Supplementary Material
A Variational Analysis of Stochastic Gradient Algorithms

Stephan Mandt SM3976@COLUMBIA.EDU
Columbia University, Data Science Institute, New York, USA

Matthew D. Hoffman MATHOFFM@ADOBE.COM
Adobe Research, San Francisco, USA

David M. Blei DAVID.BLEI@COLUMBIA.EDU
Columbia University, Departments of CS and Statistics, New York, USA

A. Stationary Covariance

The Ornstein-Uhlenbeck process has an analytic solution in terms of the stochastic integral (Gardiner et al., 1985),

$$\theta(t) = \exp(-At)\theta(0) + \sqrt{\tfrac{\epsilon}{S}} \int_0^t \exp[-A(t-t')]B dW(t') \quad (1)$$

Following Gardiner's book we derive an algebraic relation for the stationary covariance of the multivariate Ornstein-Uhlenbeck process. Define $\Sigma = \mathbb{E}[\theta(t)\theta(t)^\top]$. Using the formal solution for $\theta(t)$ given in the main paper, we find

$$\begin{aligned} A\Sigma + \Sigma A^\top &= \tfrac{\epsilon}{S} \int_{-\infty}^t A \exp[-A(t-t')]BB^\top \exp[-A^\top(t-t')]dt' \\ &+ \tfrac{\epsilon}{S} \int_{-\infty}^t \exp[-A(t-t')]BB^\top \exp[-A^\top(t-t')]dt' A^\top \\ &= \tfrac{\epsilon}{S} \int_{-\infty}^t \frac{d}{dt'} \Big(\exp[-A(t-t')]BB^\top \exp[-A^\top(t-t')]\Big) \\ &= \tfrac{\epsilon}{S} BB^\top. \end{aligned}$$

We used that the lower limit of the integral vanishes by the positivity of the eigenvalues of A.

B. Stochastic Gradient Fisher Scoring

We start from the Ornstein-Uhlenbeck process

$$\begin{aligned} d\Theta(t) &= -HA\theta(t)dt + H\left[B_{\epsilon/S} + E\right] dW(t) \\ &= -A'\theta(t)dt + B' dW(t). \end{aligned} \quad (2)$$

We defined $A' \equiv HA$ and $B' \equiv H\left[B_{\epsilon/S} + E\right]$. As derived in the paper, the variational bound is (up to a constant)

$$KL \stackrel{c}{=} \frac{N}{2}\mathrm{Tr}(A\Sigma) - \log(|NA|). \quad (3)$$

To evaluate it, the task is to remove the unknown covariance Σ from the bound. To this end, as before, we use the identity for the stationary covariance $A'\Sigma + \Sigma A'^\top = B'B'^\top$.

The criterion for the stationary covariance is equivalent to

$$\begin{aligned} HA\Sigma + \Sigma AH &= \epsilon HBB^\top H + HEE^\top H^\top \\ \Leftrightarrow A\Sigma + H^{-1}\Sigma AH &= \epsilon BB^\top H + EE^\top H \\ \Rightarrow \mathrm{Tr}(A\Sigma) &= \frac{1}{2}\mathrm{Tr}(H(\epsilon BB^\top + EE^\top)) \end{aligned} \quad (4)$$

We can re-parametrize the covariance as $\Sigma = TH$, such that T is now unknown. The KL divergence is therefore

$$\begin{aligned} KL &= -\frac{N}{2}\mathrm{Tr}(A\Sigma) + \log(|NA|) \\ &= \frac{N}{4}\mathrm{Tr}(H(\epsilon BB^\top + EE^\top)) + \frac{1}{2}\log|T| \\ &\quad + \frac{1}{2}\log|H| + \frac{1}{2}\log|NA| + \frac{D}{2}, \end{aligned} \quad (5)$$

which is the result we give in the main paper.

C. Square root preconditioning

Finally, we analyze the case where we precondition with a matrix that is proportional to the square root of the diagonal entries of the noise covariance.

We define

$$G = \sqrt{\mathrm{diag}(BB^\top)} \quad (6)$$

as the diagonal matrix that contains square roots of the diagonal elements of the noise covariance. We use an additional scalar learning rate ϵ .

Theorem (taking square roots). *Consider SGD preconditioned with G^{-1} as defined above. Under the previous assumptions, the constant learning rate which minimizes KL divergence between the stationary distribution of this process and the posterior is*

$$\epsilon^* = \frac{2DS}{N\mathrm{Tr}(BB^\top G^{-1})}. \quad (7)$$

For the proof, we read off the appropriate KL divergence from the proof of Theorem 2 with $G^{-1} \equiv H$:

$$KL(q||f) \stackrel{c}{=} \frac{\epsilon N}{2S} \mathrm{Tr}(BB^\top G^{-1}) - \mathrm{Tr}\log(G) + \frac{D}{2}\log\frac{\epsilon}{S} - \frac{1}{2}\log|\Sigma| \tag{8}$$

Minimizing this KL divergence over the learning rate ϵ yields Eq. 7 $\square$.