SUPPLEMENTARY MATERIAL FOR LIMITED-MEMORY GREEDY QUASI-NEWTON METHOD WITH NON-ASYMPTOTIC SUPERLINEAR CONVERGENCE RATE

Anonymous authors
Paper under double-blind review

A PROOF OF PROPOSITION 1

For the initial iteration $t = 0$ with the initial Hessian approximation $\mathbf{B}_0$, curvature pair $\{\mathbf{s}_0, \mathbf{r}_0\}$ and scaling factor ψ_1, consider the corrected Hessian approximation

$$\hat{\mathbf{B}}_1 = \psi_1 \mathbf{B}_1 = \psi_1 \text{BFGS}(\mathbf{B}_0, \mathbf{s}_0, \mathbf{r}_0). \tag{1}$$

Define the corrected initial Hessian approximation as $\tilde{\mathbf{B}}_0 = \psi_1 \mathbf{B}_0$ and the scaled gradient variation as $\tilde{\mathbf{r}}_0 = \psi_1 \mathbf{r}_0$. By performing the BFGS update on $\tilde{\mathbf{B}}_0$ with $\{\mathbf{s}_0, \tilde{\mathbf{r}}_0\}$, we have

$$\begin{aligned}\tilde{\mathbf{B}}_1 = \text{BFGS}(\tilde{\mathbf{B}}_0, \mathbf{s}_0, \tilde{\mathbf{r}}_0) &= \tilde{\mathbf{B}}_0 + \frac{\tilde{\mathbf{r}}_0 \tilde{\mathbf{r}}_0^\top}{\tilde{\mathbf{r}}_0^\top \mathbf{s}_0} - \frac{\tilde{\mathbf{B}}_0 \mathbf{s}_0 \mathbf{s}_0^\top \tilde{\mathbf{B}}_0^\top}{\mathbf{s}_0^\top \tilde{\mathbf{B}}_0 \mathbf{s}_t} = \psi_1 \mathbf{B}_0 + \frac{\psi_1^2 \mathbf{r}_0 \mathbf{r}_0^\top}{\psi_1 \mathbf{r}_0^\top \mathbf{s}_0} - \frac{\psi_1^2 \mathbf{B}_0 \mathbf{s}_0 \mathbf{s}_0^\top \mathbf{B}_0^\top}{\psi_1 \mathbf{s}_0^\top \mathbf{B}_0 \mathbf{s}_t} \\ &= \psi_1 \mathbf{B}_0 + \frac{\psi_1 \mathbf{r}_0 \mathbf{r}_0^\top}{\mathbf{r}_0^\top \mathbf{s}_0} - \frac{\psi_1 \mathbf{B}_0 \mathbf{s}_0 \mathbf{s}_0^\top \mathbf{B}_0^\top}{\mathbf{s}_0^\top \mathbf{B}_0 \mathbf{s}_t} = \psi_1 \Big(\mathbf{B}_0 + \frac{\mathbf{r}_0 \mathbf{r}_0^\top}{\mathbf{r}_0^\top \mathbf{s}_0} - \frac{\mathbf{B}_0 \mathbf{s}_0 \mathbf{s}_0^\top \mathbf{B}_0^\top}{\mathbf{s}_0^\top \mathbf{B}_0 \mathbf{s}_t} \Big) \\ &= \psi_1 \text{BFGS}(\mathbf{B}_0, \mathbf{s}_0, \mathbf{r}_0) = \psi_1 \mathbf{B}_1 = \hat{\mathbf{B}}_1.\end{aligned} \tag{2}$$

This indicates that scaling the Hessian approximation matrix $\mathbf{B}_1$ is equivalent to scaling the initial Hessian approximation $\mathbf{B}_0$ and the gradient variation $\mathbf{r}_0$. Thus, the proposition conclusion holds for the initial iteration $t = 1$.

For any iteration $t > 1$ with the initial Hessian approximation $\mathbf{B}_0$, stored curvature pair $\{\mathbf{s}_u, \mathbf{r}_u\}_{u=0}^{t-1}$ and the scaling factor ψ_t, consider the corrected Hessian approximation

$$\hat{\mathbf{B}}_t = \psi_t \mathbf{B}_t = \psi_{t-1} \text{BFGS}(\mathbf{B}_{t-1}, \mathbf{s}_{t-1}, \mathbf{r}_{t-1}), \ \ldots, \ \mathbf{B}_1 = \text{BFGS}(\mathbf{B}_0, \mathbf{s}_0, \mathbf{r}_0). \tag{3}$$

Define the corrected initial Hessian approximation as $\tilde{\mathbf{B}}_0 = \psi_t \mathbf{B}_0$ and the scaled gradient variations as $\tilde{\mathbf{r}}_u = \psi_t \mathbf{r}_u$ for $u = 0, \ldots, t-1$. By performing the BFGS updates on $\hat{\mathbf{B}}_0$ with $\{\mathbf{s}_u, \tilde{\mathbf{r}}_u\}_{u=0}^{t-1}$, we have

$$\tilde{\mathbf{B}}_t = \text{BFGS}(\tilde{\mathbf{B}}_{t-1}, \mathbf{s}_{t-1}, \tilde{\mathbf{r}}_{t-1}), \ \ldots, \ \tilde{\mathbf{B}}_1 = \text{BFGS}(\tilde{\mathbf{B}}_0, \mathbf{s}_0, \tilde{\mathbf{r}}_0). \tag{4}$$

We now use **induction** to prove the following statement

$$\tilde{\mathbf{B}}_k = \psi_t \mathbf{B}_k, \ \text{for all } k = 1, \ldots, t. \tag{5}$$

For the initial iteration $k = 1$, by performing the BFGS update on $\tilde{\mathbf{B}}_0$ with $\{\mathbf{s}_0, \tilde{\mathbf{r}}_0\}$, we have

$$\tilde{\mathbf{B}}_1 = \text{BFGS}(\tilde{\mathbf{B}}_0, \mathbf{s}_0, \tilde{\mathbf{r}}_0) = \psi_t \Big(\mathbf{B}_0 + \frac{\mathbf{r}_0 \mathbf{r}_0^\top}{\mathbf{r}_0^\top \mathbf{s}_0} - \frac{\mathbf{B}_0 \mathbf{s}_0 \mathbf{s}_0^\top \mathbf{B}_0^\top}{\mathbf{s}_0^\top \mathbf{B}_0 \mathbf{s}_0} \Big) = \psi_t \text{BFGS}(\mathbf{B}_0, \mathbf{s}_0, \mathbf{r}_0) = \psi_t \mathbf{B}_1. \tag{6}$$

Thus, equation 5 holds for $k = 1$. Assume equation 5 holds for iteration $k - 1 \geq 1$, i.e., $\tilde{\mathbf{B}}_{k-1} = \psi_t \mathbf{B}_{k-1}$, and consider iteration k. By performing the BFGS update on $\tilde{\mathbf{B}}_{k-1}$ with $\{\mathbf{s}_{k-1}, \tilde{\mathbf{r}}_{k-1}\}$, we have

$$\tilde{\mathbf{B}}_k = \text{BFGS}(\tilde{\mathbf{B}}_{k-1}, \mathbf{s}_{k-1}, \tilde{\mathbf{r}}_{k-1}) = \psi_t \Big(\mathbf{B}_{k-1} + \frac{\mathbf{r}_{k-1} \mathbf{r}_{k-1}^\top}{\mathbf{r}_{k-1}^\top \mathbf{s}_{k-1}} - \frac{\mathbf{B}_{k-1} \mathbf{s}_{k-1} \mathbf{s}_{k-1}^\top \mathbf{B}_{k-1}^\top}{\mathbf{s}_{k-1}^\top \mathbf{B}_{k-1} \mathbf{s}_{k-1}} \Big) = \psi_t \mathbf{B}_k \tag{7}$$

where $\tilde{\mathbf{B}}_{k-1} = \psi_t \mathbf{B}_{k-1}$ is used in the second equality. By combining equation 6 and equation 7, we prove equation 5 by induction. Thus, we get

$$\tilde{\mathbf{B}}_t = \psi_t \mathbf{B}_t = \hat{\mathbf{B}}_t. \tag{8}$$

This indicates that scaling the Hessian approximation matrix $\mathbf{B}_t$ is equivalent to scaling the initial Hessian approximation $\mathbf{B}_0$ and the gradient variations $\{\mathbf{r}_u\}_{u=0}^{t-1}$.

We conclude that at each iteration t, scaling the Hessian approximation matrix $\mathbf{B}_t$ by ψ_t is equivalent to scaling the initial Hessian approximation $\mathbf{B}_0$ and the gradient variations $\{\mathbf{r}_u\}_{u=0}^{t-1}$ by ψ_t. Therefore, we can incorporate the correction strategy into the displacement step by scaling the gradient variations and maintain the remaining unchanged, which completes the proof.

B PROOF OF PROPOSITION 2

We need the following lemmas to complete the proof.

Lemma 1. *If LG-BFGS and greedy BFGS perform the greedy selection from the same subset $\{\mathbf{e}_i\}_{i=1}^{\tau}$ of size τ and have the same initial settings, the iterates $\{\mathbf{x}_{L,t}\}_t$ generated by LG-BFGS equal to the iterates $\{\mathbf{x}_{G,t}\}_t$ generated by greedy BFGS.*

Proof. We start by noting that greedy BFGS updates the variable with (1) and the Hessian inverse approximation with (3). This is equivalent to updating the variable and the Hessian inverse approximation from the initial Hessian inverse approximation $\mathbf{H}_0$ with all historical curvature pairs $\{\mathbf{s}_k, \mathbf{r}_k\}_{k=0}^{t-1}$ at each iteration t. In this context, we can prove the lemma by proving the iterate $\mathbf{x}_{L,t}$ generated by LG-BFGS equal to the iterate $\mathbf{x}_{G,t}$ generated from the initial Hessian inverse approximation $\mathbf{H}_0$ with all historical curvature pairs $\{\mathbf{s}_k, \mathbf{r}_k\}_{k=0}^{t-1}$ for any iteration $t \geq 0$.[1]

Specifically, we use **induction** to prove the lemma. At the initial iteration $t = 0$, this conclusion holds because LG-BFGS and greedy BFGS have the same initial setting $\mathbf{x}_{L,0} = \mathbf{x}_{G,0}$. Assume that the conclusion holds at iteration $t - 1 \geq 0$, i.e., the iterate $\mathbf{x}_{L,t-1}$ generated by LG-BFGS with the limited-memory curvature pairs $\mathcal{P}_{t-1}$ equal to the iterate $\mathbf{x}_{G,t-1}$ generated by greedy BFGS with all historical curvature pairs $\{\mathbf{s}_k, \mathbf{r}_k\}_{k=0}^{t-2}$ as

$$\mathbf{x}_{L,t-1} = \mathbf{x}_{G,t-1}. \tag{9}$$

Consider iteration t with the new curvature pair $\{\mathbf{s}_{t-1}, \mathbf{r}_{t-1}\}$. Greedy BFGS updates the historical curvature pairs by adding the new curvature pair $\{\mathbf{s}_{t-1}, \mathbf{r}_{t-1}\}$ directly and form the new historical curvature pairs $\{\mathbf{s}_k, \mathbf{r}_k\}_{k=0}^{t-1}$. LG-BFGS updates the curvature pairs by incorporating the information $\{\mathbf{s}_{t-1}, \mathbf{r}_{t-1}\}$ into $\mathcal{P}_{t-1}$ and form the new curvature pairs $\mathcal{P}_t$. From Theorem 3.2 in (Berahas et al., 2022), if the Hessian inverse approximation generated from $\mathbf{H}_0$ with $\mathcal{P}_{t-1}$ equal to that generated from $\mathbf{H}_0$ with $\{\mathbf{s}_k, \mathbf{r}_k\}_{k=0}^{t-2}$, the Hessian inverse approximation generated from $\mathbf{H}_0$ with $\mathcal{P}_t$ equal to that generated from $\mathbf{H}_0$ with $\{\mathbf{s}_k, \mathbf{r}_k\}_{k=0}^{t-1}$. By using this result and equation 9, we get

$$\mathbf{x}_{L,t} = \mathbf{x}_{G,t}. \tag{10}$$

By combining equation 9 and equation 10, we prove by induction that $\mathbf{x}_{L,t} = \mathbf{x}_{G,t}$ for any iteration $t \geq 0$, which completes the proof. □

Lemma 2 (Lemma 4.3 in (Rodomanov & Nesterov, 2021))**.** *Let $\mathbf{x}$ be a decision variable and $\mathbf{B}$ the Hessian approximation satisfying*

$$\nabla^2 f(\mathbf{x}) \preceq \mathbf{B} \preceq \eta \nabla^2 f(\mathbf{x}) \tag{11}$$

for some $\eta \geq 1$. Let also $\mathbf{x}_+$ be the updated decision variable as

$$\mathbf{x}_+ = \mathbf{x} - \mathbf{B}^{-1} \nabla f(\mathbf{x}) \tag{12}$$

and $\lambda_f(\mathbf{x})$ be such that $\lambda_f(\mathbf{x}) C_M \leq 2$. Then, it holds that

$$\phi = \|\mathbf{x}_+ - \mathbf{x}\|_{\nabla^2 f(\mathbf{x})} \leq \lambda_f(\mathbf{x}) \text{ and } \lambda_f(\mathbf{x}_+) \leq \Big(1 + \frac{\lambda_f(\mathbf{x}) C_M}{2}\Big) \frac{\eta - 1 + \frac{\lambda_f(\mathbf{x}) C_M}{2}}{\eta} \lambda_f(\mathbf{x}). \tag{13}$$

Lemma 3 (Lemma 4.4 in (Rodomanov & Nesterov, 2021))**.** *Let $\mathbf{x}$ be a decision variable and $\mathbf{B}$ the Hessian approximation satisfying*

$$\nabla^2 f(\mathbf{x}) \preceq \mathbf{B} \preceq \eta \nabla^2 f(\mathbf{x}) \tag{14}$$

[1] Without loss of generality, we assume $\{\}_a^b = \emptyset$, $\sum_a^b = 0$ and $\prod_a^b = 1$ if $b < a$.

for some $\eta \geq 1$. Let also $\mathbf{x}_+$ be the updated decision variable [cf. equation 12] and $\phi = \|\mathbf{x}_+ - \mathbf{x}\|_{\nabla^2 f(\mathbf{x})}$ be the weighted update difference. Then, it holds that

$$\nabla^2 f(\mathbf{x}_+) \preceq (1 + C_M \phi)\mathbf{B} = \hat{\mathbf{B}} \tag{15}$$

and the Hessian approximation $\mathbf{B}_+$ updated by the BFGS on $\hat{\mathbf{B}}$ with the curvature pair $\{\mathbf{s}, \mathbf{r}\}$ satisfies

$$\nabla^2 f(\mathbf{x}_+) \preceq \text{BFGS}\big(\hat{\mathbf{B}}, \mathbf{s}, \mathbf{r}\big) \preceq \eta(1 + C_M \phi)^2 \nabla^2 f(\mathbf{x}_+). \tag{16}$$

Proof of Proposition 2. From Lemma 1, we know that the iterates generated by LG-BFGS is equivalent to the iterates generated by greedy BFGS, if both perform greedy selection in the same subset $\{\mathbf{e}_i\}_{i=1}^{\tau}$ of memory size τ. In this context, we can prove the linear convergence of the iterates generated by LG-BFGS by proving the linear convergence of the iterates generated by the corresponding greedy BFGS, alternatively.

We start by defining the concise notation $\lambda_t = \lambda_f(\mathbf{x}_t)$, $\phi_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\nabla^2 f(\mathbf{x}_t)}$ and

$$\eta_t = e^{2C_M \sum_{k=0}^{t-1} \lambda_k} \frac{L}{\mu} \tag{17}$$

for convenience of expression. We use **induction** to prove the following statement

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{B}_t \preceq \eta_t \nabla^2 f(\mathbf{x}_t), \tag{18}$$

$$\lambda_t \leq (1 - \frac{\mu}{2L})^t \lambda_0 \tag{19}$$

for any iteration $t \geq 0$. For the initial iteration $t = 0$ with the initial condition, we have

$$\nabla^2 f(\mathbf{x}_0) \preceq \mathbf{B}_0 \preceq \frac{L}{\mu} \nabla^2 f(\mathbf{x}_0) = \eta_0 \nabla^2 f(\mathbf{x}_0) \tag{20}$$

and

$$\lambda_0 \leq (1 - \frac{\mu}{2L})^0 \lambda_0 = \lambda_0. \tag{21}$$

Thus, equation 18 and equation 19 hold for $t = 0$.

Assume that for iteration $t - 1 \geq 0$, we have

$$\nabla^2 f(\mathbf{x}_k) \preceq \mathbf{B}_k \preceq \eta_k \nabla^2 f(\mathbf{x}_k), \tag{22}$$

$$\lambda_k \leq (1 - \frac{\mu}{2L})^k \lambda_0 \tag{23}$$

for all $0 \leq k \leq t - 1$, and consider iteration t. By using Lemma 2 with the condition equation 22, we have

$$\lambda_t \leq \Big(1 + \frac{\lambda_{t-1} C_M}{2}\Big) \frac{\eta_{t-1} - \Big(1 - \frac{\lambda_{t-1} C_M}{2}\Big)}{\eta_{t-1}} \lambda_{t-1}. \tag{24}$$

By using the fact $C_M \lambda_{t-1} \leq C_M \lambda_0 \leq 1$ from the initial condition and the inequality $1 - x \geq e^{-2t}$ for any $0 \leq x \leq 1/2$, we have

$$\frac{1 - \frac{\lambda_{t-1} C_M}{2}}{\eta_{t-1}} \geq \frac{e^{-\lambda_{t-1} C_M}}{\eta_{t-1}}. \tag{25}$$

By substituting the representation of η_{t-1} into equation 25, we get

$$\frac{1 - \frac{\lambda_{t-1} C_M}{2}}{\eta_{t-1}} \geq e^{-C_M \lambda_{t-1} - 2C_M \sum_{k=0}^{t-2} \lambda_k} \frac{\mu}{L} \geq e^{-2C_M \sum_{k=0}^{t-1} \lambda_k} \frac{\mu}{L}. \tag{26}$$

The term $2C_M \sum_{k=0}^{t-1} \lambda_k$ in equation 26 can be bounded as

$$2C_M \sum_{k=0}^{t-1} \lambda_i \leq 2C_M \sum_{k=0}^{t-1} (1 - \frac{\mu}{2L})^k \lambda_0 \leq \frac{4L}{\mu} C_M \lambda_0 \leq \ln \frac{3}{2} \tag{27}$$

where the condition equation 23 is used in the second inequality and the initial condition is used in the last inequality. By substituting equation 27 into equation 26, we have

$$\frac{1-\frac{\lambda_{t-1}C_M}{2}}{\eta_{t-1}} \geq \frac{2\mu}{3L}. \tag{28}$$

From the condition equation 23 and the initial condition, we get

$$\frac{\lambda_{t-1}C_M}{2} \leq \frac{\lambda_0 C_M}{2} \leq \frac{\ln\frac{3}{2}\,\mu}{8L} \leq \frac{\mu}{16L} \tag{29}$$

where the inequality $\ln(1+x) \leq x$ for any $x \geq 0$ is used in the last inequality. By substituting equation 29 and equation 28 into equation 24, we have

$$\lambda_t \leq \Big(1+\frac{\mu}{16L}\Big)\Big(1-\frac{2\mu}{3L}\Big)\lambda_{t-1} \leq \Big(1-\frac{\mu}{2L}\Big)\lambda_{t-1} \leq \Big(1-\frac{\mu}{2L}\Big)^t\lambda_0 \tag{30}$$

where the condition equation 23 is used in the last inequality. By using Lemma 3 with the condition equation 22, we have

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{B}_t \preceq (1+\phi_{t-1}C_M)^2\eta_{t-1}\nabla^2 f(\mathbf{x}_t). \tag{31}$$

By using the result $\phi_{t-1} \leq \lambda_{t-1}$ from Lemma 2 and the inequality $(1+x) \leq e^{2x}$, we get

$$\mathbf{B}_t \preceq (1+\lambda_{t-1}C_M)^2\eta_{t-1}\nabla^2 f(\mathbf{x}_t) \preceq e^{2C_M\lambda_{t-1}}\eta_{t-1}\nabla^2 f(\mathbf{x}_{t+1}). \tag{32}$$

By further substituting the representation of η_{t-1} into equation 32, we have

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{B}_t \preceq e^{2C_M\sum_{k=0}^{t-1}\lambda_k}\frac{L}{\mu}\nabla^2 f(\mathbf{x}_t) = \eta_t\nabla^2 f(\mathbf{x}_t). \tag{33}$$

By combining equation 20, equation 21, equation 30 and equation 33, we prove equation 18 and equation 19 by induction, which completes the proof. □

C Proof of Proposition 3

We need the following lemma to complete the proof.

Lemma 4 (Lemma 2.4 in (Rodomanov & Nesterov, 2021))**.** *Consider two positive definite matrices* $\mathbf{A} \preceq \mathbf{D}$*. For any vector* $\mathbf{s} \in \mathbb{R}^d$*, it holds that*

$$\sigma(\mathbf{A},\mathbf{D}) - \sigma(\mathbf{A},\mathrm{BFGS}(\mathbf{D},\mathbf{s},\mathbf{A}\mathbf{s})) \geq \frac{\mathbf{s}^\top(\mathbf{D}-\mathbf{A})\mathbf{s}}{\mathbf{s}^\top\mathbf{A}\mathbf{s}} \tag{34}$$

where $\mathrm{BFGS}(\mathbf{B},\mathbf{s},\mathbf{A}\mathbf{s})$ *is the BFGS update on* $\mathbf{B}$ *with the curvature pair* $\{\mathbf{s},\mathbf{A}\mathbf{s}\}$.

Proof of Proposition 3. From Lemma 3, we know that $\nabla^2 f(\mathbf{x}_+) \preceq \hat{\mathbf{B}}$. Let $\mathbf{B}_+ = \mathrm{BFGS}(\hat{\mathbf{B}},\mathbf{s},\mathbf{r})$ be the updated Hessian approximation matrix, where $\{\mathbf{s},\mathbf{r}\}$ are the curvature pair selected greedily from the subset $\{\mathbf{e}_i\}_{i=1}^\tau$, i.e.,

$$\mathbf{s} = \operatorname*{arg\,max}_{\mathbf{s}\in\{\mathbf{e}_1,\ldots,\mathbf{e}_\tau\}} \frac{\mathbf{s}^\top\hat{\mathbf{B}}\mathbf{s}}{\mathbf{s}^\top\nabla^2 f(\mathbf{x}_+)\mathbf{s}}. \tag{35}$$

Denote by $\sigma_{\mathbf{x}_+}(\mathbf{B}_+)$, $\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}})$ and $\sigma(\mathbf{B})$ the concise notation of $\sigma(\nabla^2 f(\mathbf{x}_+),\mathbf{B}_+)$, $\sigma(\nabla^2 f(\mathbf{x}_+),\hat{\mathbf{B}})$ and $\sigma(\nabla^2 f(\mathbf{x}),\mathbf{B})$. By using Lemma 4 with $\mathbf{A} = \nabla^2 f(\mathbf{x}_+)$ and $\mathbf{D} = \hat{\mathbf{B}}$, we have

$$\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) - \sigma_{\mathbf{x}_+}(\mathbf{B}_+) \geq \frac{\mathbf{s}^\top\big(\hat{\mathbf{B}}-\nabla^2 f(\mathbf{x}_+)\big)\mathbf{s}}{\mathbf{s}^\top\nabla^2 f(\mathbf{x}_+)\mathbf{s}}. \tag{36}$$

By substituting equation 35 into equation 36, we have

$$\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) - \sigma_{\mathbf{x}_+}(\mathbf{B}_+) \geq \max_{1\leq i\leq\tau} \frac{\mathbf{e}_i^\top\big(\hat{\mathbf{B}}-\nabla^2 f(\mathbf{x}_+)\big)\mathbf{e}_i}{\mathbf{e}_i^\top\nabla^2 f(\mathbf{x}_+)\mathbf{e}_i}. \tag{37}$$

Let $\mathbf{E} = \hat{\mathbf{B}} - \nabla^2 f(\mathbf{x}_+)$ be the approximation error matrix. From Assumption 1, we have

$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}_+) \preceq L\mathbf{I} \tag{38}$$

where $\mathbf{I}$ is the identity matrix. Substituting equation 38 into equation 37 yields

$$\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) - \sigma_{\mathbf{x}_+}(\mathbf{B}_+) \geq \frac{1}{L} \max_{1 \leq i \leq \tau} \mathbf{e}_i^\top \mathbf{E} \mathbf{e}_i. \tag{39}$$

Let $\beta(\mathbf{e}_i)$ be the relative condition number of the basis vector $\mathbf{e}_i$ w.r.t. $\mathbf{E}$ for $i = 1, \ldots, \tau$. From the definition of the relative condition number [Def. 1], we have

$$\mathbf{e}_i^\top \mathbf{E} \mathbf{e}_i = \frac{1}{\beta(\mathbf{e}_i)} \max_{1 \leq i \leq d} \mathbf{e}_i^\top \mathbf{E} \mathbf{e}_i. \tag{40}$$

By substituting equation 40 into equation 39, we have

$$\begin{aligned} \sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) - \sigma_{\mathbf{x}_+}(\mathbf{B}_+) &\geq \frac{1}{L} \max_{1 \leq i \leq \tau} \Big(\frac{1}{\beta(\mathbf{e}_i)} \max_{1 \leq i \leq d} \mathbf{e}_i^\top \mathbf{E} \mathbf{e}_i \Big) \\ &= \frac{1}{L \min_{1 \leq i \leq \tau} \beta(\mathbf{e}_i)} \max_{1 \leq i \leq d} \mathbf{e}_i^\top \mathbf{E} \mathbf{e}_i = \frac{1}{L\beta_\tau} \max_{1 \leq i \leq d} \mathbf{e}_i^\top \mathbf{E} \mathbf{e}_i \end{aligned} \tag{41}$$

where β_τ is the minimal relative condition number of the subset $\{\mathbf{e}_i\}_{i=1}^{\tau}$ w.r.t. $\mathbf{E}$. Since $\max_{1 \leq i \leq d} \mathbf{e}_i^\top \mathbf{E} \mathbf{e}_i \geq \mathbf{e}_i^\top \mathbf{E} \mathbf{e}_i$ for all $i = 1, \ldots, d$, we get

$$\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) - \sigma_{\mathbf{x}_+}(\mathbf{B}_+) \geq \frac{1}{L\beta_\tau d} \sum_{i=1}^{d} \mathbf{e}_i^\top \mathbf{E} \mathbf{e}_i = \frac{1}{L\beta_\tau d} \sum_{i=1}^{d} \mathrm{Tr}(\mathbf{e}_i \mathbf{e}_i^\top, \mathbf{E}) \tag{42}$$

where $\mathrm{Tr}(\cdot, \cdot)$ represents the trace operation. From the linearity of the trace operation, we get

$$\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) - \sigma_{\mathbf{x}_+}(\mathbf{B}_+) \geq \frac{1}{\beta_\tau L d} \mathrm{Tr}\Big(\sum_{i=1}^{d} \mathbf{e}_i \mathbf{e}_i^\top, \mathbf{E} \Big) = \frac{1}{\beta_\tau L d} \mathrm{Tr}(\mathbf{I}, \mathbf{E}) \geq \frac{\mu}{\beta_\tau L d} \mathrm{Tr}(\nabla^2 f(\mathbf{x}_+)^{-1}, \mathbf{E}) \tag{43}$$

where the condition equation 38 is used in the last inequality. From the definition $\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) = \mathrm{Tr}(\nabla^2 f(\mathbf{x}_+)^{-1}, \mathbf{E})$, we have

$$\sigma_{\mathbf{x}_+}(\mathbf{B}_+) \leq (1 - \frac{\mu}{\beta_\tau L d}) \sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}). \tag{44}$$

We then characterize the relationship between $\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}})$ and $\sigma_{\mathbf{x}}(\mathbf{B})$. We can represent $\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}})$ by definition as

$$\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) = \mathrm{Tr}\big(\nabla^2 f(\mathbf{x}_+)^{-1} \hat{\mathbf{B}}\big) - d = (1 + \phi C_M) \mathrm{Tr}\big(\nabla^2 f(\mathbf{x}_+)^{-1} \mathbf{B}\big) - d \tag{45}$$

where $\hat{\mathbf{B}} = (1 + \phi C_M)\mathbf{B}$ is used in the last equality. Since $1 + \phi C_M \geq 1$, we can upper bound equation 45 as

$$\sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) \leq (1 + \phi C_M)^2 \mathrm{Tr}\big(\nabla^2 f(\mathbf{x}_+)^{-1} \mathbf{B}\big) - d = (1 + \phi C_M)^2 (\sigma_{\mathbf{x}}(\mathbf{B}) + d) - d. \tag{46}$$

Expanding the terms in equation 46 yields

$$\begin{aligned} \sigma_{\mathbf{x}_+}(\hat{\mathbf{B}}) &\leq (1 + \phi C_M)^2 \sigma_{\mathbf{x}}(\mathbf{B}) + d\big((1 + \phi C_M)^2 - 1\big) \\ &= (1 + \phi C_M)^2 \sigma_{\mathbf{x}}(\mathbf{B}) + 2d\phi C_M \Big(1 + \frac{\phi C_M}{2}\Big) \\ &\leq (1 + \phi C_M)^2 \Big(\sigma_{\mathbf{x}}(\mathbf{B}) + \frac{2d\phi C_M}{1 + \phi C_M}\Big). \end{aligned} \tag{47}$$

By substituting equation 47 into equation 44, we complete the proof

$$\sigma_{\mathbf{x}_+}(\mathbf{B}_+) \leq \Big(1 - \frac{\mu}{\beta_\tau d L}\Big)(1 + \phi C_M)^2 \Big(\sigma_{\mathbf{x}}(\mathbf{B}) + \frac{2d\phi C_M}{1 + \phi C_M}\Big). \tag{48}$$

□

D PROOF OF THEOREM 1

From Lemma 1, we know that the iterates generated by LG-BFGS is equivalent to the iterates generated by greedy BFGS, if both perform greedy selection in the same subset $\{\mathbf{e}_i\}_{i=1}^{\tau}$ of memory size τ. In this context, we can prove the superlinear convergence of the iterates generated by LG-BFGS by proving the superlinear convergence of the iterates generated by the corresponding greedy BFGS, alternatively.

Denote by λ_t and σ_t the concise notation of $\lambda_f(\mathbf{x}_t)$ and $\sigma(\nabla^2 f(\mathbf{x}_t), \mathbf{B}_t)$. We start by noting that

$$\sigma_t = \mathrm{Tr}(\nabla^2 f(\mathbf{x}_t)^{-1}\mathbf{B}_t) - d = \mathrm{Tr}\big(\nabla^2 f(\mathbf{x}_t)^{-1}(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t)\big) \tag{49}$$

where $\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t)$ is positive semidefinite from Proposition 2 [cf. equation 18]. Since the maximal eigenvalue of $\nabla^2 f(\mathbf{x}_t)^{-1}\big(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t)\big)$ is bounded by the trace of $\nabla^2 f(\mathbf{x}_t)^{-1}(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t)$, i.e., the sum of eigenvalues of $\nabla^2 f(\mathbf{x}_t)^{-1}(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t)$, we have

$$\nabla^2 f(\mathbf{x}_t)^{-1}\big(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t)\big) \preceq \mathrm{Tr}\big(\nabla^2 f(\mathbf{x}_t)^{-1}(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t)\big)\mathbf{I}. \tag{50}$$

By multiplying positive definite matrix $\nabla^2 f(\mathbf{x}_t)$ on both sides of equation 50, we have

$$\mathbf{B}_t \preceq \Big(1 + \mathrm{Tr}\big(\nabla^2 f(\mathbf{x}_t)^{-1}(\mathbf{B}_t - \nabla^2 f(\mathbf{x}_t))\big)\Big)\nabla^2 f(\mathbf{x}_t) = (1+\sigma_t)\nabla^2 f(\mathbf{x}_t) \tag{51}$$

and

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{B}_t \preceq (1+\sigma_t)\nabla^2 f(\mathbf{x}_t). \tag{52}$$

By using Lemma 2 with the condition equation 52, we have

$$\lambda_{t+1} \leq \Big(1 + \frac{\lambda_t C_M}{2}\Big)\frac{\sigma_t + \frac{\lambda_t C_M}{2}}{1+\sigma_t}\lambda_t \leq \Big(1 + \frac{\lambda_t C_M}{2}\Big)\Big(\sigma_t + 2dC_M\lambda_t\Big)\lambda_t. \tag{53}$$

Consider the term $\sigma_t + 2dC_M\lambda_t$ in the bound of equation 53. We use **induction** to prove the following statement

$$\sigma_t + 2dC_M\lambda_t \leq \Phi_t \tag{54}$$

for any iteration $t \geq 0$, where

$$\Phi_t := \prod_{k=1}^{t}\big(1 - \frac{\mu}{\beta_{k,\tau}dL}\big)e^{2(2d+1)C_M\sum_{k=0}^{t-1}\lambda_k}\frac{dL}{\mu} \tag{55}$$

For the initial iteration $t = 0$, it holds that

$$\begin{aligned}\sigma_0 + 2dC_M\lambda_0 &= \mathrm{Tr}(\nabla^2 f(\mathbf{x}_0)^{-1}\mathbf{B}_0) - d + 2dC_M\lambda_0 \\ &\leq \frac{L}{\mu}\mathrm{Tr}\big(\nabla^2 f(\mathbf{x}_0)^{-1}\nabla^2 f(\mathbf{x}_0)\big) - d + 2dC_M\lambda_0 = \Big(\frac{L}{\mu} - 1\Big)d + 2dC_M\lambda_0\end{aligned} \tag{56}$$

where the initial condition of $\mathbf{B}_0$ is used in the second inequality. By substituting the initial condition of λ_0 into equation 56, we get

$$\sigma_0 + 2dC_M\lambda_0 \leq \Big(\frac{L}{\mu} - 1\Big)d + \frac{d\ln 2}{2(2d+1)} \leq \frac{dL}{\mu}. \tag{57}$$

Thus, equation 54 holds for $t = 0$. Assume that equation 54 holds for iteration $t \geq 0$, i.e.,

$$\sigma_t + 2dC_M\lambda_t \leq \Phi_t \tag{58}$$

and consider iteration $t+1$. By substituting equation 58 into equation 53 and using the inequality $1 + x \leq e^x$, we have

$$\lambda_{t+1} \leq \Big(1 + \frac{\lambda_t C_M}{2}\Big)\Phi_t\lambda_t \leq e^{\frac{\lambda_t C_M}{2}}\Phi_t\lambda_t \leq e^{2\lambda_t C_M}\Phi_t\lambda_t. \tag{59}$$

By using Lemma 2 that $\phi_t \leq \lambda_t$ and Proposition 3, we have

$$\begin{aligned}\sigma_{t+1} &\leq \Big(1-\frac{\mu}{\beta_{t+1,\tau}dL}\Big)(1+\phi_t C_M)^2\Big(\sigma_t+\frac{2d\phi_t C_M}{1+\phi_t C_M}\Big) \\ &\leq \Big(1-\frac{\mu}{\beta_{t+1,\tau}dL}\Big)(1+\lambda_t C_M)^2\Big(\sigma_t+\frac{2d\phi_t C_M}{1+\phi_t C_M}\Big) \\ &\leq \Big(1-\frac{\mu}{\beta_{t+1,\tau}dL}\Big)(1+\lambda_t C_M)^2\Big(\sigma_t+2d\phi_t C_M\Big) \leq \Big(1-\frac{\mu}{\beta_{t+1,\tau}dL}\Big)e^{2C_M\lambda_t}\Phi_t\end{aligned} \tag{60}$$

where $\beta_{t+1,\tau}$ is the minimal condition number of the subset $\{\mathbf{e}_i\}_{i=1}^{\tau}$ at iteration $t+1$ and the inequality $(1+x)^2 \leq e^{2x}$ is used in the last inequality. Since $1-\mu/(\beta_{t+1,\tau}dL) \geq 1/2$ with $d \geq 2$, combining equation 59 and equation 60 yields

$$\begin{aligned}\sigma_{t+1}+2dC_M\lambda_{t+1} &\leq (1-\frac{\mu}{\beta_{t+1,\tau}dL})e^{2C_M\lambda_t}\Phi_t + 2dC_M e^{2C_M\lambda_t}\Phi_t\lambda_t \\ &\leq (1-\frac{\mu}{\beta_{t+1,\tau}dL})e^{2C_M\lambda_t}\Phi_t + (1-\frac{\mu}{\beta_{t+1,\tau}dL})4dC_M e^{2C_M\lambda_t}\Phi_t\lambda_t \\ &= (1-\frac{\mu}{\beta_{t+1,\tau}dL})e^{2C_M\lambda_t}(1+4dC_M\lambda_t)\Phi_t.\end{aligned} \tag{61}$$

By using the inequality $1+x \leq e^x$ and substituting the representation of Φ_t into equation 61, we get

$$\sigma_{t+1}+2dC_M\lambda_{t+1} \leq (1-\frac{\mu}{\beta_{t+1,\tau}dL})e^{2C_M(2d+1)\lambda_t}\Phi_t = \Phi_{t+1}. \tag{62}$$

Thus, equation 54 holds for $t+1$. By combining equation 57, equation 58 and equation 62, we prove equation 54 by induction.

By substituting equation 54 and the representation of Φ_t into equation 59, we have

$$\lambda_{t+1} \leq e^{2\lambda_t C_M}\Phi_t\lambda_t \leq e^{2(2d+1)\lambda_t C_M}\Phi_t\lambda_t = \frac{1}{\big(1-\frac{\mu}{\beta_{t+1,\tau}dL}\big)}\Phi_{t+1}\lambda_t. \tag{63}$$

By using Proposition 2, we get

$$\begin{aligned}\Phi_{t+1} &= \prod_{k=1}^{t+1}\big(1-\frac{\mu}{\beta_{k,\tau}dL}\big)e^{2(2d+1)C_M\sum_{k=0}^{t}\lambda_k}\frac{dL}{\mu} \\ &\leq \prod_{k=1}^{t+1}\big(1-\frac{\mu}{\beta_{k,\tau}dL}\big)e^{2(2d+1)C_M\sum_{k=0}^{t}\left(1-\frac{\mu}{2L}\right)^k\lambda_0}\frac{dL}{\mu}\end{aligned} \tag{64}$$

From the fact that $\sum_{k=0}^{t}\left(1-\frac{\mu}{2L}\right)^k \leq 2L/\mu$ and the initial condition, we have

$$\Phi_{t+1} \leq \prod_{k=1}^{t+1}\big(1-\frac{\mu}{\beta_{k,\tau}dL}\big)e^{\ln 2}\frac{dL}{\mu} = \prod_{k=1}^{t+1}\big(1-\frac{\mu}{\beta_{k,\tau}dL}\big)\frac{2dL}{\mu}. \tag{65}$$

Substituting equation 65 into equation 63 yields

$$\lambda_{t+1} \leq \prod_{k=1}^{t}\big(1-\frac{\mu}{\beta_{k,\tau}dL}\big)\frac{2dL}{\mu}\lambda_t. \tag{66}$$

Let t_0 be such that

$$\prod_{k=1}^{t_0}\big(1-\frac{\mu}{\beta_{k,\tau}dL}\big)\frac{2dL}{\mu} \leq 1 \tag{67}$$

and we have

$$\lambda_{t+t_0+1} \leq \prod_{k=1}^{t+t_0}\big(1-\frac{\mu}{\beta_{k,\tau}dL}\big)\frac{2dL}{\mu}\lambda_{t+t_0} \leq \prod_{k=t_0+1}^{t+t_0}\big(1-\frac{\mu}{\beta_{k,\tau}dL}\big)\lambda_{t+t_0} \tag{68}$$

for any $t \geq 0$. By using equation 68 recursively, we get

$$\lambda_{t+t_0+1} \leq \prod_{k=t_0+1}^{t+t_0} \big(1-\frac{\mu}{\beta_{k,\tau} dL}\big)\lambda_{t+t_0} \leq \cdots \leq \prod_{k=t_0+1}^{t+t_0} \big(1-\frac{\mu}{\beta_{k,\tau} dL}\big)^{t+t_0+1-k}\lambda_{t_0}. \tag{69}$$

By further using Proposition 2, we complete the proof

$$\lambda_{t+t_0+1} \leq \prod_{k=t_0+1}^{t+t_0} \big(1-\frac{\mu}{\beta_{k,\tau} dL}\big)^{t+t_0+1-k}\Big(1-\frac{\mu}{2L}\Big)^{t_0}\lambda_0. \tag{70}$$

E Proof of Theorem 2

We start by noting that for any vector $\mathbf{e}_i$ and matrix $\mathbf{E}$, we have

$$\lambda_1 \leq \mathbf{e}_i^\top \mathbf{E}\mathbf{e}_i \leq \lambda_d \tag{71}$$

where λ_1 and λ_d are the minimal and maximal eigenvalues of $\mathbf{E}$. From Definition 1, the relative condition number of $\mathbf{e}_i$ w.r.t. $\mathbf{E}$ is bounded as

$$\beta(\mathbf{e}_i) = \frac{\max_{1\leq k\leq d} \mathbf{e}_k^\top \mathbf{E}\mathbf{e}_k}{\mathbf{e}_i^\top \mathbf{E}\mathbf{e}_i} \leq \frac{\lambda_d}{\lambda_1} = \beta \tag{72}$$

where β is the condition number of $\mathbf{E}$. By using equation 72 together with the corollary condition, we have

$$\beta_{t,\tau} \leq C_\beta \tag{73}$$

or

$$\beta_{t,\tau} = \min_{1\leq i\leq \tau} \beta(\mathbf{e}_i) \leq \beta_t \leq C_\beta \tag{74}$$

where β_t is the condition number of the approximation error matrix at iteration t. By further using equation 73 or equation 74 in the result of Theorem 1, we have

$$\lambda_{t+t_0+1} \leq \big(1-\frac{\mu}{C_\beta dL}\big)^{\frac{t(t+1)}{2}}\Big(1-\frac{\mu}{2L}\Big)^{t_0}\lambda_0 \tag{75}$$

which completes the proof.

F Bound on Condition Number β_t

We establish an upper bound on the condition number β_t of the error matrix $\hat{\mathbf{B}}_t - \nabla^2 f(\mathbf{x}_{t+1})$ with a minor modification in the correction strategy in Section 3.1. Specifically, we consider the correction strategy on the Hessian approximation $\mathbf{B}_t$ as

$$\hat{\mathbf{B}}_t = \big(1+(\phi_t C_M + \delta_t)\big)\mathbf{B}_t \tag{76}$$

where $\delta_t = q^t\delta_0 > 0$ with δ_0 a positive constant and $0 < q < 1$ a contraction factor. Since $(1+\phi_t C_M)\mathbf{B}_t \succeq \nabla^2 f(\mathbf{x}_{t+1})$ from equation 15 and $\mathbf{B}_t \succeq \nabla^2 f(\mathbf{x}_t)$ from equation 16 in Appendix B, we obtain

$$\hat{\mathbf{B}}_t - \nabla^2 f(\mathbf{x}_{t+1}) \succeq \delta_t \mathbf{B}_t \succeq \delta_t \nabla^2 f(\mathbf{x}_t). \tag{77}$$

By using Assumption 1 with $L\mathbf{I} \succeq \nabla^2 f(\mathbf{x}_t) \succeq \mu\mathbf{I}$ in equation 77, we get

$$\hat{\mathbf{B}}_t - \nabla^2 f(\mathbf{x}_{t+1}) \succeq \delta_t \mu \mathbf{I}. \tag{78}$$

From equation 18 in Appendix B, we have

$$\mathbf{B}_t \preceq \eta_t \nabla^2 f(\mathbf{x}_t) \tag{79}$$

where η_t is now defined as

$$\eta_t = e^{2C_M \sum_{k=0}^{t-1} \lambda_k + 2\sum_{k=0}^{t-1} \delta_t} \frac{L}{\mu} \tag{80}$$

with the modified correction strategy [cf. equation 76]. Since both $\{\lambda_t\}_t$ and $\{\delta_t\}_t$ are decreasing geometric sequences, η_t is upper bounded by a constant C_η. By using this fact in equation 79 and the latter in equation 76, we have

$$\hat{\mathbf{B}}_t \preceq \big(1+(\phi_t C_M + \delta_t)\big) C_\eta \nabla^2 f(\mathbf{x}_t) \preceq \big(1+(\lambda_0 C_M + \delta_0)\big) C_\eta L \mathbf{I} \tag{81}$$

where $\lambda_t \leq \lambda_0$, $\delta_t \leq \delta_0$ and $\mu\mathbf{I} \preceq \nabla^2 f(\mathbf{x}_t) \preceq L\mathbf{I}$ are used in the second inequality. By using equation 78 and equation 81, we can bound the condition number β_t of $\hat{\mathbf{B}}_t - \nabla^2 f(\mathbf{x}_{t+1})$ as

$$\beta_t \leq \frac{\big(2+(\lambda_0 C_M + \delta_0)\big) C_\eta L}{q^t \delta_0 \mu} = C_{t,\beta} \tag{82}$$

at iteration t, where λ_0 and δ_0 are initial constants. We remark that the upper bound in equation 82 is the worst-case analysis because it holds for the condition number β_t, i.e., the minimal relative condition number $\beta_{t,1}$ with memory size $\tau = 1$ [Def. 1]. Therefore, it is important to note that this bound may not be tight and the actual value of $\beta_{t,\tau}$ could be smaller.

By following similar steps as in the proofs of Proposition 2 - Theorem 2, we can establish an explicit convergence rate of LG-BFGS as

$$\lambda_f(\mathbf{x}_{t+t_0+1}) \leq \prod_{u=t_0+1}^{t+t_0} \Big(1 - \frac{\mu}{C_{u,\beta} dL}\Big)^{t+t_0+1-u} \Big(1 - \frac{\mu}{2L}\Big)^{t_0} \lambda_f(\mathbf{x}_0) \tag{83}$$

where the modified correction strategy may require a slightly more accurate initialization to derive this rate. Since this upper bound $C_{t,\beta}$ is not constant but increases with iteration t [cf. equation 82], the convergence rate in equation 83 is slower than that in Theorem 2. Specifically, we can represent $C_{t,\beta}$ in equation 82 as the form of

$$C_{t,\beta} = C_\beta q^{-t} \tag{84}$$

where C_β is a constant. By substituting equation 84 into equation 83, we get

$$\lambda_f(\mathbf{x}_{t+t_0+1}) \leq \prod_{u=t_0+1}^{t+t_0} \Big(1 - \frac{\mu}{C_\beta dL} q^u\Big)^{t+t_0+1-u} \Big(1 - \frac{\mu}{2L}\Big)^{t_0} \lambda_f(\mathbf{x}_0). \tag{85}$$

We can approximate equation 85 as

$$\begin{aligned} &\prod_{u=t_0+1}^{t+t_0} \Big(1 - \frac{\mu}{C_\beta dL} q^u\Big)^{t+t_0+1-u} \Big(1 - \frac{\mu}{2L}\Big)^{t_0} \lambda_f(\mathbf{x}_0) \\ &\approx e^{-\sum_{u=t_0+1}^{t+t_0} \frac{\mu}{C_\beta dL}(t+t_0+1-u)q^u} \Big(1 - \frac{\mu}{2L}\Big)^{t_0} \lambda_f(\mathbf{x}_0) \\ &= e^{-\frac{q^{t_0+1}\mu}{C_\beta dL} \sum_{u=0}^{t-1}(t-u)q^u} \Big(1 - \frac{\mu}{2L}\Big)^{t_0} \lambda_f(\mathbf{x}_0) \leq e^{-Ct} \Big(1 - \frac{\mu}{2L}\Big)^{t_0} \lambda_f(\mathbf{x}_0) \end{aligned} \tag{86}$$

where $C = q^{t_0+1}\mu/(C_\beta dL)$ is a constant. By combining equation 86 and the result in Proposition 2, we have

$$\lambda_f(\mathbf{x}_{t+t_0+1}) \leq \min\Big\{e^{-Ct}\Big(1 - \frac{\mu}{2L}\Big)^{t_0} \lambda_f(\mathbf{x}_0), \Big(1 - \frac{\mu}{2L}\Big)^{t+t_0+1} \lambda_f(\mathbf{x}_0)\Big\}. \tag{87}$$

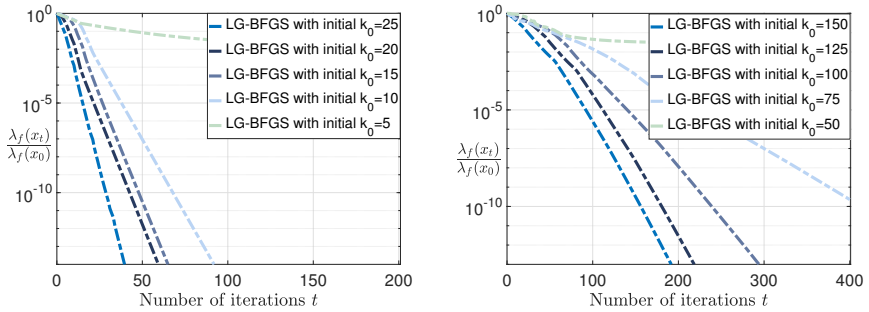
This can be considered as an improved linear rate depending on specific problem settings.

G Additional Experiments

We consider four datasets: svmguide3, connect-4, protein and mnist for classification problem, details of which are summarized in Table 1. With the local nature of superlinear convergence results

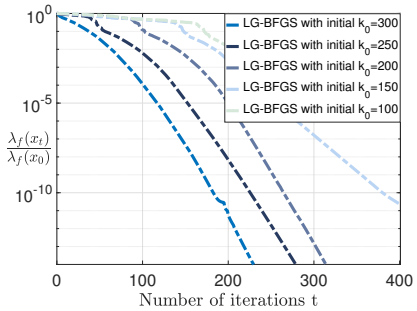
Table 1: Details of datasets: svmguide3, connect-4, protein and mnist.

Dataset	Number of samples N	Feature dimension d	Regularization parameters μ
Svmguide3	1243	21	10^{-4}
Connect-4	67557	126	10^{-4}
Protein	17766	357	10^{-4}
Mnist	60000	780	10^{-6}

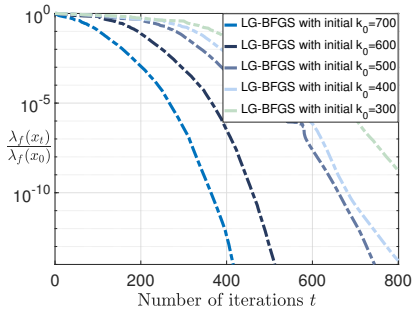


(a) Svmguide3 dataset

(b) Connect-4 dataset



(c) Protein dataset



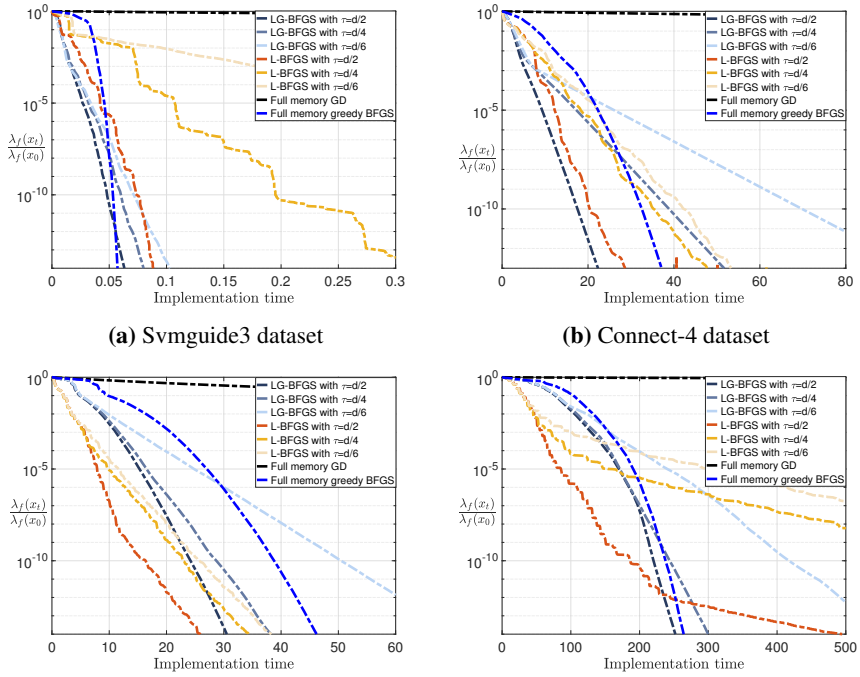
(d) Mnist dataset

Figure 1: Performance of LG-BFGS with different initialization on four datasets.

for quasi-Newton methods, we construct a setup with a warm start for all methods, i.e., the initialization is close to the solution by performing greedy BFGS for k_0 iterations. This has the practical effect of reducing the superlinear phase triggering time – see (Jin et al., 2022) for further details. Also worth mentioning is that we found it is better not to apply the correction strategy in LG-BFGS and greedy BFGS methods in practice following (Rodomanov & Nesterov, 2021; Lin et al., 2021; Jin et al., 2022), i.e., simply set $\tilde{\mathbf{r}}_u = \mathbf{r}_u$ in step 2 of Algorithm 1 for the displacement step of LG-BFGS and $\hat{\mathbf{B}}_t = \mathbf{B}_t$ in the Hessian approximation update of greedy BFGS.

Fig. 1 evaluates LG-BFGS with different initialization. We see that the performance of LG-BFGS increases with the improvement of initialization in all experiments. This relationship is expected because (i) the superlinear convergence of LG-BFGS is a local result; and (ii) the subset $\{\mathbf{e}_i\}_{i=1}^{\tau}$ being selected from good initialization roughly ensures the update progress of the Hessian approximation associated with the sparse subspace, i.e., the minimal relative condition number β_τ w.r.t. the approximation error matrix in (17) is small. These aspects manifest in the improved convergence of LG-BFGS, which corroborate our theoretical findings in Section 4.

Fig. 2 shows the convergence of LG-BFGS, L-BFGS, greedy BFGS and GD as a function of implementation time. For greedy BFGS, it has the fastest convergence rate (per iteration) but requires the most computational cost, which slows its convergence in datasets of connect-4, protein and mnist. For L-BFGS, it requires the lowest computational cost but has the slowest convergence rate, which exhibits bad performance with small memory sizes in datasets of svmguide3 and mnist. LG-BFGS strikes a balance between convergence rate of greedy BFGS and computational cost of L-BFGS,



(a) Svmguide3 dataset

(b) Connect-4 dataset

(c) Protein dataset

(d) Mnist dataset

Figure 2: Performance of LG-BFGS, L-BFGS, and greedy BFGS over implementation time on four datasets. We consider different memory sizes for LG-BFGS and L-BFGS.

i.e., it requires less computational cost than the former and obtains a faster convergence rate than the latter, corresponding to our discussions in Section 5. A final comment is that LG-BFGS and L-BFGS require less storage memory $\mathcal{O}(\tau d)$ than that required by greedy BFGS $\mathcal{O}(d^2)$.

REFERENCES

Albert S Berahas, Frank E Curtis, and Baoyu Zhou. Limited-memory bfgs with displacement aggregation. *Mathematical Programming*, 194(1-2):121–157, 2022.

Qiujiang Jin, Alec Koppel, Ketan Rajawat, and Aryan Mokhtari. Sharpened quasi-newton methods: Faster superlinear rate and larger local convergence neighborhood. In *International Conference on Machine Learning*, pp. 10228–10250. PMLR, 2022.

Dachao Lin, Haishan Ye, and Zhihua Zhang. Greedy and random quasi-newton methods with faster explicit superlinear convergence. *Advances in Neural Information Processing Systems*, 34:6646–6657, 2021.

Anton Rodomanov and Yurii Nesterov. Greedy quasi-newton methods with explicit superlinear convergence. *SIAM Journal on Optimization*, 31(1):785–811, 2021.