Stochastic Quasi-Newton Optimization in Large Dimensions Including Deep Network Training: Supplementary Material

1 Derivation of stochastic mimicry of inverse-Hessian matrix $\tilde{G}$

The evolution of state X_τ parameterized over time τ is governed by the following process dynamics' stochastic differential equation (SDE):

$$dX_\tau = \mathcal{R}_\tau dB_\tau \tag{1}$$

where, $\mathcal{R}_\tau$ is the diffusion coefficient, and B_τ is a standard Brownian motion.
The process dynamics is constrained by another measurement SDE given by:

$$dY_\tau = \nabla f d\tau + d\eta_\tau \tag{2}$$

where Y_τ is the zero-mean observation process, $\nabla f_{X_\tau} (:= \nabla f(X_\tau))$ is the gradient of cost function $f(X_\tau) : \mathbb{R}^N \to \mathbb{R}^+$, and η_τ is a measurement noise (independent of process noise). Let $X'_{\tau+1}$ denote the true state while $X_{\tau+1|\tau}$ represents the predicted state at $\tau+1$. Therefore,

$$Y_{\tau+1} = \nabla f_{X'_{\tau+1}} + \eta_{\tau+1} \tag{3}$$

The update obtained using the predicted state $X_{\tau+1|\tau}$ can be expressed as:

$$\hat{X}_{\tau+1} = X_{\tau+1|\tau} - \tilde{G}(Y_{\tau+1} - \nabla f_{X_{\tau+1|\tau}}) \tag{4}$$

It must be noted that on comparison with the Newton's-like update, the gain matrix $\tilde{G}$ resembles the negative Hessian-inverse (i.e. $\tilde{G} = -H^{-1}$). Therefore, the error can be formally presented as $e = X'_{\tau+1} - \hat{X}_{\tau+1}$. For simplicity, we avoid the subscript $\tau+1$ in later part of the derivation. Accordingly, the error covariance matrix P can be written as:

$$\begin{aligned}
& P = \mathbb{E}[(e-\bar{e})(e-\bar{e})^T] \\
\implies & P = \mathbb{E}[((X'-X-\overline{(X'-X)}) + \tilde{G}((\nabla f_{X'} - \nabla f_X) - \overline{(\nabla f_{X'} - \nabla f_X)}) + \tilde{G}\eta) \\
& \qquad ((X'-X-\overline{(X'-X)}) + \tilde{G}((\nabla f_{X'} - \nabla f_X) - \overline{(\nabla f_{X'} - \nabla f_X)}) + \tilde{G}\eta)^T] \\
\implies & P = \mathbb{E}[((X-\overline{X}) + \tilde{G}(\nabla f_X - \overline{\nabla f_X}) - \tilde{G}\eta)((X-\overline{X}) + \tilde{G}(\nabla f_X - \overline{\nabla f_X}) - \tilde{G}\eta)^T] \\
\implies & P = \mathbb{E}[(A+\tilde{G}B)(A+\tilde{G}B)^T] \\
\implies & P = \mathbb{E}[AA^T + AB^T\tilde{G}^T + \tilde{G}BA^T + \tilde{G}BB^T\tilde{G}^T]
\end{aligned} \tag{5}$$

where, $A = (X-\bar{X})$ and $B = (\nabla f_X - \overline{\nabla f_X} - \eta)$. Using the linearity of expectation, we can split this into separate expectations:

$$P = \mathbb{E}[AA^T] + \mathbb{E}[AB^T\tilde{G}^T] + \mathbb{E}[\tilde{G}BA^T] + \mathbb{E}[\tilde{G}BB^T\tilde{G}^T] \tag{6}$$

To obtain the expression of $\tilde{G}$, we minimize the trace of error covariance matrix with respect to $\tilde{G}$:

$$tr(P) = tr(C) + tr(D\tilde{G}^T) + tr(\tilde{G}D^T) + tr(\tilde{G}E\tilde{G}^T) \tag{7}$$

where, $C = \mathbb{E}[AA^T]$, $D = \mathbb{E}[AB^T]$ and $E = \mathbb{E}[BB^T]$. Thus, we get,

$$\frac{\partial tr(P)}{\partial \tilde{G}} = 2D + 2\tilde{G}E \tag{8}$$

Setting $\frac{\partial tr(P)}{\partial \hat{G}} = 0$ gives,

$$\tilde{G} = -DE^{-1} \tag{9}$$

So, the optimal $\tilde{G}$ that minimizes the $tr(P)$ is:

$$\tilde{G} = -\mathbb{E}[AB^T](\mathbb{E}[BB^T])^{-1} \tag{10}$$

Since, η is zero-mean and independent of X and ∇f, i.e. $\mathbb{E}[(X-\bar{X})\eta^T] = \mathbb{E}(X-\bar{X})\mathbb{E}[\eta^T] = 0$,

$$\mathbb{E}[AB^T] = \mathbb{E}[(X-\bar{X})(\nabla f_X - \overline{\nabla f_X})^T] \tag{11}$$

Accordingly,

$$\begin{aligned} \mathbb{E}[BB^T] &= \mathbb{E}[(\nabla f_X - \overline{\nabla f_X})(\nabla f_X - \overline{\nabla f_X})^T] + \mathbb{E}[\eta\eta^T] \\ \implies \mathbb{E}[BB^T] &= \mathbb{E}[(\nabla f_X - \overline{\nabla f_X})(\nabla f_X - \overline{\nabla f_X})^T] + R \end{aligned} \tag{12}$$

where, $R = \mathbb{E}[\eta\eta^T]$ represents the measurement noise covariance matrix. Therefore, the expression for $\tilde{G}$ can be expressed as:

$$\tilde{G} = -\mathbb{E}[(X-\bar{X})(\nabla f_X - \overline{\nabla f_X})^T]\left(\mathbb{E}[(\nabla f_X - \overline{\nabla f_X})(\nabla f_X - \overline{\nabla f_X})^T] + R\right)^{-1} \tag{13}$$

Thus, the explicit expression for stochastic mimicry of inverse-Hessian can be expressed as:

$$\tilde{G} = -\left[\int_\Omega \left(X-\bar{X}\right)\left(\nabla f - \overline{\nabla f}\right)^T dP\right]\left[\int_\Omega \left(\nabla f - \overline{\nabla f}\right)\left(\nabla f - \overline{\nabla f}\right)^T dP + R\right]^{-1} \tag{14}$$