Refining the variational posterior through iterative optimization

Marton Havasi *
Department of Engineering
University of Cambridge
mh740@cam.ac.uk

Jasper Snoek
Google Research
jsnoek@google.com

Dustin Tran
Google Research
trandustin@google.com

Jonathan Gordon
Department of Engineering
University of Cambridge
jg801@cam.ac.uk

José Miguel Hernández-Lobato
Department of Engineering
University of Cambridge,
Microsoft Research,
Alan Turning Institute
jmh233@cam.ac.uk

Abstract

Variational inference is a popular approach for approximate Bayesian inference that is particulary promising for highly parameterized models such as deep neural networks. A key challenge of variational inference is to approximate the posterior over model parameters with a distribution that is simpler and tractable yet sufficiently expressive. In this work, we propose a method for training highly flexible variational distributions by starting with a coarse approximation and iteratively refining it. Each refinement step makes cheap, local adjustments and only requires optimization of simple variational families. We demonstrate theoretically that our method always improves a bound on the approximation (the Evidence Lower BOund) and observe this empirically across a variety of benchmark tasks. In experiments, our method consistently outperforms recent variational inference methods for deep learning in terms of log-likelihood and the ELBO.

1 Introduction

Exact Bayesian inference is intractable in general for neural networks. To model epistemic uncertainty, variational inference (VI) instead approximates the true posterior with a simpler distribution. The most widely used one for neural networks is the mean-field approximation, where the posterior is represented using an independent Gaussian distribution over all the weights.

Variational inference is appealing since it reduces the problem of inference to an optimization problem, minimizing the discrepancy between the true posterior and the variational posterior. The key challenge, however, is the task of training expressive posterior approximations that can capture the true posterior without significantly increasing the computational costs. This paper describes a novel method for training highly flexible posterior approximations that do not pose a significant overhead, when compared with mean-field variational inference.

The idea is to start with a mean-field variational approximation $q(w)$ and iteratively refine it. The method is a novel variant of the auxiliary variable approaches to VI (Agakov & Barber, 2004; Ranganath et al., 2016). We augment the model parameters w using a number of auxiliary variables a_k

*Work done as a Google Brain intern.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

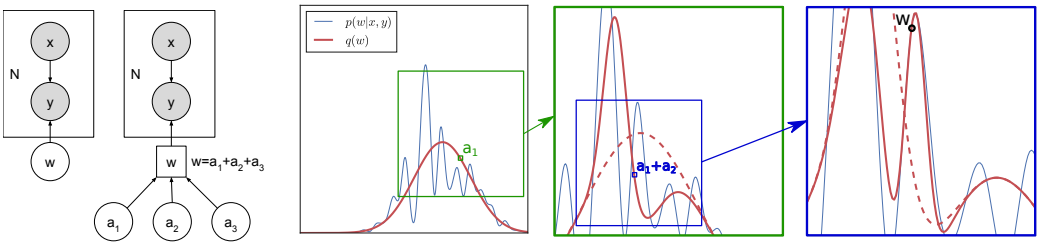

Figure 1: (Left) The supervised learning model and augmented model respectively where w is expressed as a sum of independent auxiliary variables. (Right) Illustration of the refining algorithm. In each iteration the value of an auxiliary variable is fixed and the posterior is locally adjusted. In the final iteration, a sample is drawn from w. Through the iterations, the variational distribution is able to well approximate the true posterior in a small region.

(Figure 1 shows the corresponding graphical models) for $k = 1, \ldots, K$ that leave the marginal distribution of the parameters unchanged. In each iteration, we sample the value of an auxiliary variable according to the current variational approximation $q(a_k)$ and refine the approximation by conditioning it on the newly sampled value $q(w) \approx p(w|x, y, a_{1:k})$. Each refinement step makes cheap, local adjustments to the variational posterior in the region of the sampled auxiliary variables. At the end, we draw one sample from the refined $q(w)$. The refinement iterations have to be repeated for each posterior sample. The algorithm results in samples from a highly complex distribution, starting from a simple mean-field approximation. While the distribution of the samples is difficult to quantify, it is not limited to factorized, uni-modal forms, and we show that the procedure is guaranteed to improve the resulting ELBO.

2 Methods

Variational Inference Variational inference attempts to approximate the true posterior $p(w|x, y)$ with an approximate posterior $q_\phi(w)$, typically from a simple family of distributions, for example independent Gaussians over the weights i.e. the mean-field approximation. To ensure that the approximate posterior is close to the true posterior, the parameters of $q_\phi(w)$, ϕ are optimized to maximize the Evidence Lower Bound (ELBO), which is a lower bound to the log marginal likelihood:

$$\log p(y|x) = \underbrace{D_{\mathrm{KL}}\big(q_\phi(w) \,\big\|\, p(w|x,y)\big)}_{\geq 0} + \mathcal{L}(\phi) \geq \mathcal{L}(\phi) = \mathbb{E}_{q_\phi}\big[\log p(y|x,w)\big] - D_{\mathrm{KL}}\big(q_\phi(w) \,\big\|\, p(w)\big)\,, \tag{1}$$

since the KL-divergence is non-negative.

For a new input x', the predictive distribution $p(y'|x')$ can be approximated by stochastically drawing a small number (around $M \leq 10$) of sample model parameters and averaging their prediction in an ensemble model:

$$w_{1:M} \sim q_\phi(w), \quad p(y'|x') \approx \frac{1}{M}\sum_{i=1}^{M} p(y'|x', w_i)\,. \tag{2}$$

Refining the posterior The main issue with variational inference is the inflexibility of the posterior approximation. Our idea is to refine the posterior approximation through iterative optimization. Since only a small number of samples ($M \leq 10$) are drawn for prediction, it is feasible to train a detailed posterior in the regions of these samples while relying on a coarse-grained approximation further away.

More precisely, we augment the graphical model with a finite number of auxiliary variables $a_{1:K}$ as shown on Figure 1. The constraints are that (x, y) must be conditionally independent of the auxiliary variables given w and that they must not affect the prior distribution $p(w)$. This is important in justifying the use of the initial variational approximation. While we are focusing on one

specific definition of the auxiliary variables, additive auxiliary variables, note that all of our results straight-forwardly generalize to arbitrary joint distributions $p(w, a_{1:K})$ that meet the constraints above. Given a Gaussian prior $\mathcal{N}(0, \sigma_w^2)$ over w, we express w as a sum of independent auxiliary variables

$$w = \sum_{i=1}^{K} a_i, \quad \text{with} \quad p(a_i) = \mathcal{N}(0, \sigma_{a_i}^2) \quad \text{for } i = 1, \ldots, K\,, \tag{3}$$

while ensuring that $\sum_{i=1}^{K} \sigma_{a_i}^2 = \sigma_w^2$ so that the prior $p(w) = \mathcal{N}(0, \sigma_w^2)$ is unchanged.

Locally refining the approximate posterior refers to iteratively sampling the values to the auxiliary variables and then approximating the posterior conditional on the sampled values i.e. $q_{\phi_k}(w)$ approximates $p(w|x, y, a_{1:k})$ for iterations $k = 1, \ldots, K$. Starting from the the initial mean field approximation $q_\phi(w)$, sample the value of a_1 from $q_\phi(a_1) = \int p(a_1|w) q_\phi(w)\, \mathrm{d}x$, then optimize the approximation to the conditional posterior: $q_{\phi_1}(w) \approx p(w|x, y, a_1)$. This procedure is then iteratively repeated for $a_{2:K}$. In iteration k,

$$\textbf{1)} \quad a_i \sim \int p(a_i|a_{1:k-1}, w) q_{\phi_{k-1}}(w)\, \mathrm{d}w \quad \textbf{2)} \quad \phi_k = \arg\min D_{\mathrm{KL}}\big(q_{\phi_k}(w) \,\big\|\, p(w|x, y, a_{1:k})\big)\,. \tag{4}$$

Analogously to variational inference, the KL divergence is minimized through the optimization of the conditional ELBO in each iteration: $\mathcal{L}_{|a_{1:k}}(\phi_k) = \mathbb{E}_{q_{\phi_k}}\big[\log p(y|x, w)\big] - D_{\mathrm{KL}}(q_{\phi_i}(w) \parallel p(w|a_{1:k}))$. In order to get independent samples from the variational posterior, we have to repeat the iterative refinement for each ensemble member $w_{1:M}$.

Theoretical justification Our theoretical claims are twofold. Firstly, that through this procedure, we are optimizing a lower bound to the ELBO and secondly, that the refinement cannot result in a worse posterior approximation than the initial mean-field approximation that we start with (in the ELBO sense). That is

$$\mathrm{ELBO}_{\text{ref}} \geq \mathrm{ELBO}_{\text{aux}} \geq \mathrm{ELBO}_{\text{init}}\,, \tag{5}$$

where $\mathrm{ELBO}_{\text{ref}}$ denotes the ELBO of the refined posterior, $\mathrm{ELBO}_{\text{aux}}$ refers to the objective that the refinement process is optimizing and $\mathrm{ELBO}_{\text{init}}$ is the ELBO of the initial variational approximation.

The former, $\mathrm{ELBO}_{\text{ref}} \geq \mathrm{ELBO}_{\text{aux}}$, can be shown analogously to Ranganath et al. (2016) while the latter, $\mathrm{ELBO}_{\text{aux}} \geq \mathrm{ELBO}_{\text{init}}$, holds because it can be ensured that the refinement steps do not result in a local optima worse than the initial variational approximation.

3 Experiments

To quantify the benefits of the refinement, we conducted experiments on a selection of regression and classification benchmarks using Bayesian neural networks as the underlying model. We compared the marginal log-likelihood and the ELBO to the baseline models: Deep Ensembles (Lakshminarayanan et al., 2017) and Multiplicative Normalizing Flows (Louizos & Welling, 2017) and Variational Inference trained for 30000 iterations using Adam.

Refinement In the experiments, we refine $M = 10$ ensemble members, each with $K = 5$ auxiliary variables. The means on their prior distributions are fixed at 0., and their variances form a geometric series (each auxiliary variable reduces the variance of the prior by a factor of 0.7, which roughly halves its standard deviation): $\sigma_{a_1}^2 = 0.7\sigma_w^2$, $\sigma_{a_2}^2 = 0.21\sigma_w^2$, $\sigma_{a_3}^2 = 0.063\sigma_w^2$, $\sigma_{a_4}^2 = 0.0189\sigma_w^2$, and $\sigma_{a_5}^2 = 0.0081\sigma_w^2$. σ_w was tuned with empirical Bayes. In each refinement iteration, we optimized the posterior with Adam for 200 iterations.

	Deep Ensemble	MNF	VI		Refined VI	
	MLL& Acc	MLL & Acc	MLL & Acc	ELBO	MLL & Acc	ELBO
boston_housing	-9.136 (5.719)	**-2.920** (0.133)	**-2.874** (0.151)	-668.272 (7.647)	**-2.851** (0.185)	$\geq$ -630.379 (7.716)
concrete_strength	-4.062 (0.130)	-3.202 (0.055)	**-3.138** (0.063)	-3248.137 (68.575)	**-3.131** (0.062)	$\geq$ -3071.124 (64.046)
naval_propulsion	3.995 (0.013)	3.473 (0.007)	**5.969** (0.245)	53440.701 (2047.340)	**6.128** (0.171)	$\geq$ 54882.656 (1228.361)
energy_efficiency	**-0.666** (0.058)	-0.756 (0.054)	-0.749 (0.068)	-1296.721 (66.310)	**-0.707** (0.094)	$\geq$ -1192.337 (62.089)
yacht_hydrodynamics	**-0.984** (0.104)	-1.339 (0.170)	-1.749 (0.232)	-928.758 (112.928)	-1.626 (0.231)	$\geq$ -790.052 (84.716)
kin8nm	**1.135** (0.012)	**1.125** (0.022)	1.066 (0.019)	6071.268 (61.758)	1.069 (0.018)	$\geq$ 6172.709 (67.659)
power_plant	-3.935 (0.140)	**-2.835** (0.033)	**-2.826** (0.020)	-22496.579 (130.487)	**-2.820** (0.024)	$\geq$ -22368.965 (85.308)
protein_structure	-3.687 (0.013)	**-2.928** (0.007)	**-2.926** (0.010)	-108806.007 (174.522)	**-2.923** (0.009)	$\geq$ -108597.593 (158.482)
wine	**-0.968** (0.079)	**-0.963** (0.027)	**-0.973** (0.054)	-1346.130 (18.004)	**-0.968** (0.056)	$\geq$ -1311.898 (17.487)
mnist	**-0.017** (0.001) 99.4% (0.0)	-0.034 (0.002) 99.1% (0.1)	-0.032 (0.001) 99.1% (0.1)	-7618.533 (47.589)	-0.025 (0.001) 99.2% (0.0)	$\geq$ -6310.824 (42.357)
fashion_mnist	**-0.201** (0.002) 93.1% (0.1)	-0.255 (0.004) 90.7% (0.2)	-0.255 (0.003) 90.7% (0.1)	-22830.330 (232.654)	-0.241 (0.004) 91.3% (0.2)	$\geq$ -20438.955 (79.672)
cifar10	-0.791 (0.009) 76.3% (0.3)	-0.795 (0.013) 72.8% (0.6)	-0.815 (0.004) 72.3% (0.5)	-57257.887 (299.570)	**-0.768** (0.007) 73.5% (0.5)	$\geq$ -50989.217 (238.976)

Table 1: Results on the UCI regression benchmarks with a single hidden layer containing 50 units and on image classification datasets using the LeNet-5 convolutional architecture. Metrics: marginal log-likelihood (MLL), Accuracy (where applicable) and the evidence lower bound (ELBO). The mean values and standard deviations are shown in the table.

References

Felix V Agakov and David Barber. An auxiliary variational method. In *International Conference on Neural Information Processing*, pp. 561–566. Springer, 2004.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2218–2227. JMLR. org, 2017.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pp. 324–333, 2016.