed in color. Depicted are test set predictions based on posterior predictives (**top panel**) and parameter posterior pushforwards (**bottom panel**) with four observations in the boston data set. Each column shows one observation (dashed line). The predictive distributions (histogram) and their means (solid line) for each row correspond to **standard VI**, **DVI** and **GVI**.

Chapter 6

Generalized Variational Inference, Part 3: Loss

Summary: In this and the two preceding chapters, we study the special case for the Rule of Three (RoT) for which optimization happens over a set of parameterized distributions. For its obvious relationship to previous variational methods, we call this family of algorithms Generalized Variational Inference (GVI). Broadly speaking, the current and third chapter on GVI explains how this methodology can address poorly specified likelihood functions in Bayesian methods—specifically in the context of modern Machine Learning models. We provide a short overview of some robust losses that could be used instead of the negative log likelihood, and explore their benefits on some numerical examples. We conclude with a case study for a widely popular Bayesian Machine Learning model: the Deep Gaussian Process (DGP) (Damianou and Lawrence, 2013), and show that robustifying it leads to significant improvements for predictive performance.

6.1 Robustness & Losses

It will be germane to first briefly recapitulate standard notions of robustness from the frequentist literature, and in particular the notion of influence functions. For a full treatment of influence functions and the field of robust frequentist statistics more broadly, we refer to the excellent monograph of Huber (2011).

6.1.1 Estimation & Influence Functions

Adopting notation that is standard in frequentist robust statistics, assume that $T : \mathcal{P}(\mathcal{X}) \to \boldsymbol{\Theta}$ is an estimator for some parameter of interest $\boldsymbol{\theta}$.[1] In traditional statistics, most estimators T are motivated as follows: Given the set of parameterized distributions $\{p(\cdot|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \subset \mathcal{P}(\mathcal{X})$, can we construct T so that $T(p(\cdot|\boldsymbol{\theta})) = \boldsymbol{\theta}$? In more practical terms, the question becomes whether for a sample $x_{1:n} \sim p(\cdot|\boldsymbol{\theta}_0)$ and the empirical measure $F(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$, we can guarantee that $T(F) \approx \boldsymbol{\theta}_0$, where the approximation should get exact as $n \to \infty$.

Influence functions are a way of assessing whether the function T is robust to misspecification: in other words, they help us assess how poor an estimator becomes if the data $x_{1:n}$ were generated by a distribution outside the collection $\{p(\cdot|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Specifically, they are meant to tell us what happens to T as the distribution that $x_{1:n}$ is sampled from another distribution G—and therefore look at limits of the form

$$\lim_{t \to 0} \left\{ \frac{T((1-t)H + tG) - T(H)}{t} \right\},$$

which is simply the Gateaux (i.e. functional) derivative of T at H in the direction of G. Here, H could be the true data-generating mechanism $p(\cdot|\boldsymbol{\theta}_0)$, or the empirical measure F constructed from $x_{1:n}$. Because directional derivatives (especially in function spaces) are not a particularly easy object to study, a simplified form is typically studied in practice; and usually it is this simplified form that people speak about when they study robustness.

Definition 6.1 (Influence Function)**.** Let $T : \mathcal{P}(\mathcal{X}) \to \boldsymbol{\Theta}$. The influence function of T at F in the direction of δ_x is given as

$$\mathrm{IF}(x; T, F) = \lim_{t \to 0} \left\{ \frac{T(t\delta_x + (1-t)F) - T(F)}{t} \right\}.$$

There are numerous ways in which the influence function can help us formalize robustness. Arguably the most important of these is the so-called *gross-error sensitivity* measure, which is defined as

$$S(T, F) = \sup_{x \in \mathcal{X}} |\mathrm{IF}(x; T, F)|.$$

[1] Note that this estimator is still applicable for a finite sample $x_{1:n}$ of data, since the empirical measure $\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$ is an element of $\mathcal{P}(\mathcal{X})$.

For a robust estimator, we want $S(T, F)$ to be finite—since this implies that the influence of perturbations to F (via δ_x) has a bounded and therefore limited impact on T.

In the context of the current thesis, it is important to note that Maximum Likelihood estimators—i.e., estimators based on minimizing the negative log likelihood—generally will *not* produce bounded influence functions. The result is that contaminants such as outliers or heterogeneity will severely affect how good a log likelihood based estimator T can perform on an imperfect, real-world data set. This is the main distinguishing feature between the negative log likelihood loss on the one hand, and the robust losses we will advocate for in this chapter on the other hand: while the former can produce seriously misleading estimates under misspecification, the latter have bounded influence functions and will retain desirable properties even under certain types of misspecification.

6.1.2 Robustness in the Bayesian setting

While Definition 6.1 defines robustness in frequentist procedures, it should be clear that these notions of robustness transfer straightforwardly into the Bayesian context. In this thesis, we mostly choose to not approach Bayesian robustness in an overly formal way.[2] The reason for this is relatively simple: in general, the existing formalisations of robustness in the Bayesian setting require that $P(L, D, \Pi)$ be analytically available. Since this requirement is not satisfied unless $D = \text{KLD}$ and $\Pi = \mathcal{P}(\boldsymbol{\Theta})$, the resulting theoretical analysis is limited to Gibbs posteriors—a small subset of the RoT. Moreover, common sense strongly suggests that $P(L, D, \Pi)$ will inherit the robustness properties of L so long as D and Π are not pathologically unsuitable choices. In this thesis, we will therefore mostly work with the rough hypothesis that *generally speaking, the* RoT *posterior* $P(L, D, \Pi)$ *will be robust to model misspecification if and only if the chosen loss* L *is robust to it.*

Obviously, this is not a satisfying solution from a theoretician's point of view, but we will see that it is both a reasonable and pragmatic approach—with considerable pay offs for methodology and applications.

6.1.3 A Selection of Robust Losses

While the notion of robust losses does not require a likelihood function, the current chapter (and indeed the thesis as a whole) is interested primarily in likelihood-based

[2] The exception to this will be Chapter 8, where we will more formally define robustness through adapted versions of influence functions in generalized Bayesian procedures by building on the work of Hooker and Vidyashankar (2014) and Ghosh and Basu (2016).

robust losses. The reason for this is simple: In the context of model misspecification, we are automatically talking about the relationship of a probability model $p(\cdot|\boldsymbol{\theta})$ and the true data-generating mechanism p_x of $x_{1:n}$. This means that we will ordinarily be interested in robust inferences relative to the (incorrect) model $p(\cdot|\boldsymbol{\theta})$—which necessitates a likelihood-based approach to designing robust losses.

Section 1.1.4 explained how and why **(↯L)** can severely impede the usefulness of standard Bayesian posteriors: if $p(\cdot|\boldsymbol{\theta})$ is not an accurate description of the data generating mechanism, inferences are susceptible to outliers, heterogeneity, and other adversarial aspects of the data. Recalling that the standard Bayes posterior is given by $q^*_{n,\mathrm{SB}}(\boldsymbol{\theta}) = P(-\log p(x_{1:n}|\boldsymbol{\theta}), \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, it is also clear that treating the likelihood model as (approximately) correct amounts to using the log score $L(\boldsymbol{\theta}, x_{1:n}) = -\log p(x_{1:n}|\boldsymbol{\theta})$ to assess how well $p(x_{1:n}|\boldsymbol{\theta})$ fits $\{x_i\}_{i=1}^n$. Indeed, this loss processes information about the likelihood model $p(x_{1:n}|\boldsymbol{\theta})$ contained in $x_{1:n}$ optimally within a Bayesian framework *if* this model happens to be correctly specified (Zellner, 1988).

While this implies that robust likelihood-based losses are typically less statistically efficient *under correct specification*, this tradeoff radically reverses even under mild misspecification (see e.g. Basu et al., 1998; Fujisawa and Eguchi, 2008; Hung et al., 2018; Jewson et al., 2018). For notational clarity, we will often write $L(\boldsymbol{\theta}, x_{1:n}) = L(p(\cdot|\boldsymbol{\theta}), x_{1:n})$ throughout the remainder of the current chapter to indicate a robust loss assessing the fit of likelihood parameter $\boldsymbol{\theta}$ on the sample $x_{1:n}$. The most appealing choices for $L : \boldsymbol{\Theta} \times \mathcal{X}^n \to \mathbb{R}$ are finite-sample estimators of $D(p_x(\cdot)\|p(\cdot|\boldsymbol{\theta}))$ for some robust divergence D. In other words, a natural loss is the estimated divergence between the true data-generating mechanism p_x that gave rise to $x_{1:n}$ on the one hand, and the model $p(\cdot|\boldsymbol{\theta})$ on the other hand. A notable advantage of designing losses this way is the following: even in the unlikely event that $p(\cdot|\boldsymbol{\theta})$ is correctly specified for p_x—so that there is $\boldsymbol{\theta}^*$ so that $x_{1:n}$ were drawn from the probability distribution with density given by $p(x_{1:n}|\boldsymbol{\theta}^*)$—minimizing an unbiased estimate of $D(p_x(\cdot)\|p(\cdot|\boldsymbol{\theta}))$ targets the correct value $\boldsymbol{\theta}^*$ for any statistical divergence D. So even though robust losses are less efficient than the log score under correct misspecification, they still recover the parameter value if the model happens to be correctly specified. An overview of some robust losses constructed in this way is provided in Table 6.1.

All losses presented in Table 6.1 guarantee various forms of robustness, and their main limiting factors are often of practical nature. To begin with, all except the Total Variation Distance depend on hyperparameters that are generally difficult to choose in a non-heuristic way. All of the non-additive losses in the table also

Divergence	Hyperparameters	Additive	References
α-divergence	$\alpha \in (0, 1)$	✗	Beran et al. (1977); Tamura and Boos (1986); Simpson (1987); Lindsay et al. (1994); Hooker and Vidyashankar (2014)
β-divergence	$\beta > 1$	✓	Basu et al. (1998); Ghosh and Basu (2016); Futoshi Futami et al. (2018)
γ-divergence	$\gamma > 1$	✓	Fujisawa and Eguchi (2008); Hung et al. (2018); Nakagawa and Hashimoto (2019)
Maximum Mean Discrepancy	Kernel k_ν and ν	✗	Briol et al. (2019); Chérief Abdellatif and Alquier (2022, 2020)
Kernel Stein Discrepancy	Stein Operator, kernel k_ν and ν	✗	Barp et al. (2019)
Total Variation Distance	—	✗	Yatracos (1985); Devroye and Lugosi (2012); Jeremias Knoblauch and Lara Vomfell (2020)

Table 6.1: Overview over robust likelihood-based losses derived from divergences

come with higher computational complexity, since non-additive losses do not admit unbiased estimation by sub-sampling. On top of this, such losses generally come with increased computational overhead. For example, kernel-based discrepancy measures such as the Maximum Mean Discrepancy or Kernel Stein Discrepancy are estimated using V-statistics or U-statistics. This means that evaluating these losses on a sample of size n has a computational complexity of $\mathcal{O}(n^2)$. Estimating losses based on the α-divergence or the Total Variation Distance is even more computationally demanding, since they require kernel density estimators if $\mathcal{X} = \mathbb{R}^p$. In summary, computational feasibility often makes additive losses such as those based on the β- and γ-divergences much more compelling than their non-additive alternatives. Accordingly, they are the robust losses we will study most throughout this chapter.

Additive losses based on β- and γ-divergences

At the time of writing, the only two divergence-based losses that are both robust and additive are those corresponding to the family of β- and γ-divergences first introduced by Basu et al. (1998) and Hung et al. (2018). We define them as

$L^\beta(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^n \mathcal{L}_p^\beta(\boldsymbol{\theta}, x_i)$ and $L^\gamma(\boldsymbol{\theta}, x_{1:n}) = \sum_{i=1}^n \mathcal{L}_p^\gamma(\boldsymbol{\theta}, x_i)$, where

$$\mathcal{L}_p^\beta(\boldsymbol{\theta}, x_i) = -\frac{1}{\beta - 1} p(x_i|\boldsymbol{\theta})^{\beta-1} + \frac{I_{p,\beta}(\boldsymbol{\theta})}{\beta} \tag{6.1}$$

$$\mathcal{L}_p^\gamma(\boldsymbol{\theta}, x_i) = -\frac{1}{\gamma - 1} p(x_i|\boldsymbol{\theta})^{\gamma-1} \cdot \frac{\gamma}{I_{p,\gamma}(\boldsymbol{\theta})^{\frac{\gamma-1}{\gamma}}}. \tag{6.2}$$

Here, the integral term $I_{p,c}(\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})^c d\boldsymbol{y}$ is generally available in closed form for exponential families. These losses are more robust than the log score whenever $\beta > 1$ (or $\gamma > 1$). Relative to other robust losses, they have two additional benefits: firstly and we saw in Section 4.5, they have desirable computational properties. Secondly, the hyperparameter β (or γ) has a clear interpretation since the losses recover the negative log likelihood as $\beta \to 1$ (or $\gamma \to 1$). To see this, one simply notes that $\lim_{x\to 1} \frac{z^{x-1}-1}{x-1} = \log z$ and $I_{p,1}(\boldsymbol{\theta}) = 1$. Thus—unlike the other entries in Table 6.1 except the α-divergence—the losses L^β and L^γ can be made arbitrarily close to the standard negative log likelihood. More specifically, choices of $\beta = 1+\varepsilon$ (or $\gamma = 1+\varepsilon$) for small enough $\varepsilon > 0$ will provide a loss function that is both robust *and* nearly as statistically efficient as the negative log likelihood. Unfortunately, it is generally difficult to pick the optimal degree of robustness ε because its optimal level will depend on both the scale of the data $x_{1:n}$ as well as the likelihood model. However, in numerous experiments, we found that *if the data are standardized*, values for $\varepsilon \in [0.01, 0.1]$ will yield a very favourable trade-off between robustness and efficiency across a very wide range of data sets, models, and forms of misspecification.

The estimators arising from minimizing either choice of loss can also be shown to have bounded influence functions under mild regularity conditions and for numerous statistical models (see Basu et al., 1998; Hung et al., 2018); and are consistent in the frequentist sense.

In summary, robust losses based on the β- and γ-divergences are both practical and have numerous desirable properties—both from a theoretical as well as a computational view point. For completeness, we will also elaborate upon two other important robust losses outlined in Table 6.1; both of which are attractive alternatives for robust generalized posteriors if computational complexity of order $\mathcal{O}(n^2)$ is not prohibitive (see Chérief Abdellatif and Alquier (2020) and Chapter 8).

Kernel-Stein Discrepancy Loss

While our ultimate goal is to introduce the Kernel-Stein Discrepancy (KSD), in the context of this thesis there is merit in going the extra mile to understand KSD's origin: In Chapter 8, we will extensively analyse theoretical properties pertaining

to the KSD, many of which rely on its relationship to Stein's Method. To this end, we will use the notation of that chapter, and denote for $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$ by $L^q(\mathcal{X}, \mathbb{Q})$ both the set of functions $f : \mathcal{X} \to \mathbb{R}$ for which $\|f\|_{L^q(\mathcal{X},\mathbb{Q})} := (\int_{\mathcal{X}} |f|^q \mathrm{d}\mathbb{Q})^{1/q} < \infty$ and the normed space in which two elements $f, g \in L^q(\mathcal{X}, \mathbb{Q})$ are identified if they are $\mathbb{Q}$-almost everywhere equal. If $\mathbb{Q}$ is a Lebesgue measure, we simply write $L^q(\mathcal{X})$ instead of $L^q(\mathcal{X}, \mathbb{Q})$. Further, we write $\mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$ as the set of all Borel probability measures supported on $\mathbb{R}^d$ that admit an everywhere positive probability density function (p.d.f.) $p : \mathcal{X} \to \mathbb{R}_{>0}$ with continuous partial derivatives.

Stein discrepancies were originally proposed in Gorham and Mackey (2015) as statistical divergences that are both computable and capable of providing various forms of distributional convergence control. In technical terms, the approach is based on the well-known method of Stein (1972), which requires the identification of a linear operator $\mathcal{S}_{\mathbb{Q}} : \mathcal{H} \to L^1(\mathcal{X}, \mathbb{Q})$, depending on a probability distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$ and acting on a Banach space $\mathcal{H}$, such that

$$\mathbb{E}_{X\sim\mathbb{Q}}[\mathcal{S}_{\mathbb{Q}}[h](X)] = 0 \quad \forall h \in \mathcal{H}. \tag{6.3}$$

Such an operator $\mathcal{S}_{\mathbb{Q}}$ is called a *Stein operator* and $\mathcal{H}$ is called a *Stein set*. Given a distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$, there are infinitely many operators $\mathcal{S}_{\mathbb{Q}}$ satisfying (6.3). A convenient example is the *Langevin Stein operator* (Gorham and Mackey, 2015), defined for $\mathcal{X} = \mathbb{R}^d$, $\mathbb{Q} \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$ and a Banach space $\mathcal{H}$ of differentiable functions $h : \mathbb{R}^d \to \mathbb{R}^d$, as

$$\mathcal{S}_{\mathbb{Q}}[h](x) = h(x) \cdot \nabla \log q(x) + \nabla \cdot h(x) \tag{6.4}$$

where q is the p.d.f. of $\mathbb{Q}$. Under suitable regularity conditions on $\nabla \log q$ and $\mathcal{H}$, the Langevin Stein operator satisfies Equation 6.3; see Gorham and Mackey (2015, Proposition 1). Given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$ and a Stein operator $\mathcal{S}_{\mathbb{Q}} : \mathcal{H} \to L^1(\mathcal{X}, \mathbb{Q})$ whose image is contained in $L^1(\mathcal{X}, \mathbb{P})$, the *Stein discrepancy* (SD) is defined as

$$\begin{aligned} \mathrm{SD}(\mathbb{Q}\|\mathbb{P}) &:= \sup_{\|h\|_{\mathcal{H}}\leq 1} \left| \mathbb{E}_{X\sim\mathbb{P}} \left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] - \mathbb{E}_{X\sim\mathbb{Q}} \left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] \right| \\ &= \sup_{\|h\|_{\mathcal{H}}\leq 1} \left| \mathbb{E}_{X\sim\mathbb{P}} \left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] \right|, \end{aligned} \tag{6.5}$$

where the last equality follows directly from (6.3). Under mild assumptions, a SD defines a statistical divergence between two probability distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$, meaning that $\mathrm{SD}(\mathbb{Q}\|\mathbb{P}) \geq 0$ with equality if and only if $\mathbb{P} = \mathbb{Q}$; see Proposition 1 and Theorem 2 in Barp et al. (2019). An important property of SDs that we will

exploit in Chapter 8 is that—unlike other divergences—SDs in general and the KSD in particular can be computed with an un-normalized representation of $\mathbb{Q}$.

Taking $x_{1:n}$ to be independently sampled from the true data generating process $\mathbb{P}$, letting $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be the empirical measure associated to this dataset, and letting $\mathbb{P}_{\boldsymbol{\theta}}$ denote the probability measure associated with our likelihood model $p(\cdot|\boldsymbol{\theta})$, SDs provide a natural likelihood-based loss function as $L(p(\cdot|\boldsymbol{\theta}), x_{1:n}) = \mathrm{SD}(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n)$, which is the estimated analogue of $\mathrm{SD}(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})$. This leaves one decisive question: How should we compute this loss in practice? In particular, how should we deal with the supremum of (6.5)?

Compared to other Stein discrepancies, KSDs are attractive because precisely because they correspond to a case where the supremum in (6.5) can be computed explicitly. To define the KSD, the first ingredient is a (matrix-valued) *kernel* $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}^{d\times d}$; the precise definition of which is deferred to Appendix A.3. For now, it suffices to point out that any kernel K has a uniquely associated Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}^d$, called a *vector-valued reproducing kernel Hilbert space*. This space constitutes the Stein set in KSD, and we therefore denote it as $\mathcal{H}$. Its associated norm and inner product will be denoted as $\|\cdot\|_{\mathcal{H}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, respectively. Then—and defining the action of a Stein operator $\mathcal{S}_{\mathbb{Q}}$ on both the first and second argument of a kernel K as $\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K$—the following result is a generalisation of the original KSD-construction of Chwialkowski et al. (2016) and Liu et al. (2016) to general Stein operators.

Proposition 6.1 (Closed form of SD)**.** Under Assumption 8.1 (see Chapter 8), we have

$$\mathrm{SD}^2(\mathbb{Q}\|\mathbb{P}) = \mathrm{KSD}^2(\mathbb{Q}\|\mathbb{P}) := \mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(X,X')\right]$$

where X and X' are independent.

The proof is in Appendix C.3.1. Note the immediate implication: Losses based on the KSD can be computed explicitly:

$$\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_n) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x_i,x_j), \tag{6.6}$$

where the explicit form of $\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K$ depends on $\mathcal{S}_{\mathbb{P}_\theta}$. For instance, the case of $\mathcal{X} = \mathbb{R}^d$

and the Langevin Stein operator in (6.4) is given by

$$
\begin{aligned}
&\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x') \\
&= \nabla_x \log p(x|\theta)\cdot K(x,x')\nabla_{x'}\log p(x'|\theta) + \nabla_x\cdot\left(\nabla_{x'}\cdot K(x,x')\right) \\
&\quad + \nabla_x \log p(x|\theta)\cdot\left(\nabla_{x'}\cdot K(x,x')\right) + \nabla_{x'}\log p(x'|\theta)\cdot\left(\nabla_x\cdot K(x,x')\right), \\
&= \sum_{i,j=1}^{d}\Bigg\{\frac{\partial}{\partial x_{(i)}}\log p_\theta(x)\left[K(x,x')\right]_{(i,j)}\frac{\partial}{\partial x_{(j)}}\log p_\theta(x) + \frac{\partial^2}{\partial x_{(i)}\partial x'_{(j)}}\left[K(x,x')\right]_{(i,j)} \\
&\quad + \frac{\partial}{\partial x_{(i)}}\log p_\theta(x)\frac{\partial}{\partial x'_{(j)}}\left[K(x,x')\right]_{(i,j)} + \frac{\partial}{\partial x'_{(j)}}\log p_\theta(x')\frac{\partial}{\partial x_{(i)}}\left[K(x,x')\right]_{(i,j)}\Bigg\}. \qquad (6.7)
\end{aligned}
$$

Clearly, this expression is straightforward to evaluate whenever we have access to derivatives of the kernel and the log density and can afford the corresponding $O(n^2)$ complexity. Moreover—and of particular interest for the current thesis—one can choose the kernel in ways that impart robustness in the frequentist sense (see Barp et al., 2019).

Maximum Mean Discrepancy Loss

The flexibility of choosing a kernel is a core feature of yet another likelihood-based robust loss function: the Maximum Mean Discrepancy (MMD), which is another loss that has been explored for generalized Bayesian methods (Chérief Abdellatif and Alquier, 2020). Unlike the KSD, the MMD is a *metric* on $\mathcal{P}(\mathcal{X})$ defined through a kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. Much like the KSD, the MMD exhibits an number of appealing computational and theoretical properties. In particular, for the reproducing kernel Hilbert space $\mathcal{H}$ induced by k and associated with the usual norm $\|\cdot\|_{\mathcal{H}}$,

$$
\mathrm{MMD}(\mathbb{P}\|\mathbb{Q}) := \sup_{h\in\mathcal{H}:\|h\|_{\mathcal{H}}\leq 1} \left|\mathbb{E}_{X\sim\mathbb{P}}[h(X)] - \mathbb{E}_{X\sim\mathbb{Q}}[h(X)]\right|.
$$

Much like for the MMD, we can solve this supremum in a convenient closed form. The key component for doing so is the *kernel mean embedding*—an integral transform $\mathbb{P}\mapsto\mu_{\mathbb{P}}$ that maps a distribution $\mathbb{P}\in\mathcal{P}(\mathcal{X})$ into an element of $\mathcal{H}_k$. It is defined by

$$
\mu_{\mathbb{P}}(\cdot) := \mathbb{E}_{X\sim\mathbb{P}}[k(X,\cdot)] \in \mathcal{H}, \qquad (6.8)
$$

whenever $\mathbb{E}_{X\sim\mathbb{P}}[k(X,\cdot)] < \infty$. In terms of kernel mean embeddings, the MMD between two distributions $\mathbb{P}$ and $\mathbb{Q}$ can then be written as

$$
\mathrm{MMD}(\mathbb{P}\|\mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.
$$

Conditions ensuring that $\mu_{\mathbb{P}}(\cdot)$ is well defined for any $\mathbb{P}$—so that $\text{MMD}(\mathbb{P}\|\mathbb{Q})$ is a metric—can be found in Sriperumbudur et al. (2010). An unrestrictive condition ensuring this and particularly easy to verify in practice is that the kernel be bounded.

If we can ensure that $\mu_{\mathbb{P}}(\cdot)$ is well defined for any $\mathbb{P}$, then using the well-known *reproducing property* of $\mathcal{H}$, one can show that

$$\text{MMD}(\mathbb{P}\|\mathbb{Q})^2 = \mathbb{E}_{X,X'\sim\mathbb{P}}[k(X,X')] - 2\mathbb{E}_{X\sim\mathbb{P},X'\sim\mathbb{Q}}[k(X,X')] + \mathbb{E}_{X,X'\sim\mathbb{Q}}[k(X,X')]$$

(see e.g., Section 3.5 of Krikamol Muandet et al., 2017). This is extremely convenient as it allows to compute unbiased estimators of $\text{MMD}(\mathbb{P}\|\mathbb{Q})^2$ by Monte Carlo methods: all we need are samples from $\mathbb{P}$ and $\mathbb{Q}$. More precisely, we can estimate the MMD as

$$\widehat{\text{MMD}(\mathbb{P}\|\mathbb{Q})^2} = \frac{1}{n^2}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(k(x_j,x_k) - 2k(x_j,y_k) + k(y_j,y_k)\right),$$

where $x_{1:J} \overset{i.i.d.}{\sim} \mathbb{P}$, and $y_{1:K} \overset{i.i.d.}{\sim} \mathbb{Q}$. Much like for the KSD in (8.2), this suggests the likelihood-based loss function $L(p(\cdot|\boldsymbol{\theta}), x_{1:n}) = \widehat{\text{MMD}(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n)^2}$. The resulting frequentist estimator exhibits various robustness properties that are derived and discussed by Briol et al. (2019), Alquier et al. (2020), and Pierre Alquier and Mathieu Gerber (2020).

Having studied the robust likelihood-based losses more broadly and β- and γ-divergence based losses in particular, we now turn to applying them in the context of a well-known black-box Bayesian Machine Learning model that invariably suffers under model misspecification: The Deep Gaussian Process (DGP).

6.2 Robustness for Deep Gaussian Processes

Machine Learning methods often shine when traditional modelling approaches fail. In other words, we know empirically that—at least in terms of predictive performance—we can often benefit from black-box models if it is hard to know our data-generating process exactly. This situation is also precisely what motivates the Deep Gaussian Process (DGP): with increasing depth, we attempt to provide an increasingly vague function prior. At the same time, in practice, the hyperparameters of that function prior are optimized as part of the inference procedure in an empirical-Bayes type fashion. This means that the prior will be adjusted to the data we present the DGP with, *relative to the likelihood function* that the DGP uses. For computational convenience, for the regression case this likelihood function in the DGP is typically

chosen to be a normal distribution, so that the hyperparameter optimization will try to find a latent function prior that conforms to a normal likelihood given the observations.

While this may often work in practice, remember that the motivation for using black-box models is that data-generating processes are not well-understood in the first place. Data of this type are likely to exhibit heterogeneities, outliers, and various other forms of misspecification. Why then impose the rather stringent normality requirement? Specifically, even simple forms of misspecification—such as heterogeneity via outliers—are well-understood to adversely affect posterior inferences if we do not protect ourselves against them. Fortunately, this is a problem that can straightforwardly be addressed by robust losses.

We illustrate this—using a much simpler model than the DGP for now—in Figure 6.1: As the right hand side panel shows, data are generated from a so-called ε-contaminated model: For some $\varepsilon \in (0, 1)$, a proportion ε of the data are contaminants or outliers, while the remaining $1 - \varepsilon$ come from the model specified by the user. This means that the true data-generating process is

$$p_{\text{true}}(x) = (1 - \varepsilon)p(x|\boldsymbol{\theta}_0) + \varepsilon c(x),$$

for some parameter $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ and some contamination distribution $c \in \mathcal{P}(\mathcal{X})$. In the right hand side panel of the Figure, $\varepsilon = 0.05$, the parametric model $p(\cdot|\boldsymbol{\theta}_0)$ is a standard normal; and data generated from it is displayed in light grey color. The contaminating distribution is also a normal, but with higher dispersion and with mean value of 8. Data generated from it is displayed in black. As we can also see on the right hand side panel, fitting a variational posterior with the negative log likelihood score of a normal will shift the posterior predictive distribution in the direction of this contamination. In contrast, the various GVI posteriors (constructed with robust losses based on β-divergences) do not suffer this problem.

On the left hand side of the panel, we quantify this difference in a more thorough way by displaying a Bayesian equivalent of influence functions. The displayed influence functions quantify the impact the $(n + 1)$-th observation has on the posterior distribution $q^*_{n,\text{SB}}(\boldsymbol{\theta})$ constructed from the first n observations (see Peng and Dey, 1995), where the influence is measured by computing the Fisher-Rao divergence between the posteriors based on $x_{1:n}$ and on $x_{1:(n+1)}$. On the x-axis, we have expressed the magnitude of x_{n+1} relative to the standard deviation of $q^*_{n,\text{SB}}(\boldsymbol{\theta})$. As the Figure shows, for the negative log likelihood loss with the normal distribution, the influence of x_{n+1} on the posterior belief grows stronger and stronger the more

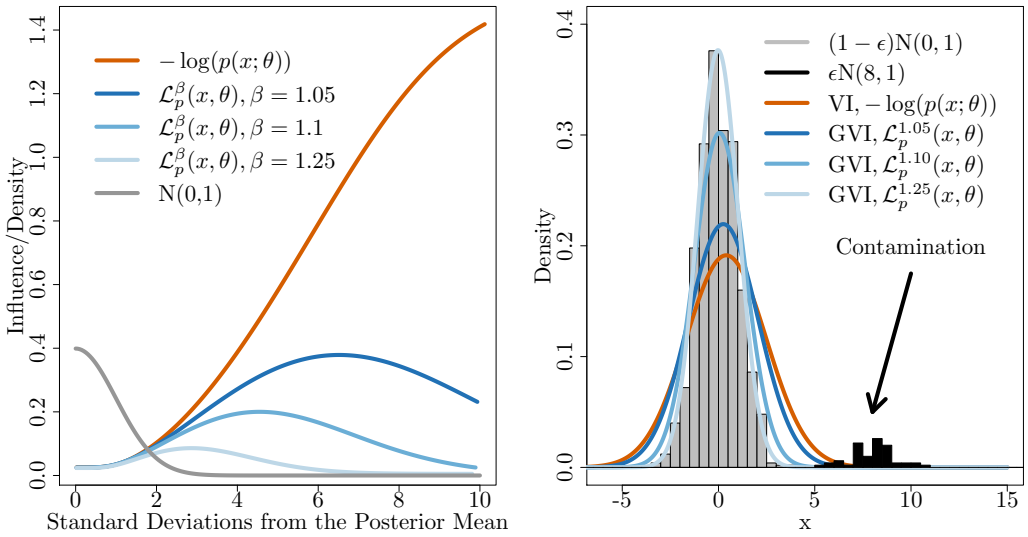


Figure 6.1: Best viewed in color. The plots compare influence functions (**Left**) and predictive posteriors (**Right**) of a **standard** Bayesian inference against a **GVI** posterior. **Left:** The influence functions of scoring the normal likelihood with a **standard** negative log likelihood against a **robust** scoring rule derived from β-divergences. Here, the influence is computed as the Fisher-Rao divergence between the posterior based on $n = 100$ and $n = 101$ observations, where we measure the magnitude by which the 101-th observation deviates from the first 100 observations through standard deviations from the posterior mean. For more details on this, see Kurtek and Bharath (2015). **Right:** A univariate normal is fitted using all the data depicted, including the outlying contamination. The posterior predictive corresponding to the **robust** scoring rule and $\beta = 1.25$ is able to ignore these outliers. This stands in contrast to the posterior predictive based on **standard Bayesian inference**, which assigns increasingly large influence to outlying observations.

untypical the observation is relative to previously observed data. This behaviour is not the same for the β-divergence, for which the outlier only gains influence up to a tipping point. Beyond said tipping point, the influence of the observation slowly but surely decays until the point where x_{n+1} has virtually no impact on the posterior anymore.

In summary, since we ordinarily apply black-box models like the DGP to data sets that are hard to model, it is reasonable to expect that these data sets may contain various contaminations and outliers. It stands to reason that in the presence of these contaminations however, one would substantially benefit from robustification of the DGP. This is what we set out to do next. To this end, we first

introduce the method and model to be improved.

6.2.1 VI for DGPs using Salimbeni & Deisenroth (2017)

While there are other variational methods that one could modify using GVI, we focus on the inference scheme introduced by Salimbeni and Deisenroth (2017). Unlike competing VI approaches for DGPs, this family encodes some part of the conditional dependence structure of the DGP. This comes at the expense of losing a tractable closed form lower bound (as in Damianou and Lawrence, 2013), but makes DGPs more practically viable by allowing for more flexible and adaptable function priors.

Deep Gaussian Processes (DGPs)

Deep Gaussian Processes (DGPs) were introduced by Damianou and Lawrence (2013) and extend the logic of deep learning to the nonparametric Bayesian setting. The principal idea is to iteratively place Gaussian Process (GP) priors over emerging latent spaces. More specifically, given matrices of observations $(\boldsymbol{X}, \boldsymbol{Y})$ where $\boldsymbol{X} \in \mathbb{R}^{n \times D}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$, a DGP of L layers introduces the additional collection of latent functions $\{\boldsymbol{F}^l\}_{l=1}^L$. Here, $\boldsymbol{F}^l$ is a matrix of dimension $D^l \times D^{l+1}$. Setting $\boldsymbol{F}^0 = \boldsymbol{X}$, $D^0 = D$ and $D^{l+1} = p$ for notational convenience, one can now write the hierarchical DGP construction as

$$
\begin{aligned}
\boldsymbol{Y}|\boldsymbol{F}^L & \sim p\left(\boldsymbol{Y}\middle|\, \boldsymbol{F}^L\right) \\
\boldsymbol{F}^L|\boldsymbol{F}^{L-1} = f^L(\boldsymbol{F}^{L-1}) & \sim \mathrm{GP}\left(\mu^L(\boldsymbol{F}^{L-1}), K^L(\boldsymbol{F}^{L-1}, \boldsymbol{F}^{L-1})\right) \\
\boldsymbol{F}^{L-1}|\boldsymbol{F}^{L-2} = f^{L-1}(\boldsymbol{F}^{L-2}) & \sim \mathrm{GP}\left(\mu^{L-1}(\boldsymbol{F}^{L-2}), K^{L-1}(\boldsymbol{F}^{L-2}, \boldsymbol{F}^{L-2})\right) \\
& \dots \\
\boldsymbol{F}^1|\boldsymbol{F}^0 = f^1(\boldsymbol{F}^0) & \sim \mathrm{GP}\left(\mu^1(\boldsymbol{F}^0), K^1(\boldsymbol{F}^0, \boldsymbol{F}^0)\right),
\end{aligned}
$$

where the mean and covariance functions are of form $\mu^l : \mathbb{R}^{D^l} \to \mathbb{R}^{D^{l+1}}$ and $K^l : \mathbb{R}^{D^l \times D^l} \to \mathbb{R}^{D^{l+1} \times D^{l+1}}$ for the collection of matrix-valued kernels $\{K^l\}_{l=1}^L$. Scalable inference in this construction is obviously a challenge. In principle, the attempts at tackling this problem rely on Variational inference (VI) strategies (Damianou and Lawrence, 2013; Dai et al., 2016; Salimbeni and Deisenroth, 2017; Hensman and Lawrence, 2014), Monte Carlo methods (Vafa, 2016; Wang et al., 2016) or more specialized approaches (Bui et al., 2016; Cutajar et al., 2017a). In the remainder, we will focus on variational strategies for DGP inference. To keep things as simple as possible, we discuss the implications of Generalized Variational Inference (GVI) only in relation to the arguably most promising VI approach of Salimbeni and Deisenroth

(2017) which encodes conditional dependence into the variational family $\mathcal{Q}$.

The conditionally dependent variational family for DGPs

Under the popular *inducing point framework* for GP inference (see Titsias; Edwin V. Bonilla et al., 2019; Matthews et al., 2016), we now introduce the exact Bayesian posterior arising from the DGP construction. First, we define the set of m additional inducing points $\boldsymbol{Z}^l = (\boldsymbol{z}_1^l, \boldsymbol{z}_2^l, \ldots, \boldsymbol{z}_m^l)^T$ and their function values $\boldsymbol{U}^l = (f^l(\boldsymbol{z}_1^l), f^l(\boldsymbol{z}_2^l), \ldots, f^l(\boldsymbol{z}_m^l))^T$ in addition to the observations $(\boldsymbol{X}, \boldsymbol{Y})$. In this context, the core idea is to choose $m << n$ in order to speed up computation via conditioning of $(\boldsymbol{Z}, \boldsymbol{U})$ on $(\boldsymbol{X}, \boldsymbol{Y})$. Throughout, we will often drop $\boldsymbol{X}$ and $\boldsymbol{Z}^l$ from the conditioning sets of the conditional probability distributions. Further, note that we will denote the i-th row of the $D^l \times D^{l+1}$ latent functions $\boldsymbol{F}^l$ as $\boldsymbol{f}_i^L$. With this in place, the joint distribution of the DGP construction is

$$p\left(\boldsymbol{Y}, \{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\right) = \underbrace{\prod_{i=1}^n p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)}_{\text{likelihood}} \times \underbrace{\prod_{l=1}^L p\left(\boldsymbol{F}^l \middle| \boldsymbol{U}^l, \boldsymbol{F}^{l-1}, \boldsymbol{Z}^{l-1}\right) p\left(\boldsymbol{U}^l \middle| \boldsymbol{Z}^{l-1}\right)}_{\text{(DGP) prior}}.$$

Thus, the posteriors $p\left(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\right)$ and $p\left(\{\boldsymbol{F}^l\}_{l=1}^L\right)$ are intractable due to the normalizing constants required for their computation. To overcome this, different variational approximations have been proposed. Here, we focus on the variational family proposed in Salimbeni and Deisenroth (2017) given by

$$q\left(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\right) = \prod_{l=1}^L p\left(\boldsymbol{F}^l \middle| \boldsymbol{U}^l, \boldsymbol{F}^{l-1}, \boldsymbol{Z}^{l-1}\right) q\left(\boldsymbol{U}^l\right), \tag{6.9}$$

$$q\left(\boldsymbol{U}^l\right) = \mathcal{N}\left(\boldsymbol{U}^l \middle| \boldsymbol{m}^l, \boldsymbol{S}_l\right). \tag{6.10}$$

The variational parameters for this posterior are $\boldsymbol{\kappa} = \left\{\{\boldsymbol{m}^l\}_{l=1}^L, \{\boldsymbol{S}_l\}_{l=1}^L\right\}$. The normal form for $q\left(\boldsymbol{U}^l\right)$ is chosen because it allows for exact integration over the inducing points $\{\boldsymbol{U}^l\}_{l=1}^L$, yielding the closed form variational posterior

$$q\left(\{\boldsymbol{F}^l\}_{l=1}^L\right) = \prod_{l=1}^L \mathcal{N}\left(\boldsymbol{F}^l \middle| \boldsymbol{\mu}^l, \boldsymbol{\Sigma}_l\right),$$

where the parameters of the posterior are available as

$$\left[\boldsymbol{\mu}^l\right]_i = \boldsymbol{\mu}^{\boldsymbol{m}^l, \boldsymbol{Z}^{l-1}}(\boldsymbol{f}_i^L) \qquad = \mu^l(\boldsymbol{f}_i^L) + \boldsymbol{a}(\boldsymbol{f}_i^L)^T\left(\boldsymbol{m}^l - \mu^l(\boldsymbol{Z}^{l-1})\right)$$

$$[\boldsymbol{\Sigma}_l]_{i,j} = \boldsymbol{\Sigma}_{\boldsymbol{S}_l, \boldsymbol{Z}^{l-1}}(\boldsymbol{f}_i^L, \boldsymbol{F}^{j,l}) = K^l(\boldsymbol{f}_i^L, \boldsymbol{F}^{j,l}) - \boldsymbol{a}(\boldsymbol{f}_i^L)^T\left(K^l(\boldsymbol{Z}^{l-1}, \boldsymbol{Z}^{l-1}) - \boldsymbol{S}_l\right)\boldsymbol{a}(\boldsymbol{F}^{j,l}),$$

and we define $\boldsymbol{a}(\boldsymbol{f}_i^L) = K^l(\boldsymbol{Z}^{l-1}, \boldsymbol{Z}^{l-1})^{-1} K^l(\boldsymbol{Z}^{l-1}, \boldsymbol{f}_i^L)$. Note the attractive feature of the family specified via eqs. (6.9) – (6.10): At each layer l, the output $\boldsymbol{f}_i^l$ only depends on the corresponding input $\boldsymbol{f}_i^{l-1}$. This property is a direct consequence of setting every layer up exactly as a sparse GP (see, e.g. Titsias; Hensman et al., 2013; Edwin V. Bonilla et al., 2019). This enables efficient probabilistic backpropagation (Hernández Lobato and Adams, 2015) with the reparameterization trick (e.g. Rezende et al., 2014; Kingma et al., 2015) and makes the approach scalable through the stochastic variational methods outlined in Chapter 4.

In particular, Salimbeni and Deisenroth (2017) propose a doubly stochastic minimization of the negative Evidence Lower Bound (ELBO) given by

$$O_{\text{VI}}(\boldsymbol{\kappa}) = -\sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{f}_i^L)} \left[\log p(\boldsymbol{y}_i | \boldsymbol{f}_i^L)\right] + \sum_{l=1}^{L} \text{KLD}(q(\boldsymbol{F}^l, \boldsymbol{U}^l) || p(\boldsymbol{F}^l, \boldsymbol{U}^l | \boldsymbol{Z}^{l-1})). \quad (6.11)$$

The Kullback-Leibler divergence (KLD) terms of this bound further simplify because by eq. (6.9), q is designed to cancel the conditional over $\boldsymbol{F}^l$ with p. This finally leads to the bound

$$O_{\text{VI}}(\boldsymbol{\kappa}) = -\sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{f}_i^L)} \left[\log p(\boldsymbol{y}_i | \boldsymbol{f}_i^L)\right] + \sum_{l=1}^{L} \text{KLD}(q(\boldsymbol{U}^l) || p(\boldsymbol{U}^l | \boldsymbol{Z}^{l-1})), \quad (6.12)$$

where for optimization the samples for $\boldsymbol{F}^l$ are drawn using the variational posteriors from the previous layers. Because $\boldsymbol{f}_i^L$ only depends on the corresponding input $\boldsymbol{F}^{i,l-1}$, this can be done using univariate Gaussians, which means that no matrix operations are required. Ultimately, it is this trick that enables the approach to produce more expressive variational approximations without losing computational efficiency.

Why is this method a so-called 'doubly stochastic' minimization? The first layer of stochasticity in the model stems from approximating the expectation over $q(\boldsymbol{\theta})$. The second layer is due to drawing a mini-batches from $\boldsymbol{X} = \boldsymbol{F}^0$ and $\boldsymbol{Y}$ at each iteration. Because of this degree of stochasticity, it is a particularly appealing feature that the expectations $\mathbb{E}_{q(\boldsymbol{f}_i^L)} \left[\log p(\boldsymbol{y}_i | \boldsymbol{f}_i^L)\right]$ in the very last layer are available in closed form for some choices of p—since this removes one layer of randomness from the procedure. Such closed forms are for instance available in the regression setting, where p is a normal likelihood. Later on, we also derive such closed forms for a new class of alternatives for p geared towards robustness and derived from normal likelihoods.

An alternative problem representation

We now decompose the components of the DGP model. Specifically, we define the collection of likelihood terms as

$$\ell_n\left(\{\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\}, \boldsymbol{Y}\right) = \sum_{i=1}^n \ell\left(\boldsymbol{f}_i^L, \boldsymbol{y}_i\right) \quad \text{for } \ell\left(\boldsymbol{f}_i^L, \boldsymbol{y}_i\right) = -\log p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)$$

and the layered DGP prior via

$$p\left(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L \middle| \{\boldsymbol{Z}^l\}_{l=1}^L\right) = \prod_{l=1}^L p_l\left(\boldsymbol{F}^l, \boldsymbol{U}^l \middle| \boldsymbol{F}^{l-1}, \boldsymbol{U}^{l-1}, \boldsymbol{Z}^{l-1}\right) \tag{6.13}$$

$$p_l\left(\boldsymbol{F}^l, \boldsymbol{U}^l \middle| \boldsymbol{F}^{l-1}, \boldsymbol{U}^{l-1}, \boldsymbol{Z}^{l-1}\right) = p\left(\boldsymbol{F}^l \middle| \boldsymbol{F}^{l-1}, \boldsymbol{U}^l, \boldsymbol{Z}^{l-1}\right) p\left(\boldsymbol{U}^l \middle| \boldsymbol{Z}^{l-1}\right). \tag{6.14}$$

With this, one can rewrite the sought-after posterior as

$$\begin{aligned} &p\left(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L \middle| \boldsymbol{Y}, \boldsymbol{X}\right) \\ &= \frac{\exp\left\{-\ell_n\left(\{\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\}, \boldsymbol{Y}\right)\right\} \pi\left(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L \middle| \{\boldsymbol{Z}^l\}_{l=1}^L\right)}{\int_{\boldsymbol{Y}} \exp\left\{-\ell_n\left(\{\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\}, \boldsymbol{Y}\right)\right\} \pi\left(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L \middle| \{\boldsymbol{Z}^l\}_{l=1}^L\right) d\boldsymbol{Y}} \end{aligned} \tag{6.15}$$

This representation gives a generalized Bayesian distribution associated with a general loss function ℓ. For the standard DGP, the loss function is the negative log likelihood $\ell(\boldsymbol{f}_i^L, \boldsymbol{y}_i) = -\log p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)$, which is the loss traditionally associated with the Bayesian paradigm. While this is the de-facto default choice for the loss, the variational methods outlined in the previous section still applies to any other additive loss ℓ—such as the β- and γ-divergence based losses introduced in Section 6.1.3.

6.2.2 GVI for DGPs

Note that throughout, we will—strictly speaking—be deriving posteriors for infinite-dimensional *latent functions* rather than finite-dimensional *parameters* $\boldsymbol{\theta}$. Yet, this distinction is somewhat semantic: In actual fact, we will not obtain posterior beliefs over all of the latent functions, but only the finitely many points of evaluation in $\{\boldsymbol{F}^l\}_{l=1}^L$ and $\{\boldsymbol{U}^l\}_{l=1}^L$. For this reason, one may think of these latent functions as de-facto parameters and think of $\{\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L\}$ as the parameter of interest.

β- and γ-divergences for DGPs

Our idea for robustifying DGPs is as simple as it is elegant: replace the negative log likelihood loss by the summable and robust β- or γ-divergence based losses $\mathcal{L}_p^\beta$ or

$\mathcal{L}_p^\gamma$, leading for some choice of divergence D to the objectives

$$O^\beta_{\text{DGP,GVI}}(\boldsymbol{\kappa}) = \sum_{i=1}^n \mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\beta(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] + \sum_{l=1}^L D(q(\boldsymbol{F}^l, \boldsymbol{U}^l)||p(\boldsymbol{F}^l, \boldsymbol{U}^l|\boldsymbol{Z}^{l-1})),$$
$$O^\gamma_{\text{DGP,GVI}}(\boldsymbol{\kappa}) = \sum_{i=1}^n \mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\gamma(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] + \sum_{l=1}^L D(q(\boldsymbol{F}^l, \boldsymbol{U}^l)||p(\boldsymbol{F}^l, \boldsymbol{U}^l|\boldsymbol{Z}^{l-1})).$$

Through tedious but straightforward calculation, one can show that the expectations $\mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\beta(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right]$ and $\mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\gamma(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right]$ are available in closed form for the regression setting when $\mathcal{L}_p^\beta$ and $\mathcal{L}_p^\gamma$ are based on a normal likelihood.

Theorem 6.1 (Closed form for robust regression)**.** If it holds that $\boldsymbol{y}_i \in \mathbb{R}^d$,

$$p(\boldsymbol{y}_i|\boldsymbol{f}_i^L) = \mathcal{N}\left(\boldsymbol{y}_i; \boldsymbol{f}_i^L, \sigma^2 I_d\right); \quad q(\boldsymbol{f}_i^L) = \mathcal{N}(\boldsymbol{f}_i^L; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{6.16}$$

then for the quantities given by

$$\widetilde{\boldsymbol{\Sigma}}^{-1} = \left(\frac{c}{\sigma^s}\boldsymbol{I}_d + \boldsymbol{\Sigma}^{-1}\right); \quad \widetilde{\boldsymbol{\mu}} = \left(\frac{c}{\sigma^2}\boldsymbol{y}_i + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right); \quad I(c) = (2\pi\sigma^2)^{-0.5dc} c^{-0.5d} \tag{6.17}$$

and for

$$E(c) = \frac{1}{c}\left(2\pi\sigma^2\right)^{-0.5dc} \frac{|\widetilde{\boldsymbol{\Sigma}}|^{0.5}}{|\boldsymbol{\Sigma}|^{0.5}} \exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}\right)\right\} \tag{6.18}$$

the following expectations are available in closed form:

$$\mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\beta(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] = -E(\beta - 1) + \frac{I(\beta)}{\beta} \tag{6.19}$$
$$\mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\beta(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] = -E(\gamma - 1) \cdot \frac{\gamma}{I(\gamma)^{\frac{\gamma}{\gamma-1}}} \tag{6.20}$$

We defer the proof of this result to Appendix B.7.1. Note that for numerical stability, $\mathcal{L}_p^\gamma$ is the preferable loss: it is multiplicative and—unlike $\mathcal{L}_p^\beta$—it never changes sign. Thus, it can be processed and stored entirely in log form, which is a desirable property for stable gradients in stochastic variational methods.

6.2.3 Varying the regularizer

Unlike their frequentist counterparts, Bayesian methods provide uncertainty quantification about the latent parameters (or functions) of interest. Specifically, uncertainty about values of $\boldsymbol{\theta}$ is quantified by penalizing how far the posterior q diverges

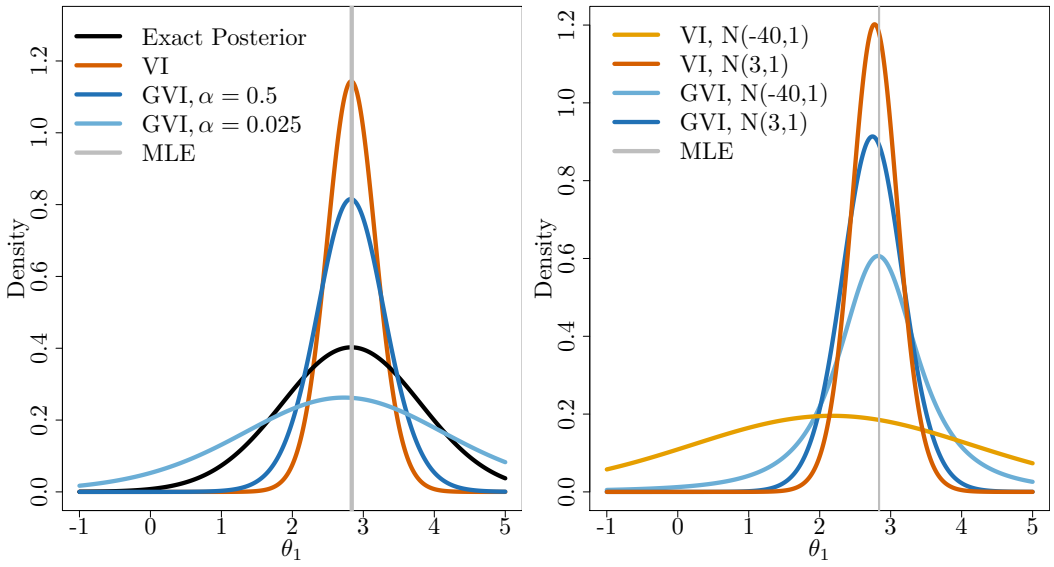


Figure 6.2: Comparing standard VI ($D = \text{KLD}$) against GVI with $D = D_{AR}^{(\alpha)}$ using posteriors with Gaussian likelihoods and mean-field Gaussian approximations. **Left:** Changing D improves marginal variances. Depicted are exact and approximate marginals. The exact posterior is correlated, causing VI to over-concentrate. GVI can avoid this. **Right:** Changing D provides prior robustness. Depicted are approximate marginals for two different priors $\pi \in \{N(-30, 2^2), N(-5, 2^2)\}$. VI is sensitive to the badly specified prior. GVI can avoid this.

from the prior π. To assess whether there is a benefit of using robust regularizers for DGPs, we focus on Rényi's α-divergence ($D_{AR}^{(\alpha)}$) as introduced in Chapter 5. This divergence is available in closed form for the variational families and priors on DGPs of (Salimbeni and Deisenroth, 2017) for $\alpha \in (0, 1)$. More importantly, it provides larger marginal variances than VI for $\alpha \in (0, 1)$, tighter marginal variances than VI for $\alpha > 1$, and is robust to badly specified priors. We refer to Figure 6.2 for an illustration of both properties. We further point to Chapter 5, which contains a much wider selection of pictorial examples that also encompass other divergences.

As a second alternative to $D = D_{AR}^{(\alpha)}$, we also consider $D = \frac{1}{w}\text{KLD}$ (see also Yang et al., 2020). Note that this has an intimate relationship to power likelihoods. In particular, using the negative power log likelihood $-\log p(\boldsymbol{y}_i|\boldsymbol{\theta})^w = -w \log p(\boldsymbol{y}_i|\boldsymbol{\theta})$ as the loss gives the same solution as using the standard log likelihood together with $D = \frac{1}{w}\text{KLD}$. For $D = \frac{1}{w}\text{KLD}$, $D(q||p)$ has closed form if both q and p are (multivariate) normal densities. To show that this discrepancy is also available in closed form for $D = D_{AR}^{(\alpha)}$, we next give a version of Proposition B.1 (see Appendix B.5).

Theorem 6.2 (closed forms for $D = D_{AR}^{(\alpha)}$)**.** For $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}^q, \boldsymbol{\Sigma}_q)$ and $p(\boldsymbol{\theta}) =$

$\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}^p, \boldsymbol{\Sigma}_p)$ and

$$(\boldsymbol{\Sigma}^*)^{-1} = \alpha\boldsymbol{\Sigma}_q^{-1} + (1-\alpha)\boldsymbol{\Sigma}_p^{-1}; \quad \boldsymbol{\mu}^* = \boldsymbol{\Sigma}^* \left(\alpha\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\mu}^q + (1-\alpha)\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}^p\right)$$

it holds that for $\alpha \in (0, 1)$,

$$D_{AR}^{(\alpha)}\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta})\right) = \frac{1}{2\alpha(1-\alpha)}\Big\{ -\alpha\left[\boldsymbol{\mu}^{q'}\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\mu}^q + \ln|\boldsymbol{\Sigma}_q|\right] - (1-\alpha)\left[\boldsymbol{\mu}^{p'}\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\mu}^p + \ln|\boldsymbol{\Sigma}_p|\right] + \left[\boldsymbol{\mu}^{*'}(\boldsymbol{\Sigma}^*)^{-1}\boldsymbol{\mu}^* + \ln|\boldsymbol{\Sigma}^*|\right]\Big\} \tag{6.21}$$

Notice that computing this is of the same order as computing the KLD. In particular, one needs to perform a Cholesky decomposition of $\boldsymbol{\Sigma}_q$ and $\boldsymbol{\Sigma}_q$ for either choice of D.

With $D_l \in \{D_{AR}^{(\alpha)}, \frac{1}{w}\text{KLD}\}$ for $l = 1, 2, \ldots L$, the final form of the GVI objectives studied in this section are given as

$$O_{\text{DGP,GVI}}^{\beta,D^{1:L}}(\boldsymbol{\kappa}) = \sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^{\beta}(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] + \sum_{l=1}^{L} D_l(q(\boldsymbol{F}^l, \boldsymbol{U}^l)||p(\boldsymbol{F}^l, \boldsymbol{U}^l|\boldsymbol{Z}^{l-1})),$$
$$O_{\text{DGP,GVI}}^{\gamma,D^{1:L}}(\boldsymbol{\kappa}) = \sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^{\gamma}(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] + \sum_{l=1}^{L} D_l(q(\boldsymbol{F}^l, \boldsymbol{U}^l)||p(\boldsymbol{F}^l, \boldsymbol{U}^l|\boldsymbol{Z}^{l-1})).$$

As shown by Theorems 6.1 and 6.2, a number of relevant quantities in this objective will be available in closed form.

Does the layer-specific divergence define a valid divergence?

The attentive reader may pause here: while the extension to general losses is straightforward, the same cannot be said about the new regularizers. In particular, two important questions arise at this point:

(I) Will the divergence term simplify to $\sum_{l=1}^{L} D_l(q(\boldsymbol{U}^l)||p(\boldsymbol{U}^l|\boldsymbol{Z}^{l-1}))$ as in eq. (6.12)?

(II) Is $\sum_{l=1}^{L} D_l(q(\boldsymbol{F}^l, \boldsymbol{U}^l)||p(\boldsymbol{F}^l, \boldsymbol{U}^l|\boldsymbol{Z}^{l-1}))$ a valid divergence between the *full* (rather than layer-specific) prior π of eq. (6.13) and the *full* (rather than layer-specific) variational posterior q of eq. (6.9)?

To see that (I) can be answered positively, one simply needs to re-examine eq. (10)–(12) in Edwin V. Bonilla et al. (2019). In particular, note that for any divergence $D'(q||p)$ that can be written as $D'(q||p) = g(D(q||p))$ for some function

$g(x)$ such that $g(x) \geq 0$ and $g(x) = 0$ if and only if $x = 0$ and for some f-divergence $D^l(q||p) = \int_{\boldsymbol{F}^l,\boldsymbol{U}} q(\boldsymbol{F}^l, \boldsymbol{U}^l) f\left(\frac{q(\boldsymbol{F}^l,\boldsymbol{U}^l)}{p(\boldsymbol{F}^l,\boldsymbol{U}^l|\boldsymbol{Z}^{l-1})}\right) d(\boldsymbol{F}^l, \boldsymbol{U}^l)$, it holds that

$$\begin{aligned} D'(q(\boldsymbol{F}^l, \boldsymbol{U}^l)||p(\boldsymbol{F}^l, \boldsymbol{U}^l|\boldsymbol{Z}^{l-1})) &= g\left(\mathbb{E}_{q(\boldsymbol{F}^l,\boldsymbol{U}^l)}\left[f\left(\frac{q(\boldsymbol{F}^l, \boldsymbol{U}^l)}{p(\boldsymbol{F}^l, \boldsymbol{U}^l|\boldsymbol{Z}^{l-1})}\right)\right]\right) \\ &= g\left(\mathbb{E}_{p(\boldsymbol{F}^l|\boldsymbol{U}^l,\boldsymbol{Z}^{l-1})q(\boldsymbol{U}^l)}\left[f\left(\frac{q(\boldsymbol{F}^l, \boldsymbol{U}^l)}{p(\boldsymbol{F}^l, \boldsymbol{U}^l|\boldsymbol{Z}^{l-1})}\right)\right]\right) \\ &= g\left(\mathbb{E}_{q(\boldsymbol{U})}\left[f\left(\frac{q(\boldsymbol{U}^l)}{p(\boldsymbol{U}^l|\boldsymbol{Z}^{l-1})}\right)\right]\right) \\ &= D'(q(\boldsymbol{U}^l)||p(\boldsymbol{U}^l|\boldsymbol{Z}^{l-1})). \end{aligned} \tag{6.22}$$

This clearly holds for the special case of the $D_{AR}^{(\alpha)}$ with $g(x) = \frac{1}{\alpha(1-\alpha)}\log(x+1)$ and $f(x) = x^{1-\alpha}$. Therefore, we can simplify the objectives further to

$$\begin{aligned} O^{\beta,D^{1:L}}_{\text{DGP,GVI}}(\boldsymbol{\kappa}) &= \sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\beta(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] + \sum_{l=1}^{L} D_l(q(\boldsymbol{U}^l)||p(\boldsymbol{U}^l|\boldsymbol{Z}^{l-1})), \\ O^{\gamma,D^{1:L}}_{\text{DGP,GVI}}(\boldsymbol{\kappa}) &= \sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{f}_i^L)}\left[\mathcal{L}_p^\gamma(\boldsymbol{f}_i^L, \boldsymbol{y}_i)\right] + \sum_{l=1}^{L} D_l(q(\boldsymbol{U}^l)||p(\boldsymbol{U}^l|\boldsymbol{Z}^{l-1})). \end{aligned}$$

Does the layer-specific divergence simplify?

The answer to (II) is less obvious and relies on a technical Lemma in Appendix B.7.2.

Corollary 6.1. In the DGP construction (see for instance eq. (6.12)), replacing the sum of KLD-terms by

$$\sum_{l=1}^{L} D^l(q(\boldsymbol{U}^l)||p(\boldsymbol{U}^l))$$

defines a valid divergence between $q(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)$ and $p(\{\boldsymbol{F}^l\}_{l=1}^L, \{\boldsymbol{U}^l\}_{l=1}^L)$ so long as D^l is an f-divergence or a divergence obtained as a monotonic transform g of an f-divergence for all $l = 1, 2, \ldots L$.

Since conditions (i) and (ii) of this Theorem are easily verified for the DGP as long as $D^l \in \{D_{AR}^{(\alpha)}, \text{KLD}\}$ for all $l = 1, 2, \ldots, L$, the answer to (I) is also positive. For details, we refer to Appendix B.7.2

6.2.4 Experimental results

We make the comparisons as fair as possible by using the `gpflow` (Matthews et al., 2017) implementation of Salimbeni and Deisenroth (2017). Further, we use the same settings, meaning that all experiments use 20,000 iterations of the ADAM optimizer (Kingma and Ba, 2014) with a learning rate of 0.01 and default settings for all other hyperparameters. We perform inference for each of the UCI data sets (Lichman, 2013) after normalization using the RBF kernel with dimension-wise lengthscales, 100 inducing points, with batch sizes of $\min(1000, n)$ and latent function dimension $D^l = \min(D, 30)$, where we remind that $D^0 = D$ was the dimension of the observation space. As Salimbeni and Deisenroth (2017), we use 50 random splits with 90% training and 10% test data to assess predictive performance in terms of negative log likelihood (NLL) and root mean square error (RMSE). With this, we compare two inference schemes:

(1) The state of the art **standard VI** techniques of Salimbeni and Deisenroth (2017);

(2) A **GVI** variant of the same inference method which replaces the log score with the robust γ-divergence based scoring rule $\mathcal{L}_p^\gamma$; and the KLD with the $D_{AR}^{(\alpha)}$.

DGPs with γ-divergence losses

As noted earlier, $\mathcal{L}_p^\gamma$ has the distinct advantage over $\mathcal{L}_p^\beta$ that it can be stored fully in log form, which is of great practical utility in the context of numerically sensitive stochastic gradient approximation. For this reason, we only study the case of losses with the γ-divergence here—though we have found the results to be near-identical for the β-divergence so long as the gradients remained numerically stable.

For choosing a value of γ, we note that inferences are robust for $\gamma > 1$ and that $\mathcal{L}_p^\gamma$ recovers the log score as $\gamma \to 1$. At the same time, the scoring rule will grow increasingly happy to ignore virtually all of the data as $\gamma \to \infty$. Accordingly, one will typically want to pick

$$\gamma = 1 + \varepsilon$$

for a small $\varepsilon > 0$. Choosing γ in this way encodes the intuition that a good robust scoring rule will behave like the log score for all but the most extreme outliers. We thus pick $\varepsilon \in \{0.01, 0.05\}$. We note that hyperparameter optimization might appear to be the natural choice for picking γ, but will not perform well in practice: Rather

than producing robust inferences, this will select for a value of γ generally producing the smallest GVI objective values across $\mathcal{Q}$[3].

The results of our empirical comparison are depicted in Figure 6.3 and confirm our two main intuitions about robustness: Firstly, the robust scoring rule provides a significant performance improvement. Secondly, the smaller value of γ (which will be closer to the log score) generally outperforms the larger value of γ, though both choices are equally good in many data sets.[4] We believe that the performance gains of the robust scoring rule is due to large parts of the latent spaces being non-informative. This implies that it is beneficial to implicitly down-weight the influence of these non-informative parts of the latent space. It is clear that robust scoring rules do exactly that (see for instance Figure 6.1), which explains their superior performance in the DGP experiments. This intuition is further bolstered by the following observation: Generally, performance *improves* with a larger number of layers L under the robust score $\mathcal{L}_p^\gamma$, but *worsens* under the log score. In other words: The more dispersed the prior over the latent space (i.e., the DGP) becomes, the more inferential outcomes benefit from implicitly ignoring its non-informative (or indeed anti-informative!) regions. Next, we provide a small batch of additional results showing that as expected, modifying D is less beneficial for DGPs than it is for BNNs. Most likely, this is due to hyperparameter optimization for the kernels of the DGP: together with the fact that Gaussian Processes are far more informative priors than fully factorized normals, careful selection of the hyperparameters ensures that unlike in the BNN case, the prior is informative.

DGPs with $D_{AR}^{(\alpha)}$ regularizers

While we showed that DGPs allow for the variation of both losses and prior regularizers, we do not want to emphasize the flexibility afforded by varying D in the current chapter. For comparison however, we showcase what happens when the regularizer is varied jointly with the loss in Figure 6.4, which compares a number of different GVI posteriors for DGPs with $L = 3$ layers. The loss is either the robust loss $\mathcal{L}_p^\gamma$ for $\gamma \in \{1.01, 1.05\}$ (top 8 entries in each row) or the standard log score (bottom 4 entries in each row). We also compare $D = \frac{1}{w}\text{KLD}$ for $w = 2.0, 1.0, 0.5$ as

[3] In practice, this means that hyperparameter optimization pushes $\gamma \to 1$ or $\gamma \to \infty$, depending on the magnitudes of $\{p(\boldsymbol{y}_i|\boldsymbol{f}_i^L)\}_{i=1}^n$.

[4] We expect this second finding about γ to generalize to new settings so long as the inputs are normalized and the outputs are not high-dimensional, which would make $\gamma = 1.01$ a decent default choice in such scenarios.

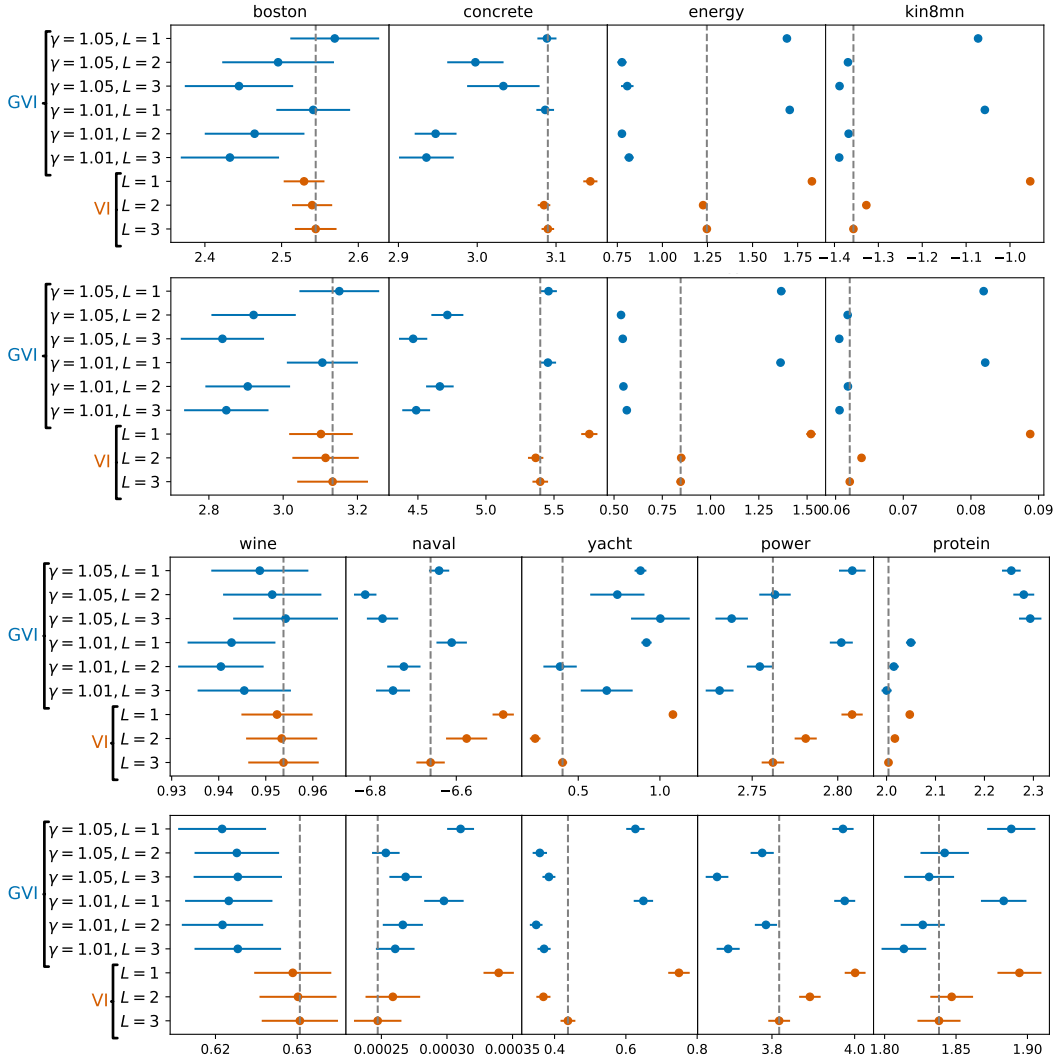


Figure 6.3: Comparing performance in DGPs with L layers for **DGP-GVI** with $\ell_n(\boldsymbol{\theta}, \boldsymbol{x}) = \sum_{i=1}^{n} \mathcal{L}_p^{\gamma}(\boldsymbol{\theta}, x_i)$ and **DGP-VI**. Benchmark performance is the DGP with three layers as in (Salimbeni and Deisenroth, 2017). **Top rows**: Negative test log likelihoods. **Bottom rows**: Test RMSE. The lower the better.

well as the composite layer-wise divergence

$$D(q\|\pi) = \sum_{l=1}^{3} D_l(q_l\|\pi_l), \quad D_1 = D_2 = \mathrm{KLD}, D_3 = D_{AR}^{(\alpha)} \text{ for } \alpha = 0.5.$$

Aligned with the intuition that the priors in DGPs are rather informative due to various hyperparameter optimization schemes, changing the prior regularizer from the KLD to the $D_{AR}^{(\alpha)}$ generally typically has either fairly little or even adverse impact.

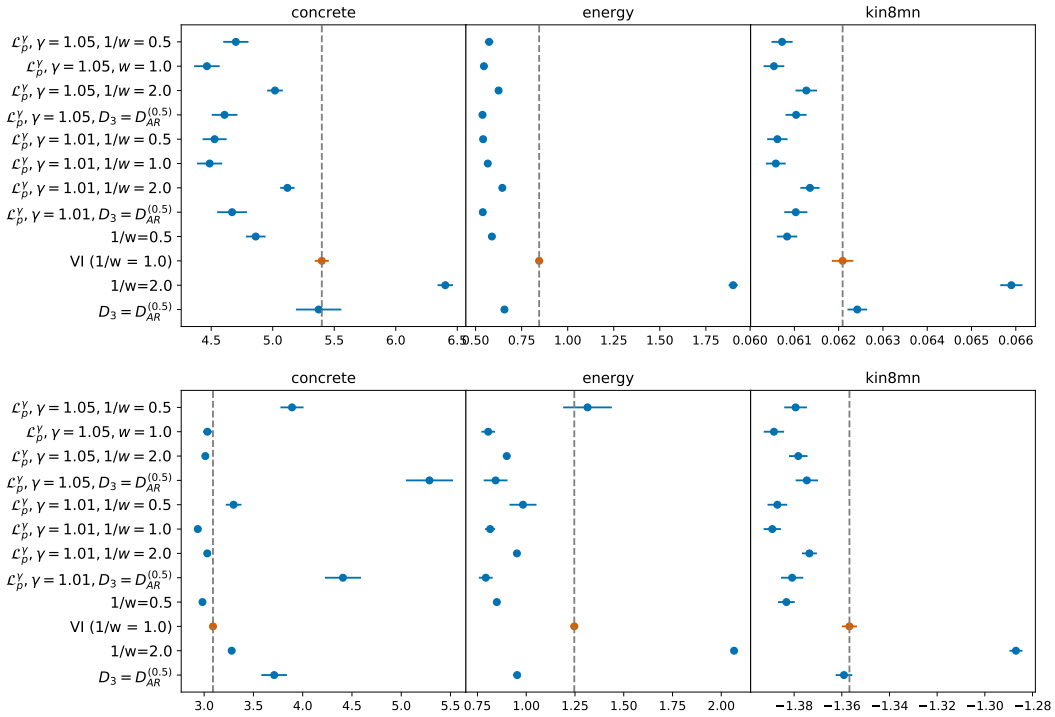


Figure 6.4: Best viewed in color. Top row depicts RMSE, bottom row the NLL across a range of data sets using DGPs with $L = 3$ layers. Dots correspond to means, whiskers the standard errors. The further to the left, the better the predictive performance.

Similarly, up- or down-weighting the KLD seems not to be beneficial across the board and will depend on the loss function. For the case of the log score however, we find a consistent improvement for down-weighting the KLD: Predictively, it improves the predictions on both metrics and across all data sets relative to standard VI. Similarly, up-weighting the KLD term is counterproductive under the log score and yields a performance deterioration across all data sets. This indicates that despite best efforts to the contrary in the form of empirical-Bayes style hyperparameter optimizations, DGPs are violating **(P)**; so that their predictive performance can be enhanced by ignoring more prior information.

Part III

Advances in Applications

Chapter 7

Robustness for on-line changepoint inference

Summary: In this chapter, we make two contributions to the body of work on Bayesian On-line Changepoint Detection (BOCPD). Firstly, we extend the framework to spatio-temporal problems and on-line model selection. Secondly, we use the Rule of Three (RoT) to robustify this important family of time series methods against outliers. We will first show how to modify the previous iterations of the algorithm to incorporate multiple models and an on-line model selection mechanism, and how this leads to richer posterior inferences—particularly in the context of spatio-temporal modelling problems. But by increasing the number of models the algorithm tracks and compares at the same time, we also amplify an existing problem of the base algorithm: a severe lack of robustness to model misspecification. Since time series data are especially hard to model correctly, this motivates the derivation of a robust version of BOCPD based on GVI with a β-divergence loss.

We will discuss the contributions of this chapter in two main sections. In Section 7.1, we will show how Bayesian On-line Changepoint Detection (BOCPD) can be extended to Bayesian On-line Changepoint Detection with Model Selection (BOCPDMS); and how BOCPDMS is particularly suitable for spatio-temporal inference problems. In Section 7.2, we will build on some of the insights into the lack of robustness in Bayesian On-line Changepoint methods that we gained from deriving BOCPDMS. In particular, we will show how BOCPD and BOCPDMS can suffer under model misspecification and in the presence of outliers; and how a GVI posterior

based on the β-divergence can rectify this issue.

7.1 Spatio-temporal Bayesian On-line Changepoint Detection with Model Selection

Real-world spatio-temporal processes are often poorly modelled by standard inference methods that assume stationarity in time and space. A variety of techniques have been developed for modelling non-stationarity in time via changepoints (CPs), ranging from methods for Gaussian Processes (GPs) (Garnett et al., 2009), the Lasso (Lin et al., 2017) or the Ising model (Fazayeli and Banerjee, 2016) over approaches using density ratio estimation (Liu et al., 2013) and kernel-based methods exploiting M-statistics (Li et al., 2015) to framing CP detection as time series clustering (Khaleghi and Ryabko, 2014). In contrast, CP inference allowing for non-stationarity in space (Herlands et al., 2016) has received comparatively little attention.

We offer the first on-line solution to this problem by modeling non-stationarity in both space and time. CPs are used to model non-stationarity in time, and the use of spatially structured Bayesian Vector Autoregressions (SSBVAR) circumvents the assumption of stationarity in space. We unify Adams and MacKay (2007) and Fearnhead and Liu (2007) into an inference procedure for **on-line prediction**, **model selection** and **CP detection**, see Fig. 7.1. Our construction exploits that both algorithms use Product Partition Models (Barry and Hartigan, 1993), which assume independence of parameters conditional on the CPs and independence of observations conditional on these parameters.

In essence, we extend the existing work on BOCPD by allowing for model uncertainty. This yields a new and powerful framework: It allows on-line inference using many existing and well-developed models simultaneously. For instance, Adams and MacKay (2007), Saatçi et al. (2010), and Fearnhead and Liu (2007) all develop different model families for BOCPD, but in their setting, one has to guess the 'best model' a priori. This is a severe issue in the on-line setting, where restarting the algorithm if the data-generating process changes drastically is not an option. We propose to solve this problem by including all relevant models in one single model universe, deciding internally and on-line which one fits the data best at which time. By letting the data speak for itself, we can also alleviate potential misspecification problems. This is practically relevant as it allows the user to deploy the algorithm without first having to extensively study idiosyncrasies of the data source. The second major contribution is an application of this very general property to complex non-stationary multivariate (spatio)temporal patterns via Spatially

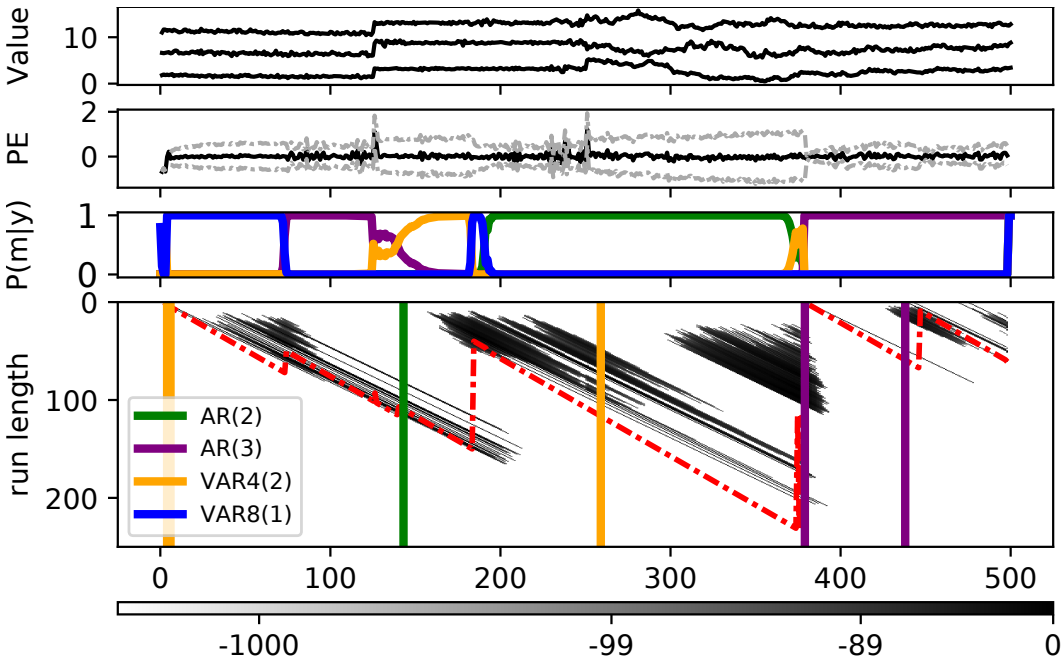


Figure 7.1: *Bayesian On-line Changepoint Detection with Model Selection (*BOCPDMS*)*: **Panel 1:** Artificial data across times $1-500$ for a regular spatial grid with 4- and 8-neighbourhood dependency structure as in Fig. 7.2, where Model universe $\mathcal{M}$ uses AR and Spatially Structured BVAR models with 4-neighbourhood and lag lengths $1-3$, see Fig 7.2. **Panel 2:** prediction error (black) and variance (gray). **Panel 3:** Model posteriors $p(m_t|x_{1:t})$. **Panel 4:** log run-length distribution (grayscale), its maximum (red) and MAP segmentation of CPs and models in corresponding colors.

Structured Bayesian Vector Autoregression (SSBVAR) models. This application is of great practical importance: For the very first time, BOCPD becomes amenable to inference in complicated multivariate spatial structures. We note that our work is also first to model multivariate data jointly within BOCPD. In other work like Saatçi et al. (2010), each time series is modeled as independently except for the common changepoints. In a sense, our extension can be seen as modified on-line version of Xuan and Murphy (2007). In their method, inference is off-line, the model universe $\mathcal{M}$ is built during execution and multivariate dependencies are restricted to decomposable graph. In contrast, our procedure specifies $\mathcal{M}$ before execution, but runs on-line and does not restrict dependencies. The closest competing on-line procedure in the literature thus far is the work of Saatçi et al. (2010), which develops GP CP models for BOCPD. Though our results suggest that parametric models may be preferable to GP models, the latter can still be integrated into our method as

elements of the model universe $\mathcal{M}$ without any further modifications.

In summary, we make three contributions: Firstly, we substantially augment the existing work on BOCPD by allowing for on-line model uncertainty. Unlike previous extensions of the algorithm (e.g. Adams and MacKay, 2007; Saatçi et al., 2010), this avoids having to guess a single best model family a priori. Secondly, we introduce SSBVARs as the first class of models for multivariate inference within BOCPD. Thirdly, we demonstrate that using a collection of parametric models can outperform nonparametric GP models in terms of prediction, CP detection and computational efficiency.

The structure of this chapter is as follows: Section 7.1.1 generalizes the BOCPD algorithm of Adams and MacKay (2007), henceforth AM, by integrating it with the approach of Fearnhead and Liu (2007), henceforth FL. In so doing, we arrive at BOCPD with Model Selection, henceforth BOCPDMS. Section 7.1.2 proposes VAR models for non-stationary processes within the BOCPD framework. This motivates populating the model universe $\mathcal{M}$ with spatially structured BVAR (SSBVAR) models. Sections 7.1.3–7.1.4 address computational aspects. Section 7.1.5 demonstrates the algorithm's advantages on real world data.

7.1.1 BOCPDMS

Let $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$ be a data stream with an unknown number of CPs. Focusing on univariate data, FL and AM tackled inference by tracking the posterior distribution for the most recent CP. While FL allow the data to be described by different model classes between CPs, AM only allow for a single model class. However, AM perform one-step-ahead predictions, whereas FL do not. Instead, they propose a Maximum A Posteriori (MAP) segmentation for CPs and models. In the remainder of this section, we unify both inference approaches. We call the resulting algorithm BOCPD with model selection (BOCPDMS), as it performs prediction, MAP segmentation and model selection all at once and on-line.

Run-length & model universe

The *run-length* r_t at time t is defined as the time since the most recent CP at time t, so $r_t = 0$ corresponds to a CP at time t. Suppose that data between successive CPs can be described by Bayesian models collected in the *model universe* $\mathcal{M}$. For the process $\{\boldsymbol{x}_t\}$ on $\mathbb{R}^S$, a model $m \in \mathcal{M}$ with finite memory of length $L \in \mathbb{N}_0$ consists of an observation density $f_m(\boldsymbol{x}_t = x_t|\boldsymbol{\theta}_m, x_{(t-L):(t-1)})$ on $\mathbb{R}^S$ and a parameter prior $\pi_m(\boldsymbol{\theta}_m)$ on $\boldsymbol{\Theta}_m$ depending on hyperparameters $\boldsymbol{\nu}_m$. The notion of $\mathcal{M}$ is due to

BOCPD with Model Selection (BOCPDMS)

Input at time 0: model universe $\mathcal{M}$; hazard h; prior q
Input at time t: next observation x_t
Output at time t: $\widehat{x}_{(t+1):(t+h_{\max})}$, S_t, $p(m_t|x_{1:t})$
5pt
for next observation x_t at time t **do**
// STEP I: Compute model-specific quantities
for $m \in \mathcal{M}$ **do**
if $t-1 =$ lag_length(m) **then**
[I.A] Initialize $p(x_{1:t}, r_t = 0, m_t = m)$ with prior
else if $t-1 >$ lag_length(m) **then**
[I.B.1] Update $p(x_{1:t}, r_t, m_t = m)$ via (7.5a), (7.5b)
[I.B.2] Prune model-specific run-length distribution
[I.B.3] Perform hyperparameter inference via (7.13)
// STEP II: Aggregate over models
if $t >=$ min(lag_length(m)) **then**
[II.1] Obtain joint distribution over $\mathcal{M}$ via (7.6a)–(7.6f)
[II.2] Compute (7.7)–(7.9)
[II.3] **Output:** $\widehat{x}_{(t+1):(t+h_{\max})}, S_t, p(m_t|x_{1:t})$

FL and allows for model uncertainty amongst models developed for BOCPD. For instance, $m \in \mathcal{M}$ could be a GP Saatçi et al. (2010), a time-deterministic regression (Fearnhead, 2005) or a mixture distribution (Caron et al., 2012).

Probabilistic formulation & recursions

Denote by m_t the model describing $x_{(t-r_t):t}$, i.e. the data since the last CP. Given hazard function $h : \mathbb{N} \to [0, 1]$, and model prior $q : \mathcal{M} \to [0, 1]$, the prior beliefs are

$$H(r_t|r_{t-1}) = \begin{cases} 1 - h(r_{t-1} + 1) & \text{if } r_t = r_{t-1} + 1 \\ h(r_{t-1} + 1) & \text{if } r_t = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7.1a}$$

$$q(m_t|m_{t-1}, r_t) = \begin{cases} \mathbf{1}_{m_{t-1}}(m_t) & \text{if } r_t = r_{t-1} + 1 \\ q(m_t) & \text{if } r_t = 0. \end{cases} \tag{7.1b}$$

Eq. (7.1b) implies that the model at time t will be equal to the model at time $t-1$ unless a CP occured at t, in which case the next model m_t will be a random draw from q. At time t, the algorithm requires for all possible models m and run-lengths

r_t the computation of the posterior predictives

$$f_m(x_t|x_{1:(t-1)}, r_t) = \int_{\Theta_m} f_m(x_t|\boldsymbol{\theta}_m)\pi_m(\boldsymbol{\theta}_m|x_{(t-L-r_t):(t-1)})d\boldsymbol{\theta}_m. \tag{7.2}$$

To make the evaluation of this integral efficient, one can use conjugate models (Xuan and Murphy, 2007) or approximations (Turner et al., 2013; Niekum et al., 2014). If the integral can in fact be computed efficiently, then the following recursion will be efficient, too:

$$\begin{aligned} p(x_{1:t}, r_t, m_t) = \sum_{m_{t-1}} \sum_{r_{t-1}} \Big\{ & f_{m_t}(x_t|x_{1:(t-1)}, r_t) q(m_t|x_{1:(t-1)}, r_t, m_{t-1}) \times \\ & p(r_t|r_{t-1}) H(x_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}. \end{aligned} \tag{7.3}$$

The recursion in AM is the special case for $|\mathcal{M}| = 1$. For $|\mathcal{M}| > 1$, $q(m_t|m_{t-1}, r_t, x_{1:(t-1)})$ arises as a new term, which for $\mathbf{1}_a$ as the indicator function of a is given by

$$q(m_t|m_{t-1}, r_t, x_{1:(t-1)}) = \begin{cases} 1_{m_{t-1}}(m_t) q(m_{t-1}|x_{1:(t-1)}, r_{t-1}) & \text{if } r_t = r_{t-1} + 1 \\ q(m_t) & \text{if } r_t = 0. \end{cases} \tag{7.4}$$

Next, define the *growth-* and *changepoint probabilities* as

$$\begin{aligned} p(x_{1:t}, r_t = r_{t-1} + 1, m_t) = & f_{m_t}(x_t|x_{1:(t-1)}, r_t) p(x_{1:(t-1)}, r_{t-1}, m_{t-1}) \times \\ & (1 - h(r_t)) q(m_{t-1}|x_{1:(t-1)}, r_t), \end{aligned} \tag{7.5a}$$

$$\begin{aligned} p(x_{1:t}, r_t = 0, m_t) = & f_{m_t}(x_t|x_{1:(t-1)}, r_t) q(m_t) \times \\ & \sum_{m_{t-1}} \sum_{r_{t-1}} \Big\{ h(r_{t-1} + 1) p(x_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\} \end{aligned} \tag{7.5b}$$

The evidence can then be calculated via (7.6a), which in turn allows calculating the joint model-and-run-length distribution (7.6b), the model posterior (7.6c), as well as the model-specific (7.6d) and global (7.6e) run-length distributions:

$$p(x_{1:t}) = \textstyle\sum_{m_t} \sum_{r_t} p(x_{1:t}, m_t, r_t) \tag{7.6a}$$

$$p(r_t, m_t|x_{1:t}) = p(x_{1:t}, r_t, m_t)/p(x_{1:t}) \tag{7.6b}$$

$$p(m_t|x_{1:t}) = \textstyle\sum_{r_t} p(r_t, m_t|x_{1:t}) \tag{7.6c}$$

$$p(r_t|m_t, x_{1:t}) = p(r_t, m_t|x_{1:t})/p(m_t|x_{1:t}) \tag{7.6d}$$

$$p(r_t|x_{1:t}) = \textstyle\sum_{m_t} p(r_t, m_t|x_{1:t}) \tag{7.6e}$$

$$q(m_{t-1}|x_{1:(t-1)}, r_{t-1}) = \frac{p(m_{t-1}, r_{t-1}|x_{1:(t-1)})}{p(r_{t-1}|x_{1:(t-1)})}. \tag{7.6f}$$

Here, (7.6f) is the conditional model posterior from (7.4). Note that (7.6e) is arrived at directly in FL and used for on-line MAP segmentation. By framing our derivations in the run-length framework of AM, we additionally obtain (7.4)–(7.6d), which enables us to additionally perform both on-line prediction and model selection at the same computational cost.

On-line algorithm outputs

Prediction: Recursive h-step-ahead forecasting uses (7.6b):

$$p(\boldsymbol{x}_{t+h}|x_{1:t}) = \sum_{r_t,m_t} \Big\{ p(\boldsymbol{x}_{t+h}|x_{1:t}, \widehat{x}_t^h, r_t, m_t) p(r_t, m_t|x_{1:t}) \Big\}, \tag{7.7}$$

where $\widehat{x}_t^h = \emptyset$ if $h = 1$ and $\widehat{x}_t^h = \widehat{x}_{(t+1):(t+h-1)}$ otherwise, with $\widehat{x}_{t+h} = \mathbb{E}(\boldsymbol{x}_{t+h}|x_{1:t}, \widehat{x}_t^h)$ the recursive forecast.

Tracking the model posterior/Bayes Factors: Another one of the novel capabilites of the algorithm is on-line monitoring of the model posterior via Eq. (7.6c). This is attractive when structural changes in the data happen slowly and are not captured well by CPs. In this case, $p(m_t|x_{1:t})$ can be used to identify periods of change, see Fig. 7.6. For pairwise comparisons, Bayes Factors can be monitored:

$$\mathrm{BF}(m_1, m_2)_t = \frac{p(m_t = m_1|x_{1:t}) \cdot q(m_2)}{p(m_t = m_2|x_{1:t}) \cdot q(m_1)}. \tag{7.8}$$

Maximum A Posteriori (MAP) segmentation: For MAP_t denoting the density of the MAP-estimate of models and CPs *before* t, and for $\mathrm{MAP}_0 = 1$, FL's recursive estimator is given by

$$\mathrm{MAP}_t = \max_{r,m} \Big\{ p(x_{1:t}, r_t = r, m_t = m) \mathrm{MAP}_{t-r-1} \Big\}. \tag{7.9}$$

For r_t^*, m_t^* maximizers for time t, the MAP segmentation is $S_t = S_{t-r_t^*-1} \cup \{(t - r_t^*, m_t^*)\}$, $S_0 = \emptyset$, where $(t', m_{t'}) \in S_t$ means a CP at $t' \leq t$, with $m_{t'} \in \mathcal{M}$ the model for $x_{t':t}$.

7.1.2 Building a spatio-temporal model universe

The last section derived BOCPDMS for arbitrary data streams $\{\boldsymbol{x}_t\}$. Next, we propose models for $\mathcal{M}$ if $\{\boldsymbol{x}_t\}$ can be mapped into a some space $\mathbb{S}$ with spatial structure. Let $\mathcal{S}$ with $|\mathcal{S}| = S$ be a set of spatial locations in $\mathbb{S}$ with measurements $\boldsymbol{x}_t = (Y_{t,1}, Y_{t,2}, \ldots, Y_{t,S})^T$ recorded at times $t = 1, 2, \ldots$

Bayesian VAR (BVAR)

Inference on $\{\boldsymbol{x}_t\}$ can be drawn using conjugate Bayesian Vector Autoregressions (BVAR) with lag length L and E additional variables $\boldsymbol{Z}_t$ as elements of model universe $\mathcal{M}$:

$$\sigma^2 \sim \text{InverseGamma}(a, b) \tag{7.10a}$$

$$\boldsymbol{\varepsilon}_t|\sigma^2 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \boldsymbol{\Omega}) \tag{7.10b}$$

$$\boldsymbol{c}|\sigma^2 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \boldsymbol{V}_c) \tag{7.10c}$$

$$\boldsymbol{x}_t = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{Z}_t + \textstyle\sum_{l=1}^{L} \boldsymbol{A}_l \boldsymbol{x}_{t-l} + \boldsymbol{\varepsilon}_t. \tag{7.10d}$$

Here, $\boldsymbol{A}_l, \boldsymbol{B}$ are $S \times S$, $S \times E$ matrices, and $\boldsymbol{c} = (\boldsymbol{\alpha}, \text{vec}(\boldsymbol{B}), \text{vec}(\boldsymbol{A}_1), \text{vec}(\boldsymbol{A}_2), \dots \text{vec}(\boldsymbol{A}_L))^T$ is a vector of $S \cdot (LS + 1 + E)$ model parameters. Scalars $a, b > 0$, matrix $\boldsymbol{V}_c$, and diagonal matrix $\boldsymbol{\Omega}$ are hyperparameters.

Approximating processes using VARs

Modelling $\{\boldsymbol{x}_t\}$ as VAR is attractive, as many complex non-linear processes have VAR representations, including Hidden Markov Models (HMMs), time-stationary GPs as well as multivariate Generalized Autoregressive Conditionally Heteroskedastic (GARCH) and fractionally integrated Vector Autoregressive Moving Average (VARMA) processes (Inoue and Kasahara, 2006; Inoue et al., 2018). Performance guarantees for VAR approximations to such processes are derived using Baxter's Inequalitiy with multivariate versions of results in Hannan and Kavalieris (1986). We formally state this in a representation theorem which heavily draws on the findings in Meyer and Kreiss (2015). For the formal statement, we need an Assumption to hold for the spectral density matrix of the process.

Denoting the spectrum of a matrix $\boldsymbol{B}$ (i.e., the set of its eigenvalues) by $\sigma(\boldsymbol{B})$, the relevant condition is a restatement of the relevant part in condition **A** of Meyer & Kreiss (2015):

Assumption 7.1. Let $\boldsymbol{W}$ be the spectral density matrix of the purely non-deterministic stochastic process $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$ satisfying the conditions of Theorem 1. We assume that the spectral density matrix is bounded, i.e. there is a constant $c > 0$ so that

$$\min\left(\sigma(\boldsymbol{W}(\lambda))\right) \geq c \tag{7.11}$$

for all frequencies $\lambda \in (-\pi, \pi]$, i.e. the eigenvalues of the spectral density matrix are uniformly bounded away from zero.

Theorem 7.1. Let $\{\boldsymbol{x}_t\}$ be a time-stationary spatio-temporal process with spectral density satisfying Assumption 7.1, $||\cdot||$ a matrix norm, $\mathbb{E}(\boldsymbol{x}_t) = 0$, $\mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^T) < \infty$, $\sum_{h=-\infty}^{\infty}(1+|h|)^3||\mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}'_{t+h}]|| < \infty$. Then (1)–(3) hold.

(1) $\boldsymbol{x}_t = \sum_{i=1}^{\infty} \boldsymbol{A}_i \boldsymbol{x}_{t-i} + \boldsymbol{\varepsilon}_t$ for matrices $\{\boldsymbol{A}_l\}_{l\in\mathbb{N}}$ and $\mathbb{E}(\varepsilon_t) = 0$, $\mathbb{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t) = \boldsymbol{D}$, $\boldsymbol{D}$ diagonal.

(2) For $\boldsymbol{x}_t = \sum_{l=1}^{L} \boldsymbol{A}_l^L \boldsymbol{x}_{t-l} + e_t$ with $\{\boldsymbol{A}_l^L\}_{l=1}^L$ the best linear projection coefficients, $\exists L_0 : \forall L > L_0$, $\sum_{l=1}^{L}(1+|l|)^3||\boldsymbol{A}_l^L - \boldsymbol{A}_l|| \leq C \cdot \sum_{l=L+1}^{\infty}(1+|l|)^3||\boldsymbol{A}_l||$ with C constant.

(3) Using T observations with $L = \mathcal{O}([T/\ln(T)]^{1/6})$ to estimate $\boldsymbol{A}_l^L$ as MAP $\widehat{\boldsymbol{A}}_l^L$ of (7.10a)–(7.10d), it holds that $L(T)^2 \sum_{l=1}^{L(T)} ||\widehat{\boldsymbol{A}}_l^{L(T)} - \boldsymbol{A}_l^{L(T)}|| \xrightarrow{P} 0$ as $T \to \infty$.

Proof. Part (1) is shown in Inoue et al. (2018), part (2) in Lemma 3.1 of Meyer and Kreiss (2015). Part (3) follows by their Remark 3.3 if we can prove that the MAP estimator $\hat{\boldsymbol{c}}(L(T))$ of $\boldsymbol{c}$ equals its Yule-Walker estimator (YWE) as $T \to \infty$. Let $\boldsymbol{B} = \boldsymbol{0}$, $\boldsymbol{\alpha} = \boldsymbol{0}$ and note that YWE equals OLS as $T \to \infty$. With $\boldsymbol{V}_{1:T}$ the regressor matrix of $\boldsymbol{x}_{t-L(T):t}$, $\hat{\boldsymbol{c}}(L(T)) = (\boldsymbol{V}'_{1:T}\boldsymbol{V}_{1:T} + \boldsymbol{V}_c^{-1})^{-1}(\boldsymbol{x}'_{1:T}\boldsymbol{x}_{1:T})$. Then, part (3) holds since under the regularity conditions of the Theorem, OLS $\xrightarrow{P} \mathbb{E}(\boldsymbol{V}'_{1:T}\boldsymbol{V}_{1:T})^{-1}\mathbb{E}(\boldsymbol{V}'_{1:T}\boldsymbol{x}_{1:T})$ and

$$\begin{aligned}
\hat{\boldsymbol{c}}(L(T)) &= (\boldsymbol{V}'_{1:T}\boldsymbol{V}_{1:T} + \boldsymbol{V}_c^{-1})^{-1}(\boldsymbol{V}'_{1:T}\boldsymbol{x}_{1:T}) \\
&= (\frac{1}{T}\boldsymbol{V}'_{1:T}\boldsymbol{V}_{1:T} + \frac{1}{T}\boldsymbol{V}_c^{-1})^{-1}\frac{1}{T}(\boldsymbol{V}'_{1:T}\boldsymbol{x}_{1:T}) \\
&\xrightarrow{P} \mathbb{E}(\boldsymbol{V}'_{1:T}\boldsymbol{V}_{1:T})^{-1}\mathbb{E}(\boldsymbol{V}'_{1:T}\boldsymbol{x}_{1:T}).,
\end{aligned}$$ □

where we have denoted convergence in probability by $\xrightarrow{P}$.

Note that in the above result, assuming $\mathbb{E}(\boldsymbol{x}_t) = \boldsymbol{0}$ is without loss of generality: If $\mathbb{E}(\boldsymbol{x}_t) = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{Z}_t$, define $\boldsymbol{x}_t^* = \boldsymbol{x}_t - (\boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{Z}_t)$ and apply the theorem to $\{\boldsymbol{x}_t^*\}$. Moreover, the results do *not* require stationarity in space. Lastly, part (3) suggests a principled way of picking lag lengths $\mathcal{L}$ for BVAR models based on functions $L(T) = C \cdot (T/\ln(T))^{1/6}$, with C a constant: If between T_1 and T_2 observations are expected between CPs, $\mathcal{L} = \{L \in \mathbb{N} : L(T_1) \leq L \leq L(T_2)\}$. In our experiments, we employ this strategy using the heuristic $T_1 = 1, T_2 = T$.

Modeling spatial dependence

While Thm. 7.1 motivates approximating spatio-temporal processes between CPs with (7.10a)–(7.10d), the matrices $\{\boldsymbol{A}_l^L\}_{l=1}^L$ have $S(LS + 1 + E)$ parameters. This

increases model complexity and ignores spatial information. We remedy both issues through neighbourhood systems on $\mathcal{S}$.

Definition 7.1 (**Neighbourhood system**). For a set of locations $\mathcal{S}$ with the sets $N_i(s) \subseteq \mathcal{S}$ as the i-th neighbourhoods of s for $0 \leq i \leq n$ and all $s \in \mathcal{S}$, let $N_i(s) \cap N_j(s) = \emptyset$, $s' \in N_i(s) \iff s \in N_i(s')$ and $N_0(s) = \{s\}$. Then, the corresponding neighbourhood system is $N(\mathcal{S}) = \{\{N_i(s)\}_{i=1}^n : s \in \mathcal{S}, 0 \leq i \leq n\}$.

In the remainder of the section, smaller indices i imply that the neighbourhoods $N_i(s)$ are closer to s. For a BVAR model of lag length L, the decay of spatial dependence is encapsulated through $\xi : \{1, \ldots, L\} \to \{0, \ldots, n\}$. In particular, only $s' \in N_i(s)$ with $i \leq \xi(l)$ are modeled as affecting s after l time periods.

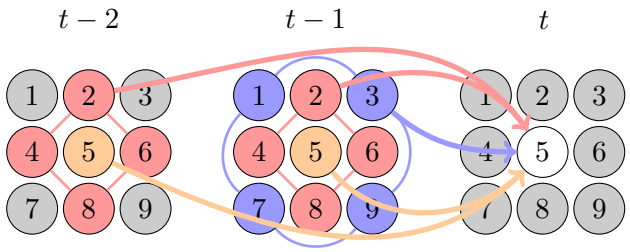


Figure 7.2: *SSBVAR modeling:* Suppose that on a regular grid of size 9, $Y_{t,5}$ depends on the past two realizations of itself and its 4- neighbourhood, and the last realization of its 8-neighbourhood. This is an SSBVAR on $\mathcal{S} = \{1, \ldots, 9\}$ with $L = 2$, $N_0(5) = \{5\}$, $N_1(5) = \{2, 4, 6, 8\}$, $N_2(5) = \{1, 3, 7, 9\}$ and function ξ with $\xi(1) = 2, \xi(2) = 1$.

Spatializing BVAR

In principle, given $N(\mathcal{S})$, sparsification of the BVAR model (7.10a)–(7.10d) is possible in two ways: As restriction on the *contemporaneous* dependence via the covariance matrix of the error term $\boldsymbol{\varepsilon}_t$, or as restriction on the *conditional* dependence via the coefficient matrices $\{\boldsymbol{A}_l\}_{l=1}^L$. We choose the latter for three reasons: Firstly, linear effects have more interesting interpretations than error covariances. Secondly, using $\{\boldsymbol{A}_l\}_{l=1}^L$ to encode spatial dependency allows us to work with arbitrary neighbourhoods. In contrast, modelling dependent errors under conjugacy limits dependencies to decomposable graphs (Xuan and Murphy, 2007). Since not even a regular grid is decomposable, this is problematic for spatial data. Thirdly, modelling errors as contemporaneous is attractive for low-frequency data where the resolution of temporal effects is coarse, but the situation reverses for high-frequency data. Since the algorithm runs on-line, we expect $\{\boldsymbol{x}_t\}$ to be observed with high frequency.

Definition 7.2 (Spatially structured BVAR (SSBVAR)). For process $\{\boldsymbol{x}_t\}$ on $\mathcal{S}$ and $(L, N(\mathcal{S}), \xi(\cdot))$, define the matrices $\{\widetilde{\boldsymbol{A}}_l\}_{l=1}^L$ by imposing that $[\widetilde{\boldsymbol{A}}_l]_{(s,s')} = 0 \iff s' \notin N_i(s)$ for any $i \leq \xi(l)$. Let $\widetilde{\boldsymbol{A}}_l^{\neq 0}$ be the vector of non-zero entries in $\widetilde{\boldsymbol{A}}_l$ and $\widetilde{\boldsymbol{c}} = (\boldsymbol{\alpha}, \text{vec}(\boldsymbol{B}), \widetilde{\boldsymbol{A}}_1^{\neq 0}, \widetilde{\boldsymbol{A}}_2^{\neq 0}, \dots \widetilde{\boldsymbol{A}}_L^{\neq 0})^T$. The SSBVAR model on $\{\boldsymbol{x}_t\}$ induced by $(L, N(\mathcal{S}), \xi(\cdot))$ is obtained by combining (7.10a)–(7.10b) with

$$\widetilde{\boldsymbol{c}}|\sigma^2 \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \cdot \boldsymbol{V}_{\widetilde{c}}) \tag{7.11a}$$

$$\boldsymbol{x}_t = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{Z}_t + \textstyle\sum_{l=1}^L \widetilde{\boldsymbol{A}}_l \boldsymbol{x}_{t-l} + \boldsymbol{\varepsilon}_t. \tag{7.11b}$$

Fig. 7.2 illustrates this idea. Further sparsification is possible by modelling neighbourhoods jointly, i.e. $[\widetilde{\boldsymbol{A}}_l]_{(s,s')} = a_i(s), \forall s' \in N_i(s)$, reducing the number of parameters to $S \cdot \sum_{l=1}^L \xi(l)$. If one imposes $a_i(s) = a_i(s') = \cdots = a_i$, this number drops to $\sum_{l=1}^L \xi(l)$.

Building SSBVARs: choosing $L, N(\mathcal{S}), \xi(\cdot)$

For the choice of lag lengths L, part *(3)* of Thm. 7.1 suggests $L \in \{L' \in \mathbb{N} : L(T_1) \leq L' \leq L(T_2)\}$ for $L(T) = \eta \cdot [T/\ln(T)]^{1/6}$ for some $\eta > 0$ if one expects T_1 to T_2 observations between CPs. For any data stream $\{\boldsymbol{x}_t\}$ on a space $\mathbb{S}$, there are different ways of constructing neighbourhood structures $N(\mathcal{S})$. For example, when analysing pollutants in London's air in section 7.1.5, $N(\mathcal{S})$ could be constructed from Euclidean or Road distances. By filling $\mathcal{M}$ with SSBVARs constructed using competing versions of $N(\mathcal{S})$, BOCPDMS provides a way of dealing with such uncertainty about spatial relations. In fact, it can dynamically discern changing spatial relationships on $\mathbb{S}$. Lastly, $\xi(\cdot)$ should usually be decreasing to reflect that measurements affect each other less when further apart.

7.1.3 Hyperparameter optimization

Hyperparameter inference on $\boldsymbol{\nu}_m$ can be addressed either by introducing an additional hierarchical layer (Wilson et al., 2010) or using and empirical Bayes procedure. The latter is obtained by maximizing the model-specific evidence

$$\log p(x_{1:T}|\boldsymbol{\nu}_m) = \sum_{t=1}^{T} \log p(x_t|\boldsymbol{\nu}_m, x_{1:(t-1)}). \tag{7.12}$$

Computation of the righthand side requires evaluating the gradients $\nabla_{\boldsymbol{\nu}_m} p(x_{1:t}, r_t|\boldsymbol{\nu}_m)$, which are obtained efficiently and recursively (Turner et al., 2009). Saatçi et al. (2010) use $x_{1:T'}$ as a test set, and run BOCPD K times to find $\widehat{\boldsymbol{\nu}}_m = \arg\max_{\boldsymbol{\nu}_m} \{p(x_{1:T'}|\boldsymbol{\nu}_m)\}$.

Most other on-line GP approaches also require substantial recomputations for hyperparameter learning (e.g., Ranganathan et al., 2011). In contrast, Caron et al. (2012) propose on-line gradient descent updates via

$$\boldsymbol{\nu}_{m,t+1} = \boldsymbol{\nu}_{m,t} + \alpha_t \nabla_{\boldsymbol{\nu}_{m,t}} \log p(x_{t+1}|x_{1:t}, \boldsymbol{\nu}_{m_{1:t}}). \tag{7.13}$$

The latter is preferable for two reasons: Firstly, inference and empirical-Bayes adjustmend of the priors are executed simultaneously (rather than sequentially) and thus enable cold-starts of BOCPDMS. Secondly, neither the on-line nature nor the computational complexity of BOCPDMS is affected.

7.1.4 Computation & Complexity

While tracking $|\mathcal{M}|$ models, BOCPDMS has linear time complexity. Step 1 in the pseudocode is the bottleneck, but looping over $\mathcal{M}$ can be parallelized: With N threads, it executes in $\mathcal{O}\left(\lceil|\mathcal{M}|/N\rceil \cdot \max_{M\in\mathcal{M}} \text{CmpTime}(M)\right)$, where $\text{CmpTime}(M)$ denotes the computation time for model M. Step 2 takes $\mathcal{O}(|R(t)||\mathcal{M}|)$, for $R(t)$ the set of all run-lengths at time t.

Pruning the run-length distribution

In a naive implementation, all run-lengths are retained and $R(t) = \{1, 2, \ldots, t\}$. This implies execution time of order $\mathcal{O}(t)$ for processing x_t, but can be made time-constant by pruning: If one discards run-lengths whose posterior probability is $\leq 1/R_{\max}$ or only keeps the $R_{\max}$ most probable ones, $|R(t)| \leq R_{\max}$ (Adams and MacKay, 2007). A third way is Stratified Rejection Control (SRC) (Fearnhead and Liu, 2007), which Caron et al. (2012) found to perform as well as the other approaches. In our experiments, we prune by keeping the $R_{\max}$ most probable model-specific run-lengths $p(r_t|m_t, x_{1:t})$ for each model.

BVAR updates

With $\boldsymbol{V}_{1:T}$ the regressor matrix of $\boldsymbol{x}_{t-L(T):t}$ as before, the bottleneck when updating a BVAR model in $\mathcal{M}$ is step I.B.1 in the pseudocode of BOCPDMS, when updating the MAP estimate $\boldsymbol{c}(r,t) = \boldsymbol{F}(r,t)\boldsymbol{W}(r,t)$ of the coefficient vector, where $\boldsymbol{F}(r,t) = (\boldsymbol{V}'_{(t-r):t}\boldsymbol{V}_{(t-r):t} + \boldsymbol{V}_{\tilde{c}})^{-1}$ and $\boldsymbol{W}(r,t) = \boldsymbol{V}'_{(t-r):t}\boldsymbol{x}_{(t-r):t}$ for all $r \in R(t)$. Since $\boldsymbol{W}(r,t) = \boldsymbol{W}(r-1,t-1) + \boldsymbol{V}'_t\boldsymbol{x}_t$, updates are $\mathcal{O}(kS)$. $\boldsymbol{F}(r-1,t-1)$ can be updated to $\boldsymbol{F}(r,t)$ using rank-k updates to its QR-decomposition in $\mathcal{O}(k^3)$ or using Woodbury's formula, in $\mathcal{O}(S^3)$, implying an overall complexity of $\mathcal{O}(|R(t)|\min\{k^3, S^3\})$ at time t.

Table 7.1: Computation time in seconds per model and per parameter in the space $\Theta = \cup_{m \in \mathcal{M}} \Theta_m$

	Nile	
	Time/$\lvert\mathcal{M}\rvert$	Time/$\lvert\Theta\rvert$
ARGPCP	42.2	21.0
BVAR	**4.03**	**0.35**
	Snowfall	
	Time/$\lvert\mathcal{M}\rvert$	Time/$\lvert\Theta\rvert$
ARGPCP	284	142
BVAR	**157**	**4.25**
	Bee	
	Time/$\lvert\mathcal{M}\rvert$	Time/$\lvert\Theta\rvert$
ARGPCP	164	23.4
BVAR	**97.3**	**0.04**
	30 Portfolios	
	Time/$\lvert\mathcal{M}\rvert$	Time/$\lvert\Theta\rvert$
ARGPCP	**12077**	403
BVAR	34183	**1.48**

Comparison with GP-based approaches

Define $k_{\max}$ as the largest number of regressors of any BVAR model in $\mathcal{M}$. From the previous paragraphs, it follows that if all models in $\mathcal{M}$ are BVARs, the overhead $C = \lceil N/|\mathcal{M}| \rceil \cdot \min\{k_{\max}^3, S^3\}$ is time-constant. Thus, BOCPDMS runs in $\mathcal{O}(TR_{\max})$ on T observations. In contrast, the models of Saatçi et al. (2010) run in $\mathcal{O}(TR_{\max}^3)$.

Empirical evaluation of computation time

The experiments in section 7.1.5 confirm our improved computational complexity relative to the GP-based competitor methods: Using the software of Turner (2012) on the Nile data, fitting their ARGPCP model takes 42 seconds compared to 12 seconds when fitting three models in BOCPDMS, so a BVAR model is $> 10\times$ faster to process. Per inferred parameter, BOCPDMS is $> 60\times$ faster than ARGPCP; and this factor is much larger for multivariate data (e.g., > 270 for the 30 Portfolio data).

For this comparison, we use the original code of Turner (2012) for the GP-models. As the MSE is smallest for ARGPCP for all data sets except for the snowfall

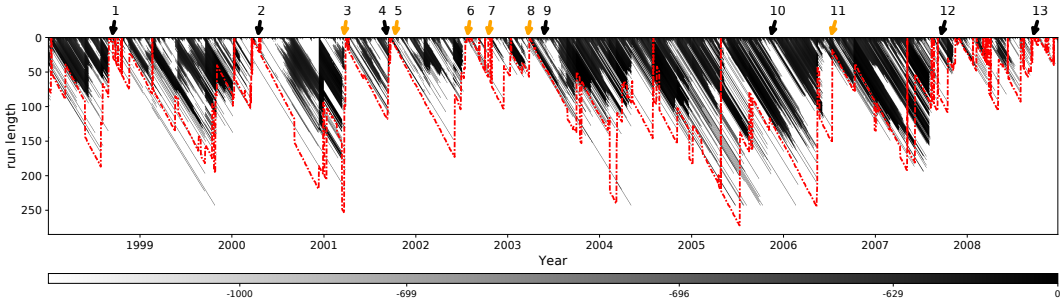


Figure 7.3: *Results for* 30 *Portfolio data set, displayed from 01/01/1998–31/12/2008*: Log run-length distribution (grayscale) and its maximum (dashed). Changepoints (CPs) found by Saatçi et al. (2010) are marked in **black**, additional CPs found by BOCPDMS in orange. Labels correspond to: (1) Asia Crisis, (2) Dot-Com bubble bursting, (3) OPEC cuts output by 4%, (4) 9/11, (5) Afghanistan war, (6) 2002 stock market crash, (7) Bombing attack in Bali, (8) Iraq war, (9) Major tax cuts under Bush, (10) US election, (11) Iran announces successful enrichment of Uranium, (12) Northern Rock bank run, (13) Lehman Brothers collapse.

data, we compare BOCPDMS against the arguably best GP CP model. We note that while NSGP performs better on the snowfall data than ARGPCP, its requirement to do Hamiltonian Monte Carlo sampling will make it significantly slower. We also note that BVAR models inside BOCPDMS outperformed the MSE of the ARGPCP model for all data sets considered. All computations were performed on a 3.1 GHz Intel i7 with 16GB RAM.

Table 7.2 summarizes the results. It is clear that BOCPDMS outperforms ARGPCP computationally: e.g., the computation time per parameter is between 60 (Nile data) and 585 (Bee data) times faster for BOCPDMS with BVAR models. Computation times are faster per model, too. The only exception to this is the 30 Portfolio data set, where the deployed SSBVAR models are orders of magnitude more parameter-rich than the ARGPCP-model. Related to this, we also note that comparing the computation time per parameter makes sense for two reasons: Firstly, BVARs model the d time series jointly, thus requiring d^2 parameters in the posterior covariance matrix of x_t. In contrast, the GP-models ignore any dependence between the series, resulting in d parameters of the (diagonal) posterior covariance matrix for x. Secondly, the parameters of the GP's kernel arguably making its parameter space Θ infinite-dimensional are not actually learnt on-line at all. Instead, they are optimized for a training period of T' observations and then fixed, see section 4 in their paper. Hence, the parameter space the GP-models can learn in is finite-dimensional.

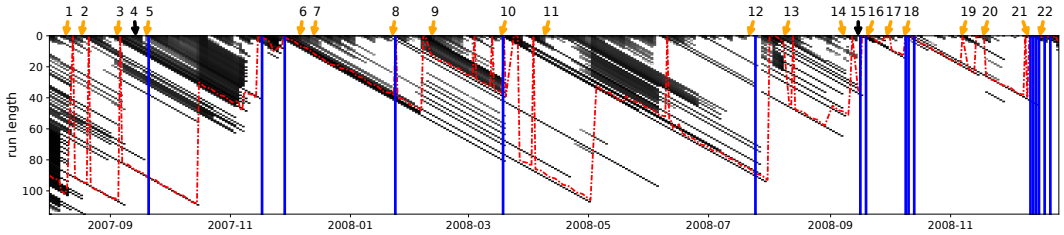


Figure 7.4: *Financial crisis 01/08/2007–31/12/2008*: Colours as in Fig 7.3, with MAP segmentation. Event labels: (1) BNP Paribas funds frozen, (2) Fed cuts lending rate, (3) IKB 1bn$ losses, (4) Northern Rock bank run, (5) Fed cuts interest rate, (6) Bush rescue plan for $>10^6$ homeowners, (7) Fed, ECB, BoE loans for banks, (8) Fed cuts funds rate, (9) G7 estimate: 400bn$ losses worldwide, (10) JP Morgan buys Bear Stearns, (11) IMF estimate: >1trn$ losses worldwide, (12) HBOS' rights issue fails, (13) ECB provides 200bn for liquidity, (14) Fannie Mae & Freddie Mac bailout, (15) Lehman collapse, (16) Russia: 500bn Roubles crisis package, (17) Fortis bailout, (18) UK: £500bn bank rescue package, (19) BoE, ECB cut interest rate, (20) G20 promise fiscal stimuli, (21) Madoff's Ponzi scheme revealed, South Korean CB sets interest rate at record low (22) Fed, Japanese central bank cut interest rates. Dates from Guillén (2009).

7.1.5 Experimental results

We evaluate the performance of the algorithm in two parts. First, we compare it to benchmark performances of GP-based models on real world data reported by Saatçi et al. (2010). This shows that as implied by Theorem 7.1, VARs are excellent approximations for a large variety of data streams. Next, we showcase BOCPDMS' novelty in the multivariate setting. We use uniform model priors q, a constant Hazard functions H and gradient descent for hyperparameter optimization as in Section 7.1.3. The lag lengths of models in $\mathcal{M}$ are chosen based on Thm. 7.1 (3) and the rates of Hannan and Kavalieris (1986) for BVARs and Bayesian Autoregressions (BARs), respectively.

Comparison with GP-based approaches

As in Saatçi et al. (2010), ARGPCP will refer to the non-linear GP-based AR model, GPTSCP to the time-deterministic model, and NSGP to the non-stationary GP allowing hyper-parameters to change at every CP. Saatçi et al. (2010) compute the mean squared error (MSE) as well as the negative log predictive likelihood (NLL) of the one-step-ahead predictions for three data sets: The water height of the Nile between $622 - 1284$ AD, the snowfall in Whistler (Canada) over a 37 year period and the 3-dimensional time series (x-, y-coordinate and headangle) of a honey bee

Table 7.2: One-step-ahead predictive MSE and NLL of BOCPDMS compared to GP-based techniques, with 95% error bars. All GP results are taken from Saatçi et al. (2010) and Turner (2012).

	Nile		Snowfall	
Method	MSE	NLL	MSE	NLL
ARGPCP	0.553	1.15	0.750	−0.604
	(0.0962)	(0.0555)	(0.0315)	(0.0385)
GPTSCP	0.583	1.19	0.689	1.17
	(0.0989)	(0.0548)	(0.0294)	(0.0183)
NSGP	0.585	1.15	**0.618**	**−1.98**
	(0.0988)	(0.0655)	(0.0242)	(0.0561)
BVAR	**0.550**	**1.13**	0.681	0.923
	(0.0948)	(0.0684)	(0.0245)	(0.0231)
	Bee Dance		30 Portfolios	
Method	MSE	NLL	MSE	NLL
ARGPCP	2.62	4.07	29.95	39.55
	(0.195)	(0.150)	(0.50)	(0.22)
GPTSCP	3.13	4.54	30.17	**39.44**
	(0.241)	(0.188)	(0.51)	(0.22)
NSGP	3.17	4.19	–	–
	(0.230)	(0.212)	–	–
BVAR	**1.74**	**3.57**	**25.93**	48.32
	(0.222)	(0.166)	(0.906)	(0.964)

during a waggle dance sequence. In Turner (2012), all of the models except NSGP were also compared on daily returns for 30 industry portfolios from 1975 − 2008. In Table 7.2, BOCPDMS is compared to these benchmarks for $\mathcal{M}$ consisting of BAR and SSBVAR models.

Designing $\mathcal{M}$ Both the Nile and the snowfall data are univariate, so $\mathcal{M}$ consists of Bayesian Autoregressions (BARs) with varying lag lengths. For the 3-dimensional bee data, $\mathcal{M}$ additionally contains unrestricted BVARs. Lastly, SSBVARs are used on the 30 Portfolio data. Two neighbourhood systems are constructed from distances in the spaces of pairwise contemporaneous correlations and autocorrelations prior to 1975, a third using the *Standard Industrial Classification (*SIC*)*, with $\xi(\cdot)$ decreasing linearly.

Predictive performance and fit: In terms of MSE, BOCPDMS clearly outperforms all GP-models on multivariate data. Even on univariate data, the only exception to this is the snowfall data, where NSGP does better. However, NSGP requires grid search or Hamiltonian Monte Carlo sampling for hyperparameter optimization at each observation (Saatçi et al., 2010). Overall, there are three main reasons why BOCPDMS performs better: Firstly, being able to change lag lengths between CPs seems more important to predictive performance than being able to model non-linear dynamics. Secondly, unlike the GP-models, we allow the individual time series to be modelled jointly via $\{\boldsymbol{A}_l^L\}$. Thirdly, the hyperparameters of the GP have a strong influence on inference. In particular, the noise variance σ is treated as a hyperparameter and optimized via type-II Maximum Likelihood/Empirical Bayes. Except for the NSGP, this is only done during a training period. Thus, the GP-models cannot adapt to the observations after training, leading to overconfident predictive distributions that are too narrow (see Turner, 2012, p. 172). This in turn leads them to be more sensitive to outliers, and to mislabel them as CPs. In contrast, (7.10a)–(7.10d) models σ as part of the inferential Bayesian hierarchy, and hyperparameter optimization is instead applied at one level higher. Consequently, our predictive distributions are wider, and the algorithm is less confident about the next observations, making it more robust to outliers. This is also responsible for the overall smaller standard errors of the GP-models in Table 7.2, since the GPs interpret outliers as CPs and immediately adapt to short-term highs or lows. Amongst other things, it is this observation that inspires our robust procedure for on-line changepoint methods introduced later in the chapter.

CP Detection: The Nile data set is also a good demonstration of this lack of robustness for the GP based models: there, BOCPDMS's MAP segmentation finds a single CP, corresponding to the installation of the nilometer around 715 CE, see Fig 7.5. In contrast, Saatçi et al. (2010) report 18 additional CPs corresponding to outliers. The same phenomenon is also reflected in the run-length distribution (RLD): While the probability mass in Figs. 7.3, 7.4 and 7.5 are spread across the retained run-lengths, the RLD reported in Saatçi et al. (2010) is more concentrated and even degenerate for the 30 Portfolio data set. On the other hand, such enhanced sensitivity to change can be advantageous. For instance, in the bee waggle dance, the GP-based techniques are better at identifying the true CPs. The reason is twofold: Firstly, the variance for the bee waggle data is homogeneous across time, so treating it as fixed helps inference. Secondly, the CPs in this data set are subtle, so having narrower predictive distributions is of great help in detecting them. However, it

adversely affects performance when changes in the error variance are essential, as for financial data: In particular, BOCPDMS finds the ground truths labelled in Saatçi et al. (2010), and discovers even more, see Fig. 7.3. This is especially apparent in times of market turmoil where changes in the variance of returns are significant. We show this using the example of the subprime mortgage financial crisis: While the RLD of Saatçi et al. (2010) identified only 2 CPs with ground truth and a third unlabelled one during the height of the crisis, BOCPDMS detects a large number of CPs corresponding to ground truths, see Fig. 7.4.

Lastly, we note that segmentations obtained off-line for both the bee waggle dance and the 30 Portfolios are reported in Xuan and Murphy (2007). Compared to the on-line segmentations produced by BOCPDMS, these are closer to the truth for the bee waggle data, but not for the 30 Portfolio data set.

Model selection: In most of the experiments where abrupt changes model the non-stationarity well, the model posterior is fairly concentrated and periods of model uncertainty are short. This is different when changes are slower, see Fig. 7.6. The implicit model complexity penalization Bayesian model selection performs provides BOCPDMS with an Occam's Razor mechanism: Simple models are typically favoured until evidence for more complex dynamics accumulates. For the bee waggle

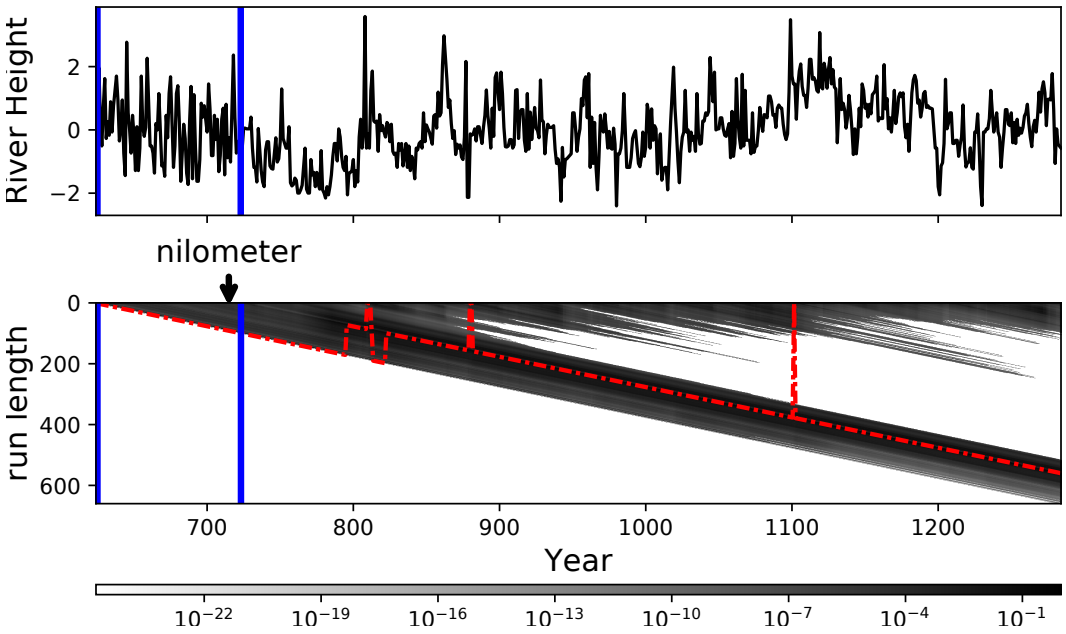


Figure 7.5: *Results for Nile data*: **Panel 1:** Nile data with structural change at 715. **Panel 2:** Both run-length distribution (grayscale with dashed maximum) and MAP segmentation detect the change.

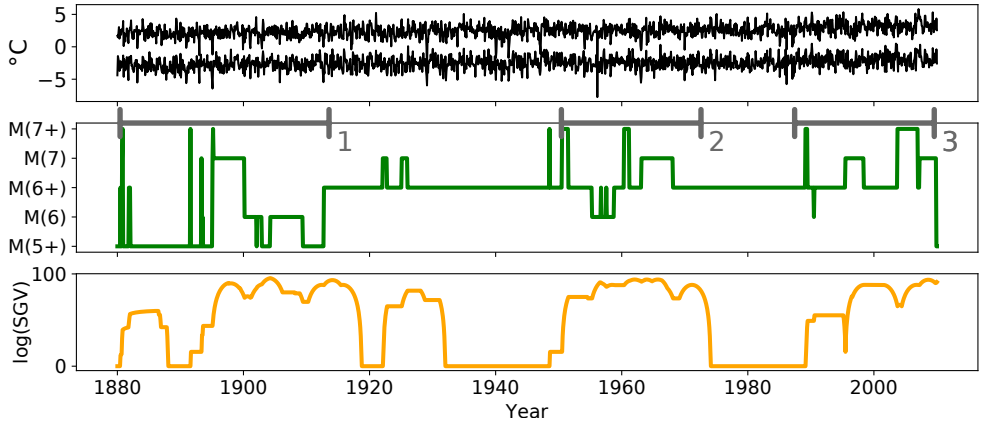


Figure 7.6: *Results for European Temperatures*: **Panel 1:** normalized temperature for Prague and Jena **Panel 2:** Model Posterior maximum, $\widehat{m}_t = \arg\max_{m_t \in \mathcal{M}} \{p(m_t|y_{1:t})\}$, model complexity decreasing top to bottom. $M(l), M(l+)$ are SSBVAR with l lags. Spatial dependence in $M(l+)$ is slower decaying. Periods of model uncertainty are (1) 2nd Industrial Revolution $1870 - 1914$, (2) Post WW2 boom $1950 - 1973$, (3) European Climate shift $1987-$present, see Luterbacher et al. (2004). **Panel 3:** To compare model uncertainty across different data and $\mathcal{M}$, the (Log) Standardized Generalized Variance (SGV) of $\widehat{m}_t$ can be used.

and the 30 Portfolio data set, BVARs are preferred to BARs. For the 30 Portfolio data, the MAP segmentation only selects SSBVARs with neighbourhoods constructed from contemporaneous correlation and autocorrelations. Neighbourhoods using SIC codes are not selected, reflecting that this classification from 1937 is out of date.

Performance on spatio-temporal data

European Temperature: Monthly temperature averages $01/01/1880-01/01/2010$ for the 21 longest-running stations across Europe are taken from http://www.ecad.eu/. We adjust for seasonality by subtracting monthly averages for each station. Station longitudes and latitudes are available, so $N(\mathcal{S})$ is based on concentric rings around the stations using Euclidean distances. Two different decay functions $\xi(\cdot), \xi^+(\cdot)$ are used, with $\xi^+(\cdot)$ using larger neighbourhoods and slower decaying. Temperature changes are poorly modeled by CPs and more likely to undergo slow transitions. Fig. 7.6 shows the way in which the model posterior captures such longer periods of change in dynamics. The values on the bottom panel are calculated by considering $\widehat{m}_t = \arg\max_{m_t \in \mathcal{M}} p(m_t|x_{1:t})$ as $|\mathcal{M}|$-dimensional multinomial random variable. Its Standardized Generalized Variance (SGV) (Wilks, 1960; SenGupta, 1987) is calculated as $|\mathcal{M}|$-th root of the covariance matrix determinant. We plot the log of the

SGV computed using the model posteriors for the last 8 years. This provides an informative summary of the model posterior dispersion.

Air Pollution: Finally, we analyze Nitrogen Oxide (NOX) observed at 29 locations across London 17/08/2002 – −17/08/2003. The quarterhourly measurements are averaged over 24 hours. Weekly seasonality is accounted for by subtracting week-day averages for each station. $\mathcal{M}$ is populated with SSBVAR models whose neighbourhoods are constructed from both road- and Euclidean distances. As 17/02/2003 marks the introduction of London's first ever congestion charge, we find structural changes in the dynamics around that date. A model with shorter lag length but identical neighbourhood structure is preferred after the congestion charge. In Fig. 7.4.2, Bayes Factors (BFs) capture the shift: Kass and Raftery (1995) classify logs of BFs as very strong evidence if their absolute value exceeds 5.

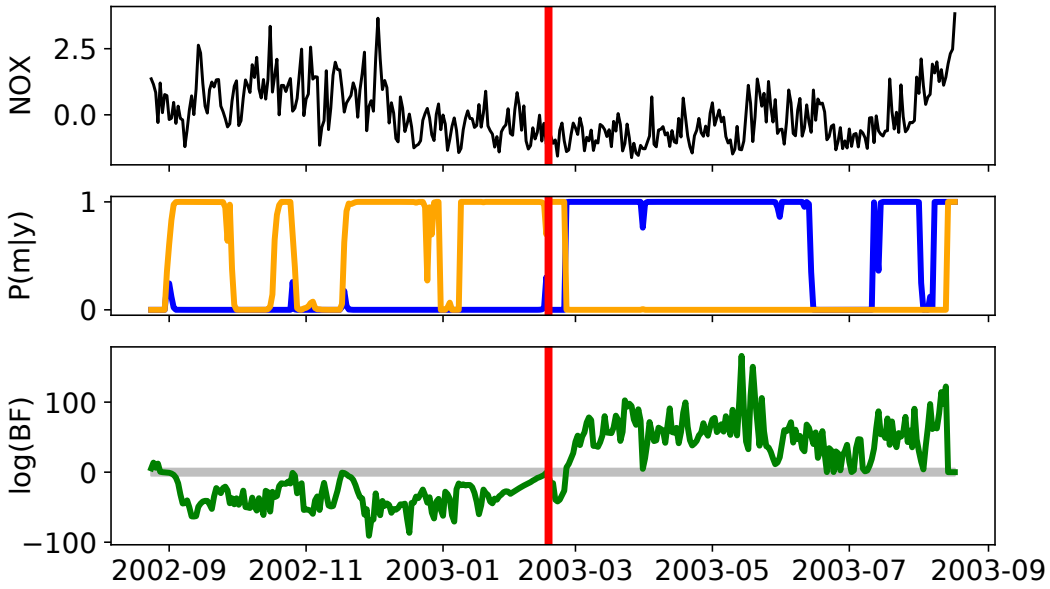


Figure 7.7: *Results for Air Pollution*: **Panel 1:** **NOX** levels for Brent, with congestion charge introduction date **Panel 2:** Model posteriors for the two best-fitting models, with Euclidean neighbourhoods. **Panel 3:** Their log Bayes Factors, $[-5, 5]$ shaded.

7.2 Doubly Robust Bayesian On-line Changepoint Detection

As we saw in Section 7.1.5, even with multiple models, BOCPD and BOCPDMS are not robust algorithms in the presence of outliers. To address this, we now

present a robust version through generalized posteriors based on β-divergence losses. The resulting inference procedure is doubly robust for both the predictive and the changepoint (CP) posterior, with linear time and constant space complexity. We provide a construction for exponential models and demonstrate it on the Bayesian Linear Regression (BLR) model. In so doing, we make two additional contributions: Firstly, we use Variational approximations that are exact as $\beta \to 1$. Secondly, we give a principled way of choosing the divergence parameter β by minimizing expected predictive loss on-line. This offers the state of the art and improves the False Discovery Rate of CPs by more than 80% on several real world data set.

In a nutshell, inference algorithms building on BOCPD (e.g. Adams and MacKay, 2007; Fearnhead and Liu, 2007; Turner et al., 2009; Xuan and Murphy, 2007; Wilson et al., 2010; Saatçi et al., 2010; Caron et al., 2012; Niekum et al., 2014; Turner et al., 2013; Ruggieri and Antonellis, 2016; Knoblauch and Damoulas, 2018) declare CPs if the posterior predictive computed from $x_{1:t}$ at time t has low density for the value of the observation x_{t+1} at time $t+1$. Naturally, this leads to a high false CP discovery rate in the presence of outliers and—as the algorithms run on-line—pre-processing is not an option. Here, we address this problem by changing the way the predictive distribution is formed: rather than integrating over the standard Bayes posterior—which as we have seen in Chapter 1 is not robust to outliers and misspecification—we design a robust posterior belief over model parameters using Generalized Variational Inference (GVI).

7.2.1 Motivation

For data $x_{1:n}$ with empirical measure $\mathbb{P}_n$ and a likelihood model $p(\cdot|\boldsymbol{\theta})$ with associated measure $\mathbb{P}_{\boldsymbol{\theta}}$, it approximately holds that

$$\arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{KLD}(\mathbb{P}_n \| \mathbb{P}_{\boldsymbol{\theta}}) \approx \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n}\sum_{i=1}^{n} -\log p(x_i|\boldsymbol{\theta}),$$

which demonstrates that standard Bayesian inference aims at minimizing the Kullback-Leibler divergence (KLD) between the fitted model and the Data Generating Mechanism. As this thesis has discussed extensively however, the negative log likelihood—or equivalently, the KLD—is not a robust way of performing inference on the model parameters under outliers or model misspecification due to the negative log likelihood's influence function (see Figure 6.1). We remedy this by instead minimizing the β-divergence ($D_B^{(\beta)}$) between the model and the data. As Figure 6.1 showed, doing so allows us to circumvent the drawbacks of the log likelihood. In addressing misspecification and outliers this way, our approach builds on the principles of the

Rule of Three (RoT), General Bayesian Inference (GBI) (see Bissiri et al., 2016; Jewson et al., 2018), Generalized Variational Inference (GVI), and robust divergences more broadly (e.g. Basu et al., 1998; Ghosh and Basu, 2016). To make our approach work, we make three contributions in separate domains that are also illustrated in Figures 7.8 and 7.10:

(1) **Robust BOCPD**: We construct the very first robust BOCPD inference procedure. The method is applicable to a wide class of (multivariate and univariate) models and is demonstrated on Bayesian Linear Regression (BLR). Unlike standard BOCPD or BOCPDMS, it is robust to confusing outliers and CPs, see Figure 7.8 **B**.

(2) **Scalable GBI:** We remedy the intractability of the proposed posterior belief using GVI. Our proposed variational family is expressive, and preserves parameter dependence. Beyond that, the GVI posterior we propose has the interpretation of an approximation to a generalized (or Gibbs-) posterior, and is exact as $\beta \to 0$, which results in a near-perfect approximation; see Figure 7.10.

(3) **Choosing β:** While Figure 6.1 shows that β regulates the degree of robustness (see also Jewson et al., 2018; Basu et al., 1998), it is unclear how to set its magnitude. For the on-line setting, we provide a way of initializing and sequentially refining β by minimizing predictive losses.

The remainder of the section is structured as follows: In Section 7.2.2, we show how to extend BOCPD (and BOCPDMS) to robust inference using the $D_B^{(\beta)}$. We quantify the degree of robustness and show that inference using the $D_B^{(\beta)}$ can be designed so that a single outlier never results in false declaration of a CP, which is impossible under the KLD. Next, we motivate an efficient GVI posterior interpretable as approximation to a generalized (or Gibbs) posterior based on the β-divergence loss. Within BOCPD, we propose using this posterior with variance-reduced Stochastic Gradient Descent. Next, Section 7.3 expands on how β can be initialized before the algorithm is run and then optimized on-line during execution time. Lastly, Section 7.4 showcases the substantial gains in performance of robust BOCPD when compared to its standard version on real world data in terms of both predictive error and CP detection.

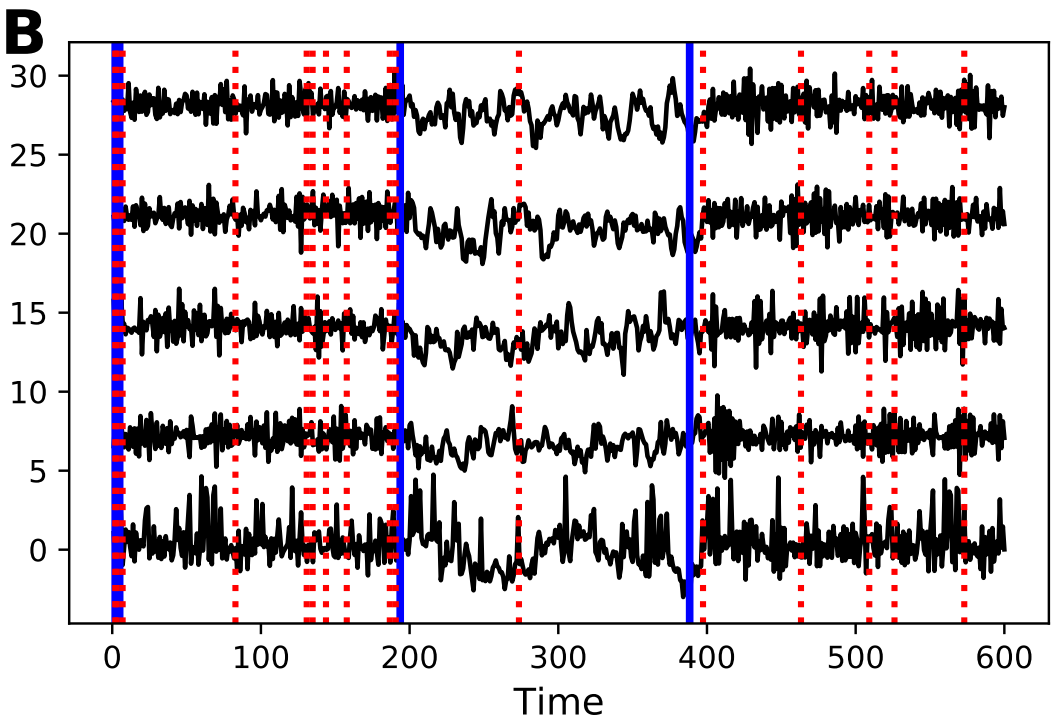


Figure 7.8: Five jointly modeled Simulated Autoregressions (ARs) with true CPs at $t = 200, 400$; bottom-most AR injected with t_4-noise. The Maximum A Posteriori (MAP) locations of CPs are shown as solid (dashed) vertical lines. The results of our robustified procedured are depicted in of blue; those of standard BOCPD in red.

7.2.2 Using Bayesian On-line Changepoint Detection with β-Divergences

BOCPD and BOCPDMS are based on the Product Partition Model (Barry and Hartigan, 1993). Recall from the previous section that the BOCPDMS model is

$$r_t|r_{t-1} \sim H(r_t, r_{t-1}) \qquad m_t|m_{t-1}, r_t \sim q(m_t|m_{t-1}, r_t) \tag{7.14a}$$

$$\boldsymbol{\theta}_m|m_t \sim \pi_{m_t}(\boldsymbol{\theta}_{m_t}) \qquad x_t|m_t, \boldsymbol{\theta}_{m_t} \sim f_{m_t}(x_t|\boldsymbol{\theta}_{m_t}) \tag{7.14b}$$

where $q(m_t|m_{t-1}, r_t) = m_{t-1}$ for $r_t > 0$ and $q(m_t)$ otherwise, and where for convenience, we have written $H(r, r') = p(r|r')$. For example, an instantation of this model with Bayesian Linear Regression (BLR) using a $d \times p$ time-varying regressor matrix $\boldsymbol{V}_t$ is given by $\boldsymbol{\theta}_m = (\sigma^2, \boldsymbol{\mu})$, $f_m(x_t|\boldsymbol{\theta}_m) = \mathcal{N}_d(x_t; \boldsymbol{V}_t\boldsymbol{\mu}, I_d)$ and $\pi_m(\boldsymbol{\theta}_m) = \mathcal{N}_d(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \sigma^2\Sigma_0)\mathcal{IG}(\sigma^2; a_0, b_0)$; where $\mathcal{N}$ denotes a normal and $\mathcal{IG}$ an inverse-gamma distribution. If the computations of the parameter posterior $\pi_m(\boldsymbol{\theta}_m|x_{1:t}, r_t)$ and the

posterior predictive

$$f_m(x_t|x_{1:(t-1)}, r_t) = \int_{\Theta_m} f_m(x_t|\boldsymbol{\theta}_m)\pi_m(\boldsymbol{\theta}_m|x_{1:(t-1)}, r_t)d\boldsymbol{\theta}_m \tag{7.15}$$

are efficient for all models $m \in \mathcal{M}$, then so is the recursive computation given in the previous section. For the BVAR and SSBVAR models discussed before, this is the case due to conjugacy—which guarantees that the predictive distributions are available in closed form.

Once we depart from the negative log likelihood function to perform robust on-line inference with a β-divergence based loss, this is no longer true. Consequently, computing the integrals and predictives becomes a challenge that needs solving. Because running samplers would blow up the computation time of the algorithm to a point where it would be on-line in name only, we choose to turn to a tool which we have studied before: GVI.

Throughout this section, we assume that the prior belief for our parameters is reasonable. Consequently, we seek to compute a posterior within the RoT that uses a robust loss function, but the conventional (and convenient) choice of $D = \text{KLD}$. Optimized over the space of all probability measures, this would entail that we are working with a posterior of the form $P(\mathcal{L}_m^{\beta_p}, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$ given as

$$\pi_m^{\beta_\text{m}}(\boldsymbol{\theta}_m|x_{(t-r_t):t}) \propto \pi_m(\theta) \exp\left\{-\textstyle\sum_{i=t-r_t}^{t} \mathcal{L}_m^{\beta_p}(x_t, \boldsymbol{\theta})\right\}, \tag{7.16}$$

where—adapting the notation of Chapter 6 to the current setting—we have that

$$\mathcal{L}_m^{\beta_p}(x_t, \boldsymbol{\theta}) = -\left(\frac{1}{\beta_\text{m} - 1} f_m(x_t|\boldsymbol{\theta}_m)^{\beta_\text{m}} - \frac{1}{\beta_\text{m}} \int_{\mathcal{X}} f_m(\boldsymbol{z}|\boldsymbol{\theta}_m)^{\beta_\text{m}} d\boldsymbol{z}\right). \tag{7.17}$$

The above equation illustrates on an intuitive level why the $D_B^{(\beta)}$ excels at robust inference: Similar to tempering, $\mathcal{L}_m^{\beta_p}$ exponentially down-weights the density for values of $\beta_\text{m} > 1$, thereby attaching less influence to observations in the tails of the model. Conversely, under the log score of KLD, *more* influence is associated with an observation the further out in the tails of the model it occurs. It is this phenomenon that we saw depicted with influence functions in Figure 6.1.

7.2.3 Robust BOCPD

The literature on robust on-line CP detection so far is sparse, and only covers limited settings without Bayesian uncertainty quantification (e.g. Pollak, 2010; Cao and Xie, 2017; Fearnhead and Rigaill, 2019). For example, the method in Fearnhead

and Rigaill (2019) only produces point estimates and is limited to fitting a piecewise constant function to univariate data. In contrast, BOCPD and BOCPDMS can be applied to multivariate data and a set of models $\mathcal{M}$ while quantifying uncertainty about these models, their parameters and potential CPs, but is not robust. Noting that for standard BOCPD the posterior expectation is given by

$$\mathbb{E}\left(x_t|x_{1:(t-1)}\right) = \sum_{r_t,m_t} \mathbb{E}\left(x_t|x_{1:(t-1)}, r_{t-1}, m_{t-1}\right) p(r_{t-1}, m_{t-1}|x_{1:(t-1)}), \quad (7.18)$$

the key observation is that prediction is driven by **two** probability distributions: The run-length and model posterior $p(r_t, m_t|x_{1:t})$ and parameter posterior distributions $\pi_m(\boldsymbol{\theta}|x_{1:t})$. This also means that we should make BOCPD robust by using **two** posteriors robustified via the $D_B^{(\beta)}$—one for the run-lengths and models we will denote $p^{\beta_{\text{rlm}}}(r_t, m_t|x_{1:t})$, and one for the parameters that we will denote $\pi_m^{\beta_{\text{m}}}(\boldsymbol{\theta}|x_{1:t})$. Here, $\beta_{\text{rlm}} > 0$, and $\beta_{\text{m}} > 0$ (and in fact, $\beta_{\text{rlm}} > 1$ and $\beta_{\text{m}} > 1$ whenever inference is supposed to be robust).

β_{rlm} prevents abrupt changes in $p^{\beta_{\text{rlm}}}(r_t, m_t|x_{1:t})$ caused by a small number of observations, see section 7.2.4. This form of robustness is easy to implement and retains the closed forms of the standard BOCPD and BOCPDMS computations: one simply replaces $f_{m_t}(x_t|x_0)$ and $f_{m_t}(x_t|x_{1:(t-1)}, r_{t-1})$ by their respective counterpart

$$\exp\{\mathcal{L}_{m_t}^{\beta_{\text{rlm}}}(r_t|x_{1:t})\} = \exp\left\{ -\left(\frac{1}{\beta_{\text{rlm}}-1} f_{m_t}(x_t|x_{1:(t-1)}, r_{t-1})^{\beta_{\text{rlm}}-1} \right.\right.$$
$$\left.\left. -\frac{1}{\beta_{\text{rlm}}} \int_{\mathcal{Y}} f_{m_t}(\boldsymbol{z}|x_{1:(t-1)}, r_{t-1})^{\beta_{\text{rlm}}} d\boldsymbol{z}\right)\right\}. \quad (7.19)$$

While $p^{\beta_{\text{rlm}}}(x_{1:t}, r_t, m_t)$ is not a normalized density anymore—and so does not generally integrate to one—$p^{\beta_{\text{rlm}}}(r_t, m_t|x_{1:t})$ still is a valid probability measure and sums to one. More importantly, $p^{\beta_{\text{rlm}}}(r_t, m_t|x_{1:t})$ ensures that the run-length distribution is robust.

Complementing this, β_{m} regulates the robustness of the parameter posteriors $\pi_m^{\beta_{\text{m}}}(\boldsymbol{\theta}|x_{1:t})$ by preventing them from being dominated by tail events. Section 7.2.5 overcomes the intractability of $\pi_m^{\beta_{\text{m}}}(\boldsymbol{\theta}|x_{1:t})$. While computation is more challenging for $\pi_m^{\beta_{\text{m}}}(\boldsymbol{\theta}|x_{1:t})$ than it is for the non-robust posterior $\pi_m(\boldsymbol{\theta}|x_{1:t})$—which was available in closed form due to conjugacy!—we will see later on how GVI based on $P(\mathcal{L}_m^{\beta_p}, \text{KLD}, \mathcal{Q})$ with $\mathcal{Q}$ being the set of all normal distributions can help us address this challenge. In particular, this GVI posterior is not only computationally efficient, but even recovers the approximated distribution $P(\mathcal{L}_m^{\beta_p}, \text{KLD}, \mathcal{P}(\boldsymbol{\Theta})) = \pi_m^{\beta_{\text{m}}}(\boldsymbol{\theta}|x_{1:t})$ exactly as $\beta_{\text{m}} \to 1$.

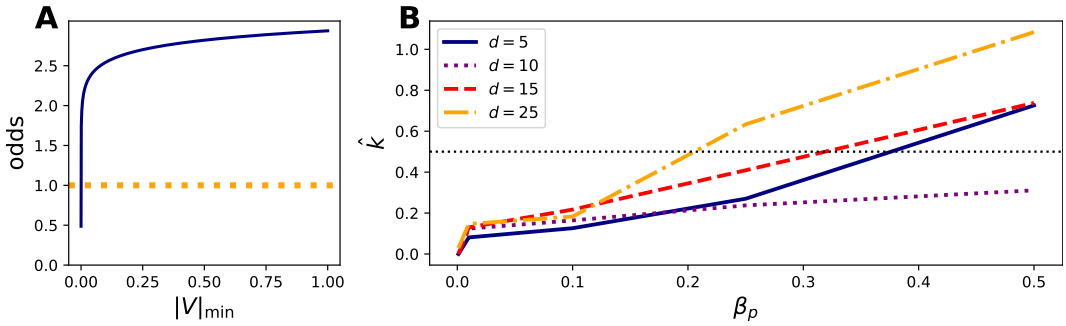


Figure 7.9: **A**: Lower bound on the odds of Thm. 7.2 for priors used for Figure 7.8 **B** and $h(r) = 1/100$. **B**: $\hat{k}$ for different choices of $\beta_p = \beta_{\text{m}} - 1$ (so that $\beta_p = 0$ corresponds to the negative log likelihood) and output (input) dimensions d ($2d$) in an autoregressive BLR.

7.2.4 Quantifying robustness

The algorithm of Fearnhead and Rigaill (2019) is robust because hyperparameters enforce that a single outlier is insufficient for declaring a CP. Analogously, we can quantify robustness by conditioning on $r_t = r$ and studying the odds of $r_{t+1} \in \{0, r+1\}$:

$$\begin{aligned} &\frac{p(r_{t+1} = r+1 | x_{1:t+1}, r_t = r, m_t)}{p(r_{t+1} = 0 | x_{1:t+1}, r_t = r, m_t)} \\ &= \frac{\cancel{p(x_{1:t}, r_t = r, m_t)} \cdot (1 - H(r_{t+1}, r_t)) f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1} | x_{1:(t-1)}, r_{t-1})}{\cancel{p(x_{1:t}, r_t = r, m_t)} \cdot H(r_{t+1}, r_t) f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1} | x_0)}. \end{aligned} \tag{7.20}$$

Here, $f_{m_t}^{\beta_{\text{rlm}}}$ denotes the negative exponential of the score under divergence D. In particular $f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1} | x_{1:(t-1)}, r_{t-1}) = \exp\{\mathcal{L}_{m_t}^{\beta_{\text{rlm}}}(r_t | x_{1:t})\}$ as in (7.19). Taking a closer look at (7.20), if x_{t+1} is an outlier with low density under $f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1} | x_{1:(t-1)}, r_{t-1})$, the odds will move in favor of a CP provided that the prior is sufficiently uninformative to make $f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1} | x_0) > f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1} | x_{1:(t-1)}, r_{t-1})$. In fact, even very small differences have a substantial impact on the odds. For BLR, Theorem 7.2 provides conditions guaranteeing that these odds never favor a CP after a single observation under the β-divergence based loss when they would under the negative log likelihood—i.e. when $f_{m_t}(x_{t+1} | x_0)$ is much larger than $f_{m_t}(x_{t+1} | x_{1:(t-1)}, r_{t-1})$.

Theorem 7.2. If m_t in (7.20) is the Bayesian Linear Regression (BLR) model with $\boldsymbol{\mu} \in \mathbb{R}^p$ and priors a_0, b_0, μ_0, Σ_0; and if the posterior predictive's variance determinant is larger than $|V|_{\min} > 0$, then one can choose any $(\beta_{\text{rlm}}, H(r_t, r_{t+1})) \in$

$S\left(p, \beta_{\text{rlm}}, a_0, b_0, \mu_0, \Sigma_0, |V|_{\min}\right)$ to guarantee that

$$\frac{(1 - H(r_{t+1}, r_t)) f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_{1:(t-1)}, r_{t-1})}{H(r_{t+1}, r_t) f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_0)} \geq 1, \tag{7.21}$$

where the set $S\left(p, \beta_{\text{rlm}}, a_0, b_0, \mu_0, \Sigma_0, |V|_{\min}\right)$ is defined by an inequality given in Appendix C.2.

Thm. 7.2 says that one can bound the odds for a CP independently of x_{t+1}. The requirement for a lower bound $|V|_{\min}$ results from the integral term in (7.19), which dominates $D_B^{(\beta)}$-based inference if $|V|$ is extremely small. In practice, this is not restrictive: E.g. for $p = 5$, $h(r) = \frac{1}{\lambda}$, $a_0 = 3, b_0 = 5, \Sigma_0 = \text{diag}(100, 5)$ used in Fig. 7.8, Thm. 7.2 holds for $(\beta_{\text{rlm}}, \lambda) = (0.15, 100)$ used for inference if $|V|_{\min} \geq 8.12 \times 10^{-6}$. Fig. 7.9 **A** plots the lower bound derived in Appendix C.2 as function of $|V|_{\min}$.

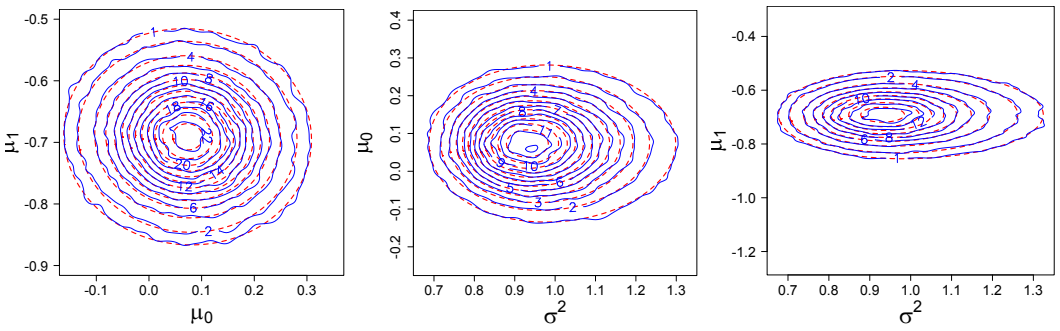


Figure 7.10: Exemplary contour plots of bivariate marginals for the approximation $\widehat{\pi}_m^{\beta_{\text{m}}}(\boldsymbol{\theta}_m)$ of (7.23) (dashed) and the target $\pi_m^{\beta_{\text{m}}}(\boldsymbol{\theta}_m|x_{(t-r_t):t})$ of (7.16) (solid) estimated and smoothed from 95,000 Hamiltonian Monte Carlo samples for the β-divergence robustified posterior of BLR with $d = 1$, two regressors and $\beta_{\text{m}} = 1.25$.

7.2.5 Structural Variational Approximation & pseudo-conjugacy

While there has been a surge in theoretical work on generalized Bayesian methods (e.g. Bissiri et al., 2016; Ghosh and Basu, 2016; Jewson et al., 2018), applications have been sparse, in large part due to intractability issues. While MCMC methods have been used successfully (Jewson et al., 2018; Ghosh and Basu, 2016), it is hard to scale them for the on-line BOCPD setting: One would have to sample from the parameter posteriors for each run-length and additionally require a second layer of sampling to evaluate the integral in (7.19). Circumventing MCMC, most work on BOCPD has focused on conjugate distributions (Adams and MacKay, 2007; Turner et al., 2009; Fearnhead and Liu, 2007) and approximations (Turner et al., 2013;

Niekum et al., 2014). We extend the latter branch of research by deploying a GVI method that does not impose independence in the parameter posterior across the dimensions of $\boldsymbol{\theta}$. For an illustration, see Figure 7.10. Further, since $D_B^{(\beta)} \to \text{KLD}$ as $\beta \to 1$, there is an especially compelling way of doing GVI based on the fact that the approximation

$$\pi_m^{\beta_\text{m}}(\boldsymbol{\theta}_m | x_{(t-r_t):t}) \approx \pi_m(\boldsymbol{\theta}_m | x_{(t-r_t):t}) \tag{7.22}$$

is exact as $\beta \to 0$. Thus, for $\mathcal{Q}$ denoting the set of all normal distributions on $\boldsymbol{\Theta}$, we approximate the β-divergence based posterior for model m and run-length r_t as

$$\widehat{\pi}_m^{\beta_\text{m}}(\boldsymbol{\theta}_m) = \underset{q \in \mathcal{Q}}{\arg\min} \left\{ \text{KLD}\left(q(\boldsymbol{\theta}_m) \middle\| \pi_m^{\beta_\text{m}}(\boldsymbol{\theta}_m | x_{(t-r_t):t}) \right) \right\}, \tag{7.23}$$

which is *exactly* equivalent to computing the GVI posterior $P(\mathcal{L}_m^{\beta_p}, \text{KLD}, \mathcal{Q})$. Note in particular that for the BLR case, this ensures that both $\widehat{\pi}_m^{\beta_\text{m}}$ and π_m^{KLD} belong to the same family—namely the family of normal distributions—even if the parameters can be very different between them. Further, for many models—and in particular for the BLR—optima of the optimization in (7.23) can be computed efficiently due to the closed form of all associated expectations. We state this in Theorem 7.3, which is almost identical to Proposition 4.2, and which we prove in Appendix B.8.1. Further, Appendix B.8.2 contains the derivation of the closed forms for the derivatives of the GVI objective for Bayesian Linear Regression (BLR) that we use in experiments.

Theorem 7.3. The objective corresponding to the posterior approximation in (7.23) as well as its derivatives with respect to the variational parameters are analytically available if they are based on an exponential family likelihood model

$$f_m(x; \theta_m) = \exp\left(\eta(\theta_m)^T T(x)\right) g(\eta(\theta_m)) A(x),$$

with conjugate prior $\pi_0(\theta_m | \nu_0, \mathcal{X}_0) = g(\eta(\theta_m))^{\nu_0} \exp\left(\nu_0 \eta(\theta_m)^T \mathcal{X}_0\right) h(\mathcal{X}_0, \nu_0)$, and variational posterior $\widehat{\pi}_m^{\beta_\text{m}}(\theta_m | \nu_m, \mathcal{X}_m) = g(\eta(\theta_m))^{\nu_m} \exp\left(\nu_m \eta(\theta_m)^T \mathcal{X}_m\right) h(\mathcal{X}_m, \nu_m)$ within the same conjugate family. Additionally, the following three quantities need to have closed form:

$$\mathbb{E}_{\widehat{\pi}_m^{\beta_\text{m}}}\left[\eta(\theta_m)\right];$$
$$\mathbb{E}_{\widehat{\pi}_m^{\beta_\text{m}}}\left[\log g(\eta(\theta_m))\right];$$
$$\int A(z)^{\beta_\text{m}} \left[h\left(\frac{(\beta_\text{m}) T(z) + \nu_m \mathcal{X}_m}{\beta_\text{m} + \nu_m}, \beta_\text{m} + \nu_m \right) \right]^{-1} dz.$$

The conditions of Theorem 7.3 are met by many exponential models, e.g. the Normal-Inverse-Gamma, the Exponential-Gamma, and the Gamma-Gamma. For a simulated autoregressive BLR, we assess the quality of $\widehat{\pi}^{\beta_\text{m}}$ following Yao et al. (2018), which proposes to estimate a difference $\hat{k}$ between $\pi_m^{\beta_\text{m}}$ and $\widehat{\pi}_m^{\beta_\text{m}}$ relative to a posterior expectation. We use this on the posterior predictive, which is an expectation relative to $\pi_m^{\beta_\text{m}}$ and drives the CP detection. Yao et al. (2018) rate $\widehat{\pi}_m^{\beta_\text{m}}$ as *close* to $\pi_m^{\beta_\text{m}}$ if $\hat{k} < 0.5$. Figs 7.10 and 7.9 **B** show that our approximation lies well below this threshold for choices of β_m decreasing reasonably fast with the dimension. Note that these are exactly the values of β_m one will want to select for inference: As d increases, the magnitude of $f_{m_t}(x_t|x_{1:(t-1)}, r_{t-1})$ decreases rapidly. Hence, β_m needs to decrease as d increases to prevent robust inference from being dominated by the integral in (7.19) and disregarding x_t (Jewson et al., 2018). This is also reflected in our experiments in section 7.4, for which we initialize $\beta_\text{m} = 0.05$ and $\beta_\text{m} = 0.005$ for $d = 1$ and $d = 29$, respectively. However, as Figures 7.10 and 7.9 **B** illustrate, the approximation is still excellent for values of β_m that are much larger than that.

7.2.6 Stochastic Variance Reduced Gradient (SVRG) for BOCPD

While highest predictive accuracy within BOCPD is achieved using full optimization of the variational parameters at each of T time periods, this has space and time complexity of $\mathcal{O}(T)$ and $\mathcal{O}(T^2)$. In comparison, Stochastic Gradient Descent (SGD) has space and time complexity of $\mathcal{O}(1)$ and $\mathcal{O}(T)$, but yields a loss in accuracy, substantially so for small run-lengths. In the BOCPD setting, there is an obvious trade-off between accuracy and scalability: Since the posterior predictive distributions $f_{m_t}(x_t|x_{1:(t-1)}, r_t)$ for all run-lengths r_t drive CP detection, SGD estimates are insufficiently accurate for small run-lengths r_t. On the other hand, once r_t is sufficiently large, the variational parameter estimates only need minor adjustments and computing an optimum is costly.

Recently, a new generation of algorithms interpolating SGD and global optimization have addressed this trade-off. They achieve substantially better convergence rates by anchoring the stochastic gradient to a point near an optimum (Johnson and Zhang, 2013; Defazio et al., 2014; Nitanda, 2014; Harikandeh et al., 2015; Lei and Jordan, 2017). We propose a memory-efficient two-stage variation of these methods tailored to BOCPD. First, the variational parameters are moved close to their global optimum using a variant of (Johnson and Zhang, 2013; Nitanda, 2014). Unlike standard versions, we anchor the gradient estimates to an optimum every m steps for the first W iterations. Compared to standard SGD or SVRG, this

Stochastic Variance Reduced Gradient (SVRG) inference for BOCPD

Input at time 0**:** Window & batch sizes W, B, b; frequency m, prior $\boldsymbol{\theta}_0$, #steps K, step size η
for next observation x_t at time t **do**
  for retained run-lengths $r \in R(t)$ **do**
    if $\tau_r = 0$ **then**
      if $r < W$ **then**
        $\boldsymbol{\theta}_r \leftarrow \boldsymbol{\theta}_r^* \leftarrow \text{FullOpt}\left(\text{ELBO}(x_{t-r:t})\right)$; $\tau_r \leftarrow m$
      else if $r \geq W$ **then**
        $\boldsymbol{\theta}_r^* \leftarrow \boldsymbol{\theta}_r$; $\tau_r \leftarrow \text{Geom}\left(B/(B+b)\right)$
      $g_r^{\text{anchor}} \leftarrow \frac{1}{B}\sum_{i\in\mathcal{I}} \nabla\text{ELBO}(\boldsymbol{\theta}_r^*, x_{t-i})$, where $\mathcal{I} \sim \text{Unif}\{0,\ldots,\min(r,W)\}$, $|\mathcal{I}| = B$
    for $i = 1, 2, \ldots, K$ **do**
      $\widetilde{\mathcal{I}} \sim \text{Unif}\{0,\ldots,\min(r,W)\}$ and $|\widetilde{\mathcal{I}}| = b$
      $g_r^{\text{old}} \leftarrow \frac{1}{b}\sum_{i\in\widetilde{\mathcal{I}}} \nabla\text{ELBO}(\boldsymbol{\theta}_r^*, x_{t-i})$, $g_r^{\text{new}} \leftarrow \frac{1}{b}\sum_{i\in\widetilde{\mathcal{I}}} \nabla\text{ELBO}(\boldsymbol{\theta}_r, x_{t-i})$
      $\boldsymbol{\theta}_r \leftarrow \boldsymbol{\theta}_r + \eta \cdot \left(g_r^{\text{new}} - g_r^{\text{old}} + g_r^{\text{anchor}}\right)$; $\tau_r \leftarrow \tau_r - 1$
  $r \leftarrow r+1$ for all $r \in R(t)$; $R(t) \leftarrow R(t) \cup \{0\}$

substantially decreases variance and increases accuracy for small r_t. Second, once $r_t > W$ we incrementally refine the estimates while keeping their variance low using a stochastic-batch variant of SVRG (Lei and Jordan, 2017; Lei et al., 2017) on a window with the W most recent observations. The resulting on-line inference has constant space and linear time complexity like SGD, but produces good estimates for small r_t and converges faster (Johnson and Zhang, 2013; Lei and Jordan, 2017; Lei et al., 2017). We provide a detailed complexity analysis of the procedure in Appendix B.8.7. Compared to MCMC-based inference, our algorithm is orders of magnitude faster: E.g. for the well-log data in section 7.4.1, an MCMC implementation in `Stan` (Carpenter et al., 2017) takes 10^5 times longer.

7.3 Choice of β

Initializing β_{m}: Losses based on the β-divergence have been used in a variety of settings (Basu et al., 1998; Ghosh and Basu, 2016; Futami et al., 2018; Yılmaz et al., 2011), but there is no principled framework for selecting the hyperparameter β.

We address this by minimizing the expected predictive loss with respect to β_{m} on-line. As the losses may not be convex in β_{m}, initial values can matter for the optimization. A priori, we pick β_{m} maximizing the $D_B^{(\beta)}$ influence for a given distance $\boldsymbol{v}^*$ between the parameter prior $\pi_m(\theta_m)$ and β_{m}. The notion of a Malahanobis Distance (MD) allows us to measure such a distance between a point on the one hand and a

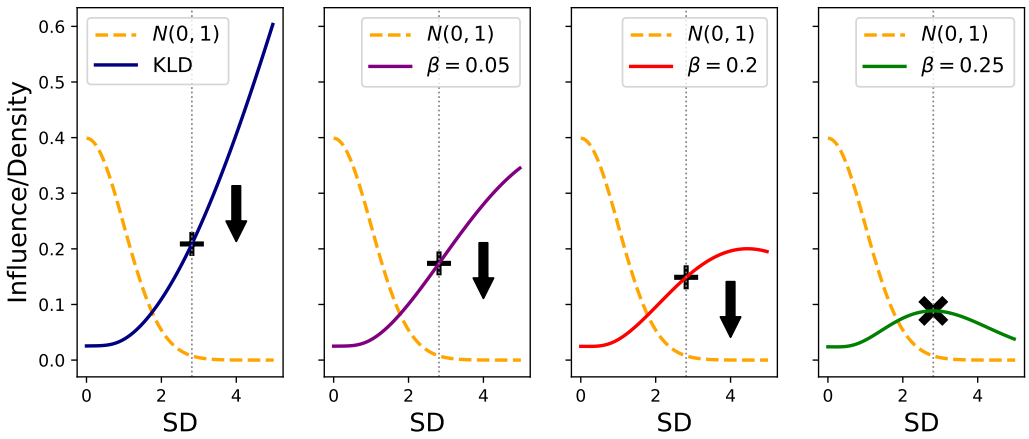


Figure 7.11: Illustration of the initialization procedure for β_{m}, from left to right.

distribution on the other hand. Denoting by $\widehat{I}(\beta_{\mathrm{m}}, \pi_m(\boldsymbol{\theta}_m)$ the (estimated) Fisher-Rao divergence depicted on the lefthand side in Figure 6.1, we define for our purposes the (estimated) MD distance as $\widehat{\mathrm{MD}}(\beta_{\mathrm{m}}, \pi_m(\theta_m)) = \arg\max_{x \in \mathbb{R}_+} \hat{I}(\beta_{\mathrm{m}}, \pi_m(\theta_m))(x)$.

The rationale behind this is as follows: As Figure 6.1 shows, $\beta_{\mathrm{m}} > 0$ induces a point of maximum influence: Points further in the tails are treated as outliers, while points closer to the mode receive similar influence as under the KLD. Our MD measure provides the value along the x-axis at which this point of maximum influence is achieved. With this in hand, we initialize β_{m} by solving the inverse problem: For a given x^*, we seek the value of β_{m} for which the point of maximum influence occurs at the MD x^*. This is illustrated in Figure 7.11. The k-th standard deviation under the prior is a good choice of x^* for low dimensions (see also Fearnhead and Rigaill, 2019), but not appropriate as delimiter for high density regions even in moderate dimensions d. Thus, we propose $x^* = \sqrt{d}$ for larger values of d. One then finds β_{m} by approximating the gradient of $\widehat{\mathrm{MD}}(\beta_{\mathrm{m}}, \pi_m(\theta_m))$ with respect to β_{m}. As β_{rlm} does not affect $\pi_m^{\beta_{\mathrm{m}}}$, its initialization matters less and generally, initializing $\beta_{\mathrm{rlm}} \in [0, 1]$ produces reasonable results.

Optimizing $\boldsymbol{\beta}$ on-line: For $\boldsymbol{\beta} = (\beta_{\mathrm{rlm}}, \beta_{\mathrm{m}})$ and prediction $\widehat{x}_t(\boldsymbol{\beta})$ of x_t obtained as posterior expectation via (7.18), define $\boldsymbol{\varepsilon}_t(\boldsymbol{\beta}) = x_t - \widehat{x}_t(\boldsymbol{\beta})$. For some predictive loss $L_p : \mathbb{R} \to \mathbb{R}_+$, we target $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \{\mathbb{E}\left(L_p(\boldsymbol{\varepsilon}_t(\boldsymbol{\beta}))\right)\}$. Replacing expected by empirical loss and deploying SGD, we seek to find the partial derivatives of $\nabla_\beta L_p\left(\varepsilon_t(\boldsymbol{\beta})\right)$. Noting that $\nabla_{\boldsymbol{\beta}} L_p\left(\varepsilon_t(\boldsymbol{\beta})\right) = L_p'\left(\varepsilon_t(\boldsymbol{\beta})\right) \cdot \nabla_{\boldsymbol{\beta}} \widehat{x}_t(\boldsymbol{\beta})$, the issue reduces to finding the partial derivatives $\nabla_{\beta_{\mathrm{rlm}}} \widehat{x}_t(\boldsymbol{\beta})$ and $\nabla_{\beta_{\mathrm{m}}} \widehat{x}_t(\boldsymbol{\beta})$. Remarkably, $\nabla_{\beta_{\mathrm{rlm}}} \widehat{x}_t(\boldsymbol{\beta})$ can be updated sequentially and efficiently by differentiating the recursion underlying BOCPD. The derivation is provided in Appendix B.8.8. The gradient

$\nabla_{\beta_{\mathrm{m}}} \widehat{x}_t(\boldsymbol{\beta})$ on the other hand is not available analytically and thus is approximated numerically. Now, β can be updated on-line via

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} - \eta \cdot \begin{bmatrix} \nabla_{\beta_{\mathrm{rlm},t}} L_p\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right) \\ \nabla_{\beta_{\mathrm{p},t}} L_p\left(\varepsilon_t(\boldsymbol{\beta}_{1:(t-1)})\right) \end{bmatrix} \tag{7.24}$$

In spirit, this procedure resembles existing approaches for model hyperparameter optimization (Caron et al., 2012). For robustness, L_p should be chosen appropriately. Thus, in our experiments we use $L_p(x) = |x|$.

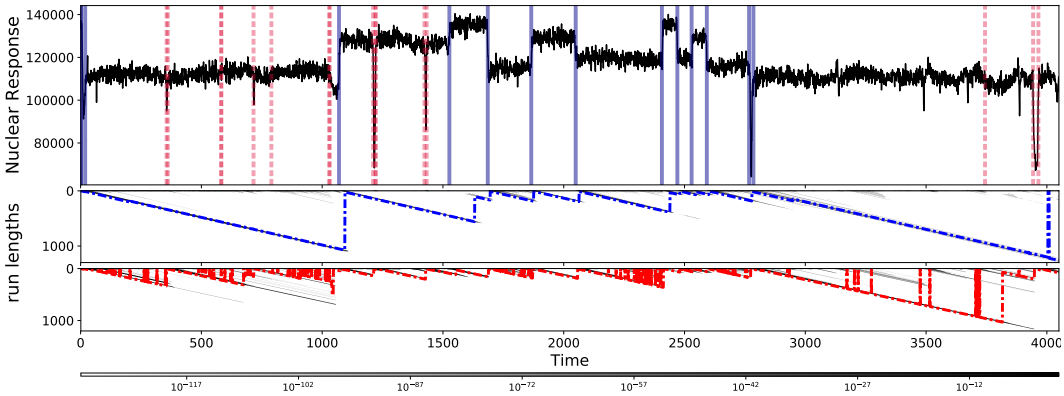


Figure 7.12: Maximum A Posteriori (MAP) segmentation and run-length distributions of the well-log data. Robust segmentation depicted using solid lines, CPs additionally declared under standard BOCPD with dashed lines. The corresponding run-length distributions for robust (middle) and standard (bottom) BOCPD are shown in grayscale. The most likely run-lengths are dashed.

7.4 Results

Next, we illustrate the most important improvements this chapter makes to BOCPD. First, we show how robust BOCPD deals with outliers on the well-log data set. Further, we show that standard BOCPD breaks down in the M-open world whilst $D_B^{(\beta)}$ yields useful inference by analyzing noisy measurements of Nitrogen Oxide (NOX) levels in London. In both experiments, we use the methods in section 7.3, online hyperparameter optimization (Caron et al., 2012) and pruning for $p(r_t, m_t|x_{1:t})$ (Adams and MacKay, 2007). More detailed information on the recursion itself can be found in Appendix B.8.8; and on the numerical treatment in Appendix A.4.3.

Additionally, Appendix A.4 provides extensive further details on the empirically studied data sets below.

7.4.1 Well-log

The well-log data set was first studied in O'Ruanaidh (1994) and has become a benchmark data set for univariate CP detection. However, except in Fearnhead and Rigaill (2019) its outliers have been removed before CP detection algorithms are run (e.g. Adams and MacKay, 2007; Levy leduc and Harchaoui, 2008; Ruggieri and Antonellis, 2016). With $\mathcal{M}$ containing one BLR model of form $y_t = \mu + \varepsilon_t$, Figure 7.12 shows that robust BOCPD deals with outliers on-line. The maximum of the run-length distribution for standard BOCPD is zero 145 times, so declaring CPs based on the run-length distribution's maximum (see e.g. Saatçi et al., 2010) yields a false discovery rate of more than 90%. This problem persists even with non-parametric, Gaussian Process, models (p. 186, Turner, 2012). Even using Maximum A Posteriori (MAP) segmentation (Fearnhead and Liu, 2007), standard BOCPD mislabels 8 outliers as CPs, making for a false discovery rate of still more than 40%. In contrast, the segmentation of our robust version does not mislabel any outliers. Further and in accordance with Thm. 7.2, its run-length distribution's maximum falsely drops to a zero run-length only once, which is in response to more than 20 consecutive outliers. A natural byproduct of the robust segmentation is a reduction in mean square (absolute) prediction error by 10% (6%) compared to the standard version. The robust version has more computational overhead than standard BOCPD, but still needs less than 0.5 seconds per observation using a 3.1 GHz Intel i7 and 16GB RAM.

Not only does robust BOCPD's segmentation in Figure 7.12 match that in Fearnhead and Rigaill (2019), but it also offers three additional on-line outputs: Firstly, it produces probabilistic (rather than point) forecasts and parameter inference. Secondly, it self-regulates its robustness via the on-line adjustment of $\boldsymbol{\beta}$. Thirdly, it can compare multiple models and produce model posteriors (see section 7.4.2). Further, unlike Fearnhead and Rigaill (2019), it is not restricted to fitting univariate data with piecewise constant functions.

7.4.2 Air Pollution

We also apply robust BOCPD to analyze Nitrogen Oxide (NOX) levels across 29 stations in London using spatially structured Bayesian Vector Autoregressions (SSBVARs). Previous robust on-line methods (e.g. Pollak, 2010; Cao and Xie, 2017; Fearnhead and Rigaill, 2019) cannot be applied to this problem because they assume univariate data or do not allow for dependent observations. As Figure 7.13 shows, robust BOCPD finds one CP corresponding to the introduction of the congestion charge,

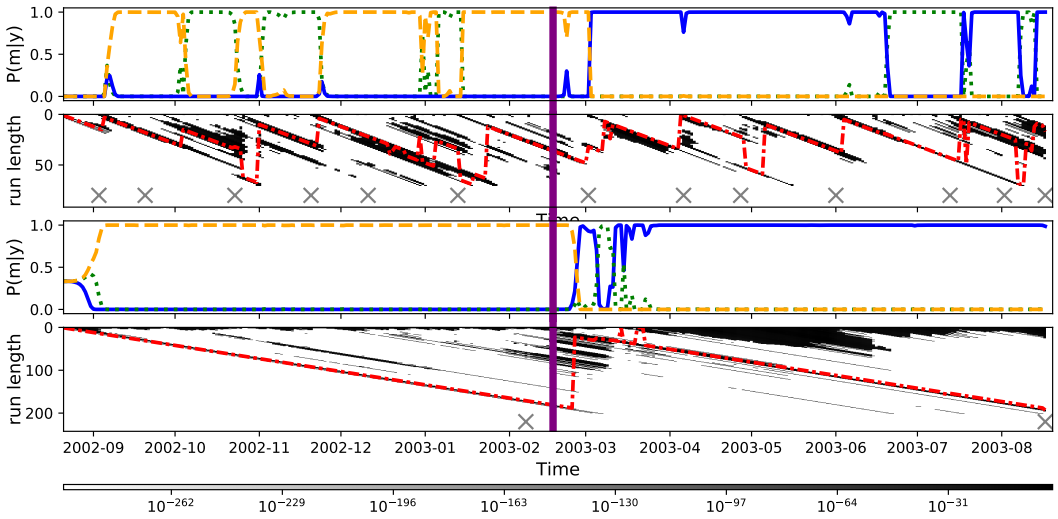


Figure 7.13: On-line model posteriors for three different VAR models (solid, dashed, dotted) and run-length distributions in grayscale with most likely run-lengths dashed for standard (top two panels) and robust (bottom two panels) BOCPD. Also marked are the congestion charge introduction, 17/02/2003 (solid vertical line) and the MAP segmentations (crosses)

while standard BOCPD produces a false discovery rateof more than 90%. Both methods find a change in dynamics (i.e. models) after the congestion charge introduction, but variance in the model posterior is substantially lower for the robust algorithm. Further, it increases the average one-step-ahead predictive likelihood by 10% compared to standard BOCPD.

Chapter 8

Robustness & Computational Convenience for Intractable Likelihoods

Summary: In this chapter, we study how the RoT can be used for the benefit of intractable likelihood models. In standard Bayesian inference, likelihood functions with intractable normalization constants incur severe computational disadvantages. By using a computationally beneficial instantiation of Stein's Method and in particular Kernel-Stein Discrepancies (KSDs) as loss functions, we circumnavigate this challenge entirely. In fact, in a wide range of settings we manage to obtain closed form posteriors for problems that would lead to doubly intractable posteriors under the standard Bayesian paradigm. Further, we show that by appropriate choice of the kernel, one can straightforwardly impart robustness on what we will call KSD-Bayes posteriors. KSD-Bayes posteriors can be written in terms of the RoT, and belong to the subclass of Gibbs posteriors. This means that they have an analytic form, which enables us to prove a range of regularity conditions that include consistency, asymptotic normality, and bias-robustness.

As we have discussed throughout this thesis, a considerable proportion of statistical modelling deviates from the idealised approach of fine-tuned, expertly-crafted descriptions of real-world phenomena. If one proceeds by naively applying Bayes' Rule, this leads to miscalibrated posteriors that concentrate at undesirable regions in the parameter space. As we have seen in the previous two chapters, one way of rectifying this is by changing how the model's parameters are scored, affecting

how "good" parameter values are discerned from "bad" ones. The current chapter considers this situation in the presence of intractable likelihoods: not only are our models of the world misspecified, but they also are of the form $p(x|\boldsymbol{\theta}) = q(x, \boldsymbol{\theta})/Z(\boldsymbol{\theta})$, where $q(x, \boldsymbol{\theta})$ is an analytically tractable—but un-normalized—function, and $Z(\boldsymbol{\theta})$ is an *intractable* normalization constant. Standard Bayesian posteriors resulting from such intractable likelihood models are sometimes called *doubly intractable*, since *two* unknown normalizers—that of the likelihood and that of the posterior itself—have to be tackled for inference (Iain Murray et al., 2006). Notably, virtually all standard Markov chain Monte Carlo (MCMC) methods cannot be used in this setting: they typically require explicit evaluation of the likelihood. Doubly intractable posteriors are not a fringe phenomenon either, and appear in many important statistical applications. This includes spatial models (Julian Besag, 1974, 1986; Peter J. Diggle, 1990), exponential random graph models (Jaewoo Park and Murali Haran, 2018), models for gene expression (Jiang et al., 2021), or hidden Potts models for satellite data (Moores et al., 2020).

In this chapter, we propose an inference approach based on the RoT for intractable likelihoods. Specifically, we employ a loss function based on a *Stein discrepancy* (Gorham and Mackey, 2015) and in particular on the minimum Stein discrepancy estimators of Barp et al. (2019). We focus on the Kernel-Stein discrepancy (KSD), and we call the resulting generalised Bayesian approach *KSD-Bayes*. In addition to dealing with intractable likelihoods, we also prove that KSD-Bayes posteriors provide robustness against misspecification, allow for a form of conjugacy that allows us to compute them in closed form, and satisfy numerous desirable theoretical properties—including frequentist consistency and Bernstein-von-Mises type results.

8.1 Background

First we provide a short summary of the relevant standing assumptions, and the notation used throughout the current chapter. Because of the substantial technical developments required for the proofs of this chapter, a lot of the notation required will be more precise and technical than in preceeding chapters.

Standing Assumption 1: The topological space $\mathcal{X}$ in which the data are contained, is locally compact and Hausdorff. The set $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$, in which parameters are contained, is Borel.

8.1.1 Notation

Measure theoretic notation: For a locally compact Hausdorff space such as $\mathcal{X}$, we let $\mathcal{P}(\mathcal{X})$ denote the set of all Borel probability measures on $\mathcal{X}$. A point mass at x is denoted $\delta_x \in \mathcal{P}(\mathcal{X})$. Similarly to what we have done in previous chapters, if $\mathcal{X}$ is equipped with a reference measure (such as the Lebesgue measure if $\mathcal{X} = \mathbb{R}^d$), then we abuse notation by writing $p \in \mathcal{P}(\mathcal{X})$ to indicate that the distribution with probability density function (p.d.f.) p is an element of $\mathcal{P}(\mathcal{X})$. As in previous chapters, we use $\mathbb{P}_{\boldsymbol{\theta}} \in \mathcal{P}(\boldsymbol{\Theta})$ as the measure induced by the p.d.f $p(\cdot|\boldsymbol{\theta})$; and we use $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ to denote the empirical measure on $\mathcal{X}$ induced by the sample $x_{1:n}$. For $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, we occasionally overload notation by denoting by $L^q(\mathcal{X}, \mathbb{P})$ both the set of functions $f : \mathcal{X} \to \mathbb{R}$ for which $\|f\|_{L^q(\mathcal{X},\mathbb{P})} := (\int_{\mathcal{X}} |f|^q \mathrm{d}\mathbb{P})^{1/q} < \infty$ and the normed space in which two elements $f, g \in L^q(\mathcal{X}, \mathbb{P})$ are identified if they are $\mathbb{P}$-almost everywhere equal. As is common practice, if $\mathbb{P}$ is a Lebesgue measure, we simply write $L^q(\mathcal{X})$ instead of $L^q(\mathcal{X}, \mathbb{P})$. Let $\mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$ be the set of all Borel probability measures $\mathbb{P}$ supported on $\mathbb{R}^d$, admitting an everywhere positive p.d.f. p and continuous partial derivatives $x \mapsto (\partial/\partial x_{(i)})p(x)$.

Real analytic notation: The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|_2$. The set of continuous functions $f : \mathcal{X} \to \mathbb{R}$ is denoted $C(\mathcal{X})$. We denote by $C_b^1(\mathbb{R}^d)$ the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that both f and the partial derivatives $x \mapsto (\partial/\partial x_{(i)})f(x)$ are bounded and continuous on $\mathbb{R}^d$. We also denote by $C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d)$ the set of bivariate functions $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that both f and the partial derivatives $(x, x') \mapsto (\partial/\partial x_{(i)})(\partial/\partial x'_{(j)})f(x, x')$ are bounded and continuous on $\mathbb{R}^d \times \mathbb{R}^d$. For an arbitrary set $\mathcal{S}(\mathcal{X})$ of functions $f : \mathcal{X} \to \mathbb{R}$, denote by $\mathcal{S}(\mathcal{X}; \mathbb{R}^k)$ the set of $\mathbb{R}^k$-valued functions whose components belong to $\mathcal{S}(\mathcal{X})$. Let ∇ and $\nabla\cdot$ be the gradient and the divergence operators in $\mathbb{R}^d$. For functions with multiple arguments, we sometimes use subscripts to indicate the argument to which the operator is applied (e.g. $\nabla_x f(x, y)$). For f an $\mathbb{R}^d$-valued function, $[\nabla f(x)]_{(i,j)} := (\partial/\partial x_{(i)})f_{(j)}(x)$ and $\nabla \cdot f(x) := \sum_{i=1}^d (\partial/\partial x_{(i)})f_{(i)}(x)$. For f an $\mathbb{R}^{d\times d}$-valued function, $[\nabla f(x)]_{(i,j,k)} := (\partial/\partial x_{(i)})f_{(j,k)}(x)$ and $[\nabla \cdot f(x)]_{(i)} := \sum_{j=1}^d (\partial/\partial x_{(j)})f_{(i,j)}(x)$.

8.1.2 Stein Discrepancy & Kernel-Stein Discrepancy (KSD)

We refer to Chapter 6 for an overview of the Stein Discrepancy generally and the Kernel-Stein Disrepancy (KSD) in particular. Essentially, we saw via Proposition 6.1 that for the Stein operator $\mathcal{S}_{\mathbb{Q}} : \mathcal{H} \to L^1(\mathcal{X}, \mathbb{Q})$ on the Vector-Valued Reproducing Kernel Hilbert Space (v-RKHS) $\mathcal{H}$ induced by the (matrix-valued) kernel $K : \mathcal{X}^2 \to$

$\mathbb{R}$, the KSD is given by

$$\text{KSD}^2(\mathbb{Q}\|\mathbb{P}) := \mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_\mathbb{Q}\mathcal{S}_\mathbb{Q}K(X,X')\right]. \tag{8.1}$$

The necessary condition required to apply Proposition 6.1 is a mild continuity assumption on the Stein operator given below.

Assumption 8.1. Let $\mathcal{H}$ be a v-RKHS with kernel $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}^{d\times d}$. For $\mathbb{Q}\in\mathcal{P}(\mathcal{X})$, let $\mathcal{S}_\mathbb{Q}$ be a Stein operator with domain $\mathcal{H}$. For each fixed $x\in\mathcal{X}$, we assume $h \mapsto \mathcal{S}_\mathbb{Q}[h](x)$ is a continuous linear functional on $\mathcal{H}$. Further, we assume that $\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_\mathbb{Q}\mathcal{S}_\mathbb{Q}K(X,X)\right] < \infty$.

Note that it is straightforward to verify that $h \mapsto \mathcal{S}_\mathbb{Q}[h](x)$ is a continuous linear functional for each fixed $x\in\mathcal{X}$ once the form of $\mathcal{S}_\mathbb{Q}$ is specified; see Appendix C.3.1. If this condition is satisfied so that the KSD-based loss takes the form of (8.1), then for samples $x_{1:n}\sim\mathbb{P}$ and for $\mathbb{P}_n$ the empirical measure of $x_{1:n}$, this loss can be computed approximately as

$$\text{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x_i,x_j), \tag{8.2}$$

where the explicit form of $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K$ depends on $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$. In (6.7), we gave provided the example for the case where the Stein Operator in question is the Langevin Stein operator and $\mathcal{X}=\mathbb{R}^d$.

For completeness, we note here that in both (6.7) as well as the rest of the chapter, we will use the following notation: we denote the j-th column of $K(x,x')\in\mathbb{R}^{d\times d}$ by $K_{-,j}(x,x')\in\mathbb{R}^d$, we define $\mathcal{S}_\mathbb{Q}K(x,x') := [\mathcal{S}_\mathbb{Q}K_{-,1}(x,x'),\ldots,\mathcal{S}_\mathbb{Q}K_{-,d}(x,x')]\in\mathbb{R}^d$ where $\mathcal{S}_\mathbb{Q}K_{-,j}(x,x') := \mathcal{S}_\mathbb{Q}[K_{-,j}(\cdot,x')](x)$ is an action of $\mathcal{S}_\mathbb{Q}$ for the $\mathbb{R}^d$-valued function $K_{-,j}(\cdot,x')$ at each $x'\in\mathcal{X}$; and we define $\mathcal{S}_\mathbb{Q}\mathcal{S}_\mathbb{Q}K(x,x') := \mathcal{S}_\mathbb{Q}[\,\mathcal{S}_\mathbb{Q}K(x,\cdot)\,](x')$ as an action of $\mathcal{S}_\mathbb{Q}$ for the $\mathbb{R}^d$-valued function $\mathcal{S}_\mathbb{Q}K(x,\cdot)$ at each $x\in\mathcal{X}$.

8.2 The KSD-Bayes posterior

Suppose we are given a prior p.d.f. $\pi\in\mathcal{P}(\boldsymbol{\Theta})$ and a statistical model $\{\mathbb{P}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta}\in\boldsymbol{\Theta}\}\subset\mathcal{P}(\mathcal{X})$ with associated p.d.f. $p(\cdot|\boldsymbol{\theta})$. For a sample $x_{1:n}$ of independent observations generated from $\mathbb{P}$ and forming the empirical measure P_n, we define the KSD-*Bayes posterior* below.

Definition 8.1 (KSD-Bayes)**.** For each $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, select a Stein Operator $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$ and denote the associated Stein discrepancy $\text{SD}(\mathbb{P}_{\boldsymbol{\theta}}\|\cdot)$. Further, let the Stein Set $\mathcal{H}$ be

a v-RKHS associated with the kernel function K and let $w \in (0, \infty)$. Writing

$$L_{\mathrm{KSD}}(x_{1:n}, \boldsymbol{\theta}) = n \cdot \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n),$$

the KSD-Bayes posterior is given as

$$\begin{aligned} \pi_n^{\mathrm{KSD}}(\boldsymbol{\theta}) &= P(L_{\mathrm{KSD}}, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta})) \\ &\propto \pi(\boldsymbol{\theta}) \exp\left\{-w L_{\mathrm{KSD}}(x_{1:n}, \boldsymbol{\theta})\right\} \end{aligned} \tag{8.3}$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

At first glance, it may seem that there is an arbitrariness to using squared discrepancy (as opposed to another power of the discrepancy), but this choice turns out to be appropriate for the KSD in a meaningful way: in particular, it ensures that fluctuations of $L_{\mathrm{KSD}}(\boldsymbol{\theta})$ about its expectation are of order $\mathcal{O}(n^{-1/2})$—exactly like the standard log likelihood loss. Beyond this, it enables tractable computation (Section 8.3) and analysis (Section 8.4).

A more difficult question is how the weight w should be selected: So far, we have considered the w- and γ-divergences in the previous two chapters. For these divergences, we can recover the standard log likelihood as $w \to 1$ ($\gamma \to 1$), and indeed we chose values of w and γ very close to 1. This meant that we had approximately well-calibrated posteriors by just choosing $w = 1$, and did not have to worry about calibration all too much.

In contrast, there is no parameterisation for the KSD that lets us recover the standard log likelihood function. Consequently, we will have to find ways of choosing w to obtain (approximately) calibrated posteriors. This is an issue affecting many generalized posteriors, and we defer our discussion for the KSD-Bayes to Section 8.5.

8.3 Conjugate Inference

The KSD-Bayes posterior can be computed in closed form for an important special case—specifically for natural exponential family models with conjugate priors. Letting $\eta : \boldsymbol{\Theta} \to \mathbb{R}^k$ and $t : \mathcal{X} \to \mathbb{R}^k$ be any sufficient statistic for some $k \in \mathbb{N}$ and letting $a : \boldsymbol{\Theta} \to \mathbb{R}$ and $b : \mathcal{X} \to \mathbb{R}$, an exponential family model has probability mass function (p.m.f.) or p.d.f. (with respect to an appropriate reference measure on $\mathcal{X}$) of the form

$$p(x|\boldsymbol{\theta}) = \exp(\eta(\boldsymbol{\theta}) \cdot t(x) - a(\boldsymbol{\theta}) + b(x)). \tag{8.4}$$

This includes a wide range of distributions with an intractable normalization constant $\exp(a(\boldsymbol{\theta}))$, used in statistical applications such as random graphs (Yang et al., 2015), spin glass models (Julian Besag, 1974) or the kernel exponential family model (Canu and Smola, 2006). The model in (8.4) is called *natural* whenever the canonical parametrisation $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$ is used.

Proposition 8.1. Consider $\mathcal{X} = \mathbb{R}^d$ and the Langevin Stein operator $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$ in (6.4), where $\mathbb{P}_{\boldsymbol{\theta}}$ is the exponential family in (8.4), and a kernel $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$. Assuming the prior has a p.d.f. π, the KSD-Bayes posterior has a p.d.f.

$$\pi_n^{\mathrm{KSD}}(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\left(-wn\{\eta(\boldsymbol{\theta}) \cdot \Lambda_n \eta(\boldsymbol{\theta}) + \eta(\boldsymbol{\theta}) \cdot \nu_n\}\right),$$

where $\Lambda_n \in \mathbb{R}^{k\times k}$ and $\nu_n \in \mathbb{R}^k$ are defined as

$$\begin{aligned}
\Lambda_n &:= \frac{1}{n^2} \sum_{i,j=1}^{n} \nabla t(x_i) \cdot K(x_i, x_j) \nabla t(x_j), \\
\nu_n &:= \frac{1}{n^2} \sum_{i,j=1}^{n} \nabla t(x_i) \cdot \left(\nabla_{x_j} \cdot K(x_i, x_j)\right) + \\
&\qquad \nabla t(x_j) \cdot \left(\nabla_{x_i} \cdot K(x_i, x_j)\right) + 2\nabla t(x_i) \cdot K(x_i, x_j) \nabla b(x_j).
\end{aligned}$$

For a natural exponential family so that we have $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$, and a prior given by $\pi(\boldsymbol{\theta}) \propto \exp(-\frac{1}{2}(\boldsymbol{\theta} - \mu) \cdot \Sigma^{-1}(\boldsymbol{\theta} - \mu))$ for a positive definite matrix Σ leads

$$\pi_n^{\mathrm{KSD}}(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \mu_n) \cdot \Sigma_n^{-1}(\boldsymbol{\theta} - \mu_n)\right), \tag{8.5}$$

where $\Sigma_n^{-1} := \Sigma^{-1} + 2wn\Lambda_n$ and $\mu_n := \Sigma_n^{-1}(\Sigma^{-1}\mu - \nu_n)$.

The proof is in Appendix C.3.2. That the Gaussian distribution will be the conjugate prior for the KSD-Bayes posterior for *all* naturally parameterised likelihoods—even in the presence of an intractable likelihood—is remarkable, and a notable difference when compared to the classical Bayesian case.

That being said, it is well known that certain minimum discrepancy estimators, such as the score matching estimator (Aapo Hyvärinen, 2005) and the minimum KSD estimator (Barp et al., 2019), have closed forms in the case of an exponential family models; and the reasoning that has led us to Proposition 8.1 is similar to the reasoning required to obtain these closed forms.

8.4 Theoretical Properties

Before showcasing the practical utility of KSD-Bayes, we study its theoretical properties. The main results are posterior consistency and a Bernstein–von Mises theorem in Section 8.4.2. Beyond that, we derive a formal guarantee for global bias-robustness of KSD-Bayes in Section 8.4.3. In obtaining these results we have developed novel intermediate results concerning an important V-statistic estimator for KSD; these are also of independent interest, and so we present them in Section 8.4.1 rather than delegating them to the Appendix. Note that as throughout the rest of the thesis, all theory is valid for the misspecified regime: We do not assume that $\mathbb{P} \in \{\mathbb{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Moreover, the results in Section 8.4.1 and Section 8.4.2 hold for general data domains $\mathcal{X}$. For convenience, we set $w = 1$ throughout the theoretical derivations—but all results follow immediately for $w \neq 1$ simply by replacing K with wK.

Standing Assumption 2: The dataset $x_{1:n}$ consists of independent samples generated from $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, with empirical distribution $\mathbb{P}_n := (1/n)\sum_{i=1}^n \delta_{x_i}$. $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is open, convex and bounded. Section 8.1 holds with $\mathbb{Q} = \mathbb{P}_{\boldsymbol{\theta}}$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Note that assuming that $\boldsymbol{\Theta}$ is bounded is done merely for simplifying presentation. In particular, there is no loss of generality here, since we can always re-parametrize a likelihood function defined on an unbounded space to ensure that its re-parameterized version is defined on a bounded space.

Notation: For shorthand, let ∂^1, ∂^2 and ∂^3 denote the partial derivatives $(\partial/\partial\boldsymbol{\theta}_{(h)})$, $(\partial^2/\partial\boldsymbol{\theta}_{(h)}\partial\boldsymbol{\theta}_{(k)})$ and $(\partial^3/\partial\boldsymbol{\theta}_{(h)}\partial\boldsymbol{\theta}_{(k)}\partial\boldsymbol{\theta}_{(l)})$ for $h, k, l \in \{1, \ldots, p\}$, where to reduce notation the indices (h, k, l) are left implicit. The gradient and Hessian operators are $[\nabla_{\boldsymbol{\theta}}]_{(h)} = (\partial/\partial\boldsymbol{\theta}_{(h)})$ and $[\nabla^2_{\boldsymbol{\theta}}]_{(h,k)} = (\partial^2/\partial\boldsymbol{\theta}_{(h)}\partial\boldsymbol{\theta}_{(k)})$.

8.4.1 Minimum KSD Estimators

First we present a novel analysis for the V-statistic in (8.2). Note that a U-statistic estimator for the KSD was analysed in Barp et al. (2019) for the so-called *diffusion* Stein operator (which itself can be seen as a generalization of the Langevin-Stein operator). In contrast, our results for the V-statistic do not depend on a specific form of $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$, and therefore are of independent interest.

It is well-known that V-statistics exhibit finite-sample bias as estimators. For our case however, this bias vanishes in the big data limit so that we can derive the following consistency result:

Lemma 8.1 (a.s. pointwise convergence)**.** For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n) - \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}) \xrightarrow{a.s.} 0. \tag{8.6}$$

The proof is contained in Appendix C.3.3. If we impose further regularity conditions, we can strengthen the pointwise convergence to a uniform one. For this purpose, it will be convenient to introduce an Assumption indexed by some integer $r_{\max} \in \mathbb{N}$ as follows:

Assumption 8.2 ($r_{\max}$-regularity)**.** For all r so that $0 \leq r \leq r_{\max}$, it holds that

(1) the map $\boldsymbol{\theta} \mapsto \partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h](x)$ exists and is continuous, for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$;

(2) $h \mapsto (\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})[h](x)$ is a continuous linear functional on $\mathcal{H}$, for each $x \in \mathcal{X}$;

(3) $\mathbb{E}_{X \sim \mathbb{P}}[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}((\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(X, X))] < \infty$,

where $(\partial^0 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}) := \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$. Note that (2) with $r = 0$ is automatically true by virtue of Standing Assumption 2.

As with $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x)$, the first and second $(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})$ are applied to the first and second argument of K respectively in the expression $(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(X, X)$ above. While these assumptions may seem arcane, they become clearly interpretable and concrete once a specific Stein operator is chosen. We showcase this with the Langevin Stein operator in Appendix C.3.1.

Lemma 8.2 (a.s. Uniform Convergence)**.** Suppose Assumption 8.2 holds for $r_{\max} = 1$. Then,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left| \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n) - \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}) \right| \xrightarrow{a.s.} 0. \tag{8.7}$$

The proof is deferred to Appendix C.3.3.

While this result shows us that the loss can be estimated well in a uniformly good way over $\boldsymbol{\Theta}$, it does not tell us anything about the minimizers of these estimated losses. This is what the next results take care of: they concern consistency and asymptotic normality of the estimator $\boldsymbol{\theta}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n)$—the minimizer for the V-statistic in (8.2). Before we can analyze the minimizer of course, we will have to assume it and its desired limit exist.

Assumption 8.3. There exist minimisers $\boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n)$ for all sufficiently large $n \in \mathbb{N}$. Further, there exists a unique $\boldsymbol{\theta}_*$ s.t. $\mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}_*} \| \mathbb{P}) < \inf_{\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \geq \varepsilon\}} \mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P})$ for any $\varepsilon > 0$.

For the well-specified case where $\exists \boldsymbol{\theta}_0$ such that $\mathbb{P}_{\boldsymbol{\theta}_0} = \mathbb{P}$, the uniqueness of $\boldsymbol{\theta}_*$ holds automatically if KSD is a proper divergence—i.e. $\text{KSD}(\mathbb{P}\|\mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$. For example, if the mild regularity conditions of Barp et al. (2019, Proposition 1) are satisfied and the parametrisation $\boldsymbol{\theta} \mapsto \mathbb{P}_{\boldsymbol{\theta}}$ is injective, the minimum is uniquely attained. Generally speaking, it is often reasonable to assume that the minimizers exist and are unique—both for minimizers in finite samples ($\boldsymbol{\theta}_n$) and in the population-sense ($\boldsymbol{\theta}_*$).

Lemma 8.3 (Strong Consistency)**.** Suppose Assumption 8.2 holds for $r_{\max} = 1$, and that Assumption 8.3 also holds. Then,

$$\boldsymbol{\theta}_n \xrightarrow{a.s.} \boldsymbol{\theta}_*. \tag{8.8}$$

The proof is deferred to Appendix C.3.3.

While the previous result tells us that the minimizer behaves as we would like in the large data limit, it tells us nothing about the rate (in n) at which this happens. This is addressed next, and we establish asymptotic normality of $\boldsymbol{\theta}_n$ (and thereby a $\sqrt{n}$-rate of convergence) that holds if we impose a small number of additional regularity conditions.

Lemma 8.4 (Asymptotic Normality)**.** Suppose Assumption 8.2 holds for $r_{\max} = 3$, and that Assumption 8.3 holds. Let $H_* := \nabla^2_{\boldsymbol{\theta}} \text{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*}$ and $J_* := \mathbb{E}_{X\sim\mathbb{P}}[S(X,\boldsymbol{\theta}_*)S(X,\boldsymbol{\theta}_*)^\top]$, where we define the column vector

$$S(x,\boldsymbol{\theta}) := \mathop{\mathbb{E}}_{X\sim\mathbb{P}}[\nabla_{\boldsymbol{\theta}}(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x,X))].$$

If H_* is non-singular,

$$\sqrt{n}\,(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) \xrightarrow{d} \mathcal{N}(0, H_*^{-1}J_*H_*^{-1}),$$

where $\xrightarrow{d}$ denotes the convergence in distribution.

The proof is given in Appendix C.3.3. These preliminaries on minimum-KSD estimation are required for our main asymptotic results on KSD-Bayes; which we present next.

8.4.2 Posterior Consistency and Bernstein-von-Mises

Armed with the technical results of Section 8.4.1, we can now establish frequentist consistency of KSD-Bayes and a Bernstein–von Mises result. Our consistency result

requires a *prior mass condition*, similar to that of Chérief Abdellatif and Alquier (2020):

Assumption 8.4. The prior is assumed to

1. admit a p.d.f. π that is continuous at $\boldsymbol{\theta}_*$, with $\pi(\boldsymbol{\theta}_*) > 0$;
2. satisfy $\int_{B_n(\alpha_1)} \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \geq e^{-\alpha_2 \sqrt{n}}$ for some constants $\alpha_1, \alpha_2 > 0$,

where we define $B_n(\alpha_1) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : |\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}) - \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}_*} \| \mathbb{P})| \leq \alpha_1 / \sqrt{n}\}$.

Assumption 8.4 specifies the amount of prior mass in a neighbourhood around the population-optimal value $\boldsymbol{\theta}_*$ that is required. This is not a strong assumption and Appendix C.3.7 demonstrates how Assumptions 8.3 ,8.2, and 8.4 can be verified in the case of an exponential family model.

Theorem 8.1 (Posterior Consistency)**.** Suppose Assumptions 8.3 and 8.4 holds. Let $\sigma(\boldsymbol{\theta}) := \mathbb{E}_{X \sim \mathbb{P}} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X) \right]$. Then, for all $\delta \in (0, 1]$,

$$\begin{aligned} &\mathbb{P}\left(\left| \int_{\boldsymbol{\Theta}} \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}) \pi_n^{\mathrm{KSD}}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}_*} \| \mathbb{P}) \right| > \delta \right) \\ &\leq \frac{\alpha_1 + \alpha_2 + 8 \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sigma(\boldsymbol{\theta})}{\delta \sqrt{n}} \end{aligned} \tag{8.9}$$

where the probability is with respect to realisations of the dataset $x_{1:n} \overset{i.i.d.}{\sim} \mathbb{P}$.

The proof is deferred to Appendix C.3.4.

While the last result shows that the posterior will ultimately collapse to a point mass at the desired point in the parameter space $\boldsymbol{\Theta}$, it does not tell us at which speed this collapse occurs. To investigate this speed, we now derive a Bernstein–von Mises result; which will show that the convergence happens at rate $\sqrt{n}$. The pioneering work of Hooker and Vidyashankar (2014) and Ghosh and Basu (2016) established Bernstein–von Mises results for Gibbs (or pseudo-) posteriors of the form $P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$ based on losses derived from the family of α- and w-divergences. Unfortunately, the form of KSD is rather different—a V-statistic instead of an average—and so different theoretical tools are required to tackle it. To this end, we turn to Jeffrey W. Miller (2021), who introduced a general approach to deriving Bernstein–von Mises results for Gibbs (or pseudo-) posteriors $P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, demonstrating how the assumptions can be verified for several additive loss functions L. Our proof builds on Jeffrey W. Miller (2021), demonstrating that the required assumptions can also be satisfied by the non-additive KSD loss function in (8.2).

Theorem 8.2 (Bernstein–von Mises)**.** Suppose Assumption 8.2 holds for $r_{\max} = 3$, and that Assumptions 8.3, and part (1) of 8.4 hold. Let $\hat{\pi}_n^{\mathrm{KSD}}$ the p.d.f. of the random variable $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_n)$ for $\boldsymbol{\theta} \sim \pi_n^{\mathrm{KSD}}$, viewed as a p.d.f. on $\mathbb{R}^p$. Let $H_* := \nabla^2_{\boldsymbol{\theta}} \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*}$. If H_* is nonsingular,

$$\int_{\mathbb{R}^p} \left| \hat{\pi}_n^{\mathrm{KSD}}(\boldsymbol{\theta}) - \frac{1}{\det(2\pi H_*^{-1})^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta} \cdot H_* \boldsymbol{\theta}\right) \right| \mathrm{d}\boldsymbol{\theta} \xrightarrow{a.s.} 0, \tag{8.10}$$

where the a.s. convergence happens as $n \to \infty$ with respect to $\mathbb{P}$.

The proof can be found in Appendix C.3.5. The theoretical results we have derived are highly encouraging: they indicate that KSD-Bayes posteriors in many ways behave like standard Bayes posteriors. We note also that the asymptotic precision matrix H_* from Theorem 8.2 differs to the precision matrix $H_* J_*^{-1} H_*$ of the minimum KSD estimator from Lemma 8.4 which would give us correct frequentist coverage. This is precisely analogous to fact that Bayesian credible sets can have asymptotically incorrect frequentist coverage if the statistical model is misspecified (Kleijn and van der Vaart, 2012), a point we will be addressing in Section 8.5.2.

8.4.3 Global Bias-Robustness of KSD-Bayes

While the main appeal of KSD-Bayes are its computational advantages in the presence of intractable normalization constants, the choice of kernel also enables us to make the KSD-Bayes robust to contamination in the dataset. Since the KSD-Bayes posterior takes the form $P(w \cdot L_{\mathrm{KSD}}, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$, it is a Gibbs posterior and can be written out in analytical form. Unlike most posteriors derived from the RoT, this allows us to establish its robustness using a fairly standard mathematical toolbox.

To this end, consider the *ε-contamination model* $\mathbb{P}_{n,\varepsilon,y} = (1-\varepsilon)\mathbb{P}_n + \varepsilon\delta_y$, where $y \in \mathcal{X}$ and $\varepsilon \in [0,1]$ (see Huber, 2011). In words, the data point y is considered to be a contaminant relative to the dataset $x_{1:n}$. Robustness to this form of contamination in the Bayesian setting has been considered in Hooker and Vidyashankar (2014); Ghosh and Basu (2016); Tomoyuki Nakagawa and Shintaro Hashimoto (2020), and we will build on this previous literature in what follows. For this, it will be convenient to overload notation and define

$$L_{\mathrm{KSD}}(\boldsymbol{\theta}; \mathbb{P}_n) := L_{\mathrm{KSD}}(\boldsymbol{\theta}, x_{1:n}).$$

With this in hand and following Ghosh and Basu (2016), we then consider a KSD-Bayes posterior based on the (contaminated) loss $L(\boldsymbol{\theta}; \mathbb{P}_{n,\varepsilon,y})$ with corresponding

density $\pi_n^{L(\boldsymbol{\theta};\mathbb{P}_{n,\varepsilon,y})}$, and define the *posterior influence function* as

$$\mathrm{PIF}(y, \boldsymbol{\theta}, \mathbb{P}_n) := \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \pi_n^{L(\boldsymbol{\theta};\mathbb{P}_{n,\varepsilon,y})}|_{\varepsilon=0}. \tag{8.11}$$

Note that unlike the influence functions we have seen in Chapter 6, this influence function is defined for every point in the parameter space; and so additionally depends on $\boldsymbol{\theta}$. We call a KSD-Bayes posterior *globally bias-robust* if

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{y\in\mathcal{X}} |\,\mathrm{PIF}(y, \boldsymbol{\theta}, \mathbb{P}_n)| < \infty,$$

meaning that the sensitivity of the generalised posterior to the contaminant y is limited.

In fact, this notion of robustness can be applied not only to the KSD-Bayes posterior, but to any Gibbs posterior $\pi_n^L = P(L, \mathrm{KLD}, \mathcal{P}(\boldsymbol{\Theta}))$ for a loss that acts on $x_{1:n}$ only through $\mathbb{P}_n$. The following lemma provides sufficient conditions for which global bias-robustness holds for any Gibbs posterior of this form:

Lemma 8.5. Let π_n^L be a generalised Bayes posterior for a fixed $n \in \mathbb{N}$ with a loss $L(\boldsymbol{\theta}; \mathbb{P}_n)$ and a prior π. Suppose $L(\boldsymbol{\theta}; \mathbb{P}_n)$ is lower-bounded and $\pi(\boldsymbol{\theta})$ is upper-bounded over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, for any $\mathbb{P}_n$. Denote $\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n) := (\mathrm{d}/\mathrm{d}\varepsilon) L(\boldsymbol{\theta}; \mathbb{P}_{n,\varepsilon,y})|_{\varepsilon=0}$. Then π_n^L is globally bias-robust if, for any $\mathbb{P}_n$,

1. $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{y\in\mathcal{X}} |\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)|\, \pi(\boldsymbol{\theta}) < \infty$, and
2. $\int_{\boldsymbol{\Theta}} \sup_{y\in\mathcal{X}} |\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)|\, \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty$.

The proof is deferred to Appendix C.3.6. It is interesting—albeit unsurprising—to note that standard Bayesian inference does not satisfy the conditions of Lemma 8.5 in general. Indeed, when $L(\boldsymbol{\theta}; \mathbb{P}_n) = n \cdot \mathbb{P}_n(-\log p(\cdot|\boldsymbol{\theta}))$ is the negative log likelihood as in the standard Bayesian case, $\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n) = \log p(y|\boldsymbol{\theta}) - \sum_{i=1}^n \log p(x_i|\boldsymbol{\theta})$, and the term $\log p(y|\boldsymbol{\theta})$ will generally be unbounded over $y \in \mathcal{X}$. This can occur even if the statistical model is not heavy-tailed, e.g. for a normal location model $p(\cdot|\boldsymbol{\theta})$ on $\mathcal{X} = \mathbb{R}^d$. In contrast, the kernel K in KSD-Bayes provides a degree of freedom which can be leveraged to ensure that the conditions of Lemma 8.5 *are* satisfied. The specific form of $\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)$ for KSD-Bayes is derived in Appendix C.3.6, and allows us to derive sufficient conditions on K for global bias-robustness of KSD-Bayes. These conditions are summarized in the following result.

Theorem 8.3 (Globally Bias-Robust)**.** For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, let $\mathbb{P}_{\boldsymbol{\theta}} \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$ and let $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$ denote the Langevin Stein operator of (6.4). Let $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$.

Suppose that π is bounded over $\boldsymbol{\Theta}$. If there exists a function $\gamma : \boldsymbol{\Theta} \to \mathbb{R}$ such that

$$\sup_{y \in \mathbb{R}^d} \Big(\nabla_y \log p_{\boldsymbol{\theta}}(y) \cdot K(y,y) \nabla_y \log p_{\boldsymbol{\theta}}(y) \Big) \leq \gamma(\boldsymbol{\theta}) \tag{8.12}$$

and, in addition, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\pi(\boldsymbol{\theta})\gamma(\boldsymbol{\theta})| < \infty$ and $\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta})\gamma(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty$, then KSD-Bayes is globally bias-robust.

The proof is outlined in Appendix C.5. The conditions of Theorem 8.3 can be satisfied through an appropriate choice of kernel K. We discuss these choices next. Unsurprisingly, the choice of kernel amounts to a robustness-efficiency trade-off, and the resulting differences in KSD-Bayes inference is explored empirically later on.

8.5 Setting Hyperparameters

Having thoroughly explored the theory of KSD-Bayes, we now turn to experimental evidence that confirms the findings of the previous section. Before doing so, we first discuss how we will set the relevant hyperparameters.

8.5.1 Setting $\mathcal{S}_{\mathcal{P}_\theta}$ and K

For Euclidean domains $\mathcal{X} = \mathbb{R}^d$, we advocate the default use of the Langevin Stein operator $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$ in (6.4) together with the kernel

$$K(x, x') = \frac{M(x) M(x')^\top}{(1 + (x - x')^\top \Sigma^{-1} (x - x'))^\gamma}, \tag{8.13}$$

where Σ is a positive definite matrix, $\gamma \in (0, 1)$ is a constant, and $M \in C_b^1(\mathbb{R}^d; \mathbb{R}^{d \times d})$ is a matrix-valued weighting function. Note that using a non-constant weighting function is equivalent to replacing the Langevin Stein operator with the *diffusion* Stein operator based on a diffusion matrix $M(x)$ as introduced by Gorham et al. (2019). For the specific choice $M(x) = I_d$, (8.13) is well-known under the name of an *inverse multi-quadratic* (IMQ) kernel.

Both the IMQ kernel and the Langevin Stein operator have appealing properties in the context of the KSD. Firstly, under mild conditions on $\mathbb{P}$, $\mathrm{KSD}(\mathbb{P} || \mathbb{P}_n) \to 0$ implies that $\mathbb{P}_n$ converges weakly to $\mathbb{P}$ (Chen et al., 2019, Theorem 4). This convergence control has conceptual relevance for our context. In particular, it ensures that small values of $\mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}} || \mathbb{P}_n)$ imply similarity between $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_n$ in the topology of weak convergence. We note that many other common kernels (e.g., Gaussian or Matérn kernels) fail to provide convergence control (Jackson Gorham and Lester

Mackey, 2017, Theorem 6). Secondly—and on a more practical level—the combination of the diffusion Stein operator and IMQ kernel with $\gamma = 1/2$ has been found to perform reliably in practice (Chen et al., 2019; Riabiz et al., 2021). In line with this, we will be treating $\gamma = 1/2$ as the de-facto default value for the IMQ and diffusion kernel.

Having fixed the general form of K and its hyperparameter γ, we are left with the weighting function $M(x)$ as the last unspecified degree of freedom in our methodology. To set $M(x)$, note that it can be chosen to regulate the efficiency-robustness trade-off: If global bias robustness is *not* required, then we recommend setting $M(x) = I_d$ as a default. Conversely, if global bias-robustness *is* required, one should select $M(x)$ such that the supremum in (8.12) exists and the preconditions of Theorem 8.3 are satisfied; see the worked examples in Section 8.6.

Lastly, while the theoretical analysis of Section 8.4 assumed that K is fixed, our experiments follow standard practice in the kernel methods community by using a data-adaptive choice of the matrix Σ. To this end, all of our experiments use the ℓ_1-regularised sample covariance matrix estimator of Ollila and Raninen (2019). The sensitivity of KSD-Bayes to the choice of kernel parameters is investigated in Section 8.6.

8.5.2 Setting w

In all our experiments, we found that the variance of the KSD-Bayes posterior with $w = 1$ is never smaller than that of the standard posterior. This provides a rough heuristic justification for a default choice of $w = 1$. For the case of a simple normal location model, this heuristic can be justified more formally, see Section 8.6.1. However, in a misspecified setting, smaller values of w are needed to avoid over-confidence in the KSD-Bayes posterior in the frequentist sense (see e.g. Pei-Shien Wu and Ryan Martin, 2020). Here, we aim to pick w to approximately calibrate the KSD-Bayes posterior. In other words, we will aim to match the scale of the asymptotic precision matrix for the KSD-posterior (namely, H_*; see Theorem 8.2) matches that of the minimum-KSD point estimator with correct frequentist coverage (namely, $H_* J_*^{-1} H_*$; see Lemma 8.4), an approach pioneered in Lyddon et al. (2019). While it is clear that no choice of w will make the asymptotic covariance matrices equal under misspecification, the hope is to at least approximately match the scales of both variance matrices. Naturally, this is complicated by the fact that $\mathbb{P}$ is unknown, so that one needs to estimate both H_* and J_* by their finite-sample equivalents H_n and J_n. With these estimates in hand, we propose the following

default for w:

$$w = \min(1, w_n) \qquad \text{where} \qquad w_n = \frac{\mathrm{tr}(H_n J_n^{-1} H_n)}{\mathrm{tr}(H_n)}, \tag{8.14}$$

where

$$\begin{aligned} H_n &:= \nabla^2_{\boldsymbol{\theta}} \, \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_n} \\ J_n &:= \frac{1}{n}\sum_{i=1}^{n} S_n(x_i, \boldsymbol{\theta}_n) S_n(x_i, \boldsymbol{\theta}_n)^\top, \\ S_n(x, \boldsymbol{\theta}) &:= \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} (\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x_i)). \end{aligned} \tag{8.15}$$

The minimum taken in (8.14) provides a safeguard against selecting a value of w that over-shrinks the posterior covariance matrix— a phenomenon that we observed for some of the experiments reported in the next section—due to poor quality of the numerical approximations H_n and J_n when n is small. The above expressions and estimators are derived for the exponential family model in Appendix C.3.7.

8.6 Experiments

Next, we present four distinct experiments. The first experiment, in Section 8.6.1, concerns a normal location model, allowing the standard posterior and the KSD-Bayes posterior to be compared and confirming our robustness results are meaningful. Section 8.6.2 presents a two-dimensional precision estimation problem, where standard Bayesian computation is challenging but computation with KSD-Bayes is available in closed form. Then, Section 8.6.3 presents a 25-dimensional kernel exponential family model, and Section 8.6.4 presents a 66-dimensional exponential graphical model; in both cases a Bayesian analysis has not, to-date, been attempted due to severe intractability of the likelihood. In addition, the kernel exponential family model allows us to explore a multi-modal dataset and to understand the potential limitations of KSD-Bayes in that context. For all experiments, we use the hyperparameter settings discussed in the previous section.

8.6.1 Normal Location Model

For pedagogical purposes, we first consider fitting a normal location model $\mathbb{P}_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}, 1)$ to a dataset $x_{1:n}$. Our aim is to illustrate the robustness properties of KSD-Bayes, and we therefore generate the dataset using a ε-contaminated normal

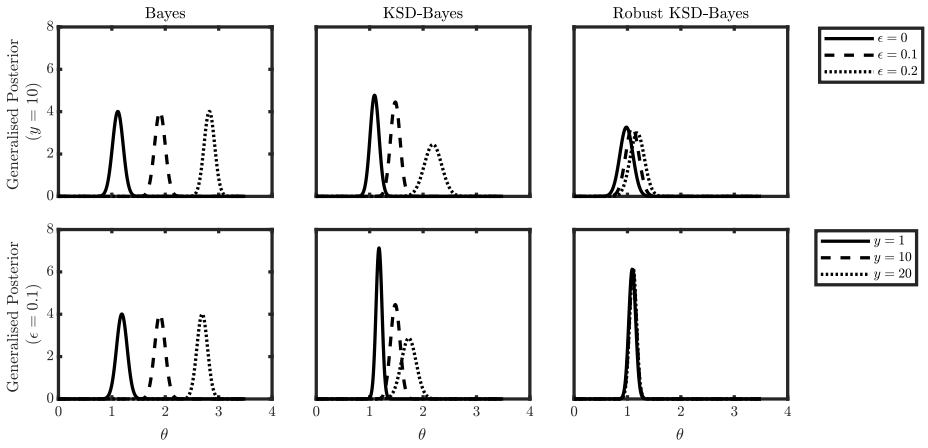


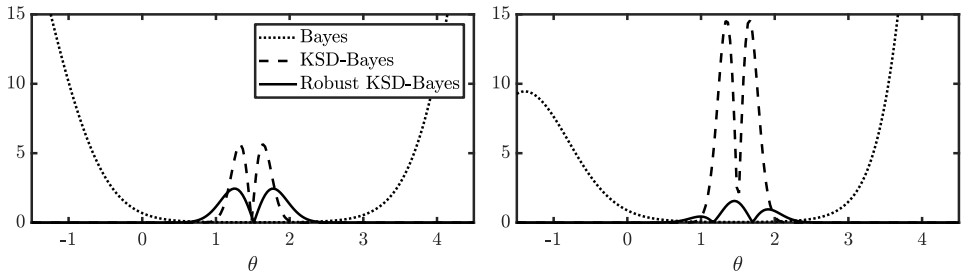
Figure 8.1: Standard Bayes and KSD-Bayes posteriors for the normal location model. The true parameter value is $\boldsymbol{\theta} = 1$, while a proportion ε of the data were contaminated by noise of the form $\mathcal{N}(y, 1)$. In the top row $y = 10$ is fixed and $\varepsilon \in \{0, 0.1, 0.2\}$ are considered, while in the bottom row $\varepsilon = 0.1$ is fixed and $y \in \{1, 10, 20\}$ are considered.

location model. Specifically, the true data-generating process is

$$(1 - \varepsilon)\mathbb{P}_{\boldsymbol{\theta}} + \varepsilon\mathbb{P}_y,$$

where we have set $\mathbb{P}_y = \mathcal{N}(y, 1)$. In this model, ε controls the extent and y the location of the contamination of $\mathbb{P}_{\boldsymbol{\theta}}$. Given data from this model, the aim is to make reliable inferences for $\boldsymbol{\theta}$ based on a contaminated dataset of size $n = 100$. The prior on $\boldsymbol{\theta}$ was $\mathcal{N}(0, 1)$.

The standard Bayesian posterior is depicted in the leftmost panels of Figure 8.1, for varying ε (top row) and varying y (bottom row). Straightforward calculation shows that the expected posterior mean is $\frac{n}{n+1}[\boldsymbol{\theta} + \varepsilon(y - \boldsymbol{\theta})]$, which increases linearly as either y or ε are increased, with the other fixed. This behaviour is also evident in the leftmost panels of Figure 8.1. The KSD-Bayes posterior is depicted in the central panels of Figure 8.1. As can be seen, it is far less sensitive to contamination compared to the standard Bayes posterior. Moreover, the variance slightly increases whenever either ε or y are increased, which is a result of estimating the weight w. In the rightmost panels of Figure 8.1, we display a robust version of the KSD-posterior using the weighting function $M(x) = (1 + x^2)^{-1/2}$. This choice will bound the influence of large values in the dataset, since $M(x)$ vanishes just fast enough as $|x| \to \infty$ to ensure that the bias-robustness conditions of Theorem 8.3 are satisfied. The effect is clear from the bottom right panel of Figure 8.1, where even



(a) $\boldsymbol{\theta} \mapsto |\,\mathrm{PIF}(y = 2.0, \boldsymbol{\theta}, \mathbb{P}_n)|$ (b) $\boldsymbol{\theta} \mapsto |\,\mathrm{PIF}(y = 20, \boldsymbol{\theta}, \mathbb{P}_n)|$

Figure 8.2: Posterior influence function for the normal location model.

for $y = 20$, the KSD-Bayes posterior remains centred very close to the true value $\boldsymbol{\theta} = 1$. While our theoretical results relate to y and do not guarantee robustness when ε is increased, the top right panel in Figure 8.1 suggests that the KSD-Bayes posterior is indeed robust in this regime as well. Figure 8.2 displays the posterior influence function (8.11) for this normal location model. It reveals that the standard Bayesian posterior is not bias-robust, since the tails of the posterior are highly sensitive to the contaminant y. In contrast, the tails of the KSD-Bayes posterior are insensitive to the contaminant y. This appears to be the case for both weighting functions, despite only one weighting function satisfying the conditions of Theorem 8.3, which is an indication that our conditions for robustness in the Theorem are much stricter than required.

8.6.2 Precision Parameters in an Intractable Likelihood Model

Our second experiment is due to Liu et al. (2019), and concerns an exponential family model $p(x|\boldsymbol{\theta}) = \exp(\boldsymbol{\theta} \cdot t(x) - a(\boldsymbol{\theta}) + b(x))$, where $\boldsymbol{\theta} \in \mathbb{R}^2$ are parameters to be inferred and $x \in \mathbb{R}^5$. The model specification is completed with

$$t(x) = (\tanh(x_{(4)}), \tanh(x_{(5)})),$$
$$b(x) = -0.5 \textstyle\sum_{i=1}^{5} x_{(i)}^2 + 0.6 x_{(1)} x_{(2)} + 0.2 \sum_{i=3}^{5} x_{(1)} x_{(i)}.$$

Despite the apparent simplicity of this model, the term $a(\boldsymbol{\theta})$, which determines the normalization constant, is analytically intractable and exact simulation from this data-generating model is not straightforward (except for the case where $\boldsymbol{\theta} = \mathbf{0}$). As a consequence, standard Bayesian analysis is not practical without, model-specific numerical methods—such as cubature rules to approximate the intractable normalization constant. In sharp contrast, the KSD-Bayes posterior is available in closed form for this model via Proposition 8.1. Our aim here is to assess robustness of

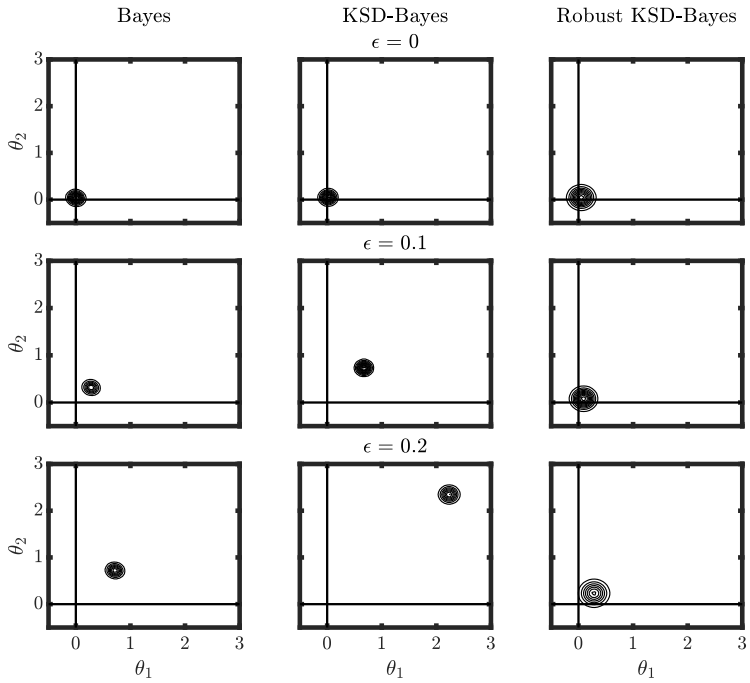


Figure 8.3: Standard Bayes posteriors and KSD-Bayes posteriors for the Liu et al. (2019) model. The true parameter value is $\boldsymbol{\theta} = 0$, while a proportion ε of the data were contaminated by being shifted by an amount $y = (10, 10)$.

the KSD-Bayes posterior, focusing on the setting where y is fixed and ε is increased, since this is the regime for which our theoretical results do *not* hold. A dataset of size $n = 500$ is generated from the model $\mathbb{P}_{\boldsymbol{\theta}}$ with true parameter $\boldsymbol{\theta} = (0, 0)$, so that $\mathbb{P}_{\boldsymbol{\theta}}$ has the form $\mathcal{N}(0, \Sigma)$ and can be sampled from exactly. Each datum x_i is, with probability ε, shifted to $x_i + y$ where $y = (10, \ldots, 10)$. The prior on $\boldsymbol{\theta}$ is $\mathcal{N}(0, 10^2 I)$.

The left column in Figure 8.3 displays the standard Bayes posterior[1], which is sensitive to contamination in the dataset—in much the same way as in the normal location model of Section 8.6.1. The KSD-Bayes posterior with $M(x) = I_d$ is depicted in the middle column of Figure 8.3, and is seen to actually be *more* sensitive to contamination compared to the standard Bayesian posterior, in the sense that the mean moves further from 0 as ε is increased. Finally, in the right column of Figure 8.3 we display a provably robust KSD-Bayes posterior obtained with weighting

[1]To obtain these results, the intractable normalization constant was approximated using a numerical cubature method. To do this, we recognise that $p(x|\boldsymbol{\theta}) = \mathcal{N}(x; 0, \Sigma) r_{\boldsymbol{\theta}}(x)/C_{\boldsymbol{\theta}}$ where $r_{\boldsymbol{\theta}}(x) = \exp(\boldsymbol{\theta}_1 \tanh(x_4) + \boldsymbol{\theta}_2 \tanh(x_5))$. Then $C_{\boldsymbol{\theta}} = \int r_{\boldsymbol{\theta}}(x) \mathrm{d}\mathcal{N}(x; 0, \Sigma)$, which is approximated using (polynomial order 10) Gauss-Hermite cubature in 2D.

function

$$M(x) = \text{diag}\Big((1 + x_{(1)}^2 + \cdots + x_{(5)}^2)^{-1/2}, (1 + x_{(1)}^2 + x_{(2)}^2)^{-1/2}, \\ \ldots, (1 + x_{(1)}^2 + x_{(5)}^2)^{-1/2}\Big),$$

As in the normal location model, this choice of $M(x)$ ensures the criteria for bias-robustness in Theorem 8.3 are satisfied. From the figure, we observe that the robustness guarantee of the Theorem is practically relevant and noticeable—even for the largest contamination proportion considered ($\varepsilon = 0.2$). We can also see that the KSD-Bayes posterior variance increases as ε does. At $\varepsilon = 0$, the spread of the robust KSD-Bayes posterior is almost twice that of the standard posterior, which is a reflection of the trade-off between robustness and efficiency inherent in the choice of K (via M).

8.6.3 Robust Nonparametric Density Estimation

Our third experiment concerns density estimation using the kernel exponential family, and explores the performance of KSD-Bayes when the dataset is multi-modal. Multi-modality is well-known to cause certain pathologies for minimum-KSD estimators (see Gorham et al. (Section 5.1 2019) and Wenliang (2020)); and here we study empirically if these pathologies carry over to the KSD-Bayes posterior. Let h denote a reference p.d.f. on $\mathbb{R}^d$, and let $\tilde{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a reproducing kernel. The *kernel exponential family* model relative to $\tilde{K}$ is given by (Canu and Smola, 2006)

$$p(x|\boldsymbol{f}) \propto h(x) \exp(\langle \boldsymbol{f}, \tilde{K}(\cdot, x)\rangle_{\mathcal{H}(\tilde{K})}), \tag{8.16}$$

and parametrised by f, an element of the v-RKHS $\mathcal{H}(\tilde{K})$. The normalization constant of (8.16) (if it exists) is typically an intractable function of $\boldsymbol{f}$. Due to this, there appears to be no Bayesian (or even generalised Bayesian) treatment of (8.16) in the literature. Indeed, we are not aware of any computational algorithm that would easily facilitate Bayesian inference for (8.16)—and so we will be unable to compare our KSD-Bayes procedure against standard Bayesian analysis. As the theory in this paper is finite-dimensional, we consider a finite-rank approximation of elements in $\mathcal{H}(\tilde{K})$ of the form $\boldsymbol{f}(x) = \sum_{i=1}^{p} \boldsymbol{\theta}_{(i)} \phi_{(i)}(x)$, with coefficients $\boldsymbol{\theta}_{(i)} \in \mathbb{R}$ and basis functions $\phi_{(i)} \in \mathcal{H}(\tilde{K})$, where we will take $\boldsymbol{\theta}$ to be p-dimensional for $p = 25$. Finite rank approximations have previously been considered for frequentist learning of kernel exponential families in (Strathmann et al., 2015; Danica J. Sutherland et al., 2018).

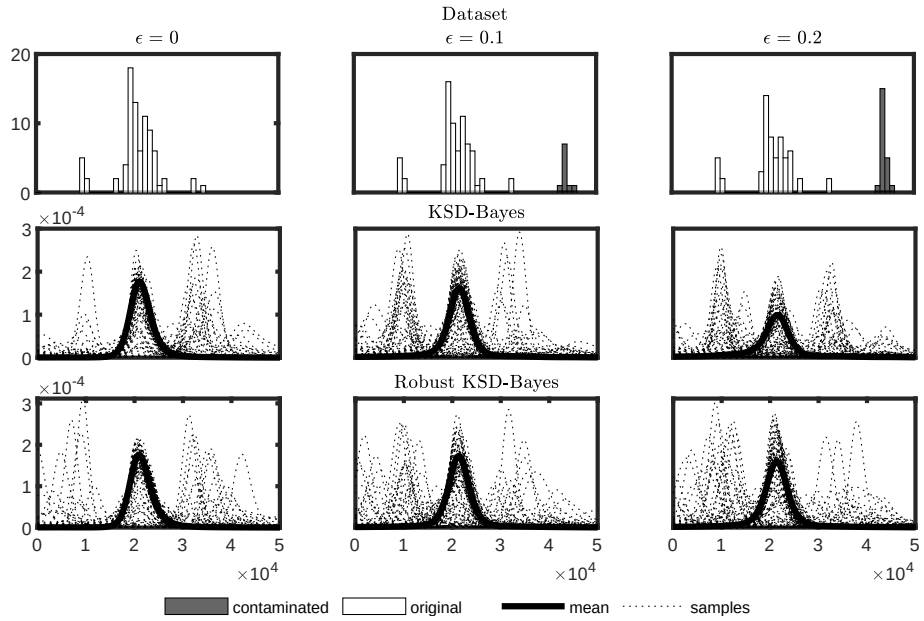


Figure 8.4: KSD-Bayes posteriors for the kernel exponential family model. A proportion ε of the data (top row) were contaminated.

In our case, the finite rank approximation ensures that any prior we induce on $\boldsymbol{f}$ via a prior on the coefficients $\boldsymbol{\theta}_{(i)}$ will be supported on $\mathcal{H}(\tilde{K})$. If one is interested in a well-defined limit as $p \to \infty$, then one will need to ensure a.s. convergence of the sum in this limit. If the ϕ_i are orthonormal in $\mathcal{H}(\tilde{K})$, and if the $\boldsymbol{\theta}_{(i)}$ are a priori independent, then $\mathbb{E}[\|\boldsymbol{f}\|^2_{\mathcal{H}(\tilde{K})}] = \sum_{i=1}^{p} \mathbb{E}[\boldsymbol{\theta}^2_{(i)}]$ so a sufficient condition, for example, is $\mathbb{E}[\boldsymbol{\theta}^2_{(i)}] = O(n^{-1-\delta})$ for some $\delta > 0$.

Our interest is in the performance of KSD-Bayes applied to a multi-modal dataset. To this end, we consider the galaxy data of Postman et al. (1986); Roeder (1990), comprising $n = 82$ velocities in km/sec of galaxies from 6 well-separated conic sections of a survey of the Corona Borealis. The data were whitened prior to computation, but results are reported with the original scale restored. For the kernel exponential family we use $h(x) = \mathcal{N}(0, 3^2)$ and the kernel $\tilde{K}(x, y) = \exp(-(x - y)^2/2)$, which ensures that (8.16) is normalizable due to Proposition 2 of Wenliang et al. (2019). For basis functions, we use $\phi_{(i+1)}(x) = (x^i/\sqrt{i!}) \exp(-x^2/2)$, $i = 0, \ldots, 24$, which are orthonormal in $\mathcal{H}(\tilde{K})$ (Steinwart et al., 2006). For our prior, we let $\boldsymbol{\theta}_{(i)} \sim \mathcal{N}(0, 10^2 i^{-1.1})$, which is weakly informative within the constraint of having a well-defined $p \to \infty$ limit. Our contamination model replaces a proportion ε of the dataset with values independently drawn from $\mathcal{N}(y, 0.1^2)$, with $y = 5$, shown as black bars in the top row of Figure 8.4.

The KSD-Bayes posterior with $M(x) = 1$ is displayed in the second row of Figure 8.4, with the bottom row presenting a robust KSD-Bayes posterior based on the weighting function $M(x) = (1 + x^2)^{-1/2}$, which ensures the conditions of Theorem 8.3 are satisfied. The results we present are for fixed y and increasing ε, since this regime is *not* covered by Theorem 8.3. The KSD-Bayes posterior mean is a uni-modal density, even though multi-modal densities are evident in sampled output. We attribute this to the insensitivity of KSD to mixture proportions, as discussed by Gorham et al. (Section 5.1 2019) and Wenliang (2020). Our results indicate that the robust weighting function reduces sensitivity to contamination in the dataset. Note in particular how the mass in the central mode of the KSD-Bayes posterior decreases when $\varepsilon = 0.2$, where the identity weighting function is used. Whether this insensitivity of KSD to well-separated regions in the dataset is desirable or not will depend on the application, but in this case it happens to be beneficial.

8.6.4 Network Inference with Exponential Graphical Models

Our final example concerns an exponential graphical model, representing negative conditional relationships among a collection of random variables $W = (W_1, \ldots, W_d)$, described in Yang et al. (2015, Sec. 2.5). The likelihood function is

$$p_{W|\boldsymbol{\theta}}(w|\boldsymbol{\theta}) \propto \exp\Big(-\sum_i \boldsymbol{\theta}_{(i)} w_{(i)} - \sum_{i<j} \boldsymbol{\theta}_{(i,j)} w_{(i)} w_{(j)} \Big), \tag{8.17}$$

where $w = (w_{(1)}, w_{(2)}, \ldots w_{(d)}) \in (0, \infty)^d$ and $\boldsymbol{\theta}_{(i)} > 0, \boldsymbol{\theta}_{(i,j)} \geq 0$. The total number of parameters is $p = d(d+1)/2$. Simulation from this model is challenging and the normalization constant is an intractable integral, so in what follows a standard Bayesian analysis is not attempted. Our aim is to fit (8.17) to a protein kinase dataset, mimicking an experiment presented by Yu et al. (2016) in the score-matching context. This dataset, originating in Sachs et al. (2005), consists of quantitative measurements of $d = 11$ phosphorylated proteins and phospholipids, simultaneously measured from single cells using a fluorescence-activated cell sorter, so the parameter $\boldsymbol{\theta}$ is 66-dimensional. Nine stimulatory or inhibitory interventional conditions were combined to give a total of $7,466$ cells in the dataset. The data were square-root transformed and samples containing values greater than 10 standard deviations from their mean were judged to be bona fide outliers and were removed. The remaining dataset of size $n = 7,449$ was normalized to have unit standard deviation. In most cases the measurement reflects the activation state of the kinases, and scientific interest lies in the mechanisms that underpin their interaction. Note that there is no scientific basis to expect only negative conditional dependencies

in the dataset; in this sense the model is likely to be misspecified. Our interest is in assessing the robustness properties of KSD-Bayes only, and no scientific conclusions will be drawn using this model. These mechanisms are often summarised as a protein signalling network, whose nodes are the d proteins and whose edges correspond to the pairs of proteins that interact. An important statistical challenge is to estimate a protein signalling network from such a dataset (Oates, 2013). However, it is known that existing approaches to network inference are non-robust in a very general sense, with the network inference community regularly highlighting the different conclusions drawn by different estimators applied to an identical dataset (Hill et al., 2016). Our interest is in determining whether networks inferred with KSD-Bayes posteriors are robust.

For our experiment, the variables $w_{(i)}$ were re-parametrised as $x_{(i)} := \log(w_{(i)})$, in order that they are unconstrained and $\mathbb{P}_{\boldsymbol{\theta}} \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$. For the contamination model, a proportion ε of the data were replaced with the fixed value $y = (10, \ldots, 10) \in \mathbb{R}^d$. Parameters were *a priori* independent with $\boldsymbol{\theta}_{(i)} \sim \mathcal{N}_{\mathrm{T}}(0, 1)$, $\boldsymbol{\theta}_{(i,j)} \sim \mathcal{N}_{\mathrm{T}}(0, 1)$, where $\mathcal{N}_{\mathrm{T}}$ is the Gaussian distribution truncated to the positive orthant of $\mathbb{R}^p$. Note that even though it is not a full normal, this prior is conjugate to the likelihood, as explained in Section 8.3, and allows the KSD-Bayes posterior to be computed in closed form. KSD-Bayes posteriors are produced both without and with the exponential weighting function $[M(x)]_{(i,i)} = \exp(-x_{(i)})$, the latter aiming to reduce sensitivity to large values in the dataset and coinciding with the identity weighting function at $x = 0$. From these, protein signalling networks were estimated using the s most significant edges, defined as the s largest values of $\bar{\boldsymbol{\theta}}_{(i,j)}/\sigma_{(i,j)}$, where the KSD-Bayes posterior marginal for $\boldsymbol{\theta}_{(i,j)}$ is $\mathcal{N}_{\mathrm{T}}(\bar{\boldsymbol{\theta}}_{(i,j)}, \sigma^2_{(i,j)})$. Results are shown in Figure 8.5; to optimise visualisation we report results for $s = 5$, though for other values of s similar conclusions hold. It is interesting to observe little agreement between the networks returned when the identity weighting function is used, which may reflect the difficulty of the network inference task. Reduced sensitivity to ε was observed when the exponential weighting function was used. In Figure 8.5 we report the number of edges that are consistent with the network reported in Sachs et al. (2005, Fig. 3A); the use of the exponential weighting function resulted in more edges being consistent with this benchmark network.

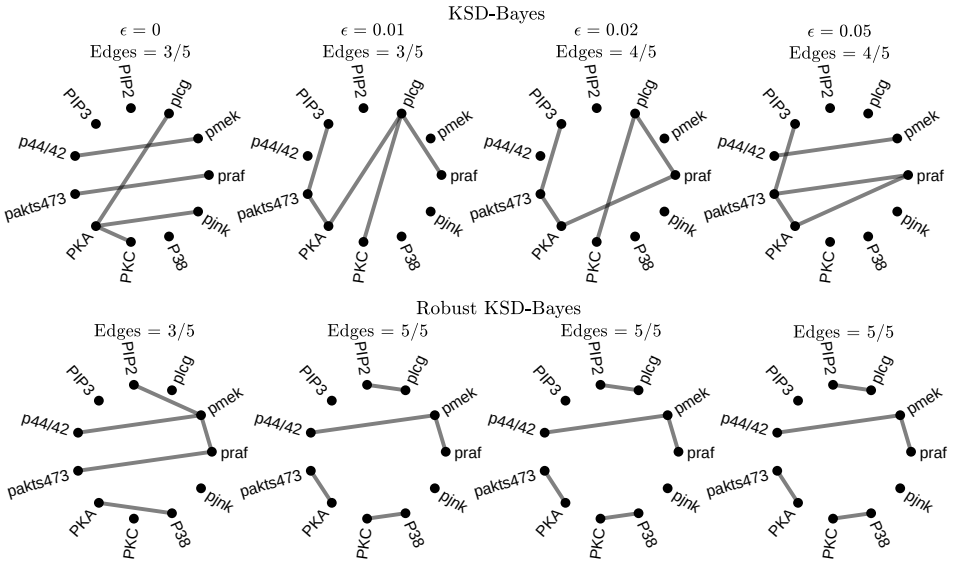


Figure 8.5: Exponential graphical model; estimated protein signalling networks as a function of the proportion ε of contamination in the dataset.

Part IV

Discussion & Appendix

Chapter 9

Discussion

In this last part of the thesis, we will first review the thesis' contributions to the field of optimization-centric posteriors and generalized posteriors more broadly in Section 9.1. In Section 9.2, we will explain some of the most important problems that remain.

9.1 Contributions of this thesis

Various generalizations of Bayesian posteriors have been developed to address the shortcomings of Bayes' Rule in the context of modern large-scale applications. This thesis contributes to the development of this branch of research in several ways: conceptually, theoretically, and methodologically. More precisely,

- Chapter 1 conceptually unified and generalized existing approaches to Bayes-like posteriors via an optimization-centric view on Bayesian methods that we call the Rule of Three (RoT). As its name suggests, the RoT allowed us to specify a belief distribution by making three independent choices: a loss, a prior regularizer, and a space over which to perform optimization. Intriguingly, each of these choices is suitable to address one of the three main assumptions associated with Bayesian inference that are often inappropriate in practice: The assumption of unlimited computational power, the assumption of a correctly specified likelihood model, and the assumption of a well-specified (or 'good') prior. The optimization-centric view on generalized Bayesian methods as expressed via the RoT is particularly appealing because it is axiomatically justified; and because it seamlessly connects both previous work on pseudo- and Gibbs-posteriors as well as that on variational methods.

- Chapter 2 answers fundamental questions about the class of generalized pos-

teriors introduced in Chapter 1. In particular, we show that under very mild regularity conditions, posteriors specified via the minimization problem underlying the RoT exist. Under slightly stronger conditions, the posterior is also unique. Beyond that, we also studied the dual form associated with the RoT. This provided invaluable insights into generalized Bayes posteriors, and enabled us to interpret them in a new light. Specifically, it allowed us to view them as adversarially robust procedures: in its dual form, the RoT took the form of an optimization problem that can be interpreted as a game. In this game, the statistician seeks to minimize a loss. Meanwhile, an adversary is allowed to perturb this loss. Crucially, the extent to which the adversary may perturb the loss depends on the choice of prior regularizer and the prior belief: together, they form a cost function that penalizes the adversary for changing the loss—with higher penalties in regions of the parameter space with large prior probability.

- Chapter 3 studied conditions under which the RoT produces posteriors that are consistent in the frequentist sense. While showing frequentist consistency for Gibbs distributions (or approximations thereof) is usually straightforward due to them being available analytically—at least up to a normalization constant—this is not true for general RoT posteriors. Notably, RoT posteriors are generally not available in any kind of closed form; and we therefore had to rely on more 'heavy-handed' tools from functional and variational analysis such as Γ-convergence. Accordingly, the proofs of this chapter really do not build on any previous work within the statistics community; and so none of the proof techniques are standard.

- Chapters 4–6 investigated Generalized Variational Inference (GVI)—one of the main methodological advances stemming from the development of the RoT. In a nutshell, GVI posteriors are the type of RoT posterior that is based on optimizing over a parameterized set of distributions; and can therefore always be (approximately) computed in practice. Chapter 4 compared GVI posteriors to standard variational inference (VI) posteriors; and illuminated how to adapt stochastic approximation techniques commonly used for VI for the computation of GVI posteriors. Chapter 5 investigated the effects of variations in the prior regularizer on the posterior. While there were a small number of limited theoretical results, most of our insights derived from empirical comparisons and confirmed that—unsurprisingly—more robust prior regularization yields posterior beliefs that are less susceptible to ill-informed prior beliefs. This was

shown for Bayesian mixture model, as well as for a Bayesian Neural Network—a black-box Bayesian Machine Learning model whose priors will invariably be ill-specified in practice. Lastly, Chapter 6 explored different robust model-based losses amenable to the GVI setting. The impact of choosing robust losses was then studied for the Deep Gaussian Process (DGP)—another black box Machine Learning model, whose likelihood should be assumed to be misspecified in most cases. The results showed that there is tangible merit to using robust losses; and the predictive performance of robust DGPs was consistently improved relative to the standard version.

- Chapter 7 applied the methodological toolkit developed in previous chapters to the setting of Bayesian On-line Changepoint Detection (BOCPD). On-line inference problems are particularly difficult to model—even more so in the presence of changepoints—and so it stands to reason that likelihoods in these algorithms will typically be misspecified. In fact, even in the canonical well-log data set, it had been common practice in previous work on on-line methods to pre-process this data set to remove outliers to avoid pathologies. We showed how RoT posteriors of the GVI families (based on robust losses derived from the β-divergence) could be used within BOCPD to eradicate these pathologies, and produce more robust and reliable statistical methods for BOCPD.

- Lastly, Chapter 8 used the versatility of the RoT to provide robustness and simplify computation in the context of intractable likelihood models. In particular, we showed that a loss based on the Kernel-Stein Discrepancy (KSD) provided not only robust posterior inferences, but also a significant computational advantage over the negative log likelihood loss associated with standard Bayesian inference: a KSD-Bayes posterior converts a doubly intractable standard Bayesian posterior into a much simpler problem. In fact, it even yields closed form posteriors in a range of settings that are of practical interest. While we studied this empirically, we also derived a number of theoretical results pertaining to the KSD-Bayes posterior's robustness as well as its asymptotic behaviour.

9.2 Open problems

While this thesis made several fundamental contributions to the field of generalized Bayesian methods, a number of key challenges remain for this collection of ideas, the most important of which we discuss below.

- **Choosing between posteriors:** While generalizing Bayesian inference can solve many problems, it raises a new problem. In particular, once Bayes' rule is abandoned as the guiding principle for performing posterior inferences, it is unclear which of the many possible alternatives one ought to choose. Note that this is *not* a model selection problem: the model selection problem revolves around the statistical model to be chosen for the data. Once this model is chosen and the prior is specified, the standard Bayes posterior follows without further any further decision. In contrast, in the generalized framework set out in this thesis, we do not even need a statistical model—a general loss function that ties the data to a parameter of interest suffices. Even in the case where we want to perform likelihood-based inference however, there are innumerable loss functions that connect a given statistical model to the data. In addition, we also have to specify a space over which to optimize the problem; and a divergence that dictates the influence the prior is allowed to have on the posterior. Throughout this thesis, we have motivated the choice of posterior through reasonable arguments—the loss should usually address concerns about model misspecification, the regularizer should be used to discount poorly-specified priors, and the space should be chosen in accordance with our computational budget—but we have not provided a general theoretically motivated recipe for making these choices. Instead, the choices throughout have been subjective. In some ways, this is somewhat dissatisfying—and certainly a missing ingredient to make the RoT a fully reliable practical tool.

- **Theory of Robustness:** A second challenge relates to quantification of robustness. While we can quantify robustness whenever the RoT posterior has an analytically available form (at least up to a normalization constant) as for instance in the theoretical analysis of Chapter 8, the standard methods for analyzing robustness are not applicable to general RoT posteriors without closed forms.[1] To advance theory of robustness in RoT posteriors, it will be necessary to overcome this hurdle and find tools of analysis that are applicable directly to optimization problems.

- **Computation:** Most RoT posteriors cannot be computed. In fact, apart from Gibbs (or pseudo-) posteriors or GVI posteriors, it is unclear how RoT posteriors should be computed in practice (apart from some naive discretization techniques that would be computationally prohibitive). To ensure more

[1] The only exception to this is the family of VI posteriors: since these posteriors can be interpreted as approximations of the standard Bayes posterior, we can sometimes prove that robustness of a posterior covers over to its approximation.

wide-spread practical use of the RoT, we will have to find ways of computing a wider class of optimization-centric posteriors. The first steps in this direction have arguably already been taken by Alquier (2021), which managed to derive the form of RoT posteriors that were prior-regularized with f-divergences.

Since the work on optimization-centric posteriors has just begun, many other open problems remain. This thesis has demonstrated that it is worth tackling these significant challenges: faced with the challenges of finding suitable priors, likelihoods, and methods of computation, it is of significant practical and theoretical merit to go beyond Bayes' Rule.

Appendix A

Additional Details

A.1 Γ-convergence

The following definition is adapted from Dal Maso (2012), and holds on general topological spaces X.

Definition A.1 (Γ-convergence)**.** We say that a function sequence $\{F_n\}_{n\in\mathbb{N}}$ with sequence members $F_n : X \to \overline{\mathbb{R}}$ Γ-converges to a function $F : X \to \overline{R}$ if the Γ-lower limit $l : X \to \overline{\mathbb{R}}$ and the Γ-upper limit $u : X \to \overline{\mathbb{R}}$ coincide. Here,

$$
\begin{aligned}
l(x) &= \sup_{U\in N(x)} \liminf_{n\to\infty} \inf_{y\in U} F_n(y) \\
u(x) &= \sup_{U\in N(x)} \limsup_{n\to\infty} \inf_{y\in U} F_n(y),
\end{aligned}
$$

where $N(x)$ is the set of all open neighbourhoods in X containing $x \in X$.

A.2 Additional BNN Experiments

While the most interesting findings of our numerical studies can be found in the main text, here we give a brief overview over two more sets of experiments for further insights into BNNs. The first set consists in three more data sets with the same settings as used in the main text. While these findings do not change the overall picture, they do require more careful analysis and dissemination. The second set of results investigates the interaction between robustifying inference relative to the loss with robustifying it relative to the prior. The results suggest a clear relationship for predictive performance as measured by the root mean square error: If robust losses are used, the KLD generally performs better. Moreover, the combination of robust loss and D = KLD outperforms VI and the investigated DVI methods on all data

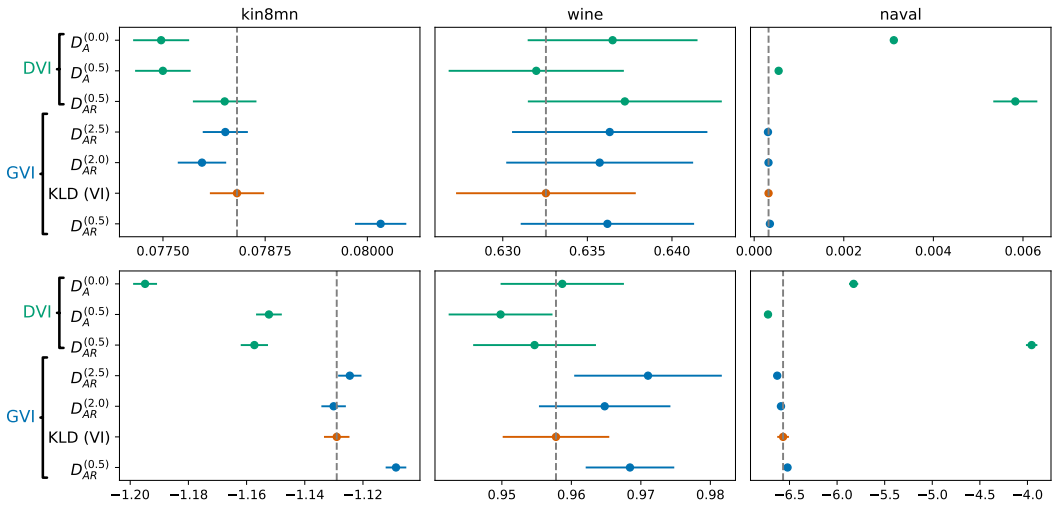


Figure A.1: Top row depicts RMSE, bottom row the NLL across a range of data sets using BNNs. Dots correspond to means, whiskers the standard errors. The further to the left, the better the predictive performance. For the depicted selection of data sets, no common pattern exists for the performance differences between **standard VI**, **DVI** and **GVI**.

sets studied. The relationship is less clear for the predictive negative log likelihood, both between loss and prior regularizer as well as between the performance to be expected under GVI, VI and DVI.

First set of additional experiments (Figure A.1)

Figure A.1 provides the predictive outcomes on three more data sets using the exact same settings and experimental setup as described in the main text. The findings generally reinforce the findings of the main text. First, while the GVI methods with $\alpha > 1$ still perform as good as or better than standard VI on the `kin8mn` data set, DVI methods show a clear performance gain relative to either of the two. Crucially, it is not clear what leads to this improvement gain, though the fact that the best-performing DVI method is the one recovering EP ($D_A^{(\alpha)}$ for $\alpha = 0$) suggests that there is tangible merit in producing mass-covering approximations to the posterior of $\boldsymbol{\theta}$ on this data set. While the deployment of DVI methods looks tempting on the `kin8mn` data set, the results on the `naval` data set are a reminder that the behaviour of these methods is in many ways unpredictable. Moreover, it shows that the risks we identified in Example 4.1 readily translate into real world applications: By using DVI methods, we may accidentally conflate the role of the loss and the role of uncertainty quantification. If the loss is well-suited for the data at hand—as the

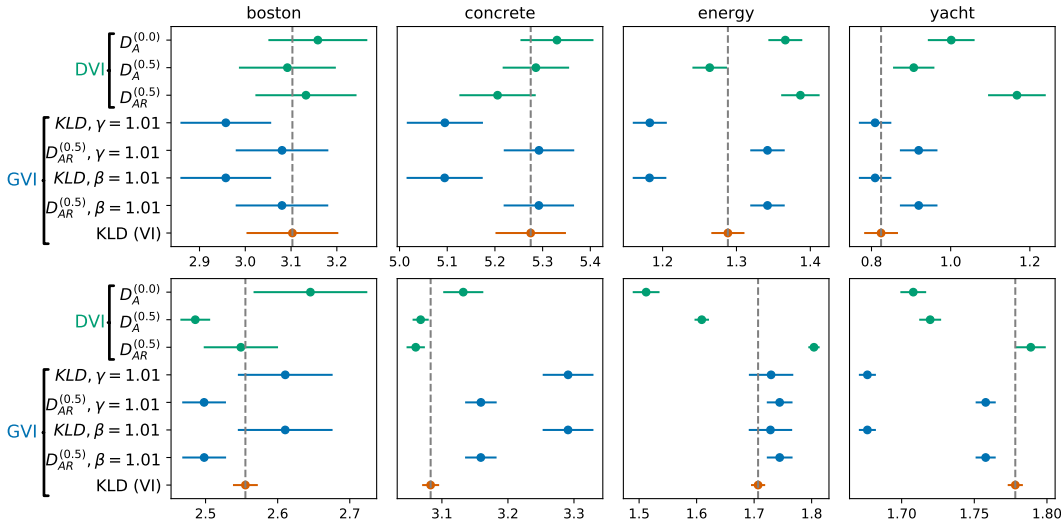


Figure A.2: Top row depicts RMSE, bottom row the NLL across a range of data sets using BNNs. Dots correspond to means, whiskers the standard errors. The further to the left, the better the predictive performance. For the depicted selection of data sets, patterns exists for the interplay between the loss and prior regularizer for **GVI**.

RMSE panel suggests it is in the `naval` case—the mass-covering behaviour of DVI methods can be extremely detrimental. Lastly, the `wine` data set provides a very similar picture to the results in Figure 5.11: Varying α introduces a banana-shaped curve for the GVI methods. As it so happens, the ideal choice of α on the `wine` data set appears to be around $\alpha = 1$ (i.e., standard VI). Taking into account the predictive uncertainty in form of the whiskers, it is doubtful if any of the methods is dominating another one on `wine`. Presumably, the reason for this is that the true posterior is relatively well approximated with the mean field normal family, yielding very similar results across all settings.

Second set of additional experiments (Figure A.2)

In a second set of additional experiments, we varied the loss function to be a robust scoring rule. Specifically, we used scoring rules based on the β-divergence and the γ-divergence. See (6.1) and (6.2) for the definition and more detail on these robust scoring rules. As for the DGP examples, we choose values of the scoring rule that are close to the log score, but sufficiently far to induce robust behaviour. All settings for optimization, initialization as well as the code are the same as for the results provided in the main text. Figure A.2 shows the results: For the RMSE, the results are unambiguous: Combining a robust scoring rule with the standard

prior regularizer $D = \text{KLD}$ appears to be the winning combination across all four data sets. The picture is less clear for the NLL: Relative to both VI and DVI, the performance gains depend on the data set. Even within the class of GVI posteriors, it is data-set dependent which prior regularizer should be chosen: For example, it is clearly beneficial to choose the $D_{AR}^{(\alpha)}$ as prior regularizer in the `boston` and `concrete` data sets, but the opposite is true on the `yacht` data set. Above all other things, this highlights the need for a good selection strategy of GVI hyperparameters: Oftentimes, intuitions about the correct prior regularizer or the appropriate loss may be incorrect.

A.3 Background on kernel methods

We provide some necessary background on matrix-valued kernels that explains them and is used in the proofs of Appendix C.3. Our main references are Carmeli et al. (2006); Caponnetto et al. (2008); Carmeli et al. (2010). For simplicity we start with the scalar-valued case and define a scalar-valued kernel:

Definition A.2 (Scalar-valued kernel)**.** A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *(scalar-valued) kernel* if

1. k is *symmetric*; i.e. $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$,
2. k is *positive semi-definite*; i.e. $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ for all $n \in \mathbb{N}$, $c_1, \ldots, c_n \in \mathbb{R}$ and all $x_1, \ldots, x_n \in \mathcal{X}$.

To every scalar-valued kernel is an associated Hilbert space $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}$, called the *reproducing kernel Hilbert space* (RKHS) of the kernel.

Definition A.3 (Reproducing kernel Hilbert space)**.** A Hilbert space $\mathcal{H}$ is said to be *reproduced* by a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if

1. $k(x, \cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$,
2. $\langle h, k(x, \cdot) \rangle_{\mathcal{H}} = h(x)$ for all $x \in \mathcal{X}$ and $h \in \mathcal{H}$.

The last item is called the *reproducing property* of k in $\mathcal{H}$.

It can be shown that, for every kernel k, there exists a unique Hilbert space $\mathcal{H}$ reproduced by k (Paulsen and Raghupathi, 2016, Theorem 2.14). These definitions can be generalised in the form of a matrix-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}$.

Definition A.4 (Matrix-valued kernel)**.** A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}$, $m > 1$, is called a *(matrix-valued) kernel* if

1. K is *symmetric*; i.e. $K(x, x') = K(x', x)$ for all $x, x' \in \mathcal{X}$,

2. k is *positive semi-definite*; i.e. $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i \cdot k(x_i, x_j) c_j \geq 0$ for all $n \in \mathbb{N}$, $c_1, \ldots, c_n \in \mathbb{R}^m$ and all $x_1, \ldots, x_n \in \mathcal{X}$.

As a direct generalisation of the scalar-valued case, there exists a uniquely associated Hilbert space $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}^m$ to every matrix-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}$. To define this Hilbert space, whose inner product we denote $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, some additional notation is required: Let F be a $\mathbb{R}^{m \times m}$-valued function and let $F_{i,-}$ denote the vector-valued function $F_{i,-} : \mathcal{X} \to \mathbb{R}^m$ defined by the the i-th row of F. Similarly, let G be a $\mathbb{R}^{m \times m}$-valued function and let $G_{-,j}$ denote the vector-valued function $G_{-,j} : \mathcal{X} \to \mathbb{R}^m$ defined by the j-th column of G. Formally define the symbols $\langle F, g \rangle_{\mathcal{H}}$, $\langle f, G \rangle_{\mathcal{H}}$ and $\langle F, G \rangle_{\mathcal{H}}$ as follows

$$\langle F, g \rangle_{\mathcal{H}} := \begin{bmatrix} \langle F_{1,-}, g \rangle_{\mathcal{H}} \\ \vdots \\ \langle F_{m,-}, g \rangle_{\mathcal{H}} \end{bmatrix} \in \mathbb{R}^m, \quad \langle f, G \rangle_{\mathcal{H}} := \begin{bmatrix} \langle f, G_{-,1} \rangle_{\mathcal{H}} \\ \vdots \\ \langle f, G_{-,m} \rangle_{\mathcal{H}} \end{bmatrix} \in \mathbb{R}^m,$$

$$\langle F, G \rangle_{\mathcal{H}} := \begin{bmatrix} \langle F_{1,-}, G_{-,1} \rangle_{\mathcal{H}} & \cdots & \langle F_{1,-}, G_{-,m} \rangle_{\mathcal{H}} \\ \vdots & & \vdots \\ \langle F_{m,-}, G_{-,1} \rangle_{\mathcal{H}} & \cdots & \langle F_{m,-}, G_{-,m} \rangle_{\mathcal{H}} \end{bmatrix} \in \mathbb{R}^{m \times m},$$

where these are to be interpreted as compound symbols only (i.e. we are not attempting to define an inner product on matrix-valued functions). Then, the generalisation of the reproducing property (see Definition A.3) to a matrix-valued kernel K is

$$h(x) = \langle h, K(x, \cdot) \rangle_{\mathcal{H}} = \begin{bmatrix} \langle h, K_{-,1}(x, \cdot) \rangle_{\mathcal{H}} \\ \vdots \\ \langle h, K_{-,m}(x, \cdot) \rangle_{\mathcal{H}} \end{bmatrix}$$

for all $x \in \mathcal{X}$ and $h \in \mathcal{H}$ (Carmeli et al., 2010). The generalisation of the symmetry property (see Definition A.3) is straight-forward; $K(x, x') = K(x', x)$ for all $x, x' \in \mathcal{X}$. A Hilbert space $\mathcal{H}$ for which these two properties are satisfied is called a vector-valued RKHS that we say is reproduced by the matrix-valued kernel K. Matrix-valued kernels and their associated vector-valued RKHS have recently been exploited in the context of Stein's method (e.g. Barp et al., 2019; Wang et al., 2019a).

A.4 Additional Details on Experiments for Robust Changepoint Detection

For all experiment, constrained Limited Memory BroydenFletcherGoldfarbShannon is used for the full optimization step, where the constraints are $\widehat{a}_n > 1, \widehat{b}_n > 1$. We use Python's `scipy.optimize` wrapper, which calls a Fortran implementation. We also tested whether inference is sensitive to different initializations of β_{m} and found that it is fairly stable as long as β_{m} is chosen reasonably. For example, for the Air Pollution data, we could recover the same changepoint (± 5 days) for initializations of β_{m} ranging from 0.005 up to 0.1. All experiments were performed on a 2017 MacBook Pro with 16 GB 2133 MHz LPDDR3 and 3.1 GHz Intel Core i7.

A.4.1 Well-log data

Hyperparameters: We set the hyperparameters for standard Bayesian On-line Changepoint Detection slightly differently, the reason being that due to the robustness guarantee of Theorem 1, we can use much less informative priors with the robust version than we can with the standard version: If priors are too flat, the standard version declares far too many changepoints. Thus, for the standard version, we use a constant CP prior (hazard) $H(r_t = r_{t-1} + 1 | r_{t-1}) = 0.01$, $a_0 = 1$, $b_0 = 10^4$, $\Sigma_0 = 0.25$, $\boldsymbol{\mu}_0 = 1.15 \cdot 10^4$, while for the robust version we can use a less informative prior by instead setting $b_0 = 10^7$. By virtue of our initialization procedure for β_p, this implies setting $\beta_{p,0} \approx 0.05$. To start out close to the KLD, we initialize $\beta_{\mathrm{rld},0} = 0.0001$.

Inferential procedure: For the robust version, we set $W = 360$, $B = 25$, $b = 10$, $m = 20$, $K = 1$. For both versions, only the 50 most likely run-lengths are kept. For the robust version, the average processing time was 0.487 per observation.

A.4.2 Air Pollution data

Preprocessing & Model Setup: The air pollution data is observed every 15 minutes across 29 stations for 365 days. We average the 96 observations made over 24 hours. This is done to move the observed data closer to a normal distribution, as the measurements have significant daily volatility variations. To account for weekly cycles, we also calculate for each station the mean for each weekday and subtract it from the raw data.. Yearly seasonality is not accounted for. Afterwards, the data is normalized station-wise. This is done only for numerical stability, because

the internal mechanisms of the used VAR models perform matrix operations (QR-decompositions and matrix multiplications in particular) that can adversely affect numerical stability for observations with large absolute value. Fig. A.3 shows some of the station's data after these preprocessing steps have been taken.

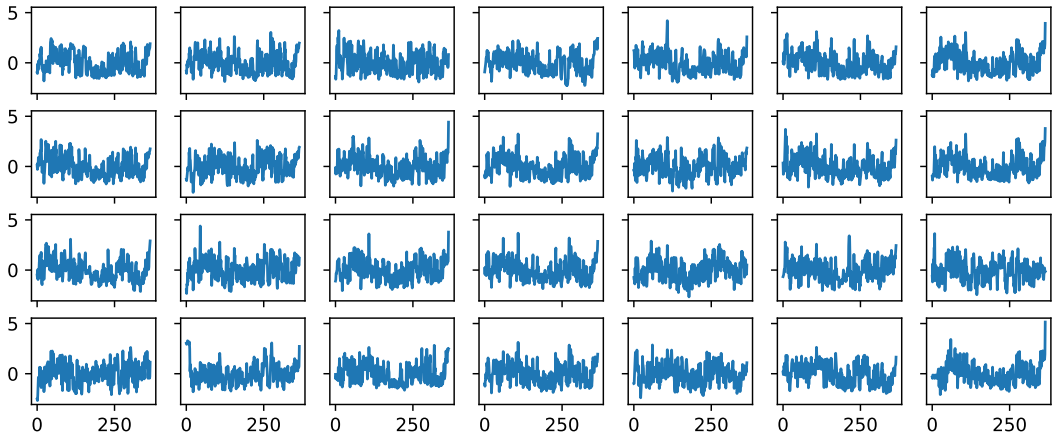


Figure A.3: Some of the stations after preprocessing steps. x-axis gives NOX level, y-axis the day.

The autoregressive models and spatially structured vector autoregressive models (VARs) are chosen to have lag lengths $1, 2, 3$. These short lag lengths are chosen to explicitly disadvantage the robust model universe: The non-robust run we compare against uses more than 20 models, with lag lengths $1, 5, 6, 7$, meaning that it is much more expressive and should be able to cope with outliers better. In spite of this, it not only declares more CPs, but also does worse than the robust version in terms of predictive performance. For both the robust and non-robust model, two spatially structured VARs are included as in Knoblauch and Damoulas (2018).

Hyperparameters: We set $H(r_t = r_{t-1}+1|r_{t-1}) = 0.001$, $a_0 = 1$, $b_0 = 25$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\Sigma_0 = I \cdot 20$, which yields initialization $\beta_{\text{m}} \approx 0.005$, $\beta_{\text{rlm}} = 0.1$. The non-robust results are directly taken from Knoblauch and Damoulas (2018) and can be replicated running the code available from
`https://github.com/alan-turing-institute/bocpdms/`

Inferential procedure: We set $W = 300$, $m = 50$, $B = 20$ and $b = 10$, $K = 25$ and retain the 50 most likely run-lengths. Processing times are more volatile

than for the well-log because the full optimization procedure is significantly more expensive to perform. Most observations take significantly less than 20 seconds to process, but some take over a minute (depending on how many of the retained run-lengths are divisible by m at each time point).

A.4.3 Optimizing β

Lastly, we investigate the trajectories for β as it is being optimized. For all trajectories, a bounded predictive absolute loss was used with threshold τ, i.e. $L(x) = \max\{|x|, \tau\}$. For β_{rld}, $\tau = 5/T$ (where T is the length of the time series) while for β_{p}, $\tau = 0.1$. The results are not sensitive to these thresholds, and they are picked with the intent that (1) a single observation should not affect β_{p} by more than 0.1 and (2) that overall, β_{rld} should not change by more than 5 in absolute magnitude. As the initialization procedure for β_{p} works very well for predictive performance, the on-line optimization never even comes close to making a step with size τ. The picture is rather different for β_{rld}, which reaches τ rather often. We note that this is because the estimated gradients for β_{rld} can be very extreme, which is why the implementation averages 50 consecutive gradients before performing a step. Overall, we note that for the well log data whose trajectories are depicted in Fig. A.4, the degrees of robustness do not change much relative to their starting points at $\beta_{\text{p}} = 0.05$ and $\beta_{\text{rld}} = 0.001$. In particular, the absolute change over more than $4,000$ observations is < 0.002 for β_{p} and < 0.015 for β_{rld}. Step sizes are $1/t$ at time t.

For the Air Pollution Data, the story is slightly different: Here, β_{p} does not change after the first iteration, where it jumps from 0.005 directly to 10^{-10}. While this seems odd, it is mainly due to the fact that for numerical stability reasons[1], one needs to ensure that $\beta_{\text{p}} > \varepsilon$ for some $\varepsilon > 0$; and in our implementation, $\varepsilon = 10^{-10}$. The interpretation of the trace graph is thus that the optimization continuously suggests less robust values for β_{p}, but that we cannot admit them due to numerical stability. The downward trend also holds for β_{rld}, which is big enough to not endanger numerical stability and hence can drift downwards.

Fig. A.4 also shows that the optimization technique used for β needs further investigation and research. For starters, the outcomes suggest that a second order method could yield better results than using a first-order SGD technique. In the future, we would like to explore this in greater detail and also explore more advanced optimization methods like line search or trust region optimization methods for this

[1] In particular, working with the $D_B^{(\beta)}$ implies that one takes the exponential of a density, i.e. e^{f^β}. So even working on a log scale now means working with the densities f^β *directly*. It should be clear that these quantities become numerically unstable for β too large or too small.

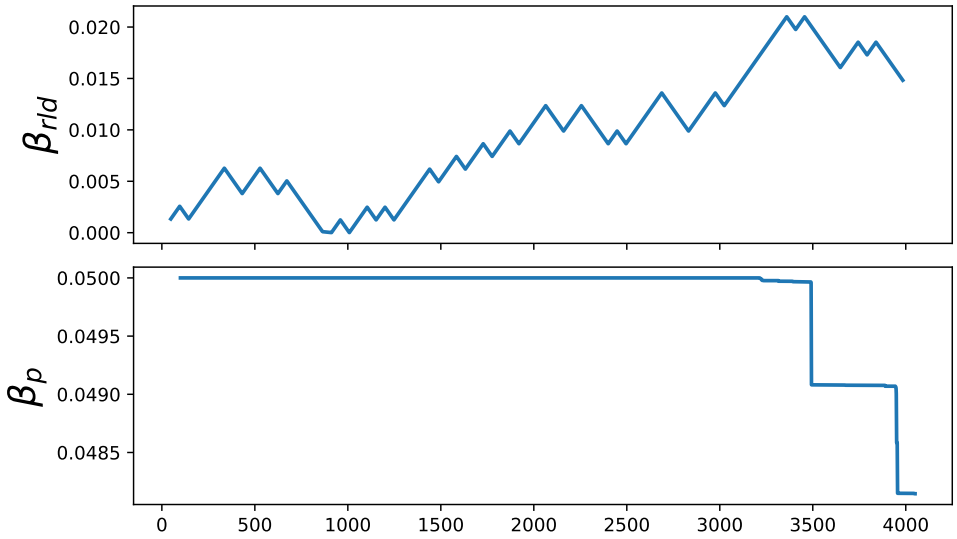


Figure A.4: β trajectories for the well-log data. For β_{rld}, steps are only taken every 50 observations to average gradient noise

problem.

Appendix B

Technical Derivations

B.1 Link to the Predictive Information Bottleneck

One can rewrite eq. (1.10) as an unconstrained optimization problem by a well-known argument. For a scalar $\beta = \beta(I_0, \boldsymbol{x}_{1:n})$ derived as in Theorem 1 of Tishby et al. (2000), we have that

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \underset{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})\in\Pi_{\text{PIB}}}{\arg\min} \left\{-I(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty}) + (1-\beta)I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n})\right\}.$$

But we can do even better: by noting that any distribution on $\boldsymbol{\theta}$ is obtained by compressing (i.e. training on) $\boldsymbol{x}_{1:n}$ only, we also know that $\boldsymbol{\theta}$ and $\boldsymbol{x}_{n+1:\infty}$ are independent once conditioned on $\boldsymbol{x}_{1:n}$. This means that $p(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty}|\boldsymbol{x}_{1:n}) = p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})p(\boldsymbol{x}_{n+1:\infty}|\boldsymbol{x}_{1:n})$, so that $I(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty}|\boldsymbol{x}_{1:n}) = 0$. By elementary operations (see Alemi, 2019), this implies that we can rewrite

$$I(\boldsymbol{\theta}, \boldsymbol{x}_{n+1:\infty}) = I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) - I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty}),$$

which we can plug into the unconstrained form to find that

$$q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \underset{p(\boldsymbol{\theta}|x_{1:n})\in\Pi_{\text{PIB}}}{\arg\min} \left\{I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty}) - \beta I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n})\right\}. \tag{B.1}$$

Though this may not be immediately obvious, eq. (B.1) has a close relationship with the RoT. To see how this conclusion can be reached, first note that

$$\begin{aligned}
\beta I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}) &= \beta \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})p(\boldsymbol{x}_{1:n})}\left[\log\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})\cancel{p(\boldsymbol{x}_{1:n})}}{p(\boldsymbol{\theta})\cancel{p(\boldsymbol{x}_{1:n})}}\right)\right] \\
&= \beta \mathbb{E}_{p(\boldsymbol{x}_{1:n})}\left[\text{KLD}\left(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})\|p(\boldsymbol{\theta})\right)\right]. \\
&=: D_{\text{PIB}}(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})\|\pi_{\text{PIB}}(\boldsymbol{\theta})),
\end{aligned}$$

where we have defined the marginal $\pi_{\text{PIB}}(\boldsymbol{\theta}) = \int_{\mathcal{X}^n} p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})p(\boldsymbol{x}_{1:n})d\boldsymbol{x}_{1:n}$. Clearly, $D_{\text{PIB}}(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})\|\pi_{\text{PIB}}(\boldsymbol{\theta})) \geq 0$ and $D_{\text{PIB}}(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})\|\pi_{\text{PIB}}(\boldsymbol{\theta})) = 0 \iff p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \pi_{\text{PIB}}(\boldsymbol{\theta})$. Notice that unlike in the Bayesian paradigm, the prior π_{PIB} here is *not* a free variable. Instead, it gives the distribution over $\boldsymbol{\theta}$ which is obtained over all possible configurations of $\boldsymbol{x}_{1:n}$, which makes this prior conceptually close to a bootstrap distribution.

Similarly, we can rewrite the first term as a loss function by noting that

$$
\begin{aligned}
&I(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty}) \\
&= \mathbb{E}_{p(\boldsymbol{x}_{n+1:\infty})}\left[\text{KLD}\left(p(\boldsymbol{\theta}, \boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty})\|p(\boldsymbol{\theta}|\boldsymbol{x}_{n+1:\infty})p(\boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty})\right)\right] \\
&= \mathbb{E}_{p(\boldsymbol{x}_{n+1:\infty})}\left[\text{KLD}\left(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})\cancel{p(\boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty})}\|p(\boldsymbol{\theta}|\boldsymbol{x}_{n+1:\infty})\cancel{p(\boldsymbol{x}_{1:n}|\boldsymbol{x}_{n+1:\infty})}\right)\right] \\
&= \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})}\Bigg[\underbrace{\mathbb{E}_{p(\boldsymbol{x}_{1:n})}\left[\log\left(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n})\right)\right] - \mathbb{E}_{p(\boldsymbol{x}_{n+1:\infty})}\left[\log p(\boldsymbol{\theta}|\boldsymbol{x}_{n+1:\infty})\right]}_{=L_{n,\text{PIB}}(p(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}))}\Bigg].
\end{aligned}
$$

While this loss is not computable in practice, it has a clear interpretation. Specifically, it jointly minimizes (i) the information that $\boldsymbol{\theta}$ loses on future data $\boldsymbol{x}_{n+1:\infty}$ and (ii) the difference between the information that $\boldsymbol{\theta}$ loses on $\boldsymbol{x}_{1:n}$ versus $\boldsymbol{x}_{n+1:\infty}$. The loss $L_{n,\text{PIB}}$ has two properties that set it apart from the losses we have considered thus far: first of, $L_{n,\text{PIB}}$ does not depend on a sample $x_{1:n}$ (but the distributions of the underlying random variables $\boldsymbol{x}_{1:n}, \boldsymbol{x}_{n+1:\infty}$). Second, $L_{n,\text{PIB}}$ is not summable. Neither of these properties affect the axiomatic development in Section **??**, since any empty sample is a finite sample and because summability was imposed for presentational purposes only.

Putting everything together, we can rewrite the PIB as

$$
q^*(\boldsymbol{\theta}|\boldsymbol{x}_{1:n}) = \underset{q \in \Pi_{\text{PIB}}}{\arg\min}\left\{\mathbb{E}_q\left[L_{n,\text{PIB}}(q)\right] + D_{\text{PIB}}(q\|\pi_{\text{PIB}})\right\}.
$$

B.2 Latent Variable Models & Variational Autoencoders

While we have thus far stated the entire development in terms of a single *global* latent variable $\boldsymbol{\theta}$, nothing stops us from extending the presented ideas to *local* latent variables. The reason for this is that none of our Axioms prohibit $\boldsymbol{\Theta}$ or Π to depend on n or indeed $x_{1:n}$. In other words, we can seamlessly transfer everything we considered thus far to the context of inference on local latent variables $\boldsymbol{z}_{1:n} \in \mathcal{Z}^n$ by taking $\boldsymbol{\Theta} = \boldsymbol{\Theta}(n) = \mathcal{Z}^n$.

To make this logic more tangible, we will explain how Variational Autoencoders (VAEs) (Kingma and Welling, 2013) can be recast in the RoT form. VAEs

use local latent variables, in our notation $\boldsymbol{\theta} = \boldsymbol{\theta}_{1:n}$, to encode lower dimensional representations of observations $x_{1:n}$ via the global parameter $\boldsymbol{\kappa}_g$. Simultaneously, they seek to probabilistically decode the latent variables back to the observation space via the global decoder model with parameters ζ. This involves an optimisation problem over a set of distributions for the latent variables. The corresponding variational family is

$$\Pi_{x_{1:n}} = \left\{ q(\boldsymbol{\theta}|\boldsymbol{\kappa}_g) = \prod_{i=1}^{n} q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i) \text{ so that } q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i) = \mathcal{N}(\boldsymbol{\theta}_i; \mu(\boldsymbol{\kappa}_g, x_i), \sigma(\boldsymbol{\kappa}_g, x_i) \right\},$$

where the parameters $\boldsymbol{\kappa}_i = (\boldsymbol{\kappa}_g, x_i)$ consist of a fixed local component observation x_i as well as the global parameter $\boldsymbol{\kappa}_g$ that is shared to be optimized over. Here, $\boldsymbol{\kappa}_g$ will define the weights of a neural network indexing a probabilistic model. The optimization problem underlying a VAE is now given by

$$\underset{\zeta, q \in \Pi_{x_{1:n}}}{\arg\min} \left\{ \sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i)} \left[-\log p_\zeta(x_i|\boldsymbol{\theta}_i) \right] + \sum_{i=1}^{n} \mathrm{KLD}\left(q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i) \| \pi(\boldsymbol{\theta}_i) \right) \right\}.$$

where $\sum_{i=1}^{n} \mathbb{E}_{q(\boldsymbol{\theta}_i|\boldsymbol{\kappa}_i)} \left[-\log p_\zeta(x_i|\boldsymbol{\theta}_i) \right]$ minimises the expected reconstruction error of decoding the probabilistic encoding and the KLD term regularises this encoding to improve the model's capacity to generate realistic pseudo-observations. Now simply note that for the fully factorized prior $\pi(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(\boldsymbol{\theta}_i)$, one can rewrite the above as

$$\underset{\zeta, q \in \Pi_{x_{1:n}}}{\arg\min} \left\{ \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa}_g)} \left[\sum_{i=1}^{n} -\log p_\zeta(x_i|\boldsymbol{\theta}_i) \right] + \mathrm{KLD}\left(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_g) \| \pi(\boldsymbol{\theta}) \right) \right\}, \tag{B.2}$$

which is a RoT form with an added optimization over the hyperparameter ζ.[1] An important distinction between this example and many of the others in Table 1.1 is that for VAEs, the variational distributions are introduced in order to regularise the latent space rather than to approximate an underlying Bayesian posterior. As a result, the VAE objective exists solely as a means to generate desirable generative distributions for a particular inference tasks.

[1] Optimizing over hyperparameters in variational objectives is very common, and our experiments in Chapter 5 make use of this technique, too. While optimizing over hyperparameters is strictly speaking not part of the RoT definition, we treat and discuss objectives of this kind essentially as members of the RoT.

B.3 Derivations for Duality Examples

B.3.1 Proof of Example 2.3

Note that when we pick D as an f-divergence, there is a standard result we can recall in the following lemma.

Lemma B.1. For any f-divergence D based on the lower-semicontinuous convex function $f : \mathbb{R} \to (-\infty, \infty]$ with $f(1) = 0$ so that $D_\pi(\cdot) = D(\cdot\|\pi)$, and $h \in \mathcal{F}_b(\mathbf{\Theta})$, it holds that

$$D_\pi^\star(h) = \inf_{b\in\mathbb{R}} \left\{\mathbb{E}_\pi[f^\star(h-b)] + b\right\}, \tag{B.3}$$

where $f^\star(t) = \sup_{t'} \{tt' - f(t')\}$ is the convex conjugate.

A proof of this result can be found in Equation (22) of (Liu and Chaudhuri, 2018). Using this, we can now prove the example for the Kullback-Leibler and χ^2 divergences.

B.3.2 Proof of Example 2.2

Noting that $f^\star(t) = \exp(t-1)$, the inner problem revolving around b can easily be solved:

$$\begin{aligned}\inf_{b\in\mathbb{R}} \{\mathbb{E}_\pi[\exp(h-b)] + b\} &= \inf_{b\in\mathbb{R}} \{\exp(-b)\cdot \mathbb{E}_\pi[\exp h] + b\} \\ &= \log \mathbb{E}_\pi[\exp h]\end{aligned}$$

B.3.3 Proof of Example 2.4

In this case we have $f^\star(t) = t + \frac{t^2}{4}$ and in particular $(w^{-1}f)^\star(t) = t + \frac{t^2}{4w^{-1}}$. The infimum problem, similar to the KLD case becomes easily tractable:

$$\begin{aligned}\inf_{b\in\mathbb{R}} \left\{\mathbb{E}_\pi[h] + \frac{1}{4w^{-1}}\mathbb{E}_\pi\left[(h-b)^2\right]\right\} &= \mathbb{E}_\pi[h] + \frac{1}{4w^{-1}} \inf_{b\in\mathbb{R}} \mathbb{E}_\pi\left[(h-b)^2\right] \\ &= \mathbb{E}_\pi[h] + \frac{1}{4w^{-1}} \operatorname{Var}_\pi[h]\end{aligned}$$

B.3.4 Proof of Example 2.5

For this case, we just invoke (Husain, 2020, Lemma 5); which in combination with our main result yields the desired result.

B.4 Proof of Proposition 4.2

Proof. Proposition 4.2 considers the following forms of the prior and likelihood

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0) &= h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}_0))\right\} \\ q(\boldsymbol{\theta}|\boldsymbol{\kappa}) &= h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\} \\ p(\boldsymbol{x}|\boldsymbol{\theta}) &= h(\boldsymbol{\theta}) \exp(g(\boldsymbol{x})^T T(\boldsymbol{\theta}) - B(\boldsymbol{x})),\end{aligned}$$

where $A(\eta(\boldsymbol{\kappa})) = \log \int h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta})\right\} d\boldsymbol{\theta}$ and $h(\boldsymbol{\theta}) = \frac{1}{\int \exp(g(\boldsymbol{x})^T T(\boldsymbol{\theta}) - B(\boldsymbol{x})) d\boldsymbol{x}}$.

The GVI objective function in this scenario is

$$\begin{aligned}O_{\text{GVI}}(\boldsymbol{\kappa}) &= \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[\sum_{i=1}^{n} \ell_G^{(\gamma)}(\boldsymbol{\theta}, \boldsymbol{x}_i)\right] + \text{KLD}(q(\boldsymbol{\theta}|\boldsymbol{\kappa})||q(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)) \\ &= \sum_{i=1}^{n} \underbrace{\int \underbrace{\ell_G^{(\gamma)}(\boldsymbol{\theta}, \boldsymbol{x}_i)}_{C_1(\boldsymbol{\kappa}, \boldsymbol{\theta}, x_i)} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\theta}_{C_2(\boldsymbol{\kappa}, \boldsymbol{x}_i)} + \underbrace{\text{KLD}(q(\boldsymbol{\theta}|\boldsymbol{\kappa})||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0))}_{C_3(\boldsymbol{\kappa}, \boldsymbol{\kappa}_0)}.\end{aligned}$$

This decomposition contains three terms that we need to check are closed forms of $\boldsymbol{\kappa}$. Firstly

$$C_1(\boldsymbol{\kappa}, \boldsymbol{\theta}, \boldsymbol{x}_i) = \ell_G^{(\gamma)}(\boldsymbol{x}_i, \boldsymbol{\theta}) = -\frac{1}{\gamma - 1} p(\boldsymbol{x}_i; \boldsymbol{\theta})^{\gamma - 1} \frac{\gamma}{\left[\int p(\boldsymbol{z}; \boldsymbol{\theta})^\gamma d\boldsymbol{z}\right]^{\frac{\gamma-1}{\gamma}}};$$

so that in order for this to be a closed form function of $\boldsymbol{\kappa}$, $\boldsymbol{\theta}$, and $\boldsymbol{x}_i$, we require that

$$I^{(\gamma)}(\boldsymbol{\theta}) = \int p(\boldsymbol{z}|\boldsymbol{\theta})^\gamma dz = \int h(\boldsymbol{\theta})^\gamma \exp(\gamma g(\boldsymbol{z})^T T(\boldsymbol{\theta}) - \gamma B(\boldsymbol{z})) dz,$$

where the theorem statement ensures that $I^{(\gamma)}(\boldsymbol{\theta})$ is a closed form function of $\boldsymbol{\theta}$. Next, consider that

$$\begin{aligned}&C_2(\boldsymbol{\kappa}, \boldsymbol{x}_i) \\ &= -\frac{\gamma}{\gamma - 1} \int h(\boldsymbol{\theta})^{\gamma - 1} \exp((\gamma - 1) g(\boldsymbol{x}_i)^T T(\boldsymbol{\theta}) - (\gamma - 1) B(\boldsymbol{x}_i)) \frac{1}{\left[h(\boldsymbol{\theta})^\gamma I^{(\gamma)}(\boldsymbol{\theta})\right]^{\frac{\gamma-1}{\gamma}}} q(\boldsymbol{\theta}|\boldsymbol{\kappa}) d\boldsymbol{\theta} \\ &= -\frac{\gamma}{\gamma - 1} \frac{\exp\left((1 - \gamma) B(\boldsymbol{x}_i) + A\left(\eta(\boldsymbol{\kappa}) + (\gamma - 1) g(\boldsymbol{x}_i)\right)\right)}{\exp\left(A(\eta(\boldsymbol{\kappa}))\right)} \mathbb{E}_{q(\boldsymbol{\theta}|(\eta(\boldsymbol{\kappa}) + (\gamma - 1) g(\boldsymbol{x}_i)))}\left[I^{(\gamma)}(\boldsymbol{\theta})^{\frac{1-\gamma}{\gamma}}\right],\end{aligned}$$

where the theorem statement ensures that $(\eta(\boldsymbol{\kappa}_n) + (\gamma - 1) g(\boldsymbol{x}_i)) \in \mathcal{N}$ for all x_i and that $F_2(\boldsymbol{\kappa}^*) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)}\left[I^{(\gamma)}(\boldsymbol{\theta})^{\frac{1-\gamma}{\gamma}}\right]$ is closed form function of $\boldsymbol{\kappa}^*$ for all $\boldsymbol{\kappa}^* \in \mathcal{N}$.

Lastly

$$C_3(\boldsymbol{\kappa}, \kappa_0) = \int h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\} \log \frac{h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\}}{h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}_0))\right\}} d\boldsymbol{\theta}$$
$$= A(\eta(\boldsymbol{\kappa}_0)) - A(\eta(\boldsymbol{\kappa})) + (\eta(\boldsymbol{\kappa}) - \eta(\boldsymbol{\kappa}_0))^T \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa})}\left[T(\boldsymbol{\theta})\right],$$

where the statement ensures that $F_1(\kappa^*) = \mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)}\left[T(\boldsymbol{\theta})\right]$ is a closed form function of $\boldsymbol{\kappa}^*$ for all $\kappa^* \in \mathcal{N}$. □

B.5 Closed forms for divergences & proof of Proposition 4.1

This section proves various closed forms for the prior regularizers in the GVI problem with the $D_A^{(\alpha)}$, $D_{AR}^{(\alpha)}$, $D_B^{(\beta)}$, and $D_G^{(\gamma)}$. We do so by proving conditions for closed forms of the $\alpha\beta\gamma$-divergence ($D_G^{(\alpha,\beta,r)}$), recovering $D_A^{(\alpha)}$, $D_{AR}^{(\alpha)}$, $D_B^{(\beta)}$, and $D_G^{(\gamma)}$ as special cases. Note that the special case of these results for the $D_{AR}^{(\alpha)}$ has been derived before (see Gil et al., 2013; Gil, 2011; Liese and Vajda, 1987). Unlike previous work, our results apply to a range of other divergences, too. We start by defining $D_G^{(\alpha,\beta,r)}$.

Definition B.1 (The $\alpha\beta\gamma$-divergence $D_G^{(\alpha,\beta,r)}$ (Cichocki and Amari, 2010))**.** The $\alpha\beta\gamma$-divergence $D_G^{(\alpha,\beta,r)}$ Cichocki and Amari (2010) takes the form

$$D_G^{(\alpha,\beta,r)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(\beta-1)(\alpha+\beta-1)r}\left[\left(\tilde{D}_G^{(\alpha,\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) + 1\right)^r - 1\right]$$

where $r > 0$, $\alpha \neq 0$, $\beta \neq 1$ and

$$\tilde{D}_G^{(\alpha,\beta)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \int \left(\alpha q(\boldsymbol{\theta})^{\alpha+\beta-1} + (\beta-1)\pi(\boldsymbol{\theta})^{\alpha+\beta-1} - (\alpha+\beta-1)q(\boldsymbol{\theta})^{\alpha}\pi(\boldsymbol{\theta})^{\beta-1}\right) d\boldsymbol{\theta}$$

Remark B.1. $D_A^{(\alpha)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ when $r = 1$ and $\beta = 2 - \alpha$. $D_{AR}^{(\alpha)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ in the limit as $r \to 0$ and $\beta = 2 - \alpha$. $D_B^{(\beta)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ when $r = \alpha = 1$. $D_G^{(\gamma)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ in the limit as $r \to 0$, $\alpha = 1$ and $\beta = \gamma$.

B.5.1 High-level overview of results and preliminaries

Summarizing some of the most important findings of this section, we find that if both $q(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ are in the same exponential variational family $\mathcal{Q}$,

- $D_{AR}^{(\alpha)}(q||\pi)$ and $D_{A}^{(\alpha)}(q|\pi)$ are always available in closed form if $\alpha \in (0,1)$ (see Corollary B.1)
- $D_{AR}^{(\alpha)}(q||\pi)$ and $D_{A}^{(\alpha)}(q|\pi)$ are available in closed form if $\alpha > 1$ for most exponential families (see again Corollary B.1)
- $D_{B}^{(\beta)}(q||\pi)$ and $D_{G}^{(\gamma)}(q||\pi)$ are available in closed form for $\beta > 1$ and $\gamma > 1$ for most exponential families (See Corollary B.5).

We note that these findings are interesting because closed forms for the divergence term drastically reduce the variance of black box GVI as introduced in Chapter 4. The remainder of this section is devoted to tedious but rigorous derivations of these findings. Before stating any results, it is useful to state the definition of an exponential family and its natural parameter space upon which the proofs rely.

Definition B.2 (Exponential families)**.** Object $\theta \in \Theta \subset \mathbb{R}^d$, $d \geq 1$ has an exponential family distribution with parameters $\kappa \in \boldsymbol{K} \subset \mathbb{R}^{p'}$, $p' \geq 1$ if there exist functions $\eta : \boldsymbol{K} \to \mathcal{N} \subset \mathbb{R}^p$, $p \geq 1$, $T : \Theta \to \mathcal{T} \subset \mathbb{R}^p$, $h : \Theta \to \mathbb{R}_{\geq 0}$ and $A : \mathcal{N} \to \mathbb{R}$ such that

$$p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa})) = h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\},$$

where $A(\eta(\boldsymbol{\kappa})) = -\log\left(\int h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta})\right\} d\theta\right)$. The set $\mathcal{N}$ is called the natural parameter space and is defined to ensure $p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}))$ is a normalized probability density, $\mathcal{N} = \{\eta(\boldsymbol{\kappa}) : A(\eta(\boldsymbol{\kappa})) < \infty\}$.

Throughout the rest of this section, we assume that the following condition holds for both the prior and the variational family $\mathcal{Q}$.

Assumption B.1 (The prior and variational families)**.** It holds that

1. the variational family $\mathcal{Q}$ is an exponential family as given in Definition B.2
2. the prior $\pi(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}_0))$ is a member of that variational family.

Amongst other things, this implies that the log-normalizing constant is a closed form function of the natural parameters and that we can derive generic conditions for closed forms by using the canonical representation of exponential families.

To showcase the implications of the derived results, we use the Mulitvariate Gaussian (MVN) to provide examples along the way.

Definition B.3 (The MVN exponential family)**.** The density of the MVN exponential family for vector $\boldsymbol{\theta}$ of dimension d is $p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa})) = h(\boldsymbol{\theta}) \exp\left\{\eta(\boldsymbol{\kappa})^T T(\boldsymbol{\theta}) - A(\eta(\boldsymbol{\kappa}))\right\}$ where

$$\eta(\boldsymbol{\kappa}) = \begin{pmatrix} \boldsymbol{V}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\boldsymbol{V}^{-1} \end{pmatrix} \qquad T(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\theta}\boldsymbol{\theta}^T \end{pmatrix}$$
$$h(\boldsymbol{\theta}) = (2\pi)^{-d/2} \qquad A(\eta(\boldsymbol{\kappa})) = \left[\frac{1}{2}\log|\boldsymbol{V}| + \frac{1}{2}\boldsymbol{\mu}\boldsymbol{V}^{-1}\boldsymbol{\mu}\right]$$

and the natural parameter space requires that $\boldsymbol{\mu}$ is a real valued vector of the same dimension as $\boldsymbol{\theta}$ and $\boldsymbol{V}$ is a $d \times d$ symmetric semi-positive definite matrix.

B.5.2 Results, proofs & examples

The remainder of this section is structued as follows: First, we give the main result for the $\alpha\beta\gamma$-divergence ($D_G^{(\alpha,\beta,r)}$) in Proposition B.1. This "master result" is then applied to various special cases for $D_G^{(\alpha,\beta,r)}$ that are of practical interest, namely the α-divergence ($D_A^{(\alpha)}$), Rényi's α-divergence ($D_{AR}^{(\alpha)}$), the β-divergence ($D_B^{(\beta)}$) as well the γ-divergence ($D_G^{(\gamma)}$).

Master result for $D_G^{(\alpha,\beta,r)}$

While the following result and corresponding proof are somewhat tedious to read, they are conceptually simple: In fact, all that is needed to derive the results is some basic algebra and the canonical form of the exponential family.

Proposition B.1 (Closed form $D_G^{(\alpha,\beta,r)}$ between exponential families)**.** The $D_G^{(\alpha,\beta,r)}$ between a variational posterior $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ and prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ is available in closed form under the following conditions

1. $\eta(\boldsymbol{\kappa}_0), \eta(\boldsymbol{\kappa}_n) \in \mathcal{N} \Rightarrow (\alpha\eta(\boldsymbol{\kappa}_0) + (\beta-1)\eta(\boldsymbol{\kappa}_n)) \in \mathcal{N}$;
2. $\mathbb{E}_{p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}))}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right]$ is a closed form function of $\eta(\boldsymbol{\kappa}) \in \mathcal{N}$.

If these conditions hold the $D_G^{(\alpha,\beta,r)}$ can be written as

$$\begin{aligned}
&\tilde{D}_G^{(\alpha,\beta)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)) \\
&= \alpha B(\boldsymbol{\kappa}_n, (\alpha+\beta-1))E(\boldsymbol{\kappa}_n, (\alpha+\beta-1)) + (\beta-1)B(\boldsymbol{\kappa}_0, (\alpha+\beta-1))E(\boldsymbol{\kappa}_0, (\alpha+\beta-1)) \\
&\quad - (\alpha+\beta-1)C(\boldsymbol{\kappa}_n, \boldsymbol{\kappa}_0, \alpha, (\beta-1))\tilde{E}(\boldsymbol{\kappa}_n, \boldsymbol{\kappa}_0, \alpha, (\beta-1))
\end{aligned}$$

where

$$B(\boldsymbol{\kappa}, \delta) = \frac{\exp\{A(\delta\eta(\boldsymbol{\kappa}))\}}{\exp\{A(\eta(\boldsymbol{\kappa}))\}^{\delta}}, \qquad C(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \delta_1, \delta_2) = \frac{\exp\{A\left(\delta_1\eta(\boldsymbol{\kappa}_1) + \delta_2)\eta(\boldsymbol{\kappa}_2)\right)\}}{\exp\{A(\eta(\boldsymbol{\kappa}_1))\}^{\delta_1}\exp\{A(\eta(\boldsymbol{\kappa}_2))\}^{\delta_2}}$$
$$E(\boldsymbol{\kappa}, \delta) = \mathbb{E}_{p(\theta|\delta\eta(\boldsymbol{\kappa}))}\left[h(\boldsymbol{\theta})^{\delta-1}\right], \qquad \tilde{E}(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \delta_1, \delta_2) = \mathbb{E}_{p(\theta|\delta_1\eta(\boldsymbol{\kappa}_1)+\delta_2\eta(\boldsymbol{\kappa}_2))}\left[h(\boldsymbol{\theta})^{\delta_1+\delta_2-1}\right]$$

we suppress the dependence of these functions on $A(\cdot)$ and $h(\cdot)$ as these derive form the definition of the exponential family (Definition B.2).

Proof. The $D_G^{(\alpha,\beta,r)}$ is a closed form function of $\tilde{D}_G^{(\alpha,\beta)}$ given in Definition B.1. Hence if $\tilde{D}_G^{(\alpha,\beta)}$ is available in closed form, then so is $D_G^{(\alpha,\beta,r)}$. In order to ensure that $\tilde{D}_G^{(\alpha,\beta)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0))$ has closed form, we need to make sure the three integrals below are available in closed form for the exponential family.

$$G_1 := \int q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)^{\alpha+\beta-1}d\boldsymbol{\theta}, \quad G_2 := \int \pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)^{\alpha+\beta-1}d\boldsymbol{\theta},$$
$$G_3 := \int q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)^{\alpha}\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)^{\beta-1}d\boldsymbol{\theta}.$$

First we tackle G_1.

$$\begin{aligned} G_1 &= \int h(\boldsymbol{\theta})^{\alpha+\beta-1}\exp\left\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta}) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_n))\right\}d\boldsymbol{\theta} \\ &= \exp\left\{A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_n))\right\}\mathbb{E}_{p(\boldsymbol{\theta}|(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right], \end{aligned}$$

where condition (1) with $\eta(\boldsymbol{\kappa}_0) = \eta(\boldsymbol{\kappa}_n)$ ensures that

$$A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)) = \int h(\boldsymbol{\theta})\exp\left\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta})\right\}d\boldsymbol{\theta} < \infty,$$

which in turn ensures that $p(\boldsymbol{\theta}|(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)$ is a normalized probability density and that
$\mathbb{E}_{p(\boldsymbol{\theta}|(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_n)}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right]$ is a valid expectation. Now, condition (2) guarantees this is a closed form function of $\eta(\boldsymbol{\kappa}_n)$. Similarly for G_2,

$$\begin{aligned} G_2 &= \int h(\boldsymbol{\theta})^{\alpha+\beta-1}\exp\left\{(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_0))\right\}d\boldsymbol{\theta} \\ &= \exp\left\{A((\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)) - (\alpha+\beta-1)A(\eta(\boldsymbol{\kappa}_0))\right\}\mathbb{E}_{p(\boldsymbol{\theta}|(\alpha+\beta-1)\eta(\boldsymbol{\kappa}_0)}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right], \end{aligned}$$

where in analogy to G_1, conditions (1) and (2) with $\eta(\boldsymbol{\kappa}_k) = \eta(\boldsymbol{\kappa}_0)$ ensure this has

a closed form. Lastly for G_3,

$$\begin{aligned} G_3 = &\int h(\boldsymbol{\theta})^{\alpha} \exp\left\{\alpha\eta(\boldsymbol{\kappa}_n)^T T(\boldsymbol{\theta}) - \alpha A(\eta(\boldsymbol{\kappa}_n))\right\} \\ &\cdot h(\boldsymbol{\theta})^{\beta-1} \exp\left\{(\beta-1)\eta(\boldsymbol{\kappa}_0)^T T(\boldsymbol{\theta}) - (\beta-1)A(\eta(\boldsymbol{\kappa}_0))\right\} d\boldsymbol{\theta} \\ = &\exp\left\{A\left(\alpha\eta(\boldsymbol{\kappa}_n) + (\beta-1)\eta(\boldsymbol{\kappa}_0)\right) - \alpha A(\eta(\boldsymbol{\kappa}_n)) - (\beta-1)A(\eta(\boldsymbol{\kappa}_0))\right\} \\ &\cdot \mathbb{E}_{p(\boldsymbol{\theta}|(\alpha\eta(\boldsymbol{\kappa}_n)+(\beta-1)\eta(\boldsymbol{\kappa}_0))}\left[h(\boldsymbol{\theta})^{\alpha+\beta-2}\right], \end{aligned}$$

where once again in analogy to G_1 and G_2, conditions (1) and (2) ensure this is a closed form function of $\eta(\boldsymbol{\kappa}_n)$ and $\eta(\boldsymbol{\kappa}_0)$.

Therefore, provided conditions (1) and(2) hold, the integrals G_1, G_2 and G_3 are available in closed form, implying that the same holds for $D_G^{(\alpha,\beta,r)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0))$. □

Remark B.2 (Conditions of Proposition B.1 for the MVN exponential family)**.** In order to illuminate the meaning and generality of the conditions of Theorem B.1, we apply them to the MVN exponential family described in Definition B.3. In this case the two conditions become:

i) For $\boldsymbol{\mu}^* := \left\{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 - \left(\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right)^{-1}\left(\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1}\boldsymbol{\mu}_2 + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\boldsymbol{\mu}_1\right)\right\}$ we require that

$$\begin{aligned} \alpha\begin{pmatrix}\boldsymbol{V}_1^{-1}\boldsymbol{\mu}_1 \\ -\frac{1}{2}\boldsymbol{V}_1^{-1}\end{pmatrix} + (\beta-1)\begin{pmatrix}\boldsymbol{V}_2^{-1}\boldsymbol{\mu}_2 \\ -\frac{1}{2}\boldsymbol{V}_2^{-1}\end{pmatrix} &= \begin{pmatrix}\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1}\boldsymbol{\mu}_1 + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\boldsymbol{\mu}_2 \\ -\frac{1}{2}\left\{\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right\}\end{pmatrix} \\ &= \begin{pmatrix}\left\{\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right\}\boldsymbol{\mu}^* \\ -\frac{1}{2}\left\{\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right\}\end{pmatrix} \in \mathcal{N} \end{aligned}$$

ii) $\mathbb{E}_{p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}))}\left[(2\pi)^{-d/2(\alpha+\beta+2)}\right] = (2\pi)^{-d/2(\alpha+\beta+2)} = f(\eta(\boldsymbol{\kappa}))$ where f is a closed form function.

Part ii) shows that the second condition is trivially satisfied for the MVN exponential family. Part i) shows that for the MVN exponential family, the first condition is satisfied provided $(V^*)^{-1} = \left\{\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right\}$ is a positive definite matrix. This condition is enough to ensure that V^* is invertible and thus that μ^* is well-defined. We elaborate further on what this means for certain parametrisations below.

Corollary: The special cases of $D_A^{(\alpha)}$, $D_{AR}^{(\alpha)}$

Next, we consider the $D_A^{(\alpha)}$ and $D_{AR}^{(\alpha)}$ special cases of the $D_G^{(\alpha,\beta,r)}$ family. Definitions 5.2 and 5.1 can be used to show that the $D_{AR}^{(\alpha)}$ is available as the following closed form function of the $D_A^{(\alpha)}$. In particular, it holds that

$$D_{AR}^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) = \frac{1}{\alpha(\alpha-1)} \log\left\{1 + \alpha(1-\alpha) D_A^{(\alpha)}(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))\right\}. \tag{B.4}$$

Thus, as demonstrated in Corollary B.2 below, the $D_A^{(\alpha)}$ being available in closed form immediately provides the $D_{AR}^{(\alpha)}$ in closed form. Before stating these results, we note that Gil et al. (2013); Gil (2011); Liese and Vajda (1987) have shown our closed form results for the $D_{AR}^{(\alpha)}$ (and thus implicitly the $D_A^{(\alpha)}$) before. We nevertheless think there is merit in stating them, since our results refer to the $D_G^{(\alpha,\beta,r)}$ and thus are more general, recovering both the $D_A^{(\alpha)}$ and $D_{AR}^{(\alpha)}$ only as a special case.

Corollary B.1 (Closed form $D_A^{(\alpha)}$ for exponential families)**.** The $D_A^{(\alpha)}$ between a variational posterior $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ and prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ is available in closed form under the following conditions

1. $(\alpha\eta(\boldsymbol{\kappa}_n) + (1-\alpha)\eta(\boldsymbol{\kappa}_0)) \in \mathcal{N}$

and in this case the $D_A^{(\alpha)}$ can be written as

$$D_A^{(\alpha)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0) = \frac{1}{\alpha(1-\alpha)}\left[1 - C(\boldsymbol{\kappa}_n, \boldsymbol{\kappa}_0, \alpha, (1-\alpha))\right],$$

where $C(\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2, \delta_1, \delta_2)$ was defined in Proposition B.1 .

Proof. Following Cichocki and Amari (2010) the single-parameter $D_A^{(\alpha)}$ is recovered as a member of the $D_G^{(\alpha,\beta,r)}$ family when $r = 1$ and $\beta = 2 - \alpha$. In this situation, Condition (2) of Theorem B.1 holds automatically and we are left with Condition (1). Substituting $\beta = 2 - \alpha$ provides Condition (1) of the Theorem above.

If $\alpha \in (0, 1)$ then the convexity of the natural parameter space ensures that providing $\eta(\boldsymbol{\kappa}_n) \in \mathcal{N}$ and $\eta(\boldsymbol{\kappa}_0) \in \mathcal{N}$ then $\alpha\eta(\boldsymbol{\kappa}_n) + (1-\alpha)\eta(\boldsymbol{\kappa}_0) \in \mathcal{N}$. If $\alpha < 0$ or $\alpha > 1$, then this can no longer be guaranteed. □

Corollary B.2 is then an immediate consequence of Corollary B.1.

Corollary B.2 (Closed form $D_{AR}^{(\alpha)}$ for exponential families)**.** The $D_{AR}^{(\alpha)}$ between a variational posterior $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ and prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ will have closed form providing the $D_A^{(\alpha)}$ between the same two densities for the same value of α has closed form.

Proof. The proof of this follows immediately from the fact that the $D_{AR}^{(\alpha)}$ can be recovered using the closed form function of the $D_{A}^{(\alpha)}$ shown in eq. (B.4) □

Remark B.3 (Conditions for Corollary B.1 for the MVN exponential family)**.** The condition that $\alpha\eta(\boldsymbol{\kappa}_n) + (1-\alpha)\eta(\boldsymbol{\kappa}_0) \in \mathcal{N}$ can only be guaranteed for $\alpha \in (0,1)$. However we can see from Remark B.2 that provided $\boldsymbol{V}^* = \left(\left(\frac{1}{\alpha}\boldsymbol{V}_1\right)^{-1} + \left(\frac{1}{\beta-1}\boldsymbol{V}_2\right)^{-1}\right)^{-1}$ is a symmetric semi-positive definite (SPD) matrix for $\beta = 2-\alpha$ then this condition will be satisfied. For $\alpha > 1$ or $\alpha < 0$ we cannot guarantee that $\boldsymbol{V}^*$ is SPD. However, we implement the $D_{AR}^{(\alpha)}$ to quantify uncertainty for $\alpha > 1$ in the main text. Corollary B.1 demonstrates that these parameters will still produce a closed form divergence provided the prior has sufficiently large variance, which can always be guaranteed to hold in practice.

Corollary: The special cases of $D_{B}^{(\beta)}$, $D_{G}^{(\gamma)}$

Next, we turn attention to the β- and γ-divergence families. Definition 5.4 shows that the $D_{G}^{(\gamma)}$ can be recovered as a closed form function of the terms of the $D_{B}^{(\beta)}$ and thus, as demonstrated in Corollary B.4 below, the $D_{B}^{(\beta)}$ being available in closed form immediately provides that the $D_{G}^{(\gamma)}$ is available in closed form While the conditions for these are slightly more restrictive than they were for the $D_{A}^{(\alpha)}$ and $D_{AR}^{(\alpha)}$, one can still obtain closed form prior regularizers for a large range of settings.

Corollary B.3 (Closed form $D_{B}^{(\beta)}$ for exponential families)**.** The $D_{B}^{(\beta)}$ between a variational posterior $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ and prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ is available in closed form under the following conditions

1. $\eta(\boldsymbol{\kappa}_1), \eta(\boldsymbol{\kappa}_2) \in \mathcal{N} \Rightarrow ((\beta-1)\eta(\boldsymbol{\kappa}_1) + \eta(\boldsymbol{\kappa}_2)) \in \mathcal{N}$
2. $\mathbb{E}_{p(\boldsymbol{\theta}|\eta(\boldsymbol{\kappa}))}\left[h(\boldsymbol{\theta})^{\beta-1}\right]$ is a closed form function of $\eta(\boldsymbol{\kappa}) \in \mathcal{N}$.

and in this case the $D_{B}^{(\beta)}$ can be written as

$$D_{B}^{(\beta)}(q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)||\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)) = \frac{1}{\beta(\beta-1)}B(\boldsymbol{\kappa}_n,\beta)E(\boldsymbol{\kappa}_n,\beta) + \frac{1}{\beta}B(\boldsymbol{\kappa}_0,\beta)E(\boldsymbol{\kappa}_0,\beta)$$
$$-\frac{1}{(\beta-1)}C(\boldsymbol{\kappa}_n,\boldsymbol{\kappa}_0,1,(\beta-1))\tilde{E}(\boldsymbol{\kappa}_n,\boldsymbol{\kappa}_0,1,(\beta-1)),$$

where the functions $B(\boldsymbol{\kappa},\delta)$, $C(\boldsymbol{\kappa}_1,\boldsymbol{\kappa}_2,\delta_1,\delta_2)$, $E(\boldsymbol{\kappa},\delta)$ and $\tilde{E}(\boldsymbol{\kappa}_1,\boldsymbol{\kappa}_2,\delta_1,\delta_2)$ are defined in Proposition B.1.

Proof. Following Cichocki and Amari (2010), the single-parameter $D_{B}^{(\beta)}$ is recovered as a member of the $D_{G}^{(\alpha,\beta,r)}$ family when $r=1$ and $\alpha=1$. In this situation, Condition (1)-(2) of Theorem B.1 become (1)-(2) above. □

Corollary B.4 is then an immediate consequence of Corollary B.3.

Corollary B.4 (Closed form $D_G^{(\gamma)}$ for exponential families)**.** The $D_G^{(\gamma)}$ between a variational posterior $q(\boldsymbol{\theta}|\boldsymbol{\kappa}_n)$ and prior $\pi(\boldsymbol{\theta}|\boldsymbol{\kappa}_0)$ will have closed form providing the $D_B^{(\beta)}$ between the same two densities with $\beta = \gamma$ has closed form.

Proof. The proof of this follows immediately from the fact that the $D_G^{(\gamma)}$ can be recovered from the $D_B^{(\beta)}$ using closed form function as outlined in Definition 5.4. □

Remark B.4 (Conditions for Corollary B.3 under the MVN exponential family)**.** Following Remark B.2, Corollary B.3 is satisfied providing $\boldsymbol{V}^* = \left((\boldsymbol{V}_n)^{-1} + \left(\frac{1}{\beta - 1} \boldsymbol{V}_0 \right)^{-1} \right)^{-1}$ is a symmetric SPD matrix. The sum of two symmetric SPD matrices is symmetric SPD and additionally the inverse of a symmetric SPD matrix is also SPD. Therefore provided $\beta > 1$ we can be sure that Condition iii) will be satisfied. Similarly to Remark B.3, when $\beta < 1$ closed forms will require that the prior has a sufficiently large variance.

In fact Remark B.4 can be extended to many other exponential families if we constrain $\beta = \gamma > 1$, this is formalised in Corollary B.5.

Corollary B.5 (Closed form $D_B^{(\beta)}$ and $D_G^{(\gamma)}$ for exponential families when $\beta = \gamma > 1$)**.** When $\beta = \gamma > 1$, the conditions for Corollary B.3 are satisfied by any exponential family whose $h(\boldsymbol{\theta})$ is a constant function of $\boldsymbol{\theta}$ and whose natural parameter space is closed under addition and scalar multiplication. This includes the Beta, Gamma, Gaussian, exponential and Laplace families.

Proof. The proof of Corollary B.5 follows straight from that of Corollary B.3. □

B.6 Log Trick (Taylor bound)

Lemma B.2 (A Taylor series bound for the natural logarithm)**.** The natural logarithm of a positive real number Z can be bounded as follows

$$\begin{cases} \log(Z) \leq \frac{Z^x - 1}{x} & \text{if } x > 0 \\ \log(Z) \geq \frac{Z^x - 1}{x} & \text{if } x < 0. \end{cases}$$

Proof. Using the series expansion of $\exp(x)$ and the Lagrange form of the remainder

we see that

$$\frac{Z^x - 1}{x} = \frac{\exp\left(x \log Z\right) - 1}{x} = \frac{\left(x \log Z\right) + \frac{1}{2!}\left(x \log Z\right)^2 + \frac{1}{3!}\left(x \log Z\right)^3 + \ldots}{x}$$
$$= \frac{\left(x \log Z\right) + \frac{1}{2}\exp(c)\left(x \log Z\right)^2}{x} = \log Z + \frac{\frac{1}{2!}\exp(c)\left(x \log Z\right)^2}{x}$$

where $c \in [0, x \log(Z)]$. Now the numerator of the remainder term $\frac{\frac{1}{2!}\exp(c)(x \log Z)^2}{x}$ is always positive and therefore the sign of x determines whether this remainder term forms an upper or lower bound for $\log(Z)$. □

B.7 Derivations for DGPs

B.7.1 Proof of Theorem 6.1

Proof. The likelihood is Gaussian with a fixed variance parameter σ^2, i.e. for $\boldsymbol{y}_i \in \mathbb{R}^d$ with $i = 1, 2, \ldots, n$

$$p(\boldsymbol{y}_i | \boldsymbol{f}_i^L) = (2\pi\sigma^2)^{-0.5d} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{f}_i^L)^T(\boldsymbol{y}_i - \boldsymbol{f}_i^L)\right\}$$

With this, note that integrating out the normal density yields

$$I_{p,c}(\boldsymbol{f}_i^L) = (2\pi\sigma^2)^{-0.5dc} c^{-0.5d}. \tag{B.5}$$

Note in particular that this is a constant and does not depend on $\boldsymbol{f}$, which makes computing the expectation over $q(\boldsymbol{f}_i^L)$ depend only on the power likelihood. Next, we show that the power likelihood is also available in closed form. This is laborious but not difficult and relies on the same algebraic tricks in the Appendix of Knoblauch et al. (2018). To simplify notation, we write $\boldsymbol{f} = \boldsymbol{f}_i^L$. Note also that the variational parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are (stochastic) functions of the draws of $\boldsymbol{f}_i^{1:L-1}$ from the

previous layers, but we suppress this dependency, again for readability.

$$
\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\frac{1}{c}p(\boldsymbol{y}_i|\boldsymbol{f})^c\right] \\
&= \frac{1}{c}(2\pi\sigma^2)^{-0.5dc}\cdot\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\exp\left\{-\frac{c}{2\sigma^2}(\boldsymbol{y}_i^T\boldsymbol{y}_i+\boldsymbol{f}^T\boldsymbol{f}-2\boldsymbol{f}^T\boldsymbol{y}_i)\right\}\right] \\
&= \frac{1}{c}(2\pi\sigma^2)^{-0.5dc}\exp\left\{-\frac{c}{2\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i\right\}\cdot\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\exp\left\{-\frac{c}{2\sigma^2}(\boldsymbol{f}^T\boldsymbol{f}-2\boldsymbol{f}^T\boldsymbol{y}_i)\right\}\right] \\
&= \frac{1}{c}(2\pi\sigma^2)^{-0.5dc}(2\pi\sigma^2)^{-0.5d}|\boldsymbol{\Sigma}|^{-0.5}\exp\left\{-\frac{c}{2\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i\right\}\times \\
&\quad\int\exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{f}-\frac{2c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{y}_i+(\boldsymbol{f}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{f}-\boldsymbol{\mu})\right)\right\}d\boldsymbol{f} \\
&= \frac{1}{c}(2\pi\sigma^2)^{-0.5dc}(2\pi)^{-0.5d}|\boldsymbol{\Sigma}|^{-0.5}\exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\}\times \\
&\quad\int\exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{f}-\frac{2c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{y}_i+\boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{f}-2\boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\}d\boldsymbol{f}
\end{aligned}
$$

The integral suggests one can obtain a closed form through the Gaussian integral by completing the squares. Defining $\widetilde{\boldsymbol{\Sigma}}^{-1} = \left(\frac{c}{\sigma^s}\boldsymbol{I}_d+\boldsymbol{\Sigma}^{-1}\right)$, $\widetilde{\boldsymbol{\mu}} = \left(\frac{c}{\sigma^2}\boldsymbol{y}_i+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$, $\widehat{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}$, one indeed has

$$
\begin{aligned}
&\frac{c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{f}-\frac{2c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{y}_i+\boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{f}-2\boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
&= \boldsymbol{f}^T\left(\boldsymbol{I}_d\frac{c}{\sigma^2}+\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{f}-2\boldsymbol{f}^T\left(\frac{c}{\sigma^2}\boldsymbol{y}_i+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\
&= (\boldsymbol{f}-\widehat{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{f}-\widehat{\boldsymbol{\mu}})-\widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}},
\end{aligned}
$$

which allows us to finally rewrite the integral as

$$
\begin{aligned}
&\int\exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{f}-\frac{2c}{\sigma^2}\boldsymbol{f}^T\boldsymbol{y}_i+\boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{f}-2\boldsymbol{f}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right\}d\boldsymbol{f} \\
&= \exp\left\{-\frac{1}{2}\widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}\right\}\int\exp\left\{-\frac{1}{2}(\boldsymbol{f}-\widehat{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{f}-\widehat{\boldsymbol{\mu}})\right\}d\boldsymbol{f} = \exp\left\{\frac{1}{2}\widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}\right\}(2\pi)^{0.5d}|\widetilde{\boldsymbol{\Sigma}}|^{0.5}.
\end{aligned}
$$

Putting everthing together and simplifying expressions, this means that

$$
\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[\frac{1}{c}p(\boldsymbol{y}_i|\boldsymbol{f})^c\right] = \frac{1}{c}\left(2\pi\sigma^2\right)^{-0.5dc}\frac{|\widetilde{\boldsymbol{\Sigma}}|^{0.5}}{|\boldsymbol{\Sigma}|^{0.5}}\exp\left\{-\frac{1}{2}\left(\frac{c}{\sigma^2}\boldsymbol{y}_i^T\boldsymbol{y}_i+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}^T\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\mu}}\right)\right\}
$$

Depending on whether one uses the β- or γ-divergence for robustifying the loss, one thus obtains the closed form expressions

$$\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-\frac{1}{\beta-1}p(\boldsymbol{y}_i|\boldsymbol{f})^{\beta-1}+\frac{I_{p,\beta}(\boldsymbol{f})}{\beta}\right]=\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-\frac{1}{\beta-1}p(\boldsymbol{y}_i|\boldsymbol{f})^{\beta-1}\right]+\frac{I_{p,\beta}(\boldsymbol{f})}{\beta}$$

$$\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-\frac{1}{\gamma-1}p(\boldsymbol{y}_i|\boldsymbol{f})^{\gamma-1}\cdot\frac{\gamma}{I_{p,\gamma}(\boldsymbol{f})^{\frac{\gamma-1}{\gamma}}}\right]=\mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\left[-\frac{1}{\gamma-1}p(\boldsymbol{y}_i|\boldsymbol{f})^{\gamma-1}\right]\cdot\frac{\gamma}{I_{p,\gamma}(\boldsymbol{f})^{\frac{\gamma-1}{\gamma}}},$$

with the expectation over $q(\boldsymbol{f}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ as in and the integrals $I_{p,\beta}(\boldsymbol{f})$, $I_{p,\gamma}(\boldsymbol{f})$ as defined above. Note that we have derived the general case for $\boldsymbol{y}_i\in\mathbb{R}^d$, where $\boldsymbol{\Sigma}$, $\boldsymbol{f}$ and $\boldsymbol{\mu}$ are matrix- and vector-valued. □

In fact, we can simplify everything even further in the univariate case. We summarize this in the next part.

Remark B.5. Since the derivation of Salimbeni and Deisenroth (2017) shows that one in fact only needs to integrate over the marginals $\boldsymbol{f}_i^L$, if $d=1$ (as in all experiments in both that paper and (Salimbeni and Deisenroth, 2017)), the computation corresponding to the expression above simplifies considerably as no matrix inverses and determinants are needed. In particular, denoting the uni-variate mean and variance parameters as μ,Σ and defining $\widetilde{\Sigma}=\frac{1}{\frac{c}{\sigma^s}+\frac{1}{\Sigma}}$ and $\widetilde{\mu}=\left(\frac{cy_i}{\sigma^2}+\frac{\mu}{\Sigma}\right)$, the expectation term over the posterior q simplifies to

$$\mathbb{E}_{q(f|\mu,\Sigma)}\left[\frac{1}{c}p(y_i|f)^c\right]=\frac{1}{c}s\left(2\pi\sigma^2\right)^{-0.5c}\sqrt{\frac{\widetilde{\Sigma}}{\Sigma}}\cdot\exp\left\{-\frac{1}{2}\left(\frac{cy_i^2}{\sigma^2}+\frac{\mu^2}{\Sigma}-\widetilde{\mu}^2\widetilde{\Sigma}\right)\right\}.$$

B.7.2 Proof of Corollary 6.1

We first prove a Lemma that plays a key role in the proof of Corollary 6.1.

Lemma B.3 (Divergence recombination)**.** Let D_l be divergences and $c_l>0$ scalars for $l=1,2,\ldots,L$. Further, denote $\boldsymbol{\theta}_{-l}=\boldsymbol{\theta}_{1:l-1,l+t:L}$ and let $q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$ and $\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$ be the conditional distributions of $\boldsymbol{\theta}_l$ for $q(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ conditioned on $\boldsymbol{\theta}_{-l}=\boldsymbol{\theta}'_{-l}$. Then, $D^{\boldsymbol{\theta}'}(q||\pi)=\sum_{l=1}^L c_l D_l\left(q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})||\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})\right)$ is a divergence between $q(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ if (i) $D^{\boldsymbol{\theta}^\circ}(q||\pi)=D^{\boldsymbol{\theta}'}(q||\pi)$ for all conditioning sets $\boldsymbol{\theta}^\circ$, $\boldsymbol{\theta}'$ and (ii) a Hammersley-Clifford Theorem holds for the collection of conditionals $\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$ and $q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$.

Proof. First, observe by definition of a divergence, $D_l(q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})||\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l}))=0$ for all l and over all potential conditioning sets $\boldsymbol{\theta}'$ holds if and only if $q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})=\pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$. Next, note that we have assumed that $D^{\boldsymbol{\theta}'}(q||\pi)=D^{\boldsymbol{\theta}^\circ}(q||\pi)$ for all

conditioning sets $\boldsymbol{\theta}'$, $\boldsymbol{\theta}^\circ$. In other words, if $D^{\boldsymbol{\theta}'}(q||\pi) = 0$ for some $\boldsymbol{\theta}'$, then it will also be 0 for *any* conditioning set $\boldsymbol{\theta}^\circ$. This immediately entails that for arbitrary $\boldsymbol{\theta}'$, $D^{\boldsymbol{\theta}'}(q||\pi) = 0$ if and only if $q_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l}) = \pi_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}'_{-l})$ for *all* l and for *any* choice of $\boldsymbol{\theta}'_{-l}$. In other words, the conditionals are the same. Since the positivity condition holds, we can then apply the Hammersley-Clifford Theorem to conclude that the conditionals fully specify the joint. This finally yields the desired result: $D^{\boldsymbol{\theta}'}(q||\pi) = 0$ if and only if $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$. □

With this technical result in hand, one can now prove the result, which shows that reverse-engineering prior regularizers inspired by eq. (**??**) is feasible so long as the layer-specific divergences D^l are f-divergences or monotonic transformations of f-divergences.

Proof. Suppressing again $\boldsymbol{Z}^l$ and $\boldsymbol{X}$ for readability, first recall that

$$q(\{\boldsymbol{U}^l\}_{l=1}^L, \{\boldsymbol{F}^l\}_{l=1}^L) = \prod_{l=1}^L p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}^{l-1}) q(\boldsymbol{U}^l)$$
$$p(\{\boldsymbol{U}^l\}_{l=1}^L, \{\boldsymbol{F}^l\}_{l=1}^L) = \prod_{l=1}^L p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}^{l-1}) p(\boldsymbol{U}^l)$$

and write for a *fixed* conditioning set $\{\boldsymbol{F}_\circ^l\}_{l=1}^L$ the new divergence

$$\begin{aligned} &D^{\{\boldsymbol{F}_\circ^l\}_{l=1}^L}\left(q(\{\boldsymbol{U}_l\}_{l=1}^L, \{\boldsymbol{F}_l\}_{l=1}^L) \| p(\{\boldsymbol{U}_l\}_{l=1}^L, \{\boldsymbol{F}_l\}_{l=1}^L)\right) \\ &= \sum_{l=1}^L D^l\left(p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) q(\boldsymbol{U}^l) \| p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) p(\boldsymbol{U}^l)\right) = \sum_{l=1}^L D^l\left(q(\boldsymbol{U}^l) \| p(\boldsymbol{U}^l)\right) \end{aligned}$$

The first equality is simply the definition of the new divergence. The second equality follows by virtue of D^l being a monotonic function of an f-divergences or an f-divergence for all l, which ensures that the l-th term is given by

$$\begin{aligned} &D^l\left(p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) q(\boldsymbol{U}^l) \| p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) p(\boldsymbol{U}^l)\right) \\ &= g\left(\mathbb{E}_{p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) p(\boldsymbol{U}^l)}\left[f\left(\frac{p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) q(\boldsymbol{U}^l)}{p(\boldsymbol{F}^l|\boldsymbol{U}^l, \boldsymbol{F}_\circ^{l-1}) p(\boldsymbol{U}^l)}\right)\right]\right). \\ &= g\left(\mathbb{E}_{p(\boldsymbol{U}^l)}\left[f\left(\frac{q(\boldsymbol{U}^l)}{p(\boldsymbol{U}^l)}\right)\right]\right) = D^l\left(q(\boldsymbol{U}^l) \| p(\boldsymbol{U}^l)\right). \end{aligned} \tag{B.6}$$

Now note that we can invoke Lemma B.3: The first condition is satisfied because the derivation was independent of the chosen $\{\boldsymbol{F}_\circ^l\}_{l=1}^L$. The second condition is satisfied as both conditionals satisfy the positivity condition required for the Hammersley-

Clifford Theorem to hold. □

B.8 Derivations for the robust GVI objective for Bayesian On-line Changepoint Detection (with Model Selection)

B.8.1 Proof of Theorem 7.3

Proof. For ease of notation and convenience, we use $\beta_{\mathrm{m}} = \beta + 1$. The model used for the inference is an exponential family model of the form

$$f(x;\theta) = \exp\left(\eta(\theta)^T T(x)\right) g(\eta(\theta)) A(x),$$

where $g(\eta(\theta)) := \left(\int \exp\left(\eta(\theta)^T T(x)\right) A(x) dx\right)^{-1}$. The posterior arising from this model and its conjugate prior is approximated by a member of the conjugate prior family. As a result, the conjugate prior and variational posterior to the above model have the form

$$\begin{aligned}\pi_0(\theta|\nu_0,\mathcal{X}_0) &= g(\eta(\theta))^{\nu_0} \exp\left(\nu_0 \eta(\theta)^T \mathcal{X}_0\right) h(\mathcal{X}_0,\nu_0)\\ \pi_n^{VB}(\theta|\nu_n,\mathcal{X}_n) &= g(\eta(\theta))^{\nu_n} \exp\left(\nu_n \eta(\theta)^T \mathcal{X}_n\right) h(\mathcal{X}_n,\nu_n),\end{aligned}$$

where $(\nu_0, \mathcal{X}_0)$ are the prior hyperparameters, $(\nu_n, \mathcal{X}_n)$ represent the variational parameters and $h(\mathcal{X}_i, \nu_i) := \left(\int g(\eta(\theta))^{\nu_i} \exp\left(\nu_i \eta(\theta)^T \mathcal{X}_i\right) d\theta\right)^{-1}$. The resulting objective function has the form

$$\begin{aligned}&O_{\mathrm{GVI}}(\nu_n,\mathcal{X}_n) = \\ &\quad \mathbb{E}_{\pi_n^{VB}}\left[\log\left(\exp\left(\sum_{i=1}^{n} -\ell^D(x;\theta)\right)\right)\right] - d_{KL}\left(\pi_n^{VB} \| \pi_0\left(\theta|\nu_0,\mathcal{X}_0\right)\right),\end{aligned}$$

where for the $D_B^{(\beta)}$ posterior

$$
\begin{aligned}
-\ell^{\beta}(x;\theta) = & \frac{1}{\beta}\left(\exp\left(\eta(\theta)^T T(x)\right) g(\eta(\theta))A(x)\right)^{\beta} - \\
& \frac{1}{\beta+1}\int \left(\exp\left(\eta(\theta)^T T(z)\right) g(\eta(\theta))A(x)\right)^{1+\beta} dz \\
= & \frac{1}{\beta}\exp\left(\beta\eta(\theta)^T T(x)\right) g(\eta(\theta))^{\beta} A(x)^{\beta} - \\
& \frac{1}{\beta+1}\int \exp\left((1+\beta)\eta(\theta)^T T(z)\right) g(\eta(\theta))^{1+\beta} A(x)^{1+\beta} dz.
\end{aligned}
$$

Therefore the $O_{\text{GVI}}(\nu_n, \mathcal{X}_n)$ has three integrals that need evaluating

$$
B_1 = \sum_{i=1}^{n} \int \frac{1}{\beta}\exp\left(\beta\eta(\theta)^T T(x_i)\right) g(\eta(\theta))^{\beta} A(x_i)^{\beta} \pi_n^{VB}(\theta|\nu_n, \mathcal{X}_n) d\theta \tag{B.7}
$$

$$
\begin{aligned}
B_2 = & \frac{n}{\beta+1}\int \left\{ \int \exp\left((1+\beta)\eta(\theta)^T T(z)\right) g(\eta(\theta))^{1+\beta} A(z))^{1+\beta} dz \right\} \times \\
& \pi_n^{VB}(\theta|\nu_n, \mathcal{X}_n) d\theta
\end{aligned} \tag{B.8}
$$

$$
B_3 = \text{KLD}\left(\pi_n^{VB}(\theta|\nu_n, \mathcal{X}_n), \pi_0(\theta|\nu_0, \mathcal{X}_0)\right). \tag{B.9}
$$

Now firstly for the term B_1 in (B.7)

$$
\begin{aligned}
B_1 = & \sum_{i=1}^{n} \int \frac{1}{\beta}\exp\left(\beta\eta(\theta)^T T(x_i)\right) g(\eta(\theta))^{\beta} A(x_i)^{\beta} g(\eta(\theta))^{\nu_n} \exp\left(\nu_n \eta(\theta)^T \mathcal{X}_n\right) h(\mathcal{X}_n, \nu_n) d\theta \\
= & \sum_{i=1}^{n} \frac{1}{\beta} A(x_i)^{\beta} h(\mathcal{X}_n, \nu_n) \int g(\eta(\theta))^{\beta+\nu_n} \exp\left(\eta(\theta)^T \left(\beta T(x_i) + \nu_n \mathcal{X}_n\right)\right) d\theta \\
= & \sum_{i=1}^{n} \frac{1}{\beta} A(x_i)^{\beta} h(\mathcal{X}_n, \nu_n) \frac{1}{h(\frac{\beta T(x_i) + \nu_n \mathcal{X}_n}{\beta+\nu_n}, \beta+\nu_n)}.
\end{aligned}
$$

Where we know that

$$
h\left(\frac{\beta T(x_i) + \nu_n \mathcal{X}_n}{\beta+\nu_n}, \beta+\nu_n\right) = \int g(\eta(\theta))^{\beta+\nu_n} \exp\left(\eta(\theta)^T \left(\beta T(x_i) + \nu_n \mathcal{X}_n\right)\right) d\theta
$$

is integrable and closed form as it represents the normalizing constant of the same exponential family as the prior and the variational posterior. Next we look at B_2 in equation (B.8). The whole integral is the product of two densities which must be positive and in order for the $O_{\text{GVI}}(\nu_n, \mathcal{X}_n)$ to be defined it must also be integrable. Therefore we can use Fubini's theorem to switch the order of integration

$$
\begin{aligned}
B_2 &= \frac{n}{\beta+1}\int\left\{\int \exp\left((1+\beta)\eta(\theta)^T T(z)\right) g(\eta(\theta))^{1+\beta}\pi_n^{VB}(\theta|\nu_n,\mathcal{X}_n)d\theta\right\}A(z)^{1+\beta}dz \\
&= \frac{n}{\beta+1}h(\mathcal{X}_n,\nu_n)\int\left\{\int \exp\left(\eta(\theta)^T\left((1+\beta)T(z)+\nu_n\mathcal{X}_n\right)\right) g(\eta(\theta))^{1+\beta+\nu_n}d\theta\right\}A(z)^{1+\beta}dz \\
&= \frac{n}{\beta+1}h(\mathcal{X}_n,\nu_n)\int \frac{A(z)^{1+\beta}}{h(\frac{(1+\beta)T(z)+\nu_n\mathcal{X}_n}{1+\beta+\nu_n}, 1+\beta+\nu_n)}dz.
\end{aligned}
$$

once again,

$$
h(\frac{(1+\beta)T(z)+\nu_n\mathcal{X}_n}{1+\beta+\nu_n}, 1+\beta+\nu_n) = \int \exp\left(\eta(\theta)^T\left((1+\beta)T(z)+\nu_n\mathcal{X}_n\right)\right) g(\eta(\theta))^{1+\beta+\nu_n}d\theta
$$

is the normalizing constant of the same exponential family as the prior and the variational posterior and is thus closed form. Lastly we look at B_3 in equation (B.9)

$$
\begin{aligned}
B_3 &= \int \pi_n^{VB}(\theta|\nu_n,\mathcal{X}_n)\log\frac{g(\eta(\theta))^{\nu_n}\exp\left(\nu_n\eta(\theta)^T\mathcal{X}_n\right)h(\mathcal{X}_n,\nu_n)}{g(\eta(\theta))^{\nu_0}\exp\left(\nu_0\eta(\theta)^T\mathcal{X}_0\right)h(\mathcal{X}_0,\nu_0)} \\
&= \log\frac{h(\mathcal{X}_n,\nu_n)}{h(\mathcal{X}_0,\nu_0)}\int \pi_n^{VB}(\theta|\nu_n,\mathcal{X}_n)\left\{(\nu_n-\nu_0)\log g(\eta(\theta)) + \left(\eta(\theta)^T\left(\nu_n\mathcal{X}_n-\nu_0\mathcal{X}_0\right)\right)\right\} \\
&= \log\frac{h(\mathcal{X}_n,\nu_n)}{h(\mathcal{X}_0,\nu_0)}\left\{(\nu_n-\nu_0)\lambda_n^{VB} + \left((\mu_n^{VB})^T\left(\nu_n\mathcal{X}_n-\nu_0\mathcal{X}_0\right)\right)\right\},
\end{aligned}
$$

where $\mu_n^{VB} = \mathbb{E}_{\pi_n^{VB}}[\eta(\theta)]$ and $\lambda_n^{VB} = \mathbb{E}_{\pi_n^{VB}}[\log g(\eta(\theta))]$.
As a result we get that

$$
\begin{aligned}
O_{\text{GVI}}(\nu_n,\mathcal{X}_n) &= B_1 - B_2 - B_3 \\
&= \sum_{i=1}^{n}\frac{1}{\beta}A(x_i)^{\beta}h(\mathcal{X}_n,\nu_n)\frac{1}{h(\frac{\beta T(x_i)+\nu_n\mathcal{X}_n}{\beta+\nu_n}, \beta+\nu_n)} \\
&\quad - \frac{n}{\beta+1}h(\mathcal{X}_n,\nu_n)\int \frac{A(z)^{1+\beta}}{h(\frac{(1+\beta)T(z)+\nu_n\mathcal{X}_n}{1+\beta+\nu_n}, 1+\beta+\nu_n)}dz \\
&\quad - \log\frac{h(\mathcal{X}_n,\nu_n)}{h(\mathcal{X}_0,\nu_0)}\left\{(\nu_n-\nu_0)\lambda_n^{VB} + \left((\mu_n^{VB})^T\left(\nu_n\mathcal{X}_n-\nu_0\mathcal{X}_0\right)\right)\right\}.
\end{aligned}
$$

□

B.8.2 Derivation of closed form GVI objective & its derivative

Because it simplifies notation and derivations, we use $1+\beta = \beta_{\mathrm{m}}$ and derive all closed forms in terms of β (rather than β_{m}). Furthermore, we will suppress conditioning on the model m since the GVI parameter posterior $\pi_m^{\beta_{\mathrm{m}}}$ has to be derived for each model m. To simplify notation, we therefore denote a generic GVI posterior to be computed as π^β.

With this notation in place, recall that we wish to approximate the posterior belief distribution $\pi^\beta(\boldsymbol{\mu}, \sigma^2|x)$ which for observations $x = (x_1, x_2, \ldots, x_n)^T$ with $x_i \in \mathbb{R}^d$, prior $\mathrm{NIG}^0(\boldsymbol{\mu}, \sigma^2|a_0, b_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, model likelihood f and density power divergence (DPD) loss

$$\ell^\beta(\boldsymbol{\mu}, \sigma^2|x_i) = \frac{1}{\beta} f(x_i|\boldsymbol{\mu}, \sigma^2)^\beta - \frac{1}{1+\beta}\int_{\mathcal{Y}} f(x_i|\boldsymbol{\mu}, \sigma^2)^{1+\beta} dx$$

is given by

$$\pi^\beta(\boldsymbol{\mu}, \sigma^2|x) \propto \mathrm{NIG}^0(\boldsymbol{\mu}, \sigma^2|a_0, b_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \cdot \exp\left\{-\sum_{i=1}^n \ell^\beta(\boldsymbol{\mu}, \sigma^2|x_i)\right\}.$$

In particular, we want to approximate it with a posterior $\mathrm{NIG}^{\mathrm{VB}}(\boldsymbol{\mu}, \sigma^2|\widehat{a}_n, \widehat{b}_n, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)$ via Variational Bayes. This can be done by minimizing the variational parameters in a Kullback-Leibler sense:

$$(a^*, b^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \underset{(\widehat{a}_n, \widehat{b}_n, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)}{\operatorname{argmin}} \left\{\mathrm{KLD}\left(\pi^\beta(\boldsymbol{\mu}, \sigma^2|x) \,\middle\|\, \mathrm{NIG}^{\mathrm{VB}}(\boldsymbol{\mu}, \sigma^2|\widehat{a}_n, \widehat{b}_n, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)\right)\right\}.$$

It is straightforward to rewrite the objective function for the above minimization as the objective targeted by the GVI posterior. Throughout, we will call this objective O_{GVI}

$$O_{\mathrm{GVI}} = \underbrace{-\mathrm{KLD}\left(\mathrm{NIG}^{\mathrm{VB}}(\boldsymbol{\mu}, \sigma^2|\widehat{a}_n, \widehat{b}_n, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n) \,\middle\|\, \mathrm{NIG}^0(\boldsymbol{\mu}, \sigma^2|a_0, b_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\right)}_{=Q_1}.$$
$$\underbrace{-\mathbb{E}_{\mathrm{VB}}\left[-\sum_{i=1}^n \ell^\beta(\boldsymbol{\mu}, \sigma^2|x_i)\right]}_{=Q_2}.$$

In what follows, closed forms are derived for both Q_1 and Q_2. Some algebraic tricks will be applied multiple times, and will be referred to by the following symbols:

■ Completion of Squares, i.e. $\boldsymbol{u}'\boldsymbol{A}\boldsymbol{u}-2\boldsymbol{v}'\boldsymbol{u} = (\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{v})'\boldsymbol{A}(\boldsymbol{u}-\boldsymbol{A}^{-1}\boldsymbol{v})-\boldsymbol{v}'\boldsymbol{A}^{-1}\boldsymbol{v}$;

$I(\mathcal{N})$ Integrating out the Normal density;

$I(\mathcal{IG})$ Integrating out the Inverse Gamma density.

Throughout, the dimensionality of $\boldsymbol{\mu}$ is $p \in \mathbb{N}$, $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ refers to a normal pdf in $\boldsymbol{\mu}$ with expectation $\boldsymbol{\mu}_0$, variance $\boldsymbol{\Sigma}_0$ and $\mathcal{IG}(\sigma^2|a, b)$ to an inverse gamma pdf in σ^2 with shape a and scale b.

B.8.3 Q_1

First, note that by definition,

$$Q_1 = \int_{\boldsymbol{\mu},\sigma^2} \underbrace{\log\left(\frac{\text{NIG}^{\text{VB}}(\boldsymbol{\mu}, \sigma^2|\widehat{a}_n, \widehat{b}_n, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n)}{\text{NIG}^{0}(\boldsymbol{\mu}, \sigma^2|a_0, b_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}\right)}_{=Q_1^{\log}} \text{NIG}^{\text{VB}}(\boldsymbol{\mu}, \sigma^2|\widehat{a}_n, \widehat{b}_n, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n) d\boldsymbol{\mu} d\sigma^2.$$

Writing out $Q_1^{\log}$, one obtains a natural sum of three components $C_1, C_2(\sigma^2), C_3(\sigma^2, \boldsymbol{\mu})$:

$$\begin{aligned}
Q_1^{\log} &= \log\left(\frac{|\widehat{\boldsymbol{\Sigma}}_n|^{-0.5}\frac{\widehat{b}_n^{\widehat{a}_n}}{\Gamma(\widehat{a}_n)}(\sigma^2)^{-0.5p-\widehat{a}_n-1}\exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n)'\widehat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n)+2\widehat{b}_n\right]\right\}}{|\boldsymbol{\Sigma}_0|^{-0.5}\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-0.5p-a_0-1}\exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)+2b_0\right]\right\}}\right) \\
&= \underbrace{\log\left(\frac{\widehat{b}_n^{\widehat{a}_n}\Gamma(a_0)}{b_0^{a_0}\Gamma(\widehat{a}_n)}\right) + 0.5\log\left|\boldsymbol{\Sigma}_0\widehat{\boldsymbol{\Sigma}}_n^{-1}\right|}_{=C_1} + \underbrace{(\widehat{a}_n - a_0)\log(\frac{1}{\sigma^2})}_{=C_2(\sigma^2)} \\
&\quad \underbrace{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n)'\widehat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n) - (\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + 2(\widehat{b}_n - b_0)\right]}_{=C_3(\sigma^2,\boldsymbol{\mu})}.
\end{aligned}$$

Next, note that $C_3(\sigma^2, \boldsymbol{\mu})$ further decomposes into

$$\underbrace{\underbrace{\frac{1}{2\sigma^2}\left[\boldsymbol{\mu}'\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}-\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}'\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right]}_{=C_4(\sigma^2,\boldsymbol{\mu})} + \underbrace{\frac{1}{\sigma^2}\underbrace{\left[\frac{1}{2}\widehat{\boldsymbol{\mu}}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n - \frac{1}{2}\boldsymbol{\mu}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + (\widehat{b}_n - b_0)\right]}_{=C_5}}_{=C_6(\sigma^2)}}_{}.$$

Notice that we have isolated the random variable $\boldsymbol{\mu}$ inside $C_4(\sigma^2, \boldsymbol{\mu})$ and that by definition, $\text{NIG}^{\text{VB}}(\boldsymbol{\mu}, \sigma^2|\widehat{a}_n, \widehat{b}_n, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n) = \mathcal{N}^{\text{VB}}(\boldsymbol{\mu}|\widehat{\boldsymbol{\mu}}_n, \sigma^2\widehat{\boldsymbol{\Sigma}}_n)\cdot\mathcal{IG}^{\text{VB}}(\sigma^2|\widehat{a}_n, \widehat{b}_n)$, meaning

that

$$Q_1 = C_1 + \int_{\sigma^2} \left\{C_2(\sigma^2) - C_6(\sigma^2)\right\} \mathcal{IG}^{\text{VB}}(\sigma^2|\widehat{a}_n, \widehat{b}_n) d\sigma^2 \\ - \int_{\sigma^2} \underbrace{\left\{ \int_{\boldsymbol{\mu}} C_4(\sigma^2, \boldsymbol{\mu}) \mathcal{N}^{\text{VB}}(\boldsymbol{\mu}|\widehat{\boldsymbol{\mu}}_n, \sigma^2 \widehat{\boldsymbol{\Sigma}}_n) d\boldsymbol{\mu} \right\}}_{=C_7(\sigma^2)} \mathcal{IG}^{\text{VB}}(\sigma^2|\widehat{a}_n, \widehat{b}_n) d\sigma^2.$$

The inner integral is available in closed form, and naturally decomposes as

$$\begin{aligned} C_7(\sigma^2) &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathcal{N}^{\text{VB}}} \left[\boldsymbol{\mu}' \left(\widehat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}_0^{-1} \right) \boldsymbol{\mu} \right] - \frac{2}{2\sigma^2} \mathbb{E}_{\mathcal{N}^{\text{VB}}} \left[\boldsymbol{\mu}' \right] \left(\widehat{\boldsymbol{\Sigma}}_n^{-1} \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathcal{N}^{\text{VB}}} \left[\text{tr} \left(\left(\widehat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}_0^{-1} \right) \boldsymbol{\mu} \boldsymbol{\mu}' \right) \right] - \frac{1}{\sigma^2} \widehat{\boldsymbol{\mu}}_n' \left(\widehat{\boldsymbol{\Sigma}}_n^{-1} \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \\ &= \frac{1}{2\sigma^2} \text{tr} \left(\left(\widehat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}_0^{-1} \right) \mathbb{E}_{\mathcal{N}^{\text{VB}}} \left[\boldsymbol{\mu} \boldsymbol{\mu}' \right] \right) - \frac{1}{\sigma^2} \widehat{\boldsymbol{\mu}}_n' \left(\widehat{\boldsymbol{\Sigma}}_n^{-1} \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \\ &= \frac{1}{2\sigma^2} \text{tr} \left(\left(\widehat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}_0^{-1} \right) \left[\sigma^2 \widehat{\boldsymbol{\Sigma}}_n - \widehat{\boldsymbol{\mu}}_n \widehat{\boldsymbol{\mu}}_n' \right] \right) - \frac{1}{\sigma^2} \widehat{\boldsymbol{\mu}}_n' \left(\widehat{\boldsymbol{\Sigma}}_n^{-1} \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \\ &= \underbrace{\frac{1}{2} \text{tr} \left(I - \boldsymbol{\Sigma}_0^{-1} \widehat{\boldsymbol{\Sigma}}_n \right)}_{=C_8} - \underbrace{\frac{1}{\sigma^2} \underbrace{\left[\frac{1}{2} \widehat{\boldsymbol{\mu}}_n' (\widehat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}_0^{-1}) \widehat{\boldsymbol{\mu}}_n - \widehat{\boldsymbol{\mu}}_n' \left(\widehat{\boldsymbol{\Sigma}}_n^{-1} \widehat{\boldsymbol{\mu}}_n - \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \right]}_{=C_9}}_{=C_{10}(\sigma^2)}. \end{aligned}$$

We may now rewrite Q_1 so as to integrate out σ^2 next:

$$Q_1 = C_1 - C_8 + \int_{\sigma^2} \left\{C_2(\sigma^2) - C_6(\sigma^2) - C_{10}(\sigma^2)\right\} \mathcal{IG}^{\text{VB}}(\sigma^2|\widehat{a}_n, \widehat{b}_n) d\sigma^2.$$

Using the additivity of integrals, we consider its three components separately and then add them up together afterwards. For $C_2(\sigma^2)$, (I) apply a change of variable with $z = \frac{\sigma^2}{\widehat{b}_n}$ and then use (II) that $\frac{d}{dx} a^{-x} = -a^x \cdot \log(a) = a^x \cdot \log(a^{-1})$ together

with Fubini's Theorem (III) to find that

$$
\begin{aligned}
C_{11} &= \int_{\sigma^2} C_2(\sigma^2)\mathcal{IG}^{\mathrm{VB}}(\sigma^2|\widehat{a}_n,\widehat{b}_n)d\sigma^2 \\
&= (\widehat{a}_n - a_0)\int_{\sigma^2}\log\left(\frac{1}{\sigma^2}\right)\frac{\widehat{b}_n^{\widehat{a}_n}}{\Gamma(\widehat{a}_n)}(\sigma^2)^{-\widehat{a}_n-1}\exp\left\{-\frac{\widehat{b}_n}{\sigma^2}\right\}d\sigma^2 \\
&\overset{\text{(I)}}{=} (\widehat{a}_n - a_0)\int_{z}\log\left(\frac{1}{z\widehat{b}_n}\right)\frac{\widehat{b}_n^{\widehat{a}_n+1}}{\Gamma(\widehat{a}_n)}\left(z\widehat{b}_n\right)^{-\widehat{a}_n-1}\exp\left\{-\frac{1}{z}\right\}dz \\
&= (\widehat{a}_n - a_0)\frac{1}{\Gamma(\widehat{a}_n)}\int_{z}\Big(-\log(z)-\log(\widehat{b}_n)\Big)\,z^{-\widehat{a}_n-1}\exp\left\{-\frac{1}{z}\right\}dz \\
&\overset{I(\mathcal{IG})}{=} (\widehat{a}_n - a_0)\left[\frac{1}{\Gamma(\widehat{a}_n)}\int_{z}(-\log(z))\,z^{-\widehat{a}_n-1}\exp\left\{-\frac{1}{z}\right\}dz - \log(\widehat{b}_n)\right] \\
&\overset{\text{(II)}}{=} (\widehat{a}_n - a_0)\left[\frac{1}{\Gamma(\widehat{a}_n)}\int_{z}\frac{d}{d\widehat{a}_n}\left\{z^{-\widehat{a}_n-1}\exp\left\{-\frac{1}{z}\right\}\right\}dz - \log(\widehat{b}_n)\right] \\
&\overset{\text{(III)}}{=} (\widehat{a}_n - a_0)\left[\frac{1}{\Gamma(\widehat{a}_n)}\frac{d}{d\widehat{a}_n}\underbrace{\left\{\int_{z}z^{-\widehat{a}_n-1}\exp\left\{-\frac{1}{z}\right\}dz\right\}}_{\overset{I(\mathcal{IG})}{=}\Gamma(\widehat{a}_n)} - \log(\widehat{b}_n)\right] \\
&= (\widehat{a}_n - a_0)\left(\frac{\Gamma'(\widehat{a}_n)}{\Gamma(\widehat{a}_n)} - \log(\widehat{b}_n)\right) \\
&= (\widehat{a}_n - a_0)\Big(\Psi(\widehat{a}_n) - \log(\widehat{b}_n)\Big),
\end{aligned}
$$

where Ψ is the digamma function. For $C_6(\sigma^2)$, one obtains the closed form as

$$
\begin{aligned}
C_{12} &= \int_{\sigma^2} C_6(\sigma^2)\mathcal{IG}^{\mathrm{VB}}(\sigma^2|\widehat{a}_n,\widehat{b}_n)d\sigma^2 \\
&= C_5\int_{\sigma^2}\frac{\widehat{b}_n^{\widehat{a}_n}}{\Gamma(\widehat{a}_n)}(\sigma^2)^{-\widehat{a}_n-1-1}\exp\left\{-\frac{\widehat{b}_n}{\sigma^2}\right\}d\sigma^2 \\
&\overset{I(\mathcal{IG})}{=} C_5\frac{\Gamma(\widehat{a}_n+1)}{\widehat{b}_n\Gamma(\widehat{a}_n)}.
\end{aligned}
$$

Using the exact same steps for $C_{10}(\sigma^2)$, one finds

$$
\begin{aligned}
C_{13} &= \int_{\sigma^2} C_{10}(\sigma^2)\mathcal{IG}^{\mathrm{VB}}(\sigma^2|\widehat{a}_n,\widehat{b}_n)d\sigma^2 \\
&\overset{I(\mathcal{IG})}{=} C_9\frac{\Gamma(\widehat{a}_n+1)}{\widehat{b}_n\Gamma(\widehat{a}_n)},
\end{aligned}
$$

finally yielding

$$
\begin{aligned}
Q_1 &= C_1 - C_8 + C_{11} - C_{12} - C_{13} \\
&= \log\left(\frac{\widehat{b}_n^{\widehat{a}_n}\Gamma(a_0)}{b_0^{a_0}\Gamma(\widehat{a}_n)}\right) + 0.5\log\left|\boldsymbol{\Sigma}_0\widehat{\boldsymbol{\Sigma}}_n^{-1}\right| - \frac{1}{2}\mathrm{tr}\left(I - \boldsymbol{\Sigma}_0^{-1}\widehat{\boldsymbol{\Sigma}}_n\right) + (\widehat{a}_n - a_0)\left(\Psi(\widehat{a}_n) - \log(\widehat{b}_n)\right) \\
&\quad - \left[\frac{1}{2}\widehat{\boldsymbol{\mu}}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n - \frac{1}{2}\boldsymbol{\mu}_0'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + (\widehat{b}_n - b_0)\right]\cdot\frac{\Gamma(\widehat{a}_n+1)}{\widehat{b}_n\Gamma(\widehat{a}_n)} \\
&\quad - \left[\frac{1}{2}\widehat{\boldsymbol{\mu}}_n'(\widehat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}_0^{-1})\widehat{\boldsymbol{\mu}}_n - \widehat{\boldsymbol{\mu}}_n'\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)\right]\cdot\frac{\Gamma(\widehat{a}_n+1)}{\widehat{b}_n\Gamma(\widehat{a}_n)} \\
&= \log\left(\frac{\widehat{b}_n^{\widehat{a}_n}\Gamma(a_0)}{b_0^{a_0}\Gamma(\widehat{a}_n)}\right) + 0.5\log\left|\boldsymbol{\Sigma}_0\widehat{\boldsymbol{\Sigma}}_n^{-1}\right| - \frac{1}{2}\mathrm{tr}\left(I - \boldsymbol{\Sigma}_0^{-1}\widehat{\boldsymbol{\Sigma}}_n\right) + (\widehat{a}_n - a_0)\left(\Psi(\widehat{a}_n) - \log(\widehat{b}_n)\right) \\
&\quad + \frac{1}{2}\left[(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n) + 2(b_0 - \widehat{b}_n)\right]\cdot\frac{\Gamma(\widehat{a}_n+1)}{\widehat{b}_n\Gamma(\widehat{a}_n)}
\end{aligned}
$$

B.8.4 Q_2

Noting that one can write Q_2 as

$$
\begin{aligned}
&= E_{\mathrm{VB}}\left[\sum_{i=1}^{n}\ell^{\beta}(\boldsymbol{\mu},\sigma^2|x_i)\right] \\
&= \int_{\boldsymbol{\mu},\sigma^2}\left\{\sum_{i=1}^{n}\left[\frac{1}{\beta}f(x_i|\boldsymbol{\mu},\sigma^2)^{\beta} - \frac{1}{1+\beta}\int_{\mathcal{Y}}f(x|\boldsymbol{\mu},\sigma^2)^{1+\beta}dx\right]\times\right. \\
&\qquad \left.\mathrm{NIG}^{\mathrm{VB}}(\boldsymbol{\mu},\sigma^2|\widehat{a}_n,\widehat{b}_n,\widehat{\boldsymbol{\mu}}_n,\widehat{\boldsymbol{\Sigma}}_n)\right\}d\sigma^2 d\boldsymbol{\mu} \\
&= \sum_{i=1}^{n}\left[\int_{\boldsymbol{\mu},\sigma^2}\left\{\frac{1}{\beta}f(x_i|\boldsymbol{\mu},\sigma^2)^{\beta} - \frac{1}{1+\beta}\int_{\mathcal{Y}}f(x|\boldsymbol{\mu},\sigma^2)^{1+\beta}dx\right\}\right. \\
&\qquad \left.\mathrm{NIG}^{\mathrm{VB}}(\boldsymbol{\mu},\sigma^2|\widehat{a}_n,\widehat{b}_n,\widehat{\boldsymbol{\mu}}_n,\widehat{\boldsymbol{\Sigma}}_n)d\sigma^2 d\boldsymbol{\mu}\right]. \qquad \text{(B.10)}
\end{aligned}
$$

The last equation implies that it is sufficient to concern ourselves with the integral for a single term. To this end, observe that the likelihood for a single observation x_i with regressor matrix $\boldsymbol{X}_i$ is given by

$$
f(x_i|\boldsymbol{\mu},\sigma^2) = \mathcal{N}(x_i|\boldsymbol{X}_i'\boldsymbol{\mu},\sigma^2 I_d), \qquad \text{(B.11)}
$$

where I_d is the identity matrix of dimension d . Looking at the likelihood terms inside ℓ^β, the β-exponentiated likelihood term can be rewritten as

$$\begin{aligned}
\frac{1}{\beta} f(x_i|\boldsymbol{\mu},\sigma^2)^\beta &= \underbrace{\frac{1}{\beta}(2\pi)^{-0.5d\beta}(\sigma^2)^{-0.5d\beta}}_{=D_1(\sigma^2)} \cdot \exp\left\{-\frac{\beta}{2\sigma^2}\left[(x_i - \boldsymbol{X}_i'\boldsymbol{\mu})'(x_i - \boldsymbol{X}_i'\boldsymbol{\mu})\right]\right\} \\
&= D_1(\sigma^2)\cdot\exp\left\{-\frac{\beta}{2\sigma^2}\left[x_i'x_i + \right.\right. \\
&\quad \left.\left. \boldsymbol{\mu}'\underbrace{(\boldsymbol{X}_i\boldsymbol{X}_i')}_{=\ddot{\boldsymbol{\Sigma}}_i^{-1}}\boldsymbol{\mu} - 2(x_i'\boldsymbol{X}_i)\boldsymbol{\mu}\right]\right\} \\
&\overset{\blacksquare}{=} D_1(\sigma^2)\cdot\exp\left\{-\tfrac{1}{2\sigma^2}\left[\beta(\boldsymbol{\mu}-\underbrace{\ddot{\boldsymbol{\Sigma}}_i(\boldsymbol{X}_i'x_i)}_{=\ddot{\boldsymbol{\mu}}_i})'\ddot{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{\mu}-\ddot{\boldsymbol{\mu}}_i) + \underbrace{\beta\left[x_i'x_i - (x_i\boldsymbol{X}_i')\ddot{\boldsymbol{\Sigma}}_i(\boldsymbol{X}_i x_i')\right]}_{=D_{2,i}}\right]\right\} \\
&= D_1(\sigma^2)\cdot\exp\left\{-\frac{1}{2\sigma^2}\left[\underbrace{\beta(\boldsymbol{\mu}-\ddot{\boldsymbol{\mu}}_i)'\ddot{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{\mu}-\ddot{\boldsymbol{\mu}}_i)}_{=D_{3,i}(\boldsymbol{\mu})} + D_{2,i}\right]\right\} \\
&= D_1(\sigma^2)\cdot\exp\left\{-\frac{1}{2\sigma^2}\left[D_{3,i}(\boldsymbol{\mu}) + D_{2,i}\right]\right\}, \qquad \text{(B.12)}
\end{aligned}$$

while the integral is available in closed form as

$$\frac{1}{1+\beta}\int_{\mathcal{Y}} f(x|\boldsymbol{\mu},\sigma^2)^{1+\beta}dx \overset{I(\mathcal{N})}{=} (\sigma^2)^{-0.5p\beta}\underbrace{(2\pi)^{-0.5d\beta}(1+\beta)^{-0.5d-1}}_{=D_4} \qquad \text{(B.13)}$$

One can see a neat separation between terms involving σ^2 and terms involving $\boldsymbol{\mu}$ again, allowing us to rewrite the integral in equation (B.10) such as to exploit the conditional structure of the normal inverse-gamma distribution in Eqs. (B.13), (B.12). Looking at integrating out σ^2 from (B.12) first, note that

$$\begin{aligned}
L_1 &= \int_{\sigma^2}\left\{\frac{1}{1+\beta}\int_{\mathcal{Y}} f(x|\boldsymbol{\mu},\sigma^2)^{1+\beta}dx\right\}\mathcal{IG}^{\text{VB}}(\sigma^2|\widehat{a}_n,\widehat{b}_n)d\sigma^2 \\
&= D_4\int_{\sigma^2}(\sigma^2)^{-0.5d\beta-\widehat{a}_n-1}\frac{\widehat{b}_n^{\widehat{a}_n}}{\Gamma(\widehat{a}_n)}\exp\left\{-\frac{\widehat{b}_n}{\sigma^2}\right\}d\sigma^2 \\
&\overset{I(\mathcal{N})}{=} D_4\cdot\frac{\Gamma(\widehat{a}_n+0.5d\beta)}{\Gamma(\widehat{a}_n)\widehat{b}_n^{0.5d\beta}}. \qquad \text{(B.14)}
\end{aligned}$$

For the β-exponentiated likelihood term, one finds that

$$\begin{aligned}
L_{2,i} &= \int_{\sigma^2,\boldsymbol{\mu}} \frac{1}{\beta} f(x_i|\boldsymbol{\mu},\sigma^2)^{\beta} \mathrm{NIG}^{\mathrm{VB}}(\boldsymbol{\mu},\sigma^2|\widehat{a}_n,\widehat{b}_n,\widehat{\boldsymbol{\mu}}_n,\widehat{\boldsymbol{\Sigma}}_n) d\sigma^2 d\boldsymbol{\mu} \\
&= \int_{\sigma^2} D_1(\sigma^2) \cdot \exp\left\{-\frac{1}{2\sigma^2} D_{2,i}\right\} \underbrace{\left[\int_{\boldsymbol{\mu}} \exp\left\{-\frac{1}{2\sigma^2} D_{3,i}(\boldsymbol{\mu})\right\} \mathcal{N}^{\mathrm{VB}}(\boldsymbol{\mu}|\widehat{\boldsymbol{\mu}}_n,\sigma^2\widehat{\boldsymbol{\Sigma}}_n) d\boldsymbol{\mu}\right]}_{=D_{5,i}(\sigma^2)} \times \\
&\qquad \mathcal{IG}^{\mathrm{VB}}(\sigma^2|\widehat{a}_n,\widehat{b}_n) d\sigma^2,
\end{aligned}$$

where we have again exploited the conditional structure of our assumed posterior. The inner integral equals

$$D_{5,i}(\sigma^2) = (2\pi)^{-0.5p} \left|\sigma^2\widehat{\boldsymbol{\Sigma}}_n\right|^{-0.5} \underbrace{\int_{\boldsymbol{\mu}} \exp\left\{-\frac{1}{2\sigma^2} \underbrace{\left[D_{3,i}(\boldsymbol{\mu}) + (\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n)'\widehat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n)\right]}_{=D_{6,i}(\boldsymbol{\mu})}\right\}}_{=D_{7,i}(\sigma^2)},$$

indicating that the closed form for the integral is available if one rewrites it as a normal density. To this end, one can use completion of squares to rewrite

$$\begin{aligned}
D_{6,i}(\boldsymbol{\mu}) &= \beta(\boldsymbol{\mu}-\ddot{\boldsymbol{\mu}}_i)'\ddot{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{\mu}-\ddot{\boldsymbol{\mu}}_i) + (\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n)'\widehat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_n) \\
&= \boldsymbol{\mu}' \underbrace{\left[\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta\ddot{\boldsymbol{\Sigma}}_i^{-1}\right]}_{=\widetilde{\boldsymbol{\Sigma}}_i^{-1}} \boldsymbol{\mu} - 2\left[\widehat{b}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta\ddot{\boldsymbol{\mu}}_i'\ddot{\boldsymbol{\Sigma}}_i^{-1}\right]\boldsymbol{\mu} + \left[\widehat{\boldsymbol{\mu}}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n + \beta\ddot{\boldsymbol{\mu}}_i'\ddot{\boldsymbol{\Sigma}}_i^{-1}\ddot{\boldsymbol{\mu}}_i\right] \\
&\overset{\blacksquare}{=} \left(\boldsymbol{\mu} - \underbrace{\widetilde{\boldsymbol{\Sigma}}_i\left[\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{b}_n + \beta\ddot{\boldsymbol{\Sigma}}_i^{-1}\ddot{\boldsymbol{\mu}}_i\right]}_{=\widetilde{\boldsymbol{\mu}}_i}\right)' \widetilde{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}_i) + \\
&\qquad \underbrace{\widehat{\boldsymbol{\mu}}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n + \beta\ddot{\boldsymbol{\mu}}_i'\ddot{\boldsymbol{\Sigma}}_i^{-1}\ddot{\boldsymbol{\mu}}_i - \left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n + \beta\ddot{\boldsymbol{\Sigma}}_i^{-1}\ddot{\boldsymbol{\mu}}_i\right)' \widetilde{\boldsymbol{\Sigma}}_i \left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n + \beta\ddot{\boldsymbol{\Sigma}}_i^{-1}\ddot{\boldsymbol{\mu}}_i\right)}_{=D_{8,i}} \\
&= (\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}_i)'\widetilde{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}_i) + D_{8,i},
\end{aligned}$$

which then allows integrating out $\boldsymbol{\mu}$ from $D_{7,i}(\sigma^2)$ using the density of a normal random variable:

$$\begin{aligned}
D_{7,i}(\sigma^2) &= \exp\left\{-\frac{1}{2\sigma^2} D_{8,i}\right\} \int_{\boldsymbol{\mu}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}_i)'\widetilde{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{\mu}-\widetilde{\boldsymbol{\mu}}_i)\right\} d\boldsymbol{\mu} \\
&\overset{I(\mathcal{N})}{=} \exp\left\{-\frac{1}{2\sigma^2} D_{8,i}\right\} (2\pi)^{0.5p} |\sigma^2\widetilde{\boldsymbol{\Sigma}}_i|^{0.5},
\end{aligned}$$

so we can finally rewrite the entire integral as

$$D_{5,i}(\sigma^2) = |\widehat{\boldsymbol{\Sigma}}_n^{-1}\widetilde{\boldsymbol{\Sigma}}_i|^{0.5}\exp\left\{-\frac{1}{2\sigma^2}D_{8,i}\right\},$$

which enables rewriting $L_{2,i}$ as

$$\begin{aligned}
L_{2,i} &= \underbrace{\frac{1}{\beta}(2\pi)^{-0.5d\beta}|\widehat{\boldsymbol{\Sigma}}_n^{-1}\widetilde{\boldsymbol{\Sigma}}_i|^{0.5}}_{=D_{9,i}}\int_{\sigma^2}(\sigma^2)^{-0.5d\beta}\exp\left\{-\frac{1}{\sigma^2}\cdot\frac{1}{2}\left[D_{2,i}+D_{8,i}\right]\right\}\mathcal{IG}^{\mathrm{VB}}(\sigma^2|\widehat{a}_n,\widehat{b}_n)d\sigma^2 \\
&\overset{I(\mathcal{IG})}{=} \frac{D_{9,i}\cdot\Gamma(\widehat{a}_n+0.5d\beta)\cdot\widehat{b}_n^{\widehat{a}_n}}{\Gamma(\widehat{a}_n)\cdot\left[\widehat{b}_n+0.5(D_{2,i}+D_{7,i})\right]^{(\widehat{a}_n+0.5d\beta)}},
\end{aligned}$$

finally implying that one may write

$$\begin{aligned}
Q_2 &= \sum_{i=1}^{n}L_{2,i} - nL_1 \\
&= \sum_{i=1}^{n}\left\{\frac{D_{9,i}\cdot\Gamma(\widehat{a}_n+0.5d\beta)\cdot\widehat{b}_n^{\widehat{a}_n}}{\Gamma(\widehat{a}_n)\cdot\left[\widehat{b}_n+0.5(D_{2,i}+D_{8,i})\right]^{(\widehat{a}_n+0.5d\beta)}}\right\} - nD_4\cdot\frac{\Gamma(\widehat{a}_n+0.5d\beta)}{\Gamma(\widehat{a}_n)\widehat{b}_n^{0.5d\beta}} \\
&= \sum_{i=1}^{n}\left\{\frac{\frac{1}{\beta}(2\pi)^{-0.5d\beta}\left|\widehat{\boldsymbol{\Sigma}}_n^{-1}\left[\widehat{\boldsymbol{\Sigma}}_n^{-1}+\beta(\boldsymbol{X}_i\boldsymbol{X}_i)\right]^{-1}\right|^{0.5}\cdot\Gamma(\widehat{a}_n+0.5d\beta)\cdot\widehat{b}_n^{\widehat{a}_n}}{\Gamma(\widehat{a}_n)\cdot\left[\widehat{b}_n+0.5\left(D_{2,i}+D_{8,i}\right)\right]^{(\widehat{a}_n+0.5d\beta)}}\right\} \\
&\qquad -n\cdot\frac{(2\pi)^{-0.5d\beta}(1+\beta)^{-0.5d-1}\cdot\Gamma(\widehat{a}_n+0.5d\beta)}{\Gamma(\widehat{a}_n)\widehat{b}_n^{0.5d\beta}}.
\end{aligned}$$

We further simplify this expression by observing that

$$\begin{aligned}
&D_{2,i}+D_{8,i} \\
&= \beta\left[x_i'x_i-(x_i\boldsymbol{X}_i')\ddot{\boldsymbol{\Sigma}}_i(\boldsymbol{X}_ix_i')\right]+\widehat{\boldsymbol{\mu}}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n+\beta\ddot{\boldsymbol{\mu}}_i'\ddot{\boldsymbol{\Sigma}}_i^{-1}\ddot{\boldsymbol{\mu}}_i \\
&\qquad -\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n+\beta\ddot{\boldsymbol{\Sigma}}_i^{-1}\ddot{\boldsymbol{\mu}}_i\right)'\widetilde{\boldsymbol{\Sigma}}_i\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n+\beta\ddot{\boldsymbol{\Sigma}}_i^{-1}\ddot{\boldsymbol{\mu}}_i\right) \\
&= \beta x_i'x_i-\beta(x_i\boldsymbol{X}_i')\ddot{\boldsymbol{\Sigma}}_i(\boldsymbol{X}_ix_i')+\widehat{\boldsymbol{\mu}}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n+\beta(x_i\boldsymbol{X}_i')\ddot{\boldsymbol{\Sigma}}_i(\boldsymbol{X}_ix_i') \\
&\qquad -\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n+\beta(\boldsymbol{X}_i'x_i)\right)'\widetilde{\boldsymbol{\Sigma}}_i\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n+\beta(\boldsymbol{X}_i'x_i)\right) \\
&= \beta x_i'x_i+\widehat{\boldsymbol{\mu}}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n-\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n+\beta(\boldsymbol{X}_i'x_i)\right)'\left[\widehat{\boldsymbol{\Sigma}}_n^{-1}+\beta(\boldsymbol{X}_i\boldsymbol{X}_i)\right]^{-1}\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n+\beta(\boldsymbol{X}_i'x_i)\right),
\end{aligned}$$

leaving us with

$$Q_2 = \frac{\Gamma(\widehat{a}_n + 0.5d\beta) \cdot \widehat{b}_n^{\widehat{a}_n} \cdot |\widehat{\boldsymbol{\Sigma}}_n^{-1}|^{0.5}}{\beta(2\pi)^{0.5d\beta}\Gamma(\widehat{a}_n)} \times$$

$$\sum_{i=1}^{n} \left\{ \frac{\left|\left[\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta(\boldsymbol{X}_i\boldsymbol{X}_i)\right]^{-1}\right|^{0.5}}{\left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + \widehat{\boldsymbol{\mu}}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n - \left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n + \beta(\boldsymbol{X}_i' x_i)\right)'\left[\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta(\boldsymbol{X}_i\boldsymbol{X}_i)\right]^{-1}\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n + \beta(\boldsymbol{X}_i' x_i)\right)\right)\right]^{(\widehat{a}_n + 0.5d\beta)}} \right\}$$

$$-n \cdot \frac{\Gamma(\widehat{a}_n + 0.5d\beta)}{\Gamma(\widehat{a}_n)\widehat{b}_n^{0.5d\beta}(2\pi)^{0.5d\beta}(1+\beta)^{0.5d+1}}.$$

B.8.5 Objective

Putting together the results of the two previous sections, the O_{GVI} is obtained as

$$\begin{aligned} O_{\text{GVI}} &= -Q_1 + Q_2 \\ &= -\log\left(\frac{\widehat{b}_n^{\widehat{a}_n}\Gamma(a_0)}{b_0^{a_0}\Gamma(\widehat{a}_n)}\right) - 0.5\log\left|\boldsymbol{\Sigma}_0\widehat{\boldsymbol{\Sigma}}_n^{-1}\right| + \frac{1}{2}\mathrm{tr}\left(I - \boldsymbol{\Sigma}_0^{-1}\widehat{\boldsymbol{\Sigma}}_n\right) - (\widehat{a}_n - a_0)\left(\Psi(\widehat{a}_n) - \log(\widehat{b}_n)\right) \\ &\quad - \left[\frac{1}{2}(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n) + (b_0 - \widehat{b}_n)\right] \cdot \frac{\Gamma(\widehat{a}_n + 1)}{\widehat{b}_n\Gamma(\widehat{a}_n)} \\ &\quad + \frac{\Gamma(\widehat{a}_n + 0.5d\beta) \cdot \widehat{b}_n^{\widehat{a}_n} \cdot |\widehat{\boldsymbol{\Sigma}}_n^{-1}|^{0.5}}{\beta(2\pi)^{0.5d\beta}\Gamma(\widehat{a}_n)} \times \end{aligned}$$

$$\sum_{i=1}^{n} \left\{ \frac{\left|\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta(\boldsymbol{X}_i'\boldsymbol{X}_i)\right|^{-0.5}}{\left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + \widehat{\boldsymbol{\mu}}_n'\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n - \left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n + \beta(\boldsymbol{X}_i' x_i)\right)'\left[\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta(\boldsymbol{X}_i'\boldsymbol{X}_i)\right]^{-1}\left(\widehat{\boldsymbol{\Sigma}}_n^{-1}\widehat{\boldsymbol{\mu}}_n + \beta(\boldsymbol{X}_i' x_i)\right)\right)\right]^{(\widehat{a}_n + 0.5d\beta)}} \right\}$$

$$-n \cdot \frac{\Gamma(\widehat{a}_n + 0.5d\beta)}{\Gamma(\widehat{a}_n)\widehat{b}_n^{0.5d\beta}(2\pi)^{0.5d\beta}(1+\beta)^{0.5d+1}}$$

B.8.6 Differentiation

In this section, we take derivatives of O_{GVI} with respect to each variational parameter, i.e. $\widehat{a}_n, \widehat{b}_n, \widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Sigma}}_n$. Observing that differentiation with respect to $\widehat{\boldsymbol{\Sigma}}_n^{-1}$ is easier than with respect to $\widehat{\boldsymbol{\Sigma}}_n$, parametrize the optimization using the Cholesky decomposition, i.e. $\widehat{\boldsymbol{\Sigma}}_n^{-1} = \mathcal{L}\mathcal{L}'$, where L is a lower triangular matrix and is unique if $\widehat{\boldsymbol{\Sigma}}_n$ (equivalently $\widehat{\boldsymbol{\Sigma}}_n^{-1}$) is positive definite[2].

[2]Note that $\mathcal{L}$ need not be unique if $\widehat{\boldsymbol{\Sigma}}_n$ is positive semi-definite, but this is of no concern for us here: Since we implicitly impose that $\widehat{\boldsymbol{\Sigma}}_n$ is non-singular (so that $\widehat{\boldsymbol{\Sigma}}_n^{-1}$ is unique and well-defined), all covariance matrices $\widehat{\boldsymbol{\Sigma}}_n$ considered have to be positive definite.

Derivative with respect to L

In what follows, we differentiate the O_{GVI} term by term with respect to the $p(p-1)\frac{1}{2}$ entries in the lower triangular part of $\mathcal{L}$ that can be summarized in the vector $\text{vech}(\mathcal{L})$. To this end, define

$$E_1 = -0.5 \log \left| \boldsymbol{\Sigma}_0 \widehat{\boldsymbol{\Sigma}}_n^{-1} \right| + \frac{1}{2} \text{tr} \left(I - \boldsymbol{\Sigma}_0^{-1} \widehat{\boldsymbol{\Sigma}}_n \right) \tag{B.15}$$

$$E_2 = \underbrace{\frac{\Gamma(\widehat{a}_n + 0.5d\beta) \cdot \widehat{b}_n^{\widehat{a}_n}}{\beta(2\pi)^{0.5d\beta}\Gamma(\widehat{a}_n)}}_{=F} |\widehat{\boldsymbol{\Sigma}}_n^{-1}|^{0.5} \tag{B.16}$$

$$E_{3,i} = \left| \widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta \left(\boldsymbol{X}_i' \boldsymbol{X}_i \right) \right|^{-0.5} \tag{B.17}$$

$$E_4 = \widehat{\boldsymbol{\mu}}_n' \widehat{\boldsymbol{\Sigma}}_n^{-1} \widehat{\boldsymbol{\mu}}_n \tag{B.18}$$

$$E_{5,i} = -\widehat{\boldsymbol{\mu}}_n' \widehat{\boldsymbol{\Sigma}}_n^{-1} \left[\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta \left(\boldsymbol{X}_i' \boldsymbol{X}_i \right) \right]^{-1} \widehat{\boldsymbol{\Sigma}}_n^{-1} \widehat{\boldsymbol{\mu}}_n \tag{B.19}$$

$$E_{6,i} = -\beta^2 (x_i' \boldsymbol{X}_i) \left[\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta \left(\boldsymbol{X}_i' \boldsymbol{X}_i \right) \right]^{-1} (\boldsymbol{X}_i' x_i), \tag{B.20}$$

$$E_{7,i} = -2\beta \widehat{\boldsymbol{\mu}}_n' \widehat{\boldsymbol{\Sigma}}_n^{-1} \left[\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta \left(\boldsymbol{X}_i' \boldsymbol{X}_i \right) \right]^{-1} (\boldsymbol{X}_i' x_i). \tag{B.21}$$

Obtaining the derivative of the O_{GVI} is equivalent to obtaining the derivatives of these newly defined quantities, as

$$\begin{aligned}
&\frac{\partial}{\partial \text{vech}(\boldsymbol{L})} \{O_{\text{GVI}}\} \\
&= \frac{\partial}{\partial \text{vech}(\boldsymbol{L})} \{E_1\} + \\
&\frac{\partial}{\partial \text{vech}(\boldsymbol{L})} \{E_2\} \cdot \sum_{i=1}^{n} \left\{ \frac{E_{3,i}}{\left[\widehat{b}_n + 0.5 \left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i} \right) \right]^{\widehat{a}_n + 0.5d\beta}} \right\} \\
&+ E_2 \cdot \sum_{i=1}^{n} \left\{ \frac{\frac{\partial}{\partial \text{vech}(\boldsymbol{L})} \{E_{3,i}\}}{\left[\widehat{b}_n + 0.5 \left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i} \right) \right]^{\widehat{a}_n + 0.5d\beta}} \right\} \\
&+ E_2 \cdot \sum_{i=1}^{n} \left\{ E_{3,i} \cdot \frac{\partial}{\partial \text{vech}(\boldsymbol{L})} \left\{ \left[\widehat{b}_n + 0.5 \left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i} \right) \right]^{-\widehat{a}_n - 0.5d\beta} \right\} \right\},
\end{aligned}$$

where the chain and sum rule imply that

$$\begin{aligned}
&\frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}\left\{\left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)\right]^{-\widehat{a}_n - 0.5d\beta}\right\} \\
&= \left(-\widehat{a}_n - 0.5d\beta\right) \cdot \left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)\right]^{-\widehat{a}_n - 0.5d\beta - 1} \times \\
&\qquad 0.5 \cdot \frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}\left\{E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right\},
\end{aligned}$$

For convenience and simplified notation when taking the derivatives of the expressions defined in (B.15)–(B.21), also define the following matrices:

$$\begin{aligned}
\mathbb{R} &= \left[\widehat{\boldsymbol{\Sigma}}_n^{-1} + \beta\left(\boldsymbol{X}_i'\boldsymbol{X}_i\right)\right] \\
\boldsymbol{\Theta} &= \widehat{\boldsymbol{\mu}}_n \widehat{\boldsymbol{\mu}}_n'.
\end{aligned}$$

Define also the following symbols to mark operations used in the derivations:

∂ Switching from differential notation $\partial\mathcal{L}$ to the derivative $\frac{\partial}{\partial \text{vech}(\boldsymbol{L})}$;

tr Properties of the trace like invariance under cyclic permutations, invariance under the transpose, additivity, and the fact that for c a scalar, $\text{tr}(c) = c$.

Note than when the differential operator ∂ is used, its scope is always limited to the next term only, unless brackets are used. Hence $\partial\mathcal{L}\mathcal{L}'$ uses the differential only with respect to $\mathcal{L}$, while $\partial\left(\mathcal{L}\mathcal{L}'\right)^{-1}$ uses it with respect to the entire expression $\left(\mathcal{L}\mathcal{L}'\right)^{-1}$. It is also worth noting that $\partial\mathcal{L}' = \left(\partial\mathcal{L}\right)'$ for any matrix $\mathcal{L}$, as this will be used in conjunction with the transpose invariance of the trace throughout to simplify terms. Using these symbols and the differential notation, proceed by noting the following:

$$\begin{aligned}
\partial(\mathcal{L}\mathcal{L}') = \partial\mathbb{R} &= \partial\mathcal{L}\mathcal{L}' + \mathcal{L}\partial\mathcal{L}' = \partial\mathcal{L}\mathcal{L}' + \mathcal{L}\partial\mathcal{L}' = \partial\mathcal{L}\mathcal{L}' + \left(\partial\mathcal{L}\mathcal{L}'\right)' \\
\partial(\mathcal{L}\mathcal{L}')^{-1} &= -(\mathcal{L}\mathcal{L}')^{-1}\left[\partial(\mathcal{L}\mathcal{L}')\right](\mathcal{L}\mathcal{L}')^{-1} \\
\partial|\mathcal{L}\mathcal{L}'| &= |\mathcal{L}\mathcal{L}'| \cdot \text{tr}\left((\mathcal{L}\mathcal{L}')^{-1}\left[\partial\mathcal{L}\mathcal{L}' + \left(\partial\mathcal{L}\mathcal{L}'\right)'\right]\right) \\
&\overset{\text{tr}}{=} 2|\mathcal{L}\mathcal{L}'| \cdot \text{tr}\left(\mathcal{L}'(\mathcal{L}\mathcal{L}')^{-1}\partial\mathcal{L}\right) \\
\partial\mathbb{R}^{-1} &= -\mathbb{R}^{-1}\partial\mathbb{R}\mathbb{R}^{-1} = -\mathbb{R}^{-1}\partial\mathcal{L}\mathcal{L}'\mathbb{R}^{-1} - \left[\mathbb{R}^{-1}\partial\mathcal{L}\mathcal{L}'\mathbb{R}^{-1}\right]'.
\end{aligned}$$

With this in place, the derivatives of the quantities defined before are obtained as

$$\begin{aligned}
\partial E_1 &= -\frac{1}{2}\partial\left\{\log|\boldsymbol{\Sigma}_0| + \log|\mathcal{L}\mathcal{L}'|\right\} - \frac{1}{2}\partial\left\{\text{tr}\left(\boldsymbol{\Sigma}_0^{-1}\widehat{\boldsymbol{\Sigma}}_n\right)\right\} \\
&= -\frac{1}{2} \cdot |\mathcal{L}\mathcal{L}'|^{-1} \cdot \partial|\mathcal{L}\mathcal{L}'| - \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_0^{-1}\partial\left(\mathcal{L}\mathcal{L}'\right)^{-1}\right) \\
&= -\frac{1}{2}\text{tr}\left(\mathcal{L}'(\mathcal{L}\mathcal{L}')^{-1}\partial\mathcal{L}\right) + \frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\mathcal{L}\mathcal{L}')^{-1}\left[\partial(\mathcal{L}\mathcal{L}')\right](\mathcal{L}\mathcal{L}')^{-1}\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{tr}}{=} -\text{tr}\left(\mathcal{L}'(\mathcal{L}\mathcal{L}')^{-1}\partial\mathcal{L}\right) + \text{tr}\left(\mathcal{L}'\left(\mathcal{L}\mathcal{L}'\right)^{-1}\boldsymbol{\Sigma}_0^{-1}\left(\mathcal{L}\mathcal{L}'\right)^{-1}\partial\mathcal{L}\right)\\
\partial E_2 &= F\cdot\partial|\mathcal{L}\mathcal{L}'|^{0.5}\\
&= \frac{F}{2}\cdot|\mathcal{L}\mathcal{L}'|^{-0.5}\cdot 2|\mathcal{L}\mathcal{L}'|\cdot\text{tr}\left(\mathcal{L}'(\mathcal{L}\mathcal{L}')^{-1}\partial\mathcal{L}\right)\\
&= F\cdot|\mathcal{L}\mathcal{L}'|^{0.5}\text{tr}\left(\mathcal{L}'(\mathcal{L}\mathcal{L}')^{-1}\partial\mathcal{L}\right)\\
\partial E_{3,i} &= \partial\mathbb{R}^{-0.5} = -\frac{1}{2}|\mathbb{R}|^{-1.5}\partial\mathbb{R}\\
&= -\frac{1}{2}|\mathbb{R}|^{-0.5}\text{tr}\left(\mathbb{R}^{-1}\partial\left(\mathcal{L}\mathcal{L}'\right)\right)\\
&\overset{\text{tr}}{=} -|\mathbb{R}|^{-0.5}\text{tr}\left(\mathcal{L}'\mathbb{R}^{-1}\partial\mathcal{L}\right)\\
\partial E_4 &\overset{\text{tr}}{=} \text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\partial(\mathcal{L}\mathcal{L}')\widehat{\boldsymbol{\mu}}_n\right)\\
&= \text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\left[\partial\mathcal{L}\mathcal{L}' + \mathcal{L}\partial\mathcal{L}'\right]\widehat{\boldsymbol{\mu}}_n\right)\\
&\overset{\text{tr}}{=} 2\cdot\text{tr}\left(\mathcal{L}'\boldsymbol{\Theta}\partial\mathcal{L}\right)\\
\partial E_{5,i} &\overset{\text{tr}}{=} -\text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\partial\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\widehat{\boldsymbol{\mu}}_n\right) - \text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\left(\mathcal{L}\mathcal{L}'\right)\partial\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\widehat{\boldsymbol{\mu}}_n\right)\\
&\quad -\text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\partial\left(\mathcal{L}\mathcal{L}'\right)\widehat{\boldsymbol{\mu}}_n\right)\\
&\overset{\text{tr}}{=} -2\cdot\text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\partial\mathcal{L}\mathcal{L}'\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\widehat{\boldsymbol{\mu}}_n\right) + 2\cdot\text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\partial\mathcal{L}\mathcal{L}'\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\widehat{\boldsymbol{\mu}}_n\right)\\
&\quad -2\cdot\text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\partial\mathcal{L}\mathcal{L}'\widehat{\boldsymbol{\mu}}_n\right)\\
&\overset{\text{tr}}{=} -2\cdot\text{tr}\left(\mathcal{L}'\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\boldsymbol{\Theta}\partial\mathcal{L}\right) + 2\cdot\text{tr}\left(\mathcal{L}'\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\boldsymbol{\Theta}\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\partial\mathcal{L}\right)\\
&\quad -2\cdot\text{tr}\left(\mathcal{L}'\boldsymbol{\Theta}\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\partial\mathcal{L}\right)\\
\partial E_{6,i} &\overset{\text{tr}}{=} -\beta^2\text{tr}\left(\left(x_i'\boldsymbol{X}_i\right)\partial\mathbb{R}^{-1}\left(\boldsymbol{X}_i'x_i\right)\right)\\
&\overset{\text{tr}}{=} 2\beta^2\text{tr}\left(\left(x_i'\boldsymbol{X}_i\right)\mathbb{R}^{-1}\partial\mathcal{L}\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_i'x_i\right)\right)\\
&\overset{\text{tr}}{=} 2\beta^2\text{tr}\left(\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_i'x_i\right)\left(x_i'\boldsymbol{X}_i\right)\mathbb{R}^{-1}\partial\mathcal{L}\right)\\
\partial E_{7,i} &\overset{\text{tr}}{=} -2\beta\cdot\left[\text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\partial\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\left(\boldsymbol{X}_ix_i\right)\right) + \text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\left(\mathcal{L}\mathcal{L}'\right)\partial\mathbb{R}^{-1}\left(\boldsymbol{X}_ix_i\right)\right)\right]\\
&\overset{\text{tr}}{=} -2\beta\cdot\Bigg[\text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\partial\mathcal{L}\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_ix_i\right)\right) + \text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\mathcal{L}\partial\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_ix_i\right)\right)\\
&\quad -\text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\partial\mathcal{L}\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_ix_i\right)\right) - \text{tr}\left(\widehat{\boldsymbol{\mu}}_n'\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\mathcal{L}\partial\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_ix_i\right)\right)\Bigg]\\
&\overset{\text{tr}}{=} -2\beta\cdot\Bigg[\text{tr}\left(\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_ix_i\right)\widehat{\boldsymbol{\mu}}_n'\partial\mathcal{L}\right) + \text{tr}\left(\mathcal{L}'\widehat{\boldsymbol{\mu}}_n\left(x_i'\boldsymbol{X}_i\right)\mathbb{R}^{-1}\partial\mathcal{L}\right)\\
&\quad -\text{tr}\left(\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_ix_i\right)\widehat{\boldsymbol{\mu}}_n'\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\partial\mathcal{L}\right) - \text{tr}\left(\mathcal{L}'\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\widehat{\boldsymbol{\mu}}_n\left(x_i'\boldsymbol{X}_i\right)\mathbb{R}^{-1}\partial\mathcal{L}\right)\Bigg]
\end{aligned}
$$

This can now be converted into derivative notation and simplified. To this end, first note that for any $p\times\partial$ matrix A which is not a function of $\mathcal{L}$,

$$
\text{tr}(Ad\mathcal{L}) = \sum_{i=1}^{p}A_{1i}dL_{i1} + \sum_{i=2}^{p}A_{2i}dL_{i2} + \cdots = \sum_{j=1}^{p}\left\{\sum_{i=j}^{p}A_{ji}dL_{ji}\right\},
$$

implying in particular that

$$\frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}\text{tr}(Ad\mathcal{L}) = \text{vech}(A^T)$$

and use this by defining $\text{vech}^T(A) = \text{vech}(A^T)$ to note that

$$\begin{aligned}
\frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}E_1 &\overset{\partial}{=} \text{vech}^T\left(-\left[\mathcal{L}'(\mathcal{L}\mathcal{L}')^{-1}\right] + \left[\mathcal{L}'\left(\mathcal{L}\mathcal{L}'\right)^{-1}\boldsymbol{\Sigma}_0^{-1}\left(\mathcal{L}\mathcal{L}'\right)^{-1}\right]\right) \\
&= \text{vech}^T\left(\mathcal{L}'\left(\mathcal{L}\mathcal{L}'\right)^{-1}\left[\boldsymbol{\Sigma}_0^{-1}\left(\mathcal{L}\mathcal{L}'\right)^{-1} - I_p\right]\right) \\
&= \text{vech}^T\left(\mathcal{L}^{-1}\left[\boldsymbol{\Sigma}_0^{-1}\left(\mathcal{L}\mathcal{L}'\right)^{-1} - I_p\right]\right) \\
&= \text{vech}\left(\left[\left(\mathcal{L}\mathcal{L}'\right)^{-1}\boldsymbol{\Sigma}_0^{-1} - I_p\right]\mathcal{L}^{-T}\right) \\
\frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}E_2 &\overset{\partial}{=} F\cdot\left|\mathcal{L}\mathcal{L}'\right|^{0.5}\cdot\text{vech}^T\left(\mathcal{L}'(\mathcal{L}\mathcal{L}')^{-1}\right) \\
&= F\cdot\left|\mathcal{L}\mathcal{L}'\right|^{0.5}\cdot\text{vech}\left(\mathcal{L}^{-T}\right) \\
\frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}E_{3,i} &\overset{\partial}{=} -|\mathbb{R}|^{-0.5}\cdot\text{vech}\left(\mathbb{R}^{-1}\mathcal{L}\right) \\
\frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}E_4 &\overset{\partial}{=} 2\cdot\text{vech}\left(\boldsymbol{\Theta}\mathcal{L}\right) \\
\frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}E_{5,i} &\overset{\partial}{=} \text{vech}^T\left(-2\mathcal{L}'\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\boldsymbol{\Theta} + 2\mathcal{L}'\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\boldsymbol{\Theta}\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1} - 2\mathcal{L}'\boldsymbol{\Theta}\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\right) \\
&= 2\cdot\text{vech}^T\left(\left[\mathcal{L}'\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\boldsymbol{\Theta}\left[\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1} - I_p\right]\right] - \left[\mathcal{L}'\boldsymbol{\Theta}\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\right]\right) \\
&= 2\cdot\text{vech}^T\left(\mathcal{L}'\left[\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\boldsymbol{\Theta}\left[\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1} - I_p\right] - \boldsymbol{\Theta}\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\right]\right) \\
&= 2\cdot\text{vech}\left(\left[\left[\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right) - I_p\right]\boldsymbol{\Theta}\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1} - \mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\boldsymbol{\Theta}\right]\mathcal{L}\right) \\
\frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}E_{6,i} &\overset{\partial}{=} 2\beta^2\cdot\text{vech}\left(\mathbb{R}^{-1}\left(\boldsymbol{X}_i'x_i\right)\left(x_i'\boldsymbol{X}_i\right)\mathbb{R}^{-1}\mathcal{L}\right) \\
\frac{\partial}{\partial \text{vech}\left(\boldsymbol{L}\right)}E_{7,i} &\overset{\partial}{=} -2\beta\cdot\text{vech}^T\Bigg(\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_i x_i\right)\widehat{\boldsymbol{\mu}}_n' + \mathcal{L}'\widehat{\boldsymbol{\mu}}_n\left(x_i'\boldsymbol{X}_i\right)\mathbb{R}^{-1} \\
&\quad -\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_i x_i\right)\widehat{\boldsymbol{\mu}}_n'\left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1} - \mathcal{L}'\mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\widehat{\boldsymbol{\mu}}_n\left(x_i'\boldsymbol{X}_i\right)\mathbb{R}^{-1}\Bigg) \\
&= -2\beta\cdot\text{vech}^T\Bigg(\mathcal{L}'\mathbb{R}^{-1}\left(\boldsymbol{X}_i'x_i\right)\widehat{\boldsymbol{\mu}}_n'\left[I_p - \left(\mathcal{L}\mathcal{L}'\right)\mathbb{R}^{-1}\right] \\
&\qquad + \left[I_p - \mathcal{L}'\mathbb{R}^{-1}\mathcal{L}\right]\mathcal{L}'\widehat{\boldsymbol{\mu}}_n\left(x_i'\boldsymbol{X}_i\right)\mathbb{R}^{-1}\Bigg) \\
&= -2\beta\cdot\text{vech}\Bigg(\left[I_p - \mathbb{R}^{-1}\left(\mathcal{L}\mathcal{L}'\right)\right]\widehat{\boldsymbol{\mu}}_n\left(x_i'\boldsymbol{X}_i\right)\mathbb{R}^{-1}\mathcal{L} \\
&\qquad +\mathbb{R}^{-1}\left(\boldsymbol{X}_i'x_i\right)\widehat{\boldsymbol{\mu}}_n'\mathcal{L}\left[I_p - \mathcal{L}'\mathbb{R}^{-1}\mathcal{L}\right]\Bigg)
\end{aligned}$$

Derivative with respect to $\widehat{\boldsymbol{\mu}}_n$

Differentiating with respect to $\widehat{\boldsymbol{\mu}}_n$ is trivial. One proceeds by the same logic as in the section before, to which end one additionally needs to define the new term

$$E_8 = -\frac{1}{2}\left[(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n) + 2(b_0 - \widehat{b}_n)\right] \cdot \frac{\Gamma(\widehat{a}_n + 1)}{\widehat{b}_n \Gamma(\widehat{a}_n)},$$

allowing us to write

$$\begin{aligned}\frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}}\{O_{\text{GVI}}\} &= \frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}}\{E_8\} + \\ &E_2 \cdot \sum_{i=1}^{n}\left\{ E_{3,i} \cdot \frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}}\left\{\left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)\right]^{-\widehat{a}_n - 0.5d\beta}\right\}\right\},\end{aligned}$$

where

$$\begin{aligned}&\frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}}\left\{\left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)\right]^{-\widehat{a}_n - 0.5d\beta}\right\} \\ &= (-\widehat{a}_n - 0.5d\beta) \cdot \left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)\right]^{-\widehat{a}_n - 0.5d\beta - 1} \times \\ &\qquad 0.5 \cdot \frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}}\left\{E_4 + E_{5,i} + E_{7,i}\right\},\end{aligned}$$

so that obtaining the derivative is achieved by finding $\frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}} E_4$, $\frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}} E_{5,i}$, $\frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}} E_{7,i}$ and $\frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}} E_8$:

$$\begin{aligned}\frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}} E_4 &= 2 \cdot \widehat{\boldsymbol{\mu}}_n' \widehat{\boldsymbol{\Sigma}}_n^{-1} \\ \frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}} E_{5,i} &= -2 \cdot \widehat{\boldsymbol{\mu}}_n' \widehat{\boldsymbol{\Sigma}}_n^{-1} \mathbb{R}^{-1} \widehat{\boldsymbol{\Sigma}}_n^{-1} \\ \frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}} E_{7,i} &= -2\beta \cdot \left(x_i' \boldsymbol{X}_i\right) \mathbb{R}^{-1} \widehat{\boldsymbol{\Sigma}}_n^{-1} \\ \frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}} E_8 &= -\frac{1}{2} \cdot \frac{\Gamma(\widehat{a}_n + 1)}{\widehat{b}_n \Gamma(\widehat{a}_n)} \left[\frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}}\left(\widehat{\boldsymbol{\mu}}_n' \boldsymbol{\Sigma}_0^{-1} \widehat{\boldsymbol{\mu}}_n\right) - 2\frac{\partial}{\partial \widehat{\boldsymbol{\mu}_n}}\left(\widehat{\boldsymbol{\mu}}_n \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right)\right] \\ &= -\frac{1}{2} \cdot \frac{\Gamma(\widehat{a}_n + 1)}{\widehat{b}_n \Gamma(\widehat{a}_n)} \left[2\widehat{\boldsymbol{\mu}}_n' \boldsymbol{\Sigma}_0^{-1} - 2\boldsymbol{\mu}_0' \boldsymbol{\Sigma}_0^{-1}\right] \\ &= \frac{\Gamma(\widehat{a}_n + 1)}{\widehat{b}_n \Gamma(\widehat{a}_n)} \left[(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n)' \boldsymbol{\Sigma}_0^{-1}\right]\end{aligned}$$

Derivative with respect to $\widehat{a}_n$

We proceed again by the same logic. Define

$$\begin{aligned}
E_9 &= -\log\left(\frac{\widehat{b}_n^{\widehat{a}_n}\Gamma(a_0)}{b_0^{a_0}\Gamma(\widehat{a}_n)}\right) \\
E_{10} &= -(\widehat{a}_n - a_0)\left(\Psi(\widehat{a}_n) - \log(\widehat{b}_n)\right) \\
E_{11} &= -n \cdot \frac{\Gamma(\widehat{a}_n + 0.5d\beta)}{\Gamma(\widehat{a}_n)\widehat{b}_n^{0.5d\beta}(2\pi)^{0.5d\beta}(1+\beta)^{0.5d+1}}.
\end{aligned}$$

Use this to write

$$\begin{aligned}
\frac{\partial}{\partial \widehat{a}_n}\{O_{\text{GVI}}\} = \frac{\partial}{\partial \widehat{a}_n}\{E_8\} + \frac{\partial}{\partial \widehat{a}_n}\{E_9\} + \frac{\partial}{\partial \widehat{a}_n}\{E_{10}\} + \frac{\partial}{\partial \widehat{a}_n}\{E_{11}\} + \\
+\frac{\partial}{\partial \widehat{a}_n}\{E_2\}\sum_{i=1}^{n}\left\{\frac{E_{3,i}}{\left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)\right]^{\widehat{a}_n + 0.5d\beta}}\right\} \\
+E_2 \cdot \sum_{i=1}^{n}\left\{E_{3,i} \cdot \frac{\partial}{\partial \widehat{a}_n}\left\{\left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)\right]^{-\widehat{a}_n - 0.5d\beta}\right\}\right\},
\end{aligned}$$

where for $\widehat{a}_n$, the inner term equals

$$\frac{\partial}{\partial \widehat{a}_n}\left\{\left[\underbrace{\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)}_{=K}\right]^{-\widehat{a}_n - 0.5d\beta}\right\} = -\log(K) \cdot K^{-\widehat{a}_n - 0.5d\beta},$$

so that the differentiation with respect to $\widehat{a}_n$ requires obtaining the following terms:

$$\begin{aligned}
\frac{\partial}{\partial \widehat{a}_n} E_2 &= \frac{|\widehat{\boldsymbol{\Sigma}}_n^{-1}|^{0.5}}{\beta(2\pi)^{0.5d\beta}} \left[\frac{\frac{\partial}{\partial \widehat{a}_n}\left\{\Gamma(\widehat{a}_n + 0.5d\beta)\right\} \widehat{b}_n^{\widehat{a}_n}}{\Gamma(\widehat{a}_n)} + \frac{\frac{\partial}{\partial \widehat{a}_n}\left\{\widehat{b}_n^{\widehat{a}_n}\right\} \Gamma\left(\widehat{a}_n + 0.5d\beta\right)}{\Gamma(\widehat{a}_n)} \right. \\
&\quad \left. + \frac{\partial}{\partial \widehat{a}_n}\left\{\Gamma(\widehat{a}_n)^{-1}\right\} \cdot \widehat{b}_n^{\widehat{a}_n} \Gamma(\widehat{a}_n + 0.5d\beta) \right] \\
&= \frac{|\widehat{\boldsymbol{\Sigma}}_n^{-1}|^{0.5} \widehat{b}_n^{\widehat{a}_n} \Gamma(\widehat{a}_n + 0.5d\beta)}{\beta(2\pi)^{0.5d\beta}\Gamma(\widehat{a}_n)} \left[\Psi(\widehat{a}_n + 0.5d\beta) + \log(\widehat{b}_n) - \Psi(\widehat{a}_n) \right] \\
\frac{\partial}{\partial \widehat{a}_n} E_8 &= -\frac{1}{2}\left[(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n)' \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n) + 2(b_0 - \widehat{b}_n) \right] \times \\
&\quad \left[\frac{\frac{\partial}{\partial \widehat{a}_n}\left\{\Gamma(\widehat{a}_n + 1)\right\}}{\widehat{b}_n \Gamma(\widehat{a}_n)} - \frac{\frac{\partial}{\partial \widehat{a}_n}\left\{\Gamma(\widehat{a}_n)\right\} \Gamma(\widehat{a}_n + 1)}{\Gamma(\widehat{a}_n)^2 \widehat{b}_n} \right] \\
&= -\frac{1}{2}\left[(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n)' \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n) + 2(b_0 - \widehat{b}_n) \right] \times \\
&\quad \frac{\Gamma(\widehat{a}_n + 1)}{\widehat{b}_n \Gamma(\widehat{a}_n)} \cdot \left[\Psi(\widehat{a}_n + 1) - \Psi(\widehat{a}_n) \right] \\
\frac{\partial}{\partial \widehat{a}_n} E_9 &= -\frac{\partial}{\partial \widehat{a}_n}\left\{ \widehat{a}_n \log(\widehat{b}_n) \right\} + \frac{\partial}{\partial \widehat{a}_n}\left\{ \log\left(\Gamma(\widehat{a}_n)\right) \right\} \\
&= -\log(\widehat{b}_n) + \Psi(\widehat{a}_n) \\
\frac{\partial}{\partial \widehat{a}_n} E_{10} &= \frac{\partial}{\partial \widehat{a}_n}\left\{ \widehat{a}_n \log(\widehat{b}_n) \right\} - \frac{\partial}{\partial \widehat{a}_n}\left\{ (\widehat{a}_n - a_0)\, \Psi(\widehat{a}_n) \right\} \\
&= \log(\widehat{b}_n) - \Psi(\widehat{a}_n) - (\widehat{a}_n - a_0)\, \Psi^{(1)}(\widehat{a}_n) \\
\frac{\partial}{\partial \widehat{a}_n} E_{11} &= -\frac{n}{\widehat{b}_n^{0.5d\beta}(2\pi)^{0.5d\beta}(1+\beta)^{0.5d+1}} \cdot \frac{\partial}{\partial \widehat{a}_n}\left\{ \frac{\Gamma(\widehat{a}_n + 0.5d\beta)}{\Gamma(\widehat{a}_n)} \right\} \\
&= -\frac{n}{\widehat{b}_n^{0.5d\beta}(2\pi)^{0.5d\beta}(1+\beta)^{0.5d+1}} \cdot \frac{\Gamma(\widehat{a}_n + 0.5d\beta)}{\Gamma(\widehat{a}_n)} \cdot \left[\Psi(\widehat{a}_n + 0.5d\beta) - \Psi(\widehat{a}_n) \right],
\end{aligned}$$

where $\Psi^{(1)}$ denotes the trigamma function.

Derivative with respect to $\widehat{b}_n$

As for the other variational parameters, note that

$$\begin{aligned}
\frac{\partial}{\partial \widehat{b}_n}\left\{O_{\text{GVI}}\right\} &= \frac{\partial}{\partial \widehat{b}_n}\left\{E_8\right\} + \frac{\partial}{\partial \widehat{b}_n}\left\{E_9\right\} + \frac{\partial}{\partial \widehat{b}_n}\left\{E_{10}\right\} + \frac{\partial}{\partial \widehat{b}_n}\left\{E_{11}\right\} + \\
&\quad + \frac{\partial}{\partial \widehat{b}_n}\left\{E_2\right\} \sum_{i=1}^{n} \left\{ \frac{E_{3,i}}{\left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)\right]^{\widehat{a}_n + 0.5d\beta}} \right\} \\
&\quad + E_2 \cdot \sum_{i=1}^{n} \left\{ E_{3,i} \cdot \frac{\partial}{\partial \widehat{b}_n} \left\{ \left[\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)\right]^{-\widehat{a}_n - 0.5d\beta} \right\} \right\},
\end{aligned}$$

where the chain rule implies that

$$\frac{\partial}{\partial \widehat{b}_n}\left\{\left[\underbrace{\widehat{b}_n + 0.5\left(\beta x_i' x_i + E_4 + E_{5,i} + E_{6,i} + E_{7,i}\right)}_{=K}\right]^{-\widehat{a}_n - 0.5d\beta}\right\} = (-\widehat{a}_n - 0.5d\beta)\cdot K^{-\widehat{a}_n - 0.5d\beta - 1}.$$

Thus one proceeds by the same logic as before.

$$\begin{aligned}
\frac{\partial}{\partial \widehat{b}_n} E_2 &= \frac{\widehat{a}_n \Gamma(\widehat{a}_n + 0.5d\beta)\cdot |\widehat{\boldsymbol{\Sigma}}_n^{-1}|^{0.5}}{\beta(2\pi)^{0.5d\beta}\Gamma(\widehat{a}_n)}\cdot \widehat{b}_n^{\widehat{a}_n - 1} \\
\frac{\partial}{\partial \widehat{b}_n} E_8 &= \frac{1}{2}\left[(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n)'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \widehat{\boldsymbol{\mu}}_n) + 2b_0\right]\frac{\Gamma(\widehat{a}_n + 1)}{\Gamma(\widehat{a}_n)}\cdot\frac{1}{\widehat{b}_n^2} \\
\frac{\partial}{\partial \widehat{b}_n} E_9 &= -\frac{\widehat{a}_n}{\widehat{b}_n} \\
\frac{\partial}{\partial \widehat{b}_n} E_{10} &= \frac{\widehat{a}_n - a_0}{\widehat{b}_n} \\
\frac{\partial}{\partial \widehat{b}_n} E_{11} &= \frac{nd\beta\cdot\Gamma(\widehat{a}_n + 0.5d\beta)}{2\cdot\Gamma(\widehat{a}_n)(2\pi)^{0.5d\beta}(1+\beta)^{0.5d+1}}\cdot \widehat{b}_n^{-0.5d\beta - 1}
\end{aligned}$$

B.8.7 Complexity Analysis of Inference

Time complexity: Our SVRG method crucially hinges on the complexity of the gradient evaluations. For BLR, we note that evaluating the complete O_{GVI}-gradient derived above for n observations has complexity $\mathcal{O}(np^3)$, where p is the number of regressors. We proceed by defining g as the (generic) complexity of a gradient evaluation, so for BLR $g = p^3$. Clearly, an SGD step using b observations is of order $\mathcal{O}(bg)$. Similarly, the computation of the anchors is $\mathcal{O}(Bg)$. Next, let the optimization routine used for full optimization have complexity $\mathcal{O}(m(n, \dim(\boldsymbol{\theta})))$. Most standard (quasi-) Newton optimization routines such as BFGS or LBFGSB (used in our implementation) are polynomial in n and $\dim(\boldsymbol{\theta})$. For such methods, since it holds that at most $W \geq n$ observations are evaluated in the full optimization, and since $\dim(\boldsymbol{\theta})$ is time-constant, $m(n, \dim(\boldsymbol{\theta}))$ is also constant in time. Thus, though these constants can be substantial, all optimization steps (whether SVRG steps or full optimization steps) are $\mathcal{O}(1)$ in time. Since one performs T of them for T observations, the computational complexity (in time) is $\mathcal{O}(T)$.

Space complexity: One needs to store observations $\boldsymbol{y}_t$ as well as gradient evaluations. Storing one of them takes $\mathcal{O}(d)$ and $\mathcal{O}(\dim(\boldsymbol{\theta}))$ space, respectively. Since we only keep a window W of the most recent observations (and gradients), this means that the space requirement is of order $\mathcal{O}(W(d+\dim(\boldsymbol{\theta})))$ and in particular constant in time.

B.8.8 Recursive On-line Optimization of $\beta_{\mathbf{rlm}}$

Recall that

$$\widehat{\boldsymbol{y}}_t(\boldsymbol{\beta}) = \sum_{r_t, m_t} \mathbb{E}\left(\boldsymbol{y}_t | \boldsymbol{y}_{1:(t-1)}, r_{t-1}, m_{t-1}, \beta_{\mathrm{m}}\right) p(r_{t-1}, m_{t-1} | \boldsymbol{y}_{1:(t-1)}, \beta_{\mathrm{rlm}}).$$

the issue reduces to finding the partial derivatives $\nabla_{\beta_{\mathrm{rlm}}} \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$ and $\nabla_{\beta_{\mathrm{m}}} \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$. Notice that for $\nabla_{\beta_{\mathrm{rlm}}} \widehat{\boldsymbol{y}}_t(\boldsymbol{\beta})$, one finds that

$$\nabla_{\beta_{\mathrm{rlm}}} \widehat{x}_t(\boldsymbol{\beta}) = \sum_{r_t, m_t} \mathbb{E}\left(x_t | x_{1:(t-1)}, r_{t-1}, m_{t-1}, \beta_{\mathrm{m}}\right) \nabla_{\beta_{\mathrm{rlm}}} p(r_{t-1}, m_{t-1} | x_{1:(t-1)}, \beta_{\mathrm{rlm}}).$$

Observe now that for $p(x_{1:t}) = \sum_{r_t, m_t} p(r_t, m_t, x_{1:t} | \beta_{\mathrm{rlm}})$,

$$\begin{aligned}
&\nabla_{\beta_{\mathrm{rlm}}} p(r_t, m_t | x_{1:t}, \beta_{\mathrm{rlm}}) \\
=&\nabla_{\beta_{\mathrm{rlm}}} \left\{ \frac{p(r_t, m_t, x_{1:t} | \beta_{\mathrm{rlm}})}{\sum_{r_t, m_t} p(r_t, m_t, x_{1:t} | \beta_{\mathrm{rlm}})} \right\} \\
=&\frac{\nabla_{\beta_{\mathrm{rlm}}} p(r_t, m_t, x_{1:t} | \beta_{\mathrm{rlm}})}{p(x_{1:t})} - \frac{p(r_t, m_t, \boldsymbol{y}_{1:t} | \beta_{\mathrm{rlm}})}{p(x_{1:t})^2} \cdot \sum_{r_t, m_t} \nabla_{\beta_{\mathrm{rlm}}} p(r_t, m_t, x_{1:t} | \beta_{\mathrm{rlm}}).
\end{aligned}$$

Thus we have reduced the problem to finding $\nabla_{\beta_{\mathrm{rlm}}} p(r_t, m_t, x_{1:t} | \beta_{\mathrm{rlm}})$. Defining for a predictive posterior distribution $f_{m_t}(x_t | x_{1:(t-1)}, r_{t-1})$ its β-divergence analogue as

$$\begin{aligned}
&f_{m_t}^{\beta_{\mathrm{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1}) \\
&= \exp\left\{ \frac{1}{\beta_{\mathrm{rlm}} - 1} f_{m_t}(x_t | x_{1:(t-1)}, r_{t-1})^{\beta_{\mathrm{rlm}} - 1} - \frac{1}{\beta_{\mathrm{rlm}}} \int_{\mathcal{Y}} f_{m_t}(x_t | x_{1:(t-1)}, r_{t-1})^{\beta_{\mathrm{rlm}}} dx_t \right\}
\end{aligned}$$

and uppressing the conditioning on β_{rlm} for convenience, one can using the recursion

$$\begin{aligned}
p(x_{1:t}, r_t, m_t) = \sum_{m_{t-1}, r_{t-1}} \Big\{ & f_{m_t}^{\beta_{\mathrm{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1}) q(m_t | x_{1:(t-1)}, r_{t-1}, m_{t-1}) \\
& H(r_t, r_{t-1}) p(x_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\},
\end{aligned}$$

compute $\nabla_{\beta_{\mathrm{rlm}}} p(r_t, m_t, x_{1:t})$ from $\nabla_{\beta_{\mathrm{rlm}}} p(r_{t-1}, m_{t-1}, x_{1:(t-1)} | \beta_{\mathrm{rlm}})$ for $r_t = r_{t-1} + 1$ as

$$\begin{aligned}
&\nabla_{\beta_{\mathrm{rlm}}} p(x_{1:t}, r_t, m_t) \\
=& \Big\{ \nabla_{\beta_{\mathrm{rlm}}} f_{m_t}^{\beta_{\mathrm{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1}) q(m_t | x_{1:(t-1)}, r_{t-1}, m_{t-1}) H(r_t, r_{t-1}) p(x_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\} + \\
& \Big\{ f_{m_t}^{\beta_{\mathrm{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1}) \nabla_{\beta_{\mathrm{rlm}}} q(m_t | x_{1:(t-1)}, r_{t-1}, m_{t-1}) H(r_t, r_{t-1}) p(x_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\} + \\
& \Big\{ f_{m_t}^{\beta_{\mathrm{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1}) q(m_t | x_{1:(t-1)}, r_{t-1}, m_{t-1}) H(r_t, r_{t-1}) \nabla_{\beta_{\mathrm{rlm}}} p(x_{1:(t-1)}, r_{t-1}, m_{t-1}) \Big\}.
\end{aligned}$$

Similarly, for $r_t = 0$ the expression becomes

$$\begin{aligned}
&\nabla_{\beta_{\text{rlm}}} p(x_{1:t}, r_t, m_t) \\
&= \nabla_{\beta_{\text{rlm}}} f_{m_t}^{\beta_{\text{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1}) \cdot q(m_t) \sum_{r_{t-1}, m_{t-1}} H(0, r_{t-1}) p(x_{1:(t-1)}, r_{t-1}, m_{t-1}) + \\
&\quad f_{m_t}^{\beta_{\text{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1}) \cdot q(m_t) \sum_{r_{t-1}, m_{t-1}} H(0, r_{t-1}) \nabla_{\beta_{\text{rlm}}} p(x_{1:(t-1)}, r_{t-1}, m_{t-1}).
\end{aligned}$$

This implies that if $f_{m_t}^{\beta_{\text{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1})$ and $q(m_t | x_{1:(t-1)}, r_{t-1}, m_{t-1})$ are differentiable with respect to β_{rlm}, then the entire expression can be updated recursively. For most exponential family likelihoods (and in particular the normal likelihood of the Bayesian Linear Regression), $\nabla_{\beta_{\text{rlm}}} f_{m_t}^{\beta_{\text{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1})$ is available analytically. In particular, as long as $\int_{\mathcal{Y}} f_{m_t}(x_t | x_{1:(t-1)}, r_{t-1})^{1+\beta_{\text{rlm}}} dx_t$ has a closed form, $\nabla_{\beta_{\text{rlm}}} f_{m_t}^{\beta_{\text{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1})$ can be found in analytic form. In the case of Bayesian Linear Regression where the d-dimensional posterior predictive takes the shape of a student-t distribution with ν degrees of freedom and posterior covariance $\frac{\nu}{\nu-2}\Sigma$, one finds that

$$\begin{aligned}
\nabla_{\beta_{\text{rlm}}} f_{m_t}^{\beta_{\text{rlm}}}(x_t | x_{1:(t-1)}, r_{t-1}) = {} & \nabla_{\beta_{\text{rlm}}} g_1(\beta_{\text{rlm}}) g_2(\beta_{\text{rlm}}) g_3(\beta_{\text{rlm}}) + \\
& g_1(\beta_{\text{rlm}}) \nabla_{\beta_{\text{rlm}}} g_2(\beta_{\text{rlm}}) g_3(\beta_{\text{rlm}}) + \\
& g_1(\beta_{\text{rlm}}) g_2(\beta_{\text{rlm}}) \nabla_{\beta_{\text{rlm}}} g_3(\beta_{\text{rlm}}),
\end{aligned}$$

where for $\eta = \nu d + d\beta_{\text{rlm}} + \nu$,

$$\begin{aligned}
g_1(\beta_{\text{rlm}}) &= \left(\frac{\Gamma(0.5[\nu + d])}{\Gamma(0.5\nu)} \right)^{\beta_{\text{rlm}}} \\
g_2(\beta_{\text{rlm}}) &= \frac{\Gamma(0.5\eta)}{\Gamma(0.5[\eta + p])} \\
g_3(\beta_{\text{rlm}}) &= (\nu\pi)^{-0.5p \cdot (\beta_{\text{rlm}} - 1)} \cdot |\Sigma|^{-(\beta_{\text{rlm}} - 1)},
\end{aligned}$$

so that their derivatives are given by

$$\begin{aligned}
\nabla_{\beta_{\text{rlm}}} g_1(\beta_{\text{rlm}}) &= -\beta_{\text{rlm}} \cdot \log(g_1(\beta_{\text{rlm}})) \cdot g_2(\beta_{\text{rlm}}) \\
\nabla_{\beta_{\text{rlm}}} g_1(\beta_{\text{rlm}}) &= 0.5(\nu + p) \left[\cdot \frac{\Gamma(0.5\eta)\Psi(0.5\eta)}{\Gamma([0.5[\eta + p])} - \frac{\Gamma(0.5[\eta])\Psi(0.5[p + \eta])}{\Gamma([0.5[\eta + p])} \right] \\
\nabla_{\beta_{\text{rlm}}} g_3(\beta_{\text{rlm}}) &= -g_3(\beta_{\text{rlm}}) \cdot \log(g_3(\beta_{\text{rlm}})) \cdot \frac{1}{\beta_{\text{rlm}} - 1}
\end{aligned}$$

As for $\nabla_{\beta_{\text{rlm}}} q(m_t|x_{1:(t-1)}, r_{t-1}, m_{t-1})$, one can again obtain it recursively, since for $r_t > 0$,

$$\begin{aligned}
&\nabla_{\beta_{\text{rlm}}} q(m_t|x_{1:(t-1)}, r_{t-1}, m_{t-1}) \\
&= \nabla_{\beta_{\text{rlm}}} \left\{ \frac{p(x_{1:(t-1)}, r_{t-1}, m_{t-1})}{\sum_{m_{t-1}} p(x_{1:(t-1)}, m_{t-1})} \right\} \\
&= \frac{\nabla_{\beta_{\text{rlm}}} p(x_{1:(t-1)}, r_{t-1}, m_{t-1})}{\sum_{m_{t-1}} p(x_{1:(t-1)}, r_{t-1}, m_{t-1})} - \frac{\sum_{m_{t-1}} \nabla_{\beta_{\text{rlm}}} p(x_{1:(t-1)}, r_{t-1}, m_{t-1})}{\left(\sum_{m_{t-1}} p(x_{1:(t-1)}, r_{t-1}, m_{t-1}) \right)^2}.
\end{aligned}$$

Appendix C

Proofs

C.1 Duality

We will be invoking general convex analysis on the space $\mathcal{F}_b(\mathbf{\Theta})$, noting that $\mathcal{F}_b(\mathbf{\Theta})$ is a Hausdorff locally convex space (through the uniform norm). Recall that $\mathcal{B}(\mathbf{\Theta})$ denotes the set of all bounded and finitely additive signed measures over Θ (with a given σ-algebra). For any set $D \subseteq \mathcal{B}(\mathbf{\Theta})$ and $h \in \mathcal{F}_b(\mathbf{\Theta})$, we use $\sigma_D(h) = \sup_{\nu \in D} \langle h, \nu \rangle$ and $\iota_D(\nu) = \infty \cdot [\![\nu \notin D]\!]$ to denote the *support* and *indicator* functions as in Rockafellar (1970).

Before we begin with the proofs, we introduce the conjugate specific to these spaces.

Definition C.1 (Rockafellar (1968))**.** For any proper convex function $F : \mathcal{B}(\mathbf{\Theta}) \to (-\infty, \infty)$, we have for any $h \in \mathcal{F}_b(\mathbf{\Theta})$ we define

$$F^\star(h) = \sup_{\mu \in \mathcal{B}(\mathbf{\Theta})} \{ \int_{\mathbf{\Theta}} h d\mu - F(\mu) \}$$

and for any $\mu \in \mathcal{B}(\mathbf{\Theta})$ we define

$$F^{\star\star}(\mu) = \sup_{h \in \mathcal{F}_b(\mathbf{\Theta})} \{ \int_{\mathbf{\Theta}} h d\mu - F^\star(h) \}.$$

Further, we recall a convenient reflexivity property that was shown to hold on these spaces and of which we will make use in the sequel.

Theorem C.1 (Zalinescu (2002) Theorem 2.3.3)**.** If X is a Hausdorff locally convex space, and $F : X \to (-\infty, \infty]$ is a proper convex lower semi-continuous function then $F^{\star\star} = F$.

Equipped with this, we can now prove the required technical Lemmas.

Lemma 2.2.1. For any $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$ and $L \in \mathcal{F}_b(\boldsymbol{\Theta})$, we have

$$\mathsf{E}_{\overline{\mathrm{co}}(\Pi)}[L] = \mathsf{E}_{\Pi}[L].$$

Proof. For any $n \in \mathbb{N}$, we denote $\Delta_n = \{\alpha \in [0,1]^n : \sum_{i=1}^n \alpha_i = 1\}$. We then have

$$\begin{aligned}
\mathsf{E}_{\overline{\mathrm{co}}(\Pi)}[L] &= \inf_{q \in \overline{\mathrm{co}}(\Pi)} \mathbb{E}_q[L] \\
&= \inf_{n \in \mathbb{N}: \alpha \in \Delta_n, q_i \in \Pi, \forall i = 1, \ldots, n} \mathbb{E}_{\sum_{i=1}^n \alpha_i q_i}[L] \\
&\overset{(1)}{=} \inf_{n \in \mathbb{N}: \alpha \in \Delta_n, q_i \in \Pi, \forall i = 1, \ldots, n} \sum_{i=1}^n \alpha_i \mathbb{E}_{q_i}[L] \\
&= \inf_{n \in \mathbb{N}: \alpha \in \Delta_n} \sum_{i=1}^n \alpha_i \inf_{q_i \in \Pi} \mathbb{E}_{q_i}[L] \\
&= \inf_{n \in \mathbb{N}: \alpha \in \Delta_n} \sum_{i=1}^n \alpha_i \mathbb{E}_{\Pi}[L] \\
&= \mathsf{E}_{\Pi}[L],
\end{aligned}$$

where (1) holds due to linearity of expectation. □

We will employ Theorem C.1 to derive the duality result and present in the form of an early Lemma for consistency in notation.

Lemma 2.2.2. For any prior $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, we have

$$D(q\|\pi) = \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \{\mathbb{E}_{q(\boldsymbol{\theta})}[\rho(\boldsymbol{\theta})] - D_\pi^\star(\rho)\}.$$

Proof. Using Theorem C.1, we have $D = D^{\star\star}$ since D is proper convex and lower-semicontinuous by assumption. From this, we now obtain the desired result simply by applying Definition C.1. □

We now require one last technical result.

Lemma 2.2.3. For any prior $\pi \in \mathcal{P}(\boldsymbol{\Theta})$, regularizer D and set $\Pi \subseteq \mathcal{P}(\boldsymbol{\Theta})$, define a function $F : \mathcal{P}(\boldsymbol{\Theta}) \times \mathcal{F}_b(\boldsymbol{\Theta}) \to \mathbb{R}$ as

$$F(q, \rho) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[L(\boldsymbol{\theta})\right] + \mathbb{E}_{q(\boldsymbol{\theta})}\left[\rho(\boldsymbol{\theta})\right] - D_\pi^\star(\rho) + \iota_{\overline{\mathrm{co}}(\Pi)}(q).$$

It holds that

$$\inf_{q \in \mathcal{P}(\boldsymbol{\Theta})} \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} F(q, \rho) = \sup_{\rho \in \mathcal{F}_b(\boldsymbol{\Theta})} \inf_{q \in \mathcal{P}(\boldsymbol{\Theta})} F(q, \rho).$$

Proof. First note that since $L, \rho \in \mathcal{F}_b(\boldsymbol{\Theta})$ and $\overline{\text{co}}\,(\Pi)$ is closed and convex (by construction), it holds that the mapping $q \mapsto F(q, \rho)$ is convex and lower-semicontinuous (Penot, 2012). Furthermore note that $D^\star_\pi(\rho)$ is convex and lower-semicontinuous for any choice of D. Now, since $q \in \mathcal{P}(\boldsymbol{\Theta}) \subset \mathcal{B}(\boldsymbol{\Theta})$, it follows that the mapping $\rho \mapsto F(q, \rho)$ is also convex and lower-semicontinuous. Next, by endowing $\mathcal{B}(\boldsymbol{\Theta})$ with the topology associated with the Banach-Alaoglu Theorem, we can use strong duality between $\mathcal{F}_b(\boldsymbol{\Theta})$ and $\mathcal{P}(\boldsymbol{\Theta})$, so that it follows that $\mathcal{P}(\boldsymbol{\Theta})$ is compact (Liu and Chaudhuri, 2018, Lemma 27(b)). Finally, noting that all conditions for Ky Fan's minimax Theorem are satisfied (Fan, 1953, Theorem 2), the result follows. □

C.2 Proof of Theorem 7.2

Proof. Conditioned on the event $\{r_t = r\}$, either $r_{t+1} = r+1$ or $r_{t+1} = 0$. The odds of these two possibilities are as in the quantity of interest in Theorem 7.2.

Now substituting the definitions of $f^{\beta_{\text{rlm}}}_{m_t}(x_{t+1}|x_{1:(t-1)}, r_{t-1})$ and $f^{\beta_{\text{rlm}}}_{m_t}(x_{t+1}|x_0)$ leaves

$$
\begin{aligned}
&\frac{f^{\beta_{\text{rlm}}}_{m_t}(x_{t+1}|x_{1:(t-1)}, r_{t-1})}{f^{\beta_{\text{rlm}}}_{m_t}(x_{t+1}|x_0)} \\
&= \frac{\exp\left(\frac{1}{\beta_{\text{rlm}}-1} p(x_{t+1}|x_{1:t})^{\beta_{\text{rlm}}-1} - \frac{1}{\beta_{\text{rlm}}} \int p(\boldsymbol{z}|x_{1:t})^{\beta_{\text{rlm}}} d\boldsymbol{z}\right)}{\exp\left(\frac{1}{\beta_{\text{rlm}}-1} p(x_{t+1}|x_0)^{\beta_{\text{rlm}}-1} - \frac{1}{\beta_{\text{rlm}}} \int p(\boldsymbol{z}|x_0)^{\beta_{\text{rlm}}} d\boldsymbol{z}\right)} \\
&= \exp\left(\frac{1}{\beta_{\text{rlm}}-1}\left(p(x_{t+1}|x_{1:t})^{\beta_{\text{rlm}}-1} - p(x_{t+1}|x_0)^{\beta_{\text{rlm}}-1}\right) - \right. \\
&\qquad \left. \frac{1}{\beta_{\text{rlm}}} \int p(\boldsymbol{z}|x_{1:t})^{\beta_{\text{rlm}}} - p(\boldsymbol{z}|x_0)^{\beta_{\text{rlm}}} d\boldsymbol{z}\right). \qquad \text{(C.1)}
\end{aligned}
$$

This proof first seeks a lower bound for this ratio. A lower bound on $\frac{1}{\beta_{\text{rlm}}-1} p(x_{t+1}|x_{1:t})^{\beta_{\text{rlm}}-1}$ is 0, while the maximal value of $\frac{1}{\beta_{\text{rlm}}-1} p(x_{t+1}|x_0)^{\beta_{\text{rlm}}-1}$ will occur at the prior mode. For the multivariate t-distribution prior predictive with NIG hyperparameters a_0, b_0, μ_0, $\boldsymbol{\Sigma}_0$ of dimensions p the prior mode has density

$$
\begin{aligned}
& p(\boldsymbol{\mu}_0|\nu_0, \boldsymbol{\mu}_0, \boldsymbol{V}_0, p) \\
&= \frac{\Gamma((\nu_0+p)/2)}{\Gamma(\nu_0/2)\nu_0^{p/2}\pi^{p/2}\left|\boldsymbol{V}_0\right|^{1/2}}\left[1+\frac{1}{\nu_0}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_0)\right]^{-(\nu_0+p)/2} && \text{(C.2)} \\
&= \frac{\Gamma((\nu_0+p)/2)}{\Gamma(\nu_0/2)\nu_0^{p/2}\pi^{p/2}\left|\boldsymbol{V}_0\right|^{1/2}} && \text{(C.3)} \\
&= \frac{\Gamma(a_0+p/2)}{\Gamma(a_0)\left(2b_0\pi\right)^{p/2}\left|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T\right|^{1/2}}. && \text{(C.4)}
\end{aligned}
$$

Hence, the only term in the lower bound of $f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_{1:(t-1)}, r_{t-1})/f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_0)$ that does not solely depend on the prior parameters is $\frac{1}{\beta_{\text{rlm}}}\int p(\boldsymbol{z}|x_{1:t})^{\beta_{\text{rlm}}}d\boldsymbol{z}$. This term appears in the negative and thus to lower bound $f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_{1:(t-1)}, r_{t-1})/f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_0)$, an upper bound for $\frac{1}{\beta_{\text{rlm}}}\int p(\boldsymbol{z}|x_{1:t})^{\beta_{\text{rlm}}}d\boldsymbol{z}$ must be found. The multivariate t-distribution can be integrated as

$$
\begin{aligned}
& \frac{1}{\beta_{\text{rlm}}}\int \text{MVSt}_\nu(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{V})^{\beta_{\text{rlm}}}d\boldsymbol{z} \\
&= \frac{\Gamma((\nu+p)/2)^{\beta_{\text{rlm}}}\Gamma((\beta_{\text{rlm}}-1\nu+\beta_{\text{rlm}}-1p+\nu)/2)}{\Gamma(\nu/2)^{\beta_{\text{rlm}}}\Gamma((\beta_{\text{rlm}}-1\nu+\beta_{\text{rlm}}-1p+\nu+p)/2)}\times \\
&\quad \frac{1}{\beta_{\text{rlm}}(\nu\pi)^{(\beta_{\text{rlm}}-1p)/2}\left|\boldsymbol{V}\right|^{\beta_{\text{rlm}}-1/2}} \\
&= \frac{\Gamma((\nu+p)/2)^{\beta_{\text{rlm}}-1}\Gamma((\nu+p)/2)\Gamma((\beta_{\text{rlm}}-1\nu+\beta_{\text{rlm}}-1p+\nu)/2)}{\Gamma(\nu/2)^{\beta_{\text{rlm}}-1}\Gamma(\nu/2)\Gamma((\beta_{\text{rlm}}-1\nu+\beta_{\text{rlm}}-1p+\nu+p)/2)}\times \\
&\quad \frac{1}{\beta_{\text{rlm}}(\pi\nu)^{(\beta_{\text{rlm}}-1p)/2}\left|\boldsymbol{V}\right|^{\beta_{\text{rlm}}-1/2}} \\
&\leq \frac{\Gamma((\nu+p)/2)^{\beta_{\text{rlm}}-1}}{\Gamma(\nu/2)^{\beta_{\text{rlm}}-1}}\frac{1}{\beta_{\text{rlm}}(\pi\nu)^{(\beta_{\text{rlm}}-1p)/2}\left|\boldsymbol{V}\right|^{\beta_{\text{rlm}}-1/2}}. && \text{(C.5)}
\end{aligned}
$$

The inequality is derived from the fact that $\frac{\Gamma\left(x+\frac{p}{2}\right)}{\Gamma(x)}$ is increasing in x and as $\beta_{\text{rlm}}-1\geq 0$ and $\nu\geq 0$ then $(\beta_{\text{rlm}}-1\nu+\beta_{\text{rlm}}-1p+\nu)/2)\geq \nu/2$ which implies $\frac{\Gamma((\nu+p)/2)\Gamma((\beta_{\text{rlm}}-1\nu+\beta_{\text{rlm}}-1p+\nu)/2)}{\Gamma(\nu/2)\Gamma((\beta_{\text{rlm}}-1\nu+\beta_{\text{rlm}}-1p+\nu+p)/2)}\leq 1$.

Now employing the well-known result using Stirling's formula to bound the gamma function

$$
(2\pi)^{1/2}x^{x-1/2}\exp(-x)\leq\Gamma(x)\leq(2\pi)^{1/2}x^{x-1/2}\exp(1/(12x)-x) \qquad \text{(C.6)}
$$

we can therefore rewrite the ratio of gamma functions leaving

$$
\begin{aligned}
&\frac{1}{\beta_{\text{rlm}}}\int \text{MVSt}-t_\nu(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{V})^{\beta_{\text{rlm}}}d\boldsymbol{z}\\
&\leq \frac{\Gamma((\nu+p)/2)^{\beta_{\text{rlm}}-1}}{\Gamma(\nu/2)^{\beta_{\text{rlm}}-1}}\frac{1}{\beta_{\text{rlm}}(\pi\nu)^{(\beta_{\text{rlm}}-1p)/2}\left|\boldsymbol{V}\right|^{\beta_{\text{rlm}}-1/2}}\\
&\leq \frac{\left(\sqrt{2\pi}((\nu+p)/2)^{(\nu+p-1)/2)}\exp(-(\nu+p)/2+1/6(\nu+p)\right)^{\beta_{\text{rlm}}-1}}{\left(\sqrt{2\pi}(\nu/2)^{(\nu-1)/2}\exp(-\nu/2)\right)^{\beta_{\text{rlm}}-1}\beta_{\text{rlm}}(\pi\nu)^{(\beta_{\text{rlm}}-1p)/2}\left|\boldsymbol{V}\right|^{\beta_{\text{rlm}}-1/2}} \qquad \text{(C.7)}\\
&= ((1+\frac{p}{\nu})^{\beta_{\text{rlm}}-1(\nu+p-1)/2)}\exp(\beta_{\text{rlm}}-1(1/(6(\nu+p))-p/2))\times\\
&\qquad \frac{1}{\beta_{\text{rlm}}(\pi)^{(\beta_{\text{rlm}}-1p)/2}\left|\boldsymbol{V}\right|^{\beta_{\text{rlm}}-1/2}}. \qquad \text{(C.8)}
\end{aligned}
$$

Clearly $\exp\left(\beta_{\text{rlm}}-1(1/(6(\nu+p))-p/2)\right)$ is decreasing in ν for all p and to demonstrate when $((1+\frac{p}{\nu})^{\beta_{\text{rlm}}-1(\nu+p-1)/2)}$ is decreasing in ν we examine its derivative

$$
\begin{aligned}
w &= \left(1+\frac{p}{\nu}\right)^{\beta_{\text{rlm}}-1(\nu+p-1)/2} \qquad \text{(C.9)}\\
&= \exp\left((\beta_{\text{rlm}}-1(\nu+p-1)/2)\log\left(\left(1+\frac{p}{\nu}\right)\right)\right) \qquad \text{(C.10)}\\
\frac{dw}{d\nu} &= \frac{\beta_{\text{rlm}}-1}{2}\left(\log\left(1+\frac{p}{\nu}\right)-(\nu+p-1)\frac{\frac{p}{\nu^2}}{1+\frac{p}{\nu}}\right)\left(1+\frac{p}{\nu}\right)^{\beta_{\text{rlm}}-1(\nu+p-1)/2} \qquad \text{(C.11)}
\end{aligned}
$$

The sign of $\frac{dw}{d\nu}$ is dictated by $\left(\log\left(1+\frac{p}{\nu}\right)-(\nu+p-1)\frac{\frac{p}{\nu^2}}{1+\frac{p}{\nu}}\right)$, which can be demonstrated to be positive always if $p=1$ and negative always if $p>1$.

Case 1: when $p>1$, $\frac{1}{\beta_{\text{rlm}}}\int p(\boldsymbol{z}|x_{1:t})^{\beta_{\text{rlm}}}d\boldsymbol{z}$ is decreasing in ν and thus we can upper bound it by substituting the smallest value of ν. Here we bound ν above 1 in order to enforce that the mean of the predictive t-distribution exists. Under the KLD posterior it is clear that a_0 rises as more data is seen and while we do not have closed forms associated with the variational approximation to the $D_B^{(\beta)}$ posterior we expect this to be the case here. As more data is seen the finite sampling uncertainty, represented by ν in the NIG case, should be decreasing. Therefore provided a_0 is set such that $2a_0>1$, then this lower bound should never be violated.

Case 2: when $p=1$, Stirling's formula has failed to provide a decreasing upper bound for $\frac{1}{\beta_{\text{rlm}}}\int p(\boldsymbol{z}|x_{1:t})^{\beta_{\text{rlm}}}dz$. However in the univariate case

$$
\begin{aligned}
&\frac{1}{\beta_{\text{rlm}}} \int \text{St}_\nu(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{V})^{\beta_{\text{rlm}}} d\boldsymbol{z} \\
&\leq \frac{\Gamma((\nu+1)/2)^{\beta_{\text{rlm}}-1}}{\Gamma(\nu/2)^{\beta_{\text{rlm}}-1}} \frac{1}{\beta_{\text{rlm}}(\nu\,|\boldsymbol{V}|)^{(\beta_{\text{rlm}}-1)/2}\pi^{(\beta_{\text{rlm}}-1)/2}} \\
&\leq \frac{1}{\beta_{\text{rlm}}\,|\boldsymbol{V}|^{(\beta_{\text{rlm}}-1)/2}\,\pi^{(\beta_{\text{rlm}}-1)/2}}
\end{aligned}
$$

Where $p = 1$ is substituted into the bound from equation (C.5) and the inequality comes from that fact that $\frac{\Gamma((x+1)/2)}{\Gamma(x/2)} \leq \sqrt{x}$. This bound conveniently does not depend on the degrees of freedom ν at all.

We can therefore lower bound $f_{m_t}^{\beta_{\text{rlm}}-1}(x_{t+1}|x_{1:(t-1)}, r_{t-1})/f_{m_t}^{\beta_{\text{rlm}}-1}(x_{t+1}|x_0)$ as

$$
\begin{aligned}
&\frac{f_{m_t}^{\beta_{\text{rlm}}-1}(x_{t+1}|x_{1:(t-1)}, r_{t-1})}{f_{m_t}^{\beta_{\text{rlm}}-1}(x_{t+1}|x_0)} \\
&\geq \begin{cases}
\exp\left\{-\frac{1}{\beta_{\text{rlm}}-1}\left(\frac{\Gamma(a_0+1/2)}{\Gamma(a_0)(2b_0\pi)^{1/2}|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{1/2}}\right)^{\beta_{\text{rlm}}-1} - \frac{1}{\beta_{\text{rlm}}|\boldsymbol{V}|^{(\beta_{\text{rlm}}-1)/2}\pi^{(\beta_{\text{rlm}}-1)/2}} + \right. \\
\left. \frac{\Gamma(a_0+1/2)^{\beta_{\text{rlm}}}\Gamma(\beta_{\text{rlm}}-1a_0+\beta_{\text{rlm}}-1/2+a_0)}{\Gamma(a_0)^{\beta_{\text{rlm}}}\Gamma(\beta_{\text{rlm}}-1a_0+\beta_{\text{rlm}}-1/2+a_0+1/2)} \frac{1}{\beta_{\text{rlm}}(2\pi b_0)^{(\beta_{\text{rlm}}-1)/2}|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{\beta_{\text{rlm}}-1/2}}\right\} & \text{if } p = 1 \\
\exp\left\{-\frac{1}{\beta_{\text{rlm}}-1}\left(\frac{\Gamma(a_0+p/2)}{\Gamma(a_0)(2b_0\pi)^{p/2}|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{1/2}}\right)^{\beta_{\text{rlm}}-1} + \right. \\
\frac{\Gamma(a_0+p/2)^{\beta_{\text{rlm}}}\Gamma(\beta_{\text{rlm}}-1a_0+\beta_{\text{rlm}}-1p/2+a_0)}{\Gamma(a_0)^{\beta_{\text{rlm}}}\Gamma(\beta_{\text{rlm}}-1a_0+\beta_{\text{rlm}}-1p/2+a_0+p/2)} \frac{1}{\beta_{\text{rlm}}(2\pi b_0)^{(\beta_{\text{rlm}}-1p)/2}|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{\beta_{\text{rlm}}-1/2}} - \\
\left. ((1+p)^{\beta_{\text{rlm}}-1p/2)}\exp(\beta_{\text{rlm}} - 1(1/(6(1+p)) - p/2))\frac{1}{\beta_{\text{rlm}}(\pi)^{(\beta_{\text{rlm}}-1p)/2}|\boldsymbol{V}|^{\beta_{\text{rlm}}-1/2}}\right\} & \text{if } p > 1
\end{cases}
\end{aligned}
$$

Now fixing $p, a_0, b_0, \mu_0, \Sigma_0$ and $|V|_{\text{min}}$ which values of $\beta_{\text{rlm}} - 1$ and $H(r_t, r_{t+1})$ would leave

$$
\frac{1 - H(r_t, r_{t+1}))}{H(r_t, r_{t+1})} \frac{f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_{1:(t-1)}, r_{t-1})}{f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_0)} \geq 1? \tag{C.12}
$$

We demonstrate this for $p > 1$ but it is straightforward to see that it extends to when $p = 1$. Rearranging the inequality in equation (C.12) gives us that (C.12) holds providing

$$
\begin{aligned}
&\frac{1}{|\boldsymbol{V}|^{\beta_{\text{rlm}}-1/2}} \leq \\
&\left[\frac{\Gamma(a_0+p/2)^{\beta_{\text{rlm}}-1}}{\Gamma(a_0)^{\beta_{\text{rlm}}-1}\,(2b_0\pi)^{\beta_{\text{rlm}}-1p/2}\,|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{\beta_{\text{rlm}}-1/2}}\times\right. \\
&\left(\frac{\Gamma(a_0+p/2)\Gamma(\beta_{\text{rlm}}-1a_0+\beta_{\text{rlm}}-1p/2+a_0)}{\Gamma(a_0)\Gamma(\beta_{\text{rlm}}-1a_0+\beta_{\text{rlm}}-1p/2+a_0+p/2)}\frac{1}{\beta_{\text{rlm}}}-\frac{1}{\beta_{\text{rlm}}-1}\right) \\
&\left.+\log\left(\frac{1-H(r_t,r_{t+1}))}{H(r_t,r_{t+1})}\right)\right]\times \\
&\frac{\beta_{\text{rlm}}(\pi)^{(\beta_{\text{rlm}}-1p)/2}}{((1+\frac{p}{2a_0})^{\alpha(2a_0+p-1)/2)}\exp(\beta_{\text{rlm}}-1(1/(6(2a_0+p))-p/2))}
\end{aligned}
$$

We define the set defined by inequality (C.13) as

$$
\begin{aligned}
&S\left(p,\beta_{\text{rlm}},a_0,b_0,\mu_0,\boldsymbol{\Sigma}_0,|\boldsymbol{V}|_{\min}\right) \\
&=\left\{(\beta_{\text{rlm}},H(r_t,r_{t+1}))\text{s.t. (C.13) is satisfied for } p,\beta_{\text{rlm}}-1,a_0,b_0,\mu_0,\boldsymbol{\Sigma}_0,|\boldsymbol{V}|_{\min}\right\}.
\end{aligned}
$$

As a result, we can see that for fixed of $a_0, b_0, \mu_0, \Sigma_0$ and $|\boldsymbol{V}| \geq |\boldsymbol{V}|_{\min}$ it is always possible to choose values of β_{rlm} and $H(r_t, r_{t+1})$ such that this holds. To see this consider fixing β_{rlm}, the the upper bound is simply increasing in $\log\left(\frac{1-H(r_t,r_{t+1})}{H(r_t,r_{t+1})}\right)$ which takes values in $\mathbb{R}$ and thus can be set large enough so that the inequality holds. □

We note that in practice this results is likely to be stronger than is necessary. The observation that is most likely to generate a change-point will have 0 mass under the predictive associated with the current segment but also appears at the prior mode. While this was necessary to demonstrate this result for all situations this is incredibly unlikely to occur. The requirement for $|\boldsymbol{V}_{\min}|$ is a result of the beta-divergence loss function depending on $\int p(z|x_{1:t})^{\beta_{\text{rlm}}}dz$. In the proof of this result we demonstrate that $f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_{1:(t-1)},r_{t-1})/f_{m_t}^{\beta_{\text{rlm}}}(x_{t+1}|x_0)$ is increasing in $|\boldsymbol{V}|$ and as a result if it is allowed to get too small the inequality in equation (C.13) would not hold. This is an undesirable consequence of the beta-divergence score not being completely local, that is to say not solely depending on the predictive probability of the observation, thus the score under the prior can be quite a lot bigger than the score under the continuing run length independent of the observations seen and solely based on the predictive covariances.

C.3 Proofs of KSD-Bayes Theoretical Results

This appendix provides proofs for all theoretical results in the main text. On occasion we refer to auxiliary theoretical results, which are given in Appendix A.3

C.3.1 Preliminaries

The following properties of the Stein operator $\mathcal{S}_{\mathbb{Q}}$ will be useful:

Lemma C.1. Under Assumption 8.1, we have, for all $x, x' \in \mathcal{X}$ and $h \in \mathcal{H}$,

(i) $\mathcal{S}_{\mathbb{Q}} K(x, \cdot) \in \mathcal{H}$,

(ii) $\mathcal{S}_{\mathbb{Q}}[h](x) = \langle h(\cdot), \mathcal{S}_{\mathbb{Q}} K(x, \cdot) \rangle_{\mathcal{H}}$,

(iii) $|\mathcal{S}_{\mathbb{Q}} \mathcal{S}_{\mathbb{Q}} K(x, x')| \leq \sqrt{\mathcal{S}_{\mathbb{Q}} \mathcal{S}_{\mathbb{Q}} K(x, x)} \sqrt{\mathcal{S}_{\mathbb{Q}} \mathcal{S}_{\mathbb{Q}} K(x', x')}$.

Proof. First of all, since $h \mapsto \mathcal{S}_{\mathbb{Q}}[h](x)$ is a continuous linear functional on $\mathcal{H}$ for each fixed $x \in \mathcal{X}$ by assumption, from the Riesz representation theorem (Steinwart and Christmann, 2008, Theorem A.5.12) there exists a so-called *representer* $g_x \in \mathcal{H}$ for each fixed $x \in \mathcal{X}$ s.t.

$$\mathcal{S}_{\mathbb{Q}}[h](x) = \langle h, g_x \rangle_{\mathcal{H}}.$$

Second of all, the reproducing property $h(x') = \langle h(\cdot), K(\cdot, x') \rangle_{\mathcal{H}}$ holds for any $h \in \mathcal{H}$, where we recall that the inner product between $h \in \mathcal{H}$ and a matrix-valued function $K(x, \cdot)$ is defined in Appendix A.3. By the reproducing property, for all $x, x' \in \mathcal{X}$,

$$g_x(x') = \langle g_x, K(\cdot, x') \rangle_{\mathcal{H}} = \mathcal{S}_{\mathbb{Q}} \left[K(\cdot, x') \right](x) = \mathcal{S}_{\mathbb{Q}} K(x, x'). \tag{C.13}$$

In particular, $\mathcal{S}_{\mathbb{Q}} K(x, \cdot) \in \mathcal{H}$ since $g_x \in \mathcal{H}$, establishing (i). Based on these two observations, we can rewrite $\mathcal{S}_{\mathbb{Q}}[h](x)$ at each fixed $x \in \mathcal{X}$ as

$$\mathcal{S}_{\mathbb{Q}}[h](x) = \langle h, g_x \rangle_{\mathcal{H}} = \langle h(\cdot), \mathcal{S}_{\mathbb{Q}} K(x, \cdot) \rangle_{\mathcal{H}}, \tag{C.14}$$

establishing (ii). We now apply (C.14) with $h(\cdot) = \mathcal{S}_{\mathbb{Q}} K(x', \cdot)$ to deduce that

$$\mathcal{S}_{\mathbb{Q}} \mathcal{S}_{\mathbb{Q}} K(x', x) = \mathcal{S}_{\mathbb{Q}} \left[\mathcal{S}_{\mathbb{Q}} K(x', \cdot) \right](x) = \langle \mathcal{S}_{\mathbb{Q}} K(x', \cdot), \mathcal{S}_{\mathbb{Q}} K(x, \cdot) \rangle_{\mathcal{H}}. \tag{C.15}$$

Applying the Cauchy-Schwarz inequality,

$$|\mathcal{S}_{\mathbb{Q}} \mathcal{S}_{\mathbb{Q}} K(x, x')| = |\langle \mathcal{S}_{\mathbb{Q}} K(x, \cdot), \mathcal{S}_{\mathbb{Q}} K(x', \cdot) \rangle_{\mathcal{H}}| \leq \|\mathcal{S}_{\mathbb{Q}} K(x, \cdot)\|_{\mathcal{H}} \|\mathcal{S}_{\mathbb{Q}} K(x', \cdot)\|_{\mathcal{H}}.$$

Here for each $x \in \mathcal{X}$ the norm term can computed using (C.15):

$$\|\mathcal{S}_{\mathbb{Q}} K(x,\cdot)\|_{\mathcal{H}} = \sqrt{\langle \mathcal{S}_{\mathbb{Q}} K(x,\cdot), \mathcal{S}_{\mathbb{Q}} K(x,\cdot)\rangle_{\mathcal{H}}} = \sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x,x)}$$

Therefore for all $x, x' \in \mathcal{X}$ we have

$$|\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x,x')| \leq \sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x,x)}\sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(x',x')},$$

establishing (iii). □

Proof of Proposition 6.1

Proof. From (ii) of Lemma C.1, for each $x \in \mathcal{X}$, $h \in \mathcal{H}$, we have

$$\mathcal{S}_{\mathbb{Q}}[h](x) = \langle h(\cdot), \mathcal{S}_{\mathbb{Q}} K(x,\cdot)\rangle_{\mathcal{H}}.$$

Taking the expectation of both sides,

$$\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] = \mathbb{E}_{X\sim\mathbb{P}}\left[\langle h(\cdot), \mathcal{S}_{\mathbb{Q}} K(X,\cdot)\rangle_{\mathcal{H}}\right] = \langle h(\cdot), \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X,\cdot)\right]\rangle_{\mathcal{H}} \quad \text{(C.16)}$$

Here since the inner product is continuous liner operator, the expectation and inner product can be exchanged if the function $x \mapsto \mathcal{S}_{\mathbb{Q}} K(x,\cdot)$ is *Bochner* $\mathbb{P}$*-integrable* (Steinwart and Christmann, 2008, A.32). This is indeed the case, since from (ii) of Lemma C.1 again, and Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{X\sim\mathbb{P}}\left[\|\mathcal{S}_{\mathbb{Q}} K(X,\cdot)\|_{\mathcal{H}}\right] &= \mathbb{E}_{X\sim\mathbb{P}}\left[\sqrt{\langle \mathcal{S}_{\mathbb{Q}} K(X,\cdot), \mathcal{S}_{\mathbb{Q}} K(X,\cdot)\rangle_{\mathcal{H}}}\right] \quad \text{(C.17)}\\ &= \mathbb{E}_{X\sim\mathbb{P}}\left[\sqrt{\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(X,X)}\right] \leq \sqrt{\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(X,X)\right]} < \infty\end{aligned}$$

where the last term is finite by Assumption 8.1. A standard argument based on the Cauchy–Schwarz inequality gives

$$\begin{aligned}\sup_{\|h\|_{\mathcal{H}}\leq 1} \left|\left\langle h(\cdot), \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X,\cdot)\right]\right\rangle_{\mathcal{H}}\right| &= \left\|\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X,\cdot)\right]\right\|_{\mathcal{H}}\\ &= \sqrt{\left\langle \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X,\cdot)\right], \mathbb{E}_{X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}} K(X',\cdot)\right]\right\rangle_{\mathcal{H}}}\\ &= \sqrt{\mathbb{E}_{X,X'\sim\mathbb{P}}\left[\left\langle \mathcal{S}_{\mathbb{Q}} K(X,\cdot), \mathcal{S}_{\mathbb{Q}} K(X',\cdot)\right\rangle_{\mathcal{H}}\right]}\\ &= \sqrt{\mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}} K(X,X')\right]} \quad \text{(C.18)}\end{aligned}$$

where X and X' are independent, and we again appeal to Bochner $\mathbb{P}$-integrability to interchange expectation and inner product. Thus from (C.16) and (C.18) we have

$$\mathrm{KSD}^2(\mathbb{Q}\|\mathbb{P}) = \left(\sup_{\|h\|_{\mathcal{H}}\leq 1} \left|\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right]\right|\right)^2 = \mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(X,X')\right],$$

as claimed. □

Verifying Assumption 8.1 for the Langevin Stein Operator

This section demonstrates how to verify the assumption that $h \mapsto \mathcal{S}_{\mathbb{Q}}[h](x)$ is a continuous linear functional on $\mathcal{H}$ for each fixed $x \in \mathcal{X}$ in the case where $\mathcal{S}_{\mathbb{Q}}$ is the Langevin Stein operator (6.4) for $\mathbb{Q} \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$. Since a linear functional is continuous if and only if it is bounded, we aim to show that, for each fixed $x \in \mathcal{X}$, there exist a constant C_x s.t. $|\mathcal{S}_{\mathbb{Q}}[h](x)| \leq C_x\|h\|_{\mathcal{H}}$ for all $h \in \mathcal{H}$.

For each fixed $x \in \mathbb{R}^d$, the Langevin Stein operator $\mathcal{S}_{\mathbb{Q}}$ is given as

$$\mathcal{S}_{\mathbb{Q}}[h](x) = \nabla \log q(x) \cdot h(x) + \nabla \cdot h(x).$$

From the reproducing property $h(x) = \langle h, K(x,\cdot)\rangle_{\mathcal{H}}$ for any $h \in \mathcal{H}$, we have

$$\begin{aligned}\mathcal{S}_{\mathbb{Q}}[h](x) &= \nabla \log q(x) \cdot \langle h, K(x,\cdot)\rangle_{\mathcal{H}} + \nabla_x \cdot \langle h, K(x,\cdot)\rangle_{\mathcal{H}} \\ &= \langle h, K(x,\cdot)\nabla \log q(x)\rangle_{\mathcal{H}} + \langle h, \nabla_x \cdot K(x,\cdot)\rangle_{\mathcal{H}}\end{aligned} \tag{C.19}$$

where the order of inner product and other operators is exchangeable by the continuity of $\langle h,\cdot\rangle_{\mathcal{H}} : \mathcal{H} \to \mathbb{R}$ (Steinwart and Christmann, 2008, Corollary 4.36). Then by the Cauchy–Schwarz inequality,

$$\begin{aligned}|\mathcal{S}_{\mathbb{Q}}[h](x)| &\leq \left(\|K(x,\cdot)\nabla \log q(x)\|_{\mathcal{H}} + \|\nabla_x \cdot K(x,\cdot)\|_{\mathcal{H}}\right)\|h\|_{\mathcal{H}} \\ &= \left(\sqrt{\nabla \log q(x) \cdot K(x,x)\nabla \log q(x)} + \sqrt{\nabla \cdot (\nabla \cdot K(x,x))}\right)\|h\|_{\mathcal{H}} =: C_x\|h\|_{\mathcal{H}}.\end{aligned} \tag{C.20}$$

where the first and second gradient of $\nabla \cdot (\nabla \cdot K(x,x))$ are taken each with respect to the first and second argument of K. For the constant C_x to exist, it is sufficient to require that $\nabla \log q(x)$, $K(x,x)$ and $\nabla \cdot (\nabla \cdot K(x,x))$ exist. This is the case when, for example, $\mathbb{Q} \in \mathcal{P}_{\mathcal{S}}(\mathbb{R}^d)$ and $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$, as assumed in Jackson Gorham and Lester Mackey (2017).

C.3.2 Proof of Proposition 8.1

Proof. From (6.7), $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K$ is given by

$$\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x,x') \overset{+C}{=} \underbrace{\nabla \log p(x|\boldsymbol{\theta}) \cdot K(x,x') \nabla \log p(x'|\boldsymbol{\theta})}_{(*_1)} + \underbrace{\nabla \log p(x|\boldsymbol{\theta}) \cdot \big(\nabla_{x'} \cdot K(x,x')\big)}_{(*_2)} + \underbrace{\nabla \log p(x'|\boldsymbol{\theta}) \cdot \big(\nabla_{x} \cdot K(x,x')\big)}_{(*_3)},$$

where $\overset{+C}{=}$ indicates equality up to an additive term that is independent of $\boldsymbol{\theta}$. The exponential family model in (8.4) satisfies $\nabla \log p(x|\boldsymbol{\theta}) = \nabla t(x)\eta(\boldsymbol{\theta}) + \nabla b(x)$. Thus for term $(*_1)$ we have

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}(*_1) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}(\nabla t(x_i)\eta(\boldsymbol{\theta})) \cdot K(x_i,x_j)\nabla t(x_j)\eta(\boldsymbol{\theta}) + \nabla b(x_i) \cdot K(x_i,x_j)\nabla t(x_j)\eta(\boldsymbol{\theta}) \\
&\qquad +(\nabla t(x_i)\eta(\boldsymbol{\theta})) \cdot K(x_i,x_j)\nabla b(x_j) + \nabla b(x_i) \cdot K(x_i,x_j)\nabla b(x_j) \\
&\overset{+C}{=} \eta(\boldsymbol{\theta}) \cdot \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \nabla t(x_i)^\top K(x_i,x_j)\nabla t(x_j)\right)\eta(\boldsymbol{\theta}) \\
&\qquad +\eta(\boldsymbol{\theta}) \cdot \left(2\sum_{i=1}^{n}\sum_{j=1}^{n} \nabla t(x_i)^\top K(x_i,x_j)\nabla b(x_j)\right)
\end{aligned} \tag{C.21}$$

where the last equality follows from symmetry of K. For terms $(*_2)$ and $(*_3)$,

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n}(*_2) &= \sum_{i=1}^{n}\sum_{j=1}^{n}(\nabla t(x_i)\eta(\boldsymbol{\theta})) \cdot (\nabla_{x'} \cdot K(x_i,x_j)) + \nabla b(x_i) \cdot (\nabla_{x'} \cdot K(x_i,x_j)) \\
&\overset{+C}{=} \eta(\boldsymbol{\theta}) \cdot \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \nabla t(x_i)^\top (\nabla_{x'} \cdot K(x_i,x_j))\right),
\end{aligned} \tag{C.22}$$

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n}(*_3) &= \sum_{i=1}^{n}\sum_{j=1}^{n}(\nabla t(x_i)\eta(\boldsymbol{\theta})) \cdot (\nabla_{x} \cdot K(x_i,x_j)) + \nabla b(x_j) \cdot (\nabla_{x} \cdot K(x_i,x_j)) \\
&\overset{+C}{=} \eta(\boldsymbol{\theta}) \cdot \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \nabla t(x_j)^\top (\nabla_{x} \cdot K(x_i,x_j))\right).
\end{aligned} \tag{C.23}$$

From 8.2, the KSD-Bayes posterior is

$$\pi_n^D(\boldsymbol{\theta}) \propto \pi(\boldsymbol{\theta}) \exp\left(-\beta n \left\{ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (*_1) + (*_2) + (*_3) \right\}\right),$$

so we may collect together terms in (C.21), (C.22), and (C.23) to obtain the expressions in Proposition 8.1. □

C.3.3 Proofs of Results in Section 8.4.1

Proof of Theorem 8.1 (a.s. Pointwise Convergence)

Proof. Let $f_n(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n)$ and $f(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P})$. Decomposing the double summation of $f_n(\boldsymbol{\theta})$ into the diagonal term $(i = j)$ and non-diagonal term $(i \neq j)$,

$$\begin{aligned} f_n(\boldsymbol{\theta}) &= \frac{1}{n^2} \sum_{i=1}^{n} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i, x_i) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} (x_i, x_j) \\ &= \frac{1}{n} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i, x_i)}_{(*_a)} + \frac{n-1}{n} \underbrace{\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i, x_i)}_{(*_b)}. \end{aligned}$$

Fix $\boldsymbol{\theta} \in \boldsymbol{\theta}$. From the strong law of large number (Durrett, 2010, Theorem 2.5.10),

$$(*_a) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i, x_i) \xrightarrow{a.s.} \mathbb{E}_{X \sim \mathbb{P}} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X) \right], \tag{C.24}$$

provided that $\mathbb{E}_{X \sim \mathbb{P}} \left[|\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X)| \right] < \infty$. From the positivity of $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x)$, we have $\mathbb{E}_{X \sim \mathbb{P}} \left[|\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X)| \right] = \mathbb{E}_{X \sim \mathbb{P}} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X) \right]$, which has been assumed to exist. The form of (b) is called an *unbiased statistic* (or *U-statistic* for short) and Wassily Hoeffding (1961) proved the strong law of large numbers

$$(*_b) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i, x_j) \xrightarrow{a.s.} \mathbb{E}_{X, X' \sim \mathbb{P}} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X') \right], \tag{C.25}$$

whenever $\mathbb{E}_{X, X' \sim \mathbb{P}} \left[|\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X')| \right] < \infty$. From item (iii) of Lemma C.1 and Jensen's inequality, we have $\mathbb{E}_{X, X' \sim \mathbb{P}} \left[|\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X')| \right] \leq \mathbb{E}_{X \sim \mathbb{P}} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X) \right]$ where the right hand side is again assumed to exist. Therefore, since $1/n \to 0$ and

$(n-1)/n \to 1$,

$$f_n(\boldsymbol{\theta}) = \frac{1}{n}(*_a) + \frac{n-1}{n}(*_b) \xrightarrow{a.s.} \mathbb{E}_{X,X' \sim \mathbb{P}}\left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(X,X')\right] = f(\boldsymbol{\theta}), \quad \text{(C.26)}$$

where the argument holds for each fixed $\boldsymbol{\theta} \in \boldsymbol{\theta}$. □

Proof of Theorem 8.2 (a.s. Uniform Convergence)

Proof. Let $f_n(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n)$ and $f(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P})$. Recall that $\boldsymbol{\theta} \subset \mathbb{R}^p$ is bounded. Theorem 21.8 in Davidson (1994) implies that $f_n \xrightarrow{a.s.} f$ uniformly on $\boldsymbol{\Theta}$ if and only if (a) $f_n \xrightarrow{a.s.} f$ pointwise on $\boldsymbol{\Theta}$ and (b) $\{f_n\}_{n=1}^{\infty}$ is strongly stochastically equicontinuous on $\boldsymbol{\Theta}$. The condition (a) is immediately implied by Theorem 8.1 and we hence show the condition (b) in the remainder.

By Davidson (1994, Theorem 21.10), $\{f_n\}_{n=1}^{\infty}$ is strongly stochastically equicontinuous on $\boldsymbol{\Theta}$ if there exists a stochastic sequence $\{\mathcal{L}_n\}_{n=1}^{\infty}$, independent of $\boldsymbol{\theta}$, s.t.

$$|f_n(\boldsymbol{\theta}) - f_n(\boldsymbol{\theta}')| \leq \mathcal{L}_n \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \ \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta} \quad \text{and} \quad \limsup_{n\to\infty} \mathcal{L}_n < \infty \text{ a.s.} \quad \text{(C.27)}$$

Since f_n is continuously differentiable on $\boldsymbol{\theta}$, and $\boldsymbol{\theta}$ is assumed to be open and convex, the mean value theorem yields

$$|f_n(\boldsymbol{\theta}) - f_n(\boldsymbol{\theta}')| \leq \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\nabla_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\|_2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \quad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}. \quad \text{(C.28)}$$

Lemma C.9 (the first of our auxiliary results, stated and proved in Appendix C.4) implies that $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\nabla_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\|_2 < \infty$ a.s. for all sufficiently large n. Therefore, setting $\mathcal{L}_n = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\nabla_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\|_2$ concludes the proof. □

Proof of Lemma 8.3 (Strong Consistency)

The following result from real analysis will be required:

Lemma C.2. Let $\boldsymbol{\theta} \subset \mathbb{R}^p$ be open and bounded. Let $f_n : \boldsymbol{\theta} \to \mathbb{R}$ and $f : \boldsymbol{\theta} \to \mathbb{R}$ be continuous functions. Assume that (i) there exists an unique $\boldsymbol{\theta}_* \in \boldsymbol{\theta}$ s.t. $f(\boldsymbol{\theta}_*) < \inf_{\{\boldsymbol{\theta} \in \boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \geq \epsilon\}} f(\boldsymbol{\theta})$ for any $\epsilon > 0$, and (ii) $\sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}} |f_n(\boldsymbol{\theta}) - f(\boldsymbol{\theta})| \to 0$ as $n \to \infty$. Let $\{\boldsymbol{\theta}_n\}_{n=1}^{\infty}$ be any sequence s.t. $\boldsymbol{\theta}_n \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\theta}} f_n(\boldsymbol{\theta})$ for all sufficiently large n. Then $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_*$ as $n \to \infty$.

Proof. The following argument is similar to that used in van der Vaart (1998, Theorem 5.7) and Whitney K. Newey and Daniel McFadden (1994, Theorem 2.1). Fix $\eta > 0$ and consider n sufficiently large that $\boldsymbol{\theta}_n$ is well-defined. From (ii), for all sufficiently large n, we have the uniform bound $|f(\boldsymbol{\theta}) - f_n(\boldsymbol{\theta})| < \eta/2$ over $\boldsymbol{\theta} \in \boldsymbol{\theta}$. Since

$\boldsymbol{\theta}_n$ is a minimiser of f_n, we therefore have $f(\boldsymbol{\theta}_n) < f_n(\boldsymbol{\theta}_n) + \eta/2 < f_n(\boldsymbol{\theta}_*) + \eta/2 < f(\boldsymbol{\theta}_*)+\eta$. Since $\eta > 0$ was arbitrary, we may take $\eta = \inf_{\{\boldsymbol{\theta}\in\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\geq\epsilon\}} f(\boldsymbol{\theta})-f(\boldsymbol{\theta}_*)$, where $\eta > 0$ from (i), to see that $f(\boldsymbol{\theta}_n) < \inf_{\{\boldsymbol{\theta}\in\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\geq\epsilon\}} f(\boldsymbol{\theta})$. Thus we have shown that $\boldsymbol{\theta}_n \in \{\boldsymbol{\theta} \in \boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 < \epsilon\}$ for all sufficiently large n. Since the argument holds for $\epsilon > 0$ arbitrarily small, the result is established. □

Now we can prove Lemma 8.3:

Proof of Lemma 8.3. Let $f_n(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n)$ and $f(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})$. From Assumption 8.3, there exists an unique $\boldsymbol{\theta}_* \in \boldsymbol{\theta}$ s.t. $f(\boldsymbol{\theta}_*) < \inf_{\{\boldsymbol{\theta}\in\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\geq\epsilon\}} f(\boldsymbol{\theta})$ for any $\epsilon > 0$, and $\boldsymbol{\theta}_n \in \boldsymbol{\theta}$ minimises f_n a.s. for all sufficiently large n. Since Assumption 8.2 ($r_{\max} = 1$) hold, f_n is continuous a.s. and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} |f_n(\boldsymbol{\theta})-f(\boldsymbol{\theta})| \overset{a.s.}{\to} 0$ by Theorem 8.2. Thus the conditions of Lemma C.2 are a.s. satisfied, from which it follows that $\boldsymbol{\theta}_n \overset{a.s.}{\longrightarrow} \boldsymbol{\theta}_*$. □

Proof of Lemma 8.4 (Asymptotic Normality)

Proof. Let $f_n(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n)$ and $f(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})$. It was assumed that, for any $h \in \mathcal{H}$ and $x \in \mathcal{X}$, the map $\boldsymbol{\theta} \mapsto \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h](x)$ is three times continuously differentiable, from which it follows that f_n is three times continuously differentiable as well. Since $\boldsymbol{\theta}_n$ minimises f_n for all sufficiently large n, we have $\nabla f_n(\boldsymbol{\theta}_n) = 0$. Hence a second order Taylor expansion around $\boldsymbol{\theta}_*$ yields

$$0 = \nabla f_n(\boldsymbol{\theta}_n) = \nabla f_n(\boldsymbol{\theta}_*) + \nabla^2 f_n(\boldsymbol{\theta}_*)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) + (\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) \cdot \nabla^3 f_n(\boldsymbol{\theta}'_n)(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*)$$

where $\boldsymbol{\theta}'_n = \alpha\boldsymbol{\theta}_* + (1-\alpha)\boldsymbol{\theta}_n$ for some $\alpha \in [0,1]$. By transposing the terms properly and scaling the both side by $\sqrt{n}$, we have

$$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) = \Big(\underbrace{\nabla^2 f_n(\boldsymbol{\theta}_*)}_{(*_1)} + \underbrace{(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*) \cdot \nabla^3 f_n(\boldsymbol{\theta}'_n)}_{(*_2)} \Big)^{-1} \Big(- \underbrace{\sqrt{n}\nabla f_n(\boldsymbol{\theta}_*)}_{(*_3)} \Big). \quad \text{(C.29)}$$

In the remainder, we show the convergence of $(*_1)$, $(*_2)$ and $(*_3)$, and apply the Slutsky's theorem to see the convergence in distribution of $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_n)$.

Term $(*_1)$: From the auxiliary result Lemma C.10 in Appendix C.4, we have that $\nabla^2 f_n(\boldsymbol{\theta}_*) \overset{a.s.}{\to} \nabla^2 f(\boldsymbol{\theta}_*) = H_*$ where H_* is positive semi-definite.

Term $(*_2)$: From the Cauchy–Schwarz inequality and auxiliary result Lemma C.9

in Section C.4,

$$\limsup_{n\to\infty} \left\|(\boldsymbol{\theta}_n - \boldsymbol{\theta}_*)\cdot \nabla^3 f_n(\boldsymbol{\theta}_n')\right\|_2 \le \limsup_{n\to\infty} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\nabla^3 f_n(\boldsymbol{\theta})\|_2 \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|_2$$
$$\le \underbrace{\limsup_{n\to\infty} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\nabla^3 f_n(\boldsymbol{\theta})\|_2}_{<\infty \text{ a.s.}} \times \limsup_{n\to\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|_2 \tag{C.30}$$

Since Lemma 8.3 implies that $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|_2 \xrightarrow{a.s.} 0$, we have $(*_2) \xrightarrow{a.s.} 0$.

Term $(*_3)$: Let $F(x,x') := \nabla_{\boldsymbol{\theta}}(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x,x'))|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} \in \mathbb{R}^p$ and recall that $S(x,\boldsymbol{\theta}_*) = \mathbb{E}_{X\sim\mathbb{P}}[F(x,X)] \in \mathbb{R}^p$. Then

$$\sqrt{n}\nabla f_n(\boldsymbol{\theta}_*) = \sqrt{n}\left(\frac{1}{n^2}\sum_{i=1}^n F(x_i,x_i) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j\neq i}^n F(x_i,x_j)\right) \tag{C.31}$$
$$= \frac{1}{\sqrt{n}}\underbrace{\frac{1}{n}\sum_{i=1}^n F(x_i,x_i)}_{(*_a)} + \frac{n-1}{n}\underbrace{\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n F(x_i,x_j)}_{(*_b)}. \tag{C.32}$$

First, it follows from the strong law of large number (Durrett, 2010, Theorem 2.5.10) that $(*_a) \xrightarrow{a.s.} \mathbb{E}_{X\sim\mathbb{P}}[F(X,X)]$ whenever $\mathbb{E}_{X\sim\mathbb{P}}[\|F(X,X)\|_2] < \infty$. Second, since $(*_b)$ is a U-statistic multiplied by $\sqrt{n}$, it follows from van der Vaart (1998, Theorem 12.3) that $(*_b) \xrightarrow{p} (1/\sqrt{n})\sum_{i=1}^n S(x_i,\boldsymbol{\theta}_*)$ whenever $\mathbb{E}_{X,X'\sim\mathbb{P}}[\|F(X,X')\|_2^2] < \infty$. (Here $\xrightarrow{p}$ denotes convergence in probability.) Both the required conditions indeed hold from the auxiliary result Lemma C.11 in Appendix C.4. Thus we have

$$\sqrt{n}\nabla f_n(\boldsymbol{\theta}_*) = \frac{1}{\sqrt{n}}(*_a) + \frac{n-1}{n}(*_b) \xrightarrow{p} \frac{1}{\sqrt{n}}\sum_{i=1}^n S(x_i,\boldsymbol{\theta}_*). \tag{C.33}$$

This convergence in probability implies that $\sqrt{n}\nabla f_n(\boldsymbol{\theta}_*)$ and $(1/\sqrt{n})\sum_{i=1}^n S(x_i,\boldsymbol{\theta}_*)$ converge in distribution to the same limit. Therefore we may apply the central limit theorem for $(1/\sqrt{n})\sum_{i=1}^n S(x_i,\boldsymbol{\theta}_*)$ to obtain the asymptotic distribution of $\sqrt{n}\nabla f_n(\boldsymbol{\theta}_*)$. Again from van der Vaart (1998, Theorem 12.3), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n S(x_i,\boldsymbol{\theta}_*) \xrightarrow{d} \mathcal{N}(0,J_*), \qquad J_* = \mathbb{E}_{X\sim\mathbb{P}}\left[S(X,\boldsymbol{\theta}_*)S(X,\boldsymbol{\theta}_*)^\top\right] \tag{C.34}$$

whenever $\mathbb{E}_{X,X'\sim\mathbb{P}}\left[\|F(X,X')\|_2^2\right] < \infty$, which implies the existence of the covariance matrix J_*. Hence $\sqrt{n}\nabla f_n(\boldsymbol{\theta}_*) \xrightarrow{d} \mathcal{N}(0,J_*)$.

Collecting together these results, we have shown that

$$(*_1) \xrightarrow{a.s.} H_*, \qquad (*_2) \xrightarrow{a.s.} 0, \qquad (*_3) \xrightarrow{d} \mathcal{N}(0, J_*). \tag{C.35}$$

Since H_* is guaranteed to be at least positive semi-definite, it is in fact strictly positive definite if H_* is non-singular, as we assumed. Finally, Slutsky's theorem allows us to conclude that $\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_n) \xrightarrow{d} \mathcal{N}\left(0, H_*^{-1} J_* H_*^{-1}\right)$ as claimed. □

Verifying Assumption 8.2 for the Langevin Stein Operator

Here we compute the quantities involved in Assumption 8.2 for the Langevin Stein operator $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$ with $\mathbb{P}_{\boldsymbol{\theta}} \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$. In this case,

$$\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h](x) = \partial^r\left(\nabla_x \log p(x|\boldsymbol{\theta}) \cdot h(x)\right) + \partial^r\left(\nabla_x \cdot h(x)\right) = \left(\partial^r \nabla_x \log p(x|\boldsymbol{\theta})\right) \cdot h(x) \tag{C.36}$$

The operator $\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$ in (C.36) is therefore well-defined and $\boldsymbol{\theta} \mapsto \partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h](x)$ is continuous whenever $\boldsymbol{\theta} \mapsto \nabla_x \log p(x|\boldsymbol{\theta})$ is r-times continuously differentiable over $\boldsymbol{\Theta}$. For each fixed $x \in \mathcal{X}$, it is clear that $h \mapsto (\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})[h](x)$ is a continuous linear functional on $\mathcal{H}$. Then the term $(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,x)$ appearing in the final part of Assumption 8.2 takes the explicit form

$$(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,x) = \left(\partial^r \nabla_x \log p(x|\boldsymbol{\theta})\right) \cdot K(x,x) \left(\partial^r \nabla_x \log p(x|\boldsymbol{\theta})\right). \tag{C.37}$$

The regularity of (C.37) therefore depends on K and $\mathbb{P}_{\boldsymbol{\theta}}$. See Appendix C.3.7, where (C.37) is computed for an exponential family model.

C.3.4 Proof of Theorem 8.1 (Posterior Consistency)

The following preliminary lemma is required, which takes inspiration from Alquier et al. (2016); **?**. Let $f_n(\boldsymbol{\theta}) = \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n)$ and $f(\boldsymbol{\theta}) = \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P})$.

Lemma C.3. Suppose Assumption 8.3 and Assumption 8.4 hold. For all $\delta \in (0, 1]$, with probability at least $1 - \delta$,

$$\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \pi_n^D(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq f(\boldsymbol{\theta}_*) + \left(\alpha_1 + \alpha_2 + \frac{8 \sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \sigma(\boldsymbol{\theta})}{\delta}\right) \frac{1}{\sqrt{n}}. \tag{C.38}$$

where the probability is taken with respect to realisations of the dataset $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

Proof. From the auxiliary result Theorem C.2 in Appendix C.4, we have a concen-

tration inequality

$$\mathbb{P}\left(|f_n(\boldsymbol{\theta}) - f(\boldsymbol{\theta})| \geq \delta\right) \leq \frac{4\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}} \leq \frac{4\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}} \tag{C.39}$$

for each $\boldsymbol{\theta} \in \boldsymbol{\theta}$, where the probability is taken with respect to the samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathbb{P}$. Taking the complement and re-scaling δ, (C.39) is equivalent to

$$\mathbb{P}\left(|f_n(\boldsymbol{\theta}) - f(\boldsymbol{\theta})| \leq \frac{4\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}}\right) \geq 1 - \delta. \tag{C.40}$$

Notice that by virtue of the absolute value, the following inequalities hold simultaneously with probability at least $1 - \delta$:

$$f(\boldsymbol{\theta}) \leq f_n(\boldsymbol{\theta}) + \frac{4\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}}. \tag{C.41}$$

$$f_n(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}) + \frac{4\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}}. \tag{C.42}$$

Taking an expectation with respect to the generalised posterior on both side of (C.41) yields, with probability at least $1 - \delta$,

$$\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\pi_n^D(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \leq \int_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\pi_n^D(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \frac{4\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}} \tag{C.43}$$

In order to apply the identity of Knoblauch et al. (2019, Theorem 1), we add the term $(1/n)\,\mathrm{KL}(\pi_n^D\|\pi) \geq 0$ in the right hand side and see that, with probability at least $1 - \delta$,

$$\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\pi_n^D(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \leq \frac{1}{n}\left\{\int_{\boldsymbol{\theta}} n f_n(\boldsymbol{\theta})\pi_n^D(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \mathrm{KL}(\pi_n^D\|\pi)\right\} + \frac{4\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}} \tag{C.44}$$

Clearly, the bracketed term on the right hand side is the solution to the following variational problem over $\mathcal{P}(\boldsymbol{\theta})$:

$$\begin{aligned}\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\pi_n^D(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} &\leq \frac{1}{n}\inf_{\rho\in\mathcal{P}(\boldsymbol{\theta})}\left\{\int_{\boldsymbol{\theta}} n f_n(\boldsymbol{\theta})\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \mathrm{KL}(\rho\|\pi)\right\} + \frac{4\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}} \\ &= \inf_{\rho\in\mathcal{P}(\boldsymbol{\theta})}\left\{\int_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \frac{1}{n}\mathrm{KL}(\rho\|\pi)\right\} + \frac{4\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}}\end{aligned} \tag{C.45}$$

Plugging (C.42) in (C.45), we have with probability at least $1 - \delta$,

$$\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\pi_n^D(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \leq \inf_{\rho\in\mathcal{P}(\boldsymbol{\theta})}\left\{\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + \frac{1}{n}\mathrm{KL}(\rho\|\pi)\right\} + \frac{8\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}} \tag{C.46}$$

Plugging the trivial bound $f(\boldsymbol{\theta}) \leq f(\boldsymbol{\theta}_*) + |f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_*)|$ into (C.46), we have

$$(\text{C.46}) \leq f(\boldsymbol{\theta}_*) + \inf_{\rho \in \mathcal{P}(\boldsymbol{\theta})} \left\{ \int_{\boldsymbol{\theta}} |f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_*)| \, \rho(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} + \frac{1}{n} \mathrm{KL}(\rho \| \pi) \right\} + \frac{8 \sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \sigma(\boldsymbol{\theta})}{\delta \sqrt{n}} \quad \text{(C.47)}$$

Notice that the infimum term can be upper bounded by any choice of $\rho \in \mathcal{P}(\boldsymbol{\theta})$. Letting $\Pi(B_n) := \int_{B_n} \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$, we take $\rho(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) / \Pi(B_n)$ for $\boldsymbol{\theta} \in B_n$ and $\rho(\boldsymbol{\theta}) = 0$ for $\boldsymbol{\theta} \notin B_n$. Then Assumption 8.4 part (2) ensures that $\int_{B_n} |f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_*)| \rho(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq \alpha_1 / \sqrt{n}$ and that $\mathrm{KL}(\rho \| \pi) = \int_{\boldsymbol{\theta}} \log\left(\rho(\boldsymbol{\theta}) / \pi(\boldsymbol{\theta})\right) \rho(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \int_{B_n} -\log(\Pi(B_n)) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} / \Pi(B_n) = -\log \Pi(B_n) \leq \alpha_2 \sqrt{n}$. Thus

$$\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \pi_n^D(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq f(\boldsymbol{\theta}_*) + \left(\alpha_1 + \alpha_2 + \frac{8 \sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \sigma(\boldsymbol{\theta})}{\delta} \right) \frac{1}{\sqrt{n}}, \quad \text{(C.48)}$$

with probability at least $1 - \delta$, as claimed. □

Now we turn to the proof of Theorem 8.1:

Proof of Theorem 8.1. Since $\boldsymbol{\theta}_*$ uniquely minimise f,

$$f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_*) \geq 0, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\theta} \qquad \Longrightarrow \qquad \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \pi_n^D(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - f(\boldsymbol{\theta}_*) \geq 0. \quad \text{(C.49)}$$

Thus, from Lemma C.3,

$$\mathbb{P}\left(\left| \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \pi_n^D(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - f(\boldsymbol{\theta}_*) \right| \leq \left(\alpha_1 + \alpha_2 + \frac{8 \sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \sigma(\boldsymbol{\theta})}{\delta} \right) \frac{1}{\sqrt{n}} \right) \geq 1 - \delta \quad \text{(C.50)}$$

Applying the simplifying upper bound

$$\alpha_1 + \alpha_2 + \frac{8 \sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \sigma(\boldsymbol{\theta})}{\delta} \leq \frac{\alpha_1 + \alpha_2 + 8 \sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \sigma(\boldsymbol{\theta})}{\delta},$$

taking complement of the probability and performing a change of variables, we obtain the stated result. □

C.3.5 Proof of Theorem 8.2 (Bernstein–von–Mises)

In this section we define the notation $f_n(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_n)$ and $f(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P})$. Similarly, denote $H_n := \nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_n)$ and $H_* = \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_*)$. Our aim is to verify the conditions of Theorem 4 in Jeffrey W. Miller (2021). The following technical lemma lists and establishes the conditions that are required:

Lemma C.4. Suppose that Assumption 8.2 ($r_{\max} = 3$), Assumption 8.3, and part (1) of Assumption 8.4 hold. Assume that H is nonsingular. Then the following statements almost surely hold:

1. the prior density π is continuous at $\boldsymbol{\theta}_*$ and $\pi(\boldsymbol{\theta}_*) > 0$,
2. $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_*$,
3. the Taylor expansion $f_n(\boldsymbol{\theta}) = f_n(\boldsymbol{\theta}_n) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_n) \cdot H_n(\boldsymbol{\theta} - \boldsymbol{\theta}_n) + r_n(\boldsymbol{\theta} - \boldsymbol{\theta}_n)$ holds on $\boldsymbol{\Theta}$, where the remainder r_n satisfies $|r_n(\vartheta)| \leq C\|\vartheta\|_2^3$ for all $\|\vartheta\|_2 \leq \epsilon$, all sufficiently large n and some C and $\epsilon > 0$,
4. $H_n \to H_*$, where H_n is symmetric and H_* is positive definite,
5. $\liminf_{n\to\infty}(\inf_{\{\boldsymbol{\theta}\in\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|_2\geq\epsilon\}} f_n(\boldsymbol{\theta}) - f_n(\boldsymbol{\theta}_n)) > 0$ for any $\epsilon > 0$.

Proof. We sequentially prove each statement in the list.

Part (1): Directly assumed in Assumption 8.4 part (1).

Part (2): Assumption 8.2 ($r_{\max} = 3$) and 8.3 are sufficient for Lemma 8.3 and hence $\boldsymbol{\theta}_n \overset{a.s.}{\to} \boldsymbol{\theta}_*$.

Part (3): From Assumption 8.2 ($r_{\max} = 3$), for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$ the map $\boldsymbol{\theta} \mapsto \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h](x)$ is three times continuously differentiable, meaning that f_n is three times continuously differentiable on $\boldsymbol{\Theta}$. Hence a second order Taylor expansion gives

$$f_n(\boldsymbol{\theta}) = f_n(\boldsymbol{\theta}_n) + \nabla f_n(\boldsymbol{\theta}_n)(\boldsymbol{\theta} - \boldsymbol{\theta}_n) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_n) \cdot H_n(\boldsymbol{\theta} - \boldsymbol{\theta}_n) + r_n(\boldsymbol{\theta} - \boldsymbol{\theta}_n), \tag{C.51}$$

where, for all sufficiently large n, $\nabla f_n(\boldsymbol{\theta}_n) = 0$ was assumed and the mean value form of the remainder term r_n in the Taylor expansion provides a bound

$$|r_n(\boldsymbol{\theta} - \boldsymbol{\theta}_n)| \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\nabla^3 f_n(\boldsymbol{\theta})\|_2 \|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|_2^3. \tag{C.52}$$

Finally, $\limsup_{n\to\infty} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\nabla^3 f_n(\boldsymbol{\theta})\|_2 < \infty$ a.s. by the auxiliary Lemma C.9 in Appendix C.4.

Part (4): H_n is symmetric since the assumed regularity of f_n allows the mixed second order partial derivatives of f_n to be interchanged. The auxiliary Lemma C.10 in Appendix C.4 establishes that $H_n \overset{a.s.}{\to} H_*$ where H_* is positive semi-definite. Thus, since we assumed H_* is nonsingular, it follows that H_* is positive definite.

Part (5): The inequality $\liminf_{n\to\infty}(a_n + b_n) \geq \liminf_{n\to\infty} a_n + \liminf_{n\to\infty} b_n$ holds for any sequences of $a_n, b_n \in \mathbb{R}$. Combining the property $\liminf_{n\to\infty}(-b_n) =$

$-\limsup_{n\to\infty} b_n$, we have that $\liminf_{n\to\infty}(a_n - b_n) \geq \liminf_{n\to\infty} a_n - \limsup_{n\to\infty} b_n$. Applying this inequality,

$$\liminf_{n\to\infty}\left(\inf_{\{\boldsymbol{\theta}\in\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|_2\geq\epsilon\}} f_n(\boldsymbol{\theta}) - f_n(\boldsymbol{\theta}_n)\right) \geq \liminf_{n\to\infty} \inf_{\{\boldsymbol{\theta}\in\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|_2\geq\epsilon\}} f_n(\boldsymbol{\theta}) - \limsup_{n\to\infty} f_n(\boldsymbol{\theta}_n) \quad \text{(C.53)}$$

Since $f_n(\cdot) \overset{a.s.}{\to} f(\cdot)$ uniformly on $\boldsymbol{\Theta}$ by Theorem 8.2 and $\boldsymbol{\theta}_n \overset{a.s.}{\to} \boldsymbol{\theta}_*$ by Lemma 8.3,

$$(*) \overset{a.s.}{=} \inf_{\{\boldsymbol{\theta}\in\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_*\|_2\geq\epsilon\}} f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_*) > 0 \quad \text{(C.54)}$$

where the last inequality follows from Assumption 8.3. □

Now we turn to the main proof:

Proof of Theorem 8.2. Our aim is to verify the conditions of Theorem 4 in Jeffrey W. Miller (2021). Note that this result in Jeffrey W. Miller (2021) views $\{f_n\}_{n=1}^{\infty}$ as a deterministic sequence; we therefore aim to show that the conditions of Theorem 4 in Jeffrey W. Miller (2021) are a.s. satisfied by our random sequence $\{f_n\}_{n=1}^{\infty}$.

Recall that the generalised posterior has p.d.f. $\pi_n^D(\boldsymbol{\theta}) \propto \exp\left(-nf_n(\boldsymbol{\theta})\right)\pi(\boldsymbol{\theta})$ defined on $\boldsymbol{\theta} \subset \mathbb{R}^p$. This p.d.f. can be trivially extended to a p.d.f. on $\mathbb{R}^p$ by defining $\pi(\boldsymbol{\theta}) = 0$ and (e.g.) $f_n(\boldsymbol{\theta}) = \inf_{\boldsymbol{\theta}\in\boldsymbol{\theta}} f_n(\boldsymbol{\theta}) + 1$ for all $\boldsymbol{\theta} \in \mathbb{R}^p \setminus \boldsymbol{\theta}$. This brings us into the setting of Jeffrey W. Miller (2021). The assumptions of Jeffrey W. Miller (2021, Theorem 4) are precisely the list in the statement of Lemma C.4, and the conclusion is that

$$\int_{\mathbb{R}^p} \left|\hat{\pi}_n^D(\boldsymbol{\theta}) - \frac{1}{\det(2\pi H_*^{-1})^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}\cdot H_*\boldsymbol{\theta}\right)\right| \mathrm{d}\boldsymbol{\theta} \to 0. \quad \text{(C.55)}$$

Thus, since from Lemma C.4 the assumptions of Jeffrey W. Miller (2021, Theorem 4) are a.s. satisfied, the conclusion in (C.55) a.s. holds, as claimed. □

C.3.6 Proof of Robustness Results

Proof of Lemma 8.5

Proof. First of all, (17) of Ghosh and Basu (2016) demonstrates that

$$\mathrm{PIF}(y, \boldsymbol{\theta}, \mathbb{P}_n) = \beta n \pi_n^L(\boldsymbol{\theta}) \left(-\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n) + \int_{\boldsymbol{\theta}} \mathrm{D}\, L(y, \boldsymbol{\theta}', \mathbb{P}_n)\pi_n^L(\boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}'\right). \quad \text{(C.56)}$$

By Jensen's inequality, we have an upper bounded

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \sup_{y\in\mathcal{X}} |\operatorname{PIF}(y, \boldsymbol{\theta}, \mathbb{P}_n)| \\
&\leq \beta n \sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \pi_n^L(\boldsymbol{\theta}) \left(\sup_{y\in\mathcal{X}} |\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)| + \int_{\boldsymbol{\theta}} \sup_{y\in\mathcal{X}} \left|\mathrm{D}\, L(y, \boldsymbol{\theta}', \mathbb{P}_n)\right| \pi_n^L(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}' \right). \quad \text{(C.57)}
\end{aligned}$$

Recall that $\pi_n^L(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \exp(-\beta n L(\boldsymbol{\theta}; \mathbb{P}_n))/Z$ where $0 < Z < \infty$ is the normalizing constant. Thus we can obtain the bound $\pi_n^L(\boldsymbol{\theta}) \leq \pi(\boldsymbol{\theta}) \exp(-\beta n \inf_{\boldsymbol{\theta}\in\boldsymbol{\theta}} L_n(\boldsymbol{\theta}; \mathbb{P}_n))/Z =: C\pi(\boldsymbol{\theta})$ for some constant $0 < C < \infty$, since $L_n(\boldsymbol{\theta}; \mathbb{P}_n)$ is lower bounded by assumption and n is fixed. From this upper bound, we have

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \sup_{y\in\mathcal{X}} |\operatorname{PIF}(y, \boldsymbol{\theta}, \mathbb{P}_n)| \\
&\leq \beta n C \sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) \left(\sup_{y\in\mathcal{X}} |\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)| + C \int_{\boldsymbol{\theta}} \sup_{y\in\mathcal{X}} \left|\mathrm{D}\, L(y, \boldsymbol{\theta}', \mathbb{P}_n)\right| \pi(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}' \right) \\
&\leq \beta n C \sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \left(\pi(\boldsymbol{\theta}) \sup_{y\in\mathcal{X}} |\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)| \right) + \\
&\qquad\qquad \beta n C^2 \left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) \right) \int_{\boldsymbol{\theta}} \sup_{y\in\mathcal{X}} \left|\mathrm{D}\, L(y, \boldsymbol{\theta}', \mathbb{P}_n)\right| \pi(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'.
\end{aligned}$$

Since $\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) < \infty$ by assumption in the statement of Lemma 8.5, it follows that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \left(\pi(\boldsymbol{\theta}) \sup_{y\in\mathcal{X}} |\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)| \right) < \infty \quad \text{and} \quad \int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) \sup_{y\in\mathcal{X}} |\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)| \, \mathrm{d}\boldsymbol{\theta} < \infty$$

are sufficient conditions for $\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \sup_{y\in\mathcal{X}} |\operatorname{PIF}(y, \boldsymbol{\theta}, \mathbb{P}_n)| < \infty$, as claimed. □

The Form of $\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)$ for KSD

The following lemma clarifies the form of $\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)$ for KSD:

Lemma C.5. For $L(\boldsymbol{\theta}; \mathbb{P}_{n,\epsilon,y}) = \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_{n,\epsilon,y})$, we have

$$\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n) = 2\mathbb{E}_{X\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, y)\right] - 2\mathbb{E}_{X,X'\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X')\right]. \quad \text{(C.58)}$$

Proof. From the definition of the ϵ-contamination model as a mixture model, and

using the symmetry of K, we have

$$\begin{aligned}
&\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_{n,\epsilon,y}) \\
&= \mathbb{E}_{X,X' \sim \mathbb{P}_{n,\epsilon,y}} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X') \right] \\
&= (1-\epsilon)^2 \mathbb{E}_{X,X' \sim \mathbb{P}_n} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X') \right] + 2(1-\epsilon)\epsilon \mathbb{E}_{X \sim \mathbb{P}_n} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, y) \right] \\
&\qquad + \epsilon^2 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(y, y). \qquad \text{(C.59)}
\end{aligned}$$

Direct differentiation then yields

$$\begin{aligned}
\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n) &= \frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}_{n,\epsilon,y}) \Big|_{\epsilon=0} \\
&= 2\mathbb{E}_{X \sim \mathbb{P}_n} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, y) \right] - 2\mathbb{E}_{X,X' \sim \mathbb{P}_n} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X') \right] \qquad \text{(C.60)}
\end{aligned}$$

as claimed. □

Proof of Theorem 8.3

Proof. From Lemma 8.5 with $\mathcal{X} = \mathbb{R}^d$, it is sufficient to show that

$$\text{(i)} \ \sup_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \left(\pi(\boldsymbol{\theta}) \sup_{y \in \mathbb{R}^d} |\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)| \right) < \infty \quad \text{and} \quad \text{(ii)} \ \int_{\boldsymbol{\theta}} \sup_{y \in \mathbb{R}^d} |\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)| \, \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty.$$

To establish (i) and (ii) we exploit the expression for $\mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n)$ in Lemma C.5. This gives us the bound

$$\left| \mathrm{D}\, L(y, \boldsymbol{\theta}, \mathbb{P}_n) \right| \leq 2\mathbb{E}_{X \sim \mathbb{P}_n} \big[\underbrace{\left| \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, y) \right|}_{=:(*_1)} \big] + 2 \underbrace{\mathbb{E}_{X,X' \sim \mathbb{P}_n} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X') \right]}_{=:(*_2)} \qquad \text{(C.61)}$$

From Lemma C.1,

$$\begin{aligned}
(*_1) &\leq \sqrt{\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(y, y)} \sqrt{\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X)}; \\
(*_2) &\leq \mathbb{E}_{X \sim \mathbb{P}_n} [\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X)].
\end{aligned}$$

Plugging these bounds into (C.61) and using Jensen's inequality gives

$$\begin{aligned}
&\text{(C.61)} \\
&\leq 2\sqrt{\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(y, y)} \sqrt{\mathbb{E}_{X \sim \mathbb{P}_n} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X) \right]} + 2\mathbb{E}_{X \sim \mathbb{P}_n} \left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X) \right] \qquad \text{(C.62)}
\end{aligned}$$

Now, observing that

$$\mathbb{E}_{X\sim\mathbb{P}_n}\left[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(X,X)\right] \leq \mathbb{E}_{X\sim\mathbb{P}_n}\left[\sup_{y\in\mathbb{R}^d}\left(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y)\right)\right]$$
$$= \sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y) \tag{C.63}$$

and taking a supremum over y in (C.62), we obtain the bound

$$\sup_{y\in\mathbb{R}^d}|\mathrm{D}\,L(y,\boldsymbol{\theta},\mathbb{P}_n)| \leq 4\sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y). \tag{C.64}$$

Therefore, from (C.64), it suffices to verify the conditions

$$\text{(I)}\ \sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\left(\pi(\boldsymbol{\theta})\sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y)\right)<\infty \quad \text{and}$$
$$\text{(II)}\ \int_{\boldsymbol{\theta}}\sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}<\infty,$$

which imply the original conditions (i) and (ii). To this end, in the remainder we (a) exploit the specific form of $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$ to derive the an explicit upper bound on $\sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y)$, then (b) verify the conditions (I) and (II) based on this upper bound.

Part (a): By the reproducing property of K, the definition of the diffusion Stein operator $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$, and the fact $(a_1+a_2)^2\leq 2(a_1^2+a_2^2)$ for $a_1,a_2\in\mathbb{R}$, we have the bound

$$\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y) = \|\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,\cdot)\|_{\mathcal{H}}^2$$
$$= \|\nabla_y\log p(y|\boldsymbol{\theta})\cdot K(y,\cdot)+\nabla_y\cdot K(y,\cdot)\|_{\mathcal{H}}^2 \tag{C.65}$$
$$\leq 2\|\nabla_y\log p(y|\boldsymbol{\theta})\cdot K(y,\cdot)\|_{\mathcal{H}}^2+2\|\nabla_y\cdot K(y,\cdot)\|_{\mathcal{H}}^2. \tag{C.66}$$

For the first term, the reproducing property of K gives that

$$\|\nabla_y\log p(y|\boldsymbol{\theta})\cdot K(y,\cdot)\|_{\mathcal{H}}^2 = \nabla_y\log p(y|\boldsymbol{\theta})\cdot K(y,y)\nabla_y\log p(y|\boldsymbol{\theta}), \tag{C.67}$$

while for the second term, and letting $R(x,x'):=\nabla_x\cdot(\nabla_{x'}\cdot K(x,x'))$, the reproducing property gives that

$$\|\nabla_y K(y,\cdot)\|_{\mathcal{H}}^2 = \left\langle\nabla_y\cdot K(y,\cdot),\nabla_y\cdot K(y,\cdot)\right\rangle_{\mathcal{H}} = R(y,y). \tag{C.68}$$

Thus, taking the supremum with respect to $y \in \mathbb{R}^d$ yields the upper bound,

$$\sup_{y\in\mathbb{R}^d} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y) \leq 2\sup_{y\in\mathbb{R}^d}\left(\nabla_y \log p(y|\boldsymbol{\theta})\cdot K(y,y)\nabla_y \log p(y|\boldsymbol{\theta})\right) + 2\sup_{y\in\mathbb{R}^d} R(y,y).$$

Since $K \in C_b^{1\times 1}(\mathbb{R}^d \times \mathbb{R}^d)$ by assumption, it follows that $C_{MK} := \sup_{y\in\mathbb{R}^d} R(y,y) < \infty$. Thus we have have that

$$\sup_{y\in\mathbb{R}^d} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y) \leq 2\gamma(\boldsymbol{\theta}) + 2C_{MK}, \tag{C.69}$$

where $\gamma(\boldsymbol{\theta})$ was defined in the statement of Theorem 8.3.

Part (b): Now we are in a position to verify conditions (I) and (II). For condition (I), we use (C.69) to obtain

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\left(\pi(\boldsymbol{\theta})\sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y)\right) \leq 2\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\pi(\boldsymbol{\theta})\gamma(\boldsymbol{\theta}) + 2C_{MK}\sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\pi(\boldsymbol{\theta}) \tag{C.70}$$

which is finite by assumption. Similarly, for condition (II), we use (C.69) to obtain

$$\int_{\boldsymbol{\theta}}\sup_{y\in\mathbb{R}^d}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(y,y)\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \leq 2\int_{\boldsymbol{\theta}}\pi(\boldsymbol{\theta})\gamma(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} + 2C_{MK}\int_{\boldsymbol{\theta}}\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \tag{C.71}$$

which is also finite by assumption. This completes the proof. □

C.3.7 Verifying Assumptions 8.3 8.2, 8.4

In this appendix we demonstrate how Assumptions 8.3 8.2, 8.4 can be verified for the exponential family model when the Langevin Stein operator is employed. For simplicity, consider the case where the data dimension is $d = 1$, the parameter dimension is $p = 1$, and the conjugate prior $\pi(\boldsymbol{\theta}) \propto \exp(-\boldsymbol{\theta}^2/2)$ is used. From (8.4), a canonical exponential family model with $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $\mathcal{X} = \mathbb{R}$ is given by

$$p(x|\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}\cdot t(x) - a(\boldsymbol{\theta}) + b(x)) \tag{C.72}$$

where $t : \mathbb{R} \to \mathbb{R}$, $a : \boldsymbol{\theta} \to \mathbb{R}$ and $b : \mathbb{R} \to \mathbb{R}$. Accordingly, the log derivative is given by $\nabla \log p(x|\boldsymbol{\theta}) = \nabla t(x)\boldsymbol{\theta} + \nabla b(x)$. Identical calculations to Proposition 8.1 show that the KSD of the exponential family model with the Langevin Stein operator takes a quadratic form

$$\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n) = C_{1,n}\boldsymbol{\theta}^2 + C_{2,n}\boldsymbol{\theta} + C_{3,n} \quad \text{and} \quad \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}) = C_1\boldsymbol{\theta}^2 + C_2\boldsymbol{\theta} + C_3.$$

where $C_{i,n} = (1/n^2)\sum_{i,j=1}^n c_i(x_i, x_j)$ and $C_i = \mathbb{E}_{X,X'\sim\mathbb{P}}[c_i(X, X')]$ and

$$
\begin{aligned}
c_1(x, x') &:= \nabla t(x) \cdot K(x, x')\nabla t(x') \\
c_2(x, x') &:= \nabla t(x) \cdot \big(\nabla_{x'} \cdot K(x, x')\big) + \nabla t(x') \cdot \big(\nabla_x \cdot K(x, x')\big) + 2\nabla t(x) \cdot K(x, x')\nabla b(x') \\
c_3(x, x') &:= b(x) \cdot K(x, x')b(x') + \nabla_x \cdot (\nabla_{x'} \cdot K(x, x')) \\
&\quad + b(x) \cdot (\nabla_{x'} \cdot K(x, x')) + b(x') \cdot (\nabla_x \cdot K(x, x')).
\end{aligned}
$$

Note that $C_{1,n} > 0$ and $C_1 > 0$ if a positive definite kernel K is used.

Verifying Assumption 8.2 ($r_{\max} = 3$): First, note that $H_* = \nabla^2_{\boldsymbol{\theta}} \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*}$ is non-singular since $\nabla_{\boldsymbol{\theta}} \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}) = \nabla^2_{\boldsymbol{\theta}}(C_1\boldsymbol{\theta}^2 + C_2\boldsymbol{\theta} + C_3) = 2C_1 > 0$. Now, as demonstrated in Section 8.4.1, when $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}$ is the Langevin Stein operator, we have $(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})[h](x) = (\partial^r \nabla_x \log p(x|\boldsymbol{\theta})) \cdot h(x)$ and $h \mapsto (\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})[h](x)$ is a continuous linear functional on $\mathcal{H}$ for each fixed $x \in \mathcal{X}$. In the exponential family case, the map $\boldsymbol{\theta} \mapsto \nabla_x \log p(x|\boldsymbol{\theta})$ is infinitely differentiable over $\boldsymbol{\Theta}$ since it is polynomial, leading to

$$
\partial^1 \nabla_x \log p(x|\boldsymbol{\theta}) = \nabla t(x), \qquad \partial^2 \nabla_x \log p(x|\boldsymbol{\theta}) = 0, \qquad \partial^3 \nabla_x \log p(x|\boldsymbol{\theta}) = 0.
$$

It is then clear that $\mathbb{E}_{X\sim\mathbb{P}}[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}((\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^r \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(X, X))] < \infty$ for $r = 2, 3$. For $r = 1$,

$$
\mathbb{E}_{X\sim\mathbb{P}}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left((\partial^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(X, X)\right)\right] = \mathbb{E}_{X\sim\mathbb{P}}\left[\nabla t(X) \cdot K(X, X)\nabla t(X)\right]. \tag{C.73}
$$

For the remaining term in Assumption 8.2, by essentially same calculations as Proposition 8.1,

$$
\begin{aligned}
&\mathbb{E}_{X\sim\mathbb{P}}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(X, X)\right)\right] \\
&= \mathbb{E}_{X\sim\mathbb{P}}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(c_1(X, X)\boldsymbol{\theta}^2 + c_2(X, X)\boldsymbol{\theta} + c_3(X, X)\right)\right] \\
&\leq \mathbb{E}_{X\sim\mathbb{P}}\left[c_1(X, X) \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \boldsymbol{\theta}^2 + c_2(X, X) \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \boldsymbol{\theta} + c_3(X, X)\right]
\end{aligned} \tag{C.74}
$$

Since $\boldsymbol{\theta}$ is a bounded set in $\mathbb{R}$, it is clear that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \boldsymbol{\theta} < \infty$. The finiteness of (C.73) and (C.74) can therefore be interpreted as finite moment conditions involving t, b, K and $\mathbb{P}$.

Verifying Assumption 8.3: If both $\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n)$ and $\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})$ are of quadratic form with $C_{1,n} > 0$ and $C_1 > 0$, the estimator $\boldsymbol{\theta}_n$ exists and the minimiser $\boldsymbol{\theta}_*$

is unique over $\mathbb{R}$. It depends on C_1, C_2, C_3 whether $\boldsymbol{\theta}_*$ is contained in $\boldsymbol{\theta}$, but if we are free to select $\boldsymbol{\theta}$ then we may select it such that $\boldsymbol{\theta}^* \in \boldsymbol{\theta}$. Since $C_1 > 0$, the *well-separated* property of $\boldsymbol{\theta}_*$ is automatically satisfied; i.e. $\mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}_*}, \mathbb{P}) < \inf_{\{\boldsymbol{\theta} \in \boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \geq \epsilon\}} \mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P})$ for all $\epsilon > 0$.

Verifying Assumption 8.4: Part (1) is immediately satisfied since the prior density $\pi(\boldsymbol{\theta}) \propto \exp(-\boldsymbol{\theta}^2/2)$ is continuous and positive on $\boldsymbol{\theta}$. For part (2), we first have

$$\begin{aligned} |\,\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}) - \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}_*} \| \mathbb{P})| &= |C_1 \boldsymbol{\theta}^2 + C_2 \boldsymbol{\theta} - C_1 \boldsymbol{\theta}_*^2 - C_2 \boldsymbol{\theta}_*| \\ &= C_1 \left| (\boldsymbol{\theta} + Z_2)^2 - Z_1 \right| \end{aligned}$$

where $Z_1 := C_2^2/(4C_1^2) + \boldsymbol{\theta}_*^2 + (C_2/C_1)\boldsymbol{\theta}_*$ and $Z_2 := C_2/(2C_1)$ by completing the square. By the simple calculation, the set $B_n(\alpha_1) = \{\boldsymbol{\theta} \in \boldsymbol{\theta} : |\,\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}} \| \mathbb{P}) - \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}_*} \| \mathbb{P})| \leq \alpha_1/\sqrt{n}\}$ is then given by

$$B_n(\alpha_1) = \left\{ \boldsymbol{\theta} \in \boldsymbol{\theta} : -\left(\frac{\alpha_1}{C_1\sqrt{n}} + Z_1 \right)^{\frac{1}{2}} - Z_2 \leq \boldsymbol{\theta} \leq \left(\frac{\alpha_1}{C_1\sqrt{n}} + Z_1 \right)^{\frac{1}{2}} - Z_2 \right\}$$

While it is difficult to derive an explicit inequality between $\Pi(B_n)$ and $\exp(-\alpha_2\sqrt{n})$, since it requires division into cases according to the values of C_1, C_2, C_3, $\boldsymbol{\theta}_*$, and the set $\boldsymbol{\theta}$, the explicit form of B_n renders it straightforward to numerically determine which values for $\alpha_1 > 0$ and $\alpha_2 > 0$ ensure that $\Pi(B_n) \geq \exp(-\alpha_2\sqrt{n})$ holds for all $n \in \mathbb{N}$.

Quantities $S_n(x, \boldsymbol{\theta})$ and J_n: Here we provide the explicit form of $S_n(x, \boldsymbol{\theta})$ and J_n used to determine the value of β for exponential family model. From the definition,

$$\begin{aligned} S_n(x, \boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \left(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x_i) \right) \\ &= 2 \left(\frac{1}{n} \sum_{i=1}^{n} c_1(x, x_i) \right) \boldsymbol{\theta} + \left(\frac{1}{n} \sum_{i=1}^{n} c_2(x, x_i) \right). \end{aligned} \tag{C.75}$$

Let $c_{1,n}(x) := (1/n) \sum_{i=1}^{n} c_1(x, x_i)$ and $c_{2,n}(x) := (1/n) \sum_{i=1}^{n} c_2(x, x_i)$. From the definition,

$$\begin{aligned} J_n &= \frac{1}{n} \sum_{i=1}^{n} S_n(x_i, \boldsymbol{\theta}_n) S_n(x_i, \boldsymbol{\theta}_n)^\top \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(2c_{1,n}(x_i)\boldsymbol{\theta}_n + c_{2,n}(x_i) \right) \left(2c_{1,n}(x_i)\boldsymbol{\theta}_n + c_{2,n}(x_i) \right)^\top. \end{aligned}$$

Together with $H_n = C_{1,n}$, the default choice of β is given by (8.14) in Section 8.5.

C.4 Auxiliary Theoretical Results for KSD-Bayes

In Appendix C.3 we exploited a number of auxiliary results, the details of which are now provided. Recall that Standing Assumptions 1 and 2 continue to hold throughout.

C.4.1 Derivative Bounds

Our auxiliary results mainly concern moments of derivative quantities, and the aim of Appendix C.4.1 is to establish the main bounds that will be used. Recall that ∂^1, ∂^2 and ∂^3 denote the partial derivatives $(\partial/\partial\boldsymbol{\theta}_h)$, $(\partial^2/\partial\boldsymbol{\theta}_h\partial\boldsymbol{\theta}_k)$ and $(\partial^3/\partial\boldsymbol{\theta}_h\partial\boldsymbol{\theta}_k\partial\boldsymbol{\theta}_l)$ respectively. For the proofs in Appendix C.4.1, we make the index explicit by re-writing them as $\partial^1_{(h)}$, $\partial^2_{(h,k)}$ and $\partial^3_{(h,k,l)}$. For $x \in \mathcal{X}$ and $(h,k,l) \in \{1,\ldots,p\}^3$, we define

$$
\begin{aligned}
m^0(x) &:= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x,x)},\\
m^1(x) &:= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\sum_{h=1}^{p}(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,x)},\\
m^2(x) &:= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\sum_{h,k=1}^{p}(\partial^2_{(h,k)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^2_{(h,k)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,x)},\\
m^3(x) &:= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\sum_{h,k,l=1}^{p}(\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,x)}.
\end{aligned}
$$

where we continue to use the convention that the first and second operator in expressions such as $(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,x')$ are respectively applied to the first and second argument of K. Further define

$$
\begin{aligned}
M^1(x,x') &:= m^1(x)m^0(x') + m^0(x)m^1(x'),\\
M^2(x,x') &:= m^2(x)m^0(x') + 2m^1(x)m^1(x') + m^0(x)m^2(x'),\\
M^3(x,x') &:= m^3(x)m^0(x') + 3m^2(x)m^1(x') + 3m^1(x)m^2(x') + m^0(x)m^3(x').
\end{aligned}
$$

Based on these quantities, we now provide three technical results, Lemmas C.6, C.8 and C.7.

Lemma C.6. Suppose Assumption 8.2 ($r_{max} = 3$) holds. For each $r = 1, 2, 3$, and for any $x, x' \in \mathcal{X}$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \nabla_{\boldsymbol{\theta}}^r \left(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x') \right) \right\|_2 \leq M^r(x, x'). \tag{C.76}$$

If instead Assumption 8.2 ($r_{max} = 1$) holds, then (C.76) holds for $r = 1$.

Proof. We first derive the upper bound for $r = 1$ and then apply the same argument for the remaining upper bound for $r = 2$ and $r = 3$. By the definition of $\nabla_{\boldsymbol{\theta}}$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \nabla_{\boldsymbol{\theta}} \left(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x') \right) \right\|_2 = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sqrt{\sum_{h=1}^{p} \left(\partial_{(h)}^1 \left(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x') \right) \right)^2}. \tag{C.77}$$

By Lemma C.1 and Standing Assumption 2, we have $\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, \cdot) \in \mathcal{H}$ for any $x \in \mathcal{X}$ and

$$(*_1) := \partial_{(h)}^1 \left(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x') \right) = \partial_{(h)}^1 \left(\left\langle \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, \cdot), \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x', \cdot) \right\rangle_{\mathcal{H}} \right). \tag{C.78}$$

From Assumption 8.2 ($r_{max} = 1$), the operator $(\partial_{(h)}^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})$ exists over $\boldsymbol{\Theta}$ and satisfies the preconditions of Lemma C.1. Hence, by setting $\mathcal{S}_{\mathbb{Q}} = (\partial_{(h)}^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})$ in Lemma C.1, we have that $(\partial_{(h)}^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}) K(x, \cdot) \in \mathcal{H}$ for each $x \in \mathcal{X}$. Let $f_{\boldsymbol{\theta}}(\cdot) = \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, \cdot)$ and $g_{\boldsymbol{\theta}}(\cdot) = \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x', \cdot)$. Then the following product rule holds:

$$\partial_{(h)}^1 \langle f_{\boldsymbol{\theta}}, g_{\boldsymbol{\theta}} \rangle_{\mathcal{H}} = \langle \partial_{(h)}^1 f_{\boldsymbol{\theta}}, g_{\boldsymbol{\theta}} \rangle_{\mathcal{H}} + \langle f_{\boldsymbol{\theta}}, \partial_{(h)}^1 g_{\boldsymbol{\theta}} \rangle_{\mathcal{H}}, \tag{C.79}$$

which is verified from definition of differentiation as a limit and continuity of the inner product. Note that $\partial_{(h)} f_{\boldsymbol{\theta}}(\cdot) = (\partial_{(h)}^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}) K(x, \cdot) \in \mathcal{H}$ and $\partial_{(h)} g_{\boldsymbol{\theta}}(\cdot) = (\partial_{(h)}^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}) K(x', \cdot) \in \mathcal{H}$. Therefore by (C.79) and the Cauchy–Schwarz inequality,

$$\begin{aligned}
(*_1) &= \left\langle \partial_{(h)}^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, \cdot), \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x', \cdot) \right\rangle_{\mathcal{H}} + \left\langle \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, \cdot), \partial_{(h)}^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x', \cdot) \right\rangle_{\mathcal{H}} \\
&\leq \underbrace{\left\| (\partial_{(h)}^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}) K(x, \cdot) \right\|_{\mathcal{H}}}_{(*_a)} \underbrace{\left\| \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x', \cdot) \right\|_{\mathcal{H}}}_{(*_b)} + \underbrace{\left\| \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, \cdot) \right\|_{\mathcal{H}}}_{(*_c)} \underbrace{\left\| (\partial_{(h)}^1 \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}) K(x', \cdot) \right\|_{\mathcal{H}}}_{(*_d)}.
\end{aligned}$$

For the original term (C.77), by the triangle inequality,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\sum_{h=1}^{p} (*_1)^2} \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\sum_{h=1}^{p} \Big((*_a)(*_b) + (*_c)(*_d)\Big)^2}$$
$$\leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\sum_{h=1}^{p} (*_a)^2(*_b)^2} + \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\sum_{h=1}^{p} (*_c)^2(*_d)^2}. \tag{C.80}$$

For the term $(*_a)$, expanding the norm yields that

$$(*_a)^2 = \Big\langle (\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,\cdot), (\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,\cdot)\Big\rangle_{\mathcal{H}} = (\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^1_{(h)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,x).$$

A similar argument applied to $(*_b)^2$, $(*_c)^2$ and $(*_d)^2$ leads to the overall bound

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \big\|\nabla_{\boldsymbol{\theta}}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x,x')\big)\big\|_2 \leq m^1(x)m^0(x') + m^0(x)m^1(x') = M^1(x,x').$$

The upper bounds for $r = 2$ and $r = 3$ are obtained by an analogous argument. Indeed, from the definition of $\nabla^2_{\boldsymbol{\theta}}$ and $\nabla^3_{\boldsymbol{\theta}}$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \big\|\nabla^2_{\boldsymbol{\theta}}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x,x')\big)\big\|_2 = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\sum_{h,k=1}^{p} \Big(\partial^2_{(h,k)}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x,x')\big)\Big)^2} =: (*''),$$
$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \big\|\nabla^3_{\boldsymbol{\theta}}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x,x')\big)\big\|_2 = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sqrt{\sum_{h,k,l=1}^{p} \Big(\partial^3_{(h,k,l)}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x,x')\big)\Big)^2} =: (*''').$$

From Assumption 8.2 ($r_{max} = 3$), the operators $(\partial^2_{(h,k)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})$ and $(\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})$ exist over $\boldsymbol{\Theta}$ and satisfy the preconditions of Lemma C.1. Hence from Lemma C.1, $\partial^2_{(h,k)}f_{\boldsymbol{\theta}}(\cdot) = (\partial^2_{(h,k)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,\cdot) \in \mathcal{H}$ and $\partial^3_{(h,k,l)}f_{\boldsymbol{\theta}}(\cdot) = (\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(x,\cdot) \in \mathcal{H}$ for any $x \in \mathcal{X}$, and in turn $\partial^2_{(h,k)}g_{\boldsymbol{\theta}}(\cdot) \in \mathcal{H}$ and $\partial^3_{(h,k,l)}g_{\boldsymbol{\theta}}(\cdot) \in \mathcal{H}$. Repeated application of the product rule (C.79) gives that

$$\begin{aligned}
&\partial^2_{(h,k)}\langle f_{\boldsymbol{\theta}}, g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} \\
&= \langle \partial^2_{(h,k)}f_{\boldsymbol{\theta}}, g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} + \langle \partial^1_{(h)}f_{\boldsymbol{\theta}}, \partial^1_{(k)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} + \langle \partial^1_{(k)}f_{\boldsymbol{\theta}}, \partial^1_{(h)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} + \langle f_{\boldsymbol{\theta}}, \partial^2_{(h,k)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}}, \\
&\partial^3_{(h,k,l)}\langle f_{\boldsymbol{\theta}}, g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} \\
&= \langle \partial^3_{(h,k,l)}f_{\boldsymbol{\theta}}, g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} + \langle \partial^2_{(h,k)}f_{\boldsymbol{\theta}}, \partial^1_{(l)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} + \langle \partial^2_{(h,l)}f_{\boldsymbol{\theta}}, \partial^1_{(k)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} + \langle \partial^2_{(k,l)}f_{\boldsymbol{\theta}}, \partial_{(h)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} \\
&\quad + \langle \partial^1_{(h)}f_{\boldsymbol{\theta}}, \partial^2_{(k,l)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} + \langle \partial^1_{(k)}f_{\boldsymbol{\theta}}, \partial^2_{(h,l)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} + \langle \partial^1_{(l)}f_{\boldsymbol{\theta}}, \partial^2_{(h,k)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}} + \langle f_{\boldsymbol{\theta}}, \partial^3_{(h,k,l)}g_{\boldsymbol{\theta}}\rangle_{\mathcal{H}}
\end{aligned} \tag{C.81}$$

Following the same argument as the preceding upper bound for $r = 1$, the triangle

inequality and Cauchy–Schwarz imply that

$$
\begin{aligned}
(*'') &\leq m^2(x)m^0(x') + m^1(x)m^1(x') + m^1(x)m^1(x') + m^0(x)m^2(x') \\
&= m^2(x)m^0(x') + 2m^1(x)m^1(x') + m^0(x)m^2(x') = M^2(x, x'), \\
(*''') &\leq m^3(x)m^0(x') + m^2(x)m^1(x') + m^2(x)m^1(x') + m^2(x)m^1(x') \\
&\qquad + m^1(x)m^2(x') + m^1(x)m^2(x') + m^1(x)m^2(x') + m^0(x)m^3(x') \\
&= m^3(x)m^0(x') + 3m^2(x)m^1(x') + 3m^1(x)m^2(x') + m^0(x)m^3(x') = M^3(x, x'),
\end{aligned}
$$

which are the claimed upper bounds for the cases $r = 2$ and $r = 3$. □

Lemma C.7. Suppose Assumption 8.2 ($r_{max} = 3$) holds. For $r = 0, 1, 2, 3$, $\mathbb{E}_{X\sim\mathbb{P}}[|m^r(X)|] < \infty$ and $\mathbb{E}_{X\sim\mathbb{P}}[|m^r(X)|^2] < \infty$. For $r = 1, 2, 3$, $\mathbb{E}_{X,X'\sim\mathbb{P}}[|M^r(X, X')|] < \infty$ and $\mathbb{E}_{X\sim\mathbb{P}}[|M^r(X, X)|] < \infty$. If instead Assumption 8.2 ($r_{max} = 1$) holds, these results hold for $0 \leq r \leq 1$.

Proof. First, note that positivity of $m^r(\cdot)$ and $M^r(\cdot)$ implies that the absolute value signs can be neglected. Moreover, from Jensen's inequality $(\mathbb{E}_{X\sim\mathbb{P}}[m^r(X)])^2 \leq \mathbb{E}_{X\sim\mathbb{P}}[m^r(X)^2]$. Thus it is sufficient to show that (a) $\mathbb{E}_{X\sim\mathbb{P}}[m^r(X)^2] < \infty$, (b) $\mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X, X')] < \infty$ and (c) $\mathbb{E}_{X\sim\mathbb{P}}[M^r(X, X)] < \infty$.

Part (a): The argument is analogous for each $r = 0, 1, 2, 3$ and we present it with $r = 3$. The bounded follows from Jensen's inequality and the triangle inequality:

$$
\begin{aligned}
&\mathbb{E}_{X\sim\mathbb{P}}\left[m^3(X)^2\right] \\
&\leq \mathbb{E}_{X\sim\mathbb{P}}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{h,k,l=1}^{p} (\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(X, X)\right] \\
&\leq \sum_{h,k,l=1}^{p} \mathbb{E}_{X\sim\mathbb{P}}\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left((\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})(\partial^3_{(h,k,l)}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}})K(X, X)\right)\right]
\end{aligned}
$$

where the terms in the sum are finite by Assumption 8.2 ($r_{max} = 3$).

Part (b): Since X, X' are independent in the expectation $\mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X, X')]$, it is clear from the definition of M^r that $\mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X, X')]$ exists if the expectation of each term $m^s(X)$, $s \leq r$, exists. Thus by part (a), $\mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X, X')] < \infty$ for $r = 1, 2, 3$.

Part (c): From the definition of $M^r(x,x)$ for $r=1,2,3$,

$$\begin{aligned}
\mathbb{E}_{X\sim\mathbb{P}}[M^1(X,X)] &= 2\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)m^0(X)],\\
\mathbb{E}_{X\sim\mathbb{P}}[M^2(X,X)] &= 2\mathbb{E}_{X\sim\mathbb{P}}[m^2(X)m^0(X)] + 2\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)m^1(X)],\\
\mathbb{E}_{X\sim\mathbb{P}}[M^3(X,X)] &= 2\mathbb{E}_{X\sim\mathbb{P}}[m^3(X)m^0(X)] + 6\mathbb{E}_{X\sim\mathbb{P}}[m^2(X)m^1(X)].
\end{aligned}$$

Applying the Cauchy–Schwarz inequality for each term,

$$\begin{aligned}
&\mathbb{E}_{X\sim\mathbb{P}}[M^1(X,X)]\\
&\le 2\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^0(X)^2]},\\
&\mathbb{E}_{X\sim\mathbb{P}}[M^2(X,X)]\\
&\le 2\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^2(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^0(X)^2]} + 2\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)^2]},\\
&\mathbb{E}_{X\sim\mathbb{P}}[M^3(X,X)]\\
&\le 2\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^3(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^0(X)^2]} + 6\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^2(X)^2]}\sqrt{\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)^2]}.
\end{aligned}$$

Since each of the latter expectations is finite by part (a), $\mathbb{E}_{X\sim\mathbb{P}}[M^r(X,X)]<\infty$ for $r=1,2,3$.

Inspection of the proof reveals that these results hold for $r=0,1$ if instead Assumption 8.2 ($r_{max}=1$) holds. □

Lemma C.8. Suppose Assumption 8.2 ($r_{max}=3$) holds. Then, for $r=1,2,3$,

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^r(x_i,x_j) \xrightarrow{a.s} \mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X,X')] < \infty. \tag{C.82}$$

If instead Assumption 8.2 ($r_{max}=1$) holds, then (C.82) holds for $r=1$.

Proof. The proof is based on the strong law of large numbers, the sufficient conditions for which are provided by Lemma C.7, which shows that $\mathbb{E}_{X\sim\mathbb{P}}\left[|m^r(X)|\right] < \infty$ for $r=0,1,2,3$ under Assumption 8.2 ($r_{max}=3$). Then the strong law of large numbers (Durrett, 2010, Theorem 2.5.10) yields that $(1/n)\sum_{i=1}^{n} m^r(x_i) \xrightarrow{a.s.} \mathbb{E}_{X\sim\mathbb{P}}\left[m^r(X)\right] =: (*_r)$ for $r=0,1,2,3$. Then, from the definition of M^1,

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^1(x_i,x_j) = \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big(m^1(x_i)m^0(x_j) + m^0(x_i)m^1(x_j)\Big)\\
&= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m^1(x_i) \times \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} m^0(x_j) + \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m^0(x_i) \times \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} m^1(x_j).
\end{aligned}$$

Since each limit in the right hand side converges a.s. to either $(*_0)$ or $(*_1)$, so that

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^1(x_i,x_j)\\
\xrightarrow{a.s.}\ &\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)]\times\mathbb{E}_{X\sim\mathbb{P}}[m^0(X)]+\mathbb{E}_{X\sim\mathbb{P}}[m^0(X)]\times\mathbb{E}_{X\sim\mathbb{P}}[m^1(X)]\\
=\ &\mathbb{E}_{X,X'\sim\mathbb{P}}[m^1(X)m^0(X')+m^0(X)m^1(X')]=\mathbb{E}_{X,X'\sim\mathbb{P}}[M^1(X,X')],
\end{aligned}$$

where X, X' are independent. An analogous argument holds for $M^2(x_i,x_j)$ and $M^3(x_i,x_j)$, giving that

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^2(x_i,x_j) &\xrightarrow{a.s.} (*_2)(*_0)+2(*_1)(*_1)+(*_0)(*_2)\\
&= \mathbb{E}_{X,X'\sim\mathbb{P}}[M^2(X,X')],\\
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} M^3(x_i,x_j) &\xrightarrow{a.s.} (*_3)(*_0)+3(*_2)(*_1)+3(*_1)(*_2)+(*_0)(*_3)\\
&= \mathbb{E}_{X,X'\sim\mathbb{P}}[M^3(X,X')].
\end{aligned}$$

Inspection of the proof reveals that (C.82) still holds for $r=1$ if Assumption 8.2 ($r_{max}=1$) holds instead. □

C.4.2 Technical Lemmas

Throughout this section we let $f_n(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n)$ and $f(\boldsymbol{\theta}) := \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})$. Similarly to $\nabla^2_{\boldsymbol{\theta}}$, we let $\nabla^3_{\boldsymbol{\theta}} := \nabla_{\boldsymbol{\theta}}\otimes\nabla_{\boldsymbol{\theta}}\otimes\nabla_{\boldsymbol{\theta}}$ denote the tensor product $\otimes$ where each component is given by $\partial^3_{h,k,l}$. For a matrix $a\in\mathbb{R}^{p\times p}$ and tensor $b\in\mathbb{R}^{p\times p\times p}$, denote their Euclidean norms by $\|a\|_2$ and $\|b\|_2$.

Lemma C.9 (Derivatives a.s. Bounded)**.** Suppose Assumption 8.2 ($r_{max}=3$) holds. Then $\limsup_{n\to\infty}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\|_2<\infty$ a.s. for $r=1,2,3$. If instead Assumption 8.2 ($r_{max}=1$) holds, then the result holds for $r=1$.

Proof. First of all, for finite n we have

$$\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}) = \nabla^r_{\boldsymbol{\theta}}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x_i,x_j) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\nabla^r_{\boldsymbol{\theta}}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x_i,x_j)\big).$$

From the triangle inequality and Lemma C.6, we further have

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\|_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\big\|\nabla^r_{\boldsymbol{\theta}}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x_i,x_j)\big)\big\|_2 \le \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}M^r(x_i,x_j).$$

It follows from Lemma C.8 that $(1/n^2)\sum_{i=1}^n\sum_{j=1}^n M^r(x_i,x_j) \xrightarrow{a.s.} \mathbb{E}_{X,X'\sim\mathbb{P}}[M^r(X,X')] < \infty$. Therefore, a.s. $\limsup_{n\to\infty}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\|\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\|_2 < \infty$. Inspection of the proof reveals that the argument still holds for $r=1$ if Assumption 8.2 ($r_{max}=1$) holds instead. □

Lemma C.10 (A.S. Convergence of Derivatives)**.** Suppose Assumption 8.2 ($r_{max}=3$) and 8.3 hold. Then we have $\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_*) \overset{a.s.}{\to} \nabla^r_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_*)$ for $r=1,2,3$. Let $H_n := \nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_n)$ and $H_* := \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_*)$. We further have $H_n \overset{a.s.}{\to} H_*$, where H_n and H_* are symmetric and H_* is semi positive definite.

Proof. The proof is structured as follows: First we show (a) $\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_*) \overset{a.s.}{\to} \nabla^r_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_*)$ for $r=1,2,3$. Then we show (b) $H_n \overset{a.s.}{\to} H_*$. Finally we show (c) H_n is symmetric and H_* is semi-positive definite.

Part (a): The argument here is analogous to that used to prove Theorem 8.1, based on the decomposition

$$\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}) = \nabla^r_{\boldsymbol{\theta}} \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i,x_j) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \nabla^r_{\boldsymbol{\theta}}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i,x_j)\big).$$

Let $F(x,x') := \nabla^r_{\boldsymbol{\theta}}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x,x')\big)$ to see that

$$\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}) = \underbrace{\frac{1}{n}\frac{1}{n}\sum_{i=1}^n F(x_i,x_i)}_{(*_1)} + \underbrace{\frac{n-1}{n}\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n F(x_i,x_j)}_{(*_2)}. \tag{C.83}$$

It follows from the strong law of large number (Durrett, 2010, Theorem 2.5.10) that $(*_1) \overset{a.s.}{\to} \mathbb{E}_{X\sim\mathbb{P}}[F(X,X)]$ provided $E_{X\sim\mathbb{P}}[\|F(X,X)\|_2] < \infty$. Similarly, it follows from the strong law of large number for U-statistics (Wassily Hoeffding, 1961) that $(*_2) \overset{a.s.}{\to} \mathbb{E}_{X,X'\sim\mathbb{P}}[F(X,X')]$ provided $E_{X,X'\sim\mathbb{P}}[\|F(X,X')\|_2] < \infty$. Both the required conditions holds by Lemma C.7 and the fact that $\|F(x,x')\|_2 \le \sup_{\boldsymbol{\theta}\in\boldsymbol{\theta}}\|\nabla^r_{\boldsymbol{\theta}}(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x,x'))\|_2 \le M^r(x,x')$ from Lemma C.6. Thus

$$\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}) \xrightarrow{a.s.} \mathbb{E}_{X,X'\sim\mathbb{P}}[F(X,X')] = \mathbb{E}_{X,X'\sim\mathbb{P}}[\nabla^r_{\boldsymbol{\theta}}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i,x_j)\big)]. \tag{C.84}$$

Since $\mathbb{E}_{X,X'\sim\mathbb{P}}[\|F(X,X')\|_2] < \infty$, we may apply the dominated convergence theorem to interchange expectation and differentiation:

$$\mathbb{E}_{X,X'\sim\mathbb{P}}[\nabla^r_{\boldsymbol{\theta}}(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X,X'))] = \nabla^r_{\boldsymbol{\theta}}\mathbb{E}_{X,X'\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X,X')] = \nabla^r_{\boldsymbol{\theta}} f(\boldsymbol{\theta}). \tag{C.85}$$

Therefore, setting $\boldsymbol{\theta} = \boldsymbol{\theta}_*$, we conclude that $\nabla^r_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_*) \overset{a.s.}{\to} \nabla^r_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_*)$.

Part (b): First of all, by the triangle inequality,

$$\left\|\nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_n) - \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_*)\right\|_2 \leq \underbrace{\left\|\nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_n) - \nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_*)\right\|_2}_{(**_1)} + \underbrace{\left\|\nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_*) - \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_*)\right\|_2}_{(**_2)}.$$

By the mean value theorem applied to $(**_1)$ and Lemma C.9 (i.e. $\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\nabla^3_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\|_2 < \infty$ a.s.), there a.s. exists a constant $0 < C < \infty$ s.t., for all sufficiently large n,

$$(**_1) = \left\|\nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_n) - \nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_*)\right\|_2 \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\nabla^3_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})\|_2 \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|_2 \leq C\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|_2 \tag{C.86}$$

Then applying Lemma 8.3 (i.e. $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_*\|_2 \overset{a.s.}{\to} 0$), we have $(**_1) \overset{a.s.}{\to} 0$. Further the preceding part (a) implied that $(**_2) \overset{a.s.}{\to} 0$. Therefore, we conclude that $\nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_n) \overset{a.s.}{\to} \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_*)$.

Part (c): Since f_n is twice continuously differentiable over $\boldsymbol{\Theta}$ by assumption, commutation of two partial derivatives $\partial_{(h)}\partial_{(k)} f_n(\boldsymbol{\theta}) = \partial_{(k)}\partial_{(h)} f_n(\boldsymbol{\theta})$ holds over $\boldsymbol{\Theta}$ by the Clairaut's theorem. Therefore the (h,k)-th entry and (k,h)-th entry of $H_n = \nabla^2_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta}_n)$ are equal. An analogous argument applies to $H_* = \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_*)$, so that both H_n and H_* are symmetric. Furthermore, the Hessian H_* is semi positive definite since $\boldsymbol{\theta}_*$ is the minimiser of f from Assumption 8.3. □

Lemma C.11 (Moment Condition for Asymptotic Normality)**.** Suppose that Assumption 8.2 ($r_{max} = 3$) holds. Let $F(x, x') := \nabla_{\boldsymbol{\theta}}(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x'))$ for any fixed $\boldsymbol{\theta} \in \boldsymbol{\theta}$. Then we have $\mathbb{E}_{X,X'\sim\mathbb{P}}\left[\|F(X, X')\|_2^2\right] < \infty$ and $\mathbb{E}_{X\sim\mathbb{P}}\left[\|F(X, X)\|_2\right] < \infty$.

Proof. First of all, it follows from Lemma C.6 that for any $x, x' \in \mathcal{X}$,

$$\|F(x, x')\|_2 \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\nabla_{\boldsymbol{\theta}}\big(\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x, x')\big)\|_2 \leq M^1(x, x'). \tag{C.87}$$

Thus for the first moment we have $\mathbb{E}_{X\sim\mathbb{P}}\left[\|F(X, X)\|_2\right] \leq \mathbb{E}_{X\sim\mathbb{P}}[M^1(X, X)] < \infty$ from Lemma C.7. For the second moment, $\mathbb{E}_{X,X'\sim\mathbb{P}}\left[\|F(X, X')\|_2^2\right] \leq \mathbb{E}_{X,X'\sim\mathbb{P}}\left[M^1(X, X')^2\right] =: (*)$. By definition,

$$\begin{aligned}(*) &= \mathbb{E}_{X,X'\sim\mathbb{P}}\left[\big(m^1(X)m^0(X') + m^0(X)m^1(X')\big)^2\right] \\ &= 4\mathbb{E}_{X\sim\mathbb{P}}\left[m^1(X)^2\right]\mathbb{E}_{X\sim\mathbb{P}}\left[m^0(X)^2\right].\end{aligned} \tag{C.88}$$

Each of these expectations is finite by Lemma C.7, which completes the proof. □

Theorem C.2 (Concentration Inequality for KSD)**.** Let $\sigma(\boldsymbol{\theta}) := \mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(X,X)]$. Then

$$\mathbb{P}\left(|f_n(\boldsymbol{\theta}) - f(\boldsymbol{\theta})| \geq \delta\right) \leq \frac{4\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}}, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\theta}, \tag{C.89}$$

where the probability is with respect to realisations of the dataset $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

Proof. Since $|a^2 - b^2| = |(a+b)(a-b)| = (a+b)|a-b|$ for all $a, b \in [0,\infty)$, we have the bound

$$\begin{aligned}&\underbrace{\left|\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n) - \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})\right|}_{=:(*)} \\ &= \underbrace{\left(\mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n) + \mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})\right)}_{=:(*_1)} \underbrace{\left|\mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n) - \mathrm{KSD}(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})\right|}_{=:(*_2)}.\end{aligned} \tag{C.90}$$

In what follows we use $\mathbb{E}$ to denote an expectation with respect to the dataset $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$. Applying Markov's inequality followed by Cauchy–Schwarz, we have

$$\mathbb{P}((*) \geq \delta) \leq \frac{1}{\delta}\mathbb{E}[(*)] = \frac{1}{\delta}\mathbb{E}[(*_1)(*_2)] \leq \frac{1}{\delta}\sqrt{\mathbb{E}[(*_1)^2]}\sqrt{\mathbb{E}[(*_2)^2]}. \tag{C.91}$$

To conclude the proof, we bound the two expectations one the right hand side.

Bounding $\mathbb{E}[(*_1)^2]$: From the fact that $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}[(*_1)^2] &\leq 2\mathbb{E}\left[\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n) + \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})\right] \\ &= 2\Big(\mathbb{E}\left[\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n)\right] + \mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P})\Big).\end{aligned}$$

The preconditions of Lemma C.1 holds due to Standing Assumption 2. Thus from Lemma C.1 part (iii), together with Jensen's inequality, we have the two bounds $\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}_n) \leq (1/n)\sum_{i=1}^n \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x_i,x_i)$ and $\mathrm{KSD}^2(\mathbb{P}_{\boldsymbol{\theta}}\|\mathbb{P}) \leq \mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(X,X)]$. Plugging these into the previous inequality, and exploiting independence of x_i and x_j whenever $i \neq j$, we have

$$\begin{aligned}\mathbb{E}[(*_1)^2] &\leq 2\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(x_i,x_i)\right] + \mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(X,X)]\right) \\ &= 2\Big(\mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(X,X)] + \mathbb{E}_{X\sim\mathbb{P}}[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}K(X,X)]\Big) = 4\sigma(\boldsymbol{\theta}),\end{aligned}$$

where existence of $\sigma(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \boldsymbol{\theta}$ is ensured by Standing Assumption 2.

Bounding $\mathbb{E}[(*_2)^2]$: From the fact $|\sup_x |f(x)| - \sup_y |g(y)|| \leq \sup_x |f(x) - g(x)|$

for functions f and g, the term $(*_2)$ is upper bounded by

$$\begin{aligned}(*_2) &= \left| \sup_{\|h\|_{\mathcal{H}} \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h](x_i) \right| - \sup_{\|h\|_{\mathcal{H}} \leq 1} \left| \mathbb{E}_{X \sim \mathbb{P}}[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h](X)] \right| \right| \\ &\leq \sup_{\|h\|_{\mathcal{H}} \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h](x_i) - \mathbb{E}_{X \sim \mathbb{P}}[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[u](X)] \right| \\ &= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E}_{X \sim \mathbb{P}}[f(X)] \right|. \end{aligned} \tag{C.92}$$

where $\mathcal{F} := \{\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h] \mid \|h\|_{\mathcal{H}} \leq 1\}$. We can see from this expression that standard arguments in the context of Rademacher complexity theory can be applied. Noting that $|\cdot|^2$ is a convex function, Proposition 4.11 in Wainwright (2019) gives that

$$\begin{aligned}\mathbb{E}\left[(*_2)^2\right] &\leq \mathbb{E}\left[\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E}_{X \sim \mathbb{P}}[f(X)] \right| \right)^2 \right] \\ &\leq \mathbb{E}\mathbb{E}_{\epsilon}\left[2^2 \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right)^2 \right] \end{aligned} \tag{C.93}$$

where $\{\epsilon_i\}_{i=1}^n$ are independent random variables taking values in $\{-1, +1\}$ with equiprobability $1/2$ and $\mathbb{E}_{\epsilon}$ is the expectation over $\{\epsilon_i\}_{i=1}^n$. From the essentially same derivation as Proposition 6.1, the following equality holds:

$$\begin{aligned}&\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \\ &= \sup_{\|h\|_{\mathcal{H}} \leq 1} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[h](x_i) \right| = \sup_{\|h\|_{\mathcal{H}} \leq 1} \left| \left\langle h, \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i, \cdot) \right\rangle_{\mathcal{H}} \right| \\ &= \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i, \cdot) \right\|_{\mathcal{H}} = \sqrt{\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_i \epsilon_j \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i, x_j)}. \end{aligned}$$

Plugging this equality into the upper bound of $\mathbb{E}\left[(*_2)^2\right]$, we have

$$\begin{aligned}\mathbb{E}\left[(*_2)^2\right] &\leq 4\mathbb{E}\mathbb{E}_{\epsilon}\left[\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \epsilon_i \epsilon_j \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(x_i, x_j) \right] \qquad \text{(C.94)} \\ &= 4\mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}}[K(X_i, X_i)] \right] \\ &= \frac{4}{n} \mathbb{E}_{X \sim \mathbb{P}}[\mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} \mathcal{S}_{\mathbb{P}_{\boldsymbol{\theta}}} K(X, X)] = \frac{4\sigma(\boldsymbol{\theta})}{n}. \qquad \text{(C.95)} \end{aligned}$$

Bounding $\mathbb{E}[(*)^2]$: Returning to (C.91), we have the overall bound

$$\mathrm{P}((*) \geq \delta) \leq \frac{\sqrt{4\sigma(\boldsymbol{\theta})}\sqrt{4\sigma(\boldsymbol{\theta})}}{\delta\sqrt{n}} \leq \frac{4\sigma(\boldsymbol{\theta})}{\delta\sqrt{n}} \tag{C.96}$$

as claimed. □

Bibliography

Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

R. P. Adams and D. J. C. MacKay. Bayesian Online Changepoint Detection. *arXiv preprint arXiv:0710.3742*, 2007.

J. Aitchison. Goodness of Prediction Fit. *Biometrika*, 62(3):547–554, 1975.

A. A. Alemi. Variational Predictive Information Bottleneck. In *Workshop on Information Theory, Advances in Neural Information Processing Systems*, 2019.

P. Alquier. Non-Exponentially Weighted Aggregation: Regret Bounds for Unbounded Loss Functions. In *International Conference on Machine Learning*, pages 207–218. PMLR, 2021.

P. Alquier and B. Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

P. Alquier, J. Ridgway, and N. Chopin. On the Properties of Variational Approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1): 8374–8414, 2016.

P. Alquier, B.-E. Chérief Abdellatif, A. Derumigny, and J.-D. Fermanian. Estimation of copulas via Maximum Mean Discrepancy. *arXiv preprint arXiv:2010.00408*, 2020.

S.-i. Amari. *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media, 2012.

M. Arbel, D. Sutherland, M. Bińkowski, and A. Gretton. On gradient regularizers for MMD GANs. In *Advances in Neural Information Processing Systems*, pages 6700–6710, 2018.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy. BoTorch: A framework for efficient Monte-Carlo Bayesian Optimization. *Advances in neural information processing systems*, 33, 2020.

A. Barp, F.-X. Briol, A. B. Duncan, M. Girolami, and L. Mackey. Minimum Stein Discrepancy estimators. In *Neural Information Processing Systems*, pages 12964–12976, 2019.

D. Barry and J. A. Hartigan. A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and Efficient Estimation by Minimising a Density Power Divergence. *Biometrika*, 85(3):549–559, 1998.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

M. J. Beal. *Variational algorithms for approximate Bayesian inference*. University College London, 2003.

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. PAC-Bayesian bounds based on the Rényi Divergence. In *Artificial Intelligence and Statistics*, pages 435–444, 2016.

R. Beran et al. Minimum Hellinger distance estimates for parametric models. *The annals of Statistics*, 5(3):445–463, 1977.

J. O. Berger. The case for objective Bayesian analysis. *Bayesian analysis*, 1(3): 385–402, 2006.

J. O. Berger and J. M. Bernardo. On the Development of the Reference Prior Method. *Bayesian statistics*, 4(4):35–60, 1992.

J. O. Berger, E. Moreno, L. R. Pericchi, M. J. Bayarri, J. M. Bernardo, J. A. Cano, J. De la Horra, J. Martín, D. Ríos Insúa, D. A. Betrò, Bruno, P. Gustafson, L. Wasserman, J. B. Kadane, C. Srinivasan, M. Lavine, A. O'Hagan, W. Polasek, C. P. Robert, C. Goutis, F. Ruggeri, G. Salinetti, and S. Sivaganesan. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128, 1979.

J. M. Bernardo and A. F. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz Serna, and A. Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

W. Bialek, I. Nemenman, and N. Tishby. Predictability, Complexity, and Learning. *Neural computation*, 13(11):2409–2463, 2001.

P. G. Bissiri, C. Holmes, and S. Walker. A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

G. E. P. Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430, 1980.

F.-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with Maximum Mean Discrepancy. *arXiv:1906.05944*, 2019.

T. Bui, D. Hernández Lobato, J. Hernandez Lobato, Y. Li, and R. Turner. Deep Gaussian Processes for regression using approximate Expectation Propagation. In *International Conference on Machine Learning*, pages 1472–1481, 2016.

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance Weighted Autoencoders. In *International Conference on Learning Representations*, 2016.

S. Canu and A. Smola. Kernel Methods and the Exponential Family. *Neurocomputing*, 69(7-9):714–720, 2006.

Y. Cao and Y. Xie. Robust sequential change-point detection by convex optimization. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1287–1291. IEEE, 2017.

A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying. Universal Multi-Task Kernels. *The Journal of Machine Learning Research*, 9:1615–1646, 2008.

P. Carbonetto, M. King, and F. Hamze. A stochastic approximation method for inference in probabilistic graphical models. *Advances in neural information processing systems*, 22:216–224, 2009.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 10:377–408, 2006.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanit. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(1):19–61, 2010.

F. Caron, A. Doucet, and R. Gottardo. On-line changepoint Eetection and parameter estimation with application to genomic data. *Statistics and Computing*, 22 (2):579–595, 2012.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A Probabilistic Programming Language. *Journal of statistical software*, 76(1), 2017.

O. Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

L. Chen, C. Tao, R. Zhang, R. Henao, and L. C. Duke. Variational Inference and Model Selection with Generalized Evidence Bounds. In *International Conference on Machine Learning*, pages 892–901, 2018.

W. Y. Chen, A. Barp, F.-X. Briol, J. Gorham, M. Girolami, L. Mackey, and C. J. Oates. Stein Point Markov Chain Monte Carlo. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1011–1021, 2019.

B.-E. Chérief Abdellatif and P. Alquier. MMD-Bayes: Robust Bayesian estimation via Maximum Mean Discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. PMLR, 2020.

B.-E. Chérief Abdellatif and P. Alquier. Finite sample properties of parametric MMD estimation: Robustness to misspecification and dependence. *Bernoulli*, 28 (1):181–213, 2022.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of Multilayer Networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015.

K. Chwialkowski, H. Strathmann, and A. Gretton. A Kernel Test of Goodness of Fit. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2606–2615, 2016.

A. Cichocki and S.-i. Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158, 1975.

K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. Random feature expansions for Deep Gaussian Processes. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 884–893. JMLR, 2017a.

K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. Random feature expansions for Deep Gaussian Processes. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 884–893. JMLR. org, 2017b.

Z. Dai, A. Damianou, J. Gonzalez, and N. Lawrence. Variational Auto-encoded Deep Gaussian Processes. In *International Conference on Learning Representations*, 2016.

G. Dal Maso. *An Introduction to* Γ*-convergence*, volume 8. Springer Science & Business Media, 2012.

A. Damianou and N. Lawrence. Deep Gaussian Processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Danica J. Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with Nyström kernel exponential families. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 652–660, 2018.

G. Darmois. Sur les lois de probabilité à estimation exhaustive. *Comptes Rendus de l'Académie des Sciences*, 200:1265–1266, 1935.

H. A. David. First (?) occurrence of common terms in probability and statisticsA second list, with corrections. *The American Statistician*, 52(1):36–40, 1998.

J. Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, 1994.

A. P. Dawid, M. Musio, and L. Ventura. Minimum Scoring Rule Inference. *Scandinavian Journal of Statistics*, 43(1):123–138, 2016.

P.-S. De Laplace. Mémoire sur la probabilité des causes par les événements. *Mém. de math. et phys. présentés à lAcad. roy. des sci*, 6:621–656, 1774.

A. Defazio, F. Bach, and S. Lacoste Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, New York, 2012.

A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. Blei. Variational Inference via χ Upper Bound Minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741, 2017.

J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. Hoffman, and R. A. Saurous. Tensorflow Distributions. *arXiv preprint arXiv:1711.10604*, 2017.

J. Domke and D. R. Sheldon. Importance Weighting and Variational Inference. In *Advances in neural information processing systems*, pages 4470–4479, 2018.

M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

R. Durrett. *Probability: Theory and Examples (4th Edition)*. Cambridge University Press, 2010.

Edwin V. Bonilla, Karl Krauth, and Amir Dezfouli. Generic Inference in Latent Gaussian Process Models. *Journal of Machine Learning Research*, 20(117):1–63, 2019.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 20:28:1–28:6, 2019.

K. Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39(1):42, 1953.

F. Farnia and D. Tse. A convex duality framework for GANs. In *Advances in Neural Information Processing Systems*, pages 5248–5258, 2018.

F. Fazayeli and A. Banerjee. Generalized direct change estimation in Ising model structure. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2281–2290, 2016.

P. Fearnhead. Exact Bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing*, 53(6):2160–2166, 2005.

P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (4):589–605, 2007.

P. Fearnhead and G. Rigaill. Changepoint Detection in the Presence of Outliers. *Journal of the American Statistical Association*, 114(525):169–183, 2019.

S. E. Fienberg. When did Bayesian inference become "Bayesian"? *Bayesian analysis*, 1(1):1–40, 2006.

R. A. Fisher. Contributions to Mathematical Statistics. 1950.

H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

F. Futami, I. Sato, and M. Sugiyama. Variational Inference based on Robust Divergences. In *Artificial Intelligence and Statistics*, 2018.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational Inference based on Robust Divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 813–822. PMLR, 2018.

K. Ganchev, J. Gillenwater, and B. Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, 2010.

J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. Gpytorch: Blackbox matrix-matrix Gaussian Process inference with GPU acceleration. In *Advances in Neural Information Processing Systems*, pages 7576–7586, 2018.

R. Garnett, M. A. Osborne, and S. J. Roberts. Sequential Bayesian prediction in the presence of changepoints. In *Proceedings of the 26th International Conference on Machine Learning*, pages 345–352. ACM, 2009.

A. Gelman, D. Simpson, and M. Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017.

P. Germain, F. Bach, A. Lacoste, and S. Lacoste Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.

Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. *Advances in neural information processing systems*, pages 507–513, 2001.

S. Ghosal. A review of consistency and convergence rates of posterior distributions. In *Proc. Varanasi Symp. on Bayesian Inference*, 1998.

S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

A. Ghosh and A. Basu. Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

M. Gil. *On Rényi Divergence measures for continuous alphabet sources*. PhD thesis, 2011.

M. Gil, F. Alajaji, and T. Linder. Rényi Divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

M. Goldstein. Influence and Belief Adjustment. *Influence Diagrams, Belief Nets and Decision Analysis*, pages 143–174, 1990.

M. Goldstein. Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, 1(3):403–420, 2006.

J. Gorham and L. Mackey. Measuring Sample Quality with Stein's Method. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring Sample Quality with Diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.

W. Grathwohl, D. Choi, Y. Wu, G. Roeder, and D. Duvenaud. Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018.

P. Grünwald. Safe Learning: Bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420, 2011.

P. Grünwald. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer, 2012.

P. Grünwald and T. Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4): 1069–1103, 2017.

B. Guedj. A Primer on PAC-Bayesian Learning. In *Proceedings of the French Mathematical Society*, volume 33, pages 391–414. Société Mathématique de France, 2019.

M. F. Guillén. The global economic & financial crisis: A timeline. *The Lauder Institute, University of Pennsylvania*, pages 1–91, 2009.

O. Hamelijnck, T. Damoulas, K. Wang, and M. Girolami. Multi-resolution Multi-task Gaussian Processes. In *Advances in Neural Information Processing Systems*, 2019.

E. Hannan and L. Kavalieris. Regression, autoregression models. *Journal of Time Series Analysis*, 7(1):27–49, 1986.

R. Harikandeh, M. O. Ahmed, A. Virani, M. Schmidt, J. Konevechcnỳ, and S. Sallinen. Stop wasting my gradients: Practical SVRG. In *Advances in Neural Information Processing Systems*, pages 2251–2259, 2015.

P. Hegde, M. Heinonen, H. Lähdesmäki, and S. Kaski. Deep Learning with Differential Gaussian Process Flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1812–1821, 2019.

J. Hensman and N. D. Lawrence. Nested Variational Compression in Deep Gaussian Processes. *stat*, 1050:3, 2014.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian Processes for Big Data. In *Uncertainty in Artificial Intelligence*, page 282, 2013.

W. Herlands, A. Wilson, H. Nickisch, S. Flaxman, D. Neill, W. Van Panhuis, and E. Xing. Scalable Gaussian Processes for characterizing multidimensional change surfaces. In *Artificial Intelligence and Statistics*, pages 1013–1021, 2016.

J. M. Hernández Lobato and R. Adams. Probabilistic Backpropagation for scalable learning of Bayesian Neural Networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

J. M. Hernández Lobato, Y. Li, M. Rowland, D. Hernández Lobato, T. D. Bui, and R. E. Turner. Black-box α-divergence minimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1511–1520, 2016.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

S. M. Hill, L. M. Heiser, T. Cokelaer, M. Unger, N. K. Nesser, D. E. Carlin, Y. Zhang, A. Sokolov, E. O. Paull, C. K. Wong, et al. Inferring Causal Molecular Networks: Empirical assessment through a community-based effort. *Nature Methods*, 13(4): 310–318, 2016.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic Variational Inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

C. Holmes and S. Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

G. Hooker and A. N. Vidyashankar. Bayesian Model Robustness via Disparities. *Test*, 23(3):556–584, 2014.

T.-C. Hu, F. Moricz, and R. Taylor. Strong laws of large numbers for arrays of rowwise independent random variables. *Acta Mathematica Hungarica*, 54(1-2): 153–162, 1989.

C.-W. Huang, S. Tan, A. Lacoste, and A. C. Courville. Improving Explorability in Variational Inference with Annealed Variational Objectives. In *Advances in Neural Information Processing Systems*, pages 9724–9734, 2018.

P. J. Huber. Robust Statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

H. Hung, Z.-Y. Jou, and S.-Y. Huang. Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, 74(1):145–154, 2018.

H. Husain. Distributional Robustness with IPMs and links to Regularization and GANs. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11816–11827, 2020.

H. Husain, R. Nock, and R. C. Williamson. A Primal-Dual link between GANs and Autoencoders. In *Advances in Neural Information Processing Systems*, pages 413–422, 2019.

Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. MCMC for Doubly-Intractable Distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2006.

A. Inoue and Y. Kasahara. Explicit representation of finite predictor coefficients and its applications. *The Annals of Statistics*, pages 973–993, 2006.

A. Inoue, Y. Kasahara, M. Pourahmadi, et al. Baxters inequality for finite predictor coefficients of multivariate long-memory stationary processes. *Bernoulli*, 24(2): 1202–1232, 2018.

Jackson Gorham and Lester Mackey. Measuring Sample Quality with Kernels. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1292–1301, 2017.

Jaewoo Park and Murali Haran. Bayesian Inference in the Presence of Intractable Normalizing Functions. *Journal of the American Statistical Association*, 113(523): 1372–1390, 2018.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge university press, 2003.

Jeffrey W. Miller. Asymptotic Normality, Concentration, and Coverage of Generalized Posteriors. *Journal of Machine Learning Research*, 22(168):1–53, 2021.

H. Jeffreys. Theory of Probability: Oxford Univ. *Press (earlier editions 1939, 1948)*, 1961.

Jeremias Knoblauch and Lara Vomfell. Robust Bayesian Inference for Discrete Outcomes with the Total Variation Distance. *ArXiv*, abs/2010.13456, 2020.

J. Jewson, J. Smith, and C. Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

X. Jiang, Q. Li, and G. Xiao. Bayesian Modeling of Spatial Transcriptomics Data via a Modified Ising Model. *arXiv:2104.13957*, 2021.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

M. Jones, N. L. Hjort, I. R. Harris, and A. Basu. A comparison of related density-based Minimum Divergence Estimators. *Biometrika*, 88(3):865–873, 2001.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine learning*, 37(2):183–233, 1999.

Julian Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

Julian Besag. On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986.

R. E. Kass and A. E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

A. Khaleghi and D. Ryabko. Asymptotically consistent estimation of the number of change points in highly dependent time series. In *Proceedings of the 31st International Conference on Machine Learning*, pages 539–547, 2014.

D. P. Kingma and J. Ba. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2013.

D. P. Kingma, T. Salimans, and M. Welling. Variational Dropout and the Local Reparameterization Trick. In *Advances in Neural Information Processing Systems*, pages 2575–2583, 2015.

B. J. Kleijn and A. W. van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

J. P. Kleijnen and R. Y. Rubinstein. Optimization and sensitivity analysis of computer simulation models by the score function method. *European Journal of Operational Research*, 88(3):413–427, 1996.

J. Knoblauch. Frequentist Consistency of Generalized Variational Inference. *arXiv preprint arXiv:1912.04946*, 2019.

J. Knoblauch and T. Damoulas. Spatio-temporal Bayesian On-line Changepoint Detection with Model Selection. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2018.

J. Knoblauch, J. Jewson, and T. Damoulas. Doubly Robust Bayesian Inference for Non-Stationary Streaming Data using β-Divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64–75, 2018.

J. Knoblauch, J. Jewson, and T. Damoulas. Generalized Variational Inference: Three arguments for deriving new posteriors. *arXiv:1904.02063*, 2019.

B. Koopman. On Distributions Admitting a Sufcient Statistic. *Transactions of the American Mathematical Society*, 39, 1936.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schlkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

S. Kurtek and K. Bharath. Bayesian sensitivity analysis with the Fisher–Rao metric. *Biometrika*, 102(3):601–616, 2015.

T. Kuśmierczyk, J. Sakaya, and A. Klami. Variational Bayesian Decision-making for Continuous Utilities. In *Advances in Neural Information Processing Systems*, 2019.

S. Lacoste Julien, F. Huszár, and Z. Ghahramani. Approximate Inference for the loss-calibrated Bayesian. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 416–424, 2011.

L. Lei and M. Jordan. Less than a Single Pass: Stochastically Controlled Stochastic Gradient. In *Artificial Intelligence and Statistics*, pages 148–156, 2017.

L. Lei, C. Ju, J. Chen, and M. I. Jordan. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems*, pages 2345–2355, 2017.

G. Lever, F. Laviolette, and J. Shawe Taylor. Tighter PAC-Bayes bounds through distribution-dependent Priors. *Theoretical Computer Science*, 473:4–28, 2013.

C. Levy leduc and Z. Harchaoui. Catching change-points with LASSO. In *Advances in Neural Information Processing Systems*, pages 617–624, 2008.

C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards deeper understanding of Moment Matching Network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.

S. Li, Y. Xie, H. Dai, and L. Song. M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems*, pages 3366–3374, 2015.

Y. Li and R. E. Turner. Rényi Divergence Variational Inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

M. Lichman. UCI machine learning repository, 2013.

F. Liese and I. Vajda. Convex statistical distances, volume 95 of Teubner Texts in Mathematics. *BSB BG Teubner Verlagsgesellschaft, Leipzig*, 1987.

K. Lin, J. L. Sharpnack, A. Rinaldo, and R. J. Tibshirani. A Sharp Error Analysis for the Fused Lasso, with Application to Approximate Changepoint Screening. In *Advances in Neural Information Processing Systems*, pages 6887–6896, 2017.

B. G. Lindsay et al. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The annals of statistics*, 22(2):1081–1114, 1994.

Q. Liu, J. Lee, and M. Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 276–284, 2016.

S. Liu and K. Chaudhuri. The inductive bias of restricted f-GANs. *arXiv preprint arXiv:1809.04542*, 2018.

S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

S. Liu, T. Kanamori, W. Jitkrittum, and Y. Chen. Fisher Efficient Inference of Intractable Models. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2019.

G. Loaiza Ganem and J. P. Cunningham. The continuous Bernoulli: fixing a pervasive error in Variational Autoencoders. In *Advances in Neural Information Processing Systems*, 2019.

Y. Lu, A. Stuart, and H. Weber. Gaussian Approximations for Probability Measures on $\mathbb{R}^d$. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):1136–1165, 2017.

J. Luterbacher, D. Dietrich, E. Xoplaki, M. Grosjean, and H. Wanner. European seasonal and annual temperature variability, trends, and extremes since 1500. *Science*, 303(5663):1499–1503, 2004.

S. P. Lyddon, C. C. Holmes, and S. G. Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

C. Ma, Y. Li, and J. M. Hernández Lobato. Variational Implicit Processes. In *International Conference on Machine Learning*, pages 4222–4233. PMLR, 2019.

D. J. C. MacKay. Bayesian methods for backpropagation networks. In *Models of neural networks III*, pages 211–254. Springer, 1996.

D. J. C. MacKay. Choice of basis for Laplace approximation. *Machine learning*, 33 (1):77–86, 1998.

T. Matsubara, J. Knoblauch, F.-X. Briol, C. Oates, et al. Robust Generalised Bayesian Inference for Intractable Likelihoods. *arXiv preprint arXiv:2104.07359*, 2021a.

T. Matsubara, C. J. Oates, and F.-X. Briol. The Ridgelet Prior: A covariance function approach to prior specification for Bayesian Neural Networks. *Journal of Machine Learning Research*, 22:1–57, 2021b.

A. G. d. G. Matthews, J. Hensman, R. Turner, and Z. Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *Journal of Machine Learning Research*, 51:231–239, 2016.

A. G. d. G. Matthews, M. Van Der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León Villagrá, Z. Ghahramani, and J. Hensman. GPflow: A Gaussian Process library using TensorFlow. *The Journal of Machine Learning Research*, 18(1): 1299–1304, 2017.

C. Mayo Wilson and A. Saraf. Qualitative Robust Bayesianism and the Likelihood Principle. *arXiv preprint arXiv:2009.03879*, 2020.

D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999a.

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM, 1999b.

M. Meyer and J.-P. Kreiss. On the Vector Autoregressive sieve bootstrap. *Journal of Time Series Analysis*, 36(3):377–397, 2015.

M. Mihoko and S. Eguchi. Robust blind source separation by β-divergence. *Neural computation*, 14(8):1859–1886, 2002.

J. W. Miller and D. B. Dunson. Robust Bayesian Inference via Coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

T. Minka. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

T. P. Minka. Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih. Monte Carlo Gradient Estimation in Machine Learning. *J. Mach. Learn. Res.*, 21(132):1–62, 2020.

M. Moores, G. Nicholls, A. Pettitt, and K. Mengersen. Scalable Bayesian inference for the inverse temperature of a hidden Potts model. *Bayesian Analysis*, 15(1): 1–27, 2020.

Y. Mroueh and T. Sercu. Fisher GAN. In *Advances in Neural Information Processing Systems*, pages 2513–2523, 2017.

Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng. Sobolev GAN. In *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview.net, 2018.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

T. Nakagawa and S. Hashimoto. Robust Bayesian Inference via γ-divergence. *Communications in Statistics-Theory and Methods*, pages 1–18, 2019.

E. Nalisnick, J. Gordon, and J. M. Hernandez Lobato. Predictive Complexity Priors. In *International Conference on Artificial Intelligence and Statistics*, pages 694–702. PMLR, 2021.

R. M. Neal. *Bayesian learning for Neural Networks*, volume 118. Springer Science & Business Media, 2012.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

S. Niekum, S. Osentoski, C. G. Atkeson, and A. G. Barto. CHAMP: Changepoint detection using approximate model parameters. Technical report, (No. CMU-RI-TR-14-10) Carnegie-Mellon University Pittsburgh PA Robotics Institute, 2014.

A. Nitanda. Stochastic Proximal Gradient Descent with Acceleration Techniques. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2014.

C. J. Oates. *Bayesian Inference for protein signalling networks*. PhD thesis, University of Warwick, 2013.

A. O'Hagan and J. E. Oakley. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering & System Safety*, 85(1):239–248, 2004.

Y. Ohnishi and J. Honorio. Novel Change of Measure Inequalities with Applications to PAC-Bayesian bounds and Monte Carlo estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 1711–1719. PMLR, 2021.

E. Ollila and E. Raninen. Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions. *IEEE Transactions on Signal Processing*, 67(10):2707–2719, 2019.

M. Opper and O. Winther. Gaussian Processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684, 2000.

J. J. K. O'Ruanaidh. *Numerical Bayesian methods applied to signal processing*. PhD thesis, University of Cambridge, 1994.

J. Paisley, D. M. Blei, and M. I. Jordan. Variational Bayesian Inference with Stochastic Search. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1363–1370, 2012.

G. Patrini, R. van den Berg, P. Forre, M. Carioni, S. Bhargav, M. Welling, T. Genewein, and F. Nielsen. Sinkhorn Autoencoders. In *Uncertainty in Artificial Intelligence*, pages 733–743. PMLR, 2020.

F. Pauli, W. Racugno, and L. Ventura. Bayesian Composite Marginal Likelihoods. *Statistica Sinica*, pages 149–164, 2011.

V. I. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016.

Pei-Shien Wu and Ryan Martin. A comparison of learning rate selection methods in Generalized Bayesian Inference. *arXiv:2012.11349*, 2020.

F. Peng and D. K. Dey. Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics*, 23(2):199–213, 1995.

J.-P. Penot. *Calculus without Derivatives*, volume 266. Springer Science & Business Media, 2012.

Peter J. Diggle. A Point Process Modelling Approach to Raised Incidence of a Rare Phenomenon in the Vicinity of a Prespecified Point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3):349–362, 1990.

D. Phan, N. Pradhan, and M. Jankowiak. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554*, 2019.

Pierre Alquier and Mathieu Gerber. Universal Robust Regression via Maximum Mean Discrepancy. *arXiv: 2006.00840*, 2020.

E. Pitman. Sufficient Statistics and Intrinsic Accuracy. *Proceedings of the Cambridge Philosophical Society*, 32, 1936.

M. Pollak. A Robust Changepoint Detection Method. *Sequential Analysis*, 29(2): 146–161, 2010.

M. Postman, J. Huchra, and M. Geller. Probes of large-scale structure in the Corona Borealis region. *The Astronomical Journal*, 92:1238–1247, 1986.

J. Quiñonero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian Process regression. *Journal of Machine Learning Research*, 6(Dec): 1939–1959, 2005.

R. Ranganath, S. Gerrish, and D. Blei. Black box Variational Inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

R. Ranganath, D. Tran, J. Altosaar, and D. Blei. Operator Variational Inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.

A. Ranganathan, M.-H. Yang, and J. Ho. Online sparse Gaussian Process regression and its applications. *IEEE Transactions on Image Processing*, 20(2):391–404, 2011.

M. D. Reid, R. M. Frongillo, R. C. Williamson, and N. Mehta. Generalized Mixability via Entropic Duality. In *Conference on Learning Theory*, pages 1501–1522, 2015.

A. Rényi. On measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

D. J. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 1530–1538, 2015.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in Deep Generative Models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

M. Riabiz, W. Chen, J. Cockayne, P. Swietach, S. A. Niederer, L. Mackey, and C. J. Oates. Optimal Thinning of MCMC Output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021. To appear.

M. Ribatet, D. Cooley, and A. C. Davison. Bayesian inference from Composite Likelihoods, with an application to spatial extremes. *Statistica Sinica*, pages 813–845, 2012.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to Langevin Diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of Random Walk Metropolis Algorithms. *The annals of applied probability*, 7(1): 110–120, 1997.

R. Rockafellar. Integrals which are Convex Functionals. *Pacific journal of mathematics*, 24(3):525–539, 1968.

R. T. Rockafellar. *Convex Analysis*. Number 28. Princeton university press, 1970.

K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411): 617–624, 1990.

S. Rossi, P. Michiardi, and M. Filippone. Good Initializations of Variational Bayes for Deep Models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5487–5497, 2019.

S. Rossi, S. Marmin, and M. Filippone. Walsh-Hadamard Variational Inference for Bayesian Deep Learning. *Advances in Neural Information Processing Systems*, 33, 2020.

R. Y. Rubinstein, A. Shapiro, and S. Uryasev. The Score Function Method. *Encyclopedia of Management Sciences*, pages 1363–1366, 1996.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

E. Ruggieri and M. Antonellis. An exact approach to Bayesian sequential change point detection. *Computational Statistics & Data Analysis*, 97:71–86, 2016.

Y. Saatçi, R. D. Turner, and C. E. Rasmussen. Gaussian Process Change Point Models. In *Proceedings of the 27th International Conference on Machine Learning*, pages 927–934, 2010.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal Protein-Signaling Networks derived from Multiparameter Single-Cell Data. *Science*, 308 (5721):523–529, 2005.

A. Saha, K. Bharath, and S. Kurtek. A Geometric Variational Approach to Bayesian Inference. *Journal of the American Statistical Association*, pages 1–25, 2019.

T. Salimans and D. A. Knowles. On using control variates with stochastic approximation for variational Bayes and its connection to stochastic linear regression. *arXiv preprint arXiv:1401.1022*, 2014.

H. Salimbeni and M. Deisenroth. Doubly Stochastic Variational Inference for Deep Gaussian Processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

A. SenGupta. Generalizations of Barlett's and Hartley's tests of homogeneity using overall variability. *Communications in Statistics-Theory and Methods*, 16(4):987–996, 1987.

J. Shawe Taylor and R. C. Williamson. A PAC analysis of a Bayesian estimator. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 2–9, 1997.

X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

J. Shi, S. Sun, and J. Zhu. Kernel Implicit Variational Inference. In *International Conference on Learning Representations*, 2018.

Z. Shun and P. McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):749–760, 1995.

D. G. Simpson. Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American statistical Association*, 82(399):802–807, 1987.

E. Snelson and Z. Ghahramani. Sparse Gaussian Processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

E. Sober. *Evidence and Evolution: The Logic behind the Science*. Cambridge University Press, 2008.

C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. Ladder Variational Autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, 1972.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.

H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabo, and A. Gretton. Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

N. Syring and R. Martin. Calibrating General Posterior Credible Regions. *Biometrika*, 106(2):479–486, 2019.

R. N. Tamura and D. D. Boos. Minimum Hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81(393):223–229, 1986.

L. C. Tiao, E. V. Bonilla, and F. Ramos. Cycle-Consistent Adversarial Learning as Approximate Bayesian Inference. *arXiv preprint arXiv:1806.01771*, 2018.

L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

N. Tishby, F. C. Pereira, and W. Bialek. The Information Bottleneck Method. *arXiv preprint physics/0004057*, 2000.

M. Titsias. Variational Learning of inducing variables in sparse Gaussian Processes.

M. Titsias and M. Lázaro Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, pages 1971–1979, 2014.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein Auto-Encoders. In *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview.net, 2018.

Tomoyuki Nakagawa and Shintaro Hashimoto. Robust Bayesian Inference via γ-divergence. *Communications in Statistics - Theory and Methods*, 49(2):343–360, 2020.

U. v. Toussaint, S. Gori, and V. Dose. Invariance priors for Bayesian feed-forward Neural Networks. *Neural Networks*, 19(10):1550–1557, 2006.

D. Tran, R. Ranganath, and D. M. Blei. The Variational Gaussian Process. In *4th International Conference on Learning Representations, ICLR 2016*, 2016.

D. Tran, R. Ranganath, and D. Blei. Hierarchical Implicit Models and Likelihood-Free Variational Inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

R. Turner, Y. Saatci, and C. E. Rasmussen. Adaptive sequential Bayesian change point detection. In *Temporal Segmentation Workshop at NIPS*, 2009.

R. D. Turner. *Gaussian processes for State Space Models and change point detection.* PhD thesis, University of Cambridge, 2012.

R. D. Turner, S. Bottone, and C. J. Stanek. Online variational approximations to non-exponential family change point models: with application to radar tracking. In *Advances in Neural Information Processing Systems*, pages 306–314, 2013.

R. E. Turner and M. Sahani. Two problems with variational Expectation Maximisation for time-series models. In *Bayesian time series models.* Cambridge University Press, 2011.

K. Vafa. Training Deep Gaussian Processes with Sampling. In *NIPS 2016 Workshop on Advances in Approximate Bayesian Inference*, 2016.

A. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 1998.

T. Van Erven and P. Harremos. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

C. Varin, N. Reid, and D. Firth. An Overview of Composite Likelihood Methods. *Statistica Sinica*, pages 5–42, 2011.

C. Villani. *Optimal transport: Old and new*, volume 338. Springer Science & Business Media, 2008.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge University Press, 2019.

S. Walker. New approaches to Bayesian consistency. *The Annals of Statistics*, 32 (5):2028–2043, 2004.

D. Wang, H. Liu, and Q. Liu. Variational Inference with Tail-adaptive f-Divergence. In *Advances in Neural Information Processing Systems*, pages 5742–5752, 2018.

D. Wang, Z. Tang, C. Bajaj, and Q. Liu. Stein Variational Gradient Descent with Matrix-Valued Kernels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2019a.

K. Wang, G. Pleiss, J. Gardner, S. Tyree, K. Q. Weinberger, and A. G. Wilson. Exact Gaussian Processes on a million data points. *Advances in Neural Information Processing Systems*, 32:14648–14659, 2019b.

Y. Wang and D. M. Blei. Frequentist Consistency of Variational Bayes. *Journal of the American Statistical Association*, pages 1–15, 2018.

Y. Wang, M. Brubaker, B. Chaib Draa, and R. Urtasun. Sequential Inference for Deep Gaussian Process. In *Artificial Intelligence and Statistics*, pages 694–703, 2016.

Wassily Hoeffding. The strong law of large numbers for U-statistics. *Institute of Statistics Mimeo Series*, 302, 1961.

S. Watanabe. *Mathematical Theory of Bayesian Statistics*. CRC Press, 2018.

L. Wenliang, D. J. Sutherland, H. Strathmann, and A. Gretton. Learning Deep Kernels for Exponential Family Densities. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6737–6746, 2019.

L. K. Wenliang. Blindness of score-based methods to isolated components and mixing proportions. *arXiv:2008.10087*, 2020.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245, 1994.

S. Wilks. Multidimensional statistical scatter. *Contributions to Probability and Statistics*, pages 486–503, 1960.

C. K. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.

R. J. Williams. Simple statistical gradient-following algorithms for Connectionist Reinforcement Learning. *Machine learning*, 8(3):229–256, 1992.

R. C. Wilson, M. R. Nassar, and J. I. Gold. Bayesian Online Learning of the Hazard rate in change-point problems. *Neural computation*, 22(9):2452–2476, 2010.

D. Wingate and T. Weber. Automated Variational Inference in Probabilistic Programming. *stat*, 1050:7, 2013.

M. Wu, N. Goodman, and S. Ermon. Differentiable Antithetic Sampling for Variance Reduction in Stochastic Variational Inference. In *Proceedings of Machine Learning Research*, volume 89, pages 2877–2886, 2019.

X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th International Conference on Machine Learning*, pages 1055–1062. ACM, 2007.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(115): 3813–3847, 2015.

Y. Yang, R. Martin, and H. Bondell. Variational approximations using Fisher divergence. *arXiv preprint arXiv:1905.05284*, 2019.

Y. Yang, D. Pati, and A. Bhattacharya. *alpha*-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Yes, but did it work?: Evaluating Variational Inference. In *International Conference on Machine Learning*, pages 5581–5590. PMLR, 2018.

Y. G. Yatracos. Rates of convergence of Minimum Distance Estimators and Kolmogorov's Entropy. *The Annals of Statistics*, pages 768–774, 1985.

K. Y. Yılmaz, A. T. Cemgil, and U. Simsekli. Generalised coupled tensor factorisation. In *Advances in neural information processing systems*, pages 2151–2159, 2011.

M. Yu, M. Kolar, and V. Gupta. Statistical Inference for Pairwise Graphical Models Using Score Matching. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2016.

C. Zalinescu. *Convex Analysis in General Vector Spaces*. World scientific, 2002.

A. Zellner. Maximal Data Information Prior Distributions. *New developments in the applications of Bayesian methods*, pages 211–232, 1977.

A. Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

G. Zhang, S. Sun, D. Duvenaud, and R. Grosse. Noisy Natural Gradient as Variational Inference. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5852–5861, 2018.

S. Zhang, Y. Gao, Y. Jiao, J. Liu, Y. Wang, and C. Yang. Wasserstein-Wasserstein Auto-Encoders. *arXiv preprint arXiv:1902.09323*, 2019.

T. Zhang. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

J. Zhu, N. Chen, and E. P. Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *The Journal of Machine Learning Research*, 15(1):1799–1847, 2014.