Variational Reference Priors

Eric Nalisnick
Department of Computer Science
University of California, Irvine
enalisni@uci.edu

Padhraic Smyth
Department of Computer Science
University of California, Irvine
smyth@ics.uci.edu

1 Introduction

Posterior distributions are useful for a broad range of tasks in machine learning ranging from model selection to reinforcement learning. Given that modern machine learning models can have millions of parameters, selecting an informative prior is typically infeasible, resulting in widespread use of priors that avoid strong assumptions. For example, recent work on deep generative models (Kingma & Welling, 2014; Rezende et al., 2014) commonly uses the standard Normal distribution for the prior on the latent space. However, just because a prior is relatively flat does not mean it is uninformative. The *Jeffreys prior* for the Bernoulli model serves as a well-known counter example: Jeffreys (1946) showed that the arcsine distribution, despite its peaks near 0 and 1, is the truly objective prior (with respect to Fisher information) and not the uniform distribution. This suggests that objective priors such as the Jeffreys or the related *Reference prior* (Bernardo, 2005) are worthy of investigation for high-dimensional, web-scale probabilistic models. However, the challenge is that these priors are difficult to derive for all but the simplest models.

In this abstract, we describe how to learn an approximate Reference prior for any probabilistic model, thereby broadening the potential use of objective priors. Our optimization framework is akin to black-box (posterior) variational inference (Ranganath et al., 2014): we propose a parametric family of distributions and perform derivation-free optimization to find the member of the family closest to the true Reference prior. The proposed method learns a Reference prior for a given model *independent* of any data source[1]. We demonstrate below that the proposed framework recovers the Jeffreys prior better than existing numerical methods and show our method's utility for a previously intractable model, a *Variational Autoencoder* (Kingma & Welling, 2014) (VAE). In an interesting case study, we see that the VAE's reference prior differs markedly from the widely used standard Normal distribution.

2 Learning Reference Priors

A *Reference prior* (Bernardo, 2005; Berger et al., 2009) is a prior $p^*(\boldsymbol{\theta})$ on the model parameters $\boldsymbol{\theta}$ that maximizes the mutual information between $\boldsymbol{\theta}$ and the data $\mathcal{D}$. We can write the mutual information as the Kullback-Liebler divergence between the posterior and prior, averaged over the data distribution:

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\, I(\boldsymbol{\theta}, \mathcal{D}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}} \int_{\mathcal{D}} p(\mathcal{D})\mathrm{KLD}[p(\boldsymbol{\theta}|\mathcal{D}) \mid\mid p(\boldsymbol{\theta})]d\mathcal{D}. \tag{1}$$

where $I(\cdot, \cdot)$ denotes the mutual information and $p(\mathcal{D}) = \int_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal likelihood (or model evidence). Reference priors are equal to the Jeffreys in one dimension, and like the Jeffreys, are invariant to reparametrization because they maximize the mutual Information, which is itself invariant. Solving the calculus of variations problem for $p^*(\theta)$ is intractable for most models and not guaranteed to be a proper normalized distribution.

Inspired by recent advances in variational inference, we use similar ideas to optimize an approximate *prior*—call it $p_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ with parameters $\boldsymbol{\lambda}$—so that it is the distribution in the family closest to the model's Reference prior $p^*(\boldsymbol{\theta})$. Most crucially, we need an optimization objective, but Equation 1 is difficult to work with since it involves the intractable quantities $p(\mathcal{D})$ and $p(\boldsymbol{\theta}|\mathcal{D})$. We can algebraically re-arrange the Reference prior definition to the following more tractable form:

$$p^*(\boldsymbol{\theta}) = \underset{p(\boldsymbol{\theta})}{\operatorname{argmax}}\, \mathbb{E}_{p(\boldsymbol{\theta})}\left[KLD[p(\mathcal{D}|\boldsymbol{\theta}) \mid\mid p(\mathcal{D})]\right]. \tag{2}$$

[1] Except when the model is for a conditional distribution, i.e. $p(y|x)$. In this case, x samples are required to learn the variational Reference prior.

Now we have the KLD between the model likelihood and evidence averaged over the prior, thus ridding ourselves of the intractable posterior term. This form is the one we use for our variational objective: $\boldsymbol{\lambda}^* = \text{argmax}_{\boldsymbol{\lambda}} \mathbb{E}_{p_{\boldsymbol{\lambda}}(\boldsymbol{\theta})} KLD[p(\mathcal{D}|\boldsymbol{\theta}) \mid\mid p(\mathcal{D})] = \text{argmax}_{\boldsymbol{\lambda}} \mathcal{J}_{\text{RP}}(\boldsymbol{\lambda})$. $\boldsymbol{\lambda}^*$ are the parameters that make the variational family closest to the true Reference prior.

Deriving a Lowerbound. Equation 2, however, still contains the difficult term $p(\mathcal{D})$, and thus we need to derive a tractable lowerbound on the mutual information. Firstly, re-writing $\mathcal{J}_{\text{RP}}$ as $\mathbb{E}_{p_{\boldsymbol{\lambda}}}\mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta})}[\log p(\mathcal{D}|\boldsymbol{\theta})] - \mathbb{E}_{p_{\boldsymbol{\lambda}}}\mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta})}[\log p(\mathcal{D})]$, we see the second term is the intractable one. To cope, we use the following *variational Rényi bound* (Li & Turner, 2016) (VR) on the marginal likelihood: $\log p(\mathcal{D}) \le \frac{1}{1-\alpha} \log \mathbb{E}_{p_{\boldsymbol{\lambda}(\boldsymbol{\theta})}} \left[p(\mathcal{D}|\boldsymbol{\theta})^{1-\alpha} \right]$ for $\alpha < 0$. To circumvent numerical difficulties, we use Li & Turner (2016)'s *VR-max* Monte Carlo estimator corresponding to $\alpha \to -\infty$: $\log \mathbb{E}_{p_{\boldsymbol{\lambda}(\boldsymbol{\theta})}} [p(\mathcal{D}|\boldsymbol{\theta})^{\infty}] \approx \max_s \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s)$ where s indexes samples $\hat{\theta}_s \sim p_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$. Combing the VR-max term with the first yields a tractable lowerbound on $\mathcal{J}_{\text{RP}}$:

$$\mathcal{J}_{\text{RP}}(\boldsymbol{\lambda}) \ge \mathcal{J}_{\text{RP}}^{LB}(\boldsymbol{\lambda}) = \mathbb{E}_{p_{\boldsymbol{\lambda}}}\mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta})}[\log p(\mathcal{D}|\boldsymbol{\theta}) - \max_{s} \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_s)]. \quad (3)$$

This is the final objective that we can maximize to learn *variational Reference priors* (VRPs). $\mathcal{J}_{\text{RP}}^{LB}(\boldsymbol{\lambda})$ can be interpreted intuitively as follows: it calculates the data's log-likelihood under the model (i.e. parameter setting) that generated this data and under several samples from the prior. We then optimize to increase the difference in the log probabilities, forcing the prior to place most of its mass on parameters that generate identifiable datasets—or in other words, datasets that have high probability under only their true generative model.

Black-Box, Gradient-Based Optimization. We can rid ourselves of the need for analytical tractability by using Monte Carlo approximations of the nested expectations in Equation 3. Moreover, we can perform fully gradient-based optimization by drawing these samples via a differentiable non-centered parametrization (DNCP) (the so-called 'reparametrization trick') (Kingma & Welling, 2014). When dealing with discrete data or parameters, we use the *Concrete distribution* (Maddison et al., 2017), a differentiable relaxation of the discrete distribution.

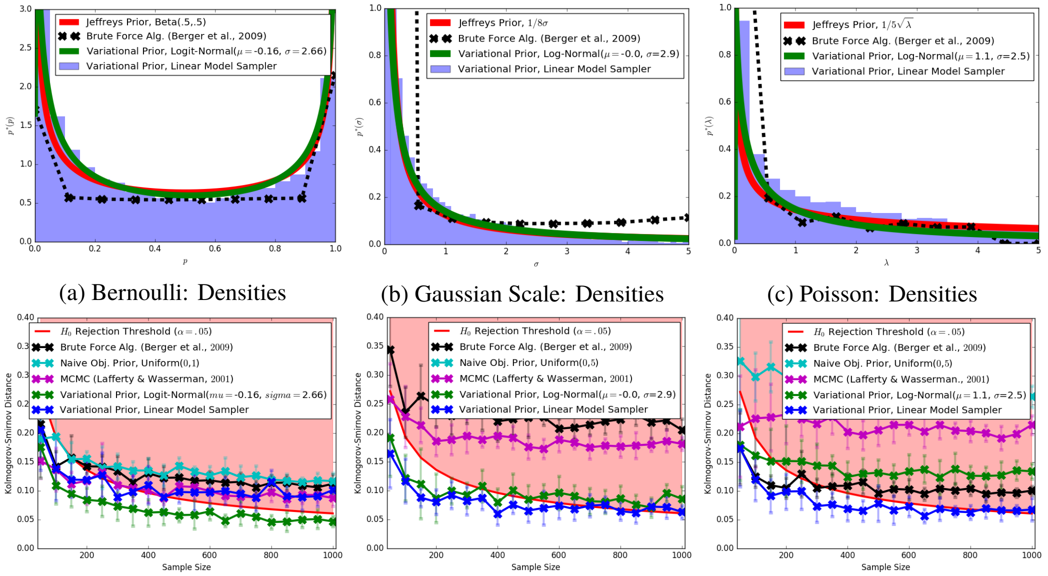
Implicit Priors. A crucial detail to note about Equation 3 is that we do not need to evaluate the prior's density. Rather, we need only to draw samples from it. This allows us to use black-box functional samplers as the variational family (Goodfellow et al., 2014; Ranganath et al., 2016), i.e. $\hat{\boldsymbol{\theta}} = f(\boldsymbol{\lambda}, \hat{\boldsymbol{\epsilon}})$ where $\boldsymbol{\epsilon} \sim p_0$ where f is some arbitrary differentiable function, such as a neural network, and p_0 is a fixed noise distribution. However, if our end goal is posterior inference, evaluation of $p_{\boldsymbol{\lambda}}$ needs to be done somehow—possibly by nonparametric density estimation techniques.

3 Experiments

Recovering Jeffreys Priors. We begin experimental evaluation by optimizing $\mathcal{J}_{\text{RP}}^{LB}(\boldsymbol{\lambda})$ (Equation 3) to recover the Jeffreys prior for three simple models: the Bernoulli mean parameter, $p^*(p) \propto \text{Beta}(.5, .5)$, the Gaussian scale parameter, $p^*(\sigma) \propto 1/\sigma$, and the Poisson rate parameter, $p^*(\lambda) = 1/\sqrt{\lambda}$. We use two types of variational approximations: an *implicit prior* (i.e. a linear model functional sampler) and a parametric density over the appropriate support space. We compare our variational approximations against three baselines: Berger et al. (2009)'s brute force algorithm for computing Reference priors, Lafferty & Wasserman (2001)'s MCMC algorithm for Reference priors, and a uniform distribution, which we chose as the naive baseline. Note that the brute force and MCMC algorithms require numerical integration and thus scale poorly to multiple dimensions.

Subfigures (a)-(c) of Figure 1 show the variational distributions plotted against the true Jeffreys prior (red) and values computed by Berger et al.'s algorithm (Berger et al., 2009) (black). We see that indeed both the samples drawn from the linear model (blue) and the learned parametric distributions (green) exhibit the same qualitative characteristics as the true Reference/Jeffreys priors.

We next quantify the gap in the approximations via a Kolmogorov-Smirnov two-sample test (KST) under the null hypothesis $H_0 : p = q$ where p is the true Jeffreys prior and q is an approximation. The KST computes the distance (KSD) between the distributions as $\text{KSD}(p, q) = \sup_x \mid \hat{F}_p(x) - \hat{F}_q(x) \mid$ where $\hat{F}_p(x)$ is the empirical CDF. Subfigures (d)-(f) of Figure 1 show the KSD between samples from the true Jeffreys prior and the various approximation techniques as sample size increases. The red line denotes the threshold at which the null hypothesis (that the distributions are equal) is rejected. The



(a) Bernoulli: Densities

(b) Gaussian Scale: Densities

(c) Poisson: Densities

(d) Bernoulli: KS Test

(e) Gaussian Scale: KS Test

(f) Poisson: KS Test

Figure 1: Optimizing Equation 3 to recover known Reference priors. Subfigures (a)-(c) show the learned density functions; Subfigures (d)-(f) show the Kolmogorov-Smirnov distance from the true distribution.

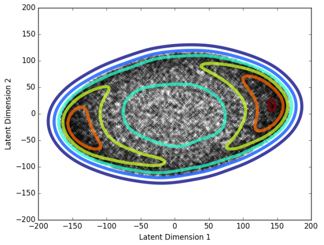


Figure 2: *The Variational Autoencoder's Reference Prior.* We learned a VRP for a VAE with a two-dimensional latent space (in order to visualize the prior). A one-hidden-layer NN sampler was used as the implicit prior. The plot to the left shows 25,000 samples from the prior. The contours are generated via kernel density estimation. We see the prior is overdispersed with two strong modes at opposite ends of the space.

VRPs are competitive to superior for all models and are the only techniques that remain statistically indistinguishable from the true Reference prior as the sample size increases.

Case Study: VRP for the VAE. Lastly, we demonstrate learning a VRP for an intractable, neural-network-based model: a *Variational Autoencoder* (Kingma & Welling, 2014). Figure 2 shows samples from the VAE's VRP, using a one-hidden-layer neural network as an implicit prior. We see that the VRP is drastically different than the standard Normal that is typically used as the prior over the latent variables: the VRP is multimodal and has a much larger variance. Yet, the difference is intuitive: placing most prior mass at opposite sides of the latent space encourages the VAE to space it's latent representations with as much distance as possible, ensuring they are as identifiable w.r.t. the model likelihood, the VAE decoder, as possible. Interestingly, recent work by Hoffman & Johnson (2016) suggests that VAEs can be improved by multimodal priors: "[T]he [VAE's] individual encoding distributions $q(\mathbf{z}_i|\mathbf{x}_i)$ do not have significant overlap. . . .then perhaps we should investigate multimodal priors that can meet $q(\mathbf{z})$ halfway." This suggests using multimodal, dispersed priors encourages flexibility and objectivity in the posterior distribution.

In conclusion, variational methods provide a potentially interesting avenue for the development of reference priors for broader classes of problems than have been traditionally explored. In this abstract we have outlined a general framework for learning such variational reference priors suggesting a number of potentially interesting directions for further exploration.

References

James O Berger, José M Bernardo, and Dongchu Sun. The formal definition of reference priors. *The Annals of Statistics*, pp. 905–938, 2009.

José M Bernardo. Reference analysis. *Handbook of statistics*, 25:17–90, 2005.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Neural Information Processing Systems (NIPS)*, 2014.

Matthew Hoffman and Matthew Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. *NIPS 2016 Workshop on Advances in Approximate Bayesian Inference*, 2016.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 186, pp. 453–461. The Royal Society, 1946.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations (ICLR)*, 2014.

John Lafferty and Larry Wasserman. Iterative markov chain monte carlo computation of reference priors and minimax risk. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 293–300, 2001.

Yingzhen Li and Richard E Turner. Variational inference with renyi divergence. *Neural Information Processing Systems (NIPS)*, 2016.

C. J. Maddison, A. Mnih, and Y. Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *International Conference on Learning Representations (ICLR)*, 2017.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 814–822, 2014.

Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pp. 496–504, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning (ICML)*, 2014.