Brain and Cognition xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

Types of approximation for probabilistic cognition: Sampling and variational

Adam N. Sanborn

Department of Psychology, University of Warwick, Coventry CV4 7AL, United Kingdom

ARTICLE INFO

Article history:
Received 2 June 2015
Accepted 17 June 2015
Available online xxxx

Keywords:
Probabilistic cognition
Rational process models
Sampling
Variational approximations

ABSTRACT

A basic challenge for probabilistic models of cognition is explaining how probabilistically correct solutions are approximated by the limited brain, and how to explain mismatches with human behavior. An emerging approach to solving this problem is to use the same approximation algorithms that were been developed in computer science and statistics for working with complex probabilistic models. Two types of approximation algorithms have been used for this purpose: sampling algorithms, such as importance sampling and Markov chain Monte Carlo, and variational algorithms, such as mean-field approximations and assumed density filtering. Here I briefly review this work, outlining how the algorithms work, how they can explain behavioral biases, and how they might be implemented in the brain. There are characteristic differences between how these two types of approximation are applied in brain and behavior, which points to how they could be combined in future research.

1. Introduction

Probabilistic cognition is a natural fit to the kind of problems posed by the environment: people are faced with noisy and ambiguous observations about the world, yet need to make good decisions. Probabilistic models allow for uncertainty and ambiguity to be dealt with appropriately, because instead of incorrectly assuming that imperfect information is known perfectly, these models can find the best possible action given that imperfect information.

These models have had broad success in explaining human data, accounting for how people are aware of their perceptual uncertainty and combine it appropriately with prior knowledge (Körding & Wolpert, 2004; Tassinari, Hudson, & Landy, 2006), and explaining how people can learn to represent an ambiguous environment in cognitive tasks (Griffiths, Steyvers, & Tenenbaum, 2007; Kemp & Tenenbaum, 2008). However despite these successes, probabilistic models have faced skepticism from two major sources: evidence of mismatches between human behavior and probabilistic cognition (Tversky & Kahneman, 1978), and the inherent computational complexity of these models. It just does not seem like we as humans can do the complex calculations necessary to arrive at the best answers, and so there must be shortcuts involved (Anderson, 1991; Simon, 1955; Van Rooij, 2008).

Fortunately the problem of working with complex probabilistic models in limited systems has received a lot of attention from computer scientists and statisticians. Researchers in these fields have developed algorithms that arrive at good solutions while minimizing computational and memory requirements. These algorithms then provide an interesting alternative to extant heuristics in psychology and neuroscience, and in cognitive science using these algorithms to explain behavior has been termed *rational process models* (Sanborn, Griffiths, & Navarro, 2010). The advantage of this approach is that when these algorithms are used in situations for which they are well-adapted, they make probabilistic cognition achievable, but when they are applied to situations for which they are poorly adapted, they can explain biases in behavior that cannot be explained by probabilistic models alone.

Computer scientists and statisticians have developed various types of approximations for probabilistic models, such as Laplace's method, sampling algorithms, variational approximations, and expectation propagation (Bishop, 2006; Doucet, de Freitas, & Gordon, 2001; Minka, 2001; Neal, 1993; Wainwright & Jordan, 2008). Here I focus on the two types that have been applied to approximate probabilistic cognition: sampling and variational approximations. Sampling algorithms are stochastic, randomly drawing samples to represent a probability distribution as a collection of points. While sampling algorithms asymptotically provide the correct answer, they are less accurate and can show biases for small numbers of samples. In contrast, variational algorithms

E-mail address: a.n.sanborn@warwick.ac.uk

http://dx.doi.org/10.1016/j.bandc.2015.06.008
0278-2626/

trade stochastic sampling for deterministic optimization. These algorithms can be very fast, but are asymptotically biased.

Researchers have used both sampling and variational algorithms as approximations to probabilistic cognition in behavior and the brain. However these investigations have tended to proceed separately, with little comparison between the work using the two types of algorithms. Below, I describe examples of both types of algorithms, how they can produce behavioral biases, and how they might be implemented in the brain. A comparison of the two types shows what each is good for, and how they could be profitably combined in future work.

2. Sampling approximations

Sampling algorithms are useful for approximating calculations that involve complex probability distributions because the collection of samples can simply stand in for the complex distribution in a calculation. These approximate calculations are asymptotically correct with an infinite number of samples, but there are generally no guarantees for smaller numbers of samples.

While it is ideal if samples can be drawn from the distribution directly, often this is not the case and more sophisticated methods are required. One commonly used variety of sampling is *importance sampling*, which avoids the problem of drawing samples directly from a complex distribution by first sampling from a similar but simpler distribution (Bishop, 2006). These samples are weighted so that they reflect the probability of the complex distribution and not the actual distribution from which they were drawn. Importance sampling works well when the simpler distribution is very similar to the complex distribution, but is inaccurate if these distributions are very different.

A generalization of importance sampling is *particle filtering* (Doucet et al., 2001). This algorithm extends importance sampling into sequential tasks in which decisions need to be made after each observation of data. The simplest version of particle filtering draws samples from the prior distribution and sequentially reweights these samples by the likelihood of the data as it is observed. However, this version of particle filtering quickly runs into trouble because it is likely that the weight for one sample will dominate all of the rest, effectively yielding only a single sample. More sophisticated particle filters add steps such as replacing the worst samples with better-performing samples or perturbing the samples to provide a better overall approximation.

Another commonly used sampling algorithm is Markov chain Monte Carlo (*MCMC*; Neal, 1993). MCMC starts at a particular set of values (the initial state) for each of the random variables and makes a series of stochastic transitions to new states. By clever choice of the transition function, the series of states produced are samples from the distribution of interest. The strength of MCMC is that not as much needs to be known about the complex distribution ahead of time, but some downsides are that the initial samples need to be discarded and that samples are autocorrelated: because most MCMC samplers preferentially transition to nearby states, transitions between far-apart states are slower.

2.1. Explaining behavioral biases

Importance sampling, particle filtering, and MCMC have all been used to explain biases in human behavior. Importance sampling has been formally linked to exemplar models, which are well-supported models of memory and categorization. This link generalizes exemplar models to new tasks and allows it to explain behavioral biases. For example, in reproduction tasks participants' responses are drawn toward the distribution of stimuli they have previously been shown. The form of the assimilative effect shows deviations from what probabilistic models predict, but these deviations can be explained by assuming participants use a restricted number of samples (Shi, Griffiths, Feldman, & Sanborn, 2010).

Particle filters have been used to explain human biases in a variety of sequential tasks. Because repeated reweighting effectively reduces the number of samples, particle filters are useful for explaining how behavior can be more strongly influenced by early than late observations: samples consistent with the early observations initially dominate, and for some types of particle filter this makes it impossible to draw samples consistent with the late observations. Particle filters have been used to explain how early observations can dominate in categorization (Sanborn et al., 2010), sentence processing (Levy, Reali, & Griffiths, 2009), and causal learning (Abbott & Griffiths, 2011). Particle filters have also been used to explain individual variability around the group mean in learning (Daw & Courville, 2008) and change point detection (Brown & Steyvers, 2009).

MCMC has been used to explain different kinds of behavioral biases. Samples generated by MCMC are autocorrelated, and this property is useful for describing how judgments change slowly over time. One application of this is to bistable perception, where the current percept of a figure can be cast as a sample from a bimodal probability distribution over interpretations, and sampling using MCMC can explain the transition times between percepts (Gershman, Vul, & Tenenbaum, 2012). Autocorrelation also means that MCMC is initially influenced by its start state, which has been used to explain how irrelevant self-generated anchors in reasoning problems can have an effect on later answers (Lieder, Griffiths, & Goodman, 2012).

2.2. Implementation in the brain

Proposals have been made for how each of the above sampling algorithms could be implemented in the brain. For importance sampling, Shi and Griffiths (2009) proposed that neural tuning curves were proportional to the likelihood and that the number of neurons with a particular tuning curve were proportional to the prior. This scheme was extended to perform inference in a hierarchical model, which the levels of the model mapped to hierarchically organized brain regions.

Lee and Mumford (2003) used a similar global organization, proposing that at each level in the cortical hierarchy probabilistic cognition was implemented with a particle filter. Messages were then passed between the levels so that the top-down effects of context and the bottom-up effects of the stimulus were both incorporated. More detailed neural implementation of particle filters are given by Huang and Rao (2014) and Legenstein and Maass (2014) using networks of spiking neurons.

Other researchers have described on how populations of neurons could implement MCMC. In these implementations, the state of the brain corresponds to a sample from a probability distribution and transitions between neural states correspond to the transitions that the MCMC algorithm makes (Fiser, Berkes, Orbán, & Lengyel, 2010). Currently there are separate kinds of MCMC implementations for sampling from continuous variables (Hennequin, Aitchison, & Lengyel, 2014; Moreno-Bote, Knill, & Pouget, 2011) and sampling from discrete variables (Buesing, Bill, Nessler, & Maass, 2011; Probst et al., 2015).

3. Variational approximations

Variational approximations are a second major type of approximation in computer science and statistics, and these algorithms trade the stochasticity of sampling for the determinism of optimization. Variational algorithms work by first defining a simpler

Please cite this article in press as: Sanborn, A. N. Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition* (2015), http://dx.doi.org/10.1016/j.bandc.2015.06.008

family of distributions that is easier to use. Given this simpler family of distributions and a chosen distance measure, such as the Kullback–Leibler divergence, the calculus of variations is used to find the family member that most closely approximates the complex distribution (Bishop, 2006).

The choice of the simpler family of distributions can be made in different ways. One common method is to start with the complex distribution and then assume independence between variables that are not actually independent. This greatly reduces computational complexity. At its maximum extent, where all the variables are assumed to be independent from one another, this is called the *mean-field* approximation, while if independence is assumed between subsets of variables this is known as a *structured mean-field* approximation. A second way to define a simpler family of distributions is to assume a particular parametric form for the family of distributions. A common choice is to find the Gaussian distribution that is closest to a true distribution because then only the means and variances need to be encoded (Friston, 2008; Hinton & Van Camp, 1993).

Variational approximations have a close connection to the concept of message passing (Wainwright & Jordan, 2008). The variables in a complex distribution can be represented as nodes on a graph and dependencies can be represented as links between the variables. If that graph is a tree, then messages can be passed along the links to exactly infer the probability distribution for each variable. However, many times the dependencies between variables are more complex, and in these cases mean-field or structured mean-field approximations can be used to reduce dependencies between variables and to produce a message passing scheme that converges to an approximate solution. Additionally, assuming a parametric family of distributions allows for a different simplification of messages: instead of passing probability distributions, sufficient statistics can be passed between nodes.

Like sampling algorithms, variational approximations can be adapted to use with sequential data, an approach called *assumed density filtering*. This type of updating is known to work well if the observed data are randomly ordered but can settle in incorrect solutions if the data are ordered in a structured way (Minka, 2001).

3.1. Explaining behavioral biases

There has been less work using variational approximations to explain human biases than there has been using sampling algorithms. One application of deterministic approximations has been to explain puzzling effects of trial order in associative learning, particularly the effects of highlighting (Kruschke, 1996; Medin & Edelson, 1988) and forward and backward blocking (Shanks, 1985). There is not enough space to detail these effects here, but they result from a structured ordering of the stimuli and in combination they have not been successfully modeled with purely probabilistic models (Kruschke, 2006a).

Instead researchers have proposed variational approximations to probabilistic models to explain these three effects. For example, a model that interpolated between a mean-field variational approximation and the full probabilistic model was used by Daw, Courville, and Dayan (2008) to produce these three trial order effects. A different approach was taken by Kruschke (2006b) who assumed that the mind is divided into modules that are each performing exact inference, but are restricted in the messages they can pass to one another. The messages chosen in this model were motivated by their fit to the data, but Sanborn and Silva (2013) showed that using messages derived from a structured mean-field approach also produced the three effects.

3.2. Implementation in the brain

The most well-known implementation of a variational approximation in the brain is the combination of a mean-field approximation with the assumption that each variable is Gaussian (Friston, 2008). It assumes variables reside at each level of the cortical hierarchy and the very restricted approximation allows for simple messages consisting of the sufficient statistics to be passed up and down the levels of the hierarchy until convergence. This implementation is also linked to the idea of predictive coding, because the messages passed down to the lower cortical areas are the predictions of the model, while the messages passed upwards to the higher cortical areas are the deviations of the data from predictions.

The variational approach was also used by Beck, Pouget, and Heller (2012) to approximate inference in a network of neurons using probabilistic population codes. Their approximation was less restrictive than the one introduced by Friston (2008), as it used a structured mean-field approximation and allowed for any exponential family distribution, rather than only Gaussian distributions.

4. Discussion

Sampling and variational algorithms each have particular strengths and weaknesses that make them appropriate for explaining different behavioral biases and which require different implementations in the brain. Sampling algorithms more easily explain variability in behavior while variational approximations more easily produce stable biases. Sampling approaches also allow for more possibilities for learning, because their representations are less restricted. In terms of implementation, sampling algorithms map to a well-coordinated stochastic system while variational algorithms have greater flexibility in how information is propagated between variables: different variational approximations result in different message passing schemes.

While sampling and variational approximations have not often been combined to approximate probabilistic cognition, brain and behavior might be better explained by using both. In terms of implementation, a natural way to combine their strengths would be to assume that probabilistic cognition is implemented by sampling algorithms in local regions, but a global variational approximation determines the message passing scheme between regions. This scheme allows for more scope for learning within local regions, while controlling the complexity of the overall model. Behaviorally, it predicts stochasticity in order effects like highlighting and neurally it predicts that the connectivity structure will reflect the messages needed for the factorized approximation, rather than the messages needed for the full distribution.

This idea is an expansion of Lee and Mumford (2003), who proposed local regions used particle filters and messages were passed between regions using a message passing scheme called loopy belief propagation, which results from an expectation propagation approximation similar to the variational mean-field approximation (Minka, 2001). Instead of restricting approximations to these two particular types, a range of both sampling and variational approximations can be explored. The effects of these approximations on models should then be compared to both behavioral biases and what is known about brain processes to better understand how probabilistic cognition is approximated.

Acknowledgements

This work was supported by funding from Economic and Social Research Council grant ES/K004948/1. The author thanks the anonymous reviewers for very helpful comments.

Please cite this article in press as: Sanborn, A. N. Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition* (2015), http://dx.doi.org/10.1016/j.bandc.2015.06.008

References

Abbott, J. T., & Griffiths, T. L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. In *Proceedings of the 33rd annual conference of the cognitive science society*.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409–429.

Beck, J., Pouget, A., & Heller, K. A. (2012). Complex inference in neural circuits with probabilistic population codes and topic models. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.). *Advances in neural information processing systems* (Vol. 25, pp. 3059–3067). Curran Associates, Inc.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology, 58*, 49–67.

Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology, 7*(11), e1002211.

Daw, N. D., & Courville, A. C. (2008). The pigeon as particle filter. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.). *Advances in neural information processing systems* (Vol. 20, pp. 369–376). Cambridge, MA: MIT Press.

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning: The case of trial order. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind* (pp. 431–452). Oxford, UK: Oxford University Press.

Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. New York: Springer.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences, 14*, 119–130.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11), e1000211.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation, 24*, 1–24.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211.

Hennequin, G., Aitchison, L., & Lengyel, M. (2014). Fast sampling-based inference in balanced neuronal networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.). *Advances in neural information processing systems* (Vol. 27, pp. 2240–2248). Curran Associates, Inc.

Hinton, G. E., & Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on computational learning theory* (pp. 5–13).

Huang, Y., & Rao, R. P. (2014). Neurons as Monte Carlo samplers: Bayesian inference and learning in spiking networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.). *Advances in neural information processing systems* (Vol. 27, pp. 1943–1951). Curran Associates, Inc.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687–10692.

Körding, K., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*, 244–247.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 3–26.

Kruschke, J. K. (2006a). Locally Bayesian learning. In R. Sun (Ed.), *Proceedings of the 28th annual meeting of the cognitive science society* (pp. 453–458). Erlbaum.

Kruschke, J. K. (2006b). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review, 113*, 677–699.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20*(7), 1434–1448.

Legenstein, R., & Maass, W. (2014). Ensembles of spiking neurons with noise support optimal probabilistic inference in a dynamically changing environment. *PLoS Computational Biology, 10*(10), e1003859.

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21, pp. 937–944).

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In *Advances in neural information processing systems* (pp. 2690–2798).

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117*, 68–85.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference*. Unpublished doctoral dissertation. MIT, Boston.

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences, 108*(30), 12491–12496.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods*. Tech. rep. no. CRG-TR-93-1. Department of Computer Science, University of Toronto.

Probst, D., Petrovici, M. A., Bytschok, I., Bill, J., Pecevski, D., Schemmel, J., et al. (2015). Probabilistic inference in discrete spaces can be implemented into networks of LIF neurons. *Frontiers in Computational Neuroscience, 9*.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to the rational model of categorization. *Psychological Review, 117*, 1144–1167.

Sanborn, A. N., & Silva, R. (2013). Constraining bridges between levels of analysis: A computational justification for locally Bayesian learning. *Journal of Mathematical Psychology*.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgment. *Quarterly Journal of Experimental Psychology, 97B*, 1–21.

Shi, L., & Griffiths, T. L. (2009). Neural implementation of hierarchical Bayesian inference by importance sampling. In *Advances in neural information processing systems* (pp. 1669–1677).

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychological Bulletin and Review, 17*, 443–464.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.

Tassinari, H., Hudson, T. E., & Landy, M. S. (2006). Combining priors and noisy visual cues in a rapid pointing task. *The Journal of Neuroscience, 26*(40), 10154–10163.

Tversky, A., & Kahneman, D. (1978). Causal schemata in judgments under uncertainty. In *Progress in social psychology*. Lawrence Erlbaum.

Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science, 32*(6), 939–984.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning, 1*(1–2), 1–305.

Please cite this article in press as: Sanborn, A. N. Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition* (2015), http://dx.doi.org/10.1016/j.bandc.2015.06.008