Robust Bayesian Inference for Simulator-based Models via the MMD Posterior Bootstrap

Charita Dellaporta University of Warwick
Jeremias Knoblauch University College London
Theodoros Damoulas University of Warwick
François-Xavier Briol University College London

Abstract

Simulator-based models are models for which the likelihood is intractable but simulation of synthetic data is possible. They are often used to describe complex real-world phenomena, and as such can often be misspecified in practice. Unfortunately, existing Bayesian approaches for simulators are known to perform poorly in those cases. In this paper, we propose a novel algorithm based on the posterior bootstrap and maximum mean discrepancy estimators. This leads to a highly-parallelisable Bayesian inference algorithm with strong robustness properties. This is demonstrated through an in-depth theoretical study which includes generalisation bounds and proofs of frequentist consistency and robustness of our posterior. The approach is then assessed on a range of examples including a g-and-k distribution and a toggle-switch model.

1 INTRODUCTION

A key assumption of standard Bayesian inference is that the data-generating mechanism lies within the family of models defined by our choice of likelihood (Bernardo et al. 2009). In reality, this assumption is often violated, and this can lead to inconsistent and misleading inference outcomes; see for example Grünwald et al. (2017) and Owhadi et al. (2015) for the cases of linear or nonparametric regression.

The need for robust inference methods is even more apparent when dealing with simulator-based models. In these models, independent sampling is possible, but evaluating the likelihood is intractable. Simulators are used in a range of fields including population genetics (Riesselman et al. 2018), ecology (Beaumont 2010) and telecommunication engineering (Bharti et al. 2021); see Cranmer et al. (2020) for a recent review. In each of these examples, the model is at best a rough approximation of a complex physical or biological phenomenon, and will most likely not capture all of the key characteristics of the underlying data generating process. Unfortunately, the main approaches to Bayesian inference for simulators, including approximate Bayesian computation (ABC) (Beaumont et al. 2002), tend to approximate the classical Bayesian posterior, which itself lacks robustness to misspecification (Frazier et al. 2020). This leads ABC posteriors to exhibit poor coverage; see Hermans et al. (2021) for an empirical study.

One promising approach for addressing model misspecification is generalised Bayesian inference (GBI; e.g. Hooker et al. (2014), Ghosh et al. (2016), Bissiri et al. (2016), Jewson et al. (2018), and Knoblauch et al. (2019)) and the closely-related fields of quasi-Bayesian inference (e.g. Chernozhukov et al. 2003), PAC-Bayesian inference (e.g. Shawe-Taylor et al. 1997; Catoni 2007) and Safe-Bayes (Grünwald 2011). These approaches typically work directly with the original model, but attempt to correct for possible misspecification by scoring observations robustly. Recent work has focused on GBI for intractable likelihood models (Schmon et al. 2020; Pacchiardi et al. 2020; Matsubara et al. 2021). While some of these techniques address robustness concerns in simulator-based methods, they have two main drawbacks: Firstly, they require setting a crucial hyperparameter that determines the relative importance of the prior for the posterior inferences. Secondly, they make use of Monte Carlo methods, which imposes a substantial computational burden.

Another recent approach for inference under misspecification is Bayesian nonparametric learning (NPL) (see Lyddon et al. 2018; Lyddon et al. 2019; Fong et al. 2019). Unlike GBI, NPL does not address misspecifica-

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

tion by robustly scoring the statistical model. Instead, robustness is achieved by obtaining a non-parametric posterior directly on the data-generating process. This posterior then implies a posterior on the parameter of interest through the use of a robust loss function.

Our paper's contribution is the first NPL-based algorithm for simulators. Specifically, we leverage the NPL framework to obtain a posterior belief distribution about the parameters that minimise the maximum mean discrepancy (MMD) (Gretton et al. 2012) between our model and the data-generating mechanism. The MMD is a probability metric that is not only robust, but also easy to approximate through simulation—and therefore suitable for simulators. Further, the MMD has numerous desirable theoretical robustness and generalisation properties (see Briol et al. 2019; Chérief-Abdellatif et al. 2019; Chérief-Abdellatif et al. 2020). One of the main achievements of this paper is to show that these hold for our method whenever we use a bounded kernel. Additionally, unlike ABC or GBI, our approach is computationally efficient: it is trivially parallelisable, and never requires discarding parameter samples or using inherently sequential Monte Carlo methods.

The paper is structured as follows. In Section 2, we recall the details of NPL and introduce MMD estimators. In Section 3, we propose our novel algorithm which combines these concepts to create a scalable approach for robust Bayesian inference with simulators. Then, Section 4 provides theoretical results including consistency and robustness. Finally, Section 5 studies the algorithm on a range of benchmark problems for simulators including the g-and-k distribution, and a toggle-switch model describing the interaction of genes through time.

2 BACKGROUND

Let $\mathcal{X}$ denote our data space, $\mathcal{P}$ the space of Borel distributions on $\mathcal{X}$, and $\mathbb{P} \in \mathcal{P}$ the true data-generating mechanism of our data. In Bayesian statistics, given observations $x_{1:n} = x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathbb{P}$, one chooses a parametric model $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\} \subset \mathcal{P}$. Given $\mathcal{P}_\Theta$, the Bayesian now places a prior on the parameter θ, and then conditions on $x_{1:n}$ to obtain a posterior on θ. For standard Bayesian inference to be well-behaved, we have to assume that $\mathcal{P}_\Theta$ is well-specified; i.e $\exists \theta_0 \in \Theta$ such that $\mathbb{P}_{\theta_0} = \mathbb{P}$. When this assumption is violated, we call the model misspecified.

Misspecification in the Bayesian context has recently seen increasing interest since the Bayesian posterior does not provide robust parameter inferences (see Grünwald 2012). This has led to GBI and NPL approaches aimed at rectifying the issue. In this paper, we combine the strengths of both approaches for robust inference with simulators.

2.1 Simulator-Based Inference

The problem of simulator-based models is a significant challenge for Bayesians. Consider some $\mathbb{P}_\theta \in \mathcal{P}_\Theta$ with fixed $\theta \in \Theta$, whose density is the likelihood $p(\cdot \mid \theta)$ and suppose we have a prior $\pi(\theta)$ on the parameter. The corresponding posterior density is given by

$$\pi(\theta \mid x_{1:n}) = \frac{\prod_{i=1}^n p(x_i \mid \theta)\pi(\theta)}{\int_\Theta \prod_{i=1}^n p(x_i \mid \theta)\pi(\theta) d\theta} \propto \prod_{i=1}^n p(x_i \mid \theta)\pi(\theta).$$

The latter is intractable whenever the likelihood $p(\cdot \mid \theta)$ cannot be evaluated pointwise. This has led to the development of simulator-based inference methods (sometimes also called likelihood-free inference methods).

In this paper, we focus on simulator-based models (also called generative models), which are parametric families. This means that $\mathbb{P}_\theta$ can be represented using a distribution $\mathbb{U}$ on a space $\mathcal{U}$ and a simulator $G_\theta : \mathcal{U} \times \mathcal{X}$ so that a sample $y \sim \mathbb{P}_\theta$ from the model can be obtained by first sampling $u \sim \mathbb{U}$, and then applying the simulator $y := G_\theta(u) \in \mathcal{X}$.

Approximate Bayesian Computation ABC algorithms are arguably the most popular family of techniques for tackling Bayesian posteriors of simulator-based models (see Beaumont et al. 2002; Sisson et al. 2018; Beaumont 2019). Most ABC algorithms are a variation of the following steps:

(i) For $b = 1, 2, \ldots B$ and prior π, sample $\theta_b \overset{\text{iid}}{\sim} \pi$;

(ii) For each θ_b, sample m realisations from $\mathbb{P}_{\theta_b}$ (i.e. simulate $u_{1:m} \overset{\text{iid}}{\sim} \mathbb{U}$ and set $y_i^{(b)} = G_{\theta_b}(u_i)\ \forall i$);

(iii) Compare $y_{1:m}^{(b)}$ with the true data $x_{1:n}$ using a discrepancy $D : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$: $D(\mathbb{P}_n, \widehat{\mathbb{P}}_{\theta_b})$ where $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\widehat{\mathbb{P}}_{\theta_b} = \frac{1}{m}\sum_{j=1}^m \delta_{y_j^{(b)}}$.

(iv) Weight θ_b as an approximate sample from the posterior according to $D(\mathbb{P}_n, \widehat{\mathbb{P}}_{\theta_b})$.

The function D is often chosen to be a discrepancy comparing summary statistics of the two datasets, but could also be a probability metric. For example, Bernton et al. (2019) used the Wasserstein distance, whilst Park et al. (2016) used the MMD. Step (iv) is usually implemented by verifying whether the discrepancy is smaller than some threshold ε, accepting the sample θ_b if so, and rejecting it otherwise. This leads to the

following approximation of the standard Bayesian posterior density:

$$\pi_{\text{ABC}}(\theta \mid x_{1:n}) \propto \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \prod_{j=1}^{m} \mathbb{1}_{f D(\mathsf{P}_n, (\mathsf{P}_\theta)_m) < \varepsilon g}(\theta) \, p(y_j \mid \theta) \pi(\theta) dy_1 \ldots dy_m.$$

Smaller values of ε imply an increased computational cost since more parameter samples will be rejected. However, as $\varepsilon \to 0$, the ABC posterior approaches the standard Bayesian posterior $\pi(\theta \mid x_{1:n})$. The latter is usually seen as a desirable property of ABC, but it can lead to poor performance because the Bayesian posterior itself lacks robustness to misspecification.

Although the prototype ABC algorithm above is parallelisable, it will tend to be inefficient because many samples will be rejected whenever ϵ is small. As a result, it is common to use inherently sequential algorithms such as population Monte Carlo (Beaumont et al. 2009) or sequential Monte Carlo with adaptive resampling (Del Moral et al. 2012).

Alternative Approaches There are a number of perhaps less prominent alternative approaches for Bayesian inference for simulators, including Bayesian synthetic likelihoods (Price et al. 2017) and techniques relying on neural density estimation (see e.g. Papamakarios et al. 2016; Papamakarios et al. 2019; Cranmer et al. 2020). Just like ABC, all these approaches are non-robust since they approximate the standard Bayesian posterior.

2.2 Generalised Bayesian Inference (GBI)

To address the robustness concern for standard Bayesian inference, GBI approaches have recently been proposed. In GBI, $l_n : \mathcal{X}^n \times \Theta \to \mathsf{R}$ denotes any (empirical) loss function and $\beta > 0$ a learning rate. Then, the associated GBI posterior's density is

$$\pi_{\text{GBI}}(\theta \mid x_{1:n}) = \frac{\exp f \beta l_n(x_{1:n}, \theta) g \pi(\theta)}{\int_\Theta \exp f \beta l_n(x_{1:n}, \theta) g \pi(\theta) d\theta}.$$

Here, the loss is typically chosen in relation to the likelihood model $p(x_{1:n} \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$. Indeed, it is easy to see that for the standard Bayesian posterior, one chooses $l_n(x_{1:n}, \theta) = -\log p(x_{1:n} \mid \theta)$. From this, it also becomes clear that the Bayes posterior's lack of robustness is intimately related to choosing the loss function $l_n(x_{1:n}, \theta) = -\log p(x_{1:n} \mid \theta)$. To address this, a host of generalised posteriors have been derived by choosing a discrepancy D with desirable robustness properties, and then seeking to find a loss so that $l_n(x_{1:n}, \theta) \propto D(\mathsf{P}, \mathsf{P}_\theta)$ (see Jewson et al. 2018).

GBI for robustness in simulator-based inference has been studied in Pacchiardi et al. (2021) and Schmon et al. (2020). The main drawback of the proposals in both papers are two-fold: firstly, the uncertainty quantification properties of $\pi_{\text{GBI}}(\theta \mid x_{1:n})$ depend on β, which can only be chosen based on rough heuristics (see Wu et al. 2020). Furthermore, the methods require techniques that are computationally more cumbersome: Pacchiardi et al. (2021) use an inherently sequential sampling method based on pseudo-marginal MCMC (Andrieu et al. 2009), and Schmon et al. (2020) relies on standard ABC methods.

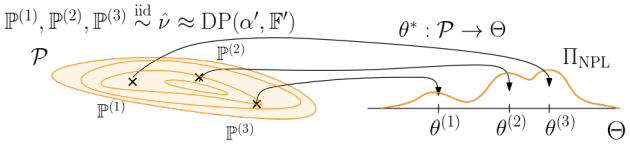


Figure 1: *Sketch of the posterior bootstrap.* Samples from the NPL posterior on Θ are obtained by mapping realisations of the posterior $\hat{\nu}$ through the map θ^*.

In this paper, we perform Bayesian inference under model misspecification through a robust discrepancy like in GBI methods—but without relying on computationally burdensome methods like in ABC.

2.3 Bayesian Nonparametric Learning (NPL)

While the Bayesian nonparametric learning (NPL) framework of Lyddon et al. (2018) was introduced to deal with misspecification, it also possesses attractive computational properties. NPL defines a nonparametric prior directly on the true data-generating process P, which in turn leads to a nonparametric posterior on P. Any posterior on the parameter space Θ of P_Θ induced by NPL thus derives from this nonparametric posterior on P. Note that this is different, but closely related to, standard Bayesian inference where conditioning occurs at the level of parameters as opposed to the level of the data-generating process.

Following Lyddon et al. (2018) and Fong et al. (2019), we use a Dirichlet Process (DP) prior on the data-generating process: $\mathsf{P} \sim \text{DP}(\alpha, \mathsf{F})$. Here, $\alpha > 0$ is a concentration parameter and $\mathsf{F} \in P$ a centering measure. Given $x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}$, it follows by conjugacy that

$$\mathsf{P} \mid x_{1:n} \sim \text{DP}\left(\alpha', \mathsf{F}'\right), \quad \alpha' = \alpha + n, \quad \mathsf{F}' = \frac{\alpha}{\alpha+n}\mathsf{F} + \frac{n}{\alpha+n}\mathsf{P}_n \tag{1}$$

where $\mathsf{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and δ_x is a Dirac measure at $x \in \mathcal{X}$. The size of α quantifies our confidence in the quality of the prior centering measure F and regulates the influence of the prior on the posterior. Accordingly, the limiting case of $\alpha = 0$ corresponds to a non-informative prior and results in the posterior $\text{DP}(n, \mathsf{P}_n)$. Hence, unlike the hyperparameter β in GBI, the heuristic $\alpha = 0$ has a clear interpretation

and yields reliable uncertainty quantification (see Fong et al. 2019; Knoblauch et al. 2020; Galvani et al. 2021).

The posterior on the data-generating process is readily translated into a posterior on a parameter space Θ. Suppose we know the *true* data-generating process P and a loss $l : X \quad \Theta \; ! \; \mathsf{R}$. Clearly, having access to P is equivalent to having access to infinitely many data points. Consequently, no Bayesian uncertainty is needed, and one can simply compute

$$\theta_l(\mathsf{P}) := \arg\inf_{\theta 2 \Theta} \mathsf{E}_{X \; \mathsf{P}}[l(X, \theta)]. \tag{2}$$

In practice, we do *not* know the true data-generating process. However, we have a posterior belief over it, so that the simple push-forward measure $(\theta_l)_{\#}(\mathrm{DP}(\alpha^0, \mathsf{F}^0))$ gives a posterior on Θ denoted by Π_{NPL}. This is illustrated in Figure 1.

For the reader who is not familiar with pushforward distributions, Π_{NPL} is best understood through the following sampling mechanism to obtain independent realisations from Π_{NPL}. At iteration j,

1. Sample $\mathsf{P}^{(j)}$ from the posterior DP in (1);
2. Compute $\theta^{(j)} = \theta_l(\mathsf{P}^{(j)})$ using (2).

This procedure is trivially parallelisable and does not discard any samples. It therefore overcomes the computational inefficiencies of ABC-based methods, whose rejection rate for samples is typically quite high—especially as ε is moved closer to zero.

Exact sampling from a DP as in step 1 above is usually infeasible. The most common approximation is the truncated stick-breaking procedure (Sethuraman 1994). This procedure in turn can be approximated by the Dirichlet approximation of the stick breaking process (Muliere et al. 1996; Ishwaran et al. 2002). In our case, this leads to

$$\begin{aligned} \tilde{x}^{(j)}_{1:T} &\overset{\text{iid}}{\sim} \mathsf{F}, \quad (w^{(j)}_{1:n}, \tilde{w}^{(j)}_{1:T}) \quad \mathrm{Dir}(1, \ldots, 1, \tfrac{\alpha}{T}, \ldots, \tfrac{\alpha}{T}). \\ \mathsf{P}^{(j)} &\quad \sum_{i=1}^n w^{(j)}_i \delta_{x_i} + \sum_{k=1}^T \tilde{w}^{(j)}_k \delta_{\tilde{x}^{(j)}_k}. \end{aligned} \tag{3}$$

$\hat{\nu}$ denotes the probability measure on P defined by (3), so that $\mathsf{P} = \sum_{i=1}^n w^{(j)} \delta_{x_i} + \sum_{k=1}^T \tilde{w}^{(j)}_k \delta_{\tilde{x}^{(j)}_k} \quad \hat{\nu}$.

It is generally not possible to obtain the minimiser in (2) in closed form, and this objective may not even be convex. This necessitates the use of numerical optimisers like stochastic gradient descent, so that step 2 above is typically only performed approximately.

In the current paper, we use the NPL framework with a loss l that corresponds to the Maximum Mean Discrepancy (MMD)—a robust discrepancy popular in GBI methods (Chérief-Abdellatif et al. 2020; Pacchiardi et al. 2021). This implies that computationally, the second step in our NPL algorithm amounts to minimum distance estimation as introduced in Briol et al. (2019).

2.4 Minimum Distance Estimation with Robust Discrepancies

Since NPL depends on a minimisation step, we will revisit a branch of frequentist statistics whose theory and methodology we extensively draw on for our algorithm: Minimum distance estimators (MDEs) (Parr et al. 1980). MDEs are a frequentist approach to parameter estimation. Given a discrepancy $D : \mathsf{P} \quad \mathsf{P} \; ! \; \mathsf{R}_+$, the MDE is

$$\hat{\theta}_n := \arg\inf_{\theta 2 \Theta} D(\mathsf{P}_n, \mathsf{P}_\theta). \tag{4}$$

MDEs can be robust to model misspecification when the underlying discrepancy is chosen with this property in mind. A common choice of discrepancy D is integral pseudo-probability metrics (Müller 1997):

$$\mathrm{IPM}(\mathsf{P}, \mathsf{Q}) = \sup_{f 2 F} \left| \int_X f(x) \, \mathsf{P}(dx) \quad \int_X f(y) \, \mathsf{Q}(dy) \right|.$$

IPMs can be thought of as comparing a family of summary statistics indexed by the class F. There are two common IPMs in the context of simulators:

(i) Wasserstein Distance Let $c : X \quad X \; ! \; [0, 1)$ be a metric and let $p \quad 1$. Furthermore, let $P_{c,p} = f\mathsf{P} \; 2 \; P \; j \; \int_X c^p(x, y) \, \mathsf{P}(dx) < 1 \; 8y \; 2 \; X g$. Then the Wasserstein distance is a map $W : P_{c,p} \quad P_{c,p} \; ! \; \mathsf{R}_+$ obtained by considering an IPM with $F_W := f f : X \; ! \; \mathsf{R} \; j \; 8x, y \; 2 \; X, j f(x) \quad f(y) j \quad c(x, y) g$. The Wasserstein distance was used for MDE by Bassetti et al. (2006) and Bernton et al. (2017) and, as previously mentioned, was used for ABC by Bernton et al. (2019).

(ii) Maximum Mean Discrepancy Let H_k be a reproducing kernel Hilbert space (RKHS) with kernel $k : X \quad X \; ! \; \mathsf{R}$ and norm $k \quad k_{H_k}$. Let $P_k = f\mathsf{P} \; 2 \; P \; j \; \int_X \sqrt{k(x, x)} \, \mathsf{P}(dx) < 1 g$. Then the MMD is a map $\mathrm{MMD} : P_k \quad P_k \; ! \; \mathsf{R}_+$ obtained by considering an IPM with $F_{\mathrm{MMD}} := f f : X \; ! \; \mathsf{R} \; j \; k f k_{H_k} \quad 1 g$. When the kernel k is characteristic (Sriperumbudur et al. 2010), the MMD is a probability metric. A common characteristic kernel is the Gaussian kernel $k(x, x^0) = \exp(\; k x \quad x^0 k_2^2 / (2l^2))$ where $l > 0$.

The MMD was first considered for MDE by Briol et al. (2019) and was further explored in Chérief-Abdellatif et al. (2019), Alquier et al. (2020), and Niu et al. (2021). It is closely related to the use of MMD in generative adversarial networks (Dziugaite et al. 2015; Li et al. 2015), and was used for goodness-of-fit testing with composite hypotheses in Key et al. (2021). As previously mentioned, it has also been used for ABC (Park et al. 2016) and GBI (Chérief-Abdellatif et al. 2020; Pacchiardi et al. 2021).

The Wasserstein distance and MMD are both popular in the context of simulators because they can both be computed or approximated for empirical mea-

sures. One can therefore use the simulator to sample m realisations from P_θ, then use $D(\widehat{\mathsf{P}}_{\theta_b}, \mathsf{P}_n)$ as an approximation of $D(\mathsf{P}_\theta, \mathsf{P}_n)$ in (4). In this context, the MMD has the advantage over the Wasserstein distance in that it is a robust distance (Briol et al. 2019; Chérief-Abdellatif et al. 2019), and can be computed in quadratic, rather than cubic, time in n and m.

3 METHODOLOGY

Our paper proposes *Bayesian nonparametric learning with the MMD* for robust and scalable inference in simulator models. Compared to ABC, our method has two main computational advantages: it does not discard any samples, and it is trivially parallelisable. Furthermore, we will show in Section 4 that the approach inherits the robustness properties of both the NPL framework and the MMD; and therefore satisfies a number of desirable properties—including finite-sample generalisation bounds, robustness guarantees, and frequentist consistency. Notably, all of these guarantees hold under model misspecification, highlighting the approach's usefulness for simulator models of complex data generating mechanisms.

Assume we have observed data $x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}$ and are interested in inference with a parametric family $\boldsymbol{P}_\Theta$ of simulator-based models. Our approach is to use the NPL framework with the loss given by the kernel scoring rule l_k (Eaton 1982; Dawid 2007):

$$l_k(x, \theta) = k(x, x) \quad 2\int_{\mathcal{X}} k(x, y)\, \mathsf{P}_\theta(dy) + \int_{\mathcal{X}} \int_{\mathcal{X}} k(y, z)\, \mathsf{P}_\theta(dy)\, \mathsf{P}_\theta(dz).$$

For this loss, the minimiser in (2) becomes the MMD estimator of Briol et al. (2019) given by:

$$\begin{aligned} \theta^*(\mathsf{P}) &:= \arg\inf_{\theta \in \Theta} \mathrm{MMD}^2(\mathsf{P}, \mathsf{P}_\theta) \\ &= \arg\inf_{\theta \in \Theta} \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y)\, \mathsf{P}_\theta(dx)\, \mathsf{P}_\theta(dy) \\ &\quad \tfrac{2}{n} \sum_{i=1}^n \int_{\mathcal{X}} k(x_i, x)\, \mathsf{P}_\theta(dx). \end{aligned} \tag{5}$$

This objective can easily be approximated by sampling from both measures and using a U-statistic. The same holds for its gradient, which naturally leads to a stochastic gradient descent algorithm. Full details are provided in Appendix C.

Pseudo code of our approach is given in Algorithm 1, and we call the resulting procedure *MMD posterior bootstrap* to pay homage to the fact that in the limiting case of $\alpha \to 0$, it is computationally similar to the Generalised Bayesian Bootstrap of Lyddon et al. (2019). From a practical standpoint, this limiting case is particularly interesting because it eliminates dependence on the centering measure F. This corresponds to a non-informative prior (see Fong et al. 2019; Knoblauch et al. 2020), and reflects our uncertainty about the model in a misspecified setting.

Algorithm 1: MMD Posterior Bootstrap

input: $x_{1:n}$, T, B, α, F, U, G_θ.
for j 1 **to** B **do**
  Sample $\tilde{x}^{(j)}_{1:T} \overset{\text{iid}}{\sim} \mathsf{F}$ and
  $(w^{(j)}_{1:n}, \tilde{w}^{(j)}_{1:T})$ Dir $\left(1, \ldots, 1, \frac{\alpha}{T}, \ldots, \frac{\alpha}{T}\right)$.
  Set $\mathsf{P}^{(j)} = \sum_{i=1}^n w_i^{(j)} \delta_{x_i} + \sum_{k=1}^T \tilde{w}_k^{(j)} \delta_{\tilde{x}_k^{(j)}}$.
  Obtain $\theta^{(j)} = \theta\ (\mathsf{P}^{(j)})$ using numerical optimisation.
end
return Posterior bootstrap sample $\theta^{(1:B)}$

4 THEORY

Before presenting our experiments, we provide a theoretical study for which we impose:

Standing Assumption 1. $|k(x, y)| \quad 1 \quad \forall x, y \in \mathcal{X}$.

While the kernel is required to be bounded, the choice of upper bound 1 is without loss of generality, since we only consider the minimiser of the MMD, which does not depend on the bound's magnitude. Note that the condition also ensures that F, P , and any element of $\boldsymbol{P}_\Theta$ are in $\boldsymbol{P}_k$. While no other assumption on the kernel is needed for our theory, it is desirable for the kernel to be characteristic—as this guarantees that the MMD is a metric on $\boldsymbol{P}_k$ and hence that P can be recovered in well-specified models.

Since the conjugacy of the DP posterior in the NPL setting relies on the assumption that the data are i.i.d, we also inherit this requirement:

Standing Assumption 2. $x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}$.

Our results provide the first generalisation, robustness, and consistency guarantees for NPL posteriors. Our first result is a generalisation bound in terms of the MMD expected under $\hat{\nu}$ (Theorem 3). Beyond that, we also prove that consistency in the frequentist sense (Theorem 4), and robustness to outliers (Theorem 5). A particular characteristic is that all of these results hold for the misspecified setting where $\mathsf{P} \notin \boldsymbol{P}_\Theta$. The well-specified counterparts are Corollaries 12, 13, and 15 in Appendix B, and are obtained by noting that in this case $\inf_{\theta \in \Theta} \mathrm{MMD}(\mathsf{P}_\theta, \mathsf{P}) = 0$.

4.1 Generalisation Error

First, we bound the generalisation error of our procedure, i.e. the error we expect based on unseen data

from the true data-generating mechanism. Here, the notion of error considered is the MMD to be expected under the (approximated) posterior, which is given by $\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})})\right]$. We therefore bound the expected value of this quantity under unseen data from the data-generating mechanism:

Theorem 3.

$$\mathsf{E}_{x_{1:n}\ \mathrm{iid}\ \mathsf{P}}\ \left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})})\right]\right] \quad \inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_\theta) + 2/\sqrt{n} + \sqrt{4\sqrt{\alpha(1+\alpha)}/\sqrt{(\alpha+n)(\alpha+n+1)}}.$$

Since the expectation over $\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})})\right]$ is trivially lower-bounded by $\inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_\theta)$, this result tells us that, in expectation, the difference between these two quantities is at most $2/\sqrt{n} + 4\sqrt{\alpha(1+\alpha)}/\sqrt{(\alpha+n)(\alpha+n+1)}$. Both terms vanish as $n\ !\ 1$, with the overall rate being a $\sqrt{n}$ rate. This is the same rate as for the MMD estimators of Briol et al. (2019) (Theorem 1), as well as the MMD-based GBI of Chérief-Abdellatif et al. (2019) (Theorem 3.1).

4.2 Posterior Consistency

To deepen our analysis, we consider consistency in the frequentist sense. This guarantees that the posterior contracts around the optimal value $\theta\ (\mathsf{P}\ )$ as $n\ !\ 1$. For standard Bayesian inference with posterior measure Π_n (whose density was previously denoted $\pi(\ j\ x_{1:n})$) defined on Θ directly, and where our model is misspecified so that $9C > 0$ such that $\inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) = C$, this would amount to

$$\Pi_n\left(\theta\ 2\ \Theta : \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) > C + \tfrac{M_n}{n^{1/2}}\right)^{\ n\ !\ 1}\ 0 \quad (6)$$

for a sequence $M_n\ !\ +1$ so that $M_n n^{\frac{1}{2}}\ !\ 0$ as $n\ !\ 1$. In other words, the posterior measure over regions of Θ that induces large values for $\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ )$ goes to C as we obtain more data so that it must ultimately concentrate around increasingly small neighbourhoods of optimal MMD-minimising values for θ.

In our case, the posterior Π_{NPL} is defined implicitly: given the function $\theta\ (\mathsf{P})$ and the approximate posterior $\hat{\nu}$ on P constructed via equation (3), our posterior on θ is obtained by the push-forward operation. Thus, the equivalent statement for our case concerns $\hat{\nu}$:

Theorem 4. *Suppose our model is misspeci ed so that for some $C > 0$ we have $\inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) = C$. Then, we have that for any $M_n\ !\ +1$:*

$$\hat{\nu}\left(\mathsf{P}\ 2\ \mathsf{P} : \mathrm{MMD}(\mathsf{P}_{\theta\ (\mathsf{P})},\mathsf{P}\ ) > C + \tfrac{M_n}{n^{1/2}}\right)^{\ n\ !\ 1}\ 0.$$

4.3 Robustness to Outliers

To assess robustness against the presence of outliers in the dataset, we consider Huber's contamination model (Huber 1992). In this setting, a proportion $1\ \ \varepsilon$ of the observed data is generated from the distribution of interest $\tilde{\mathsf{P}}$, and the rest follows a contaminating noise distribution; i.e. $\mathsf{P}\ = (1\ \ \epsilon)\tilde{\mathsf{P}} + \epsilon\mathsf{Q}$ for $\tilde{\mathsf{P}},\mathsf{Q}\ 2\ \mathcal{P}$ and $\epsilon\ 2\ [0,1]$. Here, Q is the contaminant, and so the goal is to place most posterior mass on values of θ for which $\mathsf{P}_\theta\ \ \tilde{\mathsf{P}}$, where closeness is measured via the MMD.

Corollary 5. *Suppose* $\mathsf{P}\ = (1\ \ \epsilon)\tilde{\mathsf{P}} + \epsilon\mathsf{Q}$. *Then*

$$\mathsf{E}_{x_{1:n}\ \mathrm{iid}\ \mathsf{P}}\ \left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}\left(\tilde{\mathsf{P}},\mathsf{P}_{\theta\ (\mathsf{P})}\right)\right]\right] \quad \inf_{\theta 2\Theta}\mathrm{MMD}(\tilde{\mathsf{P}},\mathsf{P}_\theta) + 4\epsilon + 2/\sqrt{n} + \sqrt{4\sqrt{\alpha(1+\alpha)}/\sqrt{(\alpha+n)(\alpha+n+1)}}.$$

Similarly to the generalisation bound discussed above, the rate at which this bound goes to zero is $\max\{n^{-\frac{1}{2}},\epsilon\}$. Since there are at most $\lceil n\epsilon\rceil$ contaminated data points in a dataset, the maximum number of outliers the dataset can have while maintaining the same rate is of order $\sqrt{n}$, which agrees with Chérief-Abdellatif et al. (2019), Corollary 3.4, who studied the frequentist minimum MMD estimator.

5 EXPERIMENTS

We now study the performance of our method using three examples. Throughout, we use the Gaussian kernel, which satisfies Standing Assumption 1. We also use a non-informative prior by setting $\alpha = 0$. Further experimental details and results are reported in Appendix C. Appendix D provides additional experiments which examine sensitivity to hyperparameters and provide comparison of our method with the MMD-Bayes method in Pacchiardi et al. (2021). We further consider an example of misspecification which is not based in a contamination model by wrongly fitting a Gaussian location model to Cauchy distributed data. The code for all experiments can be found at `https://github.com/haritadell/npl_mmd_project.git`.

5.1 Gaussian Location Model

We start by considering a toy example, the Gaussian location model. While the likelihood of this model is available, we treat it as a simulator to study the properties of our proposed method. We take $\mathsf{P}_\theta = \mathcal{N}(\theta, I_{d\ \ d})$ and use a true data generating process $\mathsf{P}\ = (1\ \ \epsilon)\,\mathsf{P}_{\theta_0} + \epsilon\,\mathsf{P}_{\theta^0}$, with $\theta_0 = (1,\ldots,1)\ 2\ \mathsf{R}^d$ and $\theta^0 = (20,\ldots,20)\ 2\ \mathsf{R}^d$ in $d = 4$. We assess robustness by considering both the well-specified case $\epsilon = 0$ and the case $\epsilon = 0.1$ and $n = 200$ realisations from $\mathsf{P}\ $.

Our simulation study will illustrate how robustness is inherited from both the NPL framework, which is more robust to model misspecification than standard Bayes or ABC, and the MMD being more robust than al-

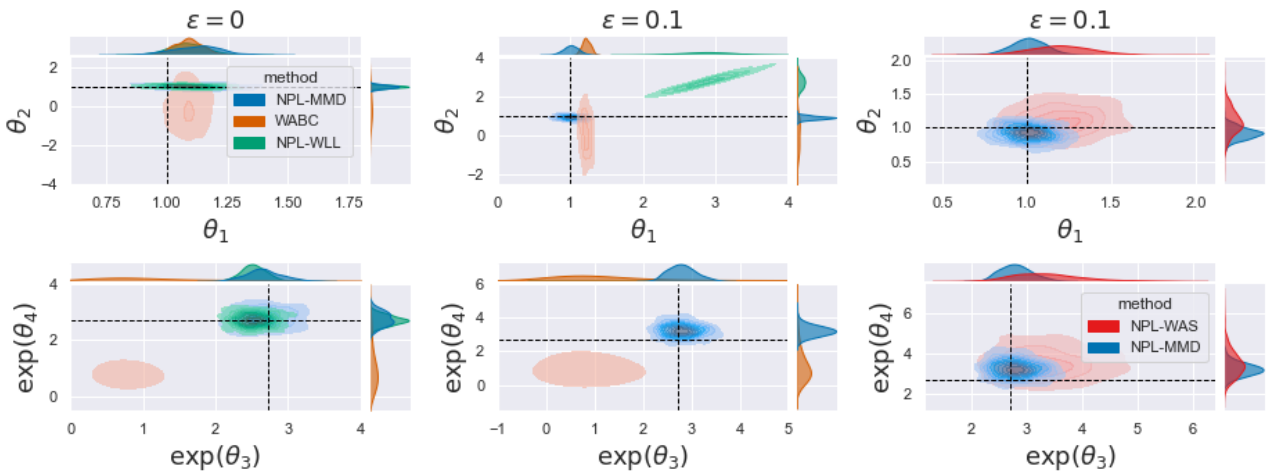


Figure 2: Posterior marginal distributions for the Gaussian location model in $d = 4$. The true parameter θ_0 is indicated by dotted lines. (Left & middle) Comparison against WABC and NPL-WLL methods for the well-specified case and $\epsilon = 0.1$. (Right) Comparison against NPL with the Wasserstein distance for $\epsilon = 0.1$. We note that in the low middle panel, the NPL-WLL method is not visible as the samples lie significantly away from θ_0.

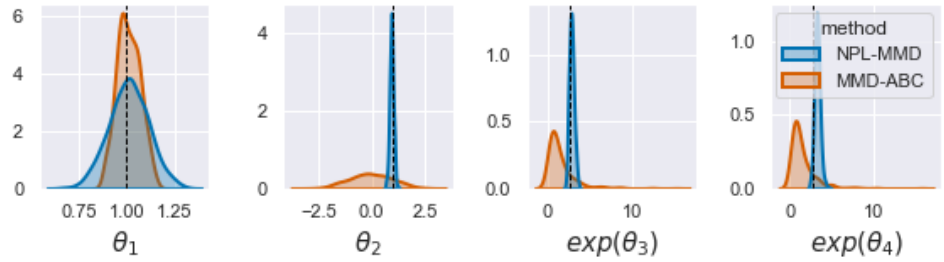


Figure 3: Marginal densities of the posterior obtained by NPL-MMD and ABC with the MMD for the Gaussian location model in $d = 4$ for $\epsilon = 0.1$. The true parameter θ_0 is indicated by dotted lines.

ternative losses such as the Wasserstein distance or negative log-likelihood.

Wasserstein ABC We first compare our method against the Wasserstein ABC (WABC) (Bernton et al. 2019) algorithm, which uses Sequential Monte Carlo (SMC). We compare against WABC since it is a popular algorithm which, unlike standard ABC, does not require hand-crafting of summary statistics. As observed in Figure 2, NPL-MMD outperforms WABC in both the well-specified and misspecified settings. The former is explained by the fact that the Wasserstein distance exhibits poor sample complexity for $d > 1$ (Fournier et al. 2015), whereas the MMD can be estimated at a $\sqrt{n}$ rate. The latter is explained by the fact that the WABC is an approximation of the exact Bayesian posterior, which is not robust.

NPL with Wasserstein distance Secondly, we consider the NPL framework using the Wasserstein distance (NPL-WAS) such that for $W : \mathsf{P}^2 \; ! \; \mathsf{R}_{\;0}$ the 2-Wasserstein distance, $\theta_W(\mathsf{P}) := \arg\inf_{\theta 2 \Theta} W(\mathsf{P}, \mathsf{P}_\theta)$. We use the `POT` package (Flamary et al. 2021) for approximating the Wasserstein distance and the *Powell* optimiser from `Scipy` (Virtanen et al. 2020). We focus specifically on the misspecified setting, and notice that NPL-MMD outperforms the NPL-WAS in that case. This clearly demonstrates the advantage of using a robust loss function, even when using a robust inference framework such as NPL.

NPL with log-likelihood Thirdly, the availability of the likelihood in this toy problem allows for comparison with the original NPL posterior bootstrap using $l(x, \theta) = \frac{1}{2\pi} \exp\frac{1}{2}(x \quad \theta)^2$ as in Lyddon et al. (2018), which we call the weighted log-likelihood NPL (NPL-WLL). Figure 2 shows that NPL-WLL and NPL-MMD perform similarly in the well-specified case, but NPL-MMD significantly outperforms NPL-WLL in the misspecified case. This is once again due

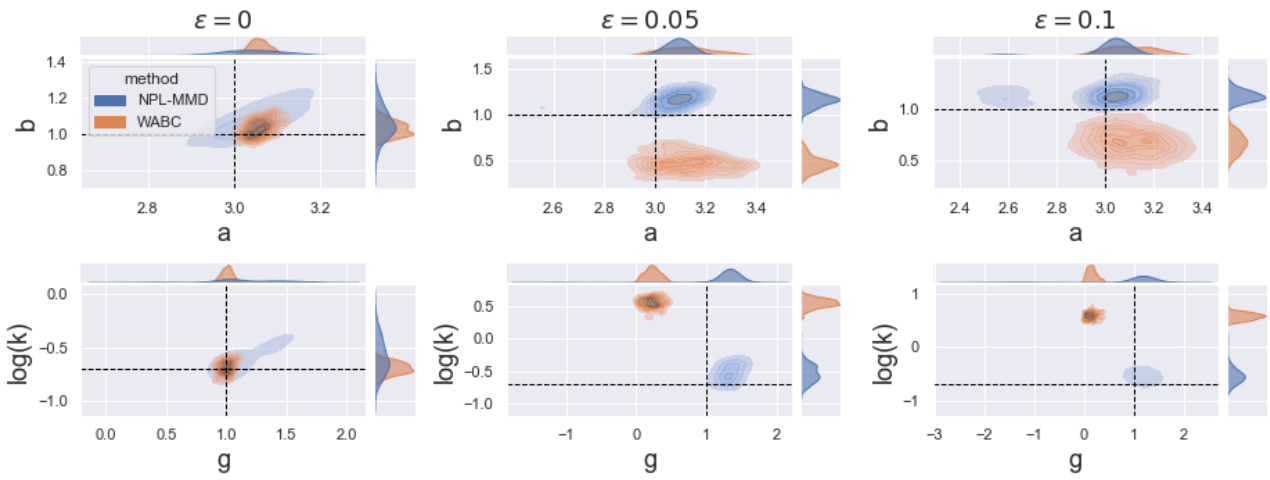


Figure 4: Posterior marginal distributions for θ in the univariate g-and-k model for an increasing percentage of outliers present in the dataset. The true parameter θ_0 is indicated by the dotted lines.

to the fact that the MMD is a robust loss, whereas the negative log-likelihood is not.

ABC with the MMD Finally, we further explore ABC method with the SMC samplers considered in Bernton et al. (2019), this time using the MMD instead of the Wasserstein distance. The ABC with the MMD has also been previously explored in Park et al. (2016). The posterior marginals are visualised in Figure 3 for $\epsilon = 0.1$. This figure clearly demonstrates the advantage of the NPL framework over ABC.

5.2 Simulator-based Models

We now consider two more complex numerical examples, for which likelihood-based inference is impossible. We compare solely against the Wasserstein ABC for simplicity by directly using the experimental setup in Bernton et al. (2019).

G-and-k Distribution Model First, consider $G_\theta : [0,1]^2 \; ! \; \mathsf{R}$ with $\theta = (a, b, g, k)$ where

$$G_\theta(u) = a + b\left(1 + 0.8\frac{1 \;\; \exp(\;\; gz(u))}{1+\exp(\;\; gz(u))}\right)(1 + z(u)^2)^k z(u),$$

$z(u) = \sqrt{\;\; 2\log(u_1)}\cos(2\pi u_2)$ and $\mathsf{U} = \mathrm{Unif}([0,1]^2)$. Parameters a, b, g and k control the location, scale, skewness and kurtosis respectively. For computational convenience, we reparametrise the last parameter by setting $k^0 = \exp(k)$. Although P_θ is one-dimensional, it is a popular baseline for simulator-based models (Prangle 2020) because of the challenge of inferring the four parameters simultaneously. It has also been used extensively in applications; for example to model the price of AirBnB rentals (Rodrigues et al. 2020), the air pollution (Rayner et al. 2002) or for non-life insurance modelling (Peters et al. 2016).

Our data consists of $n = 2^{11}$ realisations from $\mathsf{P} = (1 \quad \epsilon)\,\mathsf{P}_{\theta_0} + \epsilon\,\mathsf{Q}$ where P_{θ_0} denotes the g-and-k with $\theta_0 = (3, 1, 1, \quad \log(2))$, and Q is the shifted distribution $\mathsf{Q} = \mathsf{P}_{\theta_0} \quad 50$ with an equal number of points shifted to either direction. The resulting posteriors are shown in Figure 4 displayed as bivariate plots. The WABC method appears more sensitive to contamination in the dataset—in contrast to the NPL-MMD, which concentrates significantly closer to the true parameter values for an increasing proportion of outliers. This is particularly the case for the last two parameters g and $\log(k)$ which are well-known to be more challenging to estimate.

Toggle Switch Model Finally, we consider the toggle-switch model arising in Systems Biology (Bonassi et al. 2011; Bonassi et al. 2015). This is a dynamic model used to study cellular networks; more precisely, the network describes the interaction of two genes u and v over time. The simulator is too complex to include in the main text, but is given in Appendix C. The data is one-dimensional, but the model has 7 parameters and the latent space is 601-dimensional.

We consider inference on θ for $n = 2000$ data points simulated from the toggle-switch model with true parameter $\theta_0 = (22, 12, 4, 4.5, 325, 0.25, 0.15)$ in which 10% of the data have some added Cauchy noise of location parameter 0 and scale parameter 10. Such noise can be interpreted as measurement error in the collection of data. The posterior marginal distributions for θ are shown in Figure 5, and indicate that the NPL-MMD method is successful in concentrating around θ_0

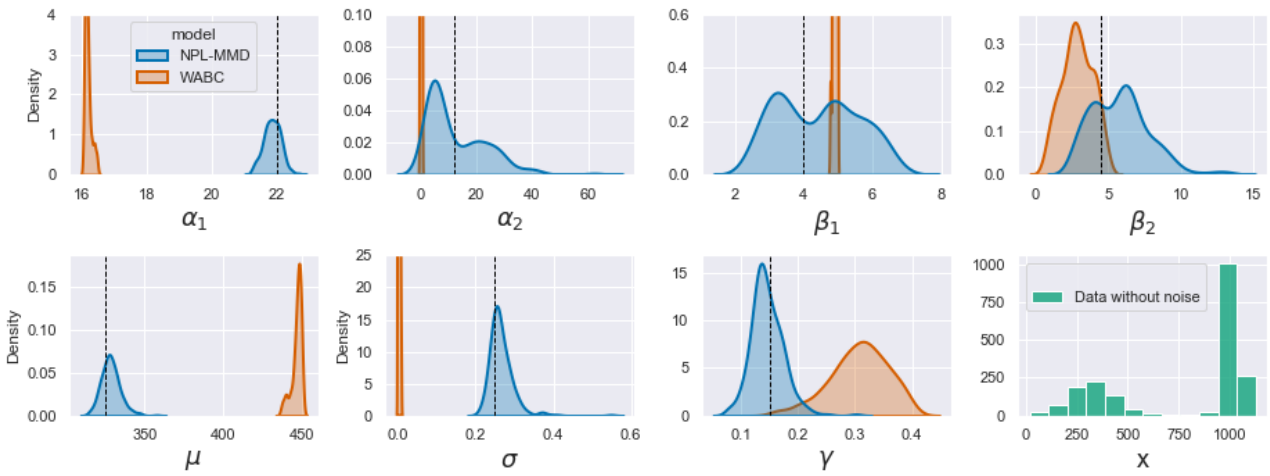


Figure 5: Posterior marginal distributions for θ in the toggle switch model with 10% of noisy data in the dataset. Bottom right figure presents 2000 samples from the well-specified toggle switch model.

despite the Cauchy noise.

5.3 Computational Time

We provide the computational time recorded for each experiment over multiple independent runs. The Gaussian location and G-and-k experiments were repeated 10 times and the Toggle-switch was repeated 5 times due to the higher computational cost. For each run, a new dataset was generated and B posterior samples were obtained for θ. The average time for the generation of B samples is recorded for each method. Our method is compatible with the use of GPU so in Table 1 we provide the average time recorded when run in `Google Colaboratory Pro`.

Table 1: Average Clock Time In Seconds.

Method	**Gaussian**		**G-and-k**		**Toggle-S**	
NPL-MMD	1.09	10^2	3.02	10	7.42	10^3
WABC	3.34	10^3	1.05	10^2	3.08	10^4

6 CONCLUSION

Our paper proposes the very first posterior bootstrap method in the simulator setting. Unlike ABC, our method does not discard any samples and can run in parallel—leading to a substantially decreased computational burden. Further, the approach is based on MMD estimators, and therefore inherits both the robustness of the NPL framework and that of the MMD. We support this claim with three theoretical results that hold even in the presence of misspecification. The results include a generalisation bound, a robustness guarantee, and a consistency result. We verified the practical utility of our theory through deploying it on three models, which highlighted both robustness and computational advantages of our method.

A particular strength of the method is that no assumption is required on $\mathcal{X}$. Hence, the results apply directly to any set $\mathcal{X}$ on which one can define kernels (e.g. graphs, strings, or discrete spaces).

In future work, we would like to tackle the challenges posed by the optimisation step. Specifically, since the MMD-objective is usually non-convex, a particularly interesting question would be if the kernel can be used to enforce a more well-behaved objective. Another direction for improvement is the cost of computing the MMD minimiser. While naively, this scales quadratically in the number of observations, it could be reduced to linear time using approaches like in Lemma 14 of Gretton et al. (2012).

Acknowledgements

CD is funded by EPSRC grant [EP/T51794X/1] as part of the Warwick CDT in Mathematics and Statistics. JK is funded through the Biometrika Fellowship courtesy of the Biometrika Trust. TD acknowledges support from a UKRI Turing AI Fellowship [EP/V02678X/1]. TD and FXB were supported by the Lloyd's Register Foundation Programme on Data-Centric Engineering and The Alan Turing Institute under the EPSRC grant [EP/N510129/1]. The authors thank all five reviewers for their useful comments as well as all reviewers of a preliminary version of the paper presented in the NeurIPS 2021 workshop "Your Model is Wrong: Robustness and misspecification in probabilistic modeling".

References

Alquier, P. and Gerber, M. (2020). "Universal robust regression via maximum mean discrepancy". In: *Arxiv:2006.00840* 1.

Andrieu, C. and Roberts, G. O. (2009). "The pseudo-marginal approach for efficient Monte Carlo computations". In: *Annals of statistics* 37.2, pp. 697–725.

Bassetti, F., Bodini, A., and Regazzini, E. (2006). "On minimum Kantorovich distance estimators". In: *Statistics and probability letters* 76, pp. 1298–1302.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). "Adaptive approximate Bayesian computation". In: *Biometrika* 96.4, pp. 983–990. ISSN: 14643510.

Beaumont, M. A. (2010). "Approximate bayesian computation in evolution and ecology". In: *Annual review of ecology, evolution, and systematics* 41, pp. 379–406.

Beaumont, M. A. (2019). "Approximate Bayesian computation". In: *Annual review of statistics and its application* 6, pp. 379–403.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). "Approximate bayesian computation in population genetics". In: *Genetics* 162.4, pp. 2025–2035.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*. Vol. 405. John Wiley & Sons.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2017). "Inference in generative models using the Wasserstein distance". In: *Information and inference: a journal of the ima* 8.4, pp. 657–676.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). "Approximate bayesian computation with the wasserstein distance". In: *Journal of the royal statistical society: series b (statistical methodology)* 81.2, pp. 235–269.

Bharti, A., Briol, F.-X., and Pedersen, T. (2021). "A general method for calibrating stochastic radio channel models with kernels". In: *Ieee transactions on antennas and propagation*.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). "A general framework for updating belief distributions". In: *Journal of the royal statistical society. series b, statistical methodology* 78.5, p. 1103.

Bonassi, F. V. and West, M. (2015). "Sequential monte carlo with adaptive weights for approximate bayesian computation". In: *Bayesian analysis* 10.1, pp. 171–187.

Bonassi, F. V., You, L., and West, M. (2011). "Bayesian learning from marginal data in bionetwork models". In: *Statistical applications in genetics and molecular biology* 10.1.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., and Wanderman-Milne, S. (2020). "Jax: composable transformations of python+ numpy programs, 2018". In: *Url http://github. com/google/jax* 4, p. 16.

Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. (2019). "Statistical inference for generative models with maximum mean discrepancy". In: *Arxiv:1906.05944*.

Catoni, O. (2007). "Pac-bayesian supervised classification: the thermodynamics of statistical learning". In: *Institute of mathematical statistics lecture notes | monograph series 56*.

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). "Finite sample properties of parametric mmd estimation: robustness to misspecification and dependence". In: *Arxiv preprint arxiv:1912.05737*.

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). "Mmd-bayes: robust bayesian estimation via maximum mean discrepancy". In: *Symposium on advances in approximate bayesian inference*. PMLR, pp. 1–21.

Chernozhukov, V. and Hong, H. (2003). "An MCMC approach to classical estimation". In: *Journal of econometrics* 115.2, pp. 293–346.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). "The frontier of simulation-based inference". In: *Proceedings of the national academy of sciences of the united states of america* 117.48.

Dawid, A. P. (2007). "The geometry of proper scoring rules". In: *Annals of the institute of statistical mathematics* 59.1, pp. 77–93.

Del Moral, P., Doucet, A., and Jasra, A. (2012). "On adaptive resampling strategies for sequential Monte Carlo methods". In: *Bernoulli* 18.1, pp. 252–278.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). "Training generative neural networks via maximum mean discrepancy optimization". In: *Proceedings of the thirty- rst conference on uncertainty in arti cial intelligence*, pp. 258–267.

Eaton, M. L. (1982). "A method for evaluating improper prior distributions". In: *Statistical decision theory and related topics iii*, pp. 329–352.

Flamary, R. et al. (2021). "Pot: python optimal transport". In: *Journal of machine learning research* 22.78, pp. 1–8.

Fong, E., Lyddon, S., and Holmes, C. (2019). "Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap". In: *International conference on machine learning*. PMLR, pp. 1952–1962.

Fournier, N. and Guillin, A. (2015). "On the rate of convergence in wasserstein distance of the empirical measure". In: *Probability theory and related elds* 162.3-4, pp. 707–738. ISSN: 1432-2064. DOI: 10.1007/s00440-014-0583-7.

Frazier, D. T., Robert, C. P., and Rousseau, J. (2020). "Model misspecification in ABC: consequences and diagnostics". In: *Journal of the royal statistical society b: statistical methodology* 82.2, pp. 421–444.

Galvani, M., Bardelli, C., Figini, S., and Muliere, P. (2021). "A bayesian nonparametric learning approach to ensemble models using the proper bayesian bootstrap". In: *Algorithms* 14.1, p. 11.

Ghosh, A. and Basu, A. (2016). "Robust bayes estimation using the density power divergence". In: *Annals of the institute of statistical mathematics* 68.2, pp. 413–437.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). "A kernel two-sample test". In: *The journal of machine learning research* 13.1, pp. 723–773.

Grünwald, P. and Van Ommen, T. (2017). "Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it". In: *Bayesian analysis* 12.4, pp. 1069–1103.

Grünwald, P. (2011). "Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity". In: *Journal of machine learning research* 19, pp. 397–419.

Grünwald, P. (2012). "The safe Bayesian". In: *International conference on algorithmic learning theory*. Springer, pp. 169–183.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., and Louppe, G. (2021). "Averting a crisis in simulation-based inference". In: *Arxiv:2110.06581*.

Hooker, G. and Vidyashankar, A. N. (2014). "Bayesian model robustness via disparities". In: *Test* 23.3, pp. 556–584.

Huber, P. J. (1992). "Robust estimation of a location parameter". In: *Breakthroughs in statistics*. Springer, pp. 492–518.

Ishwaran, H. and Zarepour, M. (2002). "Exact and approximate sum representations for the dirichlet process". In: *Canadian journal of statistics* 30.2, pp. 269–283.

Jewson, J., Smith, J. Q., and Holmes, C. (2018). "Principled Bayesian minimum divergence inference". In: *Entropy* 20.6, p. 442.

Key, O., Fernandez, T., Gretton, A., and Briol, F.-X. (2021). "Composite goodness-of-fit tests with kernels". In: *Arxiv:2111.10275*.

Kingma, D. P. and Ba, J. (2014). "Adam: a method for stochastic optimization". In: *Arxiv preprint arxiv:1412.6980*.

Knoblauch, J., Jewson, J., and Damoulas, T. (2019). "Generalized variational inference: three arguments for deriving new posteriors". In: *Arxiv:1904.02063*.

Knoblauch, J. and Vomfell, L. (2020). "Robust bayesian inference for discrete outcomes with the total variation distance". In: *Arxiv:2010.13456*.

Li, Y., Swersky, K., and Zemel, R. (2015). "Generative moment matching networks". In: *International conference on machine learning*. PMLR, pp. 1718–1727.

Lyddon, S., Walker, S., and Holmes, C. (2018). "Nonparametric learning from bayesian models with randomized objective functions". In: *Proceedings of the 32nd international conference on neural information processing systems*, pp. 2075–2085.

Lyddon, S., Holmes, C., and Walker, S. (2019). "General bayesian updating and the loss-likelihood bootstrap". In: *Biometrika* 106.2, pp. 465–478.

Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. (2021). "Robust generalised Bayesian inference for intractable likelihoods". In: *Arxiv:2104.07359*.

Muliere, P. and Secchi, P. (1996). "Bayesian nonparametric predictive inference and bootstrap techniques". In: *Annals of the institute of statistical mathematics* 48.4, pp. 663–673.

Müller, A. (1997). "Integral probability metrics and their generating classes of functions". In: *Advances in applied probability* 29.2, pp. 429–443.

Niu, Z., Meier, J., and Briol, F.-X. (2021). "Discrepancy-based inference for intractable generative models using quasi-Monte Carlo". In: *Arxiv:2106.11561*.

Owhadi, H., Scovel, C., and Sullivan, T. (2015). "Brittleness of bayesian inference under finite information in a continuous world". In: *Electronic journal of statistics* 9.1, pp. 1–79.

Pacchiardi, L. and Dutta, R. (2020). "Score matched conditional exponential families for likelihood-free inference". In: *Arxiv:2012.10903*.

Pacchiardi, L. and Dutta, R. (2021). "Generalized bayesian likelihood-free inference using scoring rules estimators". In: *Arxiv:2104.03889*.

Papamakarios, G. and Murray, I. (2016). "Fast ε-free inference of simulation models with bayesian conditional density estimation". In: *Advances in neural information processing systems*, pp. 1028–1036.

Papamakarios, G., Sterratt, D., and Murray, I. (2019). "Sequential neural likelihood: fast likelihood-free inference with autoregressive flows". In: vol. 89. International Conference on Artificial Intelligence and Statistics, pp. 837–848.

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). "K2-abc: approximate bayesian computation with

kernel embeddings". In: *Arti cial intelligence and statistics*. PMLR, pp. 398–407.

Parr, W. C. and Schucany, W. R. (1980). "Minimum distance and robust estimation". In: *Journal of the american statistical association* 75.371, pp. 616–624.

Peters, G., Chen, W., and Gerlach, R. (2016). "Estimating quantile families of loss distributions for non-life insurance modelling via L-moments". In: *Risks* 4.2, p. 14.

Prangle, D. (2020). "Gk: an r package for the g-and-k and generalised g-and-h distributions". In: *The r journal* 12.1, pp. 7–20.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2017). "Bayesian synthetic likelihood". In: *Journal of computational and graphical statistics* 8600.

Rayner, G. D. and MacGillivray, H. L. (2002). "Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions". In: *Statistics and computing* 12.1, pp. 57–75.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. (2018). "Deep generative models of genetic variation capture the effects of mutations". In: *Nature methods* 15.10, pp. 816–822.

Rodrigues, G. S., Nott, D. J., and Sisson, S. A. (2020). "Likelihood-free approximate Gibbs sampling". In: *Statistics and computing* 30.4, pp. 1057–1073.

Schmon, S. M., Cannon, P. W., and Knoblauch, J. (2020). "Generalized posteriors in approximate bayesian computation". In: *Third symposium on advances in approximate bayesian inference*.

Sethuraman, J. (1994). "A constructive definition of dirichlet priors". In: *Statistica sinica*, pp. 639–650.

Shawe-Taylor, J. and Williamson, R. C. (1997). "A PAC analysis of a Bayesian estimator". In: *Proceedings of the tenth annual conference on computational learning theory*. Vol. 6. 09, pp. 2–9.

Sisson, S. and Fan, Y. (2018). "Abc samplers". In: *Handbook of approximate bayesian computation*, pp. 87–123.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). "Hilbert space embeddings and metrics on probability measures". In: *The journal of machine learning research* 11, pp. 1517–1561.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A. J., and Gretton, A. (2017). "Generative models and model criticism via optimized maximum mean discrepancy". In: *International conference on learning representations*.

Virtanen, P. et al. (2020). "Scipy 1.0: fundamental algorithms for scientific computing in python". In: *Nature methods* 17.3, pp. 261–272.

Wu, P.-S. and Martin, R. (2020). "A comparison of learning rate selection methods in generalized bayesian inference". In: *Arxiv:2012.11349*.

Supplementary Material: Robust Bayesian Inference for Simulator-based Models via the MMD Posterior Bootstrap

In Appendix A we summarise the notation used throughout the paper. In Appendix B we prove all theoretical results. In Appendix C we provide details on the experiments introduced in the main text and in Appendix D we provide some additional experiments.

A NOTATION

In the following section, we recall the notation used in the paper:

P True data generating distribution

$x_{1:n}$ Observations $x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}$

P_n Empirical measure $\mathsf{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

F Centering measure of the Dirichlet Process prior $\mathrm{DP}(\alpha, \mathsf{F})$

F^0 Centering measure of the Dirichlet Process posterior $\mathrm{DP}(\alpha^0, \mathsf{F}^0)$ defined as $\mathsf{F}^0 = \frac{\alpha}{\alpha+n}\mathsf{F} + \frac{n}{\alpha+n}\mathsf{P}_n$

$\hat{\nu}$ Probability measure on $\mathcal{P}$ defined by the sampling procedure of samples from the Dirichlet approximation of the DP posterior $\mathrm{DP}(\alpha^0, \mathsf{F}^0)$ in (3)

ν Probability measure on $\mathcal{P}$ defined by the stick-break process representing samples from the exact DP posterior $\mathsf{P} \mid x_{1:n} \quad \mathrm{DP}(\alpha^0, \mathsf{F}^0)$

$\mathcal{X}$ Data space

$\mathcal{P}$ Space of Borel distributions on $\mathcal{X}$

P_θ Probability measure $\mathsf{P}_\theta \in \mathcal{P}$ indexed by parameter θ

$\mathcal{P}_\Theta$ Family of parametric models $\mathcal{P}_\Theta = \{\mathsf{P}_\theta : \theta \in \Theta\} \quad \mathcal{P}$

θ_l Map $\theta_l : \mathcal{P} \to \Theta$ indexed by loss l which takes as input a probability measure P and returns the value of $\theta \in \Theta$ which minimises the expected loss; i.e. $\theta_l(\mathsf{P}) = \mathrm{arginf}_{\theta \in \Theta} \mathsf{E}_{X \sim \mathsf{P}}[l(X, \theta)]$

θ^* Map $\theta^* : \mathcal{P} \to \Theta$ for the MMD-based loss which takes as input a probability measure P and returns the value of $\theta \in \Theta$ which minimises the MMD between P_θ and P; i.e. $\theta^*(\mathsf{P}) = \mathrm{arginf}_{\theta \in \Theta} \mathrm{MMD}^2(\mathsf{P}, \mathsf{P}_\theta)$

B PROOFS

We now provide proofs for all theoretical results presented in the main text as well as some additional theory. We start with the necessary background material on the MMD that we use in our proofs. We continue by proving our generalisation error bound both in the case where the DP posterior sample is represented exactly by an infinite sum of the stick-breaking process (Theorem 9) and using the approximation of algorithm 1 (Theorem 3). We also provide the corresponding corollaries for the well-specified case (Corollaries 12 and 10) for reference. We then proceed by proving the general posterior consistency result of Theorem 4; along with the special case of a well-specified model (Corollary 13). Finally, we prove the results in the case of a contaminated data generating process in both the misspecified and well-specified case (Corollaries 5 and 15). Recall that throughout all our theoretical results we impose Standing Assumptions 1 and 2.

B.1 MMD Through Kernel Mean Embeddings

In the following proofs we will use an equivalent definition of the MMD through *kernel mean embeddings*. Consider an RKHS H_k indexed by a reproducing kernel k. We say k is a reproducing kernel if (i) $k(\,,x) \in H_k$, $\forall x \in X$, (ii) $\langle f, k(\,,x)\rangle_{H_k} = f(x)$ $\forall x \in X$ and $\forall f \in H_k$ (reproducing property); and inner product $\langle\,,\,\rangle_{H_k}$; see Berlinet et al. (2011). For a function $f \in H_k$ the mean embedding $\mu_P \in H_k$ with respect to probability measure $P \in \mathcal{P}$ is defined as

$$\mu_P(\,) := \mathbb{E}_{X \sim P}[k(X,\,)] \in H_k$$

and satisfies

$$\mathbb{E}_{X \sim P}[f(X)] = \langle f, \mu_P \rangle_{H_k}.$$

The MMD between probability measures P and Q takes the form (see Gretton et al. 2012, Lemma 4)

$$\mathrm{MMD}(P, Q) = \|\mu_P - \mu_Q\|_{H_k}.$$

B.2 Generalisation Error for the Stick-Breaking Process

We start with a bound of the generalisation error for the case where the DP sample is represented *exactly*—i.e., by the infinite sum obtained from the stick-breaking process. We denote by ν the probability measure on $\mathcal{P}$ corresponding to the stick-breaking process given as

$$w_{1:\infty} \sim \mathrm{GEM}(\alpha^0), \quad \alpha^0 := \alpha + n, \tag{7}$$

$$z_{1:\infty} \overset{\mathrm{iid}}{\sim} F^0 := \tfrac{\alpha}{\alpha+n} F + \tfrac{n}{n+\alpha} P_n, \tag{8}$$

$$P = \textstyle\sum_{i=1}^{\infty} w_i \delta_{z_i} \sim \nu.$$

Note that instead of writing the expectation directly over ν, we instead often use the separate expectations $\mathbb{E}_{w_{1:\infty} \sim \mathrm{GEM}(\alpha^0)}$ and $\mathbb{E}_{z_{1:\infty} \sim F^0}$ induced by ν in the proofs below.

Before stating and proving the main result of this section, we provide three lemmas that bound the expected MMD between the true data generating mechanism P and P_n, the centering measure of the DP posterior F^0 and P and lastly F^0 and P_n. The reason for this is that the main proof will use a decomposition of the MMD using the triangle inequality; and the following technical Lemmas bound three of the trickier terms arising from said triangle inequality.

Lemma 6. *For $x_{1:n} \overset{\mathrm{iid}}{\sim} P$ we have*

$$\mathbb{E}_{x_{1:n} \overset{\mathrm{iid}}{\sim} P}\left[\mathrm{MMD}(P_n, P)\right] \leq \tfrac{1}{n}.$$

where P_n denotes the empirical measure of the sample data, i.e. $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$.

Proof. Proof follows directly from the proof of Chérief-Abdellatif et al. (2019), Lemma 7.1 and using the Jensen's inequality to obtain

$$\mathbb{E}_{x_{1:n} \overset{\mathrm{iid}}{\sim} P}\left[\mathrm{MMD}(P_n, P)\right] \leq \sqrt{\mathbb{E}_{x_{1:n} \overset{\mathrm{iid}}{\sim} P}\left[\mathrm{MMD}^2(P_n, P)\right]} \leq \tfrac{1}{n}.$$

□

Lemma 7. *Let $P = \sum_{i=1}^{\infty} w_i \delta_{z_i} \sim \nu$ denote a sample from the DP posterior, where $z_{1:\infty} \overset{\mathrm{iid}}{\sim} F^0$ and $w_{1:\infty} \sim \mathrm{GEM}(\alpha^0)$, then*

$$\mathbb{E}_{w_{1:\infty} \sim \mathrm{GEM}(\alpha^0)}\left[\mathbb{E}_{z_{1:\infty} \sim F^0}\left[\mathrm{MMD}(P, F^0)\right]\right] \leq \tfrac{1}{\alpha+n+2}.$$

Proof. First note that

$$\begin{aligned}
\mathrm{MMD}^2(P, F^0) &= \|\mu_P - \mu_{F^0}\|^2_{H_k} \\
&= \|\textstyle\sum_{i=1}^{\infty} w_i k(z_i,\,) - \mu_{F^0}\|^2_{H_k} \\
&= \|\textstyle\sum_{i=1}^{\infty} w_i \left[k(z_i,\,) - \mu_{F^0}\right]\|^2_{H_k} \\
&= \textstyle\sum_{i=1}^{\infty} w_i^2 \|k(z_i,\,) - \mu_{F^0}\|^2_{H_k} + 2\sum_{i<j} w_i w_j \langle k(z_i,\,) - \mu_{F^0}, k(z_j,\,) - \mu_{F^0}\rangle
\end{aligned}$$

Note that since $z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0$ we have that for any $i \neq j$:

$$\begin{aligned}
&\mathsf{E}_{z_i,z_j \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\left\langle k(z_i,\cdot) - \mu_{\mathsf{F}^0}, k(z_j,\cdot) - \mu_{\mathsf{F}^0}\right\rangle\right] \\
&= \mathsf{E}_{z_i,z_j \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\left\langle k(z_i,\cdot) - \mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)], k(z_j,\cdot) - \mathsf{E}_{z_j \sim \mathsf{F}^0}[(k(z_j,\cdot)]\right\rangle\right] \\
&= \mathsf{E}_{z_i,z_j \overset{\text{iid}}{\sim} \mathsf{F}^0}[k(z_i,z_j)] - \mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)]\,\mathsf{E}_{z_j \sim \mathsf{F}^0}[k(z_j,\cdot)] \\
&\qquad - \mathsf{E}_{z_j \sim \mathsf{F}^0}[k(z_j,\cdot)]\,\mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)] + \mathsf{E}_{z_j \sim \mathsf{F}^0}[k(z_j,\cdot)]\,\mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)] \\
&= \mathsf{E}_{z_i,z_j \overset{\text{iid}}{\sim} \mathsf{F}^0}[k(z_i,z_j)] - \mathsf{E}_{z_i,z_j \overset{\text{iid}}{\sim} \mathsf{F}^0}[k(z_i,z_j)] - \mathsf{E}_{z_i,z_j \overset{\text{iid}}{\sim} \mathsf{F}^0}[k(z_i,z_j)] + \mathsf{E}_{z_i,z_j \overset{\text{iid}}{\sim} \mathsf{F}^0}[k(z_i,z_j)] \\
&= 0. \qquad (9)
\end{aligned}$$

Moreover, for any $i = 1, 2, \ldots$

$$\begin{aligned}
\mathsf{E}_{z_i \sim \mathsf{F}^0}\left[k(z_i,\cdot) - \mu_{\mathsf{F}^0}, k^2_{H_k}\right] &= \mathsf{E}_{z_i \sim \mathsf{F}^0}\left[k(z_i,\cdot) - \mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)]k^2_{H_k}\right] \\
&= \mathsf{E}_{z_i \sim \mathsf{F}^0}\left[k(z_i,\cdot)k^2_{H_k} - 2\langle k(z_i,\cdot), \mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)]\rangle + k\,\mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)]k^2_{H_k}\right] \\
&= \mathsf{E}_{z_i \sim \mathsf{F}^0}\left[k(z_i,\cdot)k^2_{H_k}\right] - 2\langle \mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)], \mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)]\rangle + k\,\mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)]k^2_{H_k} \\
&= \mathsf{E}_{z_i \sim \mathsf{F}^0}\left[k(z_i,\cdot)k^2_{H_k}\right] - 2k\,\mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)]k^2_{H_k} + k\,\mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)]k^2_{H_k} \\
&= \mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)k^2_{H_k}] - k\,\mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)]k^2_{H_k} \\
&\quad \mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot)k^2_{H_k}] \\
&= \mathsf{E}_{z_i \sim \mathsf{F}^0}[\langle k(z_i,z_i), k(z_i,z_i)\rangle] \qquad (10)
\end{aligned}$$

where the last equality follows from the fact that

$$k(z_i,\cdot)k^2_{H_k} = \langle k(z_i,\cdot), k(\cdot,z_i)\rangle\langle k(z_i,z_i)\rangle$$

using the reproducing property of the RKHS which says that $\langle f, k(x,\cdot)\rangle_{H_k} = f(x)$. We then obtain:

$$\begin{aligned}
\mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}\left[\mathsf{E}_{z_{1:1} \sim \mathsf{F}^0}\left[\text{MMD}^2(\mathsf{P},\mathsf{F}^0)\right]\right] \quad &\sum_{i=1}^{1} \mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}[w_i^2]\,\mathsf{E}_{z_i \sim \mathsf{F}^0}[k(z_i,\cdot) - \mu_{\mathsf{F}^0}k^2_{H_k}] + 2\sum_{i\neq j} 0 \\
&\sum_{i=1}^{1} \mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}[w_i^2]\,\mathsf{E}_{z_i \sim \mathsf{F}^0}[\langle k(z_i,z_i)\rangle] \\
&\sum_{i=1}^{1} \mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}[w_i^2].
\end{aligned}$$

The first inequality above follows from equation (9), the second inequality follows from equation (10), and the last inequality follows from the boundedness of the kernel in standing assumption 1.

From the properties of the GEM and Beta distributions we have that since $w_k = \beta_k \prod_{i=1}^{k-1}(1-\beta_i)$ where $\beta_{1:1} \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha+n)$ then

$$\mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}[w_i] = \tfrac{(\alpha+n)^{i-1}}{(1+\alpha+n)^i} \qquad \text{and} \qquad \mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}[w_i^2] = \tfrac{2(\alpha+n)^{i-1}}{(\alpha+n+2)^i(\alpha+n+1)}.$$

Therefore

$$\begin{aligned}
\mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\text{MMD}^2(\mathsf{P},\mathsf{F}^0)\right]\right] \quad &\sum_{i=1}^{1} \mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}[w_i^2] \\
&= \tfrac{2}{(\alpha+n+2)(\alpha+n+1)} \sum_{i=1}^{1} \left(\tfrac{\alpha+n}{\alpha+n+2}\right)^{i-1} \\
&= \tfrac{2}{(\alpha+n+2)(\alpha+n+1)} \tfrac{1}{1-\frac{\alpha+n}{\alpha+n+2}} \\
&= \tfrac{1}{\alpha+n+2}
\end{aligned}$$

Finally, by Jensen's inequality,

$$\mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\text{MMD}(\mathsf{P},\mathsf{F}^0)\right]\right] \quad \sqrt{\mathsf{E}_{w_{1:1} \sim \text{GEM}(\alpha^0)}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\text{MMD}^2(\mathsf{P},\mathsf{F}^0)\right]\right]} \quad \leq \sqrt{\tfrac{1}{\alpha+n+2}}.$$

□

Lemma 8. *Let P_n denote the empirical measure $\mathsf{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ then*

$$\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\mathrm{MMD}(\mathsf{P}_n, \mathsf{F}^0)\right]\right] \quad \tfrac{2\alpha}{\alpha+n}.$$

Proof. By definition of F^0 and linearity of expectations, it follows that $\mathsf{E}_{\mathsf{F}^0}[\,] = \frac{\alpha}{\alpha+n}\mathsf{E}_{\mathsf{F}}[\,] + \frac{n}{n+\alpha}\mathsf{E}_{\mathsf{P}_n}[\,]$ and hence

$$\mu_{\mathsf{F}^0} = \mathsf{E}_{\mathsf{F}^0}[k(z,\,)] = \tfrac{\alpha}{\alpha+n}\mathsf{E}_{\mathsf{F}}[k(z,\,)] + \tfrac{n}{n+\alpha}\mathsf{E}_{\mathsf{P}_n}[k(z,\,)] = \tfrac{\alpha}{\alpha+n}\mu_{\mathsf{F}} + \tfrac{n}{n+\alpha}\mu_{\mathsf{P}_n}. \tag{11}$$

Therefore,

$$\begin{aligned}
&\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\mathrm{MMD}^2(\mathsf{P}_n, \mathsf{F}^0)\right]\right] \\
&= \mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[k\mu_{\mathsf{P}_n} \quad \mu_{\mathsf{F}^0} k^2_{H_k}\right]\right] \\
&= \mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\left\|\mu_{\mathsf{P}_n} \quad \tfrac{\alpha}{\alpha+n}\mu_{\mathsf{F}} \quad \tfrac{n}{n+\alpha}\mu_{\mathsf{P}_n}\right\|^2_{H_k}\right]\right] \\
&= \tfrac{\alpha^2}{(\alpha+n)^2}\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[k\mu_{\mathsf{P}_n} \quad \mu_{\mathsf{F}} k^2_{H_k}\right]\right] \\
&= \tfrac{\alpha^2}{(\alpha+n)^2}\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\left\|\tfrac{1}{n}\textstyle\sum_{i=1}^n k(x_i,\,)\right\|^2_{H_k} \quad \tfrac{2}{n}\textstyle\sum_{i=1}^n h k(x_i,\,), \mathsf{E}_{z\ \mathsf{F}}[k(z,\,)] i + k\mu_{\mathsf{F}} k^2_{H_k}\right]\right] \\
&= \tfrac{\alpha^2}{(\alpha+n)^2}\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\tfrac{1}{n^2}\textstyle\sum_{i=1}^n k(x_i,x_i) + \tfrac{2}{n^2}\sum_{i<j} k(x_i,x_j) \quad \tfrac{2}{n}\sum_{i=1}^n \mathsf{E}_{z\ \mathsf{F}}[k(x_i,z)]\right.\right. \\
&\qquad \left.\left. + k\,\mathsf{E}_{z\ \mathsf{F}}[k(z,\,)] k^2_{H_k}\right]\right] \\
&= \tfrac{\alpha^2}{(\alpha+n)^2}\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\tfrac{1}{n^2}\textstyle\sum_{i=1}^n k(x_i,x_i) + \tfrac{2}{n^2}\sum_{i<j} k(x_i,x_j) \quad \tfrac{2}{n}\mathsf{E}_{z\ \mathsf{F}}[\textstyle\sum_{i=1}^n k(x_i,z)]\right.\right. \\
&\qquad \left.\left. + \mathsf{E}_{z,z^0 \overset{\text{iid}}{\sim} \mathsf{F}}[k(z,z^0)]\right]\right] \\
&\quad \tfrac{\alpha^2}{(\alpha+n)^2}\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1} \overset{\text{iid}}{\sim} \mathsf{F}^0}\left[\tfrac{1}{n} + \tfrac{n(n-1)}{n^2} + 2 + 1\right]\right] \\
&= \tfrac{4\alpha^2}{(\alpha+n)^2}.
\end{aligned}$$

The first three equalities above follow from the definition of the MMD in terms of the kernel mean embeddings and equation (11). The following three equalities use the definition of the kernel mean embedding as well as the characteristic property which ensures that

$$\begin{aligned}
k\,\mathsf{E}_{z\ \mathsf{F}}[k(z,\,)] k^2_{H_k} &= h\mathsf{E}_{z\ \mathsf{F}}\, k(z,\,), \mathsf{E}_{z\ \mathsf{F}}\, k(z,\,) i_{H_k} \\
&= h\mathsf{E}_{z\ \mathsf{F}}\, k(z,\,), \mathsf{E}_{z^0\ \mathsf{F}}\, k(\,, z^0) i_{H_k} \\
&= \mathsf{E}_{z,z^0\ \mathsf{F}}\, h k(z,\,), k(\,, z^0) i_{H_k} \\
&= \mathsf{E}_{z,z^0\ \mathsf{F}}[k(z,z^0)].
\end{aligned}$$

For the inequality above we used again the standing assumption 1 about the boundedness of the kernel. To conclude, by Jensen's inequality we have

$$\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1}\ \mathsf{F}^0}\left[\mathrm{MMD}(\mathsf{P}_n, \mathsf{F}^0)\right]\right] \quad \sqrt{\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{z_{1:1}\ \mathsf{F}^0}\left[\mathrm{MMD}^2(\mathsf{P}_n, \mathsf{F}^0)\right]\right]} \quad \tfrac{2\alpha}{\alpha+n}.$$

□

We can now formulate and prove the generalisation error bound below.

Theorem 9. *Assume $x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}$ and let P be a sample from the exact DP posterior with law ν. Then*

$$\mathsf{E}_{x_{1:n} \overset{\text{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \nu}\left[\mathrm{MMD}(\mathsf{P}\ , \mathsf{P}_{\theta\ (\mathsf{P})})\right]\right] \quad \inf_{\theta 2 \Theta} \mathrm{MMD}(\mathsf{P}_\theta, \mathsf{P}\ ) + \tfrac{p^2}{n} + \tfrac{p\,2}{\alpha+n+2} + \tfrac{4\alpha}{\alpha+n}.$$

Proof. Using the triangle inequality we have that for any $\theta\ 2\ \Theta$:

$$\begin{aligned}
\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})}) \quad & \mathrm{MMD}(\mathsf{P}_{\theta\ (\mathsf{P})},\mathsf{P}) + \mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}) \\
& \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}) + \mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}) \\
& \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}) \\
& \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P}_n,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P},\mathsf{P}_n) \\
& \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P}_n,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P},\mathsf{F}^{\emptyset}) + 2\,\mathrm{MMD}(\mathsf{P}_n,\mathsf{F}^{\emptyset}).
\end{aligned}$$

For the first step above we used the triangle inequality and for the second we used the fact that since $\theta = \arg\inf_{\theta 2\Theta} \mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_\theta)$ it follows that $\mathrm{MMD}(\mathsf{P}_{\theta\ (\mathsf{P})},\mathsf{P})$ $\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P})$ for any $\theta\ 2\ \Theta$. The remaining inequalities are obtained by reapplying the triangle inequality on $\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P})$, $\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P})$ and $\mathrm{MMD}(\mathsf{P},\mathsf{P}_n)$ respectively. Since the above is true for any $\theta\ 2\ \Theta$ it follows that

$$\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})}) \quad \inf_{\theta 2\Theta} \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P}_n,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P},\mathsf{F}^{\emptyset}) + 2\,\mathrm{MMD}(\mathsf{P}_n,\mathsf{F}^{\emptyset}).$$

Taking expectations on both sides we obtain

$$\begin{aligned}
\mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\mathsf{E}_{w_{1:\ 1}\quad \mathrm{GEM}(\alpha^{\emptyset})}\left[\mathsf{E}_{z_{1:\ 1}\ ^{\mathrm{iid}}\mathsf{F}^{\emptyset}}\left[\mathrm{MMD}(\mathsf{P},\mathsf{P}_{\theta\ (\mathsf{P})})\right]\right]\right] \quad & \inf_{\theta 2\Theta} \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) + 2\,\mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\mathrm{MMD}(\mathsf{P}_n,\mathsf{P}\ )\right] \\
& + 2\,\mathsf{E}_{w_{1:\ 1}\quad \mathrm{GEM}(\alpha^{\emptyset})}\left[\mathsf{E}_{z_{1:\ 1}\ ^{\mathrm{iid}}\mathsf{F}^{\emptyset}}\left[\mathrm{MMD}(\mathsf{P},\mathsf{F}^{\emptyset})\right]\right] \\
& + 2\,\mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\mathsf{E}_{z_{1:\ 1}\ ^{\mathrm{iid}}\mathsf{F}^{\emptyset}}\left[\mathrm{MMD}(\mathsf{P}_n,\mathsf{F}^{\emptyset})\right]\right].
\end{aligned}$$

The result now follows by Lemmas 6, 7 and 8.

□

The well-specified case is an immediate consequence of the Theorem:

Corollary 10. *Suppose that the model is well-speci ed, i.e.* $\inf_{\theta 2\Theta} \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) = 0$ *then*

$$\mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \ \nu}\left[\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})})\right]\right] \quad \cancel{p}\frac{2}{n} + \cancel{p}\frac{2}{\alpha+n+2} + \frac{4\alpha}{\alpha+n}.$$

Proof. The corollary follows directly from Theorem 9 by setting $\inf_{\theta 2\Theta} \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) = 0$. □

B.3 Theorem 3

We start with a lemma bounding the expected MMD between the empirical measure and the approximated DP posterior sample. Recall that the probability measure $\hat{\nu}$ refers to the law on P induced by the sampling process:

$$(w_{1:n},\tilde{w}_{n+T}) \quad \mathrm{Dir}\left(1,\ldots,1,\tfrac{\alpha}{T},\ldots,\tfrac{\alpha}{T}\right) \tag{12}$$

$$\tilde{x}_{1:T}\ ^{\mathrm{iid}}\mathsf{F} \tag{13}$$

$$\mathsf{P} := \textstyle\sum_{i=1}^n w_i\delta_{x_i} + \sum_{k=1}^T \tilde{w}_k\delta_{\tilde{x}_k} \quad \hat{\nu}\,.$$

For clarity we use the individual expectations $\mathsf{E}_{(w_{1:n},\tilde{w}_{n+T})\ \ \mathrm{Dir}\left(1,\ldots,1,\frac{\alpha}{T},\ldots,\frac{\alpha}{T}\right)}$, which we denote by $\mathsf{E}_{w\ \ \mathrm{Dir}}$, and $\mathsf{E}_{\tilde{x}_{1:T}\ ^{\mathrm{iid}}\mathsf{F}}$ arising from Equations (12)-(13) in the proofs below—rather than a single expectation over $\hat{\nu}$.

Lemma 11. *Let* $\mathsf{P}\quad\hat{\nu}$ *denote a single draw from the approximated DP posterior, i.e.*

$$\mathsf{P} = \textstyle\sum_{i=1}^n w_i\delta_{x_i} + \sum_{k=1}^T \tilde{w}_k\delta_{\tilde{x}_k}.$$

Then

$$\mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P},\mathsf{P}_n)\right]\right] \quad 2\sqrt{\tfrac{\alpha(1+\alpha)}{(\alpha+n)(\alpha+n+1)}}$$

Proof. The mean embedding for P is $\mu_{\mathsf{P}} = \sum_{i=1}^{n} w_i k(x_i, \ ) + \sum_{j=1}^{T} \tilde{w}_j k(\tilde{x}_j, \ )$. Hence using the triangle inequality:

$$\begin{aligned}\mathrm{MMD}\left(\mathsf{P}, \mathsf{P}_n\right) &= \|\mu_{\mathsf{P}} \quad \mu_{\mathsf{P}_n}\|_{H_k} \\ &= \left\|\sum_{i=1}^{n} w_i k(x_i, \ ) + \sum_{j=1}^{T} \tilde{w}_j k(\tilde{x}_j, \ ) \quad \mu_{\mathsf{P}_n}\right\|_{H_k} \\ &\quad \|\sum_{i=1}^{n} w_i k(x_i, \ ) \quad \mu_{\mathsf{P}_n}\|_{H_k} + \left\|\sum_{j=1}^{T} \tilde{w}_j k(\tilde{x}_j, \ )\right\|_{H_k}. \end{aligned} \tag{14}$$

To bound $\mathsf{E}_{x_{1:n} \overset{\mathrm{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{\mathsf{P} \sim \hat{\nu}}\left[\|\sum_{i=1}^{n} w_i k(x_i, \ ) \quad \mu_{\mathsf{P}_n}\|_{H_k}\right]\right]$, first we note that from the moments of a Dirichlet distribution we have that for any $i, j \in \{1, \ldots, n\}$:

$$\mathsf{E}_{w \sim \mathrm{Dir}}[w_i] = \frac{1}{n+\alpha}, \qquad \mathsf{E}_{w \sim \mathrm{Dir}}[w_i^2] = \frac{2}{(\alpha+n+1)(\alpha+n)} \qquad \text{and} \qquad \mathsf{E}_{w \sim \mathrm{Dir}}[w_i w_j] = \frac{1}{(\alpha+n)(\alpha+n+1)}. \tag{15}$$

Hence

$$\begin{aligned}&\mathsf{E}_{x_{1:n} \overset{\mathrm{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{\mathsf{P} \sim \hat{\nu}}\left[\left\|\tfrac{1}{n}\sum_{i=1}^{n} k(x_i, \ ) \quad \sum_{i=1} w_i k(x_i, \ )\right\|_{H_k}^2\right]\right] \\ &= \mathsf{E}_{x_{1:n} \overset{\mathrm{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{\mathsf{P} \sim \hat{\nu}}\left[\left\|\sum_{i=1}^{n}\left(\tfrac{1}{n} \quad w_i\right) k(x_i, \ )\right\|_{H_k}^2\right]\right] \\ &= \mathsf{E}_{x_{1:n} \overset{\mathrm{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{\mathsf{P} \sim \hat{\nu}}\left[\sum_{i=1}^{n}\left(\tfrac{1}{n} \quad w_i\right)^2 k(x_i, x_i) + 2\sum_{1 \ i<j \ n}\left(\tfrac{1}{n} \quad w_i\right)\left(\tfrac{1}{n} \quad w_j\right) k(x_i, x_j)\right]\right] \\ &\quad \mathsf{E}_{\mathsf{P} \sim \hat{\nu}}\left[\sum_{i=1}^{n}\left(\tfrac{1}{n} \quad w_i\right)^2 + 2\sum_{1 \ i<j \ n}\left(\tfrac{1}{n} \quad w_i\right)\left(\tfrac{1}{n} \quad w_j\right)\right] \\ &= \sum_{i=1}^{n}\left[\tfrac{1}{n^2} \quad \tfrac{2}{n}\mathsf{E}_{w \sim \mathrm{Dir}}(w_i) + \mathsf{E}_{w \sim \mathrm{Dir}}(w_i^2)\right] + 2\sum_{1 \ i<j \ n}\left[\tfrac{1}{n^2} \quad \tfrac{1}{n}\mathsf{E}_{w \sim \mathrm{Dir}}(w_i) \quad \tfrac{1}{n}\mathsf{E}_{w \sim \mathrm{Dir}}(w_j) + \mathsf{E}_{w \sim \mathrm{Dir}}(w_i w_j)\right] \\ &= \tfrac{1}{n} \quad \tfrac{2n}{n(n+\alpha)} + \tfrac{2n}{(n+\alpha)(n+\alpha+1)} + \tfrac{n(n \ 1)}{n^2} \quad \tfrac{2(n \ 1)}{n+\alpha} + \tfrac{n(n \ 1)}{(n+\alpha)(n+\alpha+1)} \\ &= \tfrac{\alpha(\alpha+1)}{(n+\alpha)(n+\alpha+1)}. \end{aligned}$$

Here the inequality follows again from standing assumption 1 and the result is obtained by using the expectations in (15). Hence, by Jensen's inequality,

$$\mathsf{E}_{x_{1:n} \overset{\mathrm{iid}}{\sim} \mathsf{P}}\left[\mathsf{E}_{\mathsf{P} \sim \hat{\nu}}\left[\|\sum_{i=1}^{n} w_i k(x_i, \ ) \quad \mu_{\mathsf{P}_n}\|_{H_k}\right]\right] \quad \sqrt{\frac{\alpha(\alpha+1)}{(n+\alpha)(n+\alpha+1)}}. \tag{16}$$

Similarly due to standing assumption 1, to bound $\mathsf{E}_{\mathsf{P} \sim \hat{\nu}}[\|\sum_{j=1}^{T} \tilde{w}_j k(\tilde{x}_j, \ )\|_{H_k}]$, we note that

$$\begin{aligned}\left\|\sum_{j=1}^{T} \tilde{w}_j k(\tilde{x}_j, \ )\right\|_{H_k}^2 &= \sum_{j=1}^{T} \tilde{w}_j^2 k(\tilde{x}_j, \tilde{x}_j) + 2\sum_{1 \ j<k \ T} \tilde{w}_j \tilde{w}_j k(\tilde{x}_j, \tilde{x}_k) \\ &\quad \sum_{j=1}^{T} \tilde{w}_j^2 + 2\sum_{1 \ j<k \ T} \tilde{w}_j \tilde{w}_k \end{aligned}$$

hence

$$\mathsf{E}_{\mathsf{P} \sim \hat{\nu}}\left[\|\sum_{j=1}^{T} \tilde{w}_j k(\tilde{x}_j, \ )\|_{H_k}^2\right] \quad \mathsf{E}_{w \sim \mathrm{Dir}}\left[\sum_{j=1}^{T} \tilde{w}_j^2\right] + \mathsf{E}_{w \sim \mathrm{Dir}}\left[2\sum_{1 \ j<k \ T} \tilde{w}_j \tilde{w}_k\right]. \tag{17}$$

Now from the moments of a Dirichlet distribution we have that for any $j, k \in \{1, \ldots, T\}$:

$$\mathsf{E}_{w \sim \mathrm{Dir}}[\tilde{w}_j] = \frac{\alpha}{T(n+\alpha)}, \qquad \mathsf{E}_{w \sim \mathrm{Dir}}[\tilde{w}_j^2] = \frac{\alpha^2+T\alpha}{T^2(\alpha+n+1)(\alpha+n)} \qquad \text{and} \qquad \mathsf{E}_{w \sim \mathrm{Dir}}[\tilde{w}_j \tilde{w}_k] = \frac{\alpha^3+\alpha^2 n}{T^2(\alpha+n)^2(\alpha+n+1)}.$$

Substituting these in equation (17) we obtain

$$\begin{aligned}\mathsf{E}_{\mathsf{P} \sim \hat{\nu}}\left[\left\|\sum_{j=1}^{T} \tilde{w}_j k(\tilde{x}_j, \ )\right\|_{H_k}^2\right] \quad &\frac{\alpha^2+T\alpha}{T(\alpha+n+1)(\alpha+n)} + (T \quad 1)\frac{\alpha^3+\alpha^2 n}{T(\alpha+n)^2(\alpha+n+1)} \\ &= \frac{T(\alpha+n)(\alpha+1)\alpha}{T(\alpha+n+1)(\alpha+n)} \\ &= \frac{\alpha(1+\alpha)}{(\alpha+n)(\alpha+n+1)}. \end{aligned}$$

Hence, by Jensen's inequality

$$\mathsf{E}_{\mathsf{P} \sim \hat{\nu}}\left[\left\|\sum_{j=1}^{T} \tilde{w}_j k(\tilde{x}_j, \ )\right\|_{H_k}\right] \quad \sqrt{\frac{\alpha(1+\alpha)}{(\alpha+n)(\alpha+n+1)}} \tag{18}$$

for some fixed α. Therefore, by substituting equations (16) and (18) in equation (14) we obtain the required result:

$$\mathsf{E}_{x_{1:n}\overset{\text{iid}}{\sim}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P},\mathsf{P}_n)\right]\right] \quad \mathsf{E}_{x_{1:n}\overset{\text{iid}}{\sim}\mathsf{P}}\left[k\textstyle\sum_{i=1}^n w_i k(x_i,\ )\quad \mu_{\mathsf{P}_n}k_{H_k}\right] + \mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\left\|\textstyle\sum_{j=1}^T \tilde{w}_j k(\tilde{x}_j,\ )\right\|_{H_k}\right]$$
$$2\sqrt{\tfrac{\alpha(1+\alpha)}{(\alpha+n)(\alpha+n+1)}}.$$

□

We now proceed with the proof of the generalisation error in Theorem 3.

Proof. We have that for any $\theta\ 2\ \Theta$:

$$\begin{aligned}\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})}) \quad & \mathrm{MMD}(\mathsf{P}_{\theta\ (\mathsf{P})},\mathsf{P}) + \mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}) \\ & \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}) + \mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}) \\ & \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}) \\ & \mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P}_n,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P},\mathsf{P}_n)\end{aligned}$$

For the first inequality above we used the triangle inequality and for the second inequality we used the fact that $\theta\ (\mathsf{P}) = \arg\inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P},\mathsf{P}_\theta)$. The last two inequalities follow again by repeatedly applying the triangle inequality to $\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P})$ and $\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P})$. The above statements hold for any $\theta\ 2\ \Theta$ hence we have

$$\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})}) \quad \inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P}_n,\mathsf{P}\ ) + 2\,\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_n).$$

Taking double expectation on both sides we obtain

$$\mathsf{E}_{x_{1:n}\overset{\text{iid}}{\sim}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})})\right]\right] \quad \inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) + 2\,\mathsf{E}_{x_{1:n}\overset{\text{iid}}{\sim}\mathsf{P}}\left[\mathrm{MMD}(\mathsf{P}_n,\mathsf{P}\ )\right]$$
$$+ 2\,\mathsf{E}_{x_{1:n}\overset{\text{iid}}{\sim}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P},\mathsf{P}_n)\right]\right].$$

The result then follows from Lemmas 6 and 11. □

The well-specified case is an immediate consequence of the Theorem:

Corollary 12. *Suppose that the model is well-speci ed, i.e.* $\inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) = 0$ *then*

$$\mathsf{E}_{x_{1:n}\overset{\text{iid}}{\sim}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P}\ ,\mathsf{P}_{\theta\ (\mathsf{P})})\right]\right] \quad \tfrac{p^2}{n} + p\tfrac{4^{\ p\frac{}{\alpha(1+\alpha)}}}{(\alpha+n)(\alpha+n+1)}.$$

Proof. The corollary follows directly from Theorem 3 by setting $\inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) = 0$. □

B.4 Theorem 4

We can prove that the posterior consistency result indeed holds by adapting the arguments in Chérief-Abdellatif et al. (2020) (Theorem 2) from equation (6) to the required statement of Theorem 4.

Proof. Using Proposition 3 and Markov's inequality we have:

$$\mathsf{E}_{x_{1:n}\overset{\text{iid}}{\sim}\mathsf{P}}\left[\hat{\nu}\left(\mathrm{MMD}(\mathsf{P}_{\theta\ (\mathsf{P})},\mathsf{P}\ ) \quad \inf_{\theta 2\Theta}\mathrm{MMD}(\mathsf{P}_\theta,\mathsf{P}\ ) > M_n\ \ n^{\ -1/2}\right)\right]$$
$$\frac{\mathsf{E}_{x_{1:n}\overset{\text{iid}}{\sim}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P}_{\theta\ (\mathsf{P})},\mathsf{P}\ )\ \ C\right]\right]}{M_n\ n^{\ -1/2}} = \frac{\mathsf{E}_{x_{1:n}\overset{\text{iid}}{\sim}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P}_{\theta\ (\mathsf{P})},\mathsf{P}\ )\right]\right]\ \ C}{M_n\ n^{\ -1/2}}$$
$$\frac{2n^{\ -1/2}}{M_n\ n^{\ -1/2}} + \frac{4^{\ p\frac{}{n\alpha(1+\alpha)}}}{M_n(\alpha+n)(\alpha+n+1)}\ \ n\ !\ !^{\ 1}\ \ 0.$$

□

Again, the well-specified case is a special case of Theorem 4 for $C = 0$.

Corollary 13. *Suppose we have a well-speci ed model, i.e.* $\inf_{\theta 2\Theta} \mathrm{MMD}(\mathsf{P}_\theta, \mathsf{P}\ ) = 0$ *and let* $\mathsf{P} = \sum_{i=1}^n w_i \delta_{x_i} + \sum_{k=1}^T \tilde{w}_k \delta_{\tilde{x}_k}$ $\hat{\nu}$ *be a sample from the approximate posterior. Then for any* $M_n\ !\ +1$ *such that* $M_n\ n^{\ 1/2}\ !\ 0$ *as* $n\ !\ 1$

$$\hat{\nu}\left(\mathsf{P}\ 2\ P : \mathrm{MMD}(\mathsf{P}\ , \mathsf{P}_{\theta\ (\mathsf{P})}) > \tfrac{M_n}{n^{1/2}}\right)^{n\ !\ 1}\ 0.$$

Proof. Using Theorem 3 in the well specified case and Markov's inequality we have that

$$\mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\hat{\nu}\left(\mathrm{MMD}(\mathsf{P}\ , \mathsf{P}_{\theta\ (\mathsf{P})}) > M_n\ \ n^{\ 1/2}\right)\right] \quad \frac{\mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P}\ , \mathsf{P}_{\theta\ (\mathsf{P})})\right]\right]}{M_n\ n^{\ 1/2}}$$
$$\frac{2n^{\ 1/2}}{M_n\ n^{\ 1/2}} + \frac{4\sqrt{\frac{n\alpha(1+\alpha)}{(\alpha+n)(\alpha+n+1)}}}{M_n}\ ^{n\ !\ 1}\ 0.$$

□

B.5 Corollary 5

Before we present the generalisation bound we will use the following lemma from Chérief-Abdellatif et al. (2019) which bounds the absolute difference of the MMD between the parametric model and each of the contaminated and non-contaminated probability measures.

Lemma 14 (Chérief-Abdellatif et al. (2019) Lemma 3.3)**.** *For any* $\theta\ 2\ \Theta$*,*

$$|\mathrm{MMD}(\mathsf{P}_\theta, \mathsf{P}\ ) \quad \mathrm{MMD}(\mathsf{P}_\theta, \tilde{\mathsf{P}})| \quad 2\epsilon.$$

Using this lemma and the previously obtained generalisation bounds we can show the required statement of Corollary 5.

Proof. From Lemma 14 we have that for any $\theta\ 2\ \Theta$:

$$\mathrm{MMD}(\mathsf{P}_\theta, \tilde{\mathsf{P}}) \quad 2\epsilon + \mathrm{MMD}(\mathsf{P}_\theta, \mathsf{P}_0) \tag{19}$$
$$\mathrm{MMD}(\mathsf{P}_\theta, \mathsf{P}\ ) \quad 2\epsilon + \mathrm{MMD}(\mathsf{P}_\theta, \tilde{\mathsf{P}}) \tag{20}$$

Hence, using equations 19, 20 and Proposition 3 we have that

$$\mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\tilde{\mathsf{P}}, \mathsf{P}_{\theta\ (\mathsf{P})})\right]\right] \quad 2\epsilon + \mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\mathsf{P}\ , \mathsf{P}_{\theta\ (\mathsf{P})})\right]\right]$$
$$2\epsilon + \inf_{\theta 2\Theta} \mathrm{MMD}(\mathsf{P}\ , \mathsf{P}_\theta) + \frac{\sigma^2}{n} + \frac{4\sqrt{\frac{\alpha(1+\alpha)}{(\alpha+n)(\alpha+n+1)}}}{}$$
$$4\epsilon + \inf_{\theta 2\Theta} \mathrm{MMD}(\tilde{\mathsf{P}}, \mathsf{P}_\theta) + \frac{\sigma^2}{n} + \frac{4\sqrt{\frac{\alpha(1+\alpha)}{(\alpha+n)(\alpha+n+1)}}}{}$$

□

The well-specified case is an immediate consequence:

Corollary 15 (Well-specified Case)**.** *Suppose the model is well speci ed in terms of the target* $\tilde{\mathsf{P}}$*, i.e.* $9\theta\ 2\ \Theta$ *such that* $\mathsf{P}_\theta = \tilde{\mathsf{P}}$*. Then*

$$\mathsf{E}_{x_{1:n}\ ^{\mathrm{iid}}\mathsf{P}}\left[\mathsf{E}_{\mathsf{P}\ \hat{\nu}}\left[\mathrm{MMD}(\tilde{\mathsf{P}}, \mathsf{P}_{\theta\ (\mathsf{P})})\right]\right] \quad 4\epsilon + \frac{\sigma^2}{n} + \frac{4\sqrt{\frac{\alpha(1+\alpha)}{(\alpha+n)(\alpha+n+1)}}}{}.$$

Proof. The corollary follows from Corollary 5 noting that in the well-specified case $\inf_{\theta 2\Theta} \mathrm{MMD}(\mathsf{P}_\theta, \tilde{\mathsf{P}}) = 0$. □

C EXPERIMENTAL DETAILS

We now provide further experimental details of our method. We first explain how gradient-based numerical optimisation can be used to minimise the MMD objective in algorithm 1. Next, we provide the experimental setup necessary for the reproduction of the experiments presented in the main text. We then give some additional details on the G-and-k and Toggle-Switch models. We finally provide results of all experiments over a number of independent runs.

C.1 Numerical Optimisation for the MMD Objective

We first explain how the MMD objective in equation (5) can be approximated and minimized through gradient-based methods. For two probability measures P, Q *2* P recall that the squared MMD is defined as

$$\mathrm{MMD}^2(\mathsf{P},\mathsf{Q}) = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y)\,\mathsf{P}(dx)\,\mathsf{P}(dy) \quad 2\int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y)\,\mathsf{P}(dx)\,\mathsf{Q}(dy) + \int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y)\,\mathsf{Q}(dx)\,\mathsf{Q}(dy).$$

Since these integrals are usually intractable we can easily approximate this by sampling $x_{1:N} \overset{\text{iid}}{} \mathsf{P}$ and $y_{1:m} \overset{\text{iid}}{} \mathsf{Q}$ and using the U-statistic:

$$\widehat{\mathrm{MMD}}^2(\mathsf{P},\mathsf{Q}) = \frac{1}{N(N-1)} \sum_{i \neq i'}^{N} k(x_i, x_{i'}) \quad \frac{2}{Nm} \sum_{i=1}^{N}\sum_{j=1}^{m} k(x_i, y_j) + \frac{1}{m(m-1)} \sum_{j \neq j'} k(y_j, y_{j'}).$$

In our case we are interested in minimising the MMD between the parametric model P_θ *2* P_Θ defined through the simulator $G_\theta : \mathsf{U}$ *!* $\mathcal{X}$ and an approximated sample from the DP posterior P. Hence, for samples $y_{1:N} \overset{\text{iid}}{} \mathsf{P}$ and $u_{1:M} \overset{\text{iid}}{} \mathsf{U}$ the approximated squared MMD is

$$\widehat{\mathrm{MMD}}^2(\mathsf{P},\mathsf{P}_\theta) = \frac{1}{N(N-1)} \sum_{i \neq i'}^{N} k(y_i, y_{i'}) \quad \frac{2}{NM} \sum_{j=1}^{M}\sum_{i=1}^{N} k(G_\theta(u_j), y_i) + \frac{1}{M(M-1)} \sum_{j \neq j'} k(G_\theta(u_j), G_\theta(u_{j'})). \tag{21}$$

The gradient of equation (21) can also be approximated using a U-statistic as follows: assuming that the generator is differentiable with respect to θ with gradient $\nabla_\theta G_\theta$ then the U-statistic of the gradient of the squared MMD is

$$\hat{J}_\theta(u_{1:M}, y_{1:N}) := \nabla_\theta \widehat{\mathrm{MMD}}^2(\mathsf{P},\mathsf{P}_\theta) = \frac{2}{M(M-1)} \sum_{j \neq j'}^{M} \nabla_\theta G_\theta(u_j)\, \nabla_1 k(G_\theta(u_j), G_\theta(u_{j'})) \quad \frac{2}{NM} \sum_{j=1}^{M}\sum_{i=1}^{N} \nabla_\theta G_\theta(u_j)\, \nabla_1 k(G_\theta(u_j), y_i)$$

where ∇_1 denotes the partial derivative with respect to the first argument and by noting that the first term in (21) does not depend on θ. This is an unbiased statistic in the sense that

$$\mathsf{E}_{u_{1:M} \overset{\text{iid}}{} \mathsf{U}} \left[\mathsf{E}_{y_{1:N} \overset{\text{iid}}{} \mathsf{P}} [\nabla_\theta \widehat{\mathrm{MMD}}^2(\mathsf{P},\mathsf{P}_\theta)] \right] = \nabla_\theta \mathrm{MMD}^2(\mathsf{P},\mathsf{P}_\theta).$$

Using this approximation we can use gradient-based methods to minimise the required objective. For example, one can use stochastic gradient descent as follows; for learning rate η, at each iteration $k = 1, 2, \ldots$:

1. Sample $u_{1:M} \overset{\text{iid}}{} \mathsf{U}$ and $y_{1:N} \overset{\text{iid}}{} \mathsf{P}$
2. Set $\theta_k = \theta_{k-1} \quad \eta \hat{J}_{\theta_{k-1}}(u_{1:M}, y_{1:N})$

C.2 Experimental Setup

We provide details on the experimental setup required to produce figures 2, 4 and 5. For the WABC method we make use of the `winference R` package and the experimental setup provided in Bernton et al. (2019). For all experiments with our method we use `JAX` (Bradbury et al. 2020) to parallelize the bootstrap sampling and perform the optimisation step.

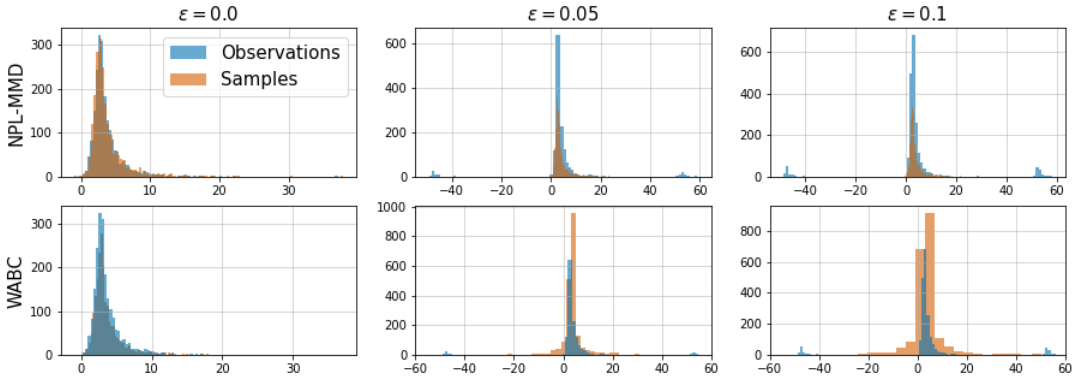


Figure 6: Histogram of 2000 samples from the g-and-k distribution with $\theta_0 = (3, 1, 1, \quad \log(2))$ (blue) and 2000 samples from the fitted models using the posterior mean (orange) for the NPL-MMD method (above) and the WABC method (below).

For all three experiments we use the Adam optimiser (Kingma et al. 2014) for the minimization of the MMD at each bootstrap estimation. The learning rate is set to $\eta = 0.1$ for the Gaussian location and g-and-k models and to $\eta = 0.04$ for the toggle-switch model. To ensure convergence, we perform 2000 optimizations steps in the toggle-switch experiment and 1000 steps for the rest.

The prior distributions set in the WABC method serve as an indication for the initialisation point of the optimization. In the Gaussian location model, a centered normal prior is imposed on the parameter, hence we initialize the optimization step at $\theta = 0$ at each bootstrap iteration. Similarly, the g-and-k model uses a Uniform prior on $[0, 10]$ for all parameters so we initialise at the midpoint $(5, 5, 5, 5)$. Finally, for the toggle-switch model we employ a random restart method as follows; 500 random values are sampled from the prior uniform distributions set in Bonassi et al. (2015) and Bernton et al. (2019). The approximated MMD loss is computed for all of them and the three initial values with the smallest loss are chosen as initial points. The optimization is performed for all three starting points for each bootstrap iteration and the estimator with the smallest loss is retained.

Finally, the length scale l of the Gaussian kernel is set using the median heuristic $l = \sqrt{\text{median}_{1 \quad 1,j \quad n}(kx_i \quad x_j k_2^2)}$ suggested in Gretton et al. (2012) and Dziugaite et al. (2015) in the Gaussian location example while for the g-and-k model we set $l = 0.15$. For the toggle-switch model, we suggest an unweighted mixture of Gaussian kernels discussed in Sutherland et al. (2017), i.e. $k(x, x^0) = \sum_{i=1}^{I} \exp(\quad kx \quad x^0 k_2^2/(2l_i^2))$ for a range of values $l_i \; 2 \; f1, 10, 20, 40, 80, 100, 130, 200, 400, 800, 1000g$.

C.3 The G-and-k Distribution

We visualise the observed data from the G-and-k distribution used in the contaminated model example of section 5 for an increasing number of outliers with $\theta_0 = (3, 1, 1, \quad \log(2))$. We further provide the density obtained by generating 2000 samples from the model using the posterior mean $\theta = \frac{1}{B}\sum_{j=1}^{B} \theta^{(j)}$ for each method. The outliers of the contaminated models are visualised on the histograms of figure 6. The densities of the fitted models show the sensitivity of the two methods to the two degrees of outliers; the densities fitted by the NPL-MMD remain similar and close to P_{θ_0} whereas the densities fitted by WABC are affected by the outliers in the misspecifed cases. Figure 7 visualises the densities obtained at each bootstrap iteration of the NPL-MMD method. Here, each density corresponds to 2000 samples from the g-and-k distribution and parameter $\theta^{(j)}$ for $j = 1, \ldots, B$. It is hence clear how each sample $\mathsf{P}^{(j)}$ from the DP posterior is mapped through $\theta \; (\;)$ to a value $\theta^{(j)}$ which gives rise to a different density. As we would expect, there is more variability in the densities for higher contamination levels. Finally, we illustrate the rate obtained in Corollary 5 in the absence of outliers for the G-and-k distribution example. Since we take $\epsilon = \alpha = 0$, the Corollary implies that the expected MMD is bounded above by $\frac{2}{n}$. To obtain an approximation of $\mathsf{E}_{x_{1:n} \; \text{iid} \; \mathsf{P}} \left[\mathsf{E}_{\mathsf{P} \quad \hat{\nu}} \left[\text{MMD}^2 \left(\tilde{\mathsf{P}}, \mathsf{P}_{\theta \; (\mathsf{P})} \right) \right] \right]$ we perform 10 runs of the algorithm and take the posterior mean $\hat{\theta} = \frac{1}{B}\sum_{i=1}^{B} \theta^{(i)}$ as the estimate of θ. We then sample 15 10^3 instances from $\mathsf{P} \quad = \mathsf{P}_{\theta_0}$ and $\mathsf{P}_{\hat{\theta}}$ and estimate the squared MMD using the U-statistic in equation (21). We repeat this experiment for ten

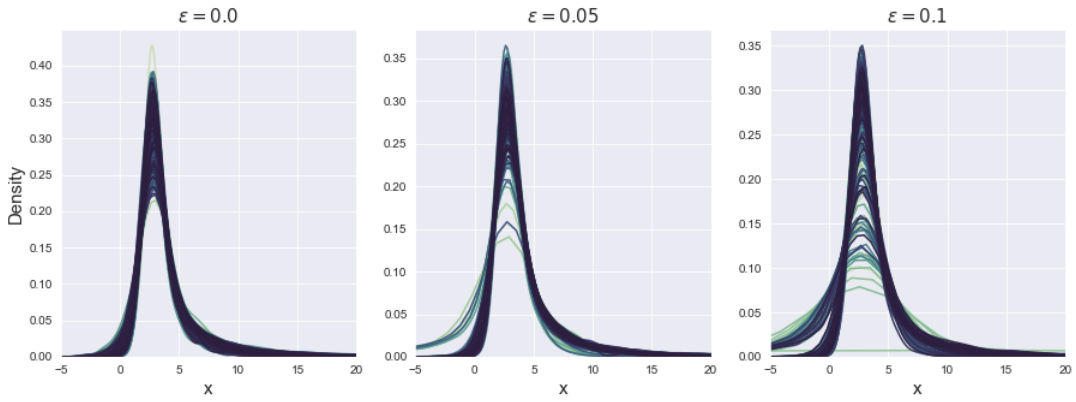


Figure 7: Realisations of the G-and-k distribution with each $\theta^{(j)}$ where $j = 1, \ldots, B$ obtained from the bootstrap iteration of the MMD posterior bootstrap algorithm for an increasing degree of contamination in the dataset.

values of sample size n 2 $[250, 4000]$ and plot $\sqrt{\mathsf{E}[\mathrm{MMD}^2(\mathsf{P}_{\hat{\theta}}, \mathsf{P}\ )]}$ against $\frac{2}{n}$ in figure 8. We observe that the estimate of the square root of the expected squared MMD is bounded above by the curve $\frac{2}{n}$ and by Jensen's inequality it is implied that the estimate of the approximate MMD will also be bounded above since:

$$\mathsf{E}[\mathrm{MMD}(\mathsf{P}_{\hat{\theta}}, \mathsf{P}\ )] \quad \sqrt{\mathsf{E}[\mathrm{MMD}^2(\mathsf{P}_{\hat{\theta}}, \mathsf{P}\ )]} \quad \frac{2}{n}.$$

C.4 The Toggle-Switch Model

C.4.1 Description of the Simulator

For cell i and unknown parameters $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \mu, \sigma, \gamma)^\top$, the simulator input is $u_i = (u_{i,1,1}, u_{i,1,2}, \ldots, u_{i,T,1}, u_{i,T,2}, u_{i,T+1,1})^\top$ Unif$([0,1]^{2T+1})$. The simulator G_θ is defined through:

$$G_\theta(u_i) = \Phi^{-1}\Big(\Phi\Big(\frac{(\mu+v_{i,T})v_{i,T}^\gamma}{\mu\sigma}\Big) + u_{i,T+1,1}\Big(1 - \Phi\Big(\frac{(\mu+v_{i,T})v_{i,T}^\gamma}{\mu\sigma}\Big)\Big)\Big)\frac{\mu\sigma}{v_{i,T}^\gamma} + (\mu + v_{i,T})$$

where for $t = 1, \ldots, T - 1$, we have

$$\begin{aligned}
\tilde{v}_{i,t+1} &= v_{i,t} + \frac{\alpha_1}{1+w_{i,t}^{\beta_1}} - (1 + 0.03v_{i,t}) \\
\tilde{w}_{i,t+1} &= w_{i,t} + \frac{\alpha_2}{1+v_{i,t}^{\beta_2}} - (1 + 0.03w_{i,t}) \\
v_{i,t+1} &= \tilde{v}_{i,t+1} + 0.5\Phi^{-1}\big(\Phi(-2\tilde{v}_{i,t+1}) + u_{i,t,1}(1 - \Phi(-2\tilde{v}_{i,t+1}))\big) \\
w_{i,t+1} &= \tilde{w}_{i,t+1} + 0.5\Phi^{-1}\big(\Phi(-2\tilde{w}_{i,t+1}) + u_{i,t,2}(1 - \Phi(-2\tilde{w}_{i,t+1}))\big)
\end{aligned}$$

and Φ denotes the CDF of the standard Gaussian distribution. We use the initial conditions $v_{i,0} = 10$, $w_{i,0} = 10$.

C.5 Results Over Multiple Independent Runs

We provide results for a number of independent runs of the experiments in section 5. For each run, a new dataset was generated and B posterior samples were obtained for each parameter θ. For each such sample, the mean estimator is recorded and the normalized mean squared error between the estimator and the true value of θ are presented with their standard deviations in tables 2, 3 and 4 for the Gaussian location, G-and-k and Toggle-Switch models respectively.

D ADDITIONAL EXPERIMENTS

We provide some additional experiments for our method by considering a type of misspecification `other` than a contamination model, comparison with MMD-Bayes in Pacchiardi et al. (2021) and exploring sensitivity to the

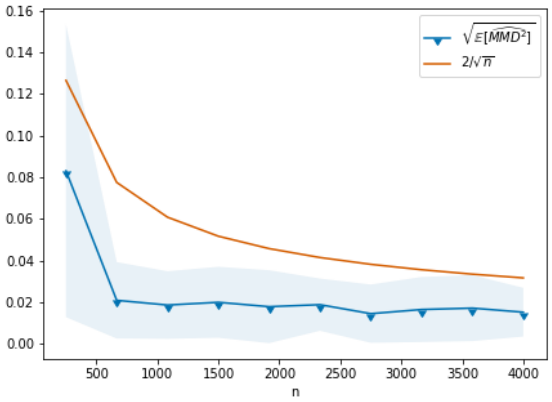


Figure 8: Illustration of the upper bound of the generalisation error in Corollary 5 in the absence of outliers for the G-and-k distribution.

Table 2: Experiment results for the Gaussian model over 10 runs

Method	**NMSE (std)**		
	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.1$
NPL-MMD	0.0107 (0.00527)	0.00889 (0.00699)	0.0113 (0.00538)
WABC	0.743 (0.0358)	0.768 (0.0624)	0.757 (0.0242)
NPL-WLL	0.00510 (0.00293)	0.945 (0.0572)	3.57 (0.0967)
NPL-WAS	0.00689 (0.00333)	0.0189 (0.00627)	0.0397 (0.0129)
MMD-ABC	0.750 (0.105)	0.755 (0.0451)	0.760 (0.0411)

Table 3: Experiment results for the G-and-k distribution over 10 runs

Method	**NMSE (std)**		
	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.1$
NPL-MMD	0.00791 (0.00524)	0.0128 (0.0117)	0.0593 (0.0255)
WABC	0.00142 (0.000965)	0.585 (0.0188)	0.532 (0.0109)

Table 4: Experiment results for the Toggle Switch model over 5 runs

Method	**NMSE (std)**
NPL-MMD	0.338 (0.277)
WABC	13.99 (13.8)

DP prior (through α and F), Gaussian kernel length scale l and truncation limit T.

D.1 Misspecified Gaussian Location Model

So far in our empirical experiments we have considered the contamination model which is canonical for analysing misspecification in robust statistics, mostly because of theoretical convenience. While the model is simple, methods that are robust against contamination models usually fare well for more practically relevant alternatives, too. Figure 9 illustrates this on a new numerical example: We generate Cauchy-distributed data, but wrongly fit a Gaussian to it. The plot shows the posterior marginals for the location parameter, and vertically marks its

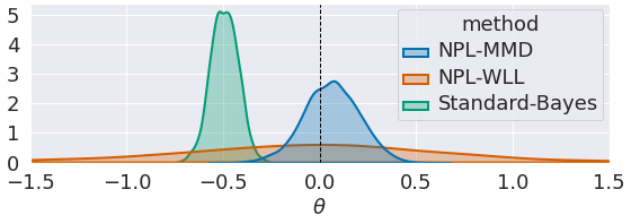


Figure 9: Marginal posterior distribution for mean of Normal location model with Cauchy data.

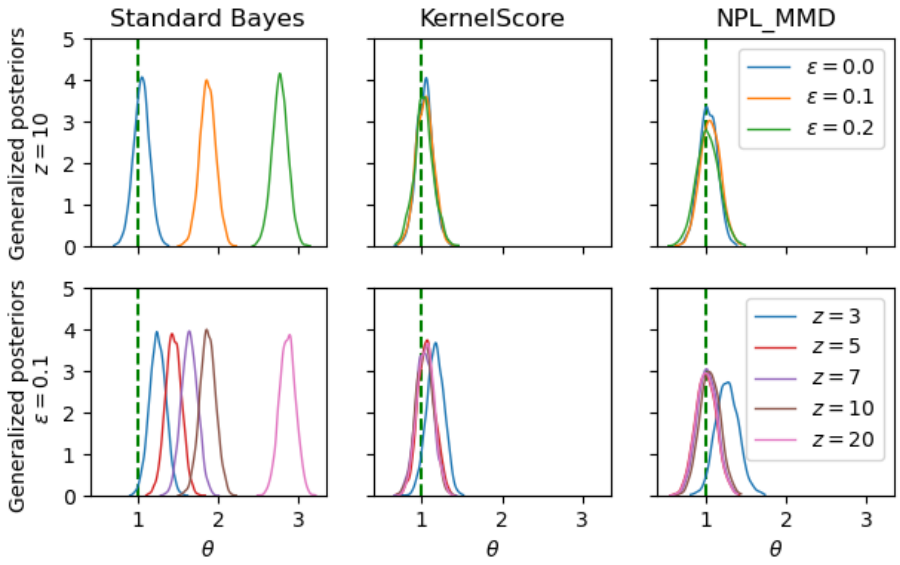


Figure 10: Contaminated Normal location model.

true location. The MSEs over 20 runs was 0.0289 (NPL-MMD), 48.2 (NPL-WLL) and 28.6 (Standard Bayes).

D.2 Comparison to MMD-Bayes for the Gaussian Location Model

We further compare our method to MMD-Bayes (`KernelScore`) in Pacchiardi et al. (2021) on an ε-contaminated Gaussian location model with outliers at location z, where the weight is chosen using the grid search as in Section 4.2 of Pacchiardi et al. (2021). Figure 10 below, shows the marginal posterior distributions for Standard Bayes, Kernel Score and the MMD Posterior bootstrap methods for an increasing number of outliers and location parameter.

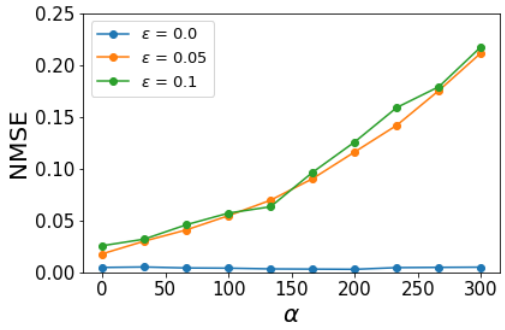
D.3 Sensitivity to Hyperparameters

In this section we empirically examine the sensitivity of the proposed method to several hyperparameters for the G-and-k distribution model.

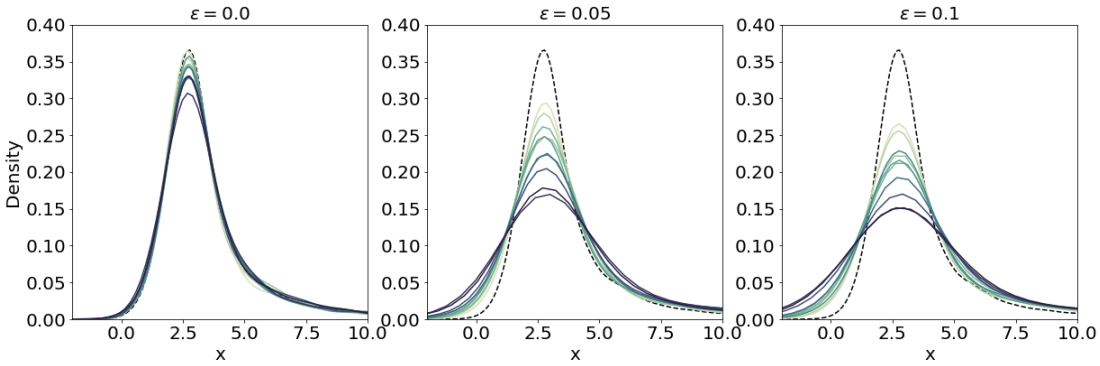
D.3.1 Sensitivity to the DP Prior

We first set $T = n$ in (3) and F to be the Normal distribution with parameters equal to the mean and standard deviation of the observed data. Here we examine the effect of the prior in two ways; first by altering the hyperparameter α of the DP prior which characterizes how much certainty we impose on F and second by choosing an increasingly 'worse' prior F, since a higher proportion of outliers leads to a worse empirical estimates of the mean and standard deviation in the Normal prior.

We generate $B = 2^9$ bootstrap samples $\theta_1, \ldots, \theta_B$ for different values of α ranging in $[0.01, 300]$ and take the mean



(a) Normalised mean squared error of the obtained estimator for an increasing value of α, corresponding to a higher confidence in the prior centering measure F.



(b) Generator samples obtained from each estimator $\hat{\theta}$. Lighter (resp. darker) curves correspond to a smaller (resp. larger) value of α. The dotted line curve corresponds to the observed dataset in the well-specified case.

Figure 11: Sensitivity to DP prior.

estimator $\hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}\theta_b$ for each value of α. In Figure 11, we plot (a) the normalised mean squared error for each estimator $\hat{\theta}$ for an increasing value of α and (b) samples from the generator with parameter $\hat{\theta}$. We observe that in the well-specified case, α has no significant effect in inference, however as we would expect, a larger value of α, in combination with a worse prior centering measure increasingly affects the parameter inference.

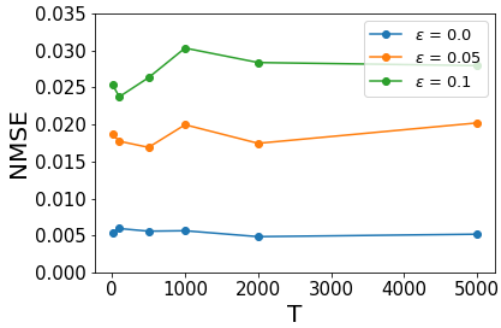
D.3.2 Sensitivity to T

We further illustrate the effect of the truncation limit T in the approximation of the DP posterior measure in (3). We fix $\alpha = 0.1$ and set F as in section D.3.1. We generate $B = 2^9$ bootstrap samples $\theta_1, \ldots, \theta_B$ for different values of T ranging in $[10, 5000]$ and take the mean estimator $\hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}\theta_b$ for each value of T. Figure 12 shows (a) the normalised mean squared error for each estimator $\hat{\theta}$ for an increasing value of T and (b) samples from the generator with parameter $\hat{\theta}$. We observe that in this example, the method is not significantly sensitive to the choice of T for all values of ϵ.

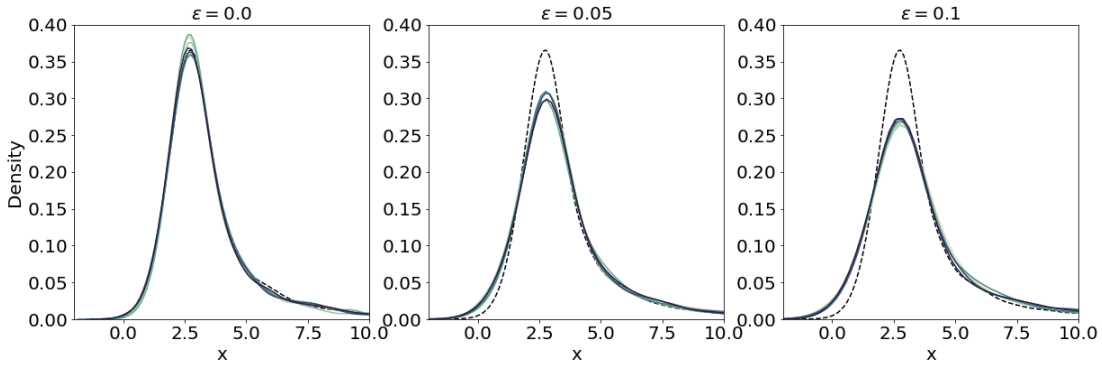
D.3.3 Sensitivity to the Hyperparameter of Gaussian Kernel

We lastly investigate the effect of the length scale l of the Gaussian kernel. We fix $\alpha = 0.1$, $T = n$ and set F as in D.3.1. We generate $B = 2^9$ bootstrap samples $\theta_1, \ldots, \theta_B$ for different values of the length scale l of the Gaussian kernel ranging in $[10^{-1}, 10^2]$ and take the mean estimator $\hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}\theta_b$ for each value of l. Figure 13 shows (a) the normalised mean squared error for each estimator $\hat{\theta}$ for an increasing value of l in logarithmic scale and (b) samples from the generator with parameter $\hat{\theta}$.

To get some more intuition we plot the MMD loss as a function of $\theta_3 = g$ around a neighborhood of the true



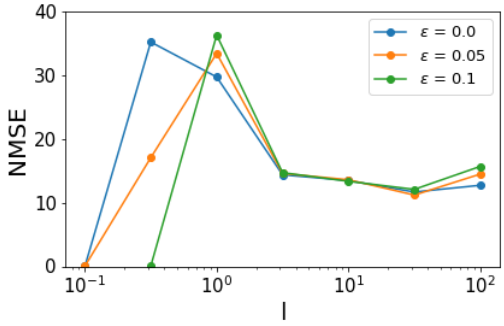
(a) Normalised mean squared error of the obtained estimator for an increasing value of T.



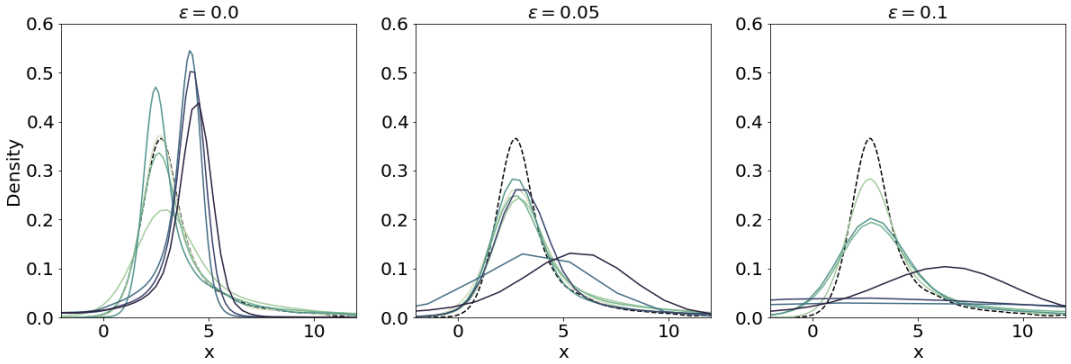
(b) Generator samples obtained from each estimator $\hat{\theta}$. Lighter (resp. darker) curves correspond to a smaller (resp. larger) value of T. The dotted line curve corresponds to the observed dataset in the well-specified case.

Figure 12: Sensitivity to parameter T, the truncation limit of the DP posterior.

parameter value $\theta_0 = 1$. Since there is noise in our estimate of the MMD, it is possible that we get a global minimum for some value of θ_3 far away from one. This is unlikely to happen for $l < 1$ because there is a much bigger dip near θ_0 as can be seen in figure 14. Of course, this is just a projection of the loss landscape during optimisation, however it gives some intuition as to why a small choice of length scale in this model leads to better results.



(a) Normalised mean squared error of the obtained estimator for an increasing value of length scale (in logarithmic scale).



(b) Generator samples obtained from each estimator $\hat{\theta}$. Lighter (resp. darker) curves correspond to a smaller (resp. larger) value of l. The dotted line curve corresponds to the observed dataset in the well-specified case.

Figure 13: Sensitivity to the length scale of the Gaussian kernel.

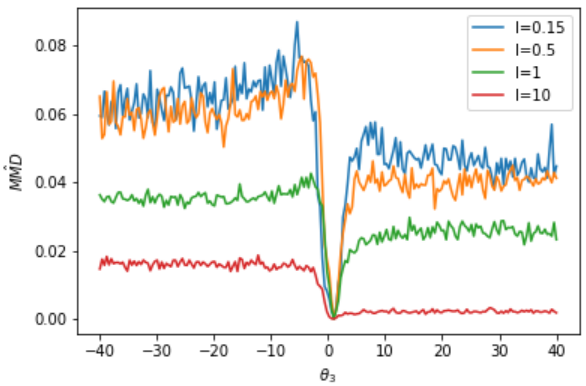


Figure 14: Estimate of the MMD loss as a function of θ_3 in the G-and-k distribution for different values of the length scale of the Gaussian kernel.