An Adaptive Learning Rate for Stochastic Variational Inference
(Supplementary Information)

Rajesh Ranganath
Princeton University, 35 Olden St., Princeton, NJ 08540
RAJESHR@CS.PRINCETON.EDU

Chong Wang
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA, 15213
CHONGW@CS.CMU.EDU

David M. Blei
Princeton University, 35 Olden St., Princeton, NJ 08540
BLEI@CS.PRINCETON.EDU

Eric P. Xing
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA, 15213
EPXING@CS.CMU.EDU

Natural gradient of the \(\lambda\)-ELBO. We can compute the natural gradient in Eq. 7 at \(\lambda\) by first finding the corresponding optimal local parameters 
\(\phi^\lambda = \arg \max_\phi \mathcal{L}(\lambda, \phi)\) and then computing the gradient of \(\mathcal{L}(\lambda, \phi^\lambda)\), i.e., the ELBO where we fix \(\phi = \phi^\lambda\).

These are equivalent because

\[ \nabla_\lambda \mathcal{L}(\lambda) = \nabla_\lambda \mathcal{L}(\lambda, \phi^\lambda) + (\nabla_\lambda \phi^\lambda)^T \nabla_\phi \mathcal{L}(\lambda, \phi^\lambda) \]

\[ = \nabla_\lambda \mathcal{L}(\lambda, \phi^\lambda). \]

The notation \(\nabla_\lambda \phi^\lambda\) is the Jacobian of \(\phi^\lambda\) as a function of \(\lambda\), and we use that \(\nabla_\phi \mathcal{L}(\lambda, \phi)\) is zero at \(\phi = \phi^\lambda\).

Derivation of the adaptive learning rate. To compute the adaptive learning rate we minimize \(E_n[\mathcal{L}(\rho_t) | \lambda_t] \) at each time \(t\). Expanding \(E_n[\mathcal{L}(\rho_t) | \lambda_t] \), we get

\[ E_n[\mathcal{L}(\rho_t) | \lambda_t] = E_n[(\lambda_t + \rho_t (\lambda_t - \hat{\lambda}_t) - \hat{\lambda}_t^*)]^T (\lambda_t + \rho_t (\hat{\lambda}_t - \lambda_t) - \hat{\lambda}_t^*). \]

We can compute this expectation in terms of the moments of the sample optimum in Eq. 15

\[ E_n[\mathcal{L}(\rho_t) | \lambda_t] = (1 - \rho_t)^2 (\lambda_t^* - \lambda_t)^T (\lambda_t^* - \lambda_t) + \rho_t^2 tr(\Sigma) \]

Setting the derivative of \(E_n[\mathcal{L}(\rho_t) | \lambda_t]\) with respect to \(\rho_t\) equal to 0 yields the optimal learning in Eq. 16.

Convergence of the idealized learning rate. We show convergence of \(\lambda_t\) to a local optima with our idealized learning rate through martingale convergence. Let \(M_{t+1} = Q(a_t^*)\), then \(M_t\) is a super-martingale with respect to the natural filtration of the sequence \(\lambda_t\).

\[ E[M_{t+1} | \lambda_t] = E[Q(a_t^*) | \lambda_t] \leq E[Q(0) | \lambda_t] = M_t. \]

Since \(M_t\) is a non-negative supermartingale by the martingale convergence theorem, we know that a finite \(M_\infty\) exists and \(M_t \to M_\infty\) almost surely. Since the \(M_t\) converge, the sequence of expected values \(E[M_t]\) converge to \(E[M_\infty]\). This means that the sequence of expected values form a Cauchy sequence, so the difference between elements of the sequence goes to zero,

\[ D_t = E[M_{t+1}] - E[M_t] \]

\[ = E[E[M_{t+1} | \lambda_t] - E[M_t | \lambda_t]] \to 0. \]

Substituting the idealized optimal learning rate into this expression gives

\[ D_t = E[ -((\lambda_t^* - \lambda_t)^T (\lambda_t^* - \lambda_t) + (\lambda_t^* - \lambda_t)^T (\lambda_t^* - \lambda_t) )^2 (\lambda_t^* - \lambda_t)^T (\lambda_t^* - \lambda_t) + tr(\Sigma))^{-1}]. \]

(1)

Since the \(D_t\)’s are a sequence of nonpositive random variables whose expectation goes to zero and that the variances are bounded (by assumption), the square portion of Eq. 1 must go to zero almost surely. This quantity going to zero implies that either \(\lambda_t \to \lambda^*\) or \(\lambda_t \to \lambda_t^*\). If \(\lambda_t = \lambda_t^*\), then \(\lambda_t\) is a local optima under the assumption that the two parameter (\(\phi\) and \(\lambda\) for the ELBO) function we are optimizing can be optimized via coordinate ascent. Putting everything together gives us that \(\lambda_t\) goes to a local optima almost surely.