Abstract

We present a black-box variational inference (BBVI) method to approximate intractable posterior distributions with an increasingly rich approximating class. Using mixture distributions as the approximating class, we first describe how to apply the re-parameterization trick and existing BBVI methods to mixtures. We then describe a method, termed **Variational Boosting**, that iteratively refines an existing approximation by defining and solving a sequence of optimization problems, allowing the practitioner to trade computation time for increased accuracy.

1 Introduction

Variational inference (VI) [7, 14, 1] is a family of methods designed to approximate an intractable target distribution (typically known only up to a constant) with a tractable surrogate distribution. VI procedures typically minimize the Kullback-Leibler (KL) divergence of the approximation to the target by maximizing an appropriately defined tractable objective. Often, the class of approximating distributions is fixed, and typically excludes the target distribution (and its neighbors), which prevents the variational approximation from becoming arbitrarily close to the true posterior.

Markov chain Monte Carlo (MCMC), an alternative class of inference methods, approximate target distributions with samples drawn from a Markov chain constructed to admit the target distribution at each marginal. MCMC methods allow a user to trade computation time for increased accuracy — drawing more samples will make the approximation closer to the true target distribution. However, MCMC algorithms typically must be run iteratively, making them difficult to parallelize. Furthermore, correctly specifying MCMC moves can be more algorithmically restrictive than optimizing an objective (e.g. data subsampling in stochastic gradient methods).

We propose a variational inference method that gradually allows the approximation to become more and more complex, affording the practitioner a trade-off between time and accuracy. Our method builds on black-box variational inference methods using the re-parameterization trick [13, 8, 10], applicable to a very general class of target distributions.

**Variational Inference** Given a target distribution with density $\pi(x)$ for a multivariate random variable $x \in \mathcal{X}$, variational inference approximates $\pi(x)$ with a tractable approximate distribution $q(x; \lambda)$, from which we can draw samples and form sample-based estimates of functions of $x$ (e.g. posterior credible intervals, Bayesian predictions, etc.). Variational methods typically minimize $KL(q||\pi)$ between the approximation and the target as a function of variational parameters $\lambda$. This is framed as an optimization problem — we optimize the KL objective (or an equivalent one) with

---

1 We assume $\pi(x)$ is known up to a constant, $\pi(x) = C \pi(x)$ for some constant $C$; we may omit $\sim$ simplicity.

2 We treat the density function as a synecdoche for the entire law, and use $q(x; \lambda)$ and $q_\lambda(x)$ interchangeably at the risk of slight notational abuse.
We define our approximate distribution to be a mixture of simpler component distributions

\[ q^{(C)}(x; \lambda) = \sum_{c=1}^{C} \rho_c q_c(x; \lambda_c) \quad \text{s.t.} \quad \sum_{c} \rho_c = 1 \quad (5) \]

where we have defined component distributions \( q_c \), mixture component parameters \( \lambda = (\lambda_1, \ldots, \lambda_C) \), and mixing proportion parameters \( \rho = (\rho_1, \ldots, \rho_C) \). The component distributions can be any distribution over \( x \) from which we can draw samples using a continuous map (e.g. multivariate normals \([6]\), or a composition of invertible maps \([11]\)).

Our method begins with a single mixture component, \( C = 1 \). We use existing VI methods to fit the first component parameter, \( \lambda_1 \), and \( \rho_1 \) is fixed to \( 1 \) by definition. At the next iteration, we fix \( \lambda_1 \) and introduce a new component into the mixture, \( q_2(x; \lambda_2) \). We then introduce a new ELBO objective as a function of new component parameters, \( \lambda_2 \), and a new mixture weight, \( \rho_2 \). We then optimize this objective until convergence. At each subsequent iteration, we introduce new component parameters and a mixing weight, and then we optimize the new objective. We refer to this procedure as \textit{variational boosting}, inspired by methods for learningstrong classifiers by weighting an ensemble of weak classifiers.\(^1\)

\(^1\)Often referred to as the \textit{marginal likelihood}, \( p(\text{data}) \), in Bayesian inference.

\(^2\)We denote full mixtures with parenthetical superscripts, \( q^{(C)} \), and components with naked subscripts, \( q_c \).
In order for our method to be applicable to a general class of target distributions, we use black-box variational inference methods and the re-parameterization trick [13, 8, 10] to fit each component and mixture weights. The re-parameterization trick is a method for obtaining unbiased estimates of the gradient of the ELBO. These gradient estimates can then be used to optimize the ELBO objective using a stochastic gradient optimization method. However, using mixtures as the variational approximation complicates the use of the re-parameterization trick. Details on the re-parameterization trick and its use in mixtures are in Appendix A.

2.1 Variational Boosting

Fitting the first component The procedure starts by fitting an approximation to \( \pi(x) \) with a distribution that consists of a single component. We do this by maximizing the first ELBO objective

\[
L^{(1)}(\lambda_1) = \mathbb{E}_q \left[ \ln \pi(x) - \ln q_1(x; \lambda_1) \right]
\]

\[
\lambda_1^* = \arg \max_{\lambda_1} L^{(1)}(\lambda_1).
\]

Depending on the forms of \( \pi \) and \( q_1 \), optimizing the ELBO can be accomplished by various methods. One general method for fitting a continuous valued component is to compute stochastic, unbiased gradients of \( L(\lambda_1) \), and use stochastic gradient optimization. See Appendix A for details. After convergence (or close to it) we fix \( \lambda_1 \) to be \( \lambda_1^* \).

Fitting component \( C + 1 \) After iteration \( C \), our current approximation to \( \pi(x) \) is a mixture distribution with \( C \) components

\[
q^{(C)}(x; \lambda) = \sum_{c=1}^{C} \rho_c q_c(x; \lambda_c)
\]

where \( \lambda = (\{\rho_c, \lambda_c\}_c) \) is a list of component parameters and mixing weights, and \( q_c(x; \lambda_c) \) is the component distribution parameterized by \( \lambda_c \). Adding a new component introduces a new component parameter, \( \lambda_{C+1} \), and a new mixing weight, \( \rho_{C+1} \). In this section, the mixing parameter \( \rho_{C+1} \in [0, 1] \) mixes between the new component, \( q_{C+1}(\cdot; \lambda_{C+1}) \) and the existing distribution, \( q^{(C)} \). The new approximate distribution is

\[
q^{(C+1)}(x; \rho_{C+1}, \lambda_{C+1}) = (1 - \rho_{C+1})q^{(C)}(x) + \rho_{C+1}q_{C+1}(x; \lambda_{C+1})
\]

The new optimization objective, as a function of \( \rho_{C+1} \) and \( \lambda_{C+1} \) is

\[
L^{(C+1)}(\rho_{C+1}, \lambda_{C+1}) = \mathbb{E}_{x \sim q^{(C+1)}} \left[ \ln \pi(x) - \ln q^{(C+1)}(x; \lambda_{C+1}, \rho_{C+1}) \right]
\]

\[
= (1 - \rho_{C+1})\mathbb{E}_{q^{(C)}} \left[ \ln \pi(x) - \ln q^{(C+1)}(x; \lambda_{C+1}, \rho_{C+1}) \right]
\]

\[
+ \rho_{C+1}\mathbb{E}_{q_{C+1}} \left[ \ln \pi(x) - \ln q^{(C+1)}(x; \lambda_{C+1}, \rho_{C+1}) \right]
\]

Above we have separated out two expectations — one with respect to the existing approximation (which is fixed), and the other with respect to the new component distribution. Because we fix the existing component distributions we only need to optimize the new component parameters \( \lambda_{C+1}, \rho_{C+1} \), allowing us to use the re-parameterization trick and Monte Carlo gradients to optimize \( L^{(C+1)} \). The details of component initialization are in Appendix B. Figure 1 depicts this procedure on a simple, two-dimensional target distribution.

2.2 Related Work

Using a mixture model as an approximating distribution in variational inference is a well-studied idea. Mixtures of mean field approximations [6] introduced mean field-like updates for a mixture approximation using a bound on the entropy term and model-specific parameter updates. Nonparametric variational inference [3] is a black-box variational inference algorithm that approximates a target distribution with a mixture of equally-weighted isotropic normals. The authors use a lower bound on the entropy term in the ELBO to make the optimization procedure tractable. Similarly,
present a method for fitting mixture distributions as an approximation. However, their method is restricted to mixture component distributions within the exponential family, and a joint optimization procedure. Another related thread of research is boosting density estimation [12], which iteratively improves unsupervised models of data. Finally, we note that [4] independently and in parallel propose a closely-related idea for an iterative “boosted” construction of a variational approximation.

3 Experiments and Analysis

Hierarchical Binomial Regression We test out our posterior approximation on a hierarchical binomial regression model[5]. We borrow an example from [2], and estimate the the binomial rates of success (batting averages) of baseball players using a hierarchical model — details of the model are in Appendix C.

To highlight the fidelity of our method, we compare Variational Boosting to mean field VI and the No-U-Turn Sampler (NUTS) [5]. The empirical distribution resulting from 20k NUTS samples is considered the “ground truth” posterior in this example. Figure 2 compares a selection of univariate and bivariate posterior marginals. We see that Variational Boosting is able to closely match the NUTS posterior estimate, allowing the user to improve upon the MFVI approximation.

4 Discussion and Conclusion

We have proposed a variational inference method that iteratively incorporates new components into the approximation. We see multiple directions for future work. The Variational Boosting framework allows for more flexible component distributions than diagonal Gaussians. For instance, compositions of invertible maps have been used to enrich variational families [11], as well as auxiliary variable variational models [9], both of which could be used as component distributions in a larger mixture. We also plan to explore alternative optimization procedures. While this work uses first-order stochastic gradient methods to fit variational approximations, natural gradient and second-order optimization methods have been shown to be effective and more efficient.

Acknowledgments

The authors would like to acknowledge Arjumand Masood, Mike Hughes, and Finale Doshi-Velez for helpful conversations.

References


A The re-parameterization trick

The re-parameterization trick is a method for computing low-variance estimates of the gradient of an objective for which we only have an unbiased estimator

\[ \mathcal{L}(\lambda) = \mathbb{E}_q [\ln \pi(x) - \ln q(x; \lambda)] \]

\[ \approx \frac{1}{L} \sum_{\ell=1}^{L} \left[ \ln \pi(x^{(\ell)}) - \ln q(x^{(\ell)}; \lambda) \right] \quad x^{(\ell)} \sim q(x; \lambda) \]

To obtain a Monte Carlo gradient of \( \mathcal{L}(\lambda) \) using the re-parameterization trick, we first separate the randomness needed to generate \( x^{(\ell)} \) from the parameters \( \lambda \), by defining a deterministic map \( f_q(x_0; \lambda) = x^{(\ell)} \) such that \( x_0 \sim q(\cdot; \lambda) \) implies \( x^{(\ell)} \sim q(x; \lambda) \). Then, we can differentiate through \( f_q \) with respect to \( \lambda \) to obtain a gradient estimator. This takes advantage of the following equivalence

\[ \mathcal{L}(\lambda) = \mathbb{E}_{x \sim q_\lambda} [\ln \pi(x) - \ln q(x; \lambda)] \]

\[ = \mathbb{E}_{x \sim q_0} \mathbb{E}_\lambda [\ln \pi(f_q(x_0; \lambda)) - \ln q(f_q(x_0; \lambda); \lambda)] \]

Now that the stochasticity has been separated from parameters \( \lambda \), we can move the gradient operator into the expectation

\[ \mathcal{L}(\lambda) = \nabla_\lambda \mathbb{E}_{x \sim q_0} [\ln \pi(f_q(x_0; \lambda)) - \ln q(f_q(x_0; \lambda); \lambda)] \]

\[ = \mathbb{E}_{x \sim q_0} \nabla_\lambda [\ln \pi(f_q(x_0; \lambda)) - \ln q(f_q(x_0; \lambda); \lambda)] \]

which directly translates the Monte Carlo objective estimator into a Monte Carlo objective gradient estimator.

A.1 The re-parameterization trick for mixtures

The re-parameterization trick when \( q \) is a mixture, however, is less straightforward. The sampling procedure for a mixture model typically contains a discrete component (i.e. sampling component identities), which is a process that cannot be differentiated through. We circumvent this complication by re-writing the variational objective as a weighted combination of expectations with respect to individual mixture components. Because of the form of the mixture, we can write the ELBO as

\[ \mathcal{L}(\lambda, \rho) = \mathbb{E}_q [\ln \pi(x) - \ln q(x; \lambda)] \]

\[ = \int \left( \sum_{c=1}^{C} \rho_c q_c(x; \lambda_c) \right) [\ln \pi(x) - \ln q(x; \lambda)] \, dx \]

\[ = \sum_{c=1}^{C} \rho_c \int q_c(x; \lambda_c) [\ln \pi(x) - \ln q(x; \lambda)] \, dx \]

\[ = \sum_{c=1}^{C} \rho_c \mathbb{E}_{q_c} [\ln \pi(x) - \ln q(x; \lambda)] \]

which is a function of expectations with respect to mixture components. If these distributions are continuous, and there exists some function \( f_c(x_0; \lambda_c) \) such that \( f_c(x_0; \lambda_c) \sim q_c(\cdot; \lambda_c) \) when \( x_0 \sim q_0 \), then we can apply the re-parameterization trick to each component to obtain gradients of the ELBO

\[ \nabla_{\lambda_c} \mathcal{L}(\lambda, \rho) = \nabla_{\lambda_c} \sum_{c=1}^{C} \rho_c \mathbb{E}_{x \sim q_c(x; \lambda_c)} [\ln \pi(x) - \ln q(x; \lambda)] \]

\[ = \sum_{c=1}^{C} \rho_c \mathbb{E}_{x \sim q_c} \left[ \nabla_{\lambda_c} \ln \pi(f_c(x_0; \lambda_c)) - \nabla_{\lambda_c} \ln q(f_c(x_0; \lambda_c)) \right]. \]

Variational Boosting uses the above fact to use the re-parameterization trick in a component-by-component manner, allowing us to improve the variational approximation as we incorporate and fit new components.

B Component Initialization

Initializing Components Introducing a new component requires initialization of component parameters. When our component distributions are mixtures of Gaussians, we found that the optimization procedure is sensitive to initialization. We found that the following importance-weighted scheme improves the optimization objective. To initialize a new component, \( \mu_{C+1} \), we first draw \( L \) samples from the existing distribution, \( x^L \sim q(C) \). For each sample, we compute an importance weight, \( \ln w^L = \ln \pi(x) - \ln q(C) \). We initialize \( \mu_{C+1} \) to the sample with the largest importance weight. We initialize this component to be small, and the new mixing weight to be small (around .01).

\(^6\)Here, \( q_0 \) is some base distribution that is, importantly, not a function of \( \lambda \).
C  Example Hierarchical Model

The model of the data is

\[
\begin{align*}
\phi & \sim \text{Unif} & \text{hyper prior} & (13) \\
\kappa & \sim \text{Pareto}(1, 1.5) & \text{hyper prior} & (14) \\
\theta_j & \sim \text{Beta}(\phi \cdot \kappa, (1 - \phi) \cdot \kappa) & \text{player } j \text{ prior} & (15) \\
y_j & \sim \text{Binomial}(K_j, \theta_j) & \text{player } j \text{ hits} & (16)
\end{align*}
\]

where \( y_j \) is the number of successes (hits) player \( j \) has attempted in \( K_j \) attempts (at bats). Each player has a latent success rate \( \theta_j \), which is governed by two global variables \( \kappa \) and \( \phi \). There are 18 players in this small example, with a total of \( D = 20 \) parameters. This model is not conjugate, and requires approximate Bayesian inference.