

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

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

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

$$dX_t = \mathbf{R}_t dB_t \quad (1)$$

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

$$dY_t = \nabla f dt + dW_t \quad (2)$$

where  $Y_t$  is the zero-mean observation process,  $\nabla f_{X_t} (:= \nabla f(X_t))$  is the gradient of cost function  $f(X_t) : \mathbb{R}^N \rightarrow \mathbb{R}^+$ , 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,

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

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

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

It must be noted that on comparison with the Newton's-like update, the gain matrix  $\tilde{\mathbf{G}}$  resembles the negative Hessian-inverse (i.e.  $\tilde{\mathbf{G}} = -\mathbf{H}^{-1}$ ). 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  $P$  can be written as:

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

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:

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

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

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

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

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

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

$$\tilde{G} = -DE^{-1} \quad (9)$$

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

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

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

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

Accordingly,

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

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

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

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

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