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

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

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

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

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

Natural gradient of the λ-ELBO. We can compute the natural gradient in Eq. 7 at λ 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

$$\begin{aligned}\nabla_\lambda \mathcal{L}(\lambda) &= \nabla_\lambda \mathcal{L}(\lambda, \phi^\lambda) + (\nabla_\lambda \phi^\lambda)^\top \nabla_\phi \mathcal{L}(\lambda, \phi^\lambda) \\ &= \nabla_\lambda \mathcal{L}(\lambda, \phi^\lambda).\end{aligned}$$

The notation $\nabla_\lambda \phi^\lambda$ is the Jacobian of ϕ^λ as a function of λ, 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 $\mathbb{E}_n[J(\rho_t)|\lambda_t]$ at each time t. Expanding $\mathbb{E}_n[J(\rho_t)|\lambda_t]$, we get

$$\begin{aligned}\mathbb{E}_n[J(\rho_t)|\lambda_t] =& \mathbb{E}_n[(\lambda_t + \rho_t(\lambda_t - \hat{\lambda}_t) - \lambda_t^*)^\top \\ & (\lambda_t + \rho_t(\lambda_t - \hat{\lambda}_t) - \lambda_t^*)].\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}_n[J(\rho_t)|\lambda_t] =& (1-\rho_t)^2(\lambda_t^* - \lambda_t)^\top(\lambda_t^* - \lambda_t) \\ & + \rho_t^2 tr(\Sigma).\end{aligned}$$

Setting the derivative of $\mathbb{E}_n[J(\rho_t)|\lambda_t]$ with respect to ρ_t equal to 0 yields the optimal learning in Eq. 16.

Convergence of the idealized learning rate. We show convergence of λ_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 λ_t,

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

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

$$\begin{aligned}D_t &\triangleq \mathbb{E}[M_{t+1}] - \mathbb{E}[M_t] \\ &= \mathbb{E}[\mathbb{E}[M_{t+1}|\lambda_t] - \mathbb{E}[M_t|\lambda_t]] \to 0.\end{aligned}$$

Substituting the idealized optimal learning rate into this expression gives

$$\begin{aligned}D_t =& \mathbb{E}[-((\lambda_t^* - \lambda_t)^\top(\lambda_t^* - \lambda_t) + (\lambda_t^* - \lambda_t)^\top(\lambda^* - \lambda_t^*))^2 \\ & ((\lambda_t^* - \lambda_t)^\top(\lambda_t^* - \lambda_t) + tr(\Sigma))^{-1}]. \qquad (1)\end{aligned}$$

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 λ_t is a local optima under the assumption that the two parameter (ϕ and λ for the ELBO) function we are optimizing can be optimized via coordinate ascent. Putting everything together gives us that λ_t goes to a local optima almost surely.