Optimisation in Probabilistic Models

Nick Robinson
Informatics, Edinburgh University
s1784599@ed.ac.uk

Abstract

Important systems, from the immune system to the process of writing literature, can seem hopelessly complex. As scientists we hope they are probabilistically simple. At least we hope they can be usefully modelled by how certain variables in the system relate to each other. Of course, probabilistic models can easily become complex themselves, with quantities within the model requiring approximation. Variational inference (VI) makes computing these approximations an optimisation task, and the use of stochastic gradient-based optimisation has improved our ability to fit complex probabilistic models. Stochastic optimisation is crucial for scaling VI to large datasets. Developing gradient estimators for arbitrary functions of random variables allows us to fit richer approximate distribution. We critically examine recent progress in VI and suggest there is a need for further theoretical work improving these techniques, and applied work demonstrating the applicability and increasing the accessibility of optimisation methods for probabilistic models.

1 Introduction

Probabilistic latent variable models are a useful way to understand, summarise or make predictions about data. Having described a model $p_\theta(x, z)$ of the data x and latent variables z with parameters θ, we need to infer the probability of unobserved quantities in our model $p_\theta(z|x) = p_\theta(x, z)/p_\theta(x)$. However, inference in complex models can be an impossible task, because $p_\theta(x)$ often has no analytic solution and numerical computation time scales exponentially, making exact inference intractable. The best we can hope for is an approximation to the posterior.

The difficulty of inference and the need to compute an approximation introduces unfortunate challenges. Ideally as scientists we would write down the model that we think best describes the data generating process, without worrying about computational limitations on inference, and then fit, evaluate and criticise the model (Blei, 2014). This would require inference to be easy to run on a given model, computationally fast even with large datasets and, of course, to be a good approximation.

1.1 Approximate inference

The standard approach to approximate inference is to frame it as a sampling problem: estimate expectations with respect to a distribution, such as $p_\theta(z|x)$, by drawing Monte Carlo samples $\hat{z} \sim p_\theta(z|x)$, most commonly by using a Markov Chain with stationary distribution $p_\theta(z|x)$. This is known as Markov Chain Monte Carlo (MCMC) (Gelman et al., 2014; Angelino et al., 2016).

An alternative method of approximate inference, and the focus of this review, formulates inference as an optimisation problem: approximate the true intractable posterior distribution $p_\theta(z|x)$ with the tractable distribution $q^*_\phi(z)$ that minimises some divergence function $\mathrm{D}(q_\phi(z)||p_\theta(z|x))$. This method is known as variational inference (VI) (Jordan et al., 1999; Wainwright et al., 2008).

VI is a promising approach to approximate inference due to recent work combining it with stochastic gradient-based optimization techniques. More specifically, work in recent years demonstrates (i) the ability to use stochastic optimization, sub-sampling large datasets, to make VI scale well (Hoffman et al., 2013), (ii) evidence that VI can be more efficient than MCMC and achieve better predictive likelihoods $q^*_\phi(x_{test}|x, z)$ (Ranganath et al., 2014), and perhaps most crucially (iii) how various gradient estimators can be used with little model-specific work and applied to a richer class of approximating distributions (Kingma & Welling, 2013; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014). The promise is faster prototyping and evaluation of probabilistic models (Duvenaud & Adams, 2015), and better models for complex, high-dimension data (Rezende et al., 2014).

1.2 Contribution

This review focuses on developments in gradient-estimation techniques in the context of variational inference, particularly use of the score function estimator and the pathwise derivative estimator (or reparamaterisation trick), and corresponding variance reduction techniques. The paper is mostly theoretical, reviewing why and how these techniques can be used, although they are ultimately motivated by the empirical results observed in the literature.

We first give a concise introduction to variational inference in section 2.[1] Recent advances in optimisation algorithms for variational inference are reviewed in section 3, and section 4 suggests potentially fruitful directions for future research given the empirical results so far. We conclude in section 5 with a discussion of related research topics in VI.

2 Variational Inference

VI approximates an intractable posterior distribution $p_\theta(z|x)$, of the unobserved variables z given the data x, with whichever distribution $q_\phi(z)$, with parameters ϕ, in some family of densities $\mathcal{Q}$ minimises the divergence $\mathrm{D}(q\|p)$, for a specified divergence function D.

We now have an optimization problem:

$$q^*_\phi = \underset{q_\phi \in \mathcal{Q}}{\arg\min}\, \mathrm{D}(q_\phi(z)\|p_\theta(z|x)) \tag{1}$$

The *de facto* divergence measure is the Kullback-Leibler (KL) Divergence (also called relative entropy):[2]

$$\mathrm{KL}(q\|p) = \mathbb{E}_q[\log q - \log p] \tag{2}$$

The KL Divergence is a measure of the inefficiency of using a code optimised for q to encode data from distribution p. It is non-negative, and zero when $q = p$. The KL Divergence cannot be minimised directly, because it requires evaluating $p_\theta(x)$. Making this explicit:

$$\begin{aligned}\mathrm{KL}(q_\phi(z)\|p_\theta(z|x)) &= \mathbb{E}_{q_\phi}[\log q_\phi(z)] - \mathbb{E}_{q_\phi}[\log p_\theta(z|x)] \\ &= \mathbb{E}_{q_\phi}[\log q_\phi(z)] - \mathbb{E}_{q_\phi}[\log p_\theta(z, x) - \log p_\theta(x)] \\ &= \mathbb{E}_{q_\phi}[\log q_\phi(z)] - \mathbb{E}_{q_\phi}[\log p_\theta(z, x)] + \log p_\theta(x)\end{aligned}$$

Since $p_\theta(x)$ is constant with reference to q_ϕ, instead of minimising the KL it is equivalent to maximising the lower bound $\mathcal{L}$ on $\log p_\theta(x)$ (known as the Evidence Lower Bound, or ELBO):

$$\begin{aligned}\mathcal{L}(q_\phi(z)) &= -\mathrm{KL}(q_\phi(z)\|p_\theta(z|x)) + \log p_\theta(x) \\ &= \mathbb{E}_{q_\phi}[\log p_\theta(z, x)] - \mathbb{E}_{q_\phi}[\log q_\phi(z)] \end{aligned} \tag{3}$$

This is sometimes interpreted as a term encouraging probability mass on latent variables z that explain the data x and a second term encouraging probability mass to spread across many configurations (Ranganath et al., 2014). The optimization objective is now $q^*_\phi = \arg\min_{q_\phi \in \mathcal{Q}} \mathcal{L}(q_\phi(z))$.

This objective is non-convex, so the q_ϕ minimising the KL Divergence will in general not be found. Of course, whether or not the fitted approximate distribution q_ϕ is good enough in practice will depend on the model and application. We first focus on the question of how to optimize $\mathcal{L}$.

[1] For a comprehensive introduction to VI see e.g. Jordan et al. (1999) or Blei et al. (2017)

[2] Alternative divergence measures for VI have recently been explored, for example the α Divergence (also called Renyi entropy) (Hernández-Lobato et al., 2016; Li & Turner, 2016) and the χ^2 Divergence (Dieng et al., 2017). We focus on the standard case of KL Divergence throughout as it is the most widely used in practice.

2.1 Stochastic optimisation

The aim is to fit the parameters ϕ of $q_\phi(z)$ to maximise the expectation (3).

If a closed-form expression for $\mathcal{L}(q_\phi(z))$ is possible, a simple coordinate ascent algorithm will find a (local) maximum. This is possible, for example, by assuming a model $p_\theta(x,z)$ in the exponential family, a conjugate prior $p(\theta)$, and the set of fully-factorised approximations: $q_\phi(z) = \prod_i q_{\phi_i}(z_i)$. In this case we can substitute expressions for p and q into the definition of the objective $\mathcal{L}$ and recover a simple expression, and subsequently compute the gradient $\nabla_\phi \mathcal{L}$ (for details see e.g. Hoffman et al. (2013); Wang & Blei (2013)).

Gradient ascent algorithms are iterative, and for this optimisation method to scale we cannot require evaluating the gradient on the whole dataset x as this is too slow of large datasets. Instead, we can compute a noisy, unbiased estimate based on a small sub-sample of the data, either mini-batches or individual data points, and this allows optimisers to converge much faster while still utilising large datasets.

The introduction of stochastic optimisation makes VI suitable for large datasets, and has been crucial to progress in machine learning more generally.[3] However, assuming the latent variables z are independent of one another in order to ensure inference is tractable limits the quality of approximation possible. Recent research has extended stochastic optimization to models where the expectation (3) cannot be solved in closed-form, and suggested gradient estimators that do not require model-specific derivations. We now turn to the general, non-conjugate case where we will simulate from the approximating distribution q_ϕ rather than computing closed-form updates. (Titsias & Lázaro-Gredilla, 2014)

3 Gradient estimators for VI

To use gradient-based optimisation requires estimating the gradient of the expectation over random variables. This section reviews recent work in this area, based on Kingma & Welling (2013); Ranganath et al. (2014); Titsias & Lázaro-Gredilla (2014); Rezende et al. (2014); Schulman et al. (2015) and Grathwohl et al. (2017).

3.1 Score function estimator

If we define

$$g_\phi(z) = \log p_\theta(z,x) - \log q_\phi(z) \tag{4}$$

which is a function only of ϕ and z given the data x is fixed, then we get an alternative expression for (3):

$$\mathcal{L}(q_\phi(z)) = \mathbb{E}_{q_\phi}[g_\phi(z)] \tag{5}$$

We need to compute $\nabla_\phi \mathcal{L}$, which we can write as an expectation with reference to q_ϕ as follows

$$\begin{aligned}
\nabla_\phi \mathbb{E}_{q_\phi}[g_\phi(z)] &= \nabla_\phi \int q_\phi(z) g_\phi(z) dz \\
&= \int g_\phi(z) \nabla_\phi q_\phi(z) + q_\phi(z) \nabla_\phi g_\phi(z)\ dz \\
&= \int g_\phi(z) \nabla_\phi q_\phi(z)\ dz \ \ (\text{since } \mathbb{E}_{q_\phi}[\nabla_\phi g_\phi(z)] = 0) \\
&= \int g_\phi(z) q_\phi(z) \nabla_\phi \log q_\phi(z)\ dz \\
&= \mathbb{E}_{q_\phi}[g_\phi(z) \nabla_\phi \log q_\phi(z)]
\end{aligned} \tag{6}$$

[3] For an extensive review of stochastic gradient optimisation in machine learning see Bottou et al. (2017).

With the gradient in this form it possible to get unbiased Monte Carlo gradient estimates using

$$\mathbb{E}_{q_\phi}[g_\phi(z)\nabla_\phi \log q_\phi(z)] \approx \frac{1}{S}\sum_{s=1}^{S} g_\phi(z^{(s)})\nabla_\phi \log q_\phi(z^{(s)}) \tag{7}$$

$$\text{where } z^{(s)} \sim q_\phi(z|x^{(i)}) \text{ and } x^{(i)} \text{ is the } i^{th} \text{ data point or data mini-batch.}$$

This is known as the score function estimator, because $\nabla_\phi \log q_\phi(z)$ is known as the score function. Clearly, to use (7) we need to be able to evaluate the score function and sample from the approximating distribution $q_\phi(z)$.

However, the crucial problem is that estimates using (7) are too high variance to be useful in practice (Kingma & Welling, 2013; Ranganath et al., 2014), motivating the development of an alternative gradient estimation technique and techniques for controlling the variance if (7) is used. We discuss each in turn.

3.2 Pathwise gradient estimator

The score function estimator is high variance, and only uses information of how ϕ influences $g_\phi(z)$, and not the influence of z directly. A lower variance estimator can be used if the approzimate distribution $q_\phi(z)$ can be written as a deterministic, differentiable function of some other random variable (Kingma & Welling, 2013; Rezende et al., 2014), that is, if we can define a differentiable function r such that

$$q_\phi(z) = q_\phi(r_\phi(\epsilon)) \text{ with } \epsilon \sim f(\epsilon) \tag{8}$$

This allows us to use samples $z \sim q_\phi(z)$ by sampling $\epsilon \sim f(\epsilon)$ and treating z a deterministic function $z = r_\phi(\epsilon)$, and it is known as the reparametrisation trick. In the simplest case, where we know the inverse CDF of q_ϕ, we can use this as $r_\phi(\epsilon)$ and sample $\epsilon \sim \text{Uniform}(0, 1)$. Reparametrisation is also possible for any location-scale family of distributions, and combinations of random variables, so long as r is differentiable with respect to ϕ, that is, for any continuous latent-variable model.

If $f(\epsilon)$ does not depend on ϕ, then we can estimate the gradient with

$$\begin{aligned}\nabla_\phi \mathbb{E}_{q_\phi}[g_\phi(z)] &= \nabla_\phi \mathbb{E}_f[g_\phi(r_\phi(\epsilon)] \\ &= \mathbb{E}_f[\nabla_\phi g_\phi(r_\phi(\epsilon)]\end{aligned} \tag{9}$$

It is then possible to get unbiased Monte Carlo gradient estimates using

$$\mathbb{E}_f[\nabla_\phi g_\phi(r_\phi(\epsilon)] \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_\phi g_\phi(r_\phi(\epsilon)) \tag{10}$$

$$\text{where } \epsilon^{(s)} \sim f(\epsilon)$$

If $f_\phi(\epsilon)$ depends on ϕ, then we can instead estimate the gradient with

$$\nabla_\phi \mathbb{E}_{q_\phi}[g_\phi(z)] = \mathbb{E}_{f_\phi}[\nabla_\phi g_\phi(r_\phi(\epsilon)) + g_\phi(r_\phi(\epsilon))\nabla_\phi \log q_\phi(z)] \tag{11}$$

Note the similarity of the second term to the score function estimator given in (7).

This pathwise gradient estimator (9, 11) has been consistently observed to give lower-variance gradient estimates than the score function estimator in practice (Rezende et al., 2014; Ruiz et al., 2016). However, as described the pathwise gradient estimator cannot be used if z is is a discrete random variable, for example indicating which distribution in a mixture model data is drawn from, because there is no differentiable reparametisation and therefore cannot take the gradient given in (9). This means we must use the score function estimator in some cases, and motivates introducing methods of controlling the variance of the gradient estimates so that they are more useful in practice.

3.3 Reducing the variance of estimators

We focus on the most common method of reducing variance, which is the use of control variates. The fundamental idea is replace the gradient estimator $\hat{e}$, for example (7), with a function that has the same expectation but lower variance Ranganath et al. (2014). This is done by subtracting from the gradient estimator $\hat{e}$, a function c that is positively correlated with the estimator and has known expectation, and then adding back on the expectation of c. Finding an appropriate function c is an area of research that is important if VI is to be able to find good enough approximation distributions (Tucker et al., 2017; Roeder et al., 2017; Grathwohl et al., 2017), and our first suggestion for more research in this area.

4 Future Work

Having discussed the recent advances in VI, we now suggest three areas where more research seems both possible and likely to be a large benefit to the research community.

4.1 Adoption

Stochastic gradient optimisation has been crucial to the advancement of deep learning on large datasets, thanks in part to how simple it is for researchers to make use of the backprogation algorithm, usually requiring no model-specific work and scaling to huge datasets. Part of the appeal of probabilistic models over deep learning is the ability to explicitly account for uncertainty, but this variability has also been a key limitation, since fitting the parameters of stochastic functions has historically required pen and paper derivations for each model. This suggests there is valuable work to be done in easing the adoption of variational methods, for example, using automatic differentiation to make building and fitting models easier (Kucukelbir et al., 2017a), and building these techniques into common programming languages used for machine learning and Bayesian statistics, such as Python and Stan (Kucukelbir et al., 2015; Duvenaud & Adams, 2015)

4.2 Evaluating performance

Of course, the recent success of deep learning is not just the ability to *fit* a complex function, it is the the excellent predictive performance of they achieve across a a wide range of datasets. Evaluating performance in Bayesian models, especially in high-dimensions where posterior summary statistics are difficult to interpret, is still an open problem (Gelman et al., 2014; Blei, 2014; Kucukelbir et al., 2017b). A crucial extension is then to demonstrate the practical usefulness of the techniques.

4.3 Second-order methods

While the gradient estimators described above can be used with any gradient optimiser, e.g. Adam, ideally the optimiser would be efficient and not require tuning hyper-parameters, and so second-order methods for VI have recently been suggested (Fan et al., 2015; Regier et al., 2017). The development of a model-agnostic second-order methods with convergence guarantees could enable the community to fit better variational approximations. Optimisation is a rich area of research but has tended not to focus on the task of variational inference.

5 Related work

Stochastic optimization in variational inference is part of broader topic of methods of approximate inference, and more general overview of advances in approximate inference is given by Angelino et al. (2016). Concurrently with this paper, Zhang et al. (2017) gives a broader review on multiple areas of recent research in VI. On gradient estimation, Schulman et al. (2015) gives a general framework for determining unbiased gradient estimators when some functions are stochastic, and Ruiz et al. (2016) recently demonstrated the use of this framework in the context of variational inference.

Variational inference is a powerful technique, undergoing fast development, and with open research problems that directly affect our ability to fit rich probabilistic models.

References

Angelino, Elaine, Johnson, Matthew James, Adams, Ryan P, et al. Patterns of scalable bayesian inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247, 2016.

Blei, David M. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.

Blei, David M, Kucukelbir, Alp, and McAuliffe, Jon D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.

Bottou, Léon, Curtis, Frank E, and Nocedal, Jorge. Optimization methods for large-scale machine learning. *stat*, 1050:2, 2017.

Dieng, Adji Bousso, Tran, Dustin, Ranganath, Rajesh, Paisley, John, and Blei, David. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, pp. 2729–2738, 2017.

Duvenaud, David and Adams, Ryan P. Black-box stochastic variational inference in five lines of python. In *NIPS Workshop on Black-box Learning and Inference*, 2015.

Fan, Kai, Wang, Ziteng, Beck, Jeff, Kwok, James, and Heller, Katherine A. Fast second order stochastic backpropagation for variational inference. In *Advances in Neural Information Processing Systems*, pp. 1387–1395, 2015.

Gelman, Andrew, Carlin, John B, Stern, Hal S, Dunson, David B, Vehtari, Aki, and Rubin, Donald B. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

Grathwohl, Will, Choi, Dami, Wu, Yuhuai, Roeder, Geoff, and Duvenaud, David. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *arXiv preprint arXiv:1711.00123*, 2017.

Hernández-Lobato, José Miguel, Li, Yingzhen, Rowland, Mark, Hernández-Lobato, Daniel, Bui, Thang, and Turner, Richard Eric. Black-box α-divergence minimization. 2016.

Hoffman, Matthew D, Blei, David M, Wang, Chong, and Paisley, John. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kucukelbir, Alp, Ranganath, Rajesh, Gelman, Andrew, and Blei, David. Automatic variational inference in stan. In *Advances in neural information processing systems*, pp. 568–576, 2015.

Kucukelbir, Alp, Tran, Dustin, Ranganath, Rajesh, Gelman, Andrew, and Blei, David M. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017a.

Kucukelbir, Alp, Wang, Yixin, and Blei, David M. Evaluating bayesian models with posterior dispersion indices. In *International Conference on Machine Learning*, pp. 1925–1934, 2017b.

Li, Yingzhen and Turner, Richard E. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pp. 1073–1081, 2016.

Ranganath, Rajesh, Gerrish, Sean, and Blei, David. Black box variational inference. In *Artificial Intelligence and Statistics*, pp. 814–822, 2014.

Regier, Jeffrey, Jordan, Michael I, and McAuliffe, Jon. Fast black-box variational inference through stochastic trust-region optimization. *arXiv preprint arXiv:1706.02375*, 2017.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Roeder, Geoffrey, Wu, Yuhuai, and Duvenaud, David K. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pp. 6928–6937, 2017.

Ruiz, Francisco R, AUEB, Michalis Titsias RC, and Blei, David. The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*, pp. 460–468, 2016.

Schulman, John, Heess, Nicolas, Weber, Theophane, and Abbeel, Pieter. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, pp. 3528–3536, 2015.

Titsias, Michalis and Lázaro-Gredilla, Miguel. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1971–1979, 2014.

Tucker, George, Mnih, Andriy, Maddison, Chris J, Lawson, John, and Sohl-Dickstein, Jascha. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pp. 2624–2633, 2017.

Wainwright, Martin J, Jordan, Michael I, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

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

Zhang, Cheng, Butepage, Judith, Kjellstrom, Hedvig, and Mandt, Stephan. Advances in variational inference. *arXiv preprint arXiv:1711.05597*, 2017.