Adaptively Setting the Learning Rate in Stochastic Variational Inference

Rajesh Ranganath
Department of Computer Science
Princeton University
rajeshr@cs.princeton.edu

Chong Wang
Machine Learning Department
Carnegie Mellon University
chongw@cs.cmu.edu

David M. Blei
Department of Computer Science
Princeton University
blei@cs.princeton.edu

Eric P. Xing
Machine Learning Department
Carnegie Mellon University
epxing@cs.cmu.edu

Abstract

Stochastic variational inference is a promising method for fitting large-scale probabilistic models with hidden structures. Different from traditional stochastic learning, stochastic variational inference uses the natural gradient, which is particularly efficient for computing probabilistic distributions. One of the issues in stochastic variational inference is to set an appropriate learning rate. Inspired by a recent approach for setting the learning rate for stochastic learning (Schaul et al., 2012), we present a strategy for setting the learning rate for stochastic variational inference and demonstrate it is effective in learning large-scale complex models.

1 Introduction

Complex probabilistic models have become a mainstay in many fields, such as text analysis, computer vision, network modeling, computational biology, and quantitative neuroscience. In each of these fields, researchers use probabilistic models with hidden variables to encode assumptions about the data. Given the model and the data, researchers seek to compute the posterior distribution to understand the hidden structure and make predictions. In general, however, computing the posterior is intractable.

Approximate methods have become an important class of tools for these models. One class of approximate methods, variational methods (Wainwright and Jordan, 2008), recast the posterior inference problem as an optimization problem by constructing an approximating class of distributions and finding the distribution in the class that is the closest. A common class of approximating distributions, the so called mean-field family, consists of the distributions that are fully factored with respect to the hidden variables of the model.

Classical optimization methods for mean-field variational inference can be slow on large datasets because they iterate over every data example before updating some shared parameters. Recently, work on stochastic variational inference has addressed this issue by moving the parameters in the direction of the noisy natural gradient based on only a couple of examples (Hoffman et al., 2012). One problem with this method is that the amount moved, the learning rate, must be specified manually. If it is too large, the parameters can bounce quite a bit and if it is too small, convergence is slow. We propose a way to get rid of this issue by setting the learning rate automatically.

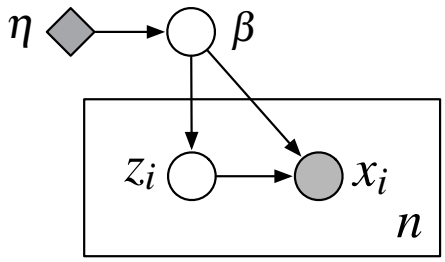

Figure 1: Graphical model for hierarchical Bayesian models with global hidden variables β, local hidden and observed variables z_i and x_i, $i = 1, \ldots, n$. Hyperparameter η is fixed, not a random variable.

2 Our Approach

ELBO The evidence lower bound (ELBO) is a lower bound on the log probability of the observations. Maximizing this bound is equivalent to minimizing the Kullback-Leibler divergence of an approximating distribution to the posterior. We consider a model depicted in Figure 1, assume it is in the exponential family and satisfies conditional conjugacy where $p(x_i, z_i \mid \beta)$ is in the exponential family and $p(\beta \mid \eta)$ is the conjugate prior (Hoffman et al., 2012). Many models of interest fall into this class, such as mixture of Gaussians, Latent Dirichlet Allocation, and probabilistic matrix factorization. The form of these distributions is

$$p(x_i, z_i \mid \beta) = h(x_i, z_i) \exp\left\{\beta^\top t(x_i, z_i) - a(\beta)\right\}, \tag{1}$$

$$p(\beta \mid \eta) = h(\beta) \exp\left\{\eta^\top t(\beta) - a(\eta)\right\}, \tag{2}$$

where we overload the notation for the base measures $h(\cdot)$, sufficient statistics $t(\cdot)$, and log normalizers $a(\cdot)$. (These will often be different for the two families.) Under the mean field assumption the variational approximation is

$$q(z_{1:n}, \beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} q(z_n \mid \phi_n). \tag{3}$$

λ and ϕ_n are the free parameters for the mean field approximating family. The goal of variational inference is to maximize the ELBO with respect to the free paramters of the approximation. We assume that each variational distribution comes from the same family as the conditional. Writing the ELBO in terms of the free parameters of the approximation gives

$$\mathcal{L}(\lambda, \phi) = \sum_{i=1}^{n} \mathrm{E}_q[\log p(x_i, z_i \mid \beta)] + \mathrm{E}_q[\log p(\beta \mid \eta)] - \mathrm{E}_q[\log q(\beta \mid \lambda)] - \sum_{i=1}^{n} \mathrm{E}_q[\log q(z_i \mid \phi_i)]. \tag{4}$$

We can write the ELBO in terms of only the global parameter λ by for any λ choosing the ϕ that maximizes the ELBO,

$$\mathcal{L}(\lambda) \triangleq \max_{\phi} \mathcal{L}(\lambda, \phi). \tag{5}$$

An optimum of this one parameter ELBO is an optimum of the full two parameter ELBO since if λ is an optimal λ, then ϕ is also optimal since it is always set to be the maximizing ϕ for a given λ. This means that if we can optimize $\mathcal{L}(\lambda)$, then we can optimize $\mathcal{L}(\lambda, \phi)$.

To optimize $\mathcal{L}(\lambda)$ we calculate its gradient. We can compute the gradient of $\mathcal{L}(\lambda)$ by first finding the corresponding optimal local parameters $\phi(\lambda)$ and then computing the (natural) gradient of $\mathcal{L}(\lambda, \phi(\lambda))$, holding $\phi(\lambda)$ fixed. The reason is that the gradient of $\mathcal{L}(\lambda)$ with respect to λ is the same as the gradient of the two-parameter ELBO $\mathcal{L}(\lambda, \phi(\lambda))$,

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla_\lambda \mathcal{L}(\lambda, \phi(\lambda)) + (\nabla_\lambda \phi(\lambda))^\top \nabla_\phi \mathcal{L}(\lambda, \phi(\lambda)) \tag{6}$$

$$= \nabla_\lambda \mathcal{L}(\lambda, \phi(\lambda)), \tag{7}$$

where $\nabla_\lambda \phi(\lambda)$ is the Jacobian of $\phi(\lambda)$ and we use the fact that the gradient of $\mathcal{L}(\lambda, \phi)$ with respect to ϕ is zero at $\phi(\lambda)$. Given the definitions of the probability distributions in Eq. 1 and Eq. 2, we can write the ELBO using the optimal $\phi(\lambda)$ as

$$\mathcal{L}(\lambda) = \eta^\top \nabla_\lambda a(\lambda) - \lambda^\top \nabla_\lambda a(\lambda) + a(\lambda) + \nabla_\lambda a(\lambda) \sum_{i=1}^{n} \bar{t}_{\phi_i}(x_i) + \text{const} \tag{8}$$

where $\bar{t}_{\phi_i}(x_i) \triangleq \mathrm{E}_{q(z_i \mid \phi_i)}[t(x_i, z_i)]$. The optimal λ is given by setting $\nabla_\lambda \mathcal{L}(\lambda) = 0$,

$$\lambda_t^* = \eta + \sum_{i=1}^{n} \bar{t}_{\phi_i}(x_i). \tag{9}$$

λ_t^* is optimal value based on $\phi(\lambda_t)$ and λ_t. Rephrasing this another way λ_t^* is the next lambda in a coordinate optimization procedure. Repeatedly setting λ_{t+1} to λ_t^* yields a coordinate ascent optimization procedure for the ELBO.

Stochastic Variational Inference Stochastic variational inference provides a way to optimize the ELBO via samples of the data. Let $I \sim \mathrm{Unif}(1, \ldots, n)$, then we can define the ELBO with respect to that sample as

$$\mathcal{L}_I(\lambda) = a(\lambda) + \nabla_\lambda a(\lambda)^\top (-\lambda + \lambda_t^I), \tag{10}$$

where we have defined the optimal λ with respect to the sample as

$$\lambda_t^I = \eta + n\bar{t}_{\phi_I}(x_I). \tag{11}$$

Generally ϕ_I the optimal ϕ for a particular λ on the sampled example is simple to calculate. Therefore λ_t^I is also simple to calculate. For example, in the mixture of gaussians models this ϕ_I represents the parameters to a variational multinomial distribution over the clusters which can be computed as a simple function of x_I and the first and second moments of the variational distriution on cluster centers.

Note that λ_t^I is a random variable with randomness coming from I, the sampling distribution over the dataset. Its expectation and variance are

$$\mathrm{E}_n[\lambda_t^I] = \lambda_t^*,$$
$$\mathrm{Var}_n[\lambda_t^I] = \mathrm{E}_n[(\lambda_t^I - \lambda_t^*)(\lambda_t^I * -\lambda_t^*)^\top] \triangleq \Sigma.$$

In general stochastic maximization proceeds by moving in the direction of the gradient, but due to the non-Euclidean nature of the parameter space for the ELBO, stochastic variational inference proceeds by moving in the optimal direction under the symmetric KL divergence metric (Hoffman et al., 2012). This is a move along the natural gradient,

$$g_t = -\lambda_t + \lambda_t^I. \tag{12}$$

The updates for stochastic variational inference with learning rate ρ_t then take the form

$$\lambda_{t+1} = \lambda_t + \rho_t g_t = \lambda_t + \rho_t(-\lambda_t + \lambda_t^I). \tag{13}$$

Adaptive learning rates One approach to formulating optimal learning rates would be to take the Taylor expansion of the ELBO around the current parameter and maximizing it with respect to the learning rate (Schaul et al., 2012). However, we found (empirically) this approach is indequate in our setting. The Taylor approximation was pretty poor when the step size is large, and the Hessian of the ELBO is not always postive definite. This led to unpredictable behaviors in this algorithm. Instead we compute adaptive learning rates by minimizing the mean squared error between the next parameter and the optimal parameter (Chien and Fu, 1967) . Let λ^* be a local optima of the variational objective, then we seek to find the learning rate that minimizes the expected squared error between the next setting of λ and the optimal λ under a positive definite metric F_t given the current setting of λ,

$$J(\rho_t) \triangleq \mathrm{E}_n[(\lambda_{t+1} - \lambda^*)^\top F_t(\lambda_{t+1} - \lambda^*) | \lambda_t]. \tag{14}$$

By adding and subtracting the per sample optima λ_t^I and substituting Equation 13 into this expression we get

$$J(\rho_t) = (1-\rho_t)^2(\lambda_t^* - \lambda_t)^\top F_t(\lambda_t^* - \lambda_t) + \rho_t^2 tr(F_t\Sigma) - 2\rho_t(\lambda_t^* - \lambda_t)^\top F_t(\lambda^* - \lambda_t^*) + const, \tag{15}$$

where const term does not depend on the learning rate. Setting the derivative of J with respect to ρ_t equal to 0 yields the optimal learning rate,

$$\rho_t = \frac{(\lambda_t^* - \lambda_t)^\top F_t(\lambda_t^* - \lambda_t) + (\lambda_t^* - \lambda_t)^\top F_t(\lambda^* - \lambda_t^*)}{(\lambda_t^* - \lambda_t)^\top F_t(\lambda_t^* - \lambda_t) + tr(F_t\Sigma)}. \tag{16}$$

This learning rate captures the intuition that if the sample optimum has high variance then the learning rate should be smaller. The second term in the numerator is a correction for the expected sample optima being the coordinate-wise optimum, λ_t^* rather than the desired optimum λ^*. Intuitively it says if the direction to the coordinate optima from the current parameter setting is the same as the direction from the coordinate optima to the final λ^*, then the learning rate should be larger. Symmetrically, if they are in opposite directions, the learning rate should be smaller.

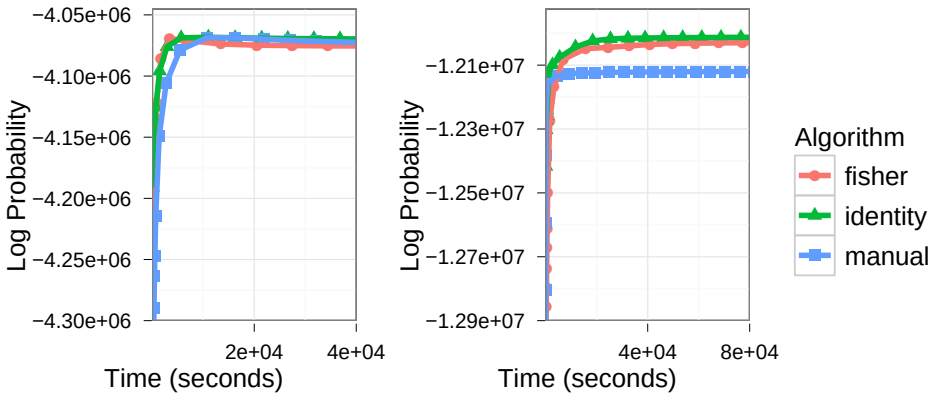

Figure 2: Held out likelihood for stochastic variational inference on LDA with adaptive learning rates and the best manual learning rate on the Nature (left panel) and Wikipedia corpora. We compute the adaptive rate with both the Fisher information and identity metrics. The adaptive learning rates do as well or better.

Choosing F Letting F be the identity matrix would be a natural choice due to computational simplicity, but we also want to account for the geometry of the parameter space. To this extent we let F_t be the Fisher information matrix at the current λ_t to account fo the geometry of the parameter space. We report results using both the Fisher information matrix at the current λ_t and the identity matrix as metrics.

Estimating the optimal learning rate While we derived an optimal learning rate, we cannot estimate it directly. We do not know the value of λ_t^* and λ^*. Since finding a good estimate for λ^* can be difficult (finding it is the goal of variational inference), we ignore the term in the optimal learning rate containing it. Ignoring the term containing λ^*, we can rewrite the optimal learning rate in terms of the noisy natural gradient as

$$\rho_t = \frac{\mathrm{E}_n[g_t]^\top \mathrm{E}_n[F_t g_t]}{\mathrm{E}_n[g_t^\top F_t g_t]}. \tag{17}$$

Similar to (Schaul et al., 2012), we approximate these expectations using exponential moving averages. Let these moving averages be denoted by $\overline{g}_t$, $\overline{\nabla}_t$, and $\overline{q}_t$ respectively, and let τ_t be the exponential window size. The updates are

$$\overline{g}_t = (1 - \tau_t^{-1})\overline{g}_{t-1} + \tau_t^{-1} g_t \tag{18}$$

$$\overline{\nabla}_t = (1 - \tau_t^{-1})\overline{\nabla}_{t-1} + \tau_t^{-1} F_t g_t \tag{19}$$

$$\overline{q}_t = (1 - \tau_t^{-1})\overline{q}_{t-1} + \tau_t^{-1} g_t^\top F_t g_t. \tag{20}$$

Using these estimates of the expectations the optimal learning rate ρ_t becomes

$$\rho_t = \frac{\overline{g}_t^\top \overline{\nabla}_t}{\overline{q}_t}. \tag{21}$$

The moving averages are less reliable after large steps, so we update our memory size using the following rule

$$\tau_{t+1} = \tau_t(1 - \rho_t) + 1. \tag{22}$$

We initialize our online estimates by computing samples at the random initialization of λ. Our estimates for the two terms in the numerator may have negative inner product. When this happens, we simply reinitialize the estimates for $\overline{g}_t$, $\overline{\nabla}_t$ with g_t and $F_t g_t$ respectively. Our estimates generalize to minibatches by noting the sample coordinate optima λ_t^I can be written as a function of the minibatch rather than just a single observation.

3 Results

In our experiments, we considered Latent Dirichlet Allocation (LDA) (Blei et al., 2003) a mixed membership model with global variables associated with the topics and local variables corresponding to the topic proportions for the

document. We ran stochastic variational inference using the best hand tuned learning rate from (Hoffman et al., 2010) against our adaptive learning rate for a batch size of 64 on two corpora. The Nature corpus and Wikipedia corpus consist of 334922 and 3.6 million documents respectively. We compared the performance of various methods by computing the likelihood on a held out test set of size 10000. Our results in Figure 2 show that the adaptive learning rate does as well as the best manual sequence found by (Hoffman et al., 2010), though the value of the metric appears limited. We hope to explore this in future work.

References

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chien, Y. and Fu, K. (1967). On Bayesian learning and stochastic approximation. *Systems Science and Cybernetics, IEEE Transactions on*, 3(1):28 –38.

Hoffman, M., Blei, D., and Bach, F. (2010). Online inference for latent Drichlet allocation. In *NIPS*.

Hoffman, M., Blei, D. M., Wang, C., and Paisley, J. (2012). Stochastic Variational Inference. *ArXiv e-prints*.

Schaul, T., Zhang, S., and LeCun, Y. (2012). No More Pesky Learning Rates. *ArXiv e-prints*.

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