Greedy and Random Quasi-Newton Methods with Faster Explicit Superlinear Convergence

Dachao Lin[1*] **Haishan Ye**[2*] **Zhihua Zhang**[3]
[1]Academy for Advanced Interdisciplinary Studies, Peking University
[2]School of Management, Xi'an Jiaotong University; Shenzhen Research Institute of Big Data
[3]School of Mathematical Sciences, Peking University; Huawei Noah's Ark Lab
lindachao@pku.edu.cn yehaishan@xjtu.edu.cn zhzhang@math.pku.edu.cn.

Abstract

In this paper, we follow Rodomanov and Nesterov [19]'s work to study quasi-Newton methods. We focus on the common SR1 and BFGS quasi-Newton methods to establish better explicit (local) superlinear convergence rates. First, based on the *greedy quasi-Newton update* which greedily selects the direction to maximize a certain measure of progress, we improve the convergence rate to a condition-number-free superlinear convergence rate. Second, based on the *random quasi-Newton update* that selects the direction randomly from a spherically symmetric distribution, we show the same superlinear convergence rate established as above. Our analysis is closely related to the approximation of a given Hessian matrix, unconstrained quadratic objective, as well as the general strongly convex, smooth and strongly self-concordant functions.

1 Introduction

We study the superlinear convergence of famous quasi-Newton methods that replace the exact Hessian applied in classical Newton methods with certain approximations. The approximation is updated in iterations based on some special formulas from the previous variation. There exist various quasi-Newton algorithms with different Hessian approximations. The three most popular versions are the *Davidon-Fletcher-Powell (DFP) method* [7, 10], the *Broyden-Fletcher-Goldfarb-Shanno (BFGS) method* [2, 3, 9, 11, 21], and the *Symmetric Rank 1 (SR1) method* [1, 7], all of which belong to the Broyden family [1] of quasi-Newton algorithms. The most attractive property of quasi-Newton methods is their superlinear convergence, which can trace back to the 1970s [4, 8, 17]. However, the superlinear convergence rates provided in prior work are asymptotic [5, 13, 15, 22, 24]. The results only show that the ratio of successive residuals tends to zero as the running iterations approach to infinity, while still lacking a specific superlinear convergence rate.

Recently, Rodomanov and Nesterov [19] gave the first explicit local superlinear convergence for their proposed new quasi-Newton methods. They introduced greedy quasi-Newton updates by greedily selecting from basis vectors to maximize a certain measure of progress, and established an explicit non-asymptotic bound on the local superlinear convergence rate correspondingly. Their proofs are mainly applicable to the DFP methods because they reduced all possible Broyden family to the DFP update based on the monotonicity property (see Lemma 2.2). However, the SR1 and BFGS updates are more popular and faster than the DFP update in practice, which also has been verified in their experiments. In addition, Rodomanov and Nesterov [19] also discovered that the randomized methods do not lose superlinear convergence but they did not provide theoretical guarantees.

[*] Equal Contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Greedy or Random Methods	Strongly Self-concordant Objective	k_0
DFP, BFGS, SR1 [19]	$\left(1-\frac{1}{n\kappa}\right)^{k(k-1)/2}\left(\frac{1}{2}\right)^k\left(1-\frac{1}{2\kappa}\right)^{k_0}$	$O\left(n\kappa\ln(n\kappa)\right)$
BFGS, SR1 (Corollary 4.4)	$\left(1-\frac{1}{n}\right)^{k(k-1)/2}\left(\frac{1}{2}\right)^k\left(1-\frac{1}{2\kappa}\right)^{k_0}$	$O\left(\max\{n,\kappa\}\ln(n\kappa)\right)$

Table 1: Comparison of the existing specific superlinear convergence rates of the random or greedy quasi-Newton methods with our results in the view of $\lambda_f(\cdot)$ (see the definition in Eq. (17)), where n is the dimension of parameters, κ is the condition number of the objective function, and k_0, k are the iteration numbers. Rodomanov and Nesterov [19]'s results are suitable for greedy quasi-Newton methods using the Broyden family update, which includes DFP, BFGS, SR1 methods.

Along this line, there are other results of local superlinear convergence analysis. Rodomanov and Nesterov [20] analyzed the well-known DFP and BFGS methods, using a standard Hessian update direction through the previous variation. They demonstrated faster initial convergence rates, while slower final rates compared to Rodomanov and Nesterov [19]'s results. Rodomanov and Nesterov [18] improved Rodomanov and Nesterov [20]'s results by reducing the dependence of the condition number κ to $\ln\kappa$, though having similar worse long-history behavior. Jin and Mokhtari [14] provided a non-asymptotic dimension-free superlinear convergence rate of the original Broyden family when the initial Hessian approximation is also good enough.

In this work, we prove faster convergence rates for greedy and random quasi-Newton algorithms, particularly on the SR1 and BFGS methods. We present our contribution as follows:

- First, in the setting of approximating an exact Hessian, we use two update methods: 1) *greedy update* modified from Rodomanov and Nesterov [19]; 2) *random update* that randomly selects the updating direction from any spherically symmetric distribution. We show that both methods share a faster condition-number-free convergence. Particularly, we obtain the superlinear convergence rate $O\left(1-\frac{k}{n}\right)$ for SR1 update, and the linear convergence rate $O((1-\frac{1}{n})^k)$ for BFGS update, where k is the current iteration, and n is the dimension of parameters. Both findings improve the original convergence rate $O((1-\frac{1}{n\kappa})^k)$ [19].
- Second, we extend our analysis to a practical scheme. We show (local) superlinear convergence under the SR1 update and BFGS update, when applied to unconstrained quadratic objective or strongly self-concordant functions. We list our results in Table 1 with the same formulation of [19]. Note that in general, the convergence goes through two periods. The first period lasts for k_0 iterations, and only has a linear convergence rate $O((1-\frac{1}{2\kappa})^{k_0})$. The second period has a superlinear convergence rate $O((1-\frac{1}{n})^{k(k-1)/2})$. Our revised bound takes fewer first-period iterations k_0 as well as a faster (condition-number-free) superlinear convergence rate in the second period compared to [19].

Notation. We denote vectors by lowercase bold letters (e.g., $\boldsymbol{u}, \boldsymbol{x}$), and matrices by capital bold letters (e.g., $\boldsymbol{W} = [w_{ij}]$). We use $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ for the n-dimensional coordinate directions. Let $\lambda_1(\boldsymbol{A}) \geq \cdots \geq \lambda_n(\boldsymbol{A})$ be the eigenvalues of a real symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Moreover, $\|\cdot\|$ denotes the ℓ_2-norm (standard Euclidean norm). For a given positive definite matrix $\boldsymbol{A}$ (i.e., $\boldsymbol{A} \succ 0$), we induce the following pair of conjugate Euclidean norms: $\|\boldsymbol{x}\|_{\boldsymbol{A}} \triangleq \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$, $\|\boldsymbol{x}\|_{\boldsymbol{A}}^* \triangleq \sqrt{\boldsymbol{x}^\top \boldsymbol{A}^{-1}\boldsymbol{x}}$. When $\boldsymbol{A} = \nabla^2 f(\boldsymbol{x}) \succ 0$, we prefer to use notation $\|\cdot\|_{\boldsymbol{x}}$ and $\|\cdot\|_{\boldsymbol{x}}^*$, provided that there is no ambiguity with the reference function f. We recall the rate of convergence used in this paper.

Definition 1.1 (Linear/Superlinear convergence) *Suppose a sequence $\{x_n\}$ converges to 0 with*

$$\lim_{n\to\infty}\frac{|x_{n+1}|}{|x_n|} = q \in [0,1).$$

Now suppose another sequence $\{y_n\}$ converges to y^* and satisfies that $|y_n - y^*| \leq |x_n|, \forall n \geq 0$. We say $\{y_n\}$ converges superlinearly if $q = 0$, linearly if $q \in (0,1)$.

Organization. In Section 2, we discuss quasi-Newton updating rules for approximating a positive definite Hessian matrix using the SR1 update (Subsection 2.1) and the BFGS update (Subsection

2.2). In Section 3, we analyze previous methods applied to the problem of minimizing a quadratic function. In Section 4, we show that similar results also hold in a more general setting of minimizing the strongly self-concordant function. We give some empirical results in Section 5. Finally, in Section 6, we present our conclusion.

2 Quasi-Newton updates

Before starting our theoretical results, we briefly review a class of quasi-Newton updating rules for approximating a positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. We follow the definition in [19], employing the following family of updates which describes the Broyden family [16, Section 6.3] of quasi-Newton updates, parameterized by a scalar $\tau \in \mathbb{R}$.

Definition 2.1 *Let $\boldsymbol{A} \preceq \boldsymbol{G}$ (that is, $\boldsymbol{G} - \boldsymbol{A}$ is positive semi-definite). For any $\boldsymbol{u} \in \mathbb{R}^n$, if $\boldsymbol{G}\boldsymbol{u} = \boldsymbol{A}\boldsymbol{u}$, define* $\mathrm{Broyd}_\tau(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \triangleq \boldsymbol{G}$. *Otherwise, we define*

$$\begin{aligned}\mathrm{Broyd}_\tau(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \triangleq \tau &\left[\boldsymbol{G} - \frac{\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{G} + \boldsymbol{G}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{A}}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}} + \left(\frac{\boldsymbol{u}^\top\boldsymbol{G}\boldsymbol{u}}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}} + 1\right)\frac{\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{A}}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}}\right] \\ &+ (1-\tau)\left[\boldsymbol{G} - \frac{(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u}\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})}{\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u}}\right].\end{aligned} \tag{1}$$

For several choices of τ, we can recover several well-known quasi-Newton methods.

For $\tau = 0$, Eq. (1) corresponds to the well-known SR1 update:

$$\mathrm{SR1}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \triangleq \boldsymbol{G} - \frac{(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u}\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})}{\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u}}. \tag{2}$$

For $\tau_{\mathrm{BFGS}} \triangleq \frac{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}}{\boldsymbol{u}^\top\boldsymbol{G}\boldsymbol{u}} \in [0,1]$, we recover the famous BFGS update:

$$\mathrm{BFGS}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \triangleq \boldsymbol{G} - \frac{\boldsymbol{G}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{G}}{\boldsymbol{u}^\top\boldsymbol{G}\boldsymbol{u}} + \frac{\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{A}}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}}. \tag{3}$$

For $\tau = 1$, it corresponds to the well-known DFP update:

$$\mathrm{DFP}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \triangleq \boldsymbol{G} - \frac{\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{G} + \boldsymbol{G}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{A}}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}} + \left(\frac{\boldsymbol{u}^\top\boldsymbol{G}\boldsymbol{u}}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}} + 1\right)\frac{\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{A}}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}}. \tag{4}$$

Additionally, the Broyden family has matrix monotonicity below, showing the relationship among these quasi-Newton methods.

Lemma 2.2 (Rodomanov and Nesterov [19] Lemmas 2.1 and 2.2) *If $\boldsymbol{A} \preceq \boldsymbol{G} \preceq \eta\boldsymbol{A}$ for some $\eta \geq 1$, then we have for any $\boldsymbol{u} \in \mathbb{R}^n$, and $\tau_1, \tau_2 \in \mathbb{R}$ with $\tau_1 \leq \tau_2$ that*

$$\mathrm{Broyd}_{\tau_1}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \preceq \mathrm{Broyd}_{\tau_2}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}).$$

And for any $\tau \in [0,1]$, we have $\boldsymbol{A} \preceq \mathrm{Broyd}_\tau(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \preceq \eta\boldsymbol{A}$.

Hence, if $\boldsymbol{A} \preceq \boldsymbol{G} \preceq \eta\boldsymbol{A}$ for some $\eta \geq 1$, it follows from Lemma 2.2 that

$$\boldsymbol{A} \preceq \mathrm{SR1}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \preceq \mathrm{BFGS}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \preceq \mathrm{DFP}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \preceq \eta\boldsymbol{A}.$$

Intuitively, the approximation produced by SR1 is better than that produced by BFGS. And both of them are better than that produced by DFP. However, Rodomanov and Nesterov [19] deduced the analysis by casting all updates described by Broyden family ($\tau \in [0,1]$) into the slowest DFP update ($\tau = 1$). Moreover, SR1 and BFGS methods also have faster numerical performance in practice. Therefore, we consider there might have faster superlinear convergence rates of SR1 and BFGS methods.

Now we consider the fundamental matrix approximation problem below to obtain the convergence rate of the SR1 and BFGS updates.

2.1 Superlinear convergence for SR1 update

We first describe the SR1 update for approximating a fixed positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and leave the proof in Appendix A. Let us now justify the efficiency of update Eq. (2) in ensuring convergence $\boldsymbol{G}$ to $\boldsymbol{A}$. For this, we introduce the following measure of progress:

$$\tau_{\boldsymbol{A}}(\boldsymbol{G}) \triangleq \operatorname{tr}(\boldsymbol{G}-\boldsymbol{A}). \tag{5}$$

Previous work [19] uses a different measure as follows

$$\sigma_{\boldsymbol{A}}(\boldsymbol{G}) \triangleq \langle \boldsymbol{A}^{-1}, \boldsymbol{G}\rangle - n = \operatorname{tr}\left[(\boldsymbol{G}-\boldsymbol{A})\,\boldsymbol{A}^{-1}\right], \tag{6}$$

which might be improper for analyzing the SR1 update. Thus we employ a more concise measure $\tau_{\boldsymbol{A}}(\cdot)$. However, we will reuse $\sigma_{\boldsymbol{A}}(\cdot)$ in the proof of BFGS update in Subsection 2.2.

According to $\tau_{\boldsymbol{A}}(\boldsymbol{G})$, one iteration update leads to

$$\tau_{\boldsymbol{A}}(\boldsymbol{G}_+) \overset{(5)}{=} \tau_{\boldsymbol{A}}(\boldsymbol{G}) - \frac{\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})^2\boldsymbol{u}}{\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u}}, \quad \boldsymbol{G}_+ = \mathrm{SR1}(\boldsymbol{G},\boldsymbol{A},\boldsymbol{u}).$$

Note that we always have $\boldsymbol{G}_+ \succeq \boldsymbol{A}$ if $\boldsymbol{G} \succeq \boldsymbol{A}$ from Lemma 2.2. Moreover, the choice of the updating direction $\boldsymbol{u}$ directly influences the decrease in the measure $\tau_{\boldsymbol{A}}(\cdot)$. In the following, we show two efficient ways for selecting $\boldsymbol{u}$ introduced in Algorithm 1.

First, we use the *greedy method* introduced in [19], that greedily selects $\boldsymbol{u}$ from the basis vectors to obtain the largest decrease of $\tau_{\boldsymbol{A}}(\boldsymbol{G}_+) - \tau_{\boldsymbol{A}}(\boldsymbol{G})$: $\bar{\boldsymbol{u}}_{\boldsymbol{A}}(\boldsymbol{G}) \triangleq \arg\max_{\boldsymbol{u}\in\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n\}} \frac{\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})^2\boldsymbol{u}}{\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u}}$. However, we may encounter numerical overflow due to divided by 0 if $\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u}$ is nearly 0. Noting that $\boldsymbol{G} \succeq \boldsymbol{A}$, we have $\sqrt{(\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})^2\boldsymbol{u})\,(\boldsymbol{u}^\top\boldsymbol{u})} \geq \boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u} \geq 0,\ \forall \boldsymbol{u} \in \mathbb{R}^n$ from Cauchy–Schwarz inequality. Thus we employ a safer adjustment:

$$(\text{Greedy SR1})\ \ \bar{\boldsymbol{u}}_{\boldsymbol{A}}(\boldsymbol{G}) \triangleq \underset{\boldsymbol{u}\in\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n\}}{\arg\max} \frac{\boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u}}{\boldsymbol{u}^\top\boldsymbol{u}} = \underset{\boldsymbol{u}\in\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n\}}{\arg\max} \boldsymbol{u}^\top(\boldsymbol{G}-\boldsymbol{A})\boldsymbol{u}. \tag{7}$$

Moreover, we only need to obtain the diagonal elements of $\boldsymbol{A}$ (the current Hessian in practice), thus the total complexity is $O(n^2)$ in each iteration[2], which is acceptable and the same as the classical quasi-Newton methods.

Second, from the proof of the *greedy method*, we discover that the *random method* by choosing $\boldsymbol{u}$ from a spherically symmetric distribution, e.g.,

$$(\text{Random SR1})\ \ \boldsymbol{u} \sim \mathcal{N}(0, I_n) \text{ or } \boldsymbol{u} \sim \mathrm{Unif}(\mathcal{S}^{n-1}), \tag{8}$$

also has similar performance (in expectation) and the same running complexity $O(n^2)$ in each iteration.

Rodomanov and Nesterov [19, Theorem 3.5] already showed that $\boldsymbol{G}_n = \boldsymbol{A}$ if the $\boldsymbol{u}_k$ for $k \in \{0,\ldots,n-1\}$ are linearly independent. Thus, we only consider $k \leq n$. Now let us estimate the decrease in the measure $\tau_{\boldsymbol{A}}(\cdot)$ based on the direction $\boldsymbol{u}$ chosen from previous strategies. In the following, the expectation considers all the randomness of the directions $\boldsymbol{u}_k$'s during iterations, and when applied to the *greedy method*, we can view it with no randomness for the same notation.

Theorem 2.3 *Under the SR1 update in Algorithm 1, we obtain that for the greedy method defined in Eq.* (7) *or the random method defined in Eq.* (8),

$$0 \leq \mathbb{E}\,\tau_{\boldsymbol{A}}(\boldsymbol{G}_k) \leq \left(1 - \frac{1}{n-k+1}\right)\mathbb{E}\,\tau_{\boldsymbol{A}}(\boldsymbol{G}_{k-1}) \leq \left(1-\frac{k}{n}\right)\tau_{\boldsymbol{A}}(\boldsymbol{G}_0),\ 1 \leq k \leq n. \tag{9}$$

Hence, $\tau_{\boldsymbol{A}}(\boldsymbol{G}_k)$ (or $\mathbb{E}\,\tau_{\boldsymbol{A}}(\boldsymbol{G}_k)$) converges to zero superlinearly.

2.2 Linear convergence for BFGS update

We now consider the classical BFGS update in the same scheme and leave the proof in Appendix A. Using the measure $\sigma_{\boldsymbol{A}}(\cdot)$, we obtain that

$$\sigma_{\boldsymbol{A}}(\boldsymbol{G}_+) = \sigma_{\boldsymbol{A}}(\boldsymbol{G}) - \frac{\boldsymbol{u}^\top\boldsymbol{G}\boldsymbol{A}^{-1}\boldsymbol{G}\boldsymbol{u}}{\boldsymbol{u}^\top\boldsymbol{G}\boldsymbol{u}} + 1,\ \boldsymbol{G}_+ = \mathrm{BFGS}(\boldsymbol{G},\boldsymbol{A},\boldsymbol{u}). \tag{10}$$

[2] Note that we can use Hessian-vector product to obtain $\boldsymbol{A}\boldsymbol{u}$ in practice.

Algorithm 1 Greedy/Random SR1 update

```
Initialization: Set $G_0 \succeq A$.
for $k = 0, \ldots, n-1$ do
    Choose $u_k$ from
    1) greedy method: $u_k = \bar{u}_A(G_k)$, or
    2) random method: $u_k \sim \mathrm{Unif}(\mathcal{S}^{n-1})$.
    Compute $G_{k+1} = \mathrm{SR1}(G_k, A, u_k)$.
end for
```

Algorithm 2 Greedy/Random BFGS update

```
Initialization: Set $G_0 \succeq A$, $L_0 = G_0^{-1/2}$.
for $k \geq 0$ do
    Compute $u_k = L_k^\top \tilde{u}_k$ with $\tilde{u}_k$ from
    1) greedy method: $\tilde{u}_k = \tilde{u}_A(L_k)$, or
    2) random method: $\tilde{u}_k \sim \mathrm{Unif}(\mathcal{S}^{n-1})$.
    Compute $G_{k+1} = \mathrm{BFGS}(G_k, A, u_k)$.
    Compute $L_{k+1}$ based on Proposition 2.6.
end for
```

If we directly apply the *greedy method* or *random method* from the previous content, we could only obtain the same linear convergence rate in [19, Theorem 2.5]. However, if we take advantage of the current $\boldsymbol{G}$, and choose a scaled direction such that $\boldsymbol{u} = \boldsymbol{L}^\top \tilde{\boldsymbol{u}}$ where $\boldsymbol{L}^\top \boldsymbol{L} = \boldsymbol{G}^{-1}$, we could simplify the formulation and obtain a faster condition-number-free linear convergence rate. Specifically, the greedy update after replacing $\boldsymbol{u}$ to $\boldsymbol{L}^\top \tilde{\boldsymbol{u}}$ is as follows:

$$\text{(Greedy BFGS)} \;\; \tilde{\boldsymbol{u}}_{\boldsymbol{A}}(\boldsymbol{L}) = \operatorname*{arg\,max}_{\tilde{\boldsymbol{u}} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}} \frac{\tilde{\boldsymbol{u}}^\top \boldsymbol{L}^{-\top} \boldsymbol{A}^{-1} \boldsymbol{L}^{-1} \tilde{\boldsymbol{u}}}{\tilde{\boldsymbol{u}}^\top \tilde{\boldsymbol{u}}}, \tag{11}$$

and the random method is the same as the SR1 update used in Eq. (8).

$$\text{(Random BFGS)} \;\; \tilde{\boldsymbol{u}} \sim \mathcal{N}(0, I_n) \text{ or } \tilde{\boldsymbol{u}} \sim \mathrm{Unif}(\mathcal{S}^{n-1}), \tag{12}$$

Now we give the linear convergence rate of our modified BFGS update.

Theorem 2.4 *Under the BFGS update in Algorithm 2, we can obtain that for the greedy method defined in Eq.* (11) *or the random method defined in Eq.* (12)*,*

$$0 \leq \mathbb{E}\, \sigma_A(\boldsymbol{G}_k) \leq \left(1 - \frac{1}{n}\right) \mathbb{E}\, \sigma_A(\boldsymbol{G}_{k-1}) \leq \left(1 - \frac{1}{n}\right)^k \sigma_A(\boldsymbol{G}_0), \; 1 \leq k. \tag{13}$$

Therefore, $\sigma_A(\boldsymbol{G}_k)$ (or $\mathbb{E}\, \sigma_A(\boldsymbol{G}_k)$) converges to zero linearly.

Remark 2.5 *Note that the complexity in Eq.* (11) *is $O(n^3)$ because we have multiply-add operations with (unknown) $\boldsymbol{A}^{-1}$. Hence we **do not** apply this greedy strategy in practice, but view it as a theoretical result similar to the random strategy. Moreover, the random method is still practical, and we show the efficiency of our scaled direction compared to the original direction in experiments.*

Finally, we can employ an efficient way (with complexity $O(n^2)$) for updating $\boldsymbol{L}_k$. The main idea is employing the rank-one update and the one-row appending update of QR decomposition with the inverse BFGS update. In Proposition 2.6, we show how to compute $\boldsymbol{L}_{k+1}$ from $\boldsymbol{L}_k$ with $O(n^2)$ flops.

Proposition 2.6 *Suppose we already have $\boldsymbol{H}_k \triangleq \boldsymbol{G}_k^{-1} = \boldsymbol{L}_k^\top \boldsymbol{L}_k$, where $\boldsymbol{L}_k$ is an upper triangular matrix. Now we construct $\boldsymbol{L}_{k+1}$ with $O(n^2)$ flops.*

Step 1: Using the formulation in Golub and Van Loan [12, Section 12.5.1], we can obtain QR decomposition of

$$\boldsymbol{L}_k \left(\boldsymbol{I}_k - \frac{\boldsymbol{A}\boldsymbol{u}_k \boldsymbol{u}_k^\top}{\boldsymbol{u}_k^\top \boldsymbol{A} \boldsymbol{u}_k} \right) = \boldsymbol{L}_k - \frac{\boldsymbol{L}_k(\boldsymbol{A}\boldsymbol{u}_k)}{\boldsymbol{u}_k^\top \boldsymbol{A} \boldsymbol{u}_k} \boldsymbol{u}_k^\top$$

with $O(n^2)$ flops because it is a rank-one change of $\boldsymbol{L}_k$.

Step 2: We have $\boldsymbol{L}_k \left(\boldsymbol{I}_n - \frac{\boldsymbol{A}\boldsymbol{u}_k \boldsymbol{u}_k^\top}{\boldsymbol{u}_k^\top \boldsymbol{A} \boldsymbol{u}_k} \right) = \boldsymbol{Q}_k \boldsymbol{R}_k$ from Step 1, with an orthogonal matrix $\boldsymbol{Q}_k \in \mathbb{R}^{n \times n}$ and an upper triangular matrix $\boldsymbol{R}_k \in \mathbb{R}^{n \times n}$. Denoting $\boldsymbol{v}_k = \frac{\boldsymbol{u}_k}{\sqrt{\boldsymbol{u}_k^\top \boldsymbol{A} \boldsymbol{u}_k}}$, we can write

$$\boldsymbol{H}_{k+1} \overset{(16)}{=} \boldsymbol{R}_k^\top \boldsymbol{Q}_k^\top \boldsymbol{Q}_k \boldsymbol{R}_k + \frac{\boldsymbol{u}_k \boldsymbol{u}_k^\top}{\boldsymbol{u}_k^\top \boldsymbol{A} \boldsymbol{u}_k} = \boldsymbol{R}_k^\top \boldsymbol{R}_k + \boldsymbol{v}_k \boldsymbol{v}_k^\top = \begin{pmatrix} \boldsymbol{v}_k & \boldsymbol{R}_k^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_k^\top \\ \boldsymbol{R}_k \end{pmatrix}.$$

Using the formulation in Golub and Van Loan [12, Section 12.5.3] for computing the QR decomposition after appending a row, we can obtain $\begin{pmatrix} \boldsymbol{v}_k^\top \\ \boldsymbol{R}_k \end{pmatrix} = \boldsymbol{Q}_{k+1} \boldsymbol{R}_{k+1}$ with only $O(n^2)$ flops, where $\boldsymbol{Q}_{k+1} \in \mathbb{R}^{(n+1) \times n}$ is a column orthogonal matrix and $\boldsymbol{R}_{k+1} \in \mathbb{R}^{n \times n}$ is an upper triangular matrix. This implies $\boldsymbol{H}_{k+1} = \boldsymbol{R}_{k+1}^\top \boldsymbol{R}_{k+1}$, which satisfies our requirements.

Algorithm 3 Greedy/Random SR1/BFGS methods for quadratic minimization

```
Initialization: Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$. Set $\boldsymbol{G}_0 = L\boldsymbol{I}_n$ (or any $\boldsymbol{G}_0 \succeq \boldsymbol{A}$).
for $k \geq 0$ do
    Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{G}_k^{-1}\nabla f(\boldsymbol{x}_k)$. Choose one of the following update rules:
    (i) SR1: Choose $\boldsymbol{u}_k$ following Algorithm 1. Compute $\boldsymbol{G}_{k+1} = \mathrm{SR1}(\boldsymbol{G}_k, \boldsymbol{A}, \boldsymbol{u}_k)$.
    (ii) BFGS: Choose $\boldsymbol{u}_k$ following Algorithm 2. Compute $\boldsymbol{G}_{k+1} = \mathrm{BFGS}(\boldsymbol{G}_k, \boldsymbol{A}, \boldsymbol{u}_k)$.
end for
```

3 Unconstrained quadratic minimization

Based on the efficiency of the greedy/random SR1 and BFGS updates in matrix approximation, we next turn to minimize the quadratic function (with a fixed Hessian):

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}, \tag{14}$$

where $\boldsymbol{A}$ is a positive definite matrix, and there exist $\mu, L > 0$, s.t., $\mu\boldsymbol{I}_n \preceq \boldsymbol{A} \preceq L\boldsymbol{I}_n$. We show the detail in Algorithm 3, which is only for theoretical analysis. In practice, we use the inverse update rules (Nocedal and Wright [16, Eq. (6.17) and Eq. (6.25)]) to update $\boldsymbol{G}_k^{-1}$ directly:

$$\boldsymbol{G}_+^{-1} = \boldsymbol{G}^{-1} + \frac{(\boldsymbol{I}_n - \boldsymbol{G}^{-1}\boldsymbol{A})\boldsymbol{u}\boldsymbol{u}^\top(\boldsymbol{I}_n - \boldsymbol{A}\boldsymbol{G}^{-1})}{\boldsymbol{u}^\top(\boldsymbol{A} - \boldsymbol{A}\boldsymbol{G}^{-1}\boldsymbol{A})\boldsymbol{u}}, \quad \boldsymbol{G}_+ = \mathrm{SR1}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}); \tag{15}$$

$$\boldsymbol{G}_+^{-1} = \left(\boldsymbol{I}_n - \frac{\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{A}}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}}\right)\boldsymbol{G}^{-1}\left(\boldsymbol{I}_n - \frac{\boldsymbol{A}\boldsymbol{u}\boldsymbol{u}^\top}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}}\right) + \frac{\boldsymbol{u}\boldsymbol{u}^\top}{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}}, \quad \boldsymbol{G}_+ = \mathrm{BFGS}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}). \tag{16}$$

To estimate the convergence rate of Scheme (14), we measure the norm of the gradient of f as

$$\lambda_f(\boldsymbol{x}) \triangleq \sqrt{\nabla f(\boldsymbol{x})^\top \nabla^2 f(\boldsymbol{x})^{-1}\nabla f(\boldsymbol{x})}, \ \boldsymbol{x} \in \mathbb{R}^n. \tag{17}$$

Note that this measure of optimality is directly related to the functional residual. Indeed, let $\boldsymbol{x}^* = \boldsymbol{A}^{-1}\boldsymbol{b}$ be the minimizer of Eq. (14). Then we obtain

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) = \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}^*\right)^\top \boldsymbol{A}\left(\boldsymbol{x} - \boldsymbol{x}^*\right) = \frac{1}{2}\left(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\right)^\top \boldsymbol{A}^{-1}\left(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\right) = \frac{1}{2}\lambda_f(\boldsymbol{x})^2.$$

The following lemma shows how $\lambda_f(\cdot)$ varies after one iteration of process in Algorithm 3.

Lemma 3.1 (Rodomanov and Nesterov [19] Lemma 3.2) *Let $k \geq 0$, and let $\eta_k \geq 1$ satisfy $\boldsymbol{A} \preceq \boldsymbol{G}_k \preceq \eta_k\boldsymbol{A}$. Then we have that $\lambda_f(\boldsymbol{x}_{k+1}) \leq \left(1 - \frac{1}{\eta_k}\right)\lambda_f(\boldsymbol{x}_k) \leq (\eta_k - 1)\lambda_f(\boldsymbol{x}_k)$.*

Thus, to estimate how fast $\lambda_f(\boldsymbol{x}_k)$ converges to 0, we need the upper bound of η_k, which is already done in Subsections 2.1 and 2.2. Thus, we can guarantee a superlinear convergence of $\lambda_f(\boldsymbol{x}_k)$ using the greedy/random SR1 or BFGS update. The proof of Theorem 3.2 can be found in Appendix A.

Theorem 3.2 *Under Algorithm 3, if we choose SR1 update, then we have*

$$\mathbb{E}\,\frac{\lambda_f(\boldsymbol{x}_{k+1})}{\lambda_f(\boldsymbol{x}_k)} \leq \left(1 - \frac{k}{n}\right)\frac{\mathrm{tr}(\boldsymbol{G}_0 - \boldsymbol{A})}{\mu}, \ 0 \leq k \leq n.$$

If we adopt BFGS update, then we have

$$\mathbb{E}\,\frac{\lambda_f(\boldsymbol{x}_{k+1})}{\lambda_f(\boldsymbol{x}_k)} \leq \left(1 - \frac{1}{n}\right)^k \mathrm{tr}\left[\left(\boldsymbol{G}_0 - \boldsymbol{A}\right)\boldsymbol{A}^{-1}\right], \ k \geq 0.$$

Thus, we see both SR1 and BFGS methods share superlinear convergence for $\{\lambda_f(\boldsymbol{x}_k)\}$. In particular, for the SR1 update, our bound recovers the classical result in [16, Theorem 6.1], showing that the update stops after finite steps because $\boldsymbol{G}_n = \boldsymbol{A}$ and $\lambda_f(\boldsymbol{x}_{n+1}) = 0$. Moreover, we also give an exact convergence rate during the entire optimization process. And the main decreasing term $\left(1 - \frac{k}{n}\right)$ for the SR1 update as well as $(1 - \frac{1}{n})^k$ for the BFGS update in the k-th iteration are independent of the condition number κ of $\boldsymbol{A}$, which improves the bound $(1 - \frac{1}{n\kappa})^k$ in [19, Theorem 3.4].

Algorithm 4 Greedy/Random SR1/BFGS methods for strongly self-concordant objective

```
1: Initialization: Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$. Set $\boldsymbol{G}_0 = L\boldsymbol{I}_n, \boldsymbol{L}_0 = \boldsymbol{I}_n/\sqrt{L}$.
2: for $k \geq 0$ do
3:     Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{G}_k^{-1}\nabla f(\boldsymbol{x}_k)$.
4:     Compute $r_k = \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{x}_k}$, $\tilde{\boldsymbol{G}}_k = (1 + Mr_k)\,\boldsymbol{G}_k$, $\tilde{\boldsymbol{L}}_k = \boldsymbol{L}_k/\sqrt{1 + Mr_k}$.
5:     (i) Greedy/Random SR1: Choose $\boldsymbol{u}_k = \bar{\boldsymbol{u}}_{\nabla^2 f(\boldsymbol{x}_{k+1})}(\tilde{\boldsymbol{G}}_k)$, or $\boldsymbol{u}_k \sim \mathrm{Unif}(\mathcal{S}^{n-1})$.
       Compute $\boldsymbol{G}_{k+1} = \mathrm{SR1}(\tilde{\boldsymbol{G}}_k, \nabla^2 f(\boldsymbol{x}_{k+1}), \boldsymbol{u}_k)$.
6:     (ii) Greedy/Random BFGS: Choose $\tilde{\boldsymbol{u}}_k = \tilde{\boldsymbol{u}}_{\nabla^2 f(\boldsymbol{x}_{k+1})}(\tilde{\boldsymbol{L}}_k)$, or $\tilde{\boldsymbol{u}}_k \sim \mathrm{Unif}(\mathcal{S}^{n-1})$.
       Compute $\boldsymbol{G}_{k+1} = \mathrm{BFGS}(\tilde{\boldsymbol{G}}_k, \nabla^2 f(\boldsymbol{x}_{k+1}), \tilde{\boldsymbol{L}}_k^\top \tilde{\boldsymbol{u}}_k)$, and $\boldsymbol{L}_{k+1}$ based on Proposition 2.6.
7: end for
```

4 Minimization of general functions

Finally, we consider the optimization of a more general machine learning objective with unfixed Hessians: $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$, where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function with positive definite Hessians. Our goal is to extend the results in the previous sections, assuming that the methods can start from a sufficiently good initial point $\boldsymbol{x}_0$. We use the same assumption *strongly self-concordant* followed by Rodomanov and Nesterov [19].

Definition 4.1 (Strongly self-concordant) *A function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly self-concordant if the Hessians of f are close to each other in the sense that there exists a constant $M \geq 0$ s.t.*

$$\nabla^2 f(\boldsymbol{y}) - \nabla^2 f(\boldsymbol{x}) \preceq M\|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{z}} \nabla^2 f(\boldsymbol{w}),\ \forall\, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^n.$$

Rodomanov and Nesterov [19] have already mentioned several properties and examples of strongly self-concordant function, such as a strongly convex function with Lipschitz continuous Hessian. Let us now make one more common assumption about the function f as Rodomanov and Nesterov [19] did, that the function f is μ-strongly convex and L-smooth, i.e., there exist $L \geq \mu > 0$ such that

$$\mu\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}_n,\ \forall\, \boldsymbol{x} \in \mathbb{R}^n$$

and $\kappa := L/\mu$. Unlike quadratic minimization, the true Hessian in each step varies. In order to make $\boldsymbol{G}_k \succeq \nabla^2 f(\boldsymbol{x}_k)$ always hold, we adjust $\boldsymbol{G}_k$ before doing quasi-Newton update. Instructed from [19], we also use the *correction strategy*, which enlarges the approximation $\boldsymbol{G}_k$ properly shown in Line 4 of Algorithm 4. Moreover, Algorithm 4 is only for theoretical analysis. We need to adopt the inverse update rules (Eqs. (15) and (16)) in practice. Additionally, we assume that the constants M and L are available for simplicity.

For the BFGS update, we can analyze how the Hessian approximation measure $\sigma_{\boldsymbol{x}}(\boldsymbol{G}) \triangleq \sigma_{\nabla^2 f(\boldsymbol{x})}(\boldsymbol{G})$ changes after one iteration following [19]. The proof can be found in Appendix A.

Theorem 4.2 *Suppose that in Algorithm 4 BFGS update is used, and that the initial point $\boldsymbol{x}_0$ is sufficiently close to the solution:*

$$M\lambda_f(\boldsymbol{x}_0) \leq \frac{\ln 2}{4\kappa(2n+1)}. \tag{18}$$

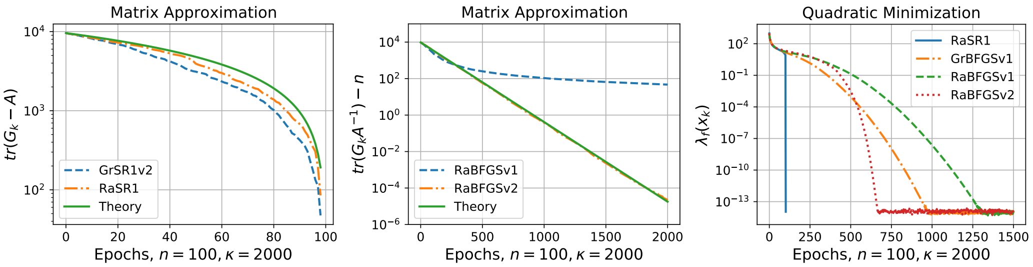
Then for all $k \geq 0$, we have

$$\nabla^2 f(\boldsymbol{x}_k) \preceq \boldsymbol{G}_k,\ \mathbb{E}\,\sigma_{\boldsymbol{x}_k}(\boldsymbol{G}_k) \leq 2n\kappa\left(1 - \frac{1}{n}\right)^k, \tag{19}$$

and

$$\mathbb{E}\,\frac{\lambda_f(\boldsymbol{x}_{k+1})}{\lambda_f(\boldsymbol{x}_k)} \leq 2n\kappa\left(1 - \frac{1}{n}\right)^k. \tag{20}$$

As for the SR1 update, the main difficulty compared to [19] is that we utilize a different measure $\tau_{\boldsymbol{A}}(\cdot)$ instead of $\sigma_{\boldsymbol{A}}(\cdot)$, resulting in a refined proof version below, and we leave the proofs in Appendix A.



(a) (b) (c)

Figure 1: (a, b) Comparison of different direction choosing methods under the SR1 or BFGS update for approximating a matrix $\boldsymbol{A}$ that $\mu \boldsymbol{I}_n \preceq \boldsymbol{A} \preceq L\boldsymbol{I}_n$ from $\boldsymbol{G}_0 = L\boldsymbol{I}_n$. (a) The variation of $\tau_{\boldsymbol{A}}(\boldsymbol{G}_k)$ during Random SR1 (RaSR1) and our Greedy SR1 (GrSR1v2) update with nearly matched upper bound. (b) The variation of $\sigma_{\boldsymbol{A}}(\boldsymbol{G}_k)$ during our Random BFGS (RaBFGSv2) update and the original random version (RaBFGSv1). (c) Comparison of SR1 and BFGS methods for quadratic objective. Here we only depict RaSR1 method, while the other SR1-type methods share similar behavior.

Theorem 4.3 *For Algorithm 4 with SR1 update, suppose that the initial point $\boldsymbol{x}_0$ is sufficiently close to the solution:*

$$M\lambda_f(\boldsymbol{x}_0) \leq \frac{\ln 2}{4\kappa(2n\kappa + 1)}. \tag{21}$$

Then for all $k \geq 0$, we have

$$\nabla^2 f(\boldsymbol{x}_k) \preceq \boldsymbol{G}_k, \mathbb{E}\,\sigma_{\boldsymbol{x}_k}(\boldsymbol{G}_k) \leq 2n\kappa^2 \left(1 - \frac{1}{n}\right)^k, \tag{22}$$

and

$$\mathbb{E}\,\frac{\lambda_f(\boldsymbol{x}_{k+1})}{\lambda_f(\boldsymbol{x}_k)} \leq 2n\kappa^2 \left(1 - \frac{1}{n}\right)^k. \tag{23}$$

Corollary 4.4 *Combining with [19, Theorem 4.7] (shown in Appendix B.1), if $\boldsymbol{x}_0$ satisfies $M\lambda_f(\boldsymbol{x}_0) \leq \frac{\ln \frac{3}{2}}{4\kappa}$, we could obtain 1) for the greedy BFGS or greedy SR1 method,*

$$\lambda_f(\boldsymbol{x}_{k_0+k}) \leq \left(1 - \frac{1}{n}\right)^{k(k-1)/2} \cdot \left(\frac{1}{2}\right)^k \cdot \left(1 - \frac{1}{2\kappa}\right)^{k_0} \cdot \lambda_f(\boldsymbol{x}_0), \textit{ for all } k \geq 0,$$

where $k_0 = O\left(\max\{n, \kappa\} \ln(n\kappa)\right)$, 2) for the random BFGS or random SR1 method, with probability $1 - \delta$ for any $\delta \in (0, 1)$,

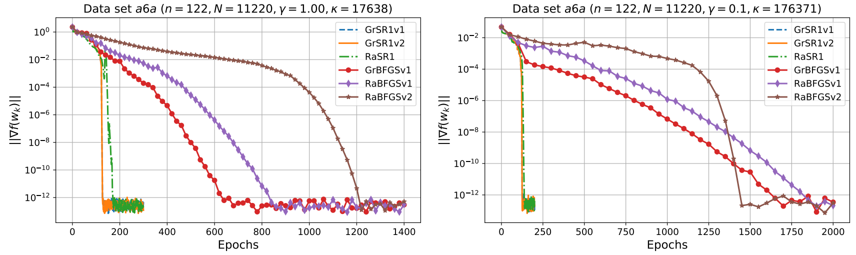
$$\lambda_f(\boldsymbol{x}_{k_0+k}) \leq \left(1 - \frac{1}{n+1}\right)^{k(k-1)/2} \cdot \left(\frac{1}{2}\right)^k \cdot \left(1 - \frac{1}{2\kappa}\right)^{k_0} \cdot \lambda_f(\boldsymbol{x}_0), \textit{ for all } k \geq 0,$$

where $k_0 = O\left(\max\{n, \kappa\} \ln(n\kappa/\delta)\right)$.

Therefore, both greedy and random methods have non-asymptotic superlinear convergence rates. Additionally, our superlinear rates are condition-number-free compared to the rates in [19].

5 Numerical experiments

In this section, we verify our theorems through numerical results for quasi-Newton methods. Rodomanov and Nesterov [19] have already compared their proposed greedy quasi-Newton method with the classical quasi-Newton methods. They showed that GrDFP, GrBFGS, GrSR1 (greedy DFP, BFGS, SR1 methods) using directions based on $\bar{\boldsymbol{u}}_{\boldsymbol{A}}(\boldsymbol{G}) = \arg\max_{\boldsymbol{u} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}} \frac{\boldsymbol{u}^\top \boldsymbol{G} \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u}}$, have quite competitive convergence with the standard versions. They also presented the results for the randomized versions RaDFP, RaBFGS, RaSR1, which choose directions directly from a standard Euclidean sphere. They discovered that the randomized methods are slightly slower than the greedy ones. However, the difference is not really significant.



(a) $\gamma = 1$

(b) $\gamma = 0.1$

Figure 2: Comparison of SR1 and BFGS update for ℓ_2-regularized logistic regression applied with 'a6a' data from the LIBSVM collection of real-world data sets. We list the name of dataset, the dimension n and the condition number κ under the corresponding γ in the title of each figure. The lines of GrSR1v1 and GrSR1v2 are overlapped in some figures.

The difference between our algorithms and their methods mainly comes from the greedy strategy for SR1 and the random strategy for BFGS[3]. Hence, we mainly focus on exhibiting our validity in these schemes. We refer to GrSR1v2 as our revised method and GrSR1v1 as the previous method. Similarly, we denote RaBFGSv2 that uses scaled directions ($\boldsymbol{L}^\top \boldsymbol{u}$) and RaBFGSv1 that directly uses random directions $\boldsymbol{u}$ correspondingly. We use $\boldsymbol{u} \sim \mathrm{Unif}(\mathcal{S}^{n-1})$ in all experiments for brevity.

Matrix approximation. When using Algorithms 1 and 2 for approximating a matrix $\boldsymbol{A} \succ 0$, we show the measure as proved by Theorems 2.3 and 2.4 in Figures 1a and 1b. As Figure 1a depicts, our greedy and random SR1 updates (GrSR1v2 and RaSR1) share superlinear convergence rate under measure $\tau_{\boldsymbol{A}}(\cdot)$, while our theoretical bound matches them well. Moreover, Figure 1b describes the behavior of the random BFGS update. Note that the greedy BFGS update we proposed does not satisfy the $O(n^2)$ complexity, thus we leave it out. Our theory totally matches the linear convergence of measure $\sigma_{\boldsymbol{A}}(\cdot)$ in our modified random BFGS update (RaBFGSv2). However, directly choosing a direction without scaling (RaBFGSv1) fails to give such bounds. Moreover, we also discover the same findings under different condition numbers in Appendix C.1. Clearly, our methods provide effective ways of approaching a positive definite Hessian matrix.

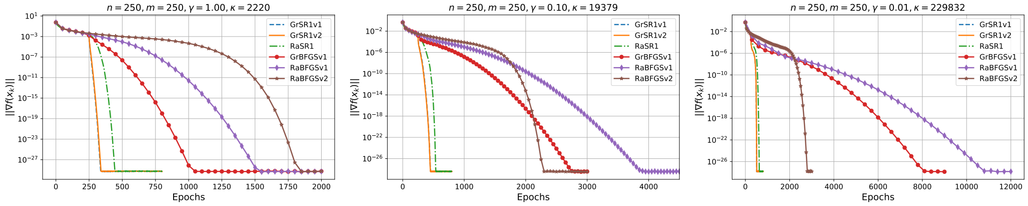
Quadratic minimization. We also consider unconstrained quadratic minimization with the same positive definite matrix $\boldsymbol{A}$ and a randomly selected vector $\boldsymbol{b} \in \mathbb{R}^n$. Running Algorithm 3 with SR1 and BFGS updates, we obtain the superlinear convergence of $\lambda_f(\cdot)$ shown in Figure 1c. Not surprisingly, our RaBFGSv2 runs faster than RaBFGSv1, while we also have the theoretical guarantee. At the same time, SR1-type methods converge to zero after $n+1$ steps because $\boldsymbol{G}_n = \boldsymbol{A}$. Here, we only depict the RaSR1 update, while the other SR1-type methods share similar behavior. Although our theoretical bound can not directly match the experiments due to the related initial terms $\tau_{\boldsymbol{A}}(\boldsymbol{G}_0)$ and $\sigma_{\boldsymbol{A}}(\boldsymbol{G}_0)$, the decay terms: $(1 - k/n)$ vs. $(1 - 1/n)^k$ already show the superiority of the SR1 method over the BFGS method in the quadratic minimization problem.

Regularized logistic regression. Next, we consider ℓ_2-regularized logistic regression:

$$f(\boldsymbol{w}) = \sum_{i=1}^{N} \ln\left(1 + e^{-y_i \boldsymbol{w}^\top \boldsymbol{x}_i}\right) + \frac{\gamma}{2}\|\boldsymbol{w}\|^2, \ \boldsymbol{w} \in \mathbb{R}^n,$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{n \times N}$ are training samples with the corresponding labels $y_1, \ldots, y_N \in \{+1, -1\}$, and $\gamma > 0$ is the regularization coefficient. We follow the same experimental design but take data from the LIBSVM collection of real-world data sets for binary classification problems [6]. And we do not apply the *correction strategy* ($\tilde{\boldsymbol{G}}_k = (1 + M r_k)\boldsymbol{G}_k$) in Algorithm 4) recommended by [19]. Other details are shown in Appendix C.2. In order to simulate the local convergence, we use the same initialization after running several standard Newton's steps to make measure $\|\nabla f(\boldsymbol{w}_0)\|$ small (around $10^{-2} \sim 10^{1}$).

[3] There is no difference in the random SR1 method compared to Rodomanov and Nesterov [19], which directly selects random directions. And our greedy BFGS method is not efficient ($O(n^3)$ in each iteration) as we mentioned in Remark 2.5. Thus we leave it out.



(a) $\gamma = 1$ (b) $\gamma = 0.1$ (c) $\gamma = 0.01$

Figure 3: Comparison of SR1 and BFGS update for Regularized Log-Sum-Exp. The dimension n, the number m of linear functions, the regularization coefficient γ and condition number κ are displayed in the title of each graph. The lines of GrSR1v1 and GrSR1v2 are overlapped in each figure.

We show the results in Figure 2. As we can see, the general picture is the same as for the previous objective. In particular, SR1-type methods are faster than BFGS-type methods, and the greedy algorithms also converge more rapidly than the random algorithms. The only difference is that our RaBFGSv2 may have slower convergence behavior than RaBFGSv1 under a small κ in Figure 2a. We consider our scaled direction is more suitable for a constant Hessian matrix as the quadratic objective has. Thus we still have a better convergence rate in the last few iterations when Hessians are nearly unchanged in Figure 2a. However, the Hessian varies drastically in the initial period. Thus there is less benefit under a more accurate Hessian approximation. When applied to the ill-conditioning setting with a large κ in Figure 2b, we find our RaBFGSv2 could be faster than GrBFGSv1 and RaBFGSv1. This implies that our proposed method has less dependence on the condition number κ.

Regularized Log-Sum-Exp. Followed by Rodomanov and Nesterov [19], we also present preliminary computational results for greedy and random quasi-Newton methods, applied to the following test function with $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m] \in \mathbb{R}^{n \times m}$, $b_1, \ldots, b_m \in \mathbb{R}$, and $\gamma > 0$:

$$f(\boldsymbol{x}) := \ln\left(\sum_{j=1}^{m} e^{\boldsymbol{c}_j^\top \boldsymbol{x} - b_j}\right) + \frac{1}{2}\sum_{j=1}^{m}\left(\boldsymbol{c}_j^\top \boldsymbol{x}\right)^2 + \frac{\gamma}{2}\left\|\boldsymbol{x}\right\|^2, \boldsymbol{x} \in \mathbb{R}^n.$$

We also use the same synthetic data in [19, Section 5.1] and leave the detail in Appendix C.2. As Figure 3 depicts, the BFGS-type methods are slower than the SR1-type methods. GrBFGSv1 and RaBFGSv1 are faster than RaBFGSv2 in a small condition number case, but they become slower than RaBFGSv2 when the condition number becomes huge. Therefore, we think our RaBFGSv2 which uses scaled direction indeed has less dependence on the condition number as our theory shows.

Overall, our proposed methods do not lose the superlinear convergence rate in the large condition number schemes, while we also present the theoretical guarantee for these algorithms.

6 Conclusion

In this work, we have studied the behavior of two famous quasi-Newton methods: the SR1 and the BFGS methods. We have presented different greedy methods in contrast to Rodomanov and Nesterov [19], and the random version of these methods. In particular, we have provided the faster Hessian approximation behavior and the condition-number-free (local) superlinear convergence rates applied to quadratic or strongly self-concordant objectives. Moreover, the experiments corroborate our analysis well. Note that our current results do not give the superlinear convergence rate of the standard quasi-Newton methods with only gradient information. However, we hope that the theoretical analysis and the related work would be useful for understanding the quasi-Newton methods in a more specific view.

Acknowledgments and Disclosure of Funding

Lin and Zhang have been supported by the National Key Research and Development Project of China (No. 2018AAA0101004), by the National Natural Science Foundation of China (No. 11771002), and Beijing Natural Science Foundation (Z190001). Ye has been supported by the National Natural Science Foundation of China (No. 12101491), and Shenzhen Research Institute of Big Data (SRIBD) (No. J00120190004).

References

[1] Charles G Broyden. Quasi-newton methods and their application to function minimisation. *Mathematics of Computation*, 21(99):368–381, 1967.

[2] Charles G Broyden. The convergence of a class of double-rank minimization algorithms: 2. the new algorithm. *IMA journal of applied mathematics*, 6(3):222–231, 1970.

[3] Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[4] Charles George Broyden, John E Dennis Jr, and Jorge J Moré. On the local and superlinear convergence of quasi-newton methods. *IMA Journal of Applied Mathematics*, 12(3):223–245, 1973.

[5] Richard H Byrd, Jorge Nocedal, and Ya-Xiang Yuan. Global convergence of a cass of quasi-newton methods on convex problems. *SIAM Journal on Numerical Analysis*, 24(5):1171–1190, 1987.

[6] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[7] William C Davidon. Variable metric method for minimization. *SIAM Journal on Optimization*, 1(1):1–17, 1991.

[8] John E Dennis and Jorge J Moré. A characterization of superlinear convergence and its application to quasi-newton methods. *Mathematics of computation*, 28(126):549–560, 1974.

[9] Roger Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13(3): 317–322, 1970.

[10] Roger Fletcher and Michael JD Powell. A rapidly convergent descent method for minimization. *The computer journal*, 6(2):163–168, 1963.

[11] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26, 1970.

[12] Gene H Golub and Charles F Van Loan. Matrix computations, 1996.

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

[14] Qiujiang Jin and Aryan Mokhtari. Non-asymptotic superlinear convergence of standard quasi-newton methods. *arXiv preprint arXiv:2003.13607*, 2020.

[15] Dmitry Kovalev, Robert M Gower, Peter Richtárik, and Alexander Rogozin. Fast linear convergence of randomized bfgs. *arXiv preprint arXiv:2002.11337*, 2020.

[16] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[17] MJD Powell. On the convergence of the variable metric algorithm. *IMA Journal of Applied Mathematics*, 7(1):21–36, 1971.

[18] Anton Rodomanov and Yurii Nesterov. New results on superlinear convergence of classical quasi-newton methods. *Journal of Optimization Theory and Applications*, 188(3):744–769, Jan 2021. ISSN 1573-2878. doi: 10.1007/s10957-020-01805-8. URL `http://dx.doi.org/10.1007/s10957-020-01805-8`.

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

[20] Anton Rodomanov and Yurii Nesterov. Rates of superlinear convergence for classical quasi-newton methods. *Mathematical Programming*, pages 1–32, 2021.

[21] David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

[22] Andrzej Stachurski. Superlinear convergence of broyden's bounded θ-class of methods. *Mathematical Programming*, 20(1):196–212, 1981.

[23] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[24] Hiroshi Yabe and Naokazu Yamaki. Local and superlinear convergence of structured quasi-newton methods for nonlinear optimization. *Journal of the Operations Research Society of Japan*, 39(4):541–557, 1996.

A Missing Proofs

A.1 Proof of Theorem 2.3

Proof: Let $\boldsymbol{G}_{k+1} = \mathrm{SR1}(\boldsymbol{G}_k, \boldsymbol{A}, \boldsymbol{u}_k)$, then

$$(\boldsymbol{G}_{k+1} - \boldsymbol{A})\boldsymbol{u}_k \overset{(2)}{=} (\boldsymbol{G}_k - \boldsymbol{A})\boldsymbol{u}_k - \frac{(\boldsymbol{G}_k - \boldsymbol{A})\boldsymbol{u}_k\boldsymbol{u}_k^\top(\boldsymbol{G}_k - \boldsymbol{A})}{\boldsymbol{u}_k^\top(\boldsymbol{G}_k - \boldsymbol{A})\boldsymbol{u}_k}\boldsymbol{u}_k = \boldsymbol{0}.$$

Moreover, for any $\boldsymbol{v}$ such that $(\boldsymbol{G}_k - \boldsymbol{A})\boldsymbol{v} = \boldsymbol{0}$, we have

$$(\boldsymbol{G}_{k+1} - \boldsymbol{A})\boldsymbol{v} = (\boldsymbol{G}_k - \boldsymbol{A})\boldsymbol{v} - \frac{(\boldsymbol{G}_k - \boldsymbol{A})\boldsymbol{u}_k\boldsymbol{u}_k^\top(\boldsymbol{G}_k - \boldsymbol{A})}{\boldsymbol{u}_k^\top(\boldsymbol{G}_k - \boldsymbol{A})\boldsymbol{u}_k}\boldsymbol{v} = \boldsymbol{0}.$$

Therefore, we obtain

$$(\boldsymbol{G}_k - \boldsymbol{A})\,\boldsymbol{u}_j = \boldsymbol{0}, \forall j < k. \tag{24}$$

For the *greedy method*, we denote $\bar{\boldsymbol{u}}_k = \bar{\boldsymbol{u}}_A(\boldsymbol{G}_k)$, then from Eq. (24), we can assume $\bar{\boldsymbol{u}}_k \neq \bar{\boldsymbol{u}}_j, \forall k > j$, and $\mathrm{rank}(\boldsymbol{G}_k - \boldsymbol{A}) \leq n - k$. At step k, without loss of generality, we assume $\bar{\boldsymbol{u}}_i = \boldsymbol{e}_i, i = 0, \dots, k-1$. Then $(\boldsymbol{G}_k - \boldsymbol{A})\boldsymbol{e}_i = \boldsymbol{0}, \forall i \leq k-1$, leading to $(\boldsymbol{G}_k - \boldsymbol{A})_{ii} = 0, \forall i \leq k-1$. From Cauchy–Schwarz inequality and $\boldsymbol{G}_k \succeq \boldsymbol{A}$, we obtain

$$\frac{\bar{\boldsymbol{u}}_k^\top(\boldsymbol{G}_k - \boldsymbol{A})^2\bar{\boldsymbol{u}}_k}{\bar{\boldsymbol{u}}_k^\top(\boldsymbol{G}_k - \boldsymbol{A})\bar{\boldsymbol{u}}_k} \geq \frac{\bar{\boldsymbol{u}}_k^\top(\boldsymbol{G}_k - \boldsymbol{A})\bar{\boldsymbol{u}}_k}{\bar{\boldsymbol{u}}_k^\top\bar{\boldsymbol{u}}_k} \overset{(7)}{=} \max_{1\leq i\leq n}(\boldsymbol{G}_k - \boldsymbol{A})_{ii} \geq \frac{1}{n-k}\mathrm{tr}(\boldsymbol{G}_k - \boldsymbol{A}),$$

where the last inequality uses $\mathrm{rank}(\boldsymbol{G}_k - \boldsymbol{A}) \leq n - k$.

Therefore, the greedy choice of $\boldsymbol{u}_k$ leads to

$$\tau_{\boldsymbol{A}}(\boldsymbol{G}_{k+1}) \leq \left(1 - \frac{1}{n-k}\right)\tau_{\boldsymbol{A}}(\boldsymbol{G}_k).$$

Consequently, we have

$$\begin{aligned}\tau_{\boldsymbol{A}}(\boldsymbol{G}_k) &\leq \left(1 - \frac{1}{n-k+1}\right)\tau_{\boldsymbol{A}}(\boldsymbol{G}_{k-1}) \leq \cdots \leq \left[\prod_{j=1}^{k}\left(1 - \frac{1}{n-j+1}\right)\right]\tau_{\boldsymbol{A}}(\boldsymbol{G}_0)\\ &= \left[\prod_{j=1}^{k}\frac{n-j}{n-j+1}\right]\tau_{\boldsymbol{A}}(\boldsymbol{G}_0) = \frac{n-k}{n}\tau_{\boldsymbol{A}}(\boldsymbol{G}_0) = \left(1 - \frac{k}{n}\right)\tau_{\boldsymbol{A}}(\boldsymbol{G}_0).\end{aligned}$$

For the *random method* at step k, since $\boldsymbol{G}_k \succeq \boldsymbol{A}$, we have $\lambda_i \triangleq \lambda_i(\boldsymbol{G}_k - \boldsymbol{A}) \geq 0, \forall i \in \{1, \dots, n\}$. We let $r_k \triangleq \mathrm{rank}(\boldsymbol{G}_k - \boldsymbol{A})$ and denote $\boldsymbol{G}_k - \boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ as the spectral decomposition of $\boldsymbol{G}_k - \boldsymbol{A}$ with an orthogonal matrix $\boldsymbol{U}$ and a diagonal matrix $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \dots, \lambda_n\}$. Let $\boldsymbol{v} = (v_1, \dots, v_n)^\top := \boldsymbol{U}^\top\boldsymbol{u}$. We can show in expectation that

$$\mathbb{E}_{\boldsymbol{u}}\frac{\boldsymbol{u}^\top(\boldsymbol{G}_k - \boldsymbol{A})^2\boldsymbol{u}}{\boldsymbol{u}^\top(\boldsymbol{G}_k - \boldsymbol{A})\boldsymbol{u}} = \mathbb{E}_{\boldsymbol{v}}\frac{\sum_{i=1}^{r_k}\lambda_i^2 v_i^2}{\sum_{i=1}^{r_k}\lambda_i v_i^2} \geq \mathbb{E}_{\boldsymbol{v}}\frac{\sum_{i=1}^{r_k}\lambda_i v_i^2}{\sum_{i=1}^{r_k} v_i^2} = \sum_{i=1}^{r_k}\lambda_i\mathbb{E}_{\boldsymbol{v}}\frac{v_i^2}{\sum_{j=1}^{r_k} v_j^2} = \frac{1}{r_k}\tau_{\boldsymbol{A}}(\boldsymbol{G}_k).$$

The first equality holds due to $(\boldsymbol{G}_k - \boldsymbol{A})^2 = \boldsymbol{U}\boldsymbol{\Lambda}^2\boldsymbol{U}^\top$, and the inequality holds due to the Cauchy–Schwarz inequality and $\boldsymbol{G}_k \succeq \boldsymbol{A}$:

$$\left(\sum_{i=1}^{r_k}\lambda_i^2 v_i^2\right)\left(\sum_{i=1}^{r_k} v_i^2\right) \geq \left(\sum_{i=1}^{r_k}\lambda_i v_i^2\right)^2;$$

the second equality uses the fact that $\boldsymbol{v}$ is still spherically symmetric, thus also permutation invariant:

$$\mathbb{E}_{\boldsymbol{v}}\frac{v_1^2}{\sum_{j=1}^{r_k} v_j^2} = \cdots = \mathbb{E}_{\boldsymbol{v}}\frac{v_{r_k}^2}{\sum_{j=1}^{r_k} v_j^2} = \frac{1}{r_k}\sum_{i=1}^{r_k}\mathbb{E}_{\boldsymbol{v}}\frac{v_i^2}{\sum_{j=1}^{r_k} v_j^2} = \frac{1}{r_k}.$$

Therefore, the random choice of $\boldsymbol{u}_k$ leads to

$$\mathbb{E}_{\boldsymbol{u}_k}\left[\tau_{\boldsymbol{A}}(\boldsymbol{G}_{k+1})|\boldsymbol{G}_k\right] \leq \left(1-\frac{1}{r_k}\right)\tau_{\boldsymbol{A}}(\boldsymbol{G}_k).$$

Note that $\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{n-1}$ are linearly independent almost surely [23], thus $r_k = \operatorname{rank}(\boldsymbol{G}_k - \boldsymbol{A}) \leq n-k, a.s.$ We finally obtain

$$\begin{aligned}\mathbb{E}\,\tau_{\boldsymbol{A}}(\boldsymbol{G}_k) &\leq \left(1-\frac{1}{n-k+1}\right)\mathbb{E}\,\tau_{\boldsymbol{A}}(\boldsymbol{G}_{k-1}) \leq \cdots \leq \prod_{j=1}^{k}\left(1-\frac{1}{n-j+1}\right)\tau_{\boldsymbol{A}}(\boldsymbol{G}_0)\\ &= \left(1-\frac{k}{n}\right)\tau_{\boldsymbol{A}}(\boldsymbol{G}_0).\end{aligned}$$

□

Remark A.1 *In fact, for the random method, when applied to $\sigma_{\boldsymbol{A}}(\cdot)$, we could not handle the expectation of the division of random variables with known expectations, because the random variables themselves are non-independent as Rodomanov and Nesterov [19] mentioned. We take $\tau_{\boldsymbol{A}}(\cdot)$ instead, which overcomes this difficulty due to the same spectral decomposition of $(\boldsymbol{G}-\boldsymbol{A})^2$ and $\boldsymbol{G}-\boldsymbol{A}$.*

A.2 Proof of Theorem 2.4

Proof: For the *greedy method* at step k, since $\boldsymbol{G}_k^{-1} = \boldsymbol{L}_k^\top \boldsymbol{L}_k$, we obtain

$$\begin{aligned}\max_{\tilde{\boldsymbol{u}}\in\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n\}} \frac{\tilde{\boldsymbol{u}}^\top \boldsymbol{L}_k^{-\top}\boldsymbol{A}^{-1}\boldsymbol{L}_k^{-1}\tilde{\boldsymbol{u}}}{\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}} &= \max_{i\in[n]}\left(\boldsymbol{L}_k^{-\top}\boldsymbol{A}^{-1}\boldsymbol{L}_k^{-1}\right)_{ii} \geq \frac{1}{n}\operatorname{tr}\left(\boldsymbol{L}_k^{-\top}\boldsymbol{A}^{-1}\boldsymbol{L}_k^{-1}\right)\\ &= \frac{1}{n}\operatorname{tr}\left(\left(\boldsymbol{L}_k^\top\boldsymbol{L}_k\right)^{-1}\boldsymbol{A}^{-1}\right) = \frac{1}{n}\operatorname{tr}(\boldsymbol{G}_k\boldsymbol{A}^{-1}).\end{aligned} \tag{25}$$

Therefore, the greedy choice of $\boldsymbol{u}_k = \boldsymbol{L}_k^\top\tilde{\boldsymbol{u}}_k$ leads to

$$\begin{aligned}\sigma_{\boldsymbol{A}}(\boldsymbol{G}_{k+1}) &\overset{(10)}{\leq} \sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) - \frac{\boldsymbol{u}_k^\top\boldsymbol{G}_k\boldsymbol{A}^{-1}\boldsymbol{G}_k\boldsymbol{u}_k}{\boldsymbol{u}_k^\top\boldsymbol{G}_k\boldsymbol{u}_k} + 1 = \sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) - \frac{\tilde{\boldsymbol{u}}_k^\top\boldsymbol{L}_k^{-\top}\boldsymbol{A}^{-1}\boldsymbol{L}_k^{-1}\tilde{\boldsymbol{u}}_k}{\tilde{\boldsymbol{u}}_k^\top\tilde{\boldsymbol{u}}_k} + 1\\ &\overset{(25)}{\leq} \sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) - \frac{1}{n}\operatorname{tr}(\boldsymbol{G}_k\boldsymbol{A}^{-1}) + 1 = \left(1-\frac{1}{n}\right)\sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) \leq \cdots \leq \left(1-\frac{1}{n}\right)^k\sigma_{\boldsymbol{A}}(\boldsymbol{G}_0).\end{aligned}$$

For the *random method*, i.e., $\tilde{\boldsymbol{u}} \sim \mathcal{N}(0, I_n)$ or $\tilde{\boldsymbol{u}} \sim \operatorname{Unif}(\mathcal{S}^{n-1})$. We have that $\mathbb{E}_{\tilde{\boldsymbol{u}}}\frac{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^\top}{\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}} = \frac{1}{n}\boldsymbol{I}_n$. Hence, we obtain

$$\begin{aligned}\mathbb{E}_{\tilde{\boldsymbol{u}}}\frac{\tilde{\boldsymbol{u}}^\top\boldsymbol{L}_k^{-\top}\boldsymbol{A}^{-1}\boldsymbol{L}_k^{-1}\tilde{\boldsymbol{u}}}{\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}} &= \operatorname{tr}\left[\boldsymbol{L}_k^{-\top}\boldsymbol{A}^{-1}\boldsymbol{L}_k^{-1}\cdot\mathbb{E}_{\tilde{\boldsymbol{u}}}\frac{\tilde{\boldsymbol{u}}\tilde{\boldsymbol{u}}^\top}{\tilde{\boldsymbol{u}}^\top\tilde{\boldsymbol{u}}}\right]\\ &= \frac{1}{n}\operatorname{tr}(\boldsymbol{L}_k^{-\top}\boldsymbol{A}^{-1}\boldsymbol{L}_k^{-1}) = \frac{1}{n}\operatorname{tr}(\boldsymbol{G}_k\boldsymbol{A}^{-1}),\end{aligned} \tag{26}$$

Therefore, the random choice of $\boldsymbol{u}_k = \boldsymbol{L}_k^\top\tilde{\boldsymbol{u}}_k$ leads to

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}_k}\left[\sigma_{\boldsymbol{A}}(\boldsymbol{G}_{k+1})|\boldsymbol{G}_k\right] &\overset{(10)}{=} \mathbb{E}_{\boldsymbol{u}_k}\sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) - \frac{\boldsymbol{u}_k^\top\boldsymbol{G}_k\boldsymbol{A}^{-1}\boldsymbol{G}_k\boldsymbol{u}_k}{\boldsymbol{u}_k^\top\boldsymbol{G}_k\boldsymbol{u}_k} + 1\\ &= \mathbb{E}_{\tilde{\boldsymbol{u}}_k}\sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) - \frac{\tilde{\boldsymbol{u}}_k^\top\boldsymbol{L}_k^{-\top}\boldsymbol{A}^{-1}\boldsymbol{L}_k^{-1}\tilde{\boldsymbol{u}}_k}{\tilde{\boldsymbol{u}}_k^\top\tilde{\boldsymbol{u}}_k} + 1\\ &\overset{(26)}{=} \sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) - \frac{1}{n}\operatorname{tr}(\boldsymbol{G}_k\boldsymbol{A}^{-1}) + 1 = \left(1-\frac{1}{n}\right)\sigma_{\boldsymbol{A}}(\boldsymbol{G}_k),\end{aligned}$$

showing that

$$\mathbb{E}\,\sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) = \left(1-\frac{1}{n}\right)\mathbb{E}\,\sigma_{\boldsymbol{A}}(\boldsymbol{G}_{k-1}) = \cdots = \left(1-\frac{1}{n}\right)^k\sigma_{\boldsymbol{A}}(\boldsymbol{G}_0).$$

□

A.3 Proof of Theorem 3.2

Proof: We denote $\eta_k = \left\|\boldsymbol{A}^{-1/2}\boldsymbol{G}_k\boldsymbol{A}^{-1/2}\right\|_2$, then $\boldsymbol{A} \preceq \boldsymbol{G}_k \preceq \eta_k\boldsymbol{A}$. Hence we get

$$\begin{aligned}
\eta_k - 1 \quad &\le \quad \left\|\boldsymbol{A}^{-1/2}\left(\boldsymbol{G}_k - \boldsymbol{A}\right)\boldsymbol{A}^{-1/2}\right\|_2 \\
&\le \quad \operatorname{tr}\left(\boldsymbol{A}^{-1/2}\left(\boldsymbol{G}_k - \boldsymbol{A}\right)\boldsymbol{A}^{-1/2}\right) \overset{(6)}{=} \sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) \qquad (27) \\
&\le \quad \frac{\operatorname{tr}\left(\boldsymbol{G}_k - \boldsymbol{A}\right)}{\mu} \overset{(5)}{=} \frac{\tau_{\boldsymbol{A}}(\boldsymbol{G}_k)}{\mu}. \qquad (28)
\end{aligned}$$

1). From Eq. (9), we know that under SR1 update,

$$\mathbb{E}(\eta_k - 1) \overset{(28)}{\le} \frac{\mathbb{E}\tau_{\boldsymbol{A}}(\boldsymbol{G}_k)}{\mu} \overset{(9)}{\le} \left(1 - \frac{k}{n}\right)\frac{\tau_{\boldsymbol{A}}(\boldsymbol{G}_0)}{\mu}.$$

By Lemma 3.1, this implies

$$\mathbb{E}\,\frac{\lambda_f(\boldsymbol{x}_{k+1})}{\lambda_f(\boldsymbol{x}_k)} \le \mathbb{E}(\eta_k - 1) \le \left(1 - \frac{k}{n}\right)\frac{\operatorname{tr}(\boldsymbol{G}_0 - \boldsymbol{A})}{\mu}.$$

2). From Eq. (13), we know that under BFGS update,

$$\mathbb{E}(\eta_k - 1) \overset{(27)}{\le} \mathbb{E}\,\sigma_{\boldsymbol{A}}(\boldsymbol{G}_k) \overset{(13))}{\le} \left(1 - \frac{1}{n}\right)^k \sigma_{\boldsymbol{A}}(\boldsymbol{G}_0).$$

By Lemma 3.1, this implies

$$\mathbb{E}\,\frac{\lambda_f(\boldsymbol{x}_{k+1})}{\lambda_f(\boldsymbol{x}_k)} \le \mathbb{E}(\eta_k - 1) \le \left(1 - \frac{1}{n}\right)^k \operatorname{tr}\left[\left(\boldsymbol{G}_0 - \boldsymbol{A}\right)\boldsymbol{A}^{-1}\right].$$

□

A.4 Proof of Theorem 4.2

We need several Lemmas from Rodomanov and Nesterov [19].

Lemma A.2 (Rodomanov and Nesterov [19] Lemma 4.2) *Let $r \triangleq \|\boldsymbol{y}-\boldsymbol{x}\|_{\boldsymbol{x}}$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then*

$$\frac{\nabla^2 f(\boldsymbol{x})}{1 + Mr} \preceq \nabla^2 f(\boldsymbol{y}) \preceq (1 + Mr)\,\nabla^2 f(\boldsymbol{x}). \qquad (29)$$

Also, for $\boldsymbol{J} \triangleq \int_0^1 \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))dt$ and $\boldsymbol{v} \in \{\boldsymbol{x}, \boldsymbol{y}\}$, we have

$$\frac{1}{1 + \frac{Mr}{2}}\nabla^2 f(\boldsymbol{v}) \preceq \boldsymbol{J} \preceq \left(1 + \frac{Mr}{2}\right)\nabla^2 f(\boldsymbol{v}).$$

Lemma A.3 (Rodomanov and Nesterov [19] Lemma 4.3 and Lemma 4.4) *Let $\boldsymbol{x} \in \mathbb{R}^n$, and a symmetric matrix $\boldsymbol{G}$, such that $\nabla^2 f(\boldsymbol{x}) \preceq \boldsymbol{G} \preceq \eta\nabla^2 f(\boldsymbol{x})$, for some $\eta \ge 1$. Let $\boldsymbol{x}_+ \in \mathbb{R}^n$, and $r = \|\boldsymbol{x}_+ - \boldsymbol{x}\|_{\boldsymbol{x}}$. Then*

$$\tilde{\boldsymbol{G}} \triangleq (1 + Mr)\,\boldsymbol{G} \succeq \nabla^2 f(\boldsymbol{x}_+),$$

and for all $\boldsymbol{u} \in \mathbb{R}^n$ and $\tau \in [0, 1]$, we have

$$\nabla^2 f(\boldsymbol{x}_+) \preceq \operatorname{Broyd}_\tau\left(\tilde{\boldsymbol{G}}, \nabla^2 f(\boldsymbol{x}_+), \boldsymbol{u}\right) \preceq \left[(1 + Mr)^2\,\eta\right]\nabla^2 f(\boldsymbol{x}_+).$$

More specifically, if $\boldsymbol{x}_+ = \boldsymbol{x} - \boldsymbol{G}^{-1}\nabla f(\boldsymbol{x})$, and letting $\lambda \triangleq \lambda_f(\boldsymbol{x})$ be such that $M\lambda \le 2$, then, $r \le \lambda$, and

$$\lambda_f(\boldsymbol{x}_+) \le \left(1 + \frac{M\lambda}{2}\right)\frac{\eta - 1 + \frac{M\lambda}{2}}{\eta}\lambda.$$

Lemma A.4 (Modified from Rodomanov and Nesterov [19] Lemma 4.8) *Let $\boldsymbol{x} \in \mathbb{R}^n$, and a symmetric matrix $\boldsymbol{G}$, such that $\nabla^2 f(\boldsymbol{x}) \preceq \boldsymbol{G}$. Let $\boldsymbol{x}_+ \in \mathbb{R}^n$, and $r = \|\boldsymbol{x}_+ - \boldsymbol{x}\|_{\boldsymbol{x}}$, $\tilde{\boldsymbol{G}} = (1+Mr)\,\boldsymbol{G}$, $\tilde{\boldsymbol{G}}^{-1} = \tilde{\boldsymbol{L}}^\top \tilde{\boldsymbol{L}}$, and $\boldsymbol{u}$ follows from Algorithm 2 applied with $\tilde{\boldsymbol{G}}, \nabla^2 f(\boldsymbol{x}_+)$. Then we have*

$$\mathbb{E}_{\tilde{\boldsymbol{u}}} \sigma_{\boldsymbol{x}_+}(\mathrm{BFGS}(\tilde{\boldsymbol{G}}, \nabla^2 f(\boldsymbol{x}_+), \tilde{\boldsymbol{L}}^\top \tilde{\boldsymbol{u}})) \le \left(1 - \frac{1}{n}\right)(1+Mr)^2 \left(\sigma_{\boldsymbol{x}}(\boldsymbol{G}) + \frac{2nMr}{1+Mr}\right),$$

Proof: We already know from Lemma A.3 that $\nabla^2 f(\boldsymbol{x}_+) \preceq \tilde{\boldsymbol{G}}$. Then by Theorem 2.4, we have

$$\mathbb{E}_{\tilde{\boldsymbol{u}}} \sigma_{\boldsymbol{x}_+}(\mathrm{BFGS}(\tilde{\boldsymbol{G}}, \nabla^2 f(\boldsymbol{x}_+), \tilde{\boldsymbol{L}}^\top \tilde{\boldsymbol{u}})) \le \left(1 - \frac{1}{n}\right) \sigma_{\boldsymbol{x}_+}(\tilde{\boldsymbol{G}}),$$

The remaining proof is same as [19, Lemma 4.8]. To make the paper self-contained, we also show the detail below.

$$\begin{aligned}
\sigma_{\boldsymbol{x}_+}(\tilde{\boldsymbol{G}}) &\overset{(6)}{=} \left\langle \nabla^2 f(\boldsymbol{x}_+)^{-1}, \tilde{\boldsymbol{G}} \right\rangle - n = (1+Mr)\left\langle \nabla^2 f(\boldsymbol{x}_+)^{-1}, \boldsymbol{G}\right\rangle - n \\
&\overset{(29)}{\le} (1+Mr)^2 \left\langle \nabla^2 f(\boldsymbol{x})^{-1}, \boldsymbol{G}\right\rangle - n \overset{(6)}{=} (1+Mr)^2 (\sigma_{\boldsymbol{x}}(\boldsymbol{G}) + n) - n \\
&= (1+Mr)^2 \sigma_{\boldsymbol{x}}(\boldsymbol{G}) + n\left((1+Mr)^2 - 1\right) \\
&= (1+Mr)^2 \sigma_{\boldsymbol{x}}(\boldsymbol{G}) + 2nMr\left(1 + \frac{Mr}{2}\right) \\
&\le (1+Mr)^2 \left(\sigma_{\boldsymbol{x}}(\boldsymbol{G}) + \frac{2nMr}{1+Mr}\right).
\end{aligned}$$

□

Now we turn to the main proof of Theorem 4.2.

Proof: The derivation is same as [19, Theorem 4.9] by using Lemma A.4. Denote $\lambda_k \triangleq \lambda_f(\boldsymbol{x}_k)$, $\sigma_k \triangleq \sigma_{\boldsymbol{x}_k}(\boldsymbol{G}_k)$ for $k \ge 0$. In view of [19, Theorem 4.7], shown in Appendix B.1, since the initial condition $\frac{\ln 2}{4\kappa(2n+1)} \le \frac{\ln \frac{3}{2}}{4\kappa}$, the first relation in Eq. (19) is indeed true, and also

$$M \sum_{i=0}^{k} \lambda_i \overset{(51)}{\le} M\lambda_0 \sum_{i=0}^{k} \left(1 - \frac{\mu}{2L}\right)^i \le \frac{2L}{\mu} M\lambda_0 \overset{(18)}{\le} \frac{\ln 2}{2(2n+1)}. \tag{30}$$

Denote $\theta_k := \sigma_k + 2nM\lambda_k$. Let us show by induction that for all $k \ge 0$, we have

$$\mathbb{E}\theta_k \le \left(1 - \frac{1}{n}\right)^k \frac{2nL}{\mu}. \tag{31}$$

Indeed, since $\boldsymbol{G}_0 = L\boldsymbol{I}_n$ and $\mu \boldsymbol{I}_n \le \nabla^2 f(\boldsymbol{x}_0) \le L\boldsymbol{I}_n$, we have

$$\begin{aligned}
\theta_0 &= \sigma_0 + 2nM\lambda_0 \overset{(6)}{=} \mathrm{tr}\left(\nabla^2 f(\boldsymbol{x}_0)^{-1}\boldsymbol{G}_0\right) - n + 2nM\lambda_0 \\
&\le \mathrm{tr}\left(\nabla^2 f(\boldsymbol{x}_0)^{-1} \cdot \frac{L}{\mu}\nabla^2 f(\boldsymbol{x}_0)\right) - n + 2nM\lambda_0 = n\left(\frac{L}{\mu} - 1\right) + \frac{n\ln 2}{2\kappa(2n+1)} \le \frac{nL}{\mu}.
\end{aligned} \tag{32}$$

Therefore, for $k = 0$, Eq. (31) is satisfied. Now suppose that it is also satisfied for some $k \ge 0$.

Since $\nabla^2 f(\boldsymbol{x}_k) \preceq \boldsymbol{G}_k$, we know that

$$\boldsymbol{G}_k - \nabla^2 f(\boldsymbol{x}_k) \preceq \sigma_k \nabla^2 f(\boldsymbol{x}_k),$$

or equivalently,

$$\boldsymbol{G}_k \preceq (1 + \sigma_k)\nabla^2 f(\boldsymbol{x}_k).$$

Since Eq. (30) shows $M\lambda_k \le 2$, we can apply Lemma A.3, leading to

$$r_k \triangleq \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{x}_k} \le \lambda_k, \tag{33}$$

and

$$\begin{aligned}\lambda_{k+1} &\le \left(1+\frac{M\lambda_k}{2}\right)\frac{\sigma_k+\frac{M\lambda_k}{2}}{1+\sigma_k}\lambda_k \le \left(1+\frac{M\lambda_k}{2}\right)\left(\sigma_k+2nM\lambda_k\right)\lambda_k \\ &\le e^{2M\lambda_k}\left(\sigma_k+2nM\lambda_k\right)\lambda_k.\end{aligned} \tag{34}$$

Further, by Lemma A.4, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{u}_k}\sigma_{k+1} &\le \left(1-\frac{1}{n}\right)\left(1+Mr_k\right)^2\left(\sigma_k+\frac{2nMr_k}{1+Mr_k}\right) \\ &\overset{(33)}{\le} \left(1-\frac{1}{n}\right)\left(1+M\lambda_k\right)^2\left(\sigma_k+2nM\lambda_k\right) \\ &\le \left(1-\frac{1}{n}\right)e^{2M\lambda_k}\left(\sigma_k+2nM\lambda_k\right).\end{aligned}$$

Note that $\frac{1}{2}\le 1-\frac{1}{n}$ because $n\ge 2$, and take expectation for all randomness. We obtain

$$\begin{aligned}&\mathbb{E}\left[\sigma_{k+1}+2nM\lambda_{k+1}\right] \\ \le\;& \mathbb{E}\left(1-\frac{1}{n}\right)e^{2M\lambda_k}\left(\sigma_k+2nM\lambda_k\right)+e^{2M\lambda_k}2nM\left(\sigma_k+2nM\lambda_k\right)\lambda_k \\ \le\;& \mathbb{E}\left(1-\frac{1}{n}\right)e^{2M\lambda_k}\left(\sigma_k+2nM\lambda_k\right)+\left(1-\frac{1}{n}\right)e^{2M\lambda_k}4nM\left(\sigma_k+2nM\lambda_k\right)\lambda_k \\ =\;& \mathbb{E}\left(1-\frac{1}{n}\right)e^{2M\lambda_k}\left(1+4nM\lambda_k\right)\left(\sigma_k+2nM\lambda_k\right) \\ \le\;& \mathbb{E}\left(1-\frac{1}{n}\right)e^{2(2n+1)M\lambda_k}\left(\sigma_k+2nM\lambda_k\right) \\ \overset{(51)}{\le}\;& \mathbb{E}\left(1-\frac{1}{n}\right)e^{2(2n+1)M\lambda_0\left(1-\frac{\mu}{2L}\right)^k}\left(\sigma_k+2nM\lambda_k\right).\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\theta_{k+1} &\le \left(1-\frac{1}{n}\right)e^{2(2n+1)M\lambda_0\left(1-\frac{\mu}{2L}\right)^k}\mathbb{E}\theta_k \\ &\le \left(1-\frac{1}{n}\right)^{k+1}e^{2(2n+1)M\lambda_0\sum_{i=0}^{k}\left(1-\frac{\mu}{2L}\right)^i}\mathbb{E}\theta_0 \qquad (35) \\ &\le \left(1-\frac{1}{n}\right)^{k+1}e^{4\kappa(2n+1)M\lambda_0}\theta_0 \\ &\overset{\substack{(18)\\(32)}}{\le} \left(1-\frac{1}{n}\right)^{k+1}2n\kappa. \qquad (36)\end{aligned}$$

Thus, Eq. (31) is proved.

Let us fix some $k\ge 0$. Since $\lambda_k\ge 0$, we have

$$\mathbb{E}\sigma_k\le\mathbb{E}\sigma_k+2nM\lambda_k=\mathbb{E}\theta_k\overset{(31)}{\le}\left(1-\frac{1}{n}\right)^k 2n\kappa.$$

This proves the second relation in Eq. (19). Finally,

$$\begin{aligned}\mathbb{E}\lambda_{k+1}/\lambda_k &\overset{(34)}{\le} \mathbb{E}e^{2M\lambda_k}\theta_k\le\mathbb{E}e^{2(2n+1)M\lambda_k}\theta_k\overset{(51)}{\le}e^{2(2n+1)M\lambda_0\left(1-\frac{\mu}{2L}\right)^k}\mathbb{E}\theta_k \\ &\overset{(35)}{\le}\left(1-\frac{1}{n}\right)^k e^{2(2n+1)M\lambda_0\sum_{i=0}^{k}\left(1-\frac{\mu}{2L}\right)^i}\theta_0\overset{(18),(32)}{\le}\left(1-\frac{1}{n}\right)^k 2n\kappa.\end{aligned}$$

Thus, Eq. (20) is proved. □

A.5 Proof of Theorem 4.3

Proof: We denote the sequence $\{\eta_k\}$ that satisfies

$$\operatorname{tr}\left(\boldsymbol{G}_k - \nabla^2 f(\boldsymbol{x}_k)\right) = \eta_k \operatorname{tr}\left(\nabla^2 f(\boldsymbol{x}_k)\right), \tag{37}$$

and

$$\lambda_k \triangleq \lambda_f(\boldsymbol{x}_{k+1}), r_k \triangleq \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{x}_k}, \sigma_k \triangleq \sigma_{\nabla^2 f(\boldsymbol{x}_k)}(\boldsymbol{G}_k).$$

From Theorem 2.3, we obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{u}_k} \operatorname{tr}\left(\boldsymbol{G}_{k+1} - \nabla^2 f(\boldsymbol{x}_{k+1})\right) &\le \left(1 - \frac{1}{n}\right) \operatorname{tr}\left(\tilde{\boldsymbol{G}}_k - \nabla^2 f(\boldsymbol{x}_{k+1})\right) \\
&\le \left(1 - \frac{1}{n}\right) \operatorname{tr}\left(\left(1 + Mr_k\right) \boldsymbol{G}_k - \frac{1}{1 + Mr_k} \nabla^2 f(\boldsymbol{x}_k)\right) \\
&\overset{(37)}{=} \left(1 - \frac{1}{n}\right) \left(\left(1 + Mr_k\right)\left(1 + \eta_k\right) - \frac{1}{1 + Mr_k}\right) \operatorname{tr}\left(\nabla^2 f(\boldsymbol{x}_k)\right) \\
&\le \left(1 - \frac{1}{n}\right) \left(\left(1 + Mr_k\right)^2 \left(1 + \eta_k\right) - 1\right) \operatorname{tr}\left(\nabla^2 f(\boldsymbol{x}_{k+1})\right),
\end{aligned}$$

where the first inequality uses Theorem 2.3 and the second inequality uses Lemma A.2. The last inequality uses Lemma A.2 again.

We can apply [19, Theorem 4.7], shown in Appendix B.1, since the initial condition $\frac{\ln 2}{4\kappa(1+2n\kappa)} \le \frac{\ln \frac{3}{2}}{4\kappa}$, and our algorithm can be viewed as a specific version of theirs. Hence $M\lambda_k \le \left(1 - \frac{\mu}{2L}\right)^k M\lambda_0 < 2$, and using Lemma A.3, we obtain $r_k \le \lambda_k$. Therefore

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{u}_k} \eta_{k+1} &\le \left(1 - \frac{1}{n}\right) \left(\left(1 + Mr_k\right)^2 \left(1 + \eta_k\right) - 1\right) \\
&\le \left(1 - \frac{1}{n}\right) \left(1 + Mr_k\right)^2 \left(\eta_k + 1 - \frac{1}{\left(1 + Mr_k\right)^2}\right) \\
&\le \left(1 - \frac{1}{n}\right) \left(1 + Mr_k\right)^2 \left(\eta_k + \frac{2Mr_k + (Mr_k)^2}{\left(1 + Mr_k\right)^2}\right) \\
&\le \left(1 - \frac{1}{n}\right) \left(1 + Mr_k\right)^2 \left(\eta_k + 2Mr_k\right) \\
&\le \left(1 - \frac{1}{n}\right) \left(1 + M\lambda_k\right)^2 \left(\eta_k + 2M\lambda_k\right).
\end{aligned} \tag{38}$$

Moreover, since $\mu \boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}_n$, we have

$$\sigma_k = \operatorname{tr}\left(\left(\boldsymbol{G}_k - \nabla^2 f(\boldsymbol{x}_k)\right) \nabla^2 f(\boldsymbol{x}_k)^{-1}\right) \le \frac{1}{\mu} \operatorname{tr}\left(\boldsymbol{G}_k - \nabla^2 f(\boldsymbol{x}_k)\right) \overset{(37)}{=} \frac{\eta_k}{\mu} \operatorname{tr}\left(\nabla^2 f(\boldsymbol{x}_k)\right) \le \eta_k n\kappa. \tag{39}$$

Applying Lemma A.3 and noting that $\boldsymbol{G}_k \preceq \left(1 + \sigma_k\right) \nabla^2 f(\boldsymbol{x}_k)$, we get

$$\lambda_{k+1} \le \left(1 + \frac{M\lambda_k}{2}\right) \frac{\sigma_k + \frac{M\lambda_k}{2}}{1 + \sigma_k} \lambda_k \overset{(39)}{\le} \left(1 + \frac{M\lambda_k}{2}\right) \left(\eta_k n\kappa + \frac{M\lambda_k}{2}\right) \lambda_k. \tag{40}$$

Combining Eq. (40) and Eq. (38), we obtain

$$\begin{aligned}
&\mathbb{E}\,\eta_{k+1} + 2M\lambda_{k+1} \\
&\le \mathbb{E}\,\left(1-\frac{1}{n}\right)(1+M\lambda_k)^2(\eta_k + 2M\lambda_k) + \left(1+\frac{M\lambda_k}{2}\right)\left(\eta_k n\kappa + \frac{M\lambda_k}{2}\right)2M\lambda_k \\
&\le \mathbb{E}\,\left(1-\frac{1}{n}\right)(1+M\lambda_k)^2(\eta_k + 2M\lambda_k) + (1+M\lambda_k)(\eta_k + 2M\lambda_k)\,2n\kappa M\lambda_k \\
&\le \mathbb{E}\,\left(1-\frac{1}{n}\right)(1+M\lambda_k)^2(\eta_k + 2M\lambda_k) + \left(1-\frac{1}{n}\right)(1+M\lambda_k)^2(\eta_k + 2M\lambda_k)\,4n\kappa M\lambda_k \\
&\le \mathbb{E}\,\left(1-\frac{1}{n}\right)(1+M\lambda_k)^2(\eta_k + 2M\lambda_k)(4n\kappa M\lambda_k + 1) \\
&\le \mathbb{E}\left(1-\frac{1}{n}\right)e^{2M\lambda_k + 4n\kappa M\lambda_k}(\eta_k + 2M\lambda_k),
\end{aligned} \tag{41}$$

where the third inequality uses $1 \le 2\left(1-\frac{1}{n}\right)$ because $n \ge 2$. Now we define

$$\theta_{k+1} \triangleq \eta_k + 2M\lambda_k.$$

In the following, we prove by induction that

$$\mathbb{E}\theta_k \le \left(1-\frac{1}{n}\right)^k 2\kappa. \tag{42}$$

1) From the initial condition: $M\lambda_0 \le \frac{\ln 2}{4\kappa(1+2n\kappa)}$, and note that $\boldsymbol{G}_0 \preceq \kappa\nabla^2 f(\boldsymbol{x}_0)$, thus

$$\theta_0 = \eta_0 + 2M\lambda_0 \le \kappa - 1 + \frac{\ln 2}{2\kappa(1+2n\kappa)} \le \kappa. \tag{43}$$

2) For $k \ge 1$, from Eq. (41), we obtain

$$\begin{aligned}
\mathbb{E}\theta_{k+1} \quad &\le \quad \mathbb{E}\left(1-\frac{1}{n}\right)e^{(2+4n\kappa)M\lambda_k}\theta_k \\
&\overset{(51)}{\le} \quad \left(1-\frac{1}{n}\right)e^{(2+4n\kappa)M\lambda_0\left(1-\frac{\mu}{2L}\right)^k}\mathbb{E}\theta_k \\
&\le \quad \left(1-\frac{1}{n}\right)^{k+1} e^{(2+4n\kappa)M\lambda_0\sum_{i=0}^{k}\left(1-\frac{\mu}{2L}\right)^i}\theta_0 \qquad (44) \\
&\le \quad \left(1-\frac{1}{n}\right)^{k+1} e^{2\kappa(2+4n\kappa)M\lambda_0}\theta_0 \\
&\overset{(21),(43)}{\le} \quad \left(1-\frac{1}{n}\right)^{k+1} 2\kappa. \qquad (45)
\end{aligned}$$

Combining 1) and 2), Eq. (42) holds for all $k \ge 0$. Therefore,

$$\mathbb{E}\eta_k \le \mathbb{E}\theta_k \le \left(1-\frac{1}{n}\right)^k 2\kappa,$$

and

$$\mathbb{E}\sigma_k \overset{(39)}{\le} \mathbb{E}\,\eta_k n\kappa \le 2n\kappa^2\left(1-\frac{1}{n}\right)^k.$$

This proves the second relation in Eq. (22). Finally, from Eq. (40),

$$\begin{aligned}
\mathbb{E}\lambda_{k+1}/\lambda_k \quad &\overset{(40)}{\le} \quad \mathbb{E}e^{2M\lambda_k}n\kappa\theta_k \le \mathbb{E}e^{(2+4n\kappa)M\lambda_k}n\kappa\theta_k \overset{(51)}{\le} e^{(2+4n\kappa)M\lambda_0\left(1-\frac{\mu}{2L}\right)^k}n\kappa\mathbb{E}\theta_k \\
&\overset{(44)}{\le} \quad \left(1-\frac{1}{n}\right)^k n\kappa e^{(2+4n\kappa)M\lambda_0\sum_{i=0}^{k}\left(1-\frac{\mu}{2L}\right)^i}\theta_0 \\
&\overset{(21),(43)}{\le} \quad \left(1-\frac{1}{n}\right)^k 2n\kappa^2.
\end{aligned}$$

Thus, Eq. (23) is proved. □

A.6 Proof of Corollary 4.4

We first give a probabilistic perspective of Theorem 4.2, particularly for the random method.

Corollary A.5 *Suppose that in Algorithm 4, the random BFGS update is used, and that the initial point $\boldsymbol{x}_0$ is sufficiently close to the solution:*

$$M\lambda_f(\boldsymbol{x}_0) \le \frac{\ln 2}{4\kappa(2n+1)}. \tag{46}$$

Then with probability $1-\delta$*, we have*

$$\lambda_f(\boldsymbol{x}_{k+1}) \le \frac{2n^3\kappa}{\delta}\left(1-\frac{1}{n+1}\right)^k \lambda_f(\boldsymbol{x}_k),\ \textit{for all } k \in \mathbb{N}. \tag{47}$$

Proof: Note that $\lambda_f(\boldsymbol{x}_k) \ge 0$. Using Markov's inequality, we have

$$\mathbb{P}\left(\frac{\lambda_f(\boldsymbol{x}_{k+1})}{\lambda_f(\boldsymbol{x}_k)} \ge \frac{2n\kappa}{\epsilon_k}\left(1-\frac{1}{n}\right)^k\right) \le \frac{\mathbb{E}\lambda_f(\boldsymbol{x}_{k+1})/\lambda_f(\boldsymbol{x}_k)}{\frac{2n\kappa}{\epsilon_k}\left(1-\frac{1}{n}\right)^k} \overset{(20)}{\le} \epsilon_k. \tag{48}$$

Choosing $\epsilon_k = \delta(1-q)q^k$ for some positive $q<1$, we have

$$\begin{aligned}\mathbb{P}\left(\frac{\lambda_f(\boldsymbol{x}_{k+1})}{\lambda_f(\boldsymbol{x}_k)} \ge \frac{2n\kappa}{\epsilon_k}\left(1-\frac{1}{n}\right)^k, \exists\, k \in \mathbb{N}\right) \quad &\le \quad \sum_{k=0}^{\infty}\mathbb{P}\left(\frac{\lambda_f(\boldsymbol{x}_{k+1})}{\lambda_f(\boldsymbol{x}_k)} \ge \frac{2n\kappa}{\epsilon_k}\left(1-\frac{1}{n}\right)^k\right) \\ &\overset{(48)}{\le} \quad \sum_{k=0}^{\infty}\epsilon_k = \sum_{k=0}^{\infty}\delta(1-q)q^k = \delta.\end{aligned}$$

Therefore, we obtain with probability $1-\delta$,

$$\lambda_f(\boldsymbol{x}_{k+1}) \le \left(\frac{1-\frac{1}{n}}{q}\right)^k \cdot \frac{2n\kappa}{(1-q)\delta}\cdot\lambda_f(\boldsymbol{x}_k), \forall k \in \mathbb{N}.$$

If we set $q = 1-1/n^2$, we could obtain with probability $1-\delta$,

$$\lambda_f(\boldsymbol{x}_{k+1}) \le \frac{2n^3\kappa}{\delta}\left(1+\frac{1}{n}\right)^{-k}\lambda_f(\boldsymbol{x}_k) = \frac{2n^3\kappa}{\delta}\left(1-\frac{1}{n+1}\right)^k\lambda_f(\boldsymbol{x}_k),\ \text{for all } k \in \mathbb{N}.$$

□

Moreover, Theorem 4.3 could also convert to such formulation with the same argument.

Corollary A.6 *For Algorithm 4 with the random SR1 update, suppose that the initial point $\boldsymbol{x}_0$ is sufficiently close to the solution:*

$$M\lambda_f(\boldsymbol{x}_0) \le \frac{\ln 2}{4\kappa(2n\kappa+1)}. \tag{49}$$

Then with probability $1-\delta$*, we have*

$$\lambda_f(\boldsymbol{x}_{k+1}) \le \frac{2n^3\kappa^2}{\delta}\left(1-\frac{1}{n+1}\right)^k\lambda_f(\boldsymbol{x}_k), \forall k \in \mathbb{N}. \tag{50}$$

The proof is totally same as the random BFGS method, so we omit it. Now we turn to the proof of Corollary 4.4.

Proof: 1) For the greedy methods, there is no randomness. We can apply Theorems 4.2 and 4.3 directly.

For the greedy BFGS method, Theorem 4.2 already shows the superlinear convergence rate of $\lambda_f(\boldsymbol{x}_k)$. Now we combine this result with [19, Theorem 4.7] to give the whole period convergence estimator. Denote by $k_1 \ge 0$ the number of the first iteration, for which

$$\left(1-\frac{1}{2\kappa}\right)^{k_1} \le \frac{1}{2n+1}.$$

Clearly, $k_1 \leq 2\kappa \ln(2n+1)$.

Since the initial point $\boldsymbol{x}_0$ is sufficiently close to the solution: $M\lambda_f(\boldsymbol{x}_0) \leq \frac{\ln\frac{3}{2}}{4\kappa}$, and reusing the result of [19, Theorem 4.7], we obtain

$$M\lambda_f(\boldsymbol{x}_{k_1}) \leq M\left(1-\frac{1}{2\kappa}\right)^{k_1}\lambda_f(\boldsymbol{x}_0) \leq \frac{\ln 2}{4\kappa\,(2n+1)},$$

which satisfies the initial condition in Theorem 4.2.

Denote by $k_2 \geq 0$ the number of the first iteration, for which

$$\left(1-\frac{1}{n}\right)^{k_2} 2n\kappa \leq \frac{1}{2}.$$

Clearly, $k_2 \leq n\ln(4n\kappa)$.

By Theorem 4.2, for all $k \geq 0$, we have

$$\lambda_f(\boldsymbol{x}_{k_1+k_2+k+1}) \leq \left(1-\frac{1}{n}\right)^{k_2+k} 2n\kappa\,\lambda_f(\boldsymbol{x}_{k_1+k_2+k}) \leq \left(1-\frac{1}{n}\right)^{k}\frac{1}{2}\,\lambda_f(\boldsymbol{x}_{k_1+k_2+k}).$$

Therefore,

$$\lambda_f(\boldsymbol{x}_{k_1+k_2+k}) \leq \left(1-\frac{1}{n}\right)^{k(k-1)/2}\left(\frac{1}{2}\right)^{k}\lambda_f(\boldsymbol{x}_{k_1+k_2}),$$

and

$$\lambda_f(\boldsymbol{x}_{k_1+k_2}) \leq \left(1-\frac{1}{2\kappa}\right)^{k_1+k_2}\lambda_f(\boldsymbol{x}_0).$$

Finally, choosing $k_0 = k_1+k_2 = O\left(\max\{n,\kappa\}\ln(n\kappa)\right)$, we obtain

$$\lambda_f(\boldsymbol{x}_{k_0+k}) \leq \left(1-\frac{1}{n}\right)^{k(k-1)/2}\left(\frac{1}{2}\right)^{k}\left(1-\frac{1}{2\kappa}\right)^{k_0}\lambda_f(\boldsymbol{x}_0).$$

We can give similar analysis for the greedy SR1 method with $k_1 \leq 2\kappa\ln(2n\kappa+1)$ and $k_2 \leq n\ln(4n\kappa^2)$. Thus $k_0 = k_1+k_2 = O\left(\max\{n,\kappa\}\ln(n\kappa)\right)$ as well.

2) For the random methods, we can apply Corollary A.5 and Corollary A.6 to give the rates.

For the random BFGS method, Corollary A.5 already shows the superlinear convergence rate of $\lambda_f(\boldsymbol{x}_k)$. Now we combine this result with [19, Theorem 4.7] to give the whole period convergence estimator. Denote by $k_1 \geq 0$ the number of the first iteration, for which

$$\left(1-\frac{1}{2\kappa}\right)^{k_1} \leq \frac{1}{2n+1}.$$

Clearly, $k_1 \leq 2\kappa\ln(2n+1)$.

Since the initial point $\boldsymbol{x}_0$ is sufficiently close to the solution: $M\lambda_f(\boldsymbol{x}_0) \leq \frac{\ln\frac{3}{2}}{4\kappa}$, and reusing the result of [19, Theorem 4.7], we obtain

$$M\lambda_f(\boldsymbol{x}_{k_1}) \leq M\left(1-\frac{1}{2\kappa}\right)^{k_1}\lambda_f(\boldsymbol{x}_0) \leq \frac{\ln 2}{4\kappa\,(2n+1)},$$

which satisfies the initial condition in Corollary A.5.

Denote by $k_2 \geq 0$ the number of the first iteration, for which

$$\frac{2n^3\kappa}{\delta}\left(1-\frac{1}{n+1}\right)^{k_2} \leq \frac{1}{2}.$$

Clearly, $k_2 \leq (n+1)\ln(4n^3\kappa/\delta)$.

By Theorem 4.3, for all $k \geq 0$, we have with probability $1-\delta$ for any $\delta \in (0,1)$

$$\lambda_f(\boldsymbol{x}_{k_1+k_2+k+1}) \leq \frac{2n^3\kappa}{\delta}\left(1-\frac{1}{n+1}\right)^{k_2+k}\lambda_f(\boldsymbol{x}_{k_1+k_2+k}) \leq \left(1-\frac{1}{n+1}\right)^{k} \cdot \frac{1}{2}\lambda_f(\boldsymbol{x}_{k_1+k_2+k}).$$

Therefore,

$$\lambda_f(\boldsymbol{x}_{k_1+k_2+k}) \leq \left(1-\frac{1}{n+1}\right)^{k(k-1)/2}\left(\frac{1}{2}\right)^{k}\lambda_f(\boldsymbol{x}_{k_1+k_2}),$$

and

$$\lambda_f(\boldsymbol{x}_{k_1+k_2}) \leq \left(1-\frac{1}{2\kappa}\right)^{k_1+k_2}\lambda_f(\boldsymbol{x}_0).$$

Finally, choosing $k_0 = k_1 + k_2 = O\left(\max\{n,\kappa\}\ln(n\kappa/\delta)\right)$, we obtain

$$\lambda_f(\boldsymbol{x}_{k_0+k}) \leq \left(1-\frac{1}{n+1}\right)^{k(k-1)/2}\left(\frac{1}{2}\right)^{k}\left(1-\frac{1}{2\kappa}\right)^{k_0}\lambda_f(\boldsymbol{x}_0).$$

We can give similar analysis for the random SR1 method with $k_1 \leq 2\kappa\ln(2n\kappa+1)$ and $k_2 \leq (n+1)\ln(4n^3\kappa^2/\delta)$. Thus $k_0 = k_1 + k_2 = O\left(\max\{n,\kappa\}\ln(n\kappa/\delta)\right)$ as well. □

B Auxiliary Theorem

Algorithm 5 Quasi-Newton Method [19, Scheme (4.17)]

```
Initialization: Choose $\boldsymbol{x}_0 \in \mathbb{R}^n$. Set $\boldsymbol{G}_0 = L\boldsymbol{I}_n$.
for $k \geq 0$ do
    Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{G}_k^{-1}\nabla f(\boldsymbol{x}_k)$.
    Compute $r_k = \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_{\boldsymbol{x}_k}$ and set $\tilde{\boldsymbol{G}}_k = (1 + Mr_k)\boldsymbol{G}_k$.
    Choose $\boldsymbol{u}_k \in \mathbb{R}^n$ and $\tau_k \in [0,1]$.
    Compute $\boldsymbol{G}_{k+1} = \mathrm{Broyd}_{\tau_k}\left(\tilde{\boldsymbol{G}}_k, \nabla^2 f(\boldsymbol{x}_{k+1}), \boldsymbol{u}_k\right)$.
end for
```

Theorem B.1 (Rodomanov and Nesterov [19] Theorem 4.7) *Under Algorithm 5, suppose the initial point $\boldsymbol{x}_0$ is sufficiently close to the solution:*

$$M\lambda_f(\boldsymbol{x}_0) \leq \frac{\ln\frac{3}{2}}{4\kappa}.$$

Then, for all $k \geq 0$, we have

$$\nabla^2 f(\boldsymbol{x}_k) \preceq \boldsymbol{G}_k \preceq e^{2M\sum_{i=0}^{k-1}\lambda_f(\boldsymbol{x}_i)}\frac{L}{\mu}\nabla^2 f(\boldsymbol{x}_k) \preceq \frac{3L}{2\mu}\nabla^2 f(\boldsymbol{x}_k),$$

and

$$\lambda_f(\boldsymbol{x}_k) \leq \left(1-\frac{\mu}{2L}\right)^{k}\lambda_f(\boldsymbol{x}_0). \tag{51}$$

Remark B.2 *Note that the choice of $\boldsymbol{u}_k$ and the update rule in Algorithm 5 are arbitrary, thus our algorithms can be viewed as a special case. Therefore, Theorem B.1 always holds as long as the initial point is sufficiently close to the solution.*

C Missing Experiments and Detail

C.1 Condition-number-free Property of Our Modified BFGS Update

For the matrix approximation task in Section 5, we list the variation of $\sigma_{\boldsymbol{A}}(\boldsymbol{G}_k)$ with various condition numbers κ shown in Figure 4. We find scaled directions ($\boldsymbol{L}^\top\boldsymbol{u}$) have the same convergence rates in different condition numbers. Thus we conclude our scaled directions give a condition-number-free convergence rate as our theory shows. However, directly selecting random directions (RaBFGSv1) converges much slower than the scaled random directions (RaBFGSv2), and the large condition number could cause slow convergence.

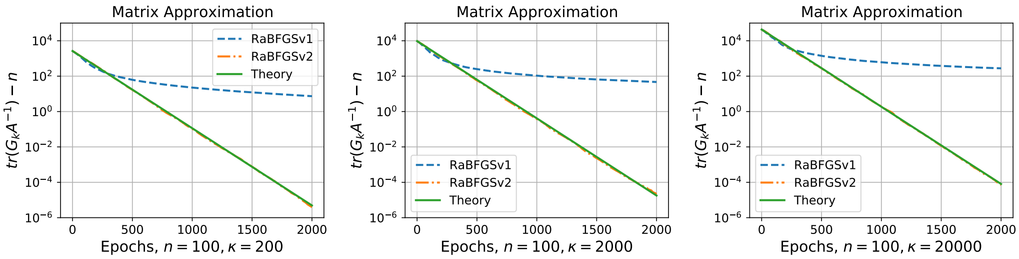


Figure 4: The variation of $\sigma_{\boldsymbol{A}}(\boldsymbol{G}_k)$ during our Random BFGS (RaBFGSv2) update and the original random version (RaBFGSv1) for approximating a matrix $\boldsymbol{A}$ that $\mu\boldsymbol{I}_n \preceq \boldsymbol{A} \preceq L\boldsymbol{I}_n$ with various κ.

C.2 Experimental Settings in Section 5

Regularized Logistic Regression. The gradient of function f is

$$\nabla f(\boldsymbol{w}) = -\sum_{i=1}^{N} \frac{e^{y_i \boldsymbol{w}^\top \boldsymbol{x}_i}}{1 + e^{y_i \boldsymbol{w}^\top \boldsymbol{x}_i}} \cdot y_i \boldsymbol{x}_i + \gamma \boldsymbol{w},\ \boldsymbol{w} \in \mathbb{R}^n,$$

and the Hessian is

$$\nabla^2 f(\boldsymbol{w}) = \sum_{i=1}^{N} \frac{e^{y_i \boldsymbol{w}^\top \boldsymbol{x}_i}}{\left(1 + e^{y_i \boldsymbol{w}^\top \boldsymbol{x}_i}\right)^2} \cdot \boldsymbol{x}_i \boldsymbol{x}_i^\top + \gamma \boldsymbol{I}_n,\ \boldsymbol{w} \in \mathbb{R}^n.$$

Thus $\gamma \boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{w}) \preceq L\boldsymbol{I}_n$ with $L = \lambda_{\max}(\boldsymbol{X}\boldsymbol{X}^\top)/4 + \gamma$, and $\kappa = L/\gamma$. In particular, the Hessian-vector product for this function can be computed with the similar complexity of that for computing the gradient, guaranteeing $O(nN)$ complexity in each step as below

$$\nabla^2 f(\boldsymbol{w})\boldsymbol{v} = \sum_{i=1}^{N} \frac{e^{y_i \boldsymbol{w}^\top \boldsymbol{x}_i}}{\left(1 + e^{y_i \boldsymbol{w}^\top \boldsymbol{x}_i}\right)^2} \cdot \left(\boldsymbol{x}_i^\top \boldsymbol{v}\right) \cdot \boldsymbol{x}_i + \gamma \boldsymbol{v},\ \boldsymbol{w} \in \mathbb{R}^n.$$

We compare $\|\nabla f(\boldsymbol{w}_k)\|$ obtained by different methods.

Regularized Log-Sum-Exp. We need access to the gradient of function f:

$$\nabla f(\boldsymbol{x}) = g(\boldsymbol{x}) + \sum_{j=1}^{m} \left(\boldsymbol{c}_j^\top \boldsymbol{x}\right) \boldsymbol{c}_j + \gamma \boldsymbol{x}, g(\boldsymbol{x}) := \sum_{j=1}^{m} \pi_j(\boldsymbol{x}) \boldsymbol{c}_j,$$

where

$$\pi_j(\boldsymbol{x}) := \frac{e^{\boldsymbol{c}_j^\top \boldsymbol{x} - b_j}}{\sum_{i=1}^{m} e^{\boldsymbol{c}_i^\top \boldsymbol{x} - b_i}} \in [0,1], j = 1, \ldots, m.$$

Thus we can see $\nabla f(\boldsymbol{x})$ can be computed in $O(mn)$ operations for a given point $\boldsymbol{x} \in \mathbb{R}^n$. Moreover, we have analytic Hessian and Hessian-vector product expression below:

$$\nabla^2 f(\boldsymbol{x}) = \sum_{j=1}^{m} \left(\pi_j(\boldsymbol{x}) + 1\right) \boldsymbol{c}_j \boldsymbol{c}_j^\top - g(\boldsymbol{x}) g(\boldsymbol{x})^\top + \gamma I_n,$$

and for a given direction $\boldsymbol{h} \in \mathbb{R}^n$,

$$\nabla^2 f(\boldsymbol{x})\boldsymbol{h} = \sum_{j=1}^{m} \left(\pi_j(\boldsymbol{x}) + 1\right) \left(\boldsymbol{c}_j^\top \boldsymbol{h}\right) \boldsymbol{c}_j - \left(g(\boldsymbol{x})^\top \boldsymbol{h}\right) g(\boldsymbol{x}) + \gamma \boldsymbol{h}.$$

Hence, the Lipschitz constant of the gradient of f can be taken as $L = 2\lambda_{\max}(CC^\top) + \gamma$, and $\kappa = L/\gamma$. As mentioned in [19, Section 5.1], the strong self-concordancy parameter is $M = 2$. So we apply the *correction strategy* ($\tilde{\boldsymbol{G}}_k = (1 + Mr_k)\boldsymbol{G}_k$).

We also adopt the same synthetic data as [19, Section 5.1]. First, we generate a collection of random vectors $\hat{\boldsymbol{c}}_1, \ldots, \hat{\boldsymbol{c}}_m$ with entries, uniformly distributed in the interval $[-1, 1]$. Then we generate $b_1, \ldots, b_m$ from the same distribution. Using this data, we form a preliminary function

$$\hat{f}(\boldsymbol{x}) := \ln\left(\sum_{j=1}^{m} e^{\hat{\boldsymbol{c}}_j^\top \boldsymbol{x} - b_j}\right),$$

and finally define

$$\boldsymbol{c}_j := \hat{\boldsymbol{c}}_j - \nabla \hat{f}(\mathbf{0}), j = 1, \ldots, m.$$

Note that by construction

$$\nabla f(\mathbf{0}) = \frac{1}{\sum_{i=1}^{m} e^{-b_i}} \sum_{j=1}^{m} e^{-b_j}\left(\hat{\boldsymbol{c}}_j - \nabla \hat{f}(\mathbf{0})\right) = \mathbf{0}.$$

So the unique minimizer of our test function is $\boldsymbol{x}^* = \mathbf{0}$. The starting point $\boldsymbol{x}_0$ for all methods is the same and generated randomly from the uniform distribution on the standard Euclidean sphere of radius $1/n$ centered at the minimizer, i.e., $\boldsymbol{x}_0 \sim \text{Unif}\left(\frac{1}{n}\mathcal{S}^{n-1}\right)$. We compare $\|\nabla f(\boldsymbol{x}_k)\|$ obtained by different methods.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See conclusion in Section 6. The main limitation is that our current results do not give the superlinear convergence rate of the standard quasi-Newton methods with only gradient information. Our proposed greedy and random quasi-Newton methods follow Rodomanov and Nesterov [19]'s work, and need to employ Hessian-vector product, though totally complexity is $O(n^2)$ as standard quasi-Newton methods.
   (c) Did you discuss any potential negative societal impacts of your work? [No] The purpose of this work is to provide a better understanding of quasi-Newton methods.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions mainly include: strongly self-concordant (Definition 4.1), smooth and strongly-convex objective.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix for detail.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code are proprietary, but we will display the code if required.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and Appendix C.2 for detail.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We wants to compare the traing dynamic, and different trials may cause different nemerical results, which can not be observed clearly in one graph.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The experiments is certainly simple and easy to run under CPUs.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] These datasets are common.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] These datasets are common.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]