Black-box $\alpha$-divergence Minimization

José Miguel Hernández-Lobato, Yingzhen Li, Daniel Hernández-Lobato, Thang Bui, Richard E. Turner, Harvard University, University of Cambridge, Universidad Autónoma de Madrid.

Abstract

We present black-box alpha (BB-$\alpha$), an approximate inference method based on the minimization of $\alpha$-divergences between probability distributions. BB-$\alpha$ scales to large datasets since it can be implemented using stochastic gradient descent. BB-$\alpha$ can be applied to complex probabilistic models with little effort since it only requires as input the likelihood function and its gradients. These gradients can be easily obtained using automatic differentiation. By tuning the parameter $\alpha$, we are able to interpolate between variational Bayes and an expectation propagation like algorithm. Experiments on probit and neural network regression problems illustrate the accuracy of the posterior approximations obtained with BB-$\alpha$.

1 Introduction

Probabilistic models are very useful tools to make predictions from data. However, they require the computation of a posterior distribution which is often intractable. To avoid this, one can use approximate inference techniques. Two examples are expectation propagation (EP) and variational Bayes (VB) \cite{Jordan1999, Wainwright2008}. These methods adjust the parameters of a tractable distribution so that it is close to the exact posterior. This is done by finding the stationary point of an energy function. In EP this involves solving a minimization problem. This enables the use of stochastic optimization methods for large scale learning. We then generalize the energy function so that it is based on the local minimization of $\alpha$-divergences.

The energy function of EP (VB) includes expectations of (the logarithm of) the likelihood factors under a tractable distribution. This complicates the use of EP and VB in complex models for which there might not exist an analytic solution for these expectations. Monte Carlo sampling has been proposed to approximate the intractable expectations in VB \cite{Besag1980}. This approach is known as black-box VB and it allows VB to be applied to almost any model. Here we propose a similar method for the more general problem of $\alpha$-divergence minimization and call it black-box alpha. For this, we simplify the energy function of EP so that the min-max problem is transformed into a minimization problem. This enables the use of stochastic optimization methods for large scale learning. We then generalize the energy function so that it is based on the local minimization of $\alpha$-divergences.

2 The energy function of expectation propagation

Let us assume that we want to use EP to approximate the posterior distribution $p(x|D)$ for some model parameters $x$, where $D$ is a dataset with $N$ data points sampled independently. Then

$$ p(x|D) \propto \left[ \prod_{n=1}^{N} f_n(x) \right] p(x), \quad (1) $$

where each $f_n(x)$ is a likelihood factor, that is, $f_n(x) = p(y_n|x)$ where $y_n$ is the $n$-th data instance, and $p(x)$ is a prior distribution, which we assume to belong to the exponential family, that is, $p(x) = \exp\{s(x)^T\nu_0 - \Phi(\nu_0)\}$, where \(\nu\) and $s(x)$ are vectors of natural parameters and sufficient statistics, respectively, and $\Phi(\nu_0)$ is the logarithm of the normalization constant of $\exp\{s(x)^T\nu_0\}$.

We can use EP to approximate the posterior $p(x|D)$ with a distribution within the same exponential family as the prior. Let $q(x) = \exp\{s(x)^T(\nu + \nu_0)\}$ be the unnormalized posterior approximation.
used by EP. We have used here the parameterization \((\nu + \nu_0)\) to make explicit the contribution of the prior (given by \(\nu_0\)) and likelihood (given by \(\nu\)) in \(q\). EP also uses a cavity distribution with the same form as \(q\) for each likelihood factor. Each cavity distribution approximates the result of removing a likelihood factor from the true posterior. Let \(q^{\nu^\top} = \exp\{s(x)^\top(\lambda_n + \nu_0)\}\) be the unnormalized cavity distribution that approximates \(q\) when the \(n\)-th likelihood factor is removed. Again, we have used here a parameterization that makes explicit the contribution of the prior to each cavity.

According to \([1]\), the EP energy function is

\[
E(\nu, \{\lambda_n\}) = \Phi(\nu_0) + (N - 1) \log \int q(x)dx - \sum_{n=1}^{N} \log \int f_n(x)q^{\nu^\top}(x)dx.  \tag{2}
\]

This energy is equal to minus the logarithm of the EP approximation of the model evidence \(p(\mathcal{D})\), that is, the normalizer of the right-hand side of \([1]\) \([3]\). Therefore, minimizing \((2)\) with respect to \(\nu\) and \(\{\lambda_n\}\) is arguably a sensible way to tune these variational parameters. In practice, EP finds a stationary solution to the constrained optimization problem

\[
\min_{\nu} \max_{\{\lambda_n\}} E(\nu, \{\lambda_n\}) \quad \text{subject to} \quad (N - 1)\nu = \sum_{n=1}^{N} \lambda_n,  \tag{3}
\]

where the constraint in \((3)\) guarantees that the \(\{\lambda_n\}\) are valid cavity parameters. That is, the product of all the cavities is equal to \(N - 1\) times the posterior approximation (leaving the prior factor aside).

By replacing \(\nu + \nu_0\) with \(1/(N - 1) \sum_{n=1}^{N} (\lambda_n + \nu_0)\) in \((2)\) and then taking derivatives with respect to each \(\lambda_n\), we obtain the conditions that a solution to \((3)\) must satisfy:

\[
E_q[s(x)] = E_{f_nq^{\nu^\top}}[s(x)], \quad n = 1, \ldots, N.  \tag{4}
\]

That is, the expectation of \(s(x)\) with respect to the posterior approximation \(q(x)\) and with respect to the product of \(f_n(x)\) and \(q^{\nu^\top}(x)\), often called the **tilted** distribution, must be the equal.

The problem in \((4)\) can be solved using a double-loop algorithm \([5, 6]\). This algorithm alternates between an optimization of the cavity parameters \(\{\lambda_n\}\) in the inner loop and an optimization of the parameters of the posterior approximation \(\{\nu\}\) in the outer loop. Each iteration of the double-loop algorithm is guaranteed to minimize the energy in \((2)\). However, the alternating optimization of \(\{\nu\}\) and \(\{\lambda_n\}\) is very inefficient and the double-loop algorithm often requires too many iterations to be useful in practice.

### 2.1 Energy optimization with tied parameters

To simplify the previous optimization problem, we assume that all the cavity parameters are equal and have a simple form. We assume that each \(\lambda_n\) is obtained as a function of \(\nu\) according to

\[
\lambda_n = \frac{\lambda_n}{N} \nu,  \tag{5}
\]

which guarantees that the constraint in the right-hand side of \((3)\) is satisfied. All the cavities are now the same and are obtained by scaling \(\nu\) in \(q\) by \((N - 1)/N\). This assumption was made in \([7]\) to obtain a version of EP called **Stochastic Expectation Propagation** that has low memory consumption. Here, we use this assumption to obtain another version of EP that is guaranteed to converge and can be implemented by optimizing an energy function without requiring a double-loop algorithm. Recall that \(\nu\) captures the contribution of the likelihood to the posterior. Thus, with \((5)\) we are assuming that each factor \(f_n(x)\) in \([1]\) has an equal contribution to the likelihood and when we remove that factor, we only have to scale \(\nu\) by \((N - 1)/N\).

After replacing \(\lambda_n + \nu_0\) in \((2)\) with the right-hand side of \((5)\) and taking derivatives with respect to \(\nu\), we obtain the new stationary conditions:

\[
E_q[s(x)] = \frac{1}{N} \sum_{n=1}^{N} E_{f_nq^{\nu^\top}}[s(x)].  \tag{6}
\]

Therefore, the expectation of \(s(x)\) with respect to \(q(x)\) is going to be equal to the average of the expectation of \(s(x)\) across the different tilted distributions \(f_n(x)q^{\nu^\top}(x)\), for \(n = 1, \ldots, N\).

The solution that minimizes \((2)\) with respect to \(\nu\) subject to \((5)\) is expected to converge to the solution of \((3)\) when more and more data are available. As \(N\) grows, we expect \(q\) and the cavities to become very peaked. When this happens, the contribution of each likelihood factor \(f_n(x)\) to the tilted distribution \(f_n(x)q^{\nu^\top}(x)\) becomes very small because \(f_n(x)\) is a rather flat function when compared to \(q^{\nu^\top}(x)\), which is very peaked. Therefore, as the amount of data \(N\) increases, we expect all the terms \(E_{f_nq^{\nu^\top}}[s(x)]\) in \((6)\) to be very similar to each other. When all of them are equal, we have that \((6)\) implies \((4)\).
The convergence of a method that minimizes (2) under (5) is guaranteed if (2) is bounded below. This is going to be the case if we assume that all the factors \{f_n(x)\} in (1) are bounded, that is, \(f_n(x) \leq c\) for some \(c > 0\). See [11] for further details.

We can optimally adjust the prior hyper-parameters \(\nu_0\) by minimizing (2) with respect to \(\nu_0\) under the new form for the cavities given by (5). After taking gradients with respect to \(\nu_0\) and equating them to zero, we obtain the stationary condition \(E_{\nu_0}[s(x)] = E_q[s(x)]\). That is, at convergence, the expectations of \(s(x)\) with respect to the posterior approximation \(q(x)\) and with respect to the prior \(p(x)\) should be the same.

### 3 Local minimization of \(\alpha\)-divergences

The energy function given in (2) can be derived as the approximation to the logarithm of the marginal likelihood given by a message passing algorithm. This algorithm locally minimizes the Kullback-Leibler (KL) distance between the tilted distributions and the posterior approximation \(q\). A generalized version of (2) is obtained by running a message passing algorithm that locally minimizes a generalization of the KL divergence called the \(\alpha\)-divergence [3]. The new energy function is

\[
E_\alpha(\nu, \{\lambda_n\}) = \Phi(\nu_0) + \left(\frac{N}{\alpha} - 1\right) \log \int q(x) dx - \sum_{n=1}^N \frac{1}{\alpha} \log \int f_n(x)^\alpha q_\alpha^n(x) dx.
\]

(7)

where \(q_\alpha^n(x) = \left[q^n(x)/q(x)\right]^\alpha q(x)\) and \(\alpha \in (-\infty, \infty) \setminus \{0\}\) is a parameter that determines the form of the \(\alpha\)-divergence that is used to measure the distance between the \(n\)-th tilted distribution \(f_n(x)q^n(x)\) and the posterior approximation \(q(x)\). Again, we can simplify the optimization of this new energy function by assuming that all the cavity parameters are equal and have the form given by (5). This results in \(q_\alpha^n(x) = \exp\{s(x)^T \beta_\alpha\}\) where \(\beta_\alpha = (N - \alpha)/N\nu + \nu_0\).

When \(\alpha\) takes value one we have that the minimization of (7) under (5) is equivalent to the minimization of (2) under (5). In the limit that \(\alpha\) approaches zero, we have that the minimizer of (7) under (5) converges to the solution that maximizes the variational lower bound. Therefore, by adjusting the \(\alpha\) parameter, we can interpolate between the solutions given by variational Bayes and the solutions given by the approximation to expectation propagation described in Section 2.1.

#### 3.1 Large scale learning

When \(N\) is very large, we can minimize (7) under (5) using stochastic optimization techniques. In particular, we can uniformly sample a mini-batch of data \(S \subseteq \{1, \ldots, N\}\) and construct the noisy estimate of the energy function given by

\[
\hat{E}_\alpha(\nu) = \Phi(\nu_0) + \left(\frac{N}{\alpha} - 1\right) \log \int q(x) dx - \frac{N}{|S|} \sum_{n \in S} \frac{1}{\alpha} \log \int f_n(x)^\alpha q_\alpha^n(x) dx.
\]

(8)

The gradients of (8) can then used to minimize the original objective by stochastic gradient descent.

#### 3.2 Black-box \(\alpha\)-divergence minimization

In complicated probabilistic models, we might not be able to analytically solve the integrals in (8) involving the likelihood factors. However, we can obtain an estimate of these integrals by Monte Carlo. For this, we draw \(K\) samples \(x_1, \ldots, x_K\) from \(q_\alpha^n(x)\) and then approximate the integrals by expectations with respect to those samples. This produces the following new noisy estimate of the energy function

\[
\hat{E}_\alpha(\nu) = \Phi(\nu_0) + \left(\frac{N}{\alpha} - 1\right) \log \int q(x) dx - \frac{N}{|S|} \sum_{n \in S} \left[\frac{1}{K} \log \frac{1}{K} \sum_{k=1}^K f_n(x_k)^\alpha\right] - \frac{N}{|S|} \Phi(\beta_\alpha),
\]

(9)

where \(\Phi(\beta_\alpha)\) is the log-normalizer of \(q_\alpha^n(x)\). Since \(q_\alpha^n(x)\) and its samples \(x_1, \ldots, x_K\) depend on \(\nu\), we can then use the reparametrization trick described in [5] to obtain the correct gradients of the Monte Carlo estimator. Note, however, that the resulting stochastic gradients will be biased because the energy function (9) applies a non-linear transformation (the logarithm) to the Monte Carlo estimator of the integrals. Nevertheless, this bias can be reduced by increasing the number of samples \(K\).

Given a new probabilistic model, one can then use the proposed approach to quickly implement, in an automatic manner, an inference algorithm based on the local minimization of \(\alpha\)-divergences. For this one only needs to write code that evaluates the likelihood factors \(f_1, \ldots, f_N\) in (9). After this, the most difficult task is the computation of the gradients of (9) so that stochastic gradient descent with minibatches can be used to optimize the energy function. However, the computation of these gradients can be easily automated by using automatic differentiation tools such as autograd (http://github.com/HIPS/autograd) or theano [9]. This approach allows us to quickly implement and test different modeling assumptions when solving applied problems.
4 Experiments

We perform experiments with a Bayesian probit regression model to validate the black-box approach described in Section 3.2. We call this method black-box alpha (BB-α). We compare BB-α with a method that uses the analytic solution to the integrals, which is possible in the probit model. This method optimizes the energy function given by (8) and it is called white-box alpha (WB-α). We also compare with a method that uses the analytic solution to the integrals, which is possible in the probit model. They also perform similarly among them, with BB-α and BB-VB are implemented by drawing K Monte Carlo samples for each data point. We plan to design automatic differentiation methods allows our technique to be applied in a straightforward manner to arbitrary probabilistic models with complex likelihood factors. Experiments with probit regression models illustrate the accuracy of the proposed approach. Experiments with neural networks show that one can obtain gains by tuning the value of α to the data. The proposed approach can be easily extended to consider a different value of α for each likelihood factor or data point. We plan to design methods for optimally choosing these α values.
Acknowledgements

YL thanks the Schlumberger Foundation Faculty for the Future fellowship on supporting her PhD study. JMHL acknowledges support from the Rafael del Pino Foundation. RET thanks EPSRC grant #EP/L000776/1 and #EP/M026957/1. TB thanks Google for funding his European Doctoral Fellowship. DHL and JMHL acknowledge support from Plan Nacional I+D+i, Grant TIN2013-42351-P, and from CAM, Grant S2013/ICE-2845 CASI-CAM-CM.

References


