A Review of Implicit Models in Approximate Bayesian Inference

Louis C. Tiao
School of Information Technologies
University of Sydney

October 16, 2017

1 Introduction

Probabilistic models play a pivotal role in machine learning. They are at the core of powerful algorithms that can uncover hidden structures, learn useful representations, and efficiently utilize these to make accurate predictions or generate realistic observations. As datasets and problems continue to grow in both volume and complexity, probabilistic models have needed to rely increasingly on approximate Bayesian inference methods in order to scale and meet their demands.

In particular, there has been a great deal of renewed interest in variational inference, an approach that reformulates the problem of approximating intractable posterior densities as an optimization problem, whereby a search is performed over a family of simpler distributions to find the member of that family closest to the posterior. In recent years, significant advances have been made toward using stochastic optimization methods to scale variational inference to large datasets, deriving generic methods to easily fit general classes of models, and using neural networks to specify flexible parametric families of approximate densities.

Concurrent with these advances, there has been tremendous research interest in the use of *implicit* probabilistic models in machine learning. Implicit models admit high fidelity to the data generating process, but do not yield tractable probability densities. As we better understand the implications of learning in implicit models, we are beginning to recognize their deep connections to density ratio estimation and approximate divergence minimization.

In the past year, the scope and applicability of variational inference has expanded dramatically. By leveraging methods of density ratio estimation and learning in implicit models, variational inference can now be applied in settings where no likelihood is available, where the family of posterior approximations is arbitrarily expressive and does not yield a density, and indeed, even where no prior density is available.

In this review, we discuss the recent methodologies for incorporating implicit models in variational inference. In section 2 we outline the general problem of inference in Bayesian models. We highlight the difficulties involved, and discuss how the methods of variational inference address these. Then, we further highlight the main intractabilities that hamper the applicability of variational inference, and review in section 3 the recent body research that has been undertaken to address these. Finally, in section 4 we present the latest research on implicit models in variational inference, discuss how these address the existing shortcomings of variational inference methods,

and further improves its potential for widespread adoption as a primary tool for statistical inference.

2 Background

Our presentation is closely based on a recent review of variational inference by Blei et al. (2017), and the PhD thesis of Kingma (2017). These should be consulted for a more complete treatment.

2.1 Bayesian modelling

We first describe the general problem of inference in Bayesian models. Let $\mathbf{x}$ be a set of observed variables and $\mathbf{z}$ be a set of latent variables, with joint density

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}). \tag{1}$$

In Bayesian models, the distribution of the observed data is assumed to be governed by the latent variables. Latent variables are drawn from a prior density $p(\mathbf{z})$ and related to the observations though the likelihood $p(\mathbf{x} \mid \mathbf{z})$. The problem of inference in Bayesian models amounts to computing the *posterior* $p(\mathbf{z} \mid \mathbf{x})$, the condition density of latent variables given the observed data. By Bayes' theorem, it can be written as

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z})}{p(\mathbf{x})}. \tag{2}$$

To compute this conditional density, we must calculate the denominator $p(\mathbf{x})$, known as the model *evidence*, which contains the marginal density of the observations. It can be calculated by marginalizing out the latent variables from the joint density,

$$p(\mathbf{x}) = \int p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z}. \tag{3}$$

For many models, this integral either requires exponential time to compute (computationally intractable), or it cannot be evaluated in closed-form (analytically intractable). Evaluating the evidence integral is the fundamental challenge for performing inference in the myriad of complex models that now pervade modern statistics and machine learning.

When it is not possible to carry out exact inference, we must turn to approximate inference methods. Currently, the two dominant paradigms for approximate inference are Markov chain Monte Carlo (MCMC), a sampling-based method, and more recently, variational inference (VI), an optimization-based method. We shall focus on the latter.

2.2 Variational inference

The basic idea of VI is to formulate inference as an optimization problem (Jordan et al., 1999; Wainwright and Jordan, 2008). We first specify a family $\mathcal{Q}$ of densities over the latent variables. Each member $q(\mathbf{z} \mid \mathbf{x}) \in \mathcal{Q}$ is a candidate approximation to the exact posterior $p(\mathbf{z} \mid \mathbf{x})$. We then

optimize over this family to find the member that minimizes the Kullback-Leibler (KL) divergence to the exact posterior,

$$q^*(\mathbf{z}|\mathbf{x}) = \underset{q(\mathbf{z}|\mathbf{x})\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}\left[q(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}|\mathbf{x})\right]. \tag{4}$$

Having found the optimal approximate density $q^*(\mathbf{z}|\mathbf{x})$, we use it as a proxy for the exact posterior density. However, a difficulty remains – expanding out all terms in the KL divergence in eq. (4) reveals its dependence on $\log p(\mathbf{x})$, the logarithm of the model evidence in eq. (3) (see appendix A). Recall that we appealed to approximate inference for the very reason that computing the evidence is intractable.

Evidence lower bound. Since we cannot compute the KL divergence without computing the evidence, directly minimizing it is not possible. Therefore, we must resort to optimizing an alternative objective function. Consider the evidence lower bound (ELBO),

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}|\mathbf{x})]. \tag{5}$$

It is easy to show that (see appendix B)

$$\log p(\mathbf{x}) = \mathrm{ELBO}(q) + \mathrm{KL}\left[q(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}|\mathbf{x})\right]. \tag{6}$$

Hence, the ELBO is the negative of the KL plus $\log p(\mathbf{x})$, a constant with respect to $q(\mathbf{z}|\mathbf{x})$. It follows that maximizing the ELBO is equivalent to minimizing the KL divergence.

Additionally, since the KL divergence is nonnegative, $\mathrm{KL}(\cdot) \geq 0$, it further follows that the ELBO is a lower bound of the log evidence, $\log p(\mathbf{x}) \geq \mathrm{ELBO}(q)$ for any q. This bound can also be shown using Jensen's inequality, as was done in the pioneering work on variational inference (Jordan et al., 1999).

Expanding out all the terms in the ELBO reveals important insight about the optimal approximate density,

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}) - \log q(\mathbf{z}|\mathbf{x})]. \tag{7}$$

We can characterize the first term as the expected log likelihood (ELL) of the data, while the second term is the negative cross-entropy between $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$, and the final term is the entropy of $q(\mathbf{z}|\mathbf{x})$. Taken together, the last two terms give us the negative KL divergence between $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$,

$$\begin{aligned}\mathrm{ELBO}(q) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \mathbb{H}[q(\mathbf{z}|\mathbf{x}), p(\mathbf{z})] + \mathbb{H}[q(\mathbf{z}|\mathbf{x})] &(8)\\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \mathrm{KL}\left[q(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})\right]. &(9)\end{aligned}$$

The ELL term in eq. (9) encourages the approximate density to place its mass on configurations of the latent variables that explain the observed data, while the negative KL divergence term encourages densities that resemble the prior. Combined, these terms form the ELBO and reflect the usual balance between the prior and likelihood.

2.2.1 Further intractabilities

Examining the terms of the ELBO also foreshadows potential sources of further intractabilities. We briefly discuss each of these now. The sections that follow will review the body of research that has been undertaken to address these.

Joint density. For the class of *conditionally conjugate* models in the exponential family, it is easy to compute the joint density in eq. (5) and optimize the ELBO using a closed-form coordinate ascent algorithm (Ghahramani and Beal, 2001). However, even models such as Bayesian logistic regression lie beyond this class. This term is generally intractable, and much tedious work is required to derive algorithms specific to each model on an individual basis (e.g. see Blei and Lafferty (2007); Braun and McAuliffe (2010); Jaakkola and Jordan (1996, 2000)).

Approximate posterior density. Traditionally, variational inference has relied on the mean-field family of approximations, which makes strong assumptions that the latent variables are independent and governed by their own parameters. These assumptions makes them convenient to work with – they yield analytically tractable densities, and are easy to optimize. However, it also means they inevitably fail to capture the posterior dependencies between latent variables. Not only are these dependencies interesting in and of themselves, they can drastically improve the expressiveness of the approximation. A major line of research in recent years has focused on developing richer families of approximations while maintaining their tractability.

Likelihood. Many models (especially in the natural sciences) do not admit a likelihood, or calculating it is intractable (finding the region to integrate over is difficult, and the integration may be prohibitively expensive). Instead, they are specified by generative process in terms of a deterministic equation, given parameters and random noise.

Prior. The idea that we may not be able to calculate the priors' density is almost unheard of in Bayesian statistics. However, an implicit prior, specified only with samples, has the potential to provide rich representation of highly-informative prior knowledge and beliefs about the latent variables. Combined with the likelihood, it can have important effects on the resulting posterior.

3 Modern variational inference

First we introduce further notation to aid our discussion. The variational family $\mathcal{Q}$ is indexed with free *variational parameters* ϕ. In other words, the approximate densities $q_\phi(\mathbf{z}\,|\,\mathbf{x}) \in \mathcal{Q}$ are parameterized by ϕ, and the optimization problem of eq. (4) is equivalent to finding the setting of parameters that minimizes the KL divergence.

3.1 Stochastic gradient-based optimization

Despite the efforts to develop generic methods for inference in non-conjugate models (Knowles and Minka, 2011; Wang and Blei, 2013), much model-specific derivation is still needed. Nowadays, the dominant approach for dealing with nonconjugate models is stochastic gradient-based optimization of the ELBO using mini-batch data subsampling, and Monte Carlo (MC) estimates of the gradient. The former scales up variational inference to massive datasets (Hoffman et al., 2013), while the latter sidesteps the cumbersome model-specific derivations to allow for generic "black-box" inference methods (Ranganath et al., 2014). The key idea is to write the gradient of the ELBO as an expectation of the gradient, approximate it with MC estimates, then perform stochastic gradient descent with repeated MC gradient estimates and mini-batches subsampled from the data.

This has been aptly referred to as "doubly stochastic" variational inference (Titsias and Lázaro-Gredilla, 2014), since we have sources of stochasticity in the gradients originating not only from the mini-batch subsamples, but also from the Monte Carlo estimation of the gradients.

Many general approaches are either based on variance reduction techniques such as Rao-Blackwellization, or Monte Carlo control variate estimators, which work for both continuous or discrete latent variables (Mnih and Gregor, 2014; Paisley et al., 2012; Wingate and Weber, 2013). We focus on the general class models with continuous latent variables that can be reparameterized by a deterministic transformation of random variables from a parameter-free base distribution (Kingma and Welling, 2013; Rezende et al., 2014).

3.1.1 Reparameterization trick

The reparameterization trick is a straightforward change of variables that expresses the random variable $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ as a deterministic transformation g_ϕ of another random variable $\boldsymbol{\epsilon}$, given $\mathbf{x}$, and parameterized by ϕ,

$$\mathbf{z} = g_\phi(\mathbf{x}, \boldsymbol{\epsilon}), \quad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}), \tag{10}$$

where the distribution $p(\boldsymbol{\epsilon})$ is independent of $\mathbf{x}$ or ϕ. The choice of transformation $g_\phi(\mathbf{x}, \boldsymbol{\epsilon})$ determines the variational family $\mathcal{Q}$. It is common to specify a location-scale family of approximate densities for which we can use a simple location-scale transformation. See appendix C for an example of specifying a family of diagonal Gaussian approximations.

Now we can write the expectation under $q_\phi(\mathbf{z}|\mathbf{x})$ as an expectation under $p(\boldsymbol{\epsilon})$, which allows the gradient and expectation to commute,

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[f(\mathbf{x}, \mathbf{z})] = \nabla_\phi \mathbb{E}_{p(\boldsymbol{\epsilon})}[f(\mathbf{x}, g_\phi(\mathbf{x}, \boldsymbol{\epsilon}))] \tag{11}$$

$$= \mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_\phi f(\mathbf{x}, g_\phi(\mathbf{x}, \boldsymbol{\epsilon}))]. \tag{12}$$

Specifying $f(\mathbf{x}, \mathbf{z}) = \log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}|\mathbf{x})$ gives us the ELBO in eq. (5). This simple reparameterization enables us to take unbiased stochastic estimates of its gradients by drawing noise samples $\boldsymbol{\epsilon}$ from $p(\boldsymbol{\epsilon})$. Importantly, for models with continuous latent variables, it can be shown to have the lowest variance among competing estimators.

3.2 Expressive approximate posteriors

It is possible to build transformations that result in highly flexible approximate posterior densities $q_\phi(\mathbf{z}|\mathbf{x})$. However, one is usually required to ensure the log posterior density is analytical and can be computed efficiently. If g_ϕ is differentiable, invertible and the density $p(\boldsymbol{\epsilon})$ is known, it is straightforward to compute the log posterior density,

$$\log q_\phi(\mathbf{z}|\mathbf{x}) = \log p(\boldsymbol{\epsilon}) - \log d_\phi(\mathbf{x}, \boldsymbol{\epsilon}). \tag{13}$$

The second term is the log absolute value of the Jacobian determinant,

$$\log d_\phi(\mathbf{x}, \boldsymbol{\epsilon}) = \log \left| \det \mathbf{J}_\phi(\mathbf{x}, \boldsymbol{\epsilon}) \right|. \tag{14}$$

Like g_ϕ, the Jacobian $\mathbf{J}_\phi(\mathbf{x}, \boldsymbol{\epsilon})$ is a function of $\boldsymbol{\epsilon}$ and $\mathbf{x}$, parameterized by ϕ,

$$\mathbf{J}_\phi(\mathbf{x}, \boldsymbol{\epsilon}) = \frac{\partial}{\partial \boldsymbol{\epsilon}} g_\phi(\mathbf{x}, \boldsymbol{\epsilon}). \tag{15}$$

Again, see appendix C for an example of these calculations for the diagonal Gaussian transformation.

3.2.1 Normalizing flows

Normalizing flows are a way to construct flexible and expressive approximate posteriors Rezende and Mohamed (2015). A normalizing flow is a sequence of smooth invertible transformations $g_t, t = 1, \ldots, T$ that map a simple initial density into successively more complex densities,

$$\mathbf{z}_0 = \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}), \tag{16}$$

$$\mathbf{z}_t = g_t(\mathbf{x}, \mathbf{z}_{t-1}). \tag{17}$$

The Jacobian of this chain of transformations factorizes as

$$\frac{\partial \mathbf{z}_T}{\partial \boldsymbol{\epsilon}} = \prod_{t=1}^{T} \frac{\partial \mathbf{z}_t}{\partial \mathbf{z}_{t-1}}. \tag{18}$$

Hence, the log Jacobian determinant factorizes as well,

$$\log \left| \det \left(\frac{\partial \mathbf{z}_T}{\partial \boldsymbol{\epsilon}} \right) \right| = \sum_{t=1}^{T} \log \left| \det \left(\frac{\partial \mathbf{z}_t}{\partial \mathbf{z}_{t-1}} \right) \right|. \tag{19}$$

Rezende and Mohamed (2015) propose a specific family of simple transformations, called the planar flows,

$$g_t(\mathbf{x}, \mathbf{z}_{t-1}) = \mathbf{z}_{t-1} + \mathbf{u} h(\mathbf{w}^T \mathbf{z}_{t-1} + b), \tag{20}$$

where the parameters of the flow are vectors $\mathbf{u}$, $\mathbf{w}$, scalar b and nonlinear activation function $h(\cdot)$. In practice, they are limited by the fact that a very long sequence of flows would be required in practice to capture dependencies in high-dimensional latent spaces. Kingma et al. (2016) address this limitation by proposing the inverse autoregressive flows, which are Gaussian autoregressive functions that are architecturally similar to LSTM cells.

While this normalizing flow approaches are an important step towards expressive posteriors, they still impose restrictions on the invertibility of transformations and the tractability of calculating Jacobian determinants.

3.2.2 Hierarchical variational models

Hierarchical variational models augment variational families by placing a prior $q(\boldsymbol{\phi} \,|\, \mathbf{x})$ on their variational parameters (Ranganath et al., 2016), and incorporating them as auxiliary latent variables (Maaløe et al., 2016). The joint density between the latent variables factorizes as

$$q(\mathbf{z}, \boldsymbol{\phi} \,|\, \mathbf{x}) = q(\mathbf{z} \,|\, \mathbf{x}, \boldsymbol{\phi}) q(\boldsymbol{\phi} \,|\, \mathbf{x}). \tag{21}$$

The goal is to marginalize out the variational parameters to obtain an expressive approximate posterior $q(\mathbf{z} \mid \mathbf{x})$ that can capture dependencies between latent variables,

$$q(\mathbf{z} \mid \mathbf{x}) = \int q(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\phi}) q(\boldsymbol{\phi} \mid \mathbf{x}) d\boldsymbol{\phi}. \tag{22}$$

The basic idea is to treat the variational family as a model of the latent variables and to expand this model hierarchically – hierarchical variational models induce dependencies between latent variables the same way that Bayesian models induce dependencies between observed variables. However, since this integral is intractable, we cannot directly maximize the ELBO. Alternatively, consider the exact posterior $q(\boldsymbol{\phi} \mid \mathbf{z}, \mathbf{x})$ over variational parameters,

$$q(\boldsymbol{\phi} \mid \mathbf{z}, \mathbf{x}) = \frac{q(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\phi}) q(\boldsymbol{\phi} \mid \mathbf{x})}{q(\mathbf{z} \mid \mathbf{x})}. \tag{23}$$

We can approximate it by introducing variational density $r_\lambda(\boldsymbol{\phi} \mid \mathbf{z}, \mathbf{x})$ with "variational hyperparameters" λ. Now we can use it to optimize a lower bound on the ELBO,

$$\begin{aligned} \text{ELBO}(q) &= \mathbb{E}_{q(\mathbf{z} \mid \mathbf{x})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \mathbf{x})] && (24) \\ &\geq \mathbb{E}_{q(\mathbf{z} \mid \mathbf{x})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\phi}) - \log q(\boldsymbol{\phi} \mid \mathbf{x}) + \log r_\lambda(\boldsymbol{\phi} \mid \mathbf{z}, \mathbf{x})] && (25) \end{aligned}$$

Ranganath et al. (2016) experiment with normalizing flows as priors on the variational parameters. Tran et al. (2016) consider nonparametric Gaussian process (GP) priors, which they prove to offer certain universal approximation guarantees. While it places no restrictions on invertibility or having tractable Jacobians, there still remains the cost of inverting the kernel matrix of the GP.

4 Implicit models in variational inference

Concurrent with the advances in variational inference, there has been tremendous research interest in *implicit* probabilistic models. This can be credited to the introduction of generative adversarial network (GAN) (Goodfellow et al., 2014), and the unprecedented success they have had with generating sharp images, synthesizing realistic natural language text and solving other similarly challenging problems.

Implicit models are specified only in terms of a simulator or some form of generative process that can produce observations of some phenomenon. In contrast with *prescribed* probabilistic models (Diggle and Gratton, 1984), which are prevalent and long-established in machine learning, implicit models do not specify explicit probability densities that can be directly evaluated on observed and latent variables within the model.

While implicit models can be more unwieldy and challenging to train, they offer greater flexibility and have the potential to describe physical phenomena with far higher fidelity than models which admit tractable densities. Indeed, implicit models are ubiquitous in physics, engineering and the natural sciences at large. They are used extensively in a diverse range of areas, including ecology, high-energy physics, climate science, geology, et cetera, where simulators are used to model real-world observations to forecast results.

Recently, implicit models have begun to emerge in various areas of machine learning research. In this section, we review the new approaches that extend variational inference to work with arbitrarily expressive implicit models.

4.1 Density ratio estimation

Recall the expanded form of the ELBO in eq. (7), which is written in terms of the log densities of the prior, likelihood and approximate posterior. In order to optimize the ELBO, we usually require these terms to be tractable. A way to relax this constraint is by directly estimating the difference of the log densities, or equivalently, the log ratio of the densities.

Learning in implicit models can be reduced to the fundamental problem of estimating density ratios by comparison of samples drawn from their probability distributions. This problem has been well-studied (Sugiyama et al., 2012), and the related approaches of *probabilistic classification*, *divergence minimization*, *ratio matching* and *moment matching* have each been developed extensively in different parts of the literature.

Mohamed and Lakshminarayanan (2017) make the connections between these approaches explicit, and relate them back to the recently proposed variants of implicit generative models – the original GANs (Goodfellow et al., 2014), f-GANs (Nowozin et al., 2016), b-GANs (Uehara et al., 2016), generative moment matching networks (Dziugaite et al., 2015; Li et al., 2015) and Wassertein GANs (Arjovsky et al., 2017), each of which are fundamentally based on one of these approaches.

Perhaps the most conceptually straightforward approach is probabilistic classification, which estimates a density ratio by building a probabilistic classifier $t(\mathbf{x})$ to distinguish between samples $\mathbf{x}$ from distributions p and q[1]. We write $t(\mathbf{x}) = \sigma(r(\mathbf{x}))$ to indicate that it is the output of some parametric function $r(\mathbf{x})$ fed through a logistic sigmoid activation $\sigma(\cdot)$. It is trained to minimize the binary cross-entropy loss (or equivalently, maximize its negative),

$$\mathcal{F}_{\text{GAN}}(r, q) = \mathbb{E}_{p(\mathbf{x})}[\log \sigma(r(\mathbf{x}))] + \mathbb{E}_{q(\mathbf{x})}[\log(1 - \sigma(r(\mathbf{x})))]. \tag{26}$$

The goal of the GAN is to learn a generative model for the implicit distribution q by finding the saddle points of this objective,

$$\min_{q \in \mathcal{Q}} \max_{r \in \mathcal{R}} \mathcal{F}_{\text{GAN}}(r, q). \tag{27}$$

It can be shown that $\mathcal{F}_{\text{GAN}}(r, q)$ attains its maximum when $r(\mathbf{x})$ is exactly the log density ratio between p and q,

$$r^*(\mathbf{x}) = \log p(\mathbf{x}) - \log q(\mathbf{x}). \tag{28}$$

Hence, we can use $r(\mathbf{x})$ as a proxy for the true log density ratio. Plugging it back in to $\mathcal{F}_{\text{GAN}}(q, r)$, we can show that

$$2\mathcal{D}_{\text{JS}}(p \parallel q) - \log(4) = \mathcal{F}_{\text{GAN}}(r^*, q) \tag{29}$$

$$\geq \max_{r \in \mathcal{R}} \mathcal{F}_{\text{GAN}}(r, q) \tag{30}$$

where $\mathcal{D}_{\text{JS}}$ is the Jensen-Shannon (JS) divergence, and we have equality in eq. (30) if family $\mathcal{R}$ is rich enough to contain the exact density ratio estimator, $r^* \in \mathcal{R}$. Therefore, it can be said that $\mathcal{F}_{\text{GAN}}(r, q)$ is a variational lower bound of divergence $\mathcal{D}_{\text{JS}}(p \parallel q)$, and importantly, one that doesn't require any explicit densities. Furthermore, finding its saddle points in the optimization problem of eq. (27) is equivalent to the approximate minimization of divergence $\mathcal{D}_{\text{JS}}(p \parallel q)$.

[1] We are overloading notation here; p and q are not (necessarily) related to the probability distributions introduced earlier, on which interested performing posterior inference.

Next we show how these ideas have been applied to performing variational inference in the presence of intractable densities.

4.2 Implicit approximate posteriors

To relax the constraint on the tractability of density $q(\mathbf{z}|\mathbf{x})$, Huszár (2017); Mescheder et al. (2017) independently proposed an approximation to the ELBO which directly estimates the log ratio of densities $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$,

$$\begin{aligned}
\widetilde{\text{ELBO}}(r,q) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}) - r(\mathbf{x},\mathbf{z})] && (31)\\
&\simeq \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}) - r^*(\mathbf{x},\mathbf{z})] && (32)\\
&= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}) - \log q(\mathbf{z}|\mathbf{x})] && (33)\\
&= \text{ELBO}(q). && (34)
\end{aligned}$$

Similar to before, we use parametric function $r(\mathbf{x},\mathbf{z})$ as an approximation to the exact ratio estimator $r^*(\mathbf{x},\mathbf{z}) = \log q(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z})$ by simultaneously maximizing the objective,

$$\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log \sigma(r(\mathbf{x},\mathbf{z}))] + \mathbb{E}_{p(\mathbf{z})}[\log(1-\sigma(r(\mathbf{x},\mathbf{z})))]. \tag{35}$$

Computing $\widetilde{\text{ELBO}}(r,q)$ has no explicit dependency on the density of approximate posterior $q(\mathbf{z}|\mathbf{x})$ (and incidentally, prior $p(\mathbf{z})$, which we discuss further in section 4.4). We only require samples from it to compute MC gradient estimates of the $\widetilde{\text{ELBO}}$.

Hence, the transformation $g_\phi(\mathbf{x},\boldsymbol{\epsilon})$ underlying $q(\mathbf{z}|\mathbf{x})$ need no longer be constrained to any known distribution. Instead, we can pick any expressive model, such as a deep neural network of arbitrarily complex architectural design, without regard for the tractability or scalability of calculating its Jacobian determinants. In a nutshell, we are relieved of the analytical burden of having to carefully craft flexible transformations in such a way that they still admit tractable densities, and incurring the cost of maximizing a cruder approximation $\widetilde{\text{ELBO}}$ to the model evidence.

Recall that the ELBO reflects the balance between prior and likelihood through the KL and ELL terms respectively. Instead of evaluating the KL through a closed-form expression, we now estimate it using the expected output of a ratio estimator. Now, we recover the optimal estimator $r^*(\mathbf{x},\mathbf{z})$ of ratios between densities $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$ from the binary classifier $\sigma(r^*(\mathbf{x},\mathbf{z}))$, which optimally discriminates between samples drawn from these distributions. Hence, we can view the bi-level optimization of the $\widetilde{\text{ELBO}}$ of eq. (31) and the objective of eq. (35) as an explicit formulation of the tension between prior and likelihood as an adversarial minimax game – the approximate posterior $q(\mathbf{z}|\mathbf{x})$ tries to produce samples $\mathbf{z}$ that "fool" the discriminator to believe it was drawn from the prior $p(\mathbf{z})$, while also trying to maximize the model-fitting ELL term. Concurrently, the discriminator is learning to perform more optimally.

Mescheder et al. (2017) showed that at the Nash equilibrium of this game, we recover the true posterior, $q(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}|\mathbf{x})$. This guarantee is analogous to that provided by MCMC, which is designed to yield an ergodic Markov chain whose stationary distribution is exactly $p(\mathbf{z}|\mathbf{x})$. In contrast, most variational inference methods are never concerned with providing such guarantees.

An approach closely related to this was first proposed by Makhzani et al. (2015), which approximately minimizes the KL between $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$ in the adversarial training method we outlined above. However, they do not explicitly maximize the ELBO, so it cannot be characterized as performing variational inference. Furthermore, their discriminator (by design) fails to incorporate the dependency on observed variables $\mathbf{x}$, so cannot guarantee $q(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}|\mathbf{x})$ even upon reaching an equilibrium.

4.3 Likelihood-free inference

Tran et al. (2017) extend on the method described in section 4.2 in two fundamental ways. First, instead of estimating the log ratio of densities $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$, they directly estimate that of the joint densities $p(\mathbf{x}, \mathbf{z})$ and $q(\mathbf{x}, \mathbf{z}) = q(\mathbf{z}|\mathbf{x})q(\mathbf{x})$, where $q(\mathbf{x})$ denotes the empirical data distribution. This takes care of all the terms in the ELBO, and relaxes the requirements on them to admit tractable densities. We now require only samples from these distribution. This is referred to by Huszár (2017) as a "joint-contrastive" approach, while the previous approach was "prior-contrastive". Some related methods where developed previously in the context of implicit generative models, and learns the inverse mapping from data to noise of a GAN (Donahue et al., 2016; Dumoulin et al., 2016). The second important generalization is the addition of global latent variables, which is important for pooling information across data point.

Having removed the requirement of a tractable likelihood, variational inference can now compete directly with the mainstay methods for likelihood-free inference, such as approximate Bayesian computation (ABC) (Beaumont, 2010; Marin et al., 2012), but at much larger scale. While their experiments demonstrate the applicability of their method on an impressive range of applications, there is still much to understand about the relative advantages and shortcomings of likelihood-free variational inference compared against "battle-tested" methods like ABC.

4.4 Implicit priors

In the section 4.2, we relaxed the constraint on the approximate posteriors to admit tractable densities by introducing a method for approximating the KL term in the ELBO. This had the side effect of also relaxing the requirement on the priors to admit tractable densities.

The prior $p(\mathbf{z})$ plays a key role in Bayesian inference. It embodies our prior knowledge and beliefs about the latent variables $\mathbf{z}$, and is combined with the likelihood to form the posterior density. As such, the prior can have an important effects on the posterior. The choice of non-informative or highly-informative priors has been subject of much philosophical debate (Gelman, 2009; Jaynes, 1968).

To our knowledge, the concept of an "implicit prior" has not been explored in the context of Bayesian modelling. An implicit prior is specified only with samples, and provides a potentially rich representation of highly-informative prior information of the latent variables. In the physical sciences for example, our knowledge is inextricably woven into the specification of our physics models and simulators. They provide a simplistic view of some physical phenomena which may not be sufficient to accurately model the real world, but which may be easier for us to reason about. Simulators could assume the role of highly-informative implicit priors, which we would reconcile with observed data from the physical world to form a posterior distribution. This is a potentially powerful method to incorporate prior scientific knowledge into the inference procedure.

We previously mentioned the work of Makhzani et al. (2015), which is closely related to variational inference with implicit priors and posteriors. To demonstrate the flexibility of their semi-supervised learning method, they contrived various implicit prior distributions specified only by samples. In particular, they experimented with the samples of latent variables that lay along a "swiss roll" shaped implicit distribution, and demonstrate the effects this had on the posterior.

Having generalized variational inference to work with implicit models in various components of the model, it is likely that many powerful methods (that don't initially appear to be Bayesian) can be subsumed as a special case of approximate Bayesian inference. A recent example of this is the adversarial domain adaption method of DiscoGAN (Kim et al., 2017), and the independently developed CycleGAN (Zhu et al., 2017). Given objects from domain A, say images of horses,

and objects from domain B, say images of zebras, the goal is to learn a mapping from domain A to B. We can relate their method to variational inference with implicit models as follows. The implicit prior $p(\mathbf{z})$ is represented by objects from domain B, and $p(\mathbf{x})$ is represented by objects from domain A. We learn the approximate posterior $q(\mathbf{z}|\mathbf{x})$ by learning the underlying mapping $\mathbf{z} = g_\phi(\mathbf{x}, \boldsymbol{\epsilon})$. We also learn the mapping corresponding to model $p(\mathbf{x}|\mathbf{z})$ to maintain cycle-consistency (this ensures we retain sufficient information from the object in the original domain to reconstruct it). Then, the adversarial training procedure described in section 4.2 approximately minimizes the divergence between $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$.

5 Conclusion

In this review, we presented the general problem of performing inference in Bayesian models. We discussed the intractability of computing the model evidence, which the exact posterior is dependent on. To address this, we then looked at variational inference, an approximate inference method based on optimization, and discussed the further sources of intractabilities that may arise when using this method. The remaining sections were devoted to a review of the literature that seeks to address each intractability. Specifically, we looked at approaches that enabled generic "black-box" methods, those that improved the expressiveness of the approximate posterior and finally, how variational inference could be generalized to implicit models by using techniques from density ratio estimation. Importantly, this line of research has enabled implicit approximate posteriors, implicit priors and likelihood-free variational inference. The many deep implications of this which are not well-understood, and this presents numerous exciting avenues for extending this research.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein Generative Adversarial Networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia. PMLR.

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annu.Rev.Ecol.Evol.Syst.*, 41(1):379–406.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M. and Lafferty, J. D. (2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics*, 1:17–35.

Braun, M. and McAuliffe, J. (2010). Variational Inference for Large-Scale Models of Discrete Choice.

Diggle, P. J. and Gratton, R. J. (1984). Monte Carlo Methods of Inference for Implicit Statistical Models.

Donahue, J., Krähenbühl, P., and Darrell, T. (2016). Adversarial Feature Learning.

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., and Courville, A. (2016). Adversarially Learned Inference.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 258–267.

Gelman, A. (2009). Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics.

Ghahramani, Z. and Beal, M. J. (2001). Propagation Algorithms for Variational Bayesian Learning. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 507–513. MIT Press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic Variational Inference. *Journal of Machine Learning Research*, 14:1303–1347.

Huszár, F. (2017). Variational Inference using Implicit Distributions.

Jaakkola, T. S. and Jordan, M. I. (1996). A variational approach to Bayesian logistic regression models and their extensions. *Aistats*, (AUGUST 2001).

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Jaynes, E. (1968). Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233.

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. (2017). Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings*

of Machine Learning Research, pages 1857–1865, International Convention Centre, Sydney, Australia. PMLR.

Kingma, D. P. (2017). *Variational Inference and Deep Learning: A New Synthesis*. PhD thesis, Universiteit van Amsterdam.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved Variational Inference with Inverse Autoregressive Flow. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4743–4751. Curran Associates, Inc.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes.

Knowles, D. A. and Minka, T. (2011). Non-conjugate Variational Message Passing for Multinomial and Binary Regression. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 1701–1709. Curran Associates, Inc.

Li, Y., Swersky, K., and Zemel, R. (2015). Generative Moment Matching Networks. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1718–1727, Lille, France. PMLR.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. (2016). Auxiliary Deep Generative Models. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1445–1453, New York, New York, USA. PMLR.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial Autoencoders.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Mescheder, L., Nowozin, S., and Geiger, A. (2017). Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2391–2400, International Convention Centre, Sydney, Australia. PMLR.

Mnih, A. and Gregor, K. (2014). Neural Variational Inference and Learning in Belief Networks. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1791–1799, Bejing, China. PMLR.

Mohamed, S. and Lakshminarayanan, B. (2017). Learning in Implicit Generative Models. In *The 5th International Conference on Learning Representations*.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 271–279. Curran Associates, Inc.

Paisley, J. W., Blei, D. M., and Jordan, M. I. (2012). Variational Bayesian Inference with Stochastic Search. In *Proceedings of the 29th International Conference on Machine Learning, {ICML} 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black Box Variational Inference. In Kaski, S. and Corander, J., editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland. PMLR.

Ranganath, R., Tran, D., and Blei, D. M. (2016). Hierarchical Variational Models. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 324–333, New York, New York, USA. PMLR.

Rezende, D. and Mohamed, S. (2015). Variational Inference with Normalizing Flows. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France. PMLR.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T., editors, *Proceedings of The 31st . . .* , volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China. PMLR.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press.

Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly Stochastic Variational Bayes for non-Conjugate Inference. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1971–1979, Bejing, China. PMLR.

Tran, D., Ranganath, R., and Blei, D. M. (2016). The Variational Gaussian Process. In *The 4th International Conference on Learning Representations*.

Tran, D., Ranganath, R., and Blei, D. M. (2017). Deep and Hierarchical Implicit Models.

Uehara, M., Sato, I., Suzuki, M., Nakayama, K., and Matsuo, Y. (2016). Generative Adversarial Nets from a Density Ratio Estimation Perspective.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(1):1005–1031.

Wingate, D. and Weber, T. (2013). Automated Variational Inference in Probabilistic Programming.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.

A Kullback-Leibler divergence

The KL divergence between the approximate and exact posterior density is

$$\begin{aligned}
\mathrm{KL}\left[q(\mathbf{z}\,|\,\mathbf{x}) \,\|\, p(\mathbf{z}\,|\,\mathbf{x})\right] &= \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}[\log q(\mathbf{z}\,|\,\mathbf{x}) - \log p(\mathbf{z}\,|\,\mathbf{x})] && (36)\\
&= \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}[\log q(\mathbf{z}\,|\,\mathbf{x}) - \log p(\mathbf{x},\mathbf{z}) + \log p(\mathbf{x})] && (37)\\
&= \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}[\log q(\mathbf{z}\,|\,\mathbf{x}) - \log p(\mathbf{x},\mathbf{z})] + \log p(\mathbf{x}) && (38)
\end{aligned}$$

This reveals how the KL divergence depends on the intractable evidence $p(\mathbf{x})$.

B Evidence lower bound

We expand the logarithm of the evidence as follows,

$$\begin{aligned}
\log p(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}[\log p(\mathbf{x})] && (39)\\
&= \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}\left[\log \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{z}\,|\,\mathbf{x})}\right] && (40)\\
&= \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}\left[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z}\,|\,\mathbf{x})}\frac{q(\mathbf{z}\,|\,\mathbf{x})}{p(\mathbf{z}\,|\,\mathbf{x})}\right] && (41)\\
&= \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}\left[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z}\,|\,\mathbf{x})}\right] + \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{x})}\left[\log \frac{q(\mathbf{z}\,|\,\mathbf{x})}{p(\mathbf{z}\,|\,\mathbf{x})}\right] && (42)\\
&= \mathrm{ELBO}(q) + \mathrm{KL}\left[q(\mathbf{z}\,|\,\mathbf{x}) \,\|\, p(\mathbf{z}\,|\,\mathbf{x})\right]. && (43)
\end{aligned}$$

This reveals the relationship between the ELBO, the KL divergence between $q(\mathbf{z}\,|\,\mathbf{x})$ and $p(\mathbf{z}\,|\,\mathbf{x})$, and the log evidence.

C Diagonal Gaussian approximation

To posit a family of diagonal Gaussian approximations

$$q_\phi(\mathbf{z}\,|\,\mathbf{x}) = \mathcal{N}(\mathbf{z}\,|\,\boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\mathbf{x})), \tag{44}$$

we use the transformation

$$g_\phi(\mathbf{x},\boldsymbol{\epsilon}) = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{45}$$

where $\odot$ denotes the element-wise product. The mean and standard deviation are functions of $\mathbf{x}$ specified by neural networks with parameters ϕ.

The Jacobian of this transformation is given by

$$\mathbf{J}_\phi(\mathbf{x},\boldsymbol{\epsilon}) = \mathrm{diag}(\boldsymbol{\sigma}_\phi(\mathbf{x})). \tag{46}$$

The determinant of a triangular matrix is the product of its diagonal entries, so we have

$$\log d_\phi(\mathbf{x}, \boldsymbol{\epsilon}) = \log \left| \det \mathbf{J}_\phi(\mathbf{x}, \boldsymbol{\epsilon}) \right| = \sum_i \log \sigma_{\phi i}(\mathbf{x}). \tag{47}$$

Since we picked $p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ the log posterior density $\log q_\phi(\mathbf{z} \mid \mathbf{x})$ is now trivial to compute.