Stochastic Variational Inference

David M. Blei
Princeton University

(DRAFT: DO NOT CITE)

December 8, 2011

We derive a stochastic optimization algorithm for mean field variational inference, which we call *online variational inference*. Our algorithm approximates the posterior distribution of a probabilistic model with hidden variables, and can handle large (or even streaming) data sets of observations.

Let $x = x_{1:n}$ be n observations, β be *global* hidden variables, and $z = z_{1:n}$ be n *local* hidden variables. We assume that the joint distribution of the hidden variables and observations is

$$p(\beta, z_{1:n}, x_{1:n}) = p(\beta \,|\, \alpha) \prod_{i=1}^{n} p(z_i \,|\, \beta) p(x_i \,|\, z_i, \beta), \tag{1}$$

where α are fixed hyperparameters. In this model, the global variables β can govern the distributions of any of the other variables. The local variables z_i only govern the distributions of their respective observations x_i. Figure 1 illustrates this model. Our goal is to approximate the posterior $p(\beta, z \,|\, x)$.

The distinction between local and global variables will be important for us to develop online inference. In Bayesian statistics, for example, think of β as parameters with a prior and $z_{1:n}$ as hidden variables which are individual to each observation. In a Bayesian mixture of Gaussians the global variables β are the mixture components and mixture proportions; the local variables z_i are the mixture assignments for each data point.

We make the assumption of *conditional conjugacy*, which means that the model satisfies two properties. The first property is that each hidden variable's factor in Equation 2 is in an exponential family,

$$\begin{aligned} p(\beta \,|\, \alpha) &= \exp\{\alpha^\top t(\beta) - a(\alpha)\} \quad (2) \\ p(z_i \,|\, \beta) &= \exp\{\beta^\top t(z_i) - a(\beta)\}, \quad (3) \end{aligned}$$

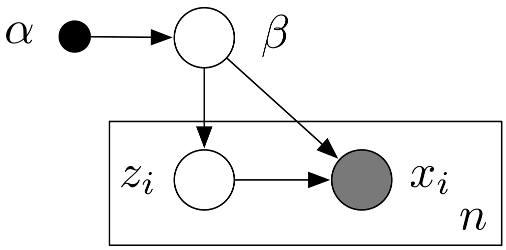

Figure 1: A graphical model with observations $x_{1:n}$, local hidden variables $z_{1:n}$ and global hidden variables β. The distribution of each observation x_i only depends on its corresponding local variable z_i and the global variables β.

where we are overloading the notation for sufficient statistics $t(\cdot)$ and log normalizer $a(\cdot)$. (These are likely different for the two families.) The second property is that each hidden variable, conditioned on all the other variables, is in the same family as the factor,

$$\begin{aligned} p(\beta \,|\, z, x) &= h(\beta)\exp\{\eta_g(z,x)^\top t(\beta) - a(\eta_g(z,x))\} && (4) \\ p(z_i \,|\, \beta, x_i) &= h(z_i)\exp\{\eta_\ell(\beta,x_i)^\top t(z_i) - a(\eta_\ell(\beta,x_i))\}. && (5) \end{aligned}$$

In these conditionals, the natural parameters are functions of the conditioning variables. For the local variables, the conditional exponential family for z_i is only a function of the global variables β and the ith data point. This follows from the factorization of the joint distribution in Equation 2.

Conditional conjugacy holds in many useful statistical models, including latent Dirichlet allocation, hierarchical Dirichlet processes, switching Kalman filters, hierarchical HMMs, Bayesian mixture models, factorial models, and the probabilistic forms of various matrix factorization models. It is the principal assumption made in many automated inference engines, such as VIBES (Bishop et al., 2003; Winn and Bishop, 2005).

The evidence lower bound. In variational inference we optimize the evidence lower bound (ELBO), which can be derived with Jensen's inequality,

$$\begin{aligned} \log p(x) &= \log \int_{\beta,z} p(\beta,z,x) && (6) \\ &= \log \int_{\beta,z} p(\beta,z,x)\frac{q(\beta,z)}{q(\beta,z)} && (7) \\ &= \log\left(\mathrm{E}_q\left[\frac{p(\beta,Z,x)}{q(\beta,Z)}\right]\right) && (8) \\ &\geq \mathrm{E}_q[\log p(\beta,Z,x)] - \mathrm{E}_q[\log q(\beta,Z)] && (9) \\ &= \mathscr{L}(q). && (10) \end{aligned}$$

This bound uses a family of distributions over the hidden variables $q(\beta,z)$. In variational inference, we restrict q to be in a tractable family, i.e., where the expectations can be taken, and we try find the member that optimizes the bound. This is equivalent to finding the member of the family that is closest in KL divergence to the posterior of interest (Jordan et al., 1999; Wainwright and Jordan, 2008).

The mean-field variational family. We specify the mean-field family for $q(\beta,z)$. In this family, each hidden variable is assumed independent and governed by its own variational parameter

$$q(\beta,z) = q(\beta\,|\,\lambda)\prod_{i=1}^{n} q(z_i\,|\,\phi_i). \tag{11}$$

The parameters λ are the global variational parameters; the parameters ϕ are the local variational parameters. The objective is now a function of the variational parameters.

Note that this family is less limiting than it might seem. Groups of hidden variables can be considered as a single "variable," allowing for various dependencies, such as time-series or spatial structure in the global variables. However, the individual groups of hidden variables are assumed to be fully factored.

The mean-field assumption will lead to many computational conveniences. At the outset, the second term of the objective in Equation 9 decomposes,

$$\mathrm{E}[\log q(\beta,z)] = \mathrm{E}_{\lambda}[\log q(\beta\,|\,\lambda)] + \sum_{i=1}^{n} \mathrm{E}_{\phi_i}[\log q(Z_i\,|\,\phi_i)], \tag{12}$$

where E_{ϕ_i} denotes an expectation with respect to $q(z_i\,|\,\phi_i)$ and similarly for E_{λ}.

The gradient of the ELBO. Our goal is to optimize the objective with respect to the variational parameters. We begin by examining the gradient.

We take the derivative of $\mathscr{L}$ with respect to each variational parameter. Consider the parameter for the variational distribution of β $q(\beta\,|\,\lambda)$. As a function of λ, we can rewrite the objective as

$$\mathscr{L}(\lambda) = \mathrm{E}[\log p(\beta\,|\,Z_{1:n},x)] - \mathrm{E}[\log q(\beta)] + \text{const}. \tag{13}$$

To see this, consider the two terms of the objective. For the entropy term (the second term of Equation 9), only $\mathrm{E}[\log q(\beta)]$ depends on λ; the other terms are absorbed in the constant. For the expected log of the joint (the first term of Equation 9), we use the chain rule

$$\mathrm{E}[\log p(\beta,Z,x)] = \mathrm{E}[\log p(Z,x)] + \mathrm{E}[\log p(\beta\,|\,Z,x)]. \tag{14}$$

The term $\mathrm{E}[\log p(Z,x)]$ is absorbed in the constant. It does not depend on $q(\beta\,|\,\lambda)$ because it is an expectation over the other variables and, in the mean-field family, β is independent of all the other variables. This completes our derivation of Equation 13.

Before we can compute the gradient, we must specify the form of each component of the factored variational family. We now specify each component to be in the same exponential family as its corresponding conditional, and that each variational parameter is its natural parameter,

$$\begin{aligned} q(\beta \mid \lambda) &= h(\beta)\exp\{\lambda^\top t(\beta) - a(\lambda)\} && (15) \\ q(z_i \mid \phi_i) &= h(z_i)\exp\{\phi_i^\top t(z_i) - a(\phi_i)\}. && (16) \end{aligned}$$

Again, we overload the sufficient statistics and log normalizers. Assuming that these exponential families are the same as their corresponding conditionals means that $t(\cdot)$ and $a(\cdot)$ in Equation 15 are the same functions as $t(\cdot)$ and $a(\cdot)$ in Equation 4. Symmetrically, $t(\cdot)$ and $a(\cdot)$ in Equation 16 are the same as in Equation 5.

With these assumptions we can obtain the final expression for the ELBO as a function of λ, the variational parameter of the global hidden variable β,

$$\mathscr{L}(\lambda) = \mathrm{E}[\eta(z,x)]^\top a'(\beta) - \lambda^\top a'(\lambda) + a(\lambda) + \text{const}, \quad (17)$$

where we used $a'(\lambda) = \mathrm{E}_\lambda[\beta]$. Note we have now absorbed $E[a(z,x)]$ into the constant. It does not depend on λ.

With the ELBO thus simplified, the gradient of $\mathscr{L}$ with respect to λ is

$$\nabla_\lambda \mathscr{L} = \nabla_\lambda^2 a(\lambda)(\mathrm{E}_\phi[\eta(Z,x)] - \lambda). \quad (18)$$

The derivative for each local variable's variational parameter ϕ_i is nearly identical. The difference is that the natural parameter of the conditional distribution only depends on the global variables β and the ith data point x_i,

$$\nabla_{\phi_i} \mathscr{L} = \nabla_{\phi_i}^2 a(\phi_i)(\mathrm{E}_\lambda[\eta(\beta, x_i)] - \phi_i). \quad (19)$$

Coordinate ascent variational inference. In most applications of mean-field variational inference, optimization proceeds by coordinate ascent.

Returning to the global variational parameter λ, its derivative equals zero when

$$\lambda = \mathrm{E}_\phi[\eta(Z,x)]. \quad (20)$$

Updating λ with this equation, holding all the other variational parameters fixed, optimizes the ELBO for λ. Notice the mean-field assumption is critical. The term $\mathrm{E}_\phi[\eta(Z,x)]$ does not depend on λ because it is an expectation of a function of the other random variables, and the mean-field assumption asserts them to be independent of β.

The derivative with respect to the local variational parameter ϕ_i equals zero when

$$\phi_i = \mathrm{E}_\lambda[\eta(\beta, x_i)]. \quad (21)$$

Mirroring the global case, notice that this expectation is only a function of the global variational parameters λ.

```
1: Initialize $\lambda^{(0)}$ randomly.
2: repeat
3:   for each data point do
4:     Update the local variational parameters, $\phi_i^{(t)} = \mathrm{E}_{\lambda^{(t-1)}}[\eta_\ell(\beta, x_i)]$.
5:   end for
6:   Update the global variational parameters, $\lambda^{(t)} = \mathrm{E}_{\phi^{(t)}}[\eta_g(Z_{1:n}, x_{1:n})]$.
7: until the ELBO converges
```

Figure 2: Coordinate ascent mean-field variational inference.

.

These updates form the coordinate ascent variational inference algorithm (see Figure 2). This algorithm is guaranteed to find a local optimum of the ELBO. It is the “classical” variational Bayes algorithm, used in many settings.

Notice that steps 3 and 4 are trivially parallelizable using a map-reduce structure. The data can be distributed across many machines and the local variational updates can be implemented in parallel. These results can then be aggregated in step 6 to find the new global variational parameters.

However, steps 3 and 4 also reveal an inefficiency in the algorithm. The algorithm begins by initializing λ randomly, where the first value of λ does not reflect any regularity in the data. However, before completing even one iteration we must analyze every data point in step 4 using these initial (random) values. This is wasteful, especially if we expect that we can learn something about the global variational parameters from only a subset of the data. Further, if the data are “infinite”, i.e., if they represent a data source where information arrives in a constant stream, then this algorithm can never complete even one iteration.

We will see that stochastic optimization of the variational objective function solves this problem. With stochastic optimization we can handle massive and streaming data sets, making progress immediately with the global variational parameters.

The efficiency of our stochastic optimization algorithm hinges on using the *natural gradient* of the variational objective. We next discuss natural gradients in general, and their role in mean-field variational inference.

The natural gradient of the ELBO. Amari (1998) discusses the natural gradient for optimization, where the natural gradient uses a Riemannian metric to better find the direction of steepest descent. In this section we describe Riemannian metrics for probability distributions and the natural gradient of the ELBO.

Consider p-vector parameters λ and $\lambda + d\lambda$. The squared length of $d\lambda$, in Euclidean space, is simply $|d\lambda|^2 = \sum_{i=1}^{p}(d\lambda_i)^2$. However, for many kinds of parameters Euclidean space is not appropriate. In general the squared length is

$$|d\lambda|^2 = \sum_{i,j} g_{ij}(\lambda) d\lambda_i d\lambda_j. \tag{22}$$

The matrix $G = (g_{ij})$ is a Riemannian metric. It depends (in general) on the parameter λ.

Intuitively, the Riemannian metric accounts for how Euclidean distance might not be appropriate for the space by stretching and shrinking the length in Equation 22. In variational inference we focus on probability distributions, where λ is the variational parameter. In this setting, we might consider the "length" between λ and $\lambda + d\lambda$ as the change in symmetrized KL divergence between the corresponding distributions. For some families of distributions, a large change in the parameter might lead to a small change in KL divergence; in this case, Equation 22 would reveal that $|d\lambda|^2$ is smaller than it would be in Euclidean space. Similarly, a small change might lead to a large change in KL divergence; in this case, Equation 22 would reveal that $|d\lambda|^2$ is larger than it would be in Euclidean space.

A Riemannian metric for the parameter of a probability distribution is the Fisher information (Amari, 1982; Kullback and Leibler, 1951),

$$G(\lambda) = \mathrm{E}\left[(\nabla_\lambda \log p(\beta \,|\, \lambda))(\nabla_\lambda \log p(\beta \,|\, \lambda))^\top \right]. \tag{23}$$

We can further simplify when $q(\beta \,|\, \lambda)$ is in the exponential family (Equation 15). In that setting, the metric is the second derivative of the log normalizer,

$$G(\lambda) = \nabla_\lambda^2 a(\lambda). \tag{24}$$

When optimizing an objective function—as we are in variational inference—the Riemannian metric is used to compute the *natural gradient*. Specifically, we obtain the natural gradient by premultiplying the usual gradient by the inverse of a Riemannian metric. Amari (1998) showed that the natural gradient is the steepest descent direction.

Returning to variational inference, consider a global variational parameter λ. The gradient is in Equation 18. Since this is a parameter to a probability distribution, a Riemannian metric is $\nabla_\lambda^2 a(\lambda)$, and note this is the first term in Equation 18. Thus, when we premultiply the gradient by the inverse of the metric we obtain a simple natural gradient,

$$\hat{\nabla}_\lambda \mathscr{L} = \mathrm{E}_\phi[\eta(Z, x)] - \lambda. \tag{25}$$

An analogous computation goes through for the local variational parameters. Researchers have used the natural gradient in variational inference for nonlinear state space models (Honkela et al., 2008) and Bayesian mixtures (Sato, 2001).[1]

The natural gradient of the ELBO opens the door to efficient gradient-based algorithms for variational inference. It is easier to compute than the classical gradient because there is

[1] Our work here—using the natural gradient in a stochastic optimization algorithm—is closest to Sato (2001), though we develop the algorithm via a different path and Sato (2001) does not address local variational parameters.

no need to premultiply by the Fisher information matrix, which can be a limiting factor for variational parameters with many components. (In the subsequent sections we will look at parameters with tens of thousands of components.) Note that the classical coordinate ascent algorithm of Figure 2 is not gradient-based; iteratively optimizing with respect to each parameter does not require computing the Fisher information because it directly zeros the second term of Equation 18.

Stochastic optimization with the natural gradient. Stochastic gradient ascent optimizes an objective function by following noisy estimates of the gradient with a decreasing step size. In their seminal paper from 1951, Robbins and Monro showed that, under certain conditions, stochastic optimization will converge to the true optimum (or, in our case, a local optimum). Noisy estimates of a gradient are often cheaper to obtain than the true gradient, and following noisy estimates of the gradient tends to find better local optima in complex objective functions. See Spall (2003) for a good overview of stochastic optimization. See Bottou (2003) for an overview of its role in machine learning.

We first review the ideas behind stochastic optimization. Suppose we are trying to optimize the objective $f(\lambda)$ and we can sample a random variable $G(\lambda)$ that has expectation equal to the derivative, $\mathrm{E}[G(\lambda)] = f'(\lambda)$. Stochastic optimization iterates an estimate of λ with

$$\lambda^t = \lambda^{t-1} + \epsilon_t g_t(\lambda^{t-1}), \tag{26}$$

where g_t are independent draws of the noisy gradient $G(\lambda)$. If the sequence of step sizes satisfies

$$\begin{aligned} \textstyle\sum \epsilon_t &= \infty && (27) \\ \textstyle\sum \epsilon_t^2 &< \infty && (28) \end{aligned}$$

then λ^t will converge to the optimal λ^* (if f is convex) or a local optimum of f (if f is not convex).

Stochastic variational inference uses stochastic gradient ascent to optimize the ELBO with respect to the global variational parameters. We obtain noisy estimates of the gradient by subsampling the data. This leads to large computational gains because of the simple form of the natural gradient in Equation 25.

The objective function is the ELBO in Equation 9. We decompose it using the grouping of the variables (and corresponding variational parameters) into local and global variables,

$$\mathscr{L} = \mathrm{E}[\log p(\beta)] - \mathrm{E}[\log q(\beta)] + \left(\sum_{i=1}^{n} \mathrm{E}[\log p(z_i \mid \beta) + \mathrm{E}[\log p(x_i \mid z_i, \beta)] - \mathrm{E}[\log q(z_i)] \right). \tag{29}$$

Consider a uniform random variable over the indices of the data set, $I \sim \mathrm{Unif}(1, \ldots, n)$. Define $\mathscr{L}_I$ to be the following (random) function of the variational parameters,

$$\mathscr{L}_I = \mathrm{E}[\log p(\beta)] - \mathrm{E}[\log q(\beta)] + n \left(\mathrm{E}[\log p(z_I \mid \beta) + \mathrm{E}[\log p(x_I \mid z_I, \beta)] - \mathrm{E}[\log q(z_I)] \right). \tag{30}$$

The expectation of $\mathscr{L}_I$ is equal to the ELBO in Equation 29. Therefore, the natural gradient of $\mathscr{L}_I$ with respect to each global variational parameter λ_j is a noisy estimate of the natural gradient of the variational objective. This process—sampling a data point and then computing the natural gradient of $\mathscr{L}_I$—will provide the noisy gradients needed to use stochastic optimimization in variational inference.

We now compute the noisy gradient. Suppose we have sampled the ith data point. Notice that Equation 30 is equivalent to the full ELBO of Equation 29 where the ith data point is observed n times. This means that we can find the natural gradient in Equation 25, computing $\mathrm{E}[\eta(\beta_{-j}, z, x)]$ for n replicates of z_i and x_i.

To proceed, we need to develop the conditional distribution $p(\beta \mid z, x)$ in more detail. The conjugacy assumptions of Equations (1)–(4) determine its form. These assumptions mean that the conditional distribution of each local variable and observation given the global variable is

$$p(z_i, x_i \mid \beta) = \exp\{\beta^\top f(z_i, x_i) - a(\beta)\}. \tag{31}$$

Because we assume that all pairs form conjugate pairs, this means that the prior distribution of the global variables β has sufficient statistics $t(\beta) = \langle \beta, -a(\beta) \rangle$. Denote the natural parameter $\alpha = \langle \alpha_1, \alpha_2 \rangle$. (Note that α_1 might be a vector, but α_2 is a scalar.)

The conditional distribution of β is

$$p(\beta \mid x, z) \propto \exp\left\{\alpha^\top t(\beta) + \left(\sum_{i=1}^n \beta^\top f(z_i, x_i) - a(\beta)\right)\right\}. \tag{32}$$

This is in the same exponential family as the prior on β and has natural parameter

$$\eta(z, x) = \langle \alpha_1 + \sum_{i=1}^n f(z_i, x_i), \alpha_2 + n \rangle. \tag{33}$$

This is an application of the Bayesian theory around conjucacy (Bernardo and Smith, 1994).

With this form in hand, we compute the full natural gradient of Equation 25,

$$\hat{\nabla}_\lambda \mathscr{L} = \langle \alpha_1 + \sum_{i=1}^n \mathrm{E}_\phi[f(Z_i, x_i)], \alpha_2 + n \rangle - \lambda. \tag{34}$$

Finally, when the ELBO contains just one data point that is replicated n times, the (noisy) natural gradient is

$$\hat{\nabla}_\lambda \mathscr{L} = \langle \alpha_1 + n\mathrm{E}_{\phi_i}[f(Z_i, x_i)], \alpha_2 + n \rangle - \lambda. \tag{35}$$

This reveals a computational advantage to using noisy gradients. The expectation in Equation 34 uses the local variational parameters for the whole data set. In the noisy natural gradient of Equation 35, we need only consider the local parameter for one data point ϕ_i to take a step in the (expected) right direction.

Define $\eta_n(z_i, x_i)$ to be the conditional natural parameter for the global variable β given a data set of n replicates of x_i,

$$\eta_n(z_i, x_i) = \langle \alpha_1 + n f(z_i, x_i)], \alpha_2 + n \rangle. \tag{36}$$

```
Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule $\epsilon_t$ appropriately.
repeat
    Sample a data point $x_t$ uniformly from the data set.
    Compute its local variational paramater,
        $\phi = \mathrm{E}_{\lambda^{(t-1)}}[\eta(\beta, x_t)]$.
    Compute "fake" global parameters as though $x_t$ is replicated $n$ times,
        $\hat{\lambda} = \mathrm{E}_{\phi}[\eta_n(Z_t, x_t)]$.
    Update the current estimate of the global variational parameters,
        $\lambda^{(t)} = (1 - \epsilon_t)\lambda^{(t-1)} + \epsilon_t \hat{\lambda}$.
until forever
```

Figure 3: Stochastic variational inference.

.

Using the Robbins-Monro algorithm, we update the global variational parameter with

$$\begin{aligned} \lambda^t &= \lambda^{(t-1)} + \epsilon_t(\mathrm{E}\left[\eta_n(z_I, x_I)\right] - \lambda^{(t-1)}) && (37) \\ &= (1 - \epsilon_t)\lambda^{(t-1)} + \epsilon_t \mathrm{E}[\eta_n(z_I, x_I)]. && (38) \end{aligned}$$

Figure 3 presents the full algorithm.

This algorithm is elegant. At each iteration, we have a current estimate of the global variational parameter $\lambda^{(t-1)}$. We sample a single data point from our data set. We compute the optimal global variational parameter as though we observed that data point n times. Finally, we set the new estimate of the global variational parameter to be a weighted average of the previous estimate and the single-data-point optimal. If ϵ_t satisfies the conditions of Robbins and Monro (1951) then this will converge to a local optimum of the ELBO.

References

Amari, S. (1982). Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Bernardo, J. and Smith, A. (1994). *Bayesian theory*. John Wiley & Sons Ltd., Chichester.

Bishop, C., Spiegelhalter, D., and Winn, J. (2003). VIBES: A variational inference engine for Bayesian networks. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 777–784. MIT Press, Cambridge, MA.

Bottou, L. (2003). Stochastic learning. In *Advanced lectures on machine learning*, pages 146–168. Springer.

Honkela, A., Tornio, M., Raiko, T., and Karhunen, J. (2008). Natural conjugate gradient in variational inference. In *Neural Information Processing*, pages 305–314. Springer.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Sato, M. (2001). Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681.

Spall, J. (2003). *Introduction to stochastic search and optimization: Estimation, simulation, and control*. John Wiley and Sons.

Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Winn, J. and Bishop, C. (2005). Variational message passing. *Journal of Machine Learning Research*, 6:661–694.