Inférence variationnelle de lois a priori de référence

Clément Gauchy [1], Antoine Van Biesbroeck [1,2], Cyril Feau [3] & Josselin Garnier [2]

[1] *Université Paris-Saclay, CEA, Service de Génie Logiciel pour la Simulation, 91191 Gif-sur-Yvette, France*
clement.gauchy@cea.fr

[2] *CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France*
antoine.van-biesbroeck@polytechnique.edu, josselin.garnier@polytechnique.edu

[3] *Université Paris-Saclay, CEA, Service d'Études Mécaniques et Thermique, 91191 Gif-sur-Yvette, France*
cyril.feau@cea.fr

Résumé. Dans de nombreux cas, les lois a priori informatives dans le cadre Bayésien sont difficiles à éliciter. Pour résoudre ce problème, la théorie des loi a priori dites de référence propose une solution agnostique consistant à choisir la loi a priori directement à partir de la vraisemblance. Cependant, ces lois a priori de référence sont souvent difficiles à calculer analytiquement pour de nombreux modèles statistiques. Nous proposons dans cet article un algorithme d'inférence variationnelle qui permet d'approximer ces lois a priori de référence. Nous illustrons la procédure sur des cas tests simples, montrant l'utilité de la méthode en approximant la loi a priori noninformative de Jeffreys dans l'asymptotique d'un grand nombre de données. Une attention particulière est accordée aux cas où la loi a priori de Jeffreys est impropre.

Mots-clés. Lois a priori de référence, Lois a priori de Jeffreys, Inférence variationnelle

Abstract. In many cases, informative Bayesian priors are difficult to elicit. To tackle this issue, reference prior theory proposes an agnostic solution consisting in deriving the prior distribution directly from the likelihood function. However, such reference priors are oftenly difficult to derive analytically for many statistical models. We propose in this article a variational inference algorithm that learns non-asymptotic reference priors. We illustrate the procedure on simple test cases, showing the usefulness of the method by recovering the Jeffreys prior in the asymptotics of a large number of data. A particular attention is given to the cases when the Jeffreys prior is improper.

Keywords. Reference prior, Jeffreys prior, Variational inference

1 Introduction

Reference priors - introduced in the seminal paper of Bernardo (1979) - are Bayesian priors that are obtained in an objective fashion: For a given likelihood function, it corresponds to

the prior distribution that maximises the influence of the data on the posterior distribution. It is consequently also the prior distribution that influences the less the posterior distribution. This behavior is of great interest when the statistician wants to express a state of ignorance on the statistical model's parameters. A more rigorous definition of the reference prior is stated in Berger et al. (2009). The main drawback of the reference prior theory is the difficulty to derive analytically the reference prior for most of the statistical models. Inspired by machine learning litterature, a solution to this issue is to learn an approximation of the non-asymptotic reference prior. The main contribution of this article is the construction of a variational inference procedure aiming to approximate the non-asymptotic reference prior in the same fashion as in Nalisnick & Smyth (2017). However, we decide to put more attention to the cases where the reference prior is improper, using an appropriate parametric study. The article is organized as follows: Section 2 details the mathematics behind reference prior theory, then a variational inference algorithm based on stochastic gradient descent and an implicit prior distributions family is presented in Section 3. Finally, Section 4 is devoted to numerical applications on simple test cases.

2 Reference prior

The goal of this section is to present the definition of the non-asymptotic reference prior. We use the following notations for the rest of the paper. The likelihood of the dataset $\mathcal{D} = \{X_i\}_{1 \leq i \leq N}$ of N i.i.d. $\mathcal{X}$-valued observations (with $\mathcal{X} \subset \mathbb{R}^d$) is $p(\mathcal{D}|\theta) = \prod_{i=1}^N p(X_i|\theta)$ where $\theta \in \Theta \subset \mathbb{R}^q$ is the model parameter. $\pi(\theta)$ is the prior, $p(\theta|\mathcal{D})$ is the posterior and $p(\mathcal{D}) = \int_\Theta p(\mathcal{D}|\theta)\pi(\theta)d\theta$ is the marginal likelihood (or model evidence). The entropy of the likelihood is given by $\mathbb{H}_{\mathcal{D}|\theta}[\mathcal{D}] = -\int_{\mathcal{X}^N} p(\mathcal{D}|\theta) \log p(\mathcal{D}|\theta) d\mathcal{D}$. The non-asymptotic reference prior of the statistical model $\mathcal{M} = \{p(\cdot|\theta), \theta \in \Theta\}$ is the prior that maximises the mutual information $\mathcal{I}_\pi(\theta, \mathcal{D})$:

$$\mathcal{I}_\pi(\theta, \mathcal{D}) = \mathbb{E}_{\theta \sim \pi}[-\mathbb{H}_{\mathcal{D}|\theta}[\mathcal{D}] - \mathbb{E}_{\mathcal{D}|\theta}[\log p(\mathcal{D})]] \ .$$

The non-asymptotic reference prior is the prior that lets the data influence the most the posterior distribution:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathcal{I}_\pi(\theta, \mathcal{D}) \ . \tag{1}$$

One can rewrite the mutual information using the Kullback-Leibler divergence:

$$\mathcal{I}_\pi(\theta, \mathcal{D}) = \int_{\mathcal{X}^N} \mathrm{KL}\left[p(\theta|\mathcal{D})||\pi(\theta)\right] p(\mathcal{D}) d\mathcal{D} \ . \tag{2}$$

With this formulation, one can see that the non-asymptotic reference prior defined in Equation (1) is the prior that maximizes the Kullback-Leibler divergence between the posterior distribution and the prior distribution, averaged on the marginal distribution of the data $\mathcal{D}$. Remark that π^* depends of the size N of the dataset $\mathcal{D}$. The reference prior according to Berger et al. (2009) can be defined - in some sense not detailed in this article - as the limit prior when $N \to +\infty$.

3 Variational inference

We remark that the mutual information $\mathcal{I}_\pi(\theta, \mathcal{D})$ depends only on samples under the prior distribution π in order to be computed. It is thus possible to define a deterministic coupling in order to transport a simple measure (such as a multidimensional Gaussian) to the complex target measure (here the non-asymptotic reference prior) as explained in Marzouk et al. (2016). It boils down to use a smooth function g such that $\theta = g(\lambda, \varepsilon)$ where $\lambda \in \Lambda \subset \mathbb{R}^p$ is a vector of parameters and ε a random variable with fixed distribution $\mathbb{P}_\varepsilon$. The probability measure $\mathbb{P}_\varepsilon$ can be chosen so that it is simple to sample from it. The function g can be chosen arbitrarly, for instance it could be a neural network (NN) with λ defining the weights and bias parameters of the NN. The idea is that the pushforward measures $\{g(\lambda, \cdot)\sharp\varepsilon, \lambda \in \Lambda\}$ form a sufficiently vast parametric family of probability distributions and hence could approximate well the non-asymptotic reference prior for an arbitrary statistical model. We denote by simplicity this pushforward measure by π_λ. In order to perform a variational inference algorithm, we have to compute the gradient

$$\frac{\partial}{\partial\lambda}\mathcal{I}_{\pi_\lambda}(\theta, \mathcal{D}) = \frac{\partial}{\partial\lambda}\Big[\int_\Theta \int_{\mathcal{X}^N} \pi_\lambda(\theta)p(\mathcal{D}|\theta)\log\frac{p(\mathcal{D}|\theta)}{p_\lambda(\mathcal{D})}d\mathcal{D}d\theta\Big] , \tag{3}$$

where

$$p_\lambda(\mathcal{D}) = \int_\Theta p(\mathcal{D}|\theta')\pi_\lambda(\theta')d\theta' .$$

We obtain

$$\begin{aligned}\frac{\partial}{\partial\lambda}\mathcal{I}_{\pi_\lambda}(\theta, \mathcal{D}) = &\int_\Theta \int_{\mathcal{X}^N} \frac{\partial\pi_\lambda}{\partial\lambda}(\theta)p(\mathcal{D}|\theta)\log\frac{p(\mathcal{D}|\theta)}{p_\lambda(\mathcal{D})}d\mathcal{D}d\theta \\ &- \int_\Theta \int_{\mathcal{X}^N} \pi_\lambda(\theta)p(\mathcal{D}|\theta)\frac{\partial\log p_\lambda(\mathcal{D})}{\partial\lambda}d\mathcal{D}d\theta .\end{aligned}$$

Remark now that

$$\begin{aligned}-\int_\Theta \int_{\mathcal{X}^N} \pi_\lambda(\theta)p(\mathcal{D}|\theta)\frac{\partial\log p_\lambda(\mathcal{D})}{\partial\lambda}d\mathcal{D}d\theta &= -\int_{\mathcal{X}^N} p_\lambda(\mathcal{D})\frac{\partial\log p_\lambda(\mathcal{D})}{\partial\lambda}d\mathcal{D} \\ &= -\int_{\mathcal{X}^N} \frac{\partial p_\lambda(\mathcal{D})}{\partial\lambda}d\mathcal{D} \\ &= -\frac{\partial}{\partial\lambda}\Big[\int_{\mathcal{X}^N} p_\lambda(\mathcal{D})d\mathcal{D}\Big] \\ &= 0 .\end{aligned}$$

This gives the following equality

$$\frac{\partial}{\partial\lambda}\mathcal{I}_{\pi_\lambda}(\theta, \mathcal{D}) = \frac{\partial}{\partial\lambda}\mathbb{E}_{\theta\sim\pi_\lambda}[h(\theta)] = \frac{\partial}{\partial\lambda}\mathbb{E}_\varepsilon[h \circ g(\lambda, \varepsilon)] , \tag{4}$$

where

$$h(\theta) = \int_{\mathcal{X}^N} p(\mathcal{D}|\theta)\log\frac{p(\mathcal{D}|\theta)}{p_\lambda(\mathcal{D})}d\mathcal{D} ,$$

and the derivative with respect to λ in (4) does not apply to the function h. In order to allow for a fully gradient-based optimization with an implicit prior formulation, we have to use the chain rule as follows:

$$\frac{\partial h \circ g}{\partial \lambda_j}(\lambda, \varepsilon) = \sum_{l=1}^{q} \frac{\partial h}{\partial \theta_l}(g(\lambda, \varepsilon)) \frac{\partial g_l}{\partial \lambda_j}(\lambda, \varepsilon) \, , \quad j = 1, \ldots, p \, .$$

The goal is now to build an estimator of the gradient $\frac{\partial h}{\partial \theta}$. Denote $\dot{p}(\mathcal{D}|\theta) = \frac{\partial p(\mathcal{D}|\theta)}{\partial \theta}$. Then,

$$\begin{aligned}
\frac{\partial h}{\partial \theta}(\theta) &= \int_{\mathcal{X}^N} \frac{\partial}{\partial \theta} \Big(p(\mathcal{D}|\theta) \log \frac{p(\mathcal{D}|\theta)}{p_\lambda(\mathcal{D})} \Big) d\mathcal{D} \\
&= \int_{\mathcal{X}^N} \dot{p}(\mathcal{D}|\theta) d\mathcal{D} + \int_{\mathcal{X}^N} \dot{p}(\mathcal{D}|\theta) \log p(\mathcal{D}|\theta) d\mathcal{D} - \int_{\mathcal{X}^N} \dot{p}(\mathcal{D}|\theta) \log p_\lambda(\mathcal{D}),
\end{aligned}$$

by the equality $\frac{\partial \log p(\mathcal{D}|\theta)}{\partial \theta} = \frac{\dot{p}(\mathcal{D}|\theta)}{p(\mathcal{D}|\theta)}$. Denoting $s(\mathcal{D}|\theta) = \frac{\partial \log p(\mathcal{D}|\theta)}{\partial \theta}$, we have

$$\int_{\mathcal{X}^N} \dot{p}(\mathcal{D}|\theta) d\mathcal{D} = \int_{\mathcal{X}^N} s(\mathcal{D}|\theta) p(\mathcal{D}|\theta) d\mathcal{D} = 0 \, .$$

Hence

$$\frac{\partial h}{\partial \theta}(\theta) = \int_{\mathcal{X}^N} s(\mathcal{D}|\theta) \log \frac{p(\mathcal{D}|\theta)}{p_\lambda(\mathcal{D})} p(\mathcal{D}|\theta) d\mathcal{D} \, .$$

A Monte-Carlo estimator of the gradient can be defined as follows:

$$\frac{\partial h}{\partial \theta}(\theta) \approx \frac{1}{J} \sum_{j=1}^{J} s(\mathcal{D}_j|\theta) \log \frac{p(\mathcal{D}_j|\theta)}{p_\lambda(\mathcal{D}_j)} \; ; \; \mathcal{D}_j \sim \bigotimes_{i=1}^{N} p(\cdot|\theta) \, . \tag{5}$$

The marginal likelihood $p_\lambda(\mathcal{D})$ can be estimated also with Monte Carlo:

$$p_\lambda(\mathcal{D}) \approx \frac{1}{T} \sum_{t=1}^{T} p(\mathcal{D}|\theta_t) \; ; \; \theta_t \sim \pi_\lambda \, . \tag{6}$$

This is the major difference with the objective function defined in Nalisnick & Smyth (2017). Indeed, the authors prefer to lower bound the mutual information $\mathcal{I}_\pi(\theta, \mathcal{D})$ using the following inequality that holds for any λ:

$$-\log(p_\lambda(\mathcal{D})) \geq -\max_{\theta \in \Theta} \log(p(\mathcal{D}|\theta)) \, .$$

The main argument of this inequality is a reduction of variance and an improvement of stability in the non-asymptotic reference prior's estimate. However, the numerical illustrations presented in this article are for low-dimensional θ's. The Monte-Carlo estimation of the marginal likelihood will then be used in the stochastic gradient descent algorithm.

4 Numerical illustrations

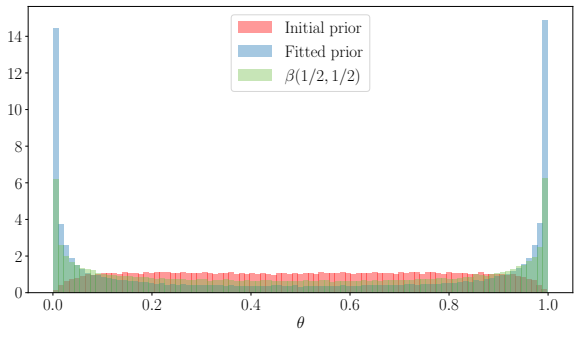
This section is dedicated to an experimental study of the variational inference algorithm based on the mutual information's gradient estimator defined in Equation (5). For the sake of simplicity and discussion about the efficiency of the non-asymptotic reference prior approximation procedure, the numerical experiments will be done on classical one-dimensional models, namely the Bernoulli model where $X_i \sim \mathcal{B}(\theta)$, $\theta \in (0,1)$, and the Gaussian model with known mean and unknown variance where $X_i \sim \mathcal{N}(0,\theta)$, $\theta \in (0,+\infty)$. These models have the main advantage to have analytical expressions for their reference priors which are the well-known Jeffreys priors.

Bernoulli model

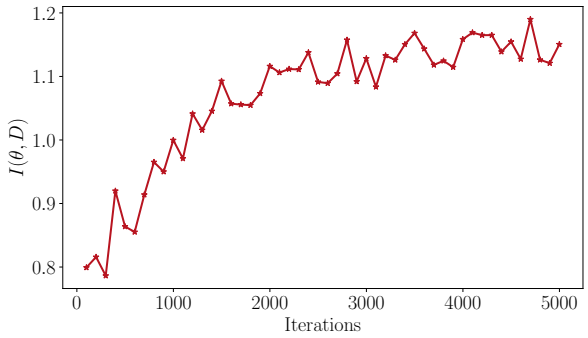
The asymptotic reference prior for this model is the Jeffreys prior $\theta \sim \text{Beta}(1/2,1/2)$. Remarkably, this reference prior is proper, which allows for a formal performance evaluation of the approximation procedure presented in this article. For computing the reference prior of the Bernoulli model, we used the Adam stochastic gradient optimization algorithm Kingma & Ba (2014) with classical settings $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a learning rate of 0.01. The deterministic coupling g in this setting is

$$g(\lambda,\varepsilon) = \frac{\tanh(\lambda^T\varepsilon)+1}{2}\ ,$$

where $\varepsilon \sim \mathcal{N}(0, I_{10\times 10})$ and $\lambda \in \mathbb{R}^{10}$. This coupling can be seen as a very simple neural network with tanh as the activation function. The optimization is initialized with a parameter λ_0 sampled from $\mathcal{N}(0, 0.2^2)$ such that the initial prior distribution π_{λ_0} is far to the reference prior. This allows to benchmark the variational inference robustness. Figure 1b represents the



(a) Histograms of the prior evaluated at a initial λ_0 value (in red), the fitted prior (in blue) and the Jeffreys prior Beta$(1/2,1/2)$ for this model.



(b) Estimation of the mutual information $\mathcal{I}(\theta,\mathcal{D})$ with $N = 10$, $J = 1000$, $T = 100$ and $S = 100$.

Figure 1: Numerical results for the Bernoulli model.

estimation of the mutual information during the iterations of the Adam stochastic gradient algorithm. Note that the mutual information reaches its maximum value after around 3000

iterations. Figure 1a represents three histograms obtained with samples of size 10^5 from the initial prior π_{λ_0}, the prior fitted using the variational inference procedure with 5000 iterations and the true reference prior $\text{Beta}(1/2, 1/2)$. The mutual information gradient defined in Equation (5) is estimated using $N = 10$ observations, $J = 1000$ replications and $T = 100$ samples of θ. Note that the fitted prior has more mass for boundary values $\theta \to 1$ or $\theta \to 0$ than the Jeffreys prior. This difference can come from the parametric approximation of the reference prior and also from the fact that the Jeffreys prior is defined as the limit of reference priors for $N \to +\infty$.

Gaussian model with known mean and unknown variance

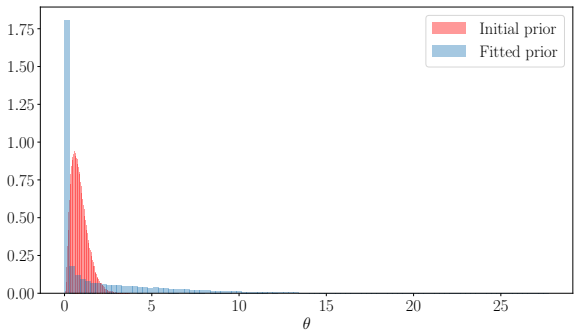
The reference prior for this model is the Jeffreys prior $\pi_{\text{J}}(\theta) \propto \frac{1}{\theta}$ which is improper. This particularity makes the variational inference procedure challenging both on a theoretical and practical ground. The approximation of the reference prior is also computed using Adam optimization algorithm with the same values of β_1 and β_2 and learning rate as the ones used in the Bernoulli model. The deterministic coupling g is here chosen as

$$g(\lambda, \tilde{\lambda}, \varepsilon) = \log(1 + \exp(\lambda^T \varepsilon + \tilde{\lambda})) \ ,$$

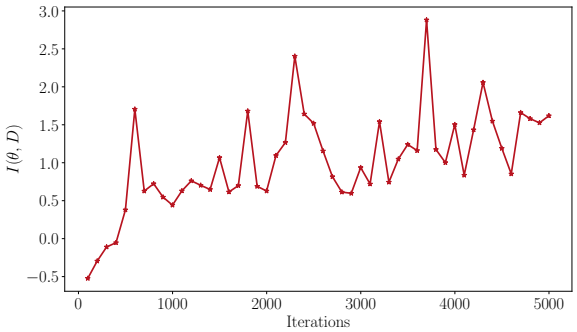
where $\varepsilon \sim \mathcal{N}(0, I_{100\times 100})$, $\lambda \in \mathbb{R}^{100}$ and $\tilde{\lambda} \in \mathbb{R}$. This coupling can also be seen as a simple neural network with a softplus activation function. Figure 2b shows an estimation of the mutual information every 100 iterations of the stochastic gradient algorithm. Note that the mutual information estimation is more noisy than for the Bernoulli model. Figure 2a represents histograms from samples of size 10^5 of the initial prior π_{λ_0} and the fitted prior after 5000 iterations of the Adam stochastic gradient algorithm. The mutual information gradient is estimated with the same numerical values as for the Bernoulli model. Because the Jeffreys prior in this case is improper, we cannot represent it graphically as for the Bernoulli model. However, note that in the Gaussian model with unknown variance, the Jeffreys prior can be seen as the limit of the conjugate inverse-gamma priors $\mathcal{IG}(\alpha, \beta)$ when $(\alpha, \beta) \to (0, 0)$ (see (Gelman et al. 2004, Section 2.8)). A parametric Maximum-Likelihood Estimation (MLE) on a sample of size 1000 of the fitted implicit prior with an inverse-gamma model is performed to assess its proximity with the Jeffreys prior. Figure 3 shows the evolution of the MLE of the (α, β) parameters during the optimization of the implicit prior π_λ every 100 iterations. The shape parameter α seems to reach an asymptotic value around 0.1 after 2000 iterations while the scale parameter β decreases to 0 faster.

5 Conclusion

This article is devoted to the development of a variational inference procedure that can be used to provide a parametric fit of the reference prior, whose analytical expression is cumbersome for most of the statistical models. After a brief review on reference prior theory, a stochastic gradient descent algorithm is proposed with an objective function derived from

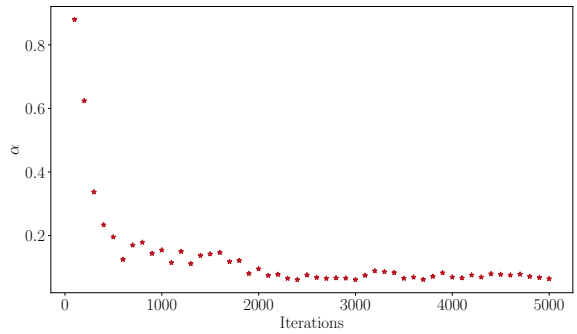


(a) Histograms of the prior evaluated at an initial λ_0 value (in red) and the fitted prior (in blue).

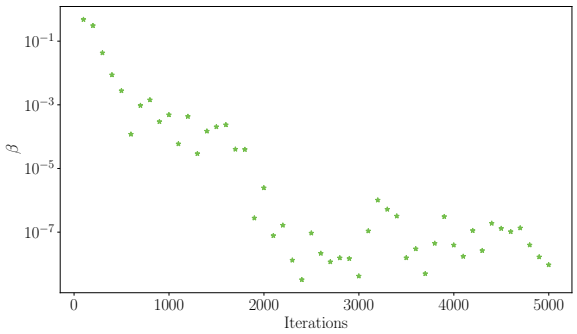


(b) Estimation of the mutual information $\mathcal{I}(\theta, \mathcal{D})$ with $N = 10$, $J = 1000$, $T = 100$ and $S = 100$.

Figure 2: Numerical results for the Gaussian model with known mean and unknown variance.



(a) Numerical values of the shape parameter α of the inverse gamma probability distribution, obtained by MLE from a sample of the implicit prior during its optimization.



(b) Numerical values of the scale parameter β of the inverse gamma probability distribution, obtained by MLE from a sample of the implicit prior during its optimization.

Figure 3: Numerical results for the Gaussian model with known mean and unknown variance.

the mutual information criterion. Remarkably, the prior density does not need to be known in order to solve the optimization problem, we only require to sample from the prior distribution. Implicit prior parametric families issued from a parametric function such as a neural network can thus be defined thanks to the pushforward measure issued from a simple measure, such as a multivariate Gaussian. In the case where the reference prior is proper the procedure proposed in this article seems to perform adequatly. However, the main limitations of this procedure come from the improprierty of reference priors in most of the statistical models, this peculiar behavior is studied on the Gaussian model with known mean and unknown variance. The perspectives of this work are numerous: First, it would be preferable to find the reference prior in the sense of Equation (1) in the space of proper probability distributions, using constrained optimization. Second, another perspective could be to enforce moments constraints in Equation (1) to plug experts knowledge in the reference prior. Finally, the definition of the non-asymptotic reference prior could be generalized with other metrics than

the one of Kullback-Leibler, such as the Maximum Mean Discrepancy Gretton et al. (2012) which has the advantage to be easily computed with expectations, leading potentially to efficient variational inference algorithms. One of the main goals of this work is similar to the one of Nalisnick & Smyth (2017), mainly deriving reference prior distributions for high-dimensional models such as Bayesian neural networks.

References

Berger, J. O., Bernardo, J. M. & Sun, D. (2009), 'The formal definition of reference priors', *The Annals of statistics* **37**(2), 905–938.

Bernardo, J. M. (1979), 'Reference posterior distributions for Bayesian inference', *Journal of the Royal Statistical Society. Series B* **41**(2), 113–147.

Gelman, A., C., J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman and Hall/CRC.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. (2012), 'A kernel two-sample test', *Journal of Machine Learning Research* **13**(25), 723–773.

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Marzouk, Y., Moselhy, T., Parno, M. & Spantini, A. (2016), 'Sampling via measure transport: An introduction', *Handbook of uncertainty quantification,* Springer Cham, pp. 1–41.

Nalisnick, E. & Smyth, P. (2017), 'Learning approximately objective priors', *arXiv preprint arXiv:1704.01168* .