Wild Variational Approximations

Yingzhen Li  
University of Cambridge  
Cambridge, CB2 1PZ, UK  
yl494@cam.ac.uk

Qiang Liu  
Dartmouth College  
Hanover, NH 03755, USA  
qiang.liu@dartmouth.edu

Abstract

We formalise the research problem of approximate inference in the wild: developing new variants of variational methods that work for arbitrary variational approximation families for which inference (e.g., sampling or calculating expectation) is tractable, but probability density function may be intractable. We provide several examples for this type of approximations, discuss energy/gradient approximation for existing techniques, and further comment on developing other variational objectives and amortising stochastic dynamics. Connections and comparisons to existing approaches are also briefly discussed.

1 Introduction

For many machine learning tasks, a probabilistic model is fitted to the underlying distribution of the observed data. In the following we discuss w.l.o.g. latent variable models denoted by \( p(x|z; \theta) \) with prior \( p_0(z) \), although the presented approaches extend to the general case. Here \( z \) denotes the latent variables that a Bayesian approach would integrate out, e.g. the latent representation of deep generative models and the weight matrices of Bayesian neural networks. The hyper-parameters are loaded in \( \theta \) which will be learned by (approximate) maximum likelihood estimation (MLE), which requires the marginal likelihood \( p(x|\theta) \). Also given an observation \( x \), inference requires computing the exact posterior \( p(z|x; \theta) = \frac{1}{p(x|\theta)} p_0(z) p(x|z; \theta) \). Throughout this paper we assume the log-likelihood terms are tractable, but even so both quantities are intractable and hence require approximations in most cases.

Practical approaches for Bayesian inference include sampling-based and optimisation-based methods. Sampling-based methods, e.g. Markov chain Monte Carlo (MCMC) [1, 2, 3, 4] approximate these quantities by drawing samples from the exact posterior \( z^k \sim p(z|x; \theta) \) and use them later for inference and prediction. However most of these approaches are unbiased only in asymptotes, and in practice they can be computationally challenging for big models. In contrast, optimisation-based methods provide fast yet powerful tools for Bayesian inference of large scale. These methods explicitly define an approximate posterior distribution \( q(z|x) \), fit it to the exact posterior by optimising some objective function \( \mathcal{L}(\theta, q; x) \), and replaces the exact posterior with \( q(z|x) \) in inference/prediction time. An example of optimisation-based methods is variational inference (VI) [5, 6], which maximises the variational lower-bound in some distribution family \( \mathcal{Q} \):

\[
\max_{q \in \mathcal{Q}} \mathcal{L}_{VI}(\theta, q; x) = \mathbb{E}_q \left[ \log \frac{p_0(z)p(x|z; \theta)}{q(z|x)} \right].
\] (1)

Maximising this lower-bound is equivalent to minimising the KL-divergence \( \text{KL}[q(z|x)||p(z|x; \theta)] \). Furthermore this lower-bound can also be used as a surrogate loss function for maximum likelihood estimation (MLE), which optimises \( \mathcal{L}_{VI}(\theta, q; x) \) w.r.t. \( \theta \) given the training data \( x \).

The first batch of VI publications utilised \( q \) distributions of simple forms, e.g. factorised Gaussians and exponential families, in order to allow a closed-form calculation for inference. Recently Monte Carlo

Submitted to the Approximate Inference workshop, NIPS 2016, Barcelona, Spain.
truly black-box' approximations, or in the language of this paper 'wild variational approximations'.

The approximation $q$ is fitted to the VI as the reparameterization trick. A wild variational approximation to the exact posterior is defined as follows.

A distribution $q(z|x)$ is said to be a wild variational approximation to $p(z|x)$ if:

(i) it is fitted to the $p(z|x)$ using an optimisation-based method;

(ii) inference with $q(z|x)$ is comparatively easier, i.e. for the function $F(z)$ in interest, it is easier to compute (or estimate with MC methods) $E_{q(z|x)}[F(z)]$ than $E_{p(z|x)}[F(z)]$;

(iii) the density $q(z|x)$ is not necessary computable in a fast way.

We provide several examples in the following, with $\phi$ denoting all the trainable parameters for the $q$ distribution. In this paper we consider gradient-based optimisation methods for learning $\phi$.

Example 1. (Generative model) Sampling $z \sim q(z|x)$ is defined by a generative model that transforms random noise $\epsilon \sim p(\epsilon)$ with a mapping $z = f(\epsilon, x)$. A prevalent example of such mapping is a (deep) neural network which takes $x$ and $\epsilon$ as input and $z$ as the output. It has been introduced to VI as the reparameterization trick [18][19][20], hence we also use reparameterizable proposal for reference. The underlying distribution $q$ is often intractable, or requires further approximations to be computed efficiently.

Example 2. (Truncated Markov chain) Here the samples $z \sim q(z|x)$ is defined by $T$-step transitions of a Markov chain. Examples include $T$-step Gibbs sampling process of a restricted Boltzmann machine [21], or $T$-step simulation of an SG-MCMC algorithm such as SGLD [4]. In the latter case the trainable parameters are the step-size and/or the preconditioning matrix and so on. Related work includes [22] which proposed an auxiliary lower-bound as the optimisation objective for this type of variational approximations.

Example 3. (Stochastic gradient descent (SGD) with constant/adaptive learning rates) It has been shown in [23][24] that the trajectory of SGD with constant/adaptive learning rates near a local optimum can be viewed as a variational approximation to the exact posterior. It has similar trainable parameters $\phi$ as for truncated SG-MCMC algorithms. These methods are more expensive when producing samples since they require evaluations of $\nabla z \log p(x, z)$, but still they can be much cheaper than sampling from the exact posterior.
1. Though in this case we train a classifier \( \tilde{A} \), a new direction for energy approximation applies density ratio estimation methods [33, 34, 35]. This straight-forward method in this class considers density estimation based on the samples generated by the gradient approximation for general \( \tilde{A} \). The auxiliary distribution \( \tilde{q} \) can use sample-based density ratio estimation methods to fit a model \( \tilde{A} \) of which the entropy term \( \tilde{A} \) is approximated as \( \tilde{A} = \log \hat{\tilde{A}} \). In the following we discuss three potential solutions to this request.

2.2 Approximations for Optimisation

Energy Approximation. Assume \( \bar{q} \) is reparameterizable, then by the chain rule the gradient \( \nabla_{\bar{q}} \mathcal{L} \) is computed as \( \nabla_{\bar{q}} \mathcal{L} = \nabla_{\bar{q}} \hat{\mathcal{L}} \). This means if we have an approximation \( \hat{\mathcal{L}} \) to the objective function, then the gradient can be approximated by \( \nabla_{\bar{q}} \mathcal{L} \approx \nabla_{\bar{q}} \hat{\mathcal{L}} \). We name this approach as energy approximation, since often the optimising objective can be interpreted as an energy function. For non-reparameterizable distributions \( \nabla_{\bar{q}} \hat{\mathcal{L}} \) can be computed with further approximations such as the generalised reparameterization trick [30].

A straight-forward method in this class considers density estimation based on the samples generated by \( \bar{q} \). In this case a density estimator \( \bar{q} \), e.g. kernel density estimators (KDE) [31] or those parametrised by a neural network [32], is fitted to the samples \( \{z^k = f(\epsilon^k, x)\} \sim \bar{q} \), and the gradient of \( \log \bar{q} \) is approximated as \( \nabla_{\bar{q}} \log \bar{q} = \nabla_{z} \hat{\bar{q}}(z|x) \nabla_{\bar{q}} f \). One might even want to directly estimate \( \log \bar{q} \) if it turns out to be more accurate. However practitioners should be careful for implementations with automatic differentiation tools, since the parameters of the density estimator \( \hat{\bar{q}} \) should not be differentiated through (even though they depend on the samples \( z^k \)).

The next proposal directly approximate (part of) the energy function, e.g. the entropy term \( \mathbb{H}[\bar{q}] \) or the KL-divergence \( \text{KL}[\bar{q}||p_0] \) in \( \mathcal{L}_{\bar{q}} \), and let the MC approach handle the rest. In MC-dropout [28] \( \text{KL}[\bar{q}||p_0] \) is approximated by the Frobenius norm of the network weights \( \frac{1}{2} \|W\|^2 \) with dropout rate \( p \) and length-scale \( l \) defined by the prior. Similar approximations are in development process for \( \alpha \)-divergence approaches, aiming at extending SRTs to those variational methods.

A new direction for energy approximation applies density ratio estimation methods [33, 34, 35]. This is done by introducing an auxiliary distribution \( \hat{\bar{q}} \) and rewrite the variational lower-bound:

\[
\mathcal{L}_{\bar{q}}(\theta; q; x) = \mathbb{E}_q \left[ \log \frac{p_0(z)p(x|z; \theta)}{\hat{\bar{q}}(z|x)} + \log \frac{\hat{\bar{q}}(z|x)}{\bar{q}(z|x)} \right].
\]

(2)

The auxiliary distribution \( \hat{\bar{q}} \) is required to have tractable density and is easy to sample from. Then one can use sample-based density ratio estimation methods to fit a model \( \tilde{R} \) for the ratio between \( \hat{\bar{q}} \) and \( \bar{q} \). The gradient approximation for general \( \hat{\bar{q}} \) distributions can be derived similarly as

\[
\nabla_{\bar{q}} \mathcal{L}_{\bar{q}} = \mathbb{E}_q \left[ \nabla_{\bar{q}} \log \frac{p_0(z)p(x|z; \theta)}{\hat{\bar{q}}(z|x)} + \nabla_{z} \tilde{R}(z) \nabla_{\bar{q}} f \right].
\]

(3)

A simple example considers \( \hat{\bar{q}} = p_0 \) and the classification approach for ratio estimation. This means we train a classifier \( D(z) \) sampled from \( p_0(z) = (1 + \exp[-\tilde{R}(z)])^{-1} \) to distinguish samples from \( p_0 \) and \( \bar{q} \). A related approach is the adversarial auto-encoder [36] which uses the prior distribution as an auxiliary. However the objective function proposed by [36] replaces the KL[\bar{q}||p_0] in the variational lower-bound with Jensen-Shannon divergence. Also the presented method can be extended to a sequence of auxiliary distributions (in similar spirit as the annealed importance sampling [37]), which can also be adapted slowly during training in order to obtain a better approximation.

Direct Gradient Approximation. The gradient of an accurate energy approximation might not necessarily be a good estimator for the exact gradient \( \nabla_{\bar{q}} \mathcal{L} \). Therefore direct gradient approximation to the exact gradient might be preferred, if one cares less about the accuracy of the approximate variational lower-bound. There exists a rich literature on (non-parametric) derivative estimation
When the step-size $\epsilon$ MCMC and particle-based approximate inference methods Amortising Stochastic Dynamics.

Acknowledgements will also be tested.

(approximate) marginalisation of latent variables is desirable. Instead we consider amortized inference (similar to GAN [45]). In contrast analytic solution for the maximiser exists if $Q$ contains it. However simply replacing the objective with say other $f$-divergences will not make the optimisation easier as $q$ has an intractable density. Neither the variational techniques for estimating $f$-divergence [43,44] as the exact posterior is difficult to sample from.

One promising direction is to replace KL divergence with the Stein discrepancy which has a special form that does not require evaluating $q(x)$ for minimisation. Briefly speaking, Stein discrepancy involves a linear functional operator $\mathcal{T}$, called Stein operator, on a set of test functions $\mathcal{G} = \{g(z)\}$ such that $\mathbb{E}_{p(z|x)}[(\mathcal{T}g)(z)] = 0$ for $\forall g \in \mathcal{G}$. Then the associated Stein discrepancy is defined as $\mathcal{S}(q \| p) = \sup_{g \in \mathcal{G}} \mathbb{E}_q[(\mathcal{T}g)(z)]$. For continuous density functions, a generic Stein operator is $\mathcal{T}g = (\nabla_x \log p(z|x), g(z)) + \nabla_z \cdot g(z)$, for which $\mathbb{E}_{p(z|x)}[(\mathcal{T}g)(z)] = 0$, called Stein’s identity, can be easily verified using integration by parts. Very recently [49] defined $g$ as parametric functions represented by neural networks, and approximate the minimax optimisation with gradient descent (similar to GAN [45]). In contrast analytic solution for the maximiser $g^\star$ exists if $\mathcal{G}$ is defined as the unit ball of a RKHS, in which case we can find the optimal $q$ by standard stochastic optimisation for minimising $\mathcal{S}(q \| p)$.

Amortising Stochastic Dynamics. MCMC and particle-based approximate inference methods [46,47], though very accurate, become inefficient when inference from multiple different distributions is repeatedly required. As an example consider learning a (deep) generative model, where fast minimising $\mathcal{Q}$ involves a linear functional operator $G$, called Stein operator, on a set of test functions $G = \{g(z)\}$ such that $\mathbb{E}_{p(z|x)}[(Gg)(z)] = 0$ for $\forall g \in G$. Then the associated Stein discrepancy is defined as $\mathcal{S}(q \| p) = \sup_{g \in G} \mathbb{E}_q[(Gg)(z)]$. For continuous density functions, a generic Stein operator is $Gg = (\nabla_x \log p(z|x), g(z)) + \nabla_z \cdot g(z)$, for which $\mathbb{E}_{p(z|x)}[(Gg)(z)] = 0$, called Stein’s identity, can be easily verified using integration by parts. Very recently [48] applied the amortized SVGD idea to learning energy-based models. Future work in this direction will consider SG-MCMC as the stochastic transition model, and alternative measure to $l_2$-norm (e.g. maximum mean discrepancy) will also be tested.

3 Discussion

In this short paper we presented the research problem of constructing wild variational approximations to the exact posterior. The development of wild variational approximation methods aims at simplifying the design and application of approximate inference methods, allowing practitioners to focus more on selecting an appropriate approximate distribution that suits the best with their needs. But still our approach encourages the control of inference procedure (through the design of approximate posterior and optimisation procedure), unlike previous research of automated methods that implemented the inference engine transparently to the users. Future studies will develop better methods, more applications and theoretical understandings for wild variational approximations, and we hope our efforts can potentially motivate new ideas in the field.

Acknowledgements: Yingzhen Li thanks the Schlumberger Foundation FFTF fellowship.

References


