Stochastic Natural Gradient Descent Convergence Analysis

Dhruv Malik

March 1st 2018

Martin's text gives a definition of variance for symmetric matrix valued random variables. But he doesn't do the non-symmetric case. I couldn't find any defintion online either. In the same vein as Martin's text, I defined the variance of a random (symmetric or non-symmetric) matrix M as follows:

$$Var(M) = \mathbb{E}\Big[(M - \mathbb{E}(M))(M - \mathbb{E}(M))^T\Big]$$

This is necessary to do because we have a matrix valued random variable for our natural gradient, which is not symmetric. Note that my definition is consistent with Martin's. Also note that if I switched the transpose from the second term to the first (for an alternative definition) then my proof would still work - in fact this would be nicer.

Lemma 1. Variance Expansion.

$$\begin{aligned}\mathbb{E}\Big[(M - \mathbb{E}(M))(M - \mathbb{E}(M))^T\Big] &= \mathbb{E}\Big[MM^T - \mathbb{E}(M)M^T - M\mathbb{E}(M)^T - \mathbb{E}(M)\mathbb{E}(M)^T\Big] \\ &= \mathbb{E}(MM^T) - \mathbb{E}(M)\mathbb{E}(M)^T - \mathbb{E}(M)\mathbb{E}(M)^T - \mathbb{E}(M)\mathbb{E}(M)^T \\ &= \mathbb{E}(MM^T) - \mathbb{E}(M)\mathbb{E}(M)^T\end{aligned}$$

Theorem 1. Stochastic Natural Gradient Descent. *Assume we have a random variable X such that $\mathbb{E}(X) = \nabla C(K)\Sigma_K^{-1} = E_K$, $Var(X) = V$ and $Var(X^T) = V'$. Then for stepsize at time t:*

$$\eta_t = \min\left\{\frac{1}{\|R\| + \frac{\|B\|^2 C(K_0)}{\mu}}, \frac{\|\Sigma_{K^*}\|(2t+1)}{\mu\sigma_{min}(R)(t+1)^2}\right\}$$

and for:

$$N > \frac{4C(K_0)Tr(V')\|\Sigma_{K^*}\|^2}{\sigma_{min}(Q)\mu^2\sigma_{min}(R)^2\epsilon}$$

natural policy gradient descent converges linearly and enjoys the following performance bound:

$$\mathbb{E}[C(K_N) - C(K^*)] < \epsilon$$

Proof. For a current policy K, the policy K' after a gradient step is $K' = K - \eta_t X$. Note that the choice of stepsize satisfies $\eta_t \leq \frac{1}{\|R+B^T P_K B\|}$. Now, by Lemma 9 in the paper:

$$\begin{aligned}
\mathbb{E}[C(K') - C(K)] &= \mathbb{E}[-2Tr(\Sigma_{K'}(K-K')^T E_K) + Tr(\Sigma_{K'}(K-K')^T(R+B^T P_K B)(K-K'))] \\
&= -2\eta_t \mathbb{E}[Tr(\Sigma_{K'} X^T E_K)] + \eta_t^2 \mathbb{E}[Tr(\Sigma_{K'} X^T (R+B^T P_K B) X)] \\
&= -2\eta_t Tr(\Sigma_{K'} E_K^T E_K) + \eta_t^2 \mathbb{E}[Tr((R+B^T P_K B) X \Sigma_{K'} X^T)] \\
&\leq -2\eta_t Tr(\Sigma_{K'} E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| \mathbb{E}[Tr(\Sigma_{K'} X^T X)] \\
&= -2\eta_t Tr(\Sigma_{K'} E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| Tr(\Sigma_{K'} \mathbb{E}[X^T X]) \\
&= -2\eta_t Tr(\Sigma_{K'} E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| Tr(\Sigma_{K'}(Var(X^T) + \mathbb{E}(X^T)\mathbb{E}(X))) \text{ by my Lemma} \\
&= -2\eta_t Tr(\Sigma_{K'} E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| Tr(\Sigma_{K'}(Var(X^T) + E_K^T E_K)) \\
&= -2\eta_t Tr(\Sigma_{K'} E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| Tr(\Sigma_{K'} Var(X^T) + \Sigma_{K'} E_K^T E_K) \\
&= -2\eta_t Tr(\Sigma_{K'} E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| Tr(\Sigma_{K'} Var(X^T)) + \eta_t^2 \left\|R+B^T P_K B\right\| Tr(\Sigma_{K'} E_K^T E \\
&\leq -2\eta_t Tr(\Sigma_{K'} E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| Tr(\Sigma_{K'} Var(X^T)) + \eta_t Tr(\Sigma_{K'} E_K^T E_K) \\
&= -\eta_t Tr(\Sigma_{K'} E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| Tr(\Sigma_{K'} Var(X^T)) \\
&\leq -\eta_t \sigma_{min}(\Sigma_{K'}) Tr(E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| \|\Sigma_{K'}\| Tr(V') \\
&\leq -\eta_t \mu Tr(E_K^T E_K) + \eta_t^2 \left\|R+B^T P_K B\right\| \frac{C(K')}{\sigma_{min}(Q)} Tr(V') \text{ by Lemma 10} \\
&\leq -\eta_t \frac{\mu \sigma_{min}(R)}{\|\Sigma_{K^*}\|}(C(K) - C(K^*)) + \eta_t^2 \left\|R+B^T P_K B\right\| \frac{C(K')}{\sigma_{min}(Q)} Tr(V') \text{ by Lemma 8}
\end{aligned}$$

We now add $C(K^*)$ on both sides of this equation, and get:

$$\mathbb{E}[C(K') - C(K^*)] \leq \left(1 - \eta_t \frac{\mu \sigma_{min}(R)}{\|\Sigma_{K^*}\|}\right)(C(K) - C(K^*)) + \eta_t \frac{C(K_0)}{\sigma_{min}(Q)} Tr(V')$$

The remainder of the proof should be identical to what is shown in the Mark Schmidt paper. The fact that I have a min in my choice for stepsize shouldn't affect the proof, because for the timesteps where the stepsize takes the first value we can pretend like we used the second value and get a looser upper bound. Another idea is to shift the values of t with a large C value such that second term is always smaller than the first term and we don't even have to take a min. SIMPLEST OF ALL: Pick a constant $C > 1$ such that the second term divided by C is smaller than first term. This eliminates the min. In the main inequality above, upper bound it to get a new main inequality where we have a C in the numerator. This cancels when multiplied by step size and we can directly apply the Mark Schmidt proof. Point is that given the inequality up above, it should be pretty simple to show linear convergence, and it would only change constants in the theorem. □