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

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

The evolution of state X_t parameterized over time t is governed by the following process dynamics' stochastic differential equation (SDE):

$$dX_t = \boldsymbol{R}_t dB_t \tag{1}$$

where, $\boldsymbol{R}_t$ is the diffusion coefficient, and B_t is a standard Brownian motion.
The process dynamics is constrained by another measurement SDE given by:

$$dY_t = \nabla f dt + dW_t \tag{2}$$

where Y_t is the zero-mean observation process, $\nabla f_{X_t} (:= \nabla f(X_t))$ is the gradient of cost function $f(X_t) : \mathbb{R}^N \to \mathbb{R}^+$, and W_t is a measurement noise (independent of process noise). Let X'_{t+1} denote the true state while $X_{t+1|t}$ represents the predicted state at $t+1$. Therefore,

$$Y_{t+1} = \nabla f_{X'_{t+1}} + W_{t+1} \tag{3}$$

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

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

It must be noted that on comparison with the Newton's-like update, the gain matrix $\tilde{\boldsymbol{G}}$ resembles the negative Hessian-inverse (i.e. $\tilde{\boldsymbol{G}} = -\boldsymbol{H}^{-1}$). Therefore, the error can be formally presented as $e = X'_{t+1} - \hat{X}_{t+1}$. For simplicity, we avoid the subscript $t+1$ in later part of the derivation. Accordingly, the error covariance matrix $\boldsymbol{P}$ can be written as:

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

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

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

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

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

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

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

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

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

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

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

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

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

Accordingly,

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

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

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

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

$$\tilde{\boldsymbol{G}} = -\left[\int_\Omega \left(X - \bar{X}\right)\left(\nabla f - \overline{\nabla f}\right)^T d\mathbb{P}\right]\left[\int_\Omega \left(\nabla f - \overline{\nabla f}\right)\left(\nabla f - \overline{\nabla f}\right)^T d\mathbb{P} + \boldsymbol{Q}\right]^{-1} \tag{14}$$