Supplementary Material: Enhance Curvature Information by Structured Stochastic Quasi-Newton Methods

Minghan Yang[1,2], Dong Xu[1,2], Hongyu Chen[1], Zaiwen Wen[2,3,4], Mengyun Chen[5]
[1] School of Mathematical Sciences, Peking University, China
[2] Beijing International Center for Mathematical Research, Peking University, China
[3] Center for Data Science, Peking University, China
[4] National Engineering Laboratory for Big Data Analysis and Applications, Peking University, China
[5] Huawei Technologies Co. Ltd, China
{yangminghan, taroxd, hongyuchen, wenzw}@pku.edu.cn ,chenmengyun1@huawei.com

Implementation Details

Logistic Regression

- The objective function considered in this part is:

$$\min_{\theta\in\mathbb{R}^n} \Psi(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log(1+\exp(-y_i\langle x_i,\theta\rangle)) + \mu\|\theta\|_2^2,$$

  where $\{x_i, y_i\} \in \mathbb{R}^n \times \{-1, 1\}$, $i \in [1, 2, \ldots, N]$.

- A description of the datasets is shown in Table 1.

- We describe the implementation details of the algorithms used in this part.

  - SGD: The batch size is set to be 1.
  - L-BFGS: The source code is downloaded from the website [1] and the default parameters are used.
  - SSN: The batch size $\mathcal{S}_H$ for the subsampled Hessian matrix is $\min\{2000, \lfloor 0.01N \rfloor\}$. The batch size of the subsampled gradient $|\mathcal{S}_g|$ is changing as $\min\{|\mathcal{S}_g| \cdot 1.1, N\}$.
  - S4QN: The set up of the subsampled Hessian H_k is the same as SSN. The matrix Λ_k is generated by the stochastic L-BFGS method and the memory size is 5.

Deep Learning

We now present the detailed implementation for the deep learning problems. The batch size for all methods is the same, i.e., 512 for Autoencoders and 256 for CNNs (ConvNet and ResNet-18). The hyper-parameters of Adam for all three architectures are tuned by using the grid search as follows.

- The initial learning rate is from {3e-2, 1e-2, 3e-3, 1e-3, 3e-4, 1e-4}.
- The parameters β_1 and β_2 are tuned in {0.9,0.99} and $\{0.99, 0.999\}$, respectively.

[1] https://www.cs.ubc.ca/~schmidtm/Software/minFunc.html

Dataset	# Data points N	# Dimension n
rcv1	20, 242	47, 236
news20	19, 996	1, 355, 191

Table 1: A description of the datasets in logistic regression.

Dataset	# Training	# Testing	Architecture	Loss
MNIST	60,000	10,000	784-1000-500-250-30-250-500-1000-784	Cross-entropy
FACES	103,500	62,100	625-1000-500-250-30-250-500-1000-625	Mean squared error
CURVES	20,000	10,000	784-1000-500-250-30-250-500-1000-784	Cross-entropy

Table 2: The corresponding information in autoencoders.

- The perturbation value ϵ is 1e-8.

The hyper-parameters of other methods are tuned for their best numerical performance depending on the network architectures. We list the experimental settings and tuning mechanisms into two parts, Autoencoders and CNNs (including ConvNet and ResNet-18), respectively.

Autoencoders

- Autoencoders are fully-connected neural networks. We test autoencoders on three datasets. The corresponding information is reported in Table 2.

- We describe the implementation details of the algorithms used in autoencoders.

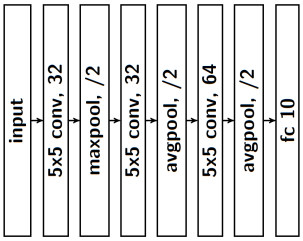
  - SGD: The stochastic gradient method with momentum 0.9. The weight decay is set to be 10^{-5} and the learning rate is fixed to be the best one from $\eta_0 \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5\}$.
  - KFAC: The learning rate is set to $\eta = \eta_0 \widehat{\beta}^{\text{epoch}}$. η_0 and $\widehat{\beta}$ is determined through grid search from $\eta_0 \in \{0.3, 0.5, 1\}$ and $\widehat{\beta} \in \{0.99, 1\}$. The damping and the momentum parameter are set to be 0.2 and 0.9, respectively.
  - SKQN-L: The learning rate is set to 1 in autoencoder for MNIST and FACES, 1.5 for CURVES. The parameter γ_k is set to $0.2 \times (\text{epoch})^{0.99}$. The momentum is set to be 0.9 and the memory size is 5.
  - SKQN-B1/SKQN-B2: The learning rate is set to 0.7 in autoencoder for MNIST, 0.4 for FACES and 0.8 for CURVES. The damping is $\gamma_0 \times (\text{epoch})^{0.99}$ with $\gamma_0 = 0.1$ for MNIST and CURVES, 0.2 for FACES. The BFGS damping is set to be 0.5 and the momentum is 0.9.

Deep CNNs

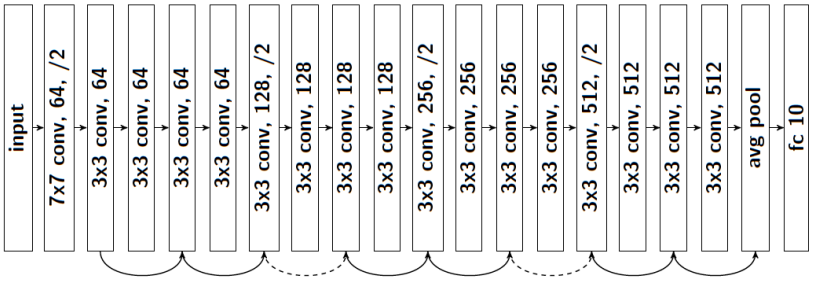
In this part, we describe the implementation details for ConvNet and ResNet-18. The loss function is cross-entropy in these two problems. The hyper-parameters of each method are the same for both case unless otherwise specified.

- The network architectures used in ConvNet and ResNet-18 are presented in Figure 1. "conv" in the figure means a sequence of convolutional kernel, Batch Normalization layer and Relu function. The numbers next to "conv" is the number of the channels of the outputs.

- SGD: The momentum is set to be 0.9. The learning rate is is set to $\eta = \eta_0(1 - \text{epoch}/\text{epoch\_end})^{\widehat{\beta}}$. The parameters are determined by grid searching for the best result from $\alpha_0 \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5\}$, epoch_end $\in \{80, 85, 90\}$ and $\widehat{\beta} \in \{4, 4.5, 5, 5.5, 6\}$.



(a) ConvNet



(b) ResNet-18

Figure 1: Network Architecture of ConvNet and ResNet-18.

- KFAC: The learning rate is $\eta = \eta_0(1 - \text{epoch}/\text{epoch\_end})^{\widehat{\beta}}$. The parameters are also determined from $\eta_0 \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$, epoch_end $\in \{70, 75, 80, 85\}$ and $\widehat{\beta} \in \{4, 5, 6\}$. The damping parameter and the momentum parameter are set to $0.7\eta_0$, 0.9, respectively. The curvature matrix is evaluated and inverted every 50 iterations.

- SKQN-L: The memory size is 1. The learning rate is set to be $\eta = \eta_0(1 - \text{epoch}/\text{epoch\_end})^{\widehat{\beta}}$. We set $\eta_0 = 0.1$, epoch_end = 85, $\widehat{\beta} = 4$ in the ConvNet and $\eta_0 = 0.15$, epoch_end = 80, $\widehat{\beta} = 6$ in the ResNet-18, respectively. The damping is $0.7 \times \eta_0(\eta/\eta_0)^{1/5}$.

- SKQN-B1: The learning rate for both cases is $\eta = 0.1 \cdot (1 - \text{epoch}/80)^5$. The damping is $0.8 \times 0.1 \cdot (\eta/0.1)^{1/5}$ in deep CNN problems.

- SKQN-B2: The learning rate is set to be $\eta = \eta_0(1 - \text{epoch}/\text{epoch\_end})^{\widehat{\beta}}$ where we set $\eta_0 = 0.12$, epoch_end = 85, $\widehat{\beta} = 5$ in the ConvNet and $\eta_0 = 0.1$, epoch_end = 85, $\widehat{\beta} = 5$ in the ResNet-18. The damping is $0.8 \times \eta_0(\eta/\eta_0)^{1/5}$.

A. Proof of Theorem 1

Proof. It follows from Assumption 2.1) that the descent property holds:

$$\Psi(y) \leq \Psi(x) + \langle \nabla\Psi(x), y - x\rangle + \frac{L_\Psi}{2}\|y - x\|^2. \tag{1}$$

Applying (1) and the Young inequality, we obtain:

$$\begin{aligned}
&\Psi(\theta_{k+1}) - \Psi(\theta_k) \\
\leq& \langle \nabla\Psi(\theta_k), \theta_{k+1} - \theta_k \rangle + \frac{L_\Psi}{2}\|\theta_{k+1} - \theta_k\|^2 \\
\leq& \left\langle \nabla_{\mathcal{S}_g^k}\Psi(\theta_k), -(\lambda_k I + B_k)^{-1}\nabla_{\mathcal{S}_g^k}\Psi(\theta_k) \right\rangle + \left\langle \nabla\Psi(\theta_k) - \nabla_{\mathcal{S}_g^k}\Psi(\theta_k), -(\lambda_k I + B_k)^{-1}\nabla_{\mathcal{S}_g^k}\Psi(\theta_k) \right\rangle \\
&+ \frac{L_\Psi}{2}\|(\lambda_k I + B_k)^{-1}\|_2^2\|\nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|_2^2 \\
\leq& -(h+\lambda_k)^{-1}\|\nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|_2^2 + \lambda_k^{-1}\|\nabla\Psi(\theta_k) - \nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|_2^2 + \frac{\lambda_k}{4}\|(\lambda_k I + B_k)^{-1}\|_2^2\|\nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|_2^2 \\
&+ \frac{L_\Psi}{2}\|(\lambda_k I + B_k)^{-1}\|_2^2\|\nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|_2^2 \\
\leq& -\left[(h+\lambda_k)^{-1} - \frac{1}{4}\lambda_k^{-1} - \frac{L_\Psi}{2}\lambda_k^{-2}\right]\|\nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|_2^2 + \lambda_k^{-1}\|\nabla\Psi(\theta_k) - \nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|^2.
\end{aligned} \tag{2}$$

Recalling that the parameter λ_k is adjusted by the norm of the stochastic gradient as follows for a given $r_1 < 1 < r_2$:

$$\lambda_k = \begin{cases} \frac{2r_1}{\|g_{k-1}\|+r_1}\alpha_k^{-1} & \|g_{k-1}\| < r_1, \\ \frac{2\|g_{k-1}\|}{\|g_{k-1}\|+r_2}\alpha_k^{-1} & \|g_{k-1}\| > r_2, \\ \alpha_k^{-1} & \text{otherwise}, \end{cases} \tag{3}$$

we prove that $(h+\lambda_k)^{-1} - \frac{1}{4}\lambda_k^{-1} - \frac{L_\Psi}{2}\lambda_k^{-2}$ is positive and bounded in all three cases.

We first consider the case when $\|g_{k-1}\| \in [r_1, r_2]$. Since $\lambda_k^{-1} = \alpha_k < \frac{r_1}{4r_2(L_\Psi+h)}$, we have $\frac{L_\Psi}{\lambda_k} < \frac{1}{4}$, and hence

$$\frac{1}{h+\lambda_k} - \frac{1}{4\lambda_k} - \frac{L_\Psi}{2\lambda_k^2} > \frac{1}{h+\lambda_k} - \frac{3}{8}\frac{1}{\lambda_k} > \frac{1}{8}\alpha_k. \tag{4}$$

The last inequality follows from $\lambda_k = \alpha_k^{-1} > \frac{4r_2(L_\Psi+h)}{r_1} > h$. As for the case when $\|g_{k-1}\| < r_1$, we have

$$\lambda_k^{-1} = \alpha_k\frac{\|g_{k-1}\| + r_1}{2r_1} \in [\frac{1}{2}\alpha_k, \alpha_k].$$

Then, we can still obtain

$$\frac{1}{h+\lambda_k} - \frac{1}{4\lambda_k} - \frac{L_\Psi}{2\lambda_k^2} > \frac{1}{h+\lambda_k} - \frac{3}{8}\frac{1}{\lambda_k} > \frac{1}{8}\frac{1}{\lambda_k} \geq \frac{1}{16}\alpha_k. \tag{5}$$

For the last case when $\|g_{k-1}\| > r_2$, it follows

$$\lambda_k^{-1} = \alpha_k\frac{\|g_{k-1}\| + r_2}{2\|g_{k-1}\|} \in [\frac{1}{2}\alpha_k, \alpha_k],$$

which implies the desired result as in (5).

Next, by using the Young inequality and taking conditional expectation based on $\mathcal{F}_{k-1}$ together with $\mathbb{E}[\nabla_{\mathcal{S}_g^k}\Psi(\theta^k)|\mathcal{F}_{k-1}] = \nabla\Psi(\theta^k)$ yields

$$\begin{aligned}
\mathbb{E}[\|\nabla_{\mathcal{S}_g^k}\Psi(\theta^k)\|^2|\mathcal{F}_{k-1}] =& \mathbb{E}[\|\nabla_{\mathcal{S}_g^k}\Psi(\theta^k) - \nabla\Psi(\theta^k) + \nabla\Psi(\theta^k)\|^2|\mathcal{F}_{k-1}] \\
=& \mathbb{E}[\|\nabla_{\mathcal{S}_g^k}\Psi(\theta^k) - \nabla\Psi(\theta^k)\|^2|\mathcal{F}_{k-1}] + \|\nabla\Psi(\theta^k)\|^2.
\end{aligned} \tag{6}$$

Taking the expectation related to $\mathcal{S}_g^k$ of (2) on both sides conditioned on $\mathcal{F}_{k-1}$ and combining (2)-(6), we obtain

$$\begin{aligned}
&\mathbb{E}[\Psi(\theta_{k+1}) - \Psi(\theta_k)|\mathcal{F}_{k-1}] \\
\leq& -\frac{1}{16}\alpha_k\|\nabla\Psi(\theta_k)\|^2 + \left[\frac{1}{\lambda_k} - \frac{1}{16}\alpha_k\right]\mathbb{E}[\|\nabla_{\mathcal{S}_g^k}\Psi(\theta^k) - \nabla\Psi(\theta^k)\|^2|\mathcal{F}_{k-1}] \\
\leq& -\frac{1}{16}\alpha_k\|\nabla\Psi(\theta_k)\|^2 + \tilde{\beta}_k\sigma_k^2,
\end{aligned} \tag{7}$$

where $\tilde{\beta}_k = \frac{1}{\lambda_k} - \frac{1}{16}\alpha_k \leq (1 - \frac{1}{16})\alpha_k$. Taking expectation, summing over the inequality and using the assumptions that there exists Ψ_{inf} such that $\Psi(\theta) \geq \Psi_{\text{inf}}, \forall \theta \in \text{dom}\Psi$, we obtain:

$$\sum_{k=1}^{\infty} \frac{1}{16}\alpha_k \mathbb{E}\|\nabla\Psi(\theta_k)\|^2 \leq \Psi(\theta_1) - \Psi^* + \sum_{k=1}^{\infty} \tilde{\beta}_k \sigma_k^2. \tag{8}$$

Therefore, we have $\sum_{k=1}^{\infty} \alpha_k \mathbb{E}\|\nabla\Psi(\theta_k)\|^2 < \infty$, which implies that $\sum_{k=1}^{\infty} \alpha_k \|\nabla\Psi(\theta_k)\|^2 < \infty$ almost surely. Consequently, we can infer

$$\lim_{k\to\infty} \inf \nabla\Psi(\theta_k) = 0 \text{ almost surely .}$$

Taking expectation, multiplying α_k on both sides of inequality (6) and summing over all k yields

$$\sum_{k=1}^{\infty} \alpha_k \mathbb{E}\|\nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|^2 = \sum_{k=1}^{\infty} \alpha_k \sigma_k^2 + \sum_{k=1}^{\infty} \alpha_k \|\nabla\Psi(\theta_k)\|^2 < \infty.$$

By the Young inequality, it implies

$$\begin{aligned}
\sum_{k=1}^{\infty} \alpha_k^{-1} \mathbb{E}\|\theta_{k+1} - \theta_k\|^2 &= \sum_{k=1}^{\infty} \alpha_k^{-1} \mathbb{E}\|(B_k + \lambda_k I)^{-1} \nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|^2 \\
&\leq \sum_{k=1}^{\infty} 2\frac{1}{\alpha_k(\lambda_k)^2} \mathbb{E}\|\nabla_{\mathcal{S}_g^k}\Psi(\theta_k)\|^2 \\
&< \infty.
\end{aligned}$$

It follows that

$$\sum_{k=1}^{\infty} \alpha_k^{-1} \mathbb{E}\|\theta_{k+1} - \theta_k\|^2 < \infty \text{ and } \sum_{k=1}^{\infty} \alpha_k^{-1} \|\theta_{k+1} - \theta_k\|^2 < \infty \text{ almost surely.}$$

On the events $\mathcal{E} = \{\|\nabla\Psi(\theta_k)\| \text{ does not converge}\}$, there exists $\epsilon > 0$ and two increasing sequences $\{p_i\}_i$, $\{q_i\}_i$ such that $p_i < q_i$ and

$$\|\nabla\Psi(\theta_{p_i})\| \geq 2\epsilon, \quad \|\nabla\Psi(\theta_{q_i})\| < \epsilon, \quad \|\nabla\Psi(\theta_k)\| \geq \epsilon,$$

for $k = p_i + 1, \ldots, q_i - 1$. Thus, it follows that

$$\epsilon^2 \sum_{i=0}^{\infty} \sum_{k=p_i}^{q_i-1} \alpha_k \leq \sum_{i=0}^{\infty} \sum_{k=p_i}^{q_i-1} \alpha_k \|\nabla\Psi(\theta_k)\|^2 \leq \sum_{k=0}^{\infty} \alpha_k \|\nabla\Psi(\theta_k)\|^2 < \infty. \tag{9}$$

Setting $\zeta_i = \sum_{k=p_i}^{q_i-1} \alpha_k$, it follows $\zeta_i \to 0$. Then by the Hölder's inequality, we obtain

$$\|\theta_{p_i} - \theta_{q_i}\| \leq \sqrt{\zeta_i}[\sum_{k=p_i}^{q_i-1} \alpha_k^{-1} \|\theta_{k+1} - \theta_k\|^2]^{1/2} \to 0.$$

Due to the Lipschitz property of $\nabla\Psi$, we have $\|\nabla\Psi(\theta_{p_i}) - \nabla\Psi(\theta_{q_i})\| \to 0$, which is a contradiction. This implies $\mathbb{P}(\mathcal{E}) = 0$. Hence, $\nabla\Psi(\theta_k)$ converges to zero almost surely.

□

B. Proof of Theorem 2

Proof. From the inequality (7) in the proof of Theorem 1, we have

$$\mathbb{E}[\Psi(\theta_{k+1}) - \Psi(\theta_k)|\mathcal{F}_{k-1}] \leq \quad -\frac{1}{16}\alpha_k \|\nabla\Psi(\theta_k)\|^2 + \frac{15}{16}\alpha_k \sigma_k^2. \tag{10}$$

Combining the Assumption 3.1) and $\sigma_k^2 \leq M_\sigma \zeta^{k-1}$, we have:

$$\mathbb{E}[\Psi(\theta_{k+1}) - \Psi_{\text{inf}}|\mathcal{F}_{k-1}] \leq (1 - \frac{1}{8}c\alpha_k)\mathbb{E}[\Psi(\theta_k) - \Psi_{\text{inf}}|\mathcal{F}_{k-1}] + \frac{15}{16}\alpha_k M_\sigma \zeta^{k-1}.$$

We prove Theorem 2 by induction. For $k = 1$, the inequality holds by the definition $\mu = \max\{\Psi(\theta_1) - \Psi_{\text{inf}}, \frac{15M_\sigma}{c}\}$. Then, we assume the inequality holds for $k \in \mathbb{N}$. Combining $\alpha_k \equiv \alpha < \min\left\{\frac{r_1}{4r_2(L_\Psi + h)}, \frac{8}{c}\right\}$, $\mu = \max\{\Psi(\theta_1) - \Psi_{\text{inf}}, \frac{15M_\sigma}{c}\}$ and $\nu = \max\{\zeta, 1 - \frac{1}{16}c\alpha\}$, we have

$$\begin{aligned}
\mathbb{E}[\Psi(\theta_{k+1}) - \Psi_{\text{inf}}|\mathcal{F}_{k-1}] &\leq (1 - \frac{1}{8}c\alpha_k)\mu\nu^{k-1} + \frac{15}{16}\alpha_k M_\sigma \zeta^{k-1} \\
&\leq \mu\nu^{k-1}\left(1 - \frac{1}{8}c\alpha_k + \frac{15}{16}M_\sigma\frac{\alpha_k}{\mu}\right) \\
&\leq \mu\nu^{k-1}\left(1 - \frac{1}{16}c\alpha_k\right) \\
&\leq \mu\nu^k,
\end{aligned} \tag{11}$$

which proves the inequality for $k + 1$. This completes the proof of Theorem 2.

□

C. Proof of Theorem 3

We first state the settings of Theorem 3. Consider the case when $\psi_i(\theta) = \ell_i(\theta) = \ell(f(x_i, \theta), y_i)$, where $f(\cdot, x) : \mathbb{R}^n \to \mathbb{R}^m$. The Hessian matrix is $\nabla^2\Psi(\theta) := H(\theta) + \Pi(\theta)$, where

$$H(\theta) = \frac{1}{N}\sum_{i=1}^N H_i(\theta) = \frac{1}{N}\sum_{i=1}^N J_f^i(\theta)\nabla_f^2\ell_i(\theta)\left(J_f^i(\theta)\right)^\top, \tag{12}$$

$$\Pi(\theta) = \frac{1}{N}\sum_{i=1}^N \Pi_i(\theta) = \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^m \nabla_{f_j}\ell_i(\theta)\nabla_\theta^2 f_j^i(\theta), \tag{13}$$

where $J_f^i(\theta) = \nabla_\theta f(x_i, \theta) \in \mathbb{R}^{n\times m}$ and $f_j^i(\theta)$ is the j-th component of $f_i(\theta)$.

Consider the iteration in the neighborhood of θ^* :

$$\theta_{k+1} = \theta_k - B_k^{-1}\nabla\Psi(\theta_k). \tag{14}$$

Here we consider the true gradient for simplicity and the conclusion also holds for stochastic gradient by adjusting relative assumptions.

The curvature matrix is

$$B_k = H_{\mathcal{S}_H^k}(\theta) + \Lambda_k,$$

where $H_{\mathcal{S}_H^k}(\theta)$ is the base matrix and Λ_k is the refinement matrix. We formulate $H_{\mathcal{S}_H^k}(\theta)$ as

$$H_{\mathcal{S}_H^k}(\theta) = \frac{1}{|\mathcal{S}_H^k|}\sum_{i\in\mathcal{S}_H^k} H_i(\theta), \tag{15}$$

and Λ_k is generated by the BFGS method. Suppose that Λ_k satisfies the following secant condition:

$$\Lambda_k u_{k-1} = v_{k-1}, \tag{16}$$

where $u_{k-1} = \theta_k - \theta_{k-1}$ and

$$\begin{aligned}
v_{k-1} =& \frac{1}{|\mathcal{S}_H^{k-1}|}\sum_{i\in\mathcal{S}_H^{k-1}}\left(J_f^i(\theta_k) - J_f^i(\theta_{k-1})\right)\nabla_f\ell_i(\theta_k) \\
=& \frac{1}{|\mathcal{S}_H^{k-1}|}\sum_{i\in\mathcal{S}_H^{k-1}}\sum_{j=1}^m\left(\nabla_\theta f_j^i(\theta_k) - \nabla_\theta f_j^i(\theta_{k-1})\right)\nabla_{f_j}\ell_i(\theta_k).
\end{aligned} \tag{17}$$

We want to prove that if the sample size is sufficiently large, then the stochastic Dennis-More condition holds. Hence, the local superlinear convergence speed can be guaranteed. A few assumptions are listed below.

Assumption 1 *1.1) The sequence $\{\theta_k\}$ satisfies $\sum_k \|\theta_k - \theta^*\| < \infty$ a.s. for an optimal point θ^* where $\nabla^2\Psi(\theta^*)$ are positive definite and there exists $\widetilde{\lambda} > 0$ such that for $i = 1, \ldots, n$, $\Pi_i(\theta^*) \succeq \widetilde{\lambda} I$.*

1.2) The gradient $\nabla_f \ell_i(\theta)$ is bounded, the Hessian $\nabla_f^2 \ell_i(\theta)$ is bounded and Lipschitz continuous near θ^* with Lipschitz constant L_ℓ, $\forall i = 1, \ldots, N$, i.e., $\|\nabla_f \ell_i(\theta)\|_2 \le \kappa_\ell$, $\|\nabla_f^2 \ell_i(\theta)\|_2 \le \widetilde{\kappa}_\ell$ and $\|\nabla_f^2 \ell_i(\theta_1) - \nabla_f^2 \ell_i(\theta_1)\|_2 \le L_\ell \|\theta_1 - \theta_2\|$, for any θ_1, θ_2 near θ^*.

1.3) The gradient $\nabla f_j^i(\theta)$ is bounded, the Hessian $\nabla^2 f_j^i(\theta)$ is bounded and Lipschitz continuous near θ^* with Lipschitz constant L_f, $\forall i = 1, \ldots, N$ and $\forall j = 1, \ldots, m$, i.e., $\|\nabla f_j^i(\theta)\|_2 \le \kappa_f$, $\|\nabla^2 f_j^i(\theta)\|_2 \le \tilde{\kappa}_f$ and $\|\nabla^2 f_j^i(\theta_1) - \nabla^2 f_j^i(\theta_2)\|_2 \le L_f \|\theta_1 - \theta_2\|_2$ for any θ_1, θ_2 near θ^*.

Lemma 1 *Under Assumption 1, the following conclusions hold:*

- *The GGN matrix $H_i(\theta)$ and $\Pi_i(\theta)$ are bounded for $i = 1, \ldots, n$, i.e., there exists constants κ_H and κ_Π such that for all θ near θ^*,*
$$\|H_i(\theta)\| \le \kappa_H \quad \text{and} \quad \|\Pi_i(\theta)\| \le \kappa_\Pi.$$
- *$J_f^i(\theta)$ is Lipschitz continuous near θ^* with Lipschitz constant L_J, $\forall i = 1, \ldots, N$, i.e.,*
$$\|J_f^i(\theta_1) - J_f^i(\theta_2)\| \le L_J \|\theta_1 - \theta_2\|,$$
for any θ_1, θ_2 near θ^*.
- *$H(\theta)$ is Lipschitz continuous near θ^* with Lipschitz constant L_H, $\forall i = 1, \ldots, N$, i.e.,*
$$\|H(\theta_1) - H(\theta_2)\| \le L_H \|\theta_1 - \theta_2\|,$$
for any θ_1, θ_2 near θ^*.

Proof. We first prove that $H_i(\theta)$ and $\Pi_i(\theta)$ are bounded. Recalling $H_i(\theta) = J_f^i(\theta) \nabla_f^2 \ell_i(\theta) \left(J_f^i(\theta)\right)^\top$, we have

$$\begin{aligned} \|J_f^i(\theta) \nabla_f^2 \ell_i(\theta) (J_f^i(\theta))^\top\|_2 &\le \|J_f^i(\theta)\|_2 \|\nabla_f^2 \ell_i(\theta)\|_2 \| \left(J_f^i(\theta)\right)^\top \|_2 \\ &\le \|J_f^i(\theta)\|_F \|\nabla_f^2 \ell_i(\theta)\|_2 \| \left(J_f^i(\theta)\right)^\top \|_F. \end{aligned} \tag{18}$$

Since $J_f^i(\theta) = [\nabla f_1^i(\theta), \ldots, \nabla f_m^i(\theta)]$ and $\|\nabla f_j^i(\theta)\|_2 \le \kappa_f$, we have $\|J_f^i(\theta)\|_F \le \sqrt{m}\kappa_f$ which implies $\|H_i(\theta)\|_2 \le m\kappa_f^2 \tilde{\kappa}_\ell := \kappa_H$. Similarly, we have

$$\begin{aligned} \|\Pi_i(\theta)\|_2 &= \| \sum_{j=1}^m \nabla_{f_j} \ell_i(\theta) \nabla_\theta^2 f_j^i(\theta)\|_2 \le \sum_{j=1}^m \|\nabla_{f_j} \ell_i(\theta) \nabla_\theta^2 f_j^i(\theta)\|_2 \\ &\le \sum_{j=1}^m |\nabla_{f_j} \ell_i(\theta)| \|\nabla_\theta^2 f_j^i(\theta)\|_2 \le m\kappa_\ell \tilde{\kappa}_f := \kappa_\Pi. \end{aligned} \tag{19}$$

We next show that $J_f^i(\theta)$ is Lipschitz continuous. For each row of $J_f^i(\theta)$, we get:

$$\nabla f_j^i(\theta_2) - \nabla f_j^i(\theta_1) = \int_0^1 \nabla^2 f_j^i \left((1-t)\theta_1 + t(\theta_2)\right) (\theta_2 - \theta_1)\, dt.$$

This implies $\|\nabla f_j^i(\theta_2) - \nabla f_j^i(\theta_1)\|_2 \le \frac{1}{2}\tilde{\kappa}_f \|\theta_2 - \theta_1\|_2$ and

$$\begin{aligned} \|J_f^i(\theta_1) - J_f^i(\theta_2)\|_2 &\le \|J_f^i(\theta_1) - J_f^i(\theta_2)\|_F \\ &\le \frac{\sqrt{m}}{2} \tilde{\kappa}_f \|\theta_1 - \theta_2\|_2 \\ &:= L_J \|\theta_1 - \theta_2\|_2. \end{aligned} \tag{20}$$

Finally, we show that $H_i(\theta)$ is Lipschitz continuous near θ^* with Lipschitz constant L_H:

$$
\begin{aligned}
&\|H_i(\theta_1) - H_i(\theta_2)\|_2 \\
=&\|J_f^i(\theta_2)\nabla_f^2\ell_i(\theta_2)(J_f^i(\theta_2))^\top - J_f^i(\theta_1)\nabla_f^2\ell_i(\theta_2)(J_f^i(\theta_2))^\top \\
&+ J_f^i(\theta_1)\nabla_f^2\ell_i(\theta_2)(J_f^i(\theta_2))^\top - J_f^i(\theta_1)\nabla_f^2\ell_i(\theta_1)(J_f^i(\theta_1))^\top\|_2 \\
\leq&\|J_f^i(\theta_2) - J_f^i(\theta_1)\|_2\|\nabla_f^2\ell_i(\theta_2)\|_2\|(J_f^i(\theta_2))^\top\|_2 \\
&+ \|J_f^i(\theta_1)\nabla_f^2\ell_i(\theta_2)(J_f^i(\theta_2))^\top - J_f^i(\theta_1)\nabla_f^2\ell_i(\theta_1)(J_f^i(\theta_1))^\top\|_2 \\
\leq&L_f\|\theta_2 - \theta_1\|_2\tilde{\kappa}_\ell\sqrt{m}\kappa_f + \sqrt{m}\kappa_f\left(\|\nabla_f^2\ell_i(\theta_2)(J_f^i(\theta_2))^\top - \nabla_f^2\ell_i(\theta_1)(J_f^i(\theta_1))^\top\|_2\right) \\
\leq&L_f\|\theta_2 - \theta_1\|_2\tilde{\kappa}_\ell\sqrt{m}\kappa_f + \sqrt{m}\kappa_f\left(\tilde{\kappa}_\ell\|(J_f^i(\theta_2))^\top - (J_f^i(\theta_1))^\top\|_2 + \sqrt{m}\kappa_f\|\nabla_f^2\ell_i(\theta_2) - \nabla_f^2\ell_i(\theta_1)\|_2\right) \\
\leq&L_f\|\theta_2 - \theta_1\|_2\tilde{\kappa}_\ell\sqrt{m}\kappa_f + \sqrt{m}\kappa_f\left(\tilde{\kappa}_J L_\ell\|\theta_1 - \theta_2\|_2 + \sqrt{m}\kappa_f L_\ell\|\theta_1 - \theta_2\|_2\right) \\
:=&L_H\|\theta_1 - \theta_2\|_2.
\end{aligned}
\tag{21}
$$

□

Our local results are summarized in the following theorem.

Theorem. Suppose that Assumption 1 is satisfied. If the sample size $|S_H^k|$ increases superlinearly, then the sequence $\{\theta_k\}$ generated by (14) converges to θ^* superlinearly almost surely.

Proof. The proof is divided into two parts. The first part is to show that the stochastic Dennis-Morë condition holds almost surely, i.e.,

$$
\lim_{k\to\infty}\frac{\|(B_k - \nabla^2\Psi(\theta^*)s_k)\|_2}{\|s_k\|_2} = 0 \ a.s.. \tag{22}
$$

The second part is to show that we can obtain the superlinear convergence rate from (22).

(1.) By Lemma 1, we have $\|H_{S_H^k}(\theta) - H(\theta)\| \leq 2\kappa_H$, $\|\Pi_{S_H^k}(\theta) - \Pi(\theta)\| \leq 2\kappa_\Pi$. The matrix Bernstein's inequality yields

$$
\mathbb{P}(\|H_{S_H^k}(\theta) - H(\theta)\|_2 \geq \epsilon_k) \leq 2n\exp\{-\frac{\epsilon_k^2|S_H^k|}{16\kappa_1^2}\} \text{ and } \mathbb{P}(\|\Pi_{S_H^k}(\theta) - \Pi(\theta)\|_2 \geq \epsilon_k) \leq 2n\exp\{-\frac{\epsilon_k^2|S_H^k|}{16\kappa_2^2}\}.
$$

By construction, let $\sum_{k=1}^\infty \epsilon_k < \infty$ and the sample size grow so that $\sum_{k=1}^\infty 2n\exp\{-\frac{\epsilon_k^2|S_H^k|}{16\kappa^2}\} < \infty$. This can be guaranteed, for example, if we choose $\epsilon_k = O(\frac{1}{k^{1+\delta_1}})$ and $|S_H^k| = O(k^{3+3\delta_1})$.

By Borel-Cantelli Lemma, there exists k_0 such that $\forall k > k_0$, $\|H_{S_H^k}(\theta) - H(\theta)\|_2 \leq \epsilon_k$ a.s. and $\|\Pi_{S_H^k}(\theta) - \Pi(\theta)\|_2 \leq \epsilon_k$ a.s.. Define the space where $\sum_k \|\theta_k - \theta^*\| < \infty$, $\|H_{S_H^k}(\theta) - H(\theta)\|_2 \leq \epsilon_k$ and $\|\Pi_{S_H^k}(\theta) - \Pi(\theta)\|_2 \leq \epsilon_k$ by Ξ. It is easy to know that $\mathbb{P}(\Xi) = 1$. Denote $e_k = \max\{\|\theta_{k+1} - \theta^*\|, \|\theta_k - \theta^*\|\}$, and $\sum_{i=1}^\infty e_k < \infty$ in space Ξ.

Define two hypothetical sequences:

$$
\begin{aligned}
\widehat{\Lambda}_{k+1} &= \Lambda_k - \frac{\Lambda_k u_k u_k^\top \Lambda_k}{u_k^\top\Lambda_k u_k} + \frac{\Pi_{S_H^k}(\theta^*)u_k u_k^\top \Pi_{S_H^k}(\theta^*)}{u_k^\top\Pi_{S_H^k}(\theta^*)u_k}, \\
\widetilde{\Lambda}_{k+1} &= \Lambda_k - \frac{\Lambda_k u_k u_k^\top \Lambda_k}{u_k^\top\Lambda_k u_k} + \frac{\Pi(\theta^*)u_k u_k^\top \Pi(\theta^*)}{u_k^\top\Pi(\theta^*)u_k}.
\end{aligned}
$$

From Lemma C.14 [2], we have:

$$
\|\widetilde{\Lambda}_{k+1} - I\|_F^2 - \|\Lambda_k - I\|_F^2 = -\left[\left(1 - \frac{u_k^\top\Lambda_k\Lambda_k u_k}{u_k^\top\Lambda_k u_k}\right)^2 + 2\left(\frac{u_k^\top\Lambda_k\Lambda_k\Lambda_k u_k}{u_k^\top\Lambda_k u_k} - \left(\frac{u_k^\top\Lambda_k\Lambda_k u_k}{u_k^\top\Lambda_k u_k}\right)^2\right)\right].
$$

Without loss of generality, we assume that $\Pi(\theta^*) = I$, otherwise do linear transformation for variables by $\widetilde{\theta} = \Pi(\theta^*)^{1/2}\theta$. We next need to show that $\|\Lambda_k - I\| - \|\widetilde{\Lambda}_{k+1} - I\| \to 0$.

From section 4 in [1], this is required to prove that

$$
\|\Lambda_{k+1} - \widetilde{\Lambda}_{k+1}\| \leq O(\epsilon_k + e_k).
$$

From Lemma C.15 in [2], this is required to prove that there exists constants c_1, c_2, c_3, c_4 such that:

a.1.1) $c_1 u_k^\top u_k \le v_k^\top u_k \le c_2 u_k^\top u_k$,

a.1.2) $\|\delta_k\| \le c_3 \|u_k\| e_k$,

a.1.3) $\frac{v_k^\top \delta_k}{u_k^\top v_k} \le c_4 e_k$,

where $\delta_k = \Pi_{S_H^k}(\theta^*) u_k - v_k$.

From Assumption 1, we can obtain that when θ_k nears θ^*, there exists $c_1 < \frac{1}{2}\widetilde{\lambda}$ such that $v_k^\top u_k \ge c_1 u_k^\top u_k$. By Lemma 1, it is easy to know that $v_k^\top u_k \le \|u_k\|\|v_k\| \le L_J \kappa_\ell \|u_k\|^2$. Let $c_2 = L_J \kappa_\ell$ and we prove a.1.1). Note that each f_j^i is twice continuously differentiable, we have

$$\begin{aligned}\delta_k &= \Pi_{S_H^k}(\theta^*) u_k - v_k \\ &= \frac{1}{|\mathcal{S}_H^k|} \sum_{i \in \mathcal{S}_H^k} \sum_{j=1}^{m} \int_0^1 \left(\nabla_{f_j} \ell_i(\theta^*) \nabla_\theta^2 f_j^i(\theta^*) - \nabla_{f_j} \ell_i(\theta_{k+1}) \nabla^2 f_j^i \left((1-t)\theta_k + t(\theta_{k+1}) \right) \right) u_k dt. \end{aligned} \tag{23}$$

Since $\|\nabla_{f_j} \ell_i(\theta)\|_2 \le \kappa_\ell$, $\|\nabla_{f_j}^2 \ell_i(\theta)\|_2 \le \tilde{\kappa}_\ell$ and $\|\nabla^2 f_j^i(\theta_1) - \nabla^2 f_j^i(\theta_2))\|_2 \le L_f \|\theta_1 - \theta_2\|_2, \forall i, j$, we conclude that there exists constant c_3, such that a.1.2 holds. a.1.3 follows from a.1.1 and a.1.2 immediately by Cauchy–Schwarz inequality.

By a.1.1), a.1.2) and a.1.3), following from Lemma C.15 in [2], we can prove that

$$\begin{aligned}\|\Lambda_{k+1} - \widehat{\Lambda}_{k+1}\| &= \left\| -\frac{v_k \delta^\top + \delta v_k^\top + \delta\delta^\top}{u_k^\top v_k} + \frac{v_k^\top \delta (v_k v_k^\top + v_k \delta^\top + \delta v_k^\top + \delta\delta^\top)}{u_k^\top v_k + v_k^\top \delta} \right\| \le O(e_k), \\ \|\widehat{\Lambda}_{k+1} - \widetilde{\Lambda}_{k+1}\| &= \left\| -\frac{\widetilde{v}_k \hat{\delta}^\top + \hat{\delta} \widetilde{v}_k^\top + \hat{\delta}\hat{\delta}^\top}{u_k^\top \widetilde{v}_k} + \frac{\widetilde{v}_k^\top \hat{\delta} (\widetilde{v}_k \widetilde{v}_k^\top + \widetilde{v}_k \hat{\delta}^\top + \hat{\delta} \widetilde{v}_k^\top + \hat{\delta}\hat{\delta}^\top)}{u_k^\top \widetilde{v}_k + v_k^\top \hat{\delta}} \right\| \le O(\epsilon_k), \end{aligned} \tag{24}$$

where $\widehat{v}_k = \Pi_{S_H^k}(\theta^*) u_k$, $\widetilde{v}_k = \Pi(\theta^*) u_k$ and $\hat{\delta} = \widehat{v}_k - \widetilde{v}_k$. This shows that

$$\|\Lambda_{k+1} - \widetilde{\Lambda}_k\| \le O(e_k + \epsilon_k).$$

Following the same idea of section 4 in [1], we have

$$\lim_{k \to \infty} \frac{\|(\Lambda_k - I) u_k\|}{\|u_k\|} = 0 \quad a.s..$$

Our previous results yield that:

$$\begin{aligned} &\lim_{k \to \infty} \frac{\|(B_k - \nabla^2 \Psi(\theta^*) s_k)\|}{\|s_k\|} \\ = &\lim_{k \to \infty} \frac{\|(H_{S_H^k} + \Lambda_k - \nabla \Psi(\theta^*)) u_k\|}{\|u_k\|} \\ = &\lim_{k \to \infty} \frac{\|(H_{S_H^k}(\theta_k) - H(\theta_k) + H(\theta_k) - H(\theta^*) + \Lambda_k - \Pi(\theta^*)) u_k\|}{\|u_k\|} \\ \le &\lim_{k \to \infty} \frac{\|(H_{S_H^k}(\theta_k) - H(\theta_k)\|\|u_k\| + \|H(\theta_k) - H(\theta^*)\|\|u_k\| + \|(\Lambda_k - \Pi(\theta^*)) u_k\|}{\|u_k\|} = 0. \end{aligned} \tag{25}$$

The result (25) is actually the stochastic Dennis-Möre condition.

(2.) The next step is to show that superlinear convergence results are guaranteed if (25) holds. For simplicity of notations, we set

$$\begin{aligned} w_1^k &= (B_k - \nabla^2 \Psi(\theta^*))(\theta^{k+1} - \theta^k), \\ w_2^k &= \nabla \Psi(\theta^{k+1}) - \nabla \Psi(\theta^k) - \nabla^2 \Psi(\theta^*)(\theta^{k+1} - \theta^k). \end{aligned}$$

Then by (14), we have

$$B_k(\theta^{k+1}-\theta^k)-\nabla^2\Psi(\theta^*)(\theta^{k+1}-\theta^k)=-\nabla\Psi(\theta^k)-\nabla^2\Psi(\theta^*)(\theta^{k+1}-\theta^k).$$

It follows that

$$w_1^k-w_2^k=-\nabla\Psi(\theta^{k+1}).$$

Due to Assumptions 1-2, we have that $\|w_1^k\|/\|\theta^{k+1}-\theta^k\|$ and $\|w_2^k\|/\|\theta^{k+1}-\theta^k\|$ converges to 0 almost surely. It follows that

$$m_k:=\frac{\|-\nabla\Psi(\theta^{k+1})\|}{\|\theta^{k+1}-\theta^k\|}\to 0 \text{ almost surely.} \tag{26}$$

By the nonsingularity of $\nabla^2\Psi(x^*)$ and the convergence of $\{\theta^k\}$, with probability 1, there exists a constant ξ such that

$$\|\nabla\Psi(\theta^{k+1})\|\geq\xi\|\theta^{k+1}-\theta^*\|.$$

It implies that

$$m_k\geq\frac{\xi\|\theta^{k+1}-\theta^*\|}{\|\theta^{k+1}-\theta^*\|+\|\theta^k-\theta^*\|}.$$

Hence, it follows that

$$\frac{\|\theta^{k+1}-\theta^*\|}{\|\theta^k-\theta^*\|}\leq\frac{m_k}{\xi-m_k}\to 0.$$

This finishes the proof. □

References

[1] Andreas Griewank and Ph L Toint. Local convergence analysis for partitioned quasi-newton updates. *Numerische Mathematik*, 39(3):429–448, 1982. 8, 9

[2] Chaoxu Zhou, Wenbo Gao, and Donald Goldfarb. Stochastic adaptive quasi-newton methods for minimizing expected values. In *International Conference on Machine Learning*, pages 4150–4159, 2017. 8, 9