Journal of Machine Learning Research 27 (2026) 1-33 Submitted 1/26; Revised 6/26; Published 7/26

Symmetric Rank-k Methods

Chengchang Liu LIUCHENGCHANG@WESTLAKE.EDU.CN
Department of Artificial Intelligence, Westlake University
Hangzhou, China

Cheng Chen CHCHEN@SEI.ECNU.EDU.CN
Software Engineering Institute, East China Normal University
Shanghai, China

Luo Luo* LUOLUO@FUDAN.EDU.CN
School of Data Science, Fudan University
Shanghai, China
and
Shanghai Key Laboratory for Contemporary Applied Mathematics
Shanghai, China

Editor: Michael Mahoney

Abstract

This paper proposes a novel class of block quasi-Newton methods for convex optimization which we call symmetric rank-k (SR-k) methods. Each iteration of SR-k incorporates the curvature information with k Hessian-vector products achieved from the greedy or random strategy. We prove that SR-k methods have the local superlinear convergence rate of $\mathcal{O}\big((1-k/d)^{t(t-1)/2}\big)$ for minimizing smooth and strongly convex functions, where d is the problem dimension and t is the iteration counter. This is the first explicit superlinear convergence rate for block quasi-Newton methods, and it successfully explains why block quasi-Newton methods converge faster than ordinary quasi-Newton methods in practice. We also leverage the idea of SR-k methods to study the block BFGS and block DFP methods, showing their superior convergence rates.

Keywords: block quasi-Newton methods, Broyden family updates, superlinear convergence

1. Introduction

We study quasi-Newton methods for solving the minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}), \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is smooth and strongly convex. Quasi-Newton methods (Broyden, 1970b,a; Shanno, 1970; Broyden, 1967; Davidon, 1991; Byrd et al., 1987; Yuan, 1991; Asl et al., 2024) are widely recognized for their fast convergence rates and efficient updates, which attract growing attention in many fields such as statistics (Jamshidian and Jennrich, 1997; Zhang and Sutton, 2011; Bishwal, 2008), economics (Ludwig, 2007; Li et al., 2013) and machine learning (Goldfarb et al., 2020; Hennig and Kiefel, 2013; Liu and Luo, 2022; Liu et al., 2022;

$*$. The corresponding author

Lee et al., 2018; Qiu et al., 2024). Unlike standard Newton methods, which need to compute the Hessian and its inverse, quasi-Newton methods go along the descent direction by the following scheme

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t),$$

where $\mathbf{G}_t \in \mathbb{R}^{d\times d}$ is an estimator of the Hessian $\nabla^2 f(\mathbf{x}_t)$. The most popular ways to construct the Hessian estimator are the Broyden family updates, including the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method (Broyden, 1970b,a; Shanno, 1970), the Davidon–Fletcher–Powell (DFP) method (Davidon, 1991; Fletcher and Powell, 1963), and the symmetric rank-one (SR1) method (Broyden, 1967; Davidon, 1991).

The classical quasi-Newton methods with Broyden family updates (Broyden, 1970a,b) find the Hessian estimator $\mathbf{G}_{t+1}$ for the next round by the secant equation

$$\mathbf{G}_{t+1}(\mathbf{x}_{t+1} - \mathbf{x}_t) = \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t). \tag{2}$$

These methods have been proven to exhibit local superlinear convergence in the 1970s (Powell, 1971; Dennis and Moré, 1974; Broyden et al., 1973; Dai, 2002), and their non-asymptotic superlinear rates were established in recent years (Rodomanov and Nesterov, 2022, 2021b; Ye et al., 2023; Jin and Mokhtari, 2023; Jin et al., 2022). For example, Rodomanov and Nesterov (Rodomanov and Nesterov, 2022) showed the classical BFGS method enjoys the local superlinear rates of $\mathcal{O}\big((d\varkappa/t)^{t/2}\big)$, where $\varkappa > 1$ is the condition number. They also improved this result to $\mathcal{O}\big((\exp(d\ln(\varkappa)/t) - 1)^{t/2}\big)$ in consequent work (Rodomanov and Nesterov, 2021b). Later, Ye et al. (2023) showed the classical SR1 method converges with a local superlinear rate of $\mathcal{O}\big((d\ln(\varkappa)/t)^{t/2}\big)$.

Some recent work (Gower and Richtárik, 2017; Rodomanov and Nesterov, 2021a; Lin et al., 2022; Ji and Dai, 2023) proposed another type of quasi-Newton methods, which construct the Hessian estimator by the following equation

$$\mathbf{G}_{t+1} = \nabla^2 f(\mathbf{x}_{t+1})\mathbf{u}_t \tag{3}$$

where $\mathbf{u}_t \in \mathbb{R}^d$ is chosen by a greedy or randomized strategy. Rodomanov and Nesterov (2021a) established a local superlinear rate of $\mathcal{O}\big((1-1/(\varkappa d))^{t(t-1)/2}\big)$ for greedy quasi-Newton methods with Broyden family updates. Later, Lin et al. (2022) provided a condition-number free superlinear rate of $\mathcal{O}\big((1-1/d)^{t(t-1)/2}\big)$ for the greedy and randomized quasi-Newton methods with the specific BFGS and SR1 updates.

Block quasi-Newton methods construct the Hessian estimator along multiple directions at each iteration. The study of these methods dates back to the 1980s. Schnabel (Schnabel, 1983) proposed the first block BFGS method by extending equation (2) to multiple secant equations, i.e., $\mathbf{G}_{t+1}(\mathbf{x}_{t+1} - \mathbf{x}_{t+1-j}) = \nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_{t+1-j})$ for all $j = 1, \ldots, k$. However, the resulting update of this method cannot guarantee that the Hessian estimator is symmetric, and it requires additionally modification to ensure the positive definiteness (Schnabel, 1983; O'Leary and Yeremin, 1994; Gao and Goldfarb, 2018; Lee and Sun, 2024). Another line of research in 2010s (Gao and Goldfarb, 2018; Gower et al., 2016; Gower and Richtárik, 2017; Kovalev et al., 2020; Boutet et al., 2020) considers the variants of block BFGS methods that generalize condition (3) to

$$\mathbf{G}_{t+1}\mathbf{U}_t = \nabla^2 f(\mathbf{x}_{t+1})\mathbf{U}_t, \tag{4}$$

where $\mathbf{U}_t = [\mathbf{u}_t^{(1)}, \cdots, \mathbf{u}_t^{(k)}] \in \mathbb{R}^{d\times k}$ indicates multiple directions constructed by deterministic or randomized methods. Since the Hessian-matrix product $\nabla^2 f(\mathbf{x}_{t+1})\mathbf{U}_t$ can be efficiently achieved by accessing the Hessian-vector products $\{\nabla^2 f(\mathbf{x}_{t+1})\mathbf{u}_t^{(j)}\}_{j=1}^k$ in parallel (Gao and Goldfarb, 2018), these block BFGS methods usually exhibit better empirical performance than the ordinary quasi-Newton methods without block fashion updates. Gao and Goldfarb (2018); Kovalev et al. (2020) provided asymptotic local superlinear rates for block BFGS methods based on the condition (4), but the advantage of block quasi-Newton methods over ordinary quasi-Newton methods is still unclear in theory. This naturally leads to the following question:

Can we provide a block quasi-Newton method with explicit superior convergence rate over the ordinary ones?

In this paper, we give an affirmative answer by proposing symmetric rank-k (SR-k) methods. We design a novel block update to construct the Hessian estimator $\mathbf{G}_{t+1} \in \mathbb{R}^{d\times d}$ that satisfies equation (4). We also provide the randomized and greedy strategies to determine directions in $\mathbf{U}_t \in \mathbb{R}^{d\times k}$ for SR-k methods, which lead to the explicit local superlinear convergence rate of $\mathcal{O}\big((1-k/d)^{t(t-1)/2}\big)$, where k is the number of directions used to approximate the Hessian at each iteration. For the case of $k = 1$, our methods reduce to randomized and greedy SR1 methods (Lin et al., 2022). For the case of $k \geq 2$, it is clear that SR-k methods have faster superlinear rates than existing greedy and randomized quasi-Newton methods (Lin et al., 2022; Rodomanov and Nesterov, 2021a). For the case of $k = d$, it recovers the quadratic convergence rate like standard Newton's method. We also follow the design of SR-k to propose the variants of randomized block BFGS methods (Gower et al., 2016; Gower and Richtárik, 2017; Kovalev et al., 2020) and a randomized block DFP method, resulting in faster explicit superlinear rates. We compare the proposed methods with existing quasi-Newton methods in Table 1.

Paper Organization The remainder of this paper is organized as follows. In Section 2, we introduce notations and assumptions throughout this paper. In Section 3, we propose the SR-k update in the view of matrix approximation. In Section 4, we propose the quasi-Newton methods with SR-k updates for minimizing smooth strongly convex functions and provide their superior local superlinear convergence rates. In Section 5, we propose the new randomized block BFGS methods and randomized block DFP method with explicit local superlinear convergence rates. In Section 6, we conduct numerical experiments to show the outperformance of the proposed methods. Finally, we conclude our work in Section 7.

2. Preliminaries

We first introduce the notations used in this paper. We use $\{\mathbf{e}_1, \cdots, \mathbf{e}_d\}$ to present the standard basis in space $\mathbb{R}^d$ and let $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ be the identity matrix. We use $\mathbf{S}^\dagger$ to denote the Moore-Penrose inverse of a matrix $\mathbf{S}$. We denote the trace of a square matrix by $\mathrm{tr}(\cdot)$. We use $\|\cdot\|$ to present the spectral norm of a matrix and the Euclidean norm of a vector. Given a positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, we denote the corresponding weighted norm as $\|\mathbf{x}\|_{\mathbf{A}} \triangleq (\mathbf{x}^\top \mathbf{A}\mathbf{x})^{1/2}$ for some $\mathbf{x} \in \mathbb{R}^d$. We use the notation $\|\mathbf{x}\|_{\mathbf{z}}$ to present $\|\mathbf{x}\|_{\nabla^2 f(\mathbf{z})}$ for positive definite Hessian $\nabla^2 f(\mathbf{z})$, if there is no ambiguity for the reference function $f(\cdot)$. We

Table 1: We summarize local convergence rates of existing and proposed quasi-Newton methods for strongly convex optimization, where $\varkappa$ denotes the condition number of the objective and t is the iteration counter. The second column displays the number of secant equations used to construct the Hessian estimator.

Methods	**#Equations**	**Convergence**
Classical BFGS (Rodomanov and Nesterov, 2022, 2021b)	1	$\mathcal{O}\Big(\Big(\frac{d\varkappa}{t}\Big)^{t/2}\Big)$ $\mathcal{O}\Big(\Big(\exp\Big(\frac{d\ln(\varkappa)}{t}\Big)-1\Big)^{t/2}\Big)$
Classical BFGS/DFP (Jin and Mokhtari, 2023)	1	$\mathcal{O}\Big(\Big(\frac{1}{t}\Big)^{t/2}\Big)$ (*)
Classical DFP (Rodomanov and Nesterov, 2022)	1	$\mathcal{O}\Big(\Big(\frac{d\varkappa^2}{t}\Big)^{t/2}\Big)$
Classical SR1 (Ye et al., 2023)	1	$\mathcal{O}\Big(\Big(\frac{d\ln(\varkappa)}{t}\Big)^{t/2}\Big)$
Greedy/Randomized Broyden (Rodomanov and Nesterov, 2021a; Lin et al., 2022)	1	$\mathcal{O}\Big(\Big(1-\frac{1}{\varkappa d}\Big)^{t(t-1)/2}\Big)$
Randomized BFGS (Lin et al., 2022)	1	$\mathcal{O}\Big(\Big(1-\frac{1}{d}\Big)^{t(t-1)/2}\Big)$
Greedy/Randomized SR1 (Lin et al., 2022)	1	$\mathcal{O}\Big(\Big(1-\frac{1}{d}\Big)^{t(t-1)/2}\Big)$
Block-BFGS (Gao and Goldfarb, 2018)	$k\in[d]$	asymptotic superlinear
Randomized Block-BFGS (v1) (Kovalev et al., 2020; Gower and Richtárik, 2017)	$k\in[d]$	asymptotic superlinear
Greedy/Randomized SR-k Algorithm 1	$k\in[d-1]$ $k=d$	$\mathcal{O}\Big(\Big(1-\frac{k}{d}\Big)^{t(t-1)/2}\Big)$ quadratic
Randomized Block-BFGS/DFP Algorithm 2	$k\in[d-1]$ $k=d$	$\mathcal{O}\Big(\Big(1-\frac{k}{d\varkappa}\Big)^{t(t-1)/2}\Big)$ quadratic
Faster Randomized Block-BFGS Algorithm 3	$k\in[d-1]$ $k=d$	$\mathcal{O}\Big(\Big(1-\frac{k}{d}\Big)^{t(t-1)/2}\Big)$ quadratic

(*) This convergence rate requires an additional assumption that the initial Hessian estimator be sufficiently close to the exact Hessian (Jin and Mokhtari, 2023, Corollary 3).

also define $\mathbf{E}_k(\mathbf{A}) \triangleq [\mathbf{e}_{i_1};\cdots;\mathbf{e}_{i_k}] \in \mathbb{R}^{d\times k}$, where $i_1,\ldots,i_k$ are the indices for the largest k diagonal entries of matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$.

Throughout this paper, we suppose the objective in problem (1) satisfies the following assumptions.

Assumption 2.1 *We assume the objective $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, i.e., there exists some constant $L \geq 0$ such that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Assumption 2.2 *We assume the objective $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex, i.e., there exists some constant $\mu > 0$ such that*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) - \frac{\lambda(1-\lambda)\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\lambda \in [0, 1]$.

We define the condition number as $\varkappa \triangleq L/\mu$. The following proposition shows the twice differentiable function has a bounded Hessian under Assumptions 2.1 and 2.2 (Nesterov, 2018).

Proposition 2.3 *Suppose the objective $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable and satisfies Assumptions 2.1 and 2.2, then it holds $\mu\mathbf{I}_d \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}_d$ for any $\mathbf{x} \in \mathbb{R}^d$.*

We also impose the assumption of strongly self-concordance (Rodomanov and Nesterov, 2021a; Lin et al., 2022) as follows.

Assumption 2.4 *We assume the objective $f : \mathbb{R}^d \to \mathbb{R}$ is M-strongly self-concordant, i.e., there exists some constant $M \geq 0$ such that*

$$\nabla^2 f(\mathbf{y}) - \nabla^2 f(\mathbf{x}) \preceq M\|\mathbf{y} - \mathbf{x}\|_{\mathbf{z}} \nabla^2 f(\mathbf{w})$$

holds for all $\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{z} \in \mathbb{R}^d$.

Noticing that the strongly convex function with Lipschitz continuous Hessian is strongly self-concordant (Rodomanov and Nesterov, 2021a).

Proposition 2.5 *Suppose the objective $f : \mathbb{R}^d \to \mathbb{R}$ satisfies Assumption 2.2 and has L_2-Lipschitz continuous Hessian, i.e., for some $L_2 \geq 0$, $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq L_2\|\mathbf{x} - \mathbf{y}\|$ holds for all $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^d$, then the function $f(\cdot)$ is M-strongly self-concordant with $M = L_2/\mu^{3/2}$.*

3. Symmetric Rank-k Updates

We propose the symmetric rank-k (SR-k) update for matrix approximation as follows.

Definition 3.1 (SR-k Update) *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{G} \in \mathbb{R}^{d\times d}$ be two positive-definite matrices with $\mathbf{A} \preceq \mathbf{G}$. For any matrix $\mathbf{U} \in \mathbb{R}^{d\times k}$, we define*

$$\text{SR-}k(\mathbf{G}, \mathbf{A}, \mathbf{U}) \triangleq \mathbf{G} - (\mathbf{G} - \mathbf{A})\mathbf{U}\big(\mathbf{U}^\top(\mathbf{G} - \mathbf{A})\mathbf{U}\big)^\dagger \mathbf{U}^\top(\mathbf{G} - \mathbf{A}). \tag{5}$$

Noticing that the formula of SR-k update contains a term of Moore-Penrose inverse, since the matrix $\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U}$ is possibly singular even if matrices $\mathbf{A},\mathbf{G}\in\mathbb{R}^{d\times d}$, and $\mathbf{U}\in\mathbb{R}^{d\times k}$ are full rank. This leads to its design and analysis be quite different from other block-type updates, i.e., block BFGS/DFP updates we study in Section 5.

In this section, we always use $\mathbf{G}_+$ to denote the output of SR-k update such that

$$\mathbf{G}_+ \triangleq \text{SR-}k(\mathbf{G},\mathbf{A},\mathbf{U}). \tag{6}$$

We provide the following lemma to show that SR-k update does not increase the deviation from target matrix $\mathbf{A}$.

Lemma 3.2 *Given any positive-definite matrices $\mathbf{A}\in\mathbb{R}^{d\times d}$, $\mathbf{G}\in\mathbb{R}^{d\times d}$ with $\mathbf{A}\preceq\mathbf{G}\preceq\eta\mathbf{A}$ for some $\eta\geq 1$, then $\mathbf{G}_+$ defined by (6) holds that*

$$\mathbf{A}\preceq\mathbf{G}_+\preceq\eta\mathbf{A}. \tag{7}$$

Proof According to the update rule (5), we have

$$\begin{aligned}
&\mathbf{G}_+-\mathbf{A}=(\mathbf{G}-\mathbf{A})-(\mathbf{G}-\mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U})^\dagger\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\\
&=\left(\mathbf{I}_d-(\mathbf{G}-\mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U})^\dagger\mathbf{U}^\top\right)(\mathbf{G}-\mathbf{A})\left(\mathbf{I}_d-\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U})^\dagger\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\right)\\
&\succeq\mathbf{0},
\end{aligned}$$

which indicates $\mathbf{G}_+\succeq\mathbf{A}$. On the other hand, the condition $\mathbf{G}\preceq\eta\mathbf{A}$ indicates

$$\mathbf{G}_+\preceq\eta\mathbf{A}-(\mathbf{G}-\mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U})^\dagger\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\preceq\eta\mathbf{A},$$

where the last inequality is by the condition $\mathbf{G}\succeq\mathbf{A}$. Hence, we finish the proof. ■

To evaluate the convergence of SR-k update, we introduce the quantity (Lin et al., 2022; Ye et al., 2023)

$$\tau_{\mathbf{A}}(\mathbf{G})\triangleq\operatorname{tr}(\mathbf{G}-\mathbf{A}), \tag{8}$$

to describe the difference between the target matrix $\mathbf{A}$ and the current estimator $\mathbf{G}$.

In the remainder of this section, we aim to establish the following convergence guarantee

$$\mathbb{E}[\tau_{\mathbf{A}}(\mathbf{G}_+)]\leq\left(1-\frac{k}{d}\right)\tau_{\mathbf{A}}(\mathbf{G}) \tag{9}$$

for approximating the target matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$ by SR-k iteration (6) with the appropriate choice of (random) matrix $\mathbf{U}\in\mathbb{R}^{d\times k}$. Note that if we take $k=1$, the equation (9) will reduce to $\mathbb{E}[\tau_{\mathbf{A}}(\mathbf{G}_+)]\leq(1-1/d)\,\tau_{\mathbf{A}}(\mathbf{G})$, which corresponds to the convergence result of randomized and greedy SR1 updates Lin et al. (2022).

Observe that we can split $\tau_{\mathbf{A}}(\mathbf{G}_+)$ as follows

$$\begin{aligned}
\tau_{\mathbf{A}}(\mathbf{G}_+)&\overset{(5),(6)}{=}\operatorname{tr}\big(\mathbf{G}-\mathbf{A}-(\mathbf{G}-\mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U})^\dagger\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\big)\\
&=\underbrace{\operatorname{tr}(\mathbf{G}-\mathbf{A})}_{\text{Part I}}-\underbrace{\operatorname{tr}\big((\mathbf{G}-\mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U})^\dagger\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\big)}_{\text{Part II}}.
\end{aligned} \tag{10}$$

Since Part I is equal to $\tau_{\mathbf{A}}(\mathbf{G})$ and Part II is nonnegative, the SR-k update can reduce the estimation error with regard to $\tau_{\mathbf{A}}(\cdot)$. To obtain (9), we only need to prove

$$\mathbb{E}\big[\mathrm{tr}\big((\mathbf{G}-\mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U})^\dagger\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\big)\big] \geq \frac{k}{d}\mathrm{tr}(\mathbf{G}-\mathbf{A}).$$

We first provide the following lemma to bound Part II (in the view of $\mathbf{R}=\mathbf{G}-\mathbf{A}$).

Lemma 3.3 *For a symmetric positive semi-definite matrix $\mathbf{R}\in\mathbb{R}^{d\times d}$ and a full rank matrix $\mathbf{U}\in\mathbb{R}^{d\times k}$ with $k\leq d$, it holds that*

$$\mathrm{tr}\big(\mathbf{R}\mathbf{U}\left(\mathbf{U}^\top\mathbf{R}\mathbf{U}\right)^\dagger\mathbf{U}^\top\mathbf{R}\big) \geq \mathrm{tr}\big(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{R}\big). \tag{11}$$

Proof Let $\mathbf{U}=\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top$ be the reduced SVD of $\mathbf{U}\in\mathbb{R}^{d\times k}$, where $\mathbf{Q}\in\mathbb{R}^{d\times k}$, $\mathbf{V}\in\mathbb{R}^{k\times k}$ are (column) orthonormal (i.e. $\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}_k$ and $\mathbf{V}^\top\mathbf{V}=\mathbf{I}_k$) and $\mathbf{\Sigma}\in\mathbb{R}^{k\times k}$ is diagonal, then the right-hand side of inequality (11) can be written as

$$\mathrm{tr}\big(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{R}\big) = \mathrm{tr}\big(\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\left(\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top\right)^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\big) = \mathrm{tr}\big(\mathbf{Q}\mathbf{Q}^\top\mathbf{R}\big).$$

Consequently, we upper bound the left-hand side of inequality (11) as follows

$$\begin{aligned}
&\mathrm{tr}\big(\mathbf{R}\mathbf{U}\left(\mathbf{U}^\top\mathbf{R}\mathbf{U}\right)^\dagger\mathbf{U}^\top\mathbf{R}\big) = \mathrm{tr}\big(\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\left(\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\right)^\dagger\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\big)\\
&= \mathrm{tr}\big(\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\left(\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\right)^\dagger\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\big)\\
&\quad + \mathrm{tr}\big(\big(\mathbf{I}_d-\mathbf{Q}\mathbf{Q}^\top\big)^{1/2}\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\left(\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\right)^\dagger\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\big(\mathbf{I}_d-\mathbf{Q}\mathbf{Q}^\top\big)^{1/2}\big)\\
&\geq \mathrm{tr}\big(\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\left(\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\right)^\dagger\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\big)\\
&= \mathrm{tr}\big((\mathbf{V}\mathbf{\Sigma})^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\left(\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\right)^\dagger\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{\Sigma}\mathbf{V}^\top)^{-1}\big)\\
&= \mathrm{tr}\big((\mathbf{V}\mathbf{\Sigma})^{-1}\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{\Sigma}\mathbf{V}^\top)^{-1}\big) = \mathrm{tr}\big(\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\big),
\end{aligned}$$

where the inequality is due to $\mathbf{Q}$ is column orthonormal (thus $\mathbf{I}_d-\mathbf{Q}\mathbf{Q}^\top\succeq\mathbf{0}$) and we have $\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\left(\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}\mathbf{Q}\mathbf{\Sigma}\mathbf{V}^\top\right)^\dagger\mathbf{V}\mathbf{\Sigma}\mathbf{Q}^\top\mathbf{R}$ is positive semi-definite. We connect above results to finish the proof. ■

We provide two strategies for selecting matrix $\mathbf{U}\in\mathbb{R}^{d\times k}$ of the SR-k update:

(a) For the randomized strategy, we construct matrix $\mathbf{U}\in\mathbb{R}^{d\times k}$ by sampling each of its entries according to the standard normal distribution independently, i.e., $[\mathbf{U}]_{ij}\overset{\text{i.i.d}}{\sim}\mathcal{N}(0,1)$ for all $i\in[d]$ and $j\in[k]$.

(b) For the greedy strategy, we construct matrix $\mathbf{U}\in\mathbb{R}^{d\times k}$ as $\mathbf{U}=\mathbf{E}_k(\mathbf{G}-\mathbf{A})$, where $\mathbf{E}_k(\cdot)$ follows the definition in Section 2.

The following lemma indicates that the term of $\mathrm{tr}\big(\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}\mathbf{R}\big)$ in inequality (11) can be further lower bounded when we choose $\mathbf{U}\in\mathbb{R}^{d\times k}$ by the above randomized or greedy strategy, which guarantees a sufficient decrease of $\tau_{\mathbf{A}}(\cdot)$ for SR-k updates.

Lemma 3.4 SR-k *updates with randomized and greedy strategies have the following properties:*

(a) If $\mathbf{U} \in \mathbb{R}^{d\times k}$ *is chosen as* $[\mathbf{U}]_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$*, then we have*

$$\mathbb{E}\big[\mathrm{tr}\big(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{R}\big)\big] = \frac{k}{d}\mathrm{tr}(\mathbf{R}) \tag{12}$$

for any matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$*.*

(b) If $\mathbf{U} \in \mathbb{R}^{d\times k}$ *is chosen as* $\mathbf{U} = \mathbf{E}_k(\mathbf{R})$*, then we have*

$$\mathrm{tr}\big(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{R}\big) \geq \frac{k}{d}\mathrm{tr}(\mathbf{R}) \tag{13}$$

for any matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$*.*

Proof We first consider the randomized strategy such that each entry of the matrix $\mathbf{U} \in \mathbb{R}^{d\times k}$ is independently sampled from $\mathcal{N}(0,1)$. We use $\mathcal{V}_{d,k}$ to present the Stiefel manifold which is the set of all $d \times k$ column orthogonal matrices and denote $\mathcal{P}_{k,d-k}$ as the set of all $m \times m$ orthogonal projection matrices of rank k.

According to Theorem 2.2.1 (iii) and Theorem 2.2.2 (iii) of Chikuse (2003), the random matrix $\mathbf{Z} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1/2}$ is uniformly distributed on $\mathcal{V}_{d,k}$ and the random matrix $\mathbf{Z}\mathbf{Z}^\top = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$ is uniformly distributed on $\mathcal{P}_{k,d-k}$. Then, applying Theorem 2.2.2 (i) of Chikuse (2003) on matrix $\mathbf{Z}\mathbf{Z}^\top$ achieves

$$\mathbb{E}\left[\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\right] = \frac{k}{d}\mathbf{I}_d. \tag{14}$$

Consequently, we have

$$\mathbb{E}\left[\mathrm{tr}\big(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\mathbf{R}\big)\right] = \mathrm{tr}\big(\mathbb{E}\left[\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top\right]\mathbf{R}\big) \overset{(14)}{=} \frac{k}{d}\mathrm{tr}(\mathbf{R}).$$

Then we consider the greedy strategy such that $\mathbf{U} = \mathbf{E}_k(\mathbf{R})$. We use $\{r_i\}_{i=1}^k$ to denote k largest diagonal entries of $\mathbf{R}$ with $r_1 \geq r_2 \geq \cdots \geq r_k$, then we have

$$\sum_{i=1}^k r_i \geq \frac{k}{d}\mathrm{tr}(\mathbf{R}). \tag{15}$$

We let $\mathbf{u}_i \in \mathbb{R}^d$ be the i-th column of $\mathbf{U} \in \mathbb{R}^{d\times k}$, then we have

$$\begin{aligned}&\mathrm{tr}\big(\mathbf{U}\big(\mathbf{U}^\top\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{R}\big) = \mathrm{tr}\big(\big(\mathbf{U}^\top\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{R}\mathbf{U}\big) = \mathrm{tr}\big(\mathbf{I}_k\mathbf{U}^\top\mathbf{R}\mathbf{U}\big) = \mathrm{tr}\big(\mathbf{U}^\top\mathbf{R}\mathbf{U}\big)\\ &= \sum_{i=1}^k \mathbf{u}_i^\top\mathbf{R}\mathbf{u}_i = \sum_{i=1}^k r_i \overset{(15)}{\geq} \frac{k}{d}\mathrm{tr}(\mathbf{R}).\end{aligned}$$

■

Remark 3.5 *The proof of result (12) in Lemma 3.4 uses some statistical results on manifold (Chikuse, 2003). For readers who are not familiar with manifold theory, we also provide an elementary proof for equation (12) in Appendix A.*

Now, we formally present the convergence result (9) for SR-k update.

Theorem 3.6 *Let* $\mathbf{G}_+ = \text{SR-}k(\mathbf{G}, \mathbf{A}, \mathbf{U})$ *with* $\mathbf{G}, \mathbf{A} \in \mathbb{R}^{d\times d}$ *such that* $\mathbf{G} \succeq \mathbf{A}$ *and select* $\mathbf{U} \in \mathbb{R}^{d\times k}$*, where* $k \leq d$*, by the randomized strategy* $[\mathbf{U}]_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$ *or the greedy strategy* $\mathbf{U} = \mathbf{E}_k(\mathbf{G} - \mathbf{A})$*. Then we have* $\mathbb{E}\left[\tau_{\mathbf{A}}(\mathbf{G}_+)\right] \leq (1 - k/d)\,\tau_{\mathbf{A}}(\mathbf{G})$.

Proof Applying Lemma 3.3 and Lemma 3.4 by taking $\mathbf{R} = \mathbf{G} - \mathbf{A}$, we have

$$\begin{aligned}
\mathbb{E}[\tau_{\mathbf{A}}(\mathbf{G}_+)] &\overset{(10)}{=} \tau_{\mathbf{A}}(\mathbf{G}) - \mathbb{E}\big[\mathrm{tr}\big((\mathbf{G}-\mathbf{A})\mathbf{U}(\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U}\big)^\dagger \mathbf{U}^\top(\mathbf{G}-\mathbf{A})\big)\big] \\
&\overset{(11)}{\leq} \tau_{\mathbf{A}}(\mathbf{G}) - \mathbb{E}\big[\mathrm{tr}\big(\mathbf{U}\left(\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\big)\big] \\
&\overset{(12),(13)}{\leq} \tau_{\mathbf{A}}(\mathbf{G}) - \frac{k}{d}\mathrm{tr}\big((\mathbf{G}-\mathbf{A})\big) = \left(1 - \frac{k}{d}\right)\tau_{\mathbf{A}}(\mathbf{G}).
\end{aligned}$$

■

Theorem 3.6 reveals the advantage of the SR-k update, i.e., we can achieve faster convergence with respect to $\tau_{\mathbf{A}}(\cdot)$ by increasing the block size k. As a comparison, the results of ordinary randomized or greedy SR1 updates (Lin et al., 2022) are exactly the special case of Theorem 3.6 when $k = 1$.

4. Minimizing Strongly Convex Function

By leveraging the proposed SR-k updates, we introduce a novel block quasi-Newton method, referred to as the SR-k method. The specifics of the SR-k method are outlined in Algorithm 1, where $M > 0$ is the self-concordant parameter that follows the notation in Assumption 2.4.

We shall consider the convergence rate of SR-k method (Algorithm 1) and show its superiority to existing quasi-Newton methods. Our convergence analysis is based on the measure of local gradient norm (Nesterov, 2018; Rodomanov and Nesterov, 2021a; Lin et al., 2022), which is defined as

$$\lambda(\mathbf{x}) \triangleq \sqrt{\nabla f(\mathbf{x})^\top (\nabla^2 f(\mathbf{x}))^{-1} \nabla f(\mathbf{x})}.$$

The analysis starts from the following result for quasi-Newton iterations.

Lemma 4.1 (Rodomanov and Nesterov, 2021a, Lemma 4.3)) *Suppose that the twice differentiable objective* $f : \mathbb{R}^d \to \mathbb{R}$ *is strongly self-concordant with constant* $M \geq 0$ *and the positive definite matrix* $\mathbf{G}_t \in \mathbb{R}^{d\times d}$ *satisfies* $\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{G}_t \preceq \eta_t \nabla^2 f(\mathbf{x}_t)$ *for some* $\eta_t \geq 1$ *and* $M\lambda(\mathbf{x}_t) \leq 2$*. Then the update formula*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t) \tag{16}$$

holds that

$$r_t \leq \lambda(\mathbf{x}_t) \quad \textit{and} \quad \lambda(\mathbf{x}_{t+1}) \leq \left(1 - \frac{1}{\eta_t}\right)\lambda(\mathbf{x}_t) + \frac{M(\lambda(\mathbf{x}_t))^2}{2} + \frac{M^2(\lambda(\mathbf{x}_t))^3}{4\eta_t}, \tag{17}$$

where $r_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{x}_t}$.

Algorithm 1 Symmetric Rank-k Method

Input: $\mathbf{x}_0, \mathbf{G}_0$, and k
for $t = 0, 1 \ldots$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t)$
  $r_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{x}_t}$
  $\tilde{\mathbf{G}}_t = (1 + Mr_t)\mathbf{G}_t$
  Construct $\mathbf{U}_t \in \mathbb{R}^{d\times k}$ by
    (a) randomized strategy: $[\mathbf{U}_t]_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$
    (b) greedy strategy: $\mathbf{U}_t = \mathbf{E}_k(\tilde{\mathbf{G}}_t - \nabla^2 f(\mathbf{x}_{t+1}))$
  $\mathbf{G}_{t+1} = \text{SR-}k(\tilde{\mathbf{G}}_t, \nabla^2 f(\mathbf{x}_{t+1}), \mathbf{U}_t)$
end for

Note that the value of η_t in equation (17) is crucial to describe the local convergence rates of different types of quasi-Newton methods. Applying Lemma 4.1 by setting $\eta_t = 3\eta_0/2$ and combining with Lemma 3.2, we can establish the linear convergence rate of SR-k methods as follows (see Appendix C for the detailed proof).

Theorem 4.2 *Under Assumptions 2.1, 2.2, and 2.4, if we run SR-k method (Algorithm 1) with* $\mathbf{x}_0 \in \mathbb{R}^d$, $\mathbf{G}_0 \in \mathbb{R}^{d\times d}$ *such that* $\nabla^2 f(\mathbf{x}_0) \preceq \mathbf{G}_0 \preceq \eta_0 \nabla^2 f(\mathbf{x}_0)$ *and* $M\lambda(\mathbf{x}_0) \leq \ln(3/2)/(4\eta_0)$ *for some* $\eta_0 \geq 1$. *Then it holds that*

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{G}_t \preceq \frac{3\eta_0}{2}\nabla^2 f(\mathbf{x}_t) \qquad \text{and} \qquad \lambda(\mathbf{x}_t) \leq \left(1 - \frac{1}{2\eta_0}\right)^t \lambda(\mathbf{x}_0). \tag{18}$$

Furthermore, we can obtain the superlinear rate for iteration (16) if there exists some sequence $\{\eta_t\}$ such that $\eta_t \geq 1$ for all $t \geq 1$ and $\lim_{t\to\infty} \eta_t = 1$. For example, the randomized and greedy SR1 methods (Lin et al., 2022) lead to some $\{\eta_t\}$ such that $\eta_t \geq 1$ and $\mathbb{E}[\eta_t - 1] \leq \mathcal{O}((1 - 1/d)^t)$. As the results shown in Theorem 3.6, the proposed SR-k updates have superiority in matrix approximation. So we desire to construct some $\{\eta_t\}$ for SR-k methods with the tighter bound $\mathbb{E}[\eta_t - 1] \leq \mathcal{O}((1 - k/d)^t)$.

Based on the above intuition, we derive the faster local superlinear convergence rate for SR-k methods in the following theorem (see Appendix D for the detailed proof), which is explicitly sharper than the convergence rates of existing randomized and greedy quasi-Newton methods (Lin et al., 2022; Rodomanov and Nesterov, 2021a).

Theorem 4.3 *Under Assumptions 2.1, 2.2, and 2.4, we run SR-k method (Algorithm 1) with* $\mathbf{x}_0 \in \mathbb{R}^d$ *and* $\mathbf{G}_0 \in \mathbb{R}^{d\times d}$ *such that*

$$M\lambda(\mathbf{x}_0) \leq \frac{\ln 2}{2} \cdot \frac{d-k}{\eta_0 d^2 \varkappa} \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}_0) \preceq \mathbf{G}_0 \preceq \eta_0 \nabla^2 f(\mathbf{x}_0) \tag{19}$$

for some $k < d$ *and* $\eta_0 \geq 1$. *Then it holds that*

$$\mathbb{E}\left[\frac{\lambda(\mathbf{x}_{t+1})}{\lambda(\mathbf{x}_t)}\right] \leq 2d\varkappa\eta_0 \left(1 - \frac{k}{d}\right)^t, \tag{20}$$

which naturally indicates the following two-stage convergence behaviors:

(a) For SR-k *method with randomized strategy, we have*

$$\lambda(\mathbf{x}_{t_0+t}) \leq \left(1-\frac{k}{d+k}\right)^{t(t-1)/2} \cdot \left(\frac{1}{2}\right)^{t} \cdot \left(1-\frac{1}{2\eta_0}\right)^{t_0} \lambda(\mathbf{x}_0),$$

with probability at least $1-\delta$ *for some* $\delta \in (0,1)$*, where* $t_0 = \mathcal{O}(d\ln(\varkappa d\eta_0/\delta)/k)$.

(b) For SR-k *method with greedy strategy, we have*

$$\lambda(\mathbf{x}_{t_0+t}) \leq \left(1-\frac{k}{d}\right)^{t(t-1)/2} \cdot \left(\frac{1}{2}\right)^{t} \cdot \left(1-\frac{1}{2\eta_0}\right)^{t_0} \lambda(\mathbf{x}_0),$$

where $t_0 = \mathcal{O}\left(d\ln(\eta_0 d\varkappa)/k\right)$.

Remark 4.4 *The condition* $\nabla^2 f(\mathbf{x}_0) \preceq \mathbf{G}_0 \preceq \eta_0 \nabla^2 f(\mathbf{x}_0)$ *in Theorem 4.2 and 4.3 can be satisfied by simply setting* $\mathbf{G}_0 = L\mathbf{I}_d$*, then we can efficiently implement the update on* $\mathbf{G}_t$ *by Woodbury identity (Woodbury, 1950). In a follow-up work, Liu et al. (2023) analyzed block quasi-Newton methods for solving nonlinear equations. However, their convergence guarantees require the Jacobian estimator at the initial point to be sufficiently accurate, which leads to potentially expensive costs in the step of initialization.*

Remark 4.5 *Theorem 4.3 also indicates the theoretical difference between the randomized and greedy strategies. The greedy strategy enjoys a deterministic convergence guarantee with respect to the local gradient norm* $\lambda(\mathbf{x}_t)$*, while the randomized strategy achieves the corresponding guarantee with high probability and introduces an additional logarithmic dependence on the failure probability* δ*. However, the greedy strategy requires extra computation to identify the top-k diagonal entries of the matrix* $\tilde{\mathbf{G}}_t - \nabla^2 f(\mathbf{x}_{t+1})$ *at each iteration, e.g., line 6(b) of Algorithm 1. In contrast, the randomized strategy is easier to implement since it only needs construct* $\mathbf{U}_t$ *by following the normal distribution, which can make it preferable in large-scale applications.*

We can also set $k = d$ for SR-k methods, which leads to $\eta_t = 1$ almost surely for all $t \geq 1$ and achieves the quadratic convergence rate like standard Newton methods.

Corollary 4.6 *Under Assumptions 2.1, 2.2, and 2.4, we run the SR-k method (Algorithm 1) with* $k = d$ *and* $\mathbf{x}_0 \in \mathbb{R}^d$ *such that* $M\lambda(\mathbf{x}_0) \leq 2$*, then* $\lambda(\mathbf{x}_{t+1}) \leq M(\lambda(\mathbf{x}_t))^2$ *is held almost surely for all* $t \geq 1$.

Proof The update rule $\mathbf{G}_t = \text{SR-}k(\tilde{\mathbf{G}}_{t-1}, \nabla^2 f(\mathbf{x}_t), \mathbf{U}_{t-1})$ implies the equality holds with $\mathbf{U}_{t-1}^\top \mathbf{G}_t \mathbf{U}_{t-1} = \mathbf{U}_{t-1}^\top \nabla^2 f(\mathbf{x}_t)\mathbf{U}_{t-1}$. Our choice of $\mathbf{U}_t \in \mathbb{R}^{d\times d}$ guarantees that it is nonsingular almost surely, so we have $\mathbf{G}_t = \nabla^2 f(\mathbf{x}_t)$ for all $t \geq 1$. Lemma 4.1 with $\eta_t = 1$ implies that $\lambda(\mathbf{x}_{t+1}) \leq M(\lambda(\mathbf{x}_t))^2/2 + M^2(\lambda(\mathbf{x}_t))^3/2 \leq M(\lambda(\mathbf{x}_t))^2$ holds for all $t \geq 1$ almost surely, which finishes the proof. ■

5. Improved Results for Block BFGS and DFP Methods

Following our our investigation on SR-k methods, we can also achieve the non-asymptotic superlinear convergence rates of randomized block BFGS and randomized block DFP methods (Gower et al., 2016; Gower and Richtárik, 2017). The block BFGS (Schnabel, 1983; Gower and Richtárik, 2017; Gower et al., 2016) and DFP (Gower and Richtárik, 2017) updates are defined as follows.

Definition 5.1 *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{G} \in \mathbb{R}^{d\times d}$ be two positive-definite symmetric matrices with $\mathbf{A} \preceq \mathbf{G}$. For any full rank matrix $\mathbf{U} \in \mathbb{R}^{d\times k}$ with $k \leq d$, we define*

$$\text{BlockBFGS}(\mathbf{G}, \mathbf{A}, \mathbf{U}) \triangleq \mathbf{G} - \mathbf{G}\mathbf{U}\big(\mathbf{U}^\top\mathbf{G}\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{G} + \mathbf{A}\mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{A}, \tag{21}$$

and

$$\begin{aligned}&\text{BlockDFP}(\mathbf{G}, \mathbf{A}, \mathbf{U})\\ &\triangleq \mathbf{A}\mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{A} + \big(\mathbf{I}_d - \mathbf{A}\mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big)\mathbf{G}\big(\mathbf{I}_d - \mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{A}\big).\end{aligned} \tag{22}$$

In previous work, Gower et al. (2016) and Kovalev et al. (2020) proposed a randomized block BFGS method by constructing the Hessian estimator with formula (21) and showed it has an asymptotic local superlinear convergence rate. On the other hand, Gower and Richtárik (2017) studied the randomized block DFP update (22) for matrix approximation, but they did not consider solving optimization problems.

To achieve the explicit superlinear convergence rate of block BFGS and block DFP methods, we provide some properties of block BFGS and block DFP updates which are similar to our observations on SR-k update. We first show that block BFGS and DFP updates also have the non-increasing deviation from the target matrix.

Lemma 5.2 *Given any positive-definite matrices $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{G} \in \mathbb{R}^{d\times d}$ which satisfy $\mathbf{A} \preceq \mathbf{G} \preceq \eta\mathbf{A}$ for some $\eta \geq 1$, both of the updates $\mathbf{G}_+ = \text{BlockBFGS}(\mathbf{G}, \mathbf{A}, \mathbf{U})$ and $\mathbf{G}_+ = \text{BlockDFP}(\mathbf{G}, \mathbf{A}, \mathbf{U})$ for full rank matrix $\mathbf{U} \in \mathbb{R}^{d\times k}$ hold that $\mathbf{A} \preceq \mathbf{G}_+ \preceq \eta\mathbf{A}$.*

Proof We first consider the block BFGS update $\mathbf{G}_+ = \text{BlockBFGS}(\mathbf{G}, \mathbf{A}, \mathbf{U})$. According to the Woodbury identity (Woodbury, 1950), we have

$$\mathbf{G}_+^{-1} = \mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top + \big(\mathbf{I}_d - \mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{A}\big)\mathbf{G}^{-1}\big(\mathbf{I}_d - \mathbf{A}\mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big). \tag{23}$$

The condition $\mathbf{A} \preceq \mathbf{G} \preceq \eta\mathbf{A}$ means $\eta^{-1}\mathbf{A}^{-1} \preceq \mathbf{G}^{-1} \preceq \mathbf{A}^{-1}$, which implies

$$\mathbf{G}_+^{-1} \overset{(23)}{\preceq} \mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top + \big(\mathbf{I}_d - \mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{A}\big)\mathbf{A}^{-1}\big(\mathbf{I}_d - \mathbf{A}\mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big) = \mathbf{A}^{-1}$$

and

$$\begin{aligned}\mathbf{G}_+^{-1} &\overset{(23)}{\succeq} \mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top + \eta^{-1}\big(\mathbf{I}_d - \mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{A}\big)\mathbf{A}^{-1}\big(\mathbf{I}_d - \mathbf{A}\mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big)\\ &= \eta^{-1}\mathbf{A}^{-1} + (1-\eta^{-1})\mathbf{U}\big(\mathbf{U}^\top\mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top \succeq \eta^{-1}\mathbf{A}^{-1}.\end{aligned}$$

Thus we have $\eta^{-1}\mathbf{A}^{-1} \preceq \mathbf{G}_+^{-1} \preceq \mathbf{A}^{-1}$, which finishes the proof for the block BFGS update. We then consider the DFP update $\mathbf{G}_+ = \text{BlockDFP}(\mathbf{G}, \mathbf{A}, \mathbf{U})$. The condition $\mathbf{A} \preceq \mathbf{G} \preceq \eta\mathbf{A}$ means

$$\mathbf{G}_+ \overset{(22)}{\succeq} \mathbf{AU}\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\mathbf{A} + \big(\mathbf{I}_d - \mathbf{AU}\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\big)\mathbf{A}\big(\mathbf{I}_d - \mathbf{U}\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\mathbf{A}\big) = \mathbf{A}$$

and

$$\begin{aligned}\mathbf{G}_+ &\overset{(22)}{\preceq} \mathbf{AU}\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\mathbf{A} + \eta\big(\mathbf{I}_d - \mathbf{AU}\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\big)\mathbf{A}\big(\mathbf{I}_d - \mathbf{U}\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\mathbf{A}\big)\\ &= \eta\mathbf{A} + (1-\eta)\mathbf{AU}\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\mathbf{A} \preceq \eta\mathbf{A},\end{aligned}$$

where the last inequality is due to the facts $\eta \geq 1$ and $\mathbf{AU}\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\mathbf{A} \succeq \mathbf{0}$, which finishes the proof for the block DFP update. ■

Then we introduce the following quantity (Rodomanov and Nesterov, 2021a)

$$\sigma_{\mathbf{A}}(\mathbf{G}) \triangleq \text{tr}\big(\mathbf{A}^{-1}(\mathbf{G} - \mathbf{A})\big) \tag{24}$$

to measure the difference between the target matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and the current estimator $\mathbf{G} \in \mathbb{R}^{d\times d}$. In the following theorem, we show that randomized block BFGS and DFP updates converge to the target matrix with a faster rate than the ordinary randomized BFGS and DFP updates (Rodomanov and Nesterov, 2021a; Lin et al., 2022).

Theorem 5.3 *Consider the randomized block update*

$$\mathbf{G}_+ = \text{BlockBFGS}(\mathbf{G}, \mathbf{A}, \mathbf{U}) \qquad \textit{or} \qquad \mathbf{G}_+ = \text{BlockDFP}(\mathbf{G}, \mathbf{A}, \mathbf{U}), \tag{25}$$

where $\mathbf{G}, \mathbf{A} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{G} \succeq \mathbf{A}$. *If* $\mathbf{U} \in \mathbb{R}^{d\times k}$ *is chosen as* $[\mathbf{U}]_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$ *and it satisfies that* $\mu\mathbf{I}_d \preceq \mathbf{A} \preceq L\mathbf{I}_d$, *then we have*

$$\mathbb{E}\left[\sigma_{\mathbf{A}}(\mathbf{G}_+)\right] \leq \left(1 - \frac{k}{d\varkappa}\right)\sigma_{\mathbf{A}}(\mathbf{G}). \tag{26}$$

Proof The condition $\mu\mathbf{I}_d \preceq \mathbf{A} \preceq L\mathbf{I}_d$ means we have

$$\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1} \succeq \frac{1}{L}\cdot\big(\mathbf{U}^\top\mathbf{U}\big)^{-1} \qquad \text{and} \qquad \frac{1}{\mu}\cdot\mathbf{I}_d - \mathbf{A}^{-1} \succeq \mathbf{0}, \tag{27}$$

which leads to the following inequality

$$\begin{aligned}&\text{tr}\big(\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\mathbf{GU}\big) - \text{tr}\big(\mathbf{AU}\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top\big)\\ &= \text{tr}\big(\big(\mathbf{U}^\top\mathbf{AU}\big)^{-1}\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U}\big) \overset{(27)}{\geq} \frac{1}{L}\text{tr}\big(\big(\mathbf{U}^\top\mathbf{U}\big)^{-1}\mathbf{U}^\top(\mathbf{G}-\mathbf{A})\mathbf{U}\big)\\ &= \frac{1}{L}\text{tr}\big((\mathbf{G}-\mathbf{A})^{1/2}\mathbf{U}\big(\mathbf{U}^\top\mathbf{U}\big)^{-1}\mathbf{U}^\top(\mathbf{G}-\mathbf{A})^{1/2}\big)\\ &\overset{(27)}{\geq} \frac{\mu}{L}\text{tr}\big(\mathbf{A}^{-1}(\mathbf{G}-\mathbf{A})^{1/2}\mathbf{U}\big(\mathbf{U}^\top\mathbf{U}\big)^{-1}\mathbf{U}^\top(\mathbf{G}-\mathbf{A})^{1/2}\big).\end{aligned} \tag{28}$$

For the block BFGS update, we set $\mathbf{P} = \mathbf{A}^{1/2}\mathbf{U}$. Then it holds

$$\begin{aligned}
&\mathbf{U}^\top \mathbf{G}\big(\mathbf{A}^{-1} - \mathbf{U}\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big)\mathbf{G}\mathbf{U} \\
&= \mathbf{U}^\top \mathbf{G}\mathbf{A}^{-1/2}\big(\mathbf{I}_d - \mathbf{A}^{1/2}\mathbf{U}\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top \mathbf{A}^{1/2}\big)\mathbf{A}^{-1/2}\mathbf{G}\mathbf{U} \\
&= \mathbf{U}^\top \mathbf{G}\mathbf{A}^{-1/2}\big(\mathbf{I}_d - \mathbf{P}\big(\mathbf{P}^\top \mathbf{P}\big)^{-1}\mathbf{P}^\top\big)\mathbf{A}^{-1/2}\mathbf{G}\mathbf{U} \succeq \mathbf{0},
\end{aligned} \tag{29}$$

which implies

$$\begin{aligned}
&\mathrm{tr}\big(\mathbf{U}\mathbf{G}\big(\mathbf{U}^\top \mathbf{G}\mathbf{U}\big)^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{A}^{-1}\big) - \mathrm{tr}\big(\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{U}\big) \\
&= \mathrm{tr}\big(\big(\mathbf{U}^\top \mathbf{G}\mathbf{U}\big)^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{A}^{-1}\mathbf{G}\mathbf{U}\big) - \mathrm{tr}\big(\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{U}\big) \\
&= \mathrm{tr}\big(\big(\mathbf{U}^\top \mathbf{G}\mathbf{U}\big)^{-1}\big(\mathbf{U}^\top \mathbf{G}\mathbf{A}^{-1}\mathbf{G}\mathbf{U} - \mathbf{U}^\top \mathbf{G}\mathbf{U}\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{U}\big)\big) \\
&= \mathrm{tr}\big(\big(\mathbf{U}^\top \mathbf{G}\mathbf{U}\big)^{-1/2}\mathbf{U}^\top \mathbf{G}\big(\mathbf{A}^{-1} - \mathbf{U}\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big)\mathbf{G}\mathbf{U}\big(\mathbf{U}^\top \mathbf{G}\mathbf{U}\big)^{-1/2}\big) \overset{(29)}{\geq} 0.
\end{aligned} \tag{30}$$

The equation (21) implies

$$\begin{aligned}
&\mathrm{tr}\big((\mathbf{G}_+ - \mathbf{A})\mathbf{A}^{-1}\big) \\
&\overset{(21)}{=} \mathrm{tr}\big((\mathbf{G} - \mathbf{A})\mathbf{A}^{-1}\big) - \mathrm{tr}\big(\mathbf{G}\mathbf{U}\big(\mathbf{U}^\top \mathbf{G}\mathbf{U}\big)^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{A}^{-1}\big) + \mathrm{tr}\big(\mathbf{A}\mathbf{U}\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big) \\
&\overset{(30)}{\leq} \mathrm{tr}\big((\mathbf{G} - \mathbf{A})\mathbf{A}^{-1}\big) - \big(\mathrm{tr}\big(\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{U}\big) - \mathrm{tr}\big(\mathbf{A}\mathbf{U}\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big)\big).
\end{aligned}$$

Taking expectation on both sides of the above result, we obtain

$$\begin{aligned}
\mathbb{E}[\sigma_{\mathbf{A}}(\mathbf{G}_+)] &\overset{(24)}{\leq} \sigma_{\mathbf{A}}(\mathbf{G}) - \mathbb{E}\big[\mathrm{tr}\big(\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{U} - \mathbf{A}\mathbf{U}\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big)\big] \\
&\overset{(28)}{\leq} \sigma_{\mathbf{A}}(\mathbf{G}) - \mathbb{E}\left[\frac{\mu}{L}\mathrm{tr}\big(\mathbf{A}^{-1}(\mathbf{G} - \mathbf{A})^{1/2}\mathbf{U}\big(\mathbf{U}^\top \mathbf{U}\big)^{-1}\mathbf{U}^\top(\mathbf{G} - \mathbf{A})^{1/2}\big)\right] \\
&= \sigma_{\mathbf{A}}(\mathbf{G}) - \frac{\mu}{L}\mathbb{E}\big[\mathrm{tr}\big((\mathbf{G} - \mathbf{A})^{1/2}\mathbf{A}^{-1}(\mathbf{G} - \mathbf{A})^{1/2}\mathbf{U}(\mathbf{U}^\top \mathbf{U})^{-1}\mathbf{U}^\top\big)\big] \\
&= \sigma_{\mathbf{A}}(\mathbf{G}) - \frac{k}{d\varkappa}\mathrm{tr}\big((\mathbf{G} - \mathbf{A})^{1/2}\mathbf{A}^{-1}(\mathbf{G} - \mathbf{A})^{1/2}\big) = \left(1 - \frac{k}{d\varkappa}\right)\sigma_{\mathbf{A}}(\mathbf{G}),
\end{aligned} \tag{31}$$

where the last line is by applying Lemma 3.4(a) with $\mathbf{R} = (\mathbf{G} - \mathbf{A})^{1/2}\mathbf{A}^{-1}(\mathbf{G} - \mathbf{A})^{1/2}$.

For the block DFP update, the equation (22) indicates

$$\begin{aligned}
&\mathrm{tr}\big((\mathbf{G}_+ - \mathbf{A})\mathbf{A}^{-1}\big) \\
&\overset{(22)}{=} \mathrm{tr}\big(\mathbf{A}\mathbf{U}(\mathbf{U}^\top \mathbf{A}\mathbf{U})^{-1}\mathbf{U}^\top\big) + \mathrm{tr}\big((\mathbf{G} - \mathbf{A})\mathbf{A}^{-1}\big) + \mathrm{tr}\big(\mathbf{A}\mathbf{U}(\mathbf{U}^\top \mathbf{A}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{U}(\mathbf{U}^\top \mathbf{A}\mathbf{U})^{-1}\mathbf{U}^\top\big) \\
&\quad - \mathrm{tr}\big(\mathbf{A}\mathbf{U}(\mathbf{U}^\top \mathbf{A}\mathbf{U})^{-1}\mathbf{U}^\top \mathbf{G}\mathbf{A}^{-1}\big) - \mathrm{tr}\big(\mathbf{G}\mathbf{U}(\mathbf{U}^\top \mathbf{A}\mathbf{U})^{-1}\mathbf{U}^\top\big) \\
&= \mathrm{tr}\big((\mathbf{G} - \mathbf{A})\mathbf{A}^{-1}\big) - \big(\mathrm{tr}\big((\mathbf{U}^\top \mathbf{A}\mathbf{U})^{-1}(\mathbf{U}^\top \mathbf{G}\mathbf{U})\big) - \mathrm{tr}\big(\mathbf{A}\mathbf{U}\big(\mathbf{U}^\top \mathbf{A}\mathbf{U}\big)^{-1}\mathbf{U}^\top\big)\big).
\end{aligned}$$

Taking the expectation on both sides of the above equation, then we can follow the analysis of the block BFGS update from the step of (31) to prove the desired result for the block DFP update. Hence, we finish the proof. ■

Algorithm 2 Randomized Block BFGS/DFP

Input: $\mathbf{x}_0$, $\mathbf{G}_0$, and k
for $t = 0, 1 \ldots$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t)$
  $r_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{x}_t}$
  $\tilde{\mathbf{G}}_t = (1 + Mr_t)\mathbf{G}_t$
  Construct $\mathbf{U}_t$ by $[\mathbf{U}_t]_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$
  Update $\mathbf{G}_t$ by
    (a) $\mathbf{G}_{t+1} = \text{BlockBFGS}(\tilde{\mathbf{G}}_t, \nabla^2 f(\mathbf{x}_{t+1}), \mathbf{U}_t)$
    (b) $\mathbf{G}_{t+1} = \text{BlockDFP}(\tilde{\mathbf{G}}_t, \nabla^2 f(\mathbf{x}_{t+1}), \mathbf{U}_t)$
end for

Remark 5.4 *Notice that Gower and Richtárik (2017) also established convergence rates for randomized block BFGS and DFP updates, but their results do not reveal the relationship between the convergence rates and the block size k.*

We propose randomized block BFGS and DFP methods in Algorithm 2. Using the results of Theorem 5.3, we establish the explicit superlinear convergence rate for our methods as follows (see Appendix E for the detailed proof).

Theorem 5.5 *Under Assumption 2.1, 2.2, and 2.4, we run randomized block BFGS/DFP method (Algorithm 2) with $\mathbf{x}_0 \in \mathbb{R}^d$ and $\mathbf{G}_0 \in \mathbb{R}^{d\times d}$ such that*

$$M\lambda(\mathbf{x}_0) \leq \frac{\ln 2}{4} \cdot \frac{d\varkappa - k}{\eta_0 d^2 \varkappa} \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}_0) \preceq \mathbf{G}_0 \preceq \eta_0 \nabla^2 f(\mathbf{x}_0) \tag{32}$$

for some $k < d$ and $\eta_0 \geq 1$. Then we have

$$\mathbb{E}\left[\frac{\lambda(\mathbf{x}_{t+1})}{\lambda(\mathbf{x}_t)}\right] \leq 2d\eta_0 \left(1 - \frac{k}{d\varkappa}\right)^t.$$

For randomized block BFGS method, we can introduce the scaled directions to eliminate the condition number in the results of Theorem 5.3 and 5.5. We first improve the result for matrix approximation as follows.

Theorem 5.6 *Consider the block BFGS update*

$$\mathbf{G}_+ = \text{BlockBFGS}(\mathbf{G}, \mathbf{A}, \mathbf{L}^\top \mathbf{U}) \tag{33}$$

where $\mathbf{G} \succeq \mathbf{A} \in \mathbb{R}^{d\times d}$. If $\mu\mathbf{I} \preceq \mathbf{A} \preceq L\mathbf{I}$, $\mathbf{L} \in \mathbb{R}^{d\times d}$ satisfies that $\mathbf{L}^\top \mathbf{L} = \mathbf{G}^{-1}$, and $\mathbf{U} \in \mathbb{R}^{d\times k}$ is chosen as $[\mathbf{U}]_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$. Then we have

$$\mathbb{E}\left[\sigma_{\mathbf{A}}(\mathbf{G}_+)\right] \leq \left(1 - \frac{k}{d}\right)\sigma_{\mathbf{A}}(\mathbf{G}). \tag{34}$$

Proof According to the block-BFGS update (21), we have

$$\begin{aligned}
&\operatorname{tr}\big((\mathbf{G}_+ - \mathbf{A})\mathbf{A}^{-1}\big) \\
&= \operatorname{tr}\big((\mathbf{G} - \mathbf{A})\mathbf{A}^{-1}\big) - \operatorname{tr}\big(\mathbf{L}^{-1}\mathbf{U}\big(\mathbf{U}^\top\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{L}^{-\top}\mathbf{A}^{-1}\big) + \operatorname{tr}\big(\mathbf{A}\mathbf{L}^\top\mathbf{U}\big(\mathbf{U}\mathbf{L}^\top\mathbf{A}\mathbf{L}^\top\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{L}\big) \\
&= \operatorname{tr}\big((\mathbf{G} - \mathbf{A})\mathbf{A}^{-1}\big) - \operatorname{tr}\big(\mathbf{L}^{-1}\mathbf{U}\big(\mathbf{U}^\top\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{L}^{-\top}\mathbf{A}^{-1}\big) + k.
\end{aligned}$$

The second term on the right-hand side holds that

$$\mathbb{E}\big[\operatorname{tr}\big(\mathbf{L}^{-1}\mathbf{U}\big(\mathbf{U}^\top\mathbf{U}\big)^{-1}\mathbf{U}^\top\mathbf{L}^{-\top}\mathbf{A}^{-1}\big)\big] = \operatorname{tr}\big(\mathbb{E}\big[\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\big]\mathbf{L}^{-\top}\mathbf{A}^{-1}\mathbf{L}^{-1}\big) \overset{(14)}{=} \frac{k}{d}\operatorname{tr}\big(\mathbf{A}^{-1}\mathbf{G}\big).$$

Thus, we combine the above results and get

$$\mathbb{E}\left[\sigma_{\mathbf{A}}(\mathbf{G}_+)\right] = \sigma_{\mathbf{A}}(\mathbf{G}) - \frac{k}{d}\operatorname{tr}\big(\mathbf{A}^{-1}\mathbf{G}\big) + k = \sigma_{\mathbf{A}}(\mathbf{G}) - \frac{k}{d}\operatorname{tr}\big(\mathbf{A}^{-1}(\mathbf{G} - \mathbf{A})\big) = \left(1 - \frac{k}{d}\right)\sigma_{\mathbf{A}}(\mathbf{G}).$$

■

For given matrices $\mathbf{G}, \mathbf{A}, \mathbf{L}, \mathbf{U} \in \mathbb{R}^{d\times d}$ satisfying the condition of Theorem 5.6, we can access $\mathbf{L}_+ \in \mathbb{R}^{d\times d}$ such that $\mathbf{G}_+^{-1} = \mathbf{L}_+^\top\mathbf{L}_+$ by a closed form expression (Gower and Richtárik, 2017, Section 9).

Proposition 5.7 *Under the setting of Theorem 5.6, the matrix*

$$\begin{aligned}
\mathbf{L}_+ &= \operatorname{UpdateL}(\mathbf{L}, \mathbf{A}, \mathbf{U}) \\
&\triangleq \mathbf{L} + \big(\mathbf{U}\big(\mathbf{U}^\top\mathbf{U}\big)^{-1/2} - \mathbf{L}\mathbf{A}\mathbf{L}^\top\mathbf{U}\big(\mathbf{U}^\top\mathbf{L}\mathbf{A}\mathbf{L}^\top\mathbf{U}\big)^{-1/2}\big)\big(\mathbf{U}^\top\mathbf{L}\mathbf{A}\mathbf{L}^\top\mathbf{U}\big)^{-1/2}\mathbf{U}^\top\mathbf{L},
\end{aligned}$$

holds that $\mathbf{G}_+^{-1} = \mathbf{L}_+^\top\mathbf{L}_+$.

We can achieve Proposition 5.7 by applying equation (9.11) of Gower and Richtárik (2017) with $\tilde{\mathbf{S}}_k = \mathbf{U}$ and $\mathbf{L}_k = \mathbf{L}^\top$, where $\tilde{\mathbf{S}}_k$ and $\mathbf{L}_k$ follow notations of Gower and Richtárik (2017). This result means we can access the scaling matrix $\mathbf{L}_+$ efficiently during iterations.

Based on the BFGS update with scaling directions shown in Theorem 5.6 and the efficient update rule shown in Proposition 5.7, we propose the faster randomized block BFGS method in Algorithm 3, which has the following local convergence rate matching SR-k methods (see Appendix F for the detailed proof).

Theorem 5.8 *Under Assumption 2.1, 2.2, and 2.4, we run faster randomized block BFGS method (Algorithm 3) with* $\mathbf{x}_0 \in \mathbb{R}^d$ *and* $\mathbf{L}_0, \mathbf{G}_0 \in \mathbb{R}^{d\times d}$ *such that*

$$M\lambda_0 \leq \frac{\ln 2}{4}\cdot\frac{d-k}{\eta_0 d^2}, \quad \nabla^2 f(\mathbf{x}_0) \preceq \mathbf{G}_0 \preceq \eta_0\nabla^2 f(\mathbf{x}_0), \quad \textit{and} \quad \mathbf{G}_0^{-1} = \mathbf{L}_0^\top\mathbf{L}_0 \tag{35}$$

for some $k < d$ *and* $\eta_0 \geq 1$. *Then we have*

$$\mathbb{E}\left[\frac{\lambda(\mathbf{x}_{t+1})}{\lambda(\mathbf{x}_t)}\right] \leq 2d\eta_0\left(1 - \frac{k}{d}\right)^t.$$

Algorithm 3 Faster Randomized Block BFGS

Input: $\mathbf{x}_0$, $\mathbf{G}_0$, $\mathbf{L}_0$, and k
for $t = 0, 1 \dots$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \mathbf{G}_t^{-1}\nabla f(\mathbf{x}_t)$
  $r_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{x}_t}$
  $\tilde{\mathbf{G}}_t = (1 + Mr_t)\mathbf{G}_t$
  $\tilde{\mathbf{L}}_t = \mathbf{L}_t/\sqrt{1 + Mr_t}$
  Construct $\mathbf{U}_t$ by $[\mathbf{U}_t]_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$
  $\mathbf{G}_{t+1} = \text{BlockBFGS}(\tilde{\mathbf{G}}_t, \nabla^2 f(\mathbf{x}_{t+1}), \tilde{\mathbf{L}}_t^\top \mathbf{U}_t)$
  $\mathbf{L}_{t+1} = \text{UpdateL}(\tilde{\mathbf{L}}_t, \nabla^2 f(\mathbf{x}_{t+1}), \mathbf{U}_t)$
end for

Remark 5.9 *The condition on $\mathbf{G}_0 \in \mathbb{R}^{d\times d}$ in Theorem 5.5 and 5.8 can be satisfied by simply setting $\mathbf{G}_0 = L\mathbf{I}_d$. In addition, we can set $\mathbf{L}_0 = L^{-1/2}\mathbf{I}_d$ to satisfy the condition on $\mathbf{L}_0 \in \mathbb{R}^{d\times d}$ in Theorem 5.8, then we can efficiently implement the update on $\mathbf{L}_t$ by Proposition 5.7.*

Remark 5.10 *For $k = 1$, the convergence rates provided by Theorem 5.5 and 5.8 match the results of the ordinary randomized BFGS/DFP methods (Rodomanov and Nesterov, 2021a; Lin et al., 2022) and fast randomized BFGS methods (Lin et al., 2022) respectively. For $k = d$, the updates* (25) *and* (33) *holds $\mathbf{G}_+ = \mathbf{A}$ almost surely. This leads to the local quadratic convergence rate of our (faster) randomized block BFGS/DFP methods, which is similar to the behaviors of* SR-k *methods shown in Corollary 4.6.*

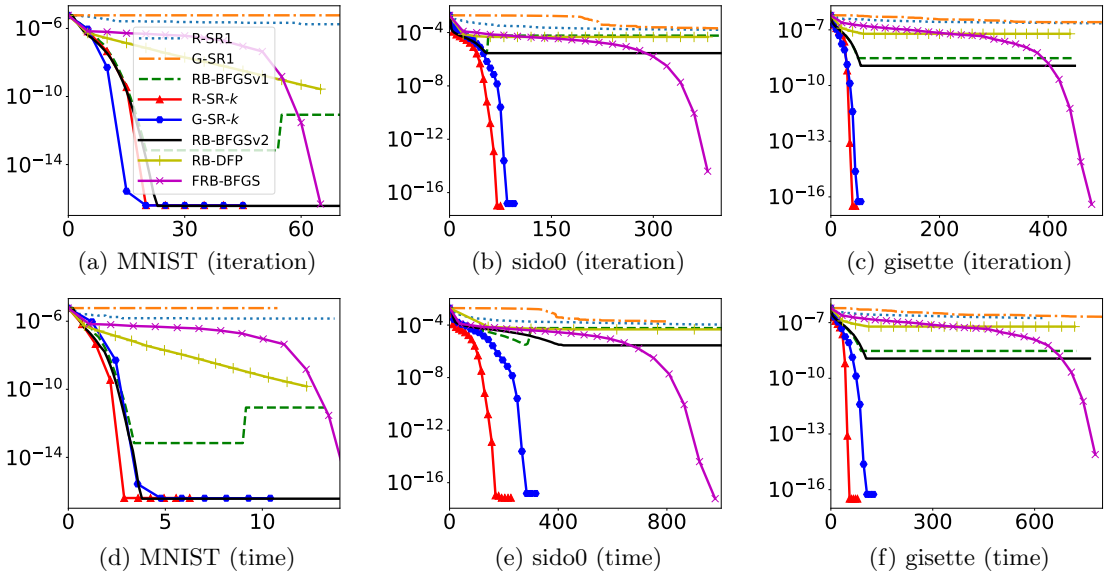
6. Numerical Experiments

We conduct the experiments on the model of regularized logistic regression, which can be formulated as

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp\left(-b_i \mathbf{a}_i^\top \mathbf{x}\right)\right) + \frac{\gamma}{2}\|\mathbf{x}\|^2, \tag{36}$$

where $\mathbf{a}_i \in \mathbb{R}^d$ and $b_i \in \{-1, +1\}$ are the feature and the corresponding label of the i-th sample respectively, and $\gamma > 0$ is the regularization hyperparameter.

We refer to SR-k methods with randomized/greedy strategies (Algorithm 1) as R-SR-k/G-SR-k. The SR1 methods with randomized/greedy strategies are referred to as R-SR1/G-SR1 (Lin et al., 2022, Algorithm 4). We refer to the existing randomized block BFGS method (Kovalev et al., 2020, Algorithm 1) as RB-BFGSv1 and our randomized block BFGS/DFP methods (Algorithm 2) as RB-BFGSv2/RB-DFP. We denote our faster block BFGS method (Algorithm 3) as FRB-BFGS. We compare the proposed R-SR-k, G-SR-k, RB-BFGSv2, RB-DFP, and FRB-BFGS with baselines on problem (36). We do not include the empirical results of classical SR1/BFGS/DFP methods since G-SR1 has competitive performance with them Rodomanov and Nesterov (2021a). For all methods, we tune $\mathbf{G}_0$ and M from $\{\mathbf{I}_d, 10\mathbf{I}_d, 10^2\mathbf{I}_d, 10^3\mathbf{I}_d, 10^4\mathbf{I}_d\}$ and $\{1, 10, 10^2, 10^3, 10^4\}$ respectively.



(a) MNIST (iteration) (b) sido0 (iteration) (c) gisette (iteration)

(d) MNIST (time) (e) sido0 (time) (f) gisette (time)

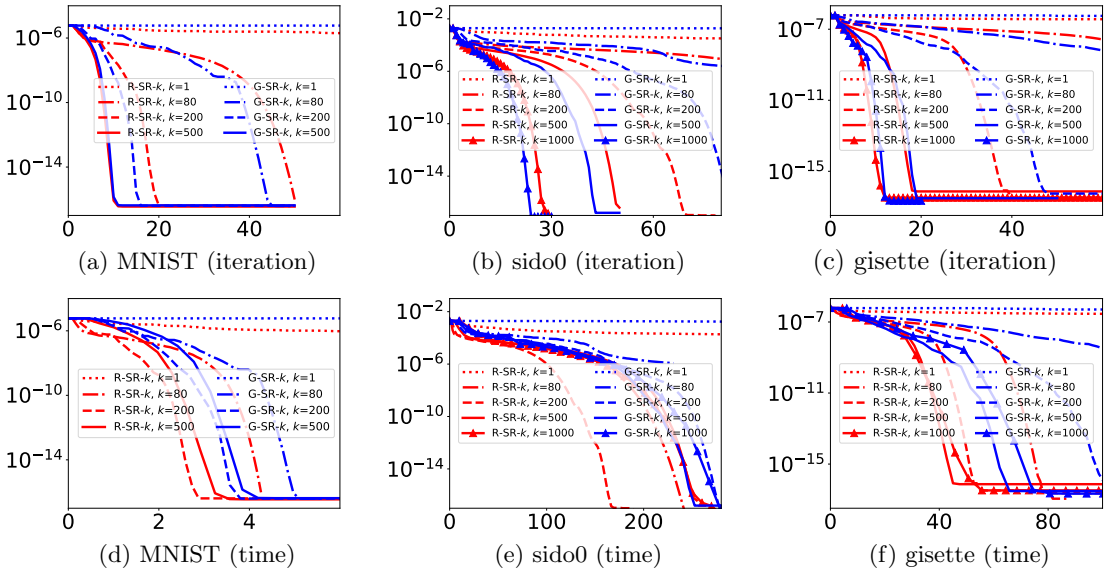
Figure 1: We show "#iteration vs. $\|\nabla f(\mathbf{x})\|$" and "running time (s) vs. $\|\nabla f(\mathbf{x})\|$" on datasets "MNIST", "sido0", and "gisette", where we take $k = 200$ for all of block quasi-Newton methods.

We evaluate the performance on datasets "MNIST" ($d = 780$), "sido0" ($d = 4,932$), and "gisette" ($d = 5,000$). We conduct experiments using Python 3.8.12 on a PC with Apple M1.

We present the results of "iteration numbers vs. gradient norm" and "running time (seconds) vs. gradient norm" in Figure 1, where we take $k = 200$ for block quasi-Newton methods (R-SR-k, G-SR-k, RB-BFGSv1, RB-BFGSv2, and FRB-BFGS). We observe the proposed R-SR-k and G-SR-k significantly outperform other methods. The proposed block BFGS methods (RB-BFGSv2 and FRB-BFGS) outperform the block DFP method (RB-DFP), which is similar to the advantage of the BFGS method over the DFP method (Rodomanov and Nesterov, 2021a).

We also demonstrate the impact of parameter k for SR-k methods. It is natural that Figure 2(a), 2(b), and 2(c) show the larger k leads to faster convergence in terms of iteration numbers, which validates our theoretical analysis in Section 4. In addition, Figure 2(d), 2(e), and 2(f) show SR-k methods with $k > 1$ significantly outperform SR1 in terms of the running time. This is because the block update can reduce the cache miss rate and take advantage of parallel computing. However, increasing k does not always result in less running time because the acceleration caused by the block update is limited by the cache size.

Figure 2 further illustrates the practical trade-off between the randomized and greedy strategies. In most cases, the greedy strategy requires fewer iterations than the randomized strategy, which is consistent with its sharper theoretical guaranty. However, the greedy strategy may take more CPU time because it needs to compute the top-k diagonal entries of matrix $\tilde{\mathbf{G}}_t - \nabla^2 f(\mathbf{x}_{t+1})$ to form $\mathbf{U}_t$.



(a) MNIST (iteration) (b) sido0 (iteration) (c) gisette (iteration)

(d) MNIST (time) (e) sido0 (time) (f) gisette (time)

Figure 2: We show "#iteration vs. $\|\nabla f(\mathbf{x})\|$" and "running time (s) vs. $\|\nabla f(\mathbf{x})\|$" for proposed R-SR-k and G-SR-k with different $k \in \{1, 80, 200, 500\}$ on datasets "MNIST", and $k \in \{1, 80, 200, 500, 1000\}$ on datasets "sido0" and "gisette".

7. Conclusion

In this paper, we have proposed rank-k (SR-k) methods for convex optimization. We have proved that SR-$k$methods enjoy the explicit local superlinear rate of $\mathcal{O}\left((1-k/d)^{t(t-1)/2}\right)$. Our result successfully reveals the advantage of block-type updates in quasi-Newton methods, building a bridge between the theories of ordinary quasi-Newton methods and the standard Newton method. We also provide the convergence rate of $\mathcal{O}\left((1-k/(\varkappa d))^{t(t-1)/2}\right)$ for randomized block BFGS/DFP methods and design a faster randomized block BFGS method to match the rate of SR-$k$methods. In future work, it would be interesting to study the global convergence of block quasi-Newton methods (Jiang et al., 2023; Jiang and Mokhtari, 2023; Rodomanov, 2024; Jin et al., 2025, 2024) and their limited memory (Berahas et al., 2022; Liu and Nocedal, 1989; Gao et al., 2023) as well as stochastic variants (Mokhtari et al., 2018; Wang et al., 2017, 2019; Yang et al., 2022).

Acknowledgments

We are grateful to Peter Richtárik for sharing with us his work on block quasi-Newton methods, which inspired some of the ideas developed in this paper. Chengchang Liu is supported by the National Natural Science Foundation of China (No. 624B2125). Luo Luo is supported by the National Natural Science Foundation of China (No. 12571557) and the Shanghai Basic Research Program (23JC1401000). Cheng Chen is supported by the National Natural Science Foundation of China (No. 62306116) and the Shanghai Special Fund for Promoting High-Quality Industrial Development (No. 20250307).

Appendix A. An Elementary Proof of Lemma 3.4(a)

Before proving Lemma 3.4(a), we first provide the following lemma.

Lemma A.1 *Assume matrix $\mathbf{P} \in \mathbb{R}^{d\times r}$ is column orthonormal with $r \leq d$ and the d-dimensional random vector $\mathbf{v}$ is distributed according to $\mathcal{N}_d(\mathbf{0}, \mathbf{P}\mathbf{P}^\top)$, then it holds that $\mathbb{E}\left[\mathbf{v}\mathbf{v}^\top/(\mathbf{v}^\top\mathbf{v})\right] = \mathbf{P}\mathbf{P}^\top/r$.*

Proof The distribution $\mathbf{v} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{P}\mathbf{P}^\top)$ implies there exists a r-dimensional normally distributed random vector $\mathbf{w} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$ such that $\mathbf{v} = \mathbf{P}\mathbf{w}$. Thus, we have

$$\mathbb{E}\left[\frac{\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top\mathbf{v}}\right] = \mathbb{E}\left[\frac{(\mathbf{P}\mathbf{w})(\mathbf{P}\mathbf{w})^\top}{(\mathbf{P}\mathbf{w})^\top(\mathbf{P}\mathbf{w})}\right] = \mathbb{E}\left[\frac{\mathbf{P}\mathbf{w}\mathbf{w}^\top\mathbf{P}^\top}{\mathbf{w}^\top\mathbf{P}^\top\mathbf{P}\mathbf{w}}\right] = \mathbf{P}\mathbb{E}\left[\frac{\mathbf{w}\mathbf{w}^\top}{\mathbf{w}^\top\mathbf{w}}\right]\mathbf{P}^\top = \frac{1}{r}\mathbf{P}\mathbf{P}^\top,$$

where the last step is because $\mathbf{w}/\|\mathbf{w}\|$ is uniformly distributed on the r-dimensional unit sphere and its covariance matrix is $\mathbf{I}_r/r$. ■

Now we prove statement (a) of Lemma 3.4.

Proof We only need to prove equation (14) by using induction on k. The induction base $k = 1$ have been verified by Lemma A.1. Now we assume equation (14) holds for any $\mathbf{U} \in \mathbb{R}^{d\times k}$ such that each of its entries are independently distributed according to $\mathcal{N}(0, 1)$. We define the random matrix

$$\hat{\mathbf{U}} = \begin{bmatrix}\mathbf{U} & \mathbf{v}\end{bmatrix} \in \mathbb{R}^{d\times(k+1)},$$

where $\mathbf{v} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}_d)$ is independent distributed to $\mathbf{U}$. We define $\mathbf{B} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$, then we use block matrix inversion formula to rewrite $(\hat{\mathbf{U}}^\top\hat{\mathbf{U}})^{-1}$ as follows

$$\begin{aligned}(\hat{\mathbf{U}}^\top\hat{\mathbf{U}})^{-1} &= \left(\begin{bmatrix}\mathbf{U}^\top\\ \mathbf{v}^\top\end{bmatrix}\begin{bmatrix}\mathbf{U} & \mathbf{v}\end{bmatrix}\right)^{-1} = \left(\begin{bmatrix}\mathbf{U}^\top\mathbf{U} & \mathbf{U}^\top\mathbf{v}\\ \mathbf{v}^\top\mathbf{U} & \mathbf{v}^\top\mathbf{v}\end{bmatrix}\right)^{-1}\\ &= \begin{bmatrix}\mathbf{S} & -(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{v}(\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v})^{-1}\\ -(\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v})^{-1}\mathbf{v}^\top\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1} & (\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v})^{-1}\end{bmatrix},\end{aligned}$$

where $\mathbf{S} = (\mathbf{U}^\top\mathbf{U})^{-1} + (\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{v}(\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v})^{-1}\mathbf{v}^\top\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}$. Thus, we have

$$\begin{aligned}\hat{\mathbf{U}}(\hat{\mathbf{U}}^\top\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top &= \begin{bmatrix}\mathbf{U} & \mathbf{v}\end{bmatrix}\left(\begin{bmatrix}\mathbf{U}^\top\\ \mathbf{v}^\top\end{bmatrix}\begin{bmatrix}\mathbf{U} & \mathbf{v}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{U}^\top\\ \mathbf{v}^\top\end{bmatrix}\\ &= \mathbf{B} + \frac{\mathbf{B}\mathbf{v}\mathbf{v}^\top\mathbf{B}}{\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v}} - \frac{\mathbf{B}\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v}} - \frac{\mathbf{v}\mathbf{v}^\top\mathbf{B}}{\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v}} + \frac{\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v}}\\ &= \mathbf{B} + \frac{(\mathbf{I}_d-\mathbf{B})\mathbf{v}\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})}{\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v}}.\end{aligned}$$

Since the rank of the projection matrix $\mathbf{I}_d - \mathbf{B}$ is $d-k$, we can write $\mathbf{I}_d - \mathbf{B} = \mathbf{Q}\mathbf{Q}^\top$, where $\mathbf{Q} \in \mathbb{R}^{d\times(d-k)}$ is column orthonormal. Thus, we achieve

$$\begin{aligned}
&\mathbb{E}\big[\hat{\mathbf{U}}(\hat{\mathbf{U}}^\top\hat{\mathbf{U}})\hat{\mathbf{U}}^\top\big] = \frac{k}{d}\mathbf{I}_d + \mathbb{E}_{\mathbf{U}}\left[\mathbb{E}_{\mathbf{v}}\left[\frac{(\mathbf{I}_d-\mathbf{B})\mathbf{v}\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})}{\mathbf{v}^\top(\mathbf{I}_d-\mathbf{B})\mathbf{v}}\,\Big|\,\mathbf{U}\right]\right] \\
&= \frac{k}{d}\mathbf{I}_d + \mathbb{E}_{\mathbf{U}}\left[\mathbb{E}_{\mathbf{v}}\left[\frac{(\mathbf{Q}\mathbf{Q}^\top\mathbf{v})(\mathbf{v}^\top\mathbf{Q}\mathbf{Q}^\top)}{(\mathbf{v}^\top\mathbf{Q}\mathbf{Q}^\top)(\mathbf{Q}\mathbf{Q}^\top\mathbf{v})}\,\Big|\,\mathbf{U}\right]\right] = \frac{k}{d}\mathbf{I}_d + \frac{1}{d-k}\mathbb{E}_{\mathbf{U}}[\mathbf{Q}\mathbf{Q}^\top] \\
&= \frac{k}{d}\mathbf{I}_d + \frac{1}{d-k}\mathbb{E}_{\mathbf{U}}[\mathbf{I}_d-\mathbf{B}] = \frac{k}{d}\mathbf{I}_d + \frac{1}{d-k}\left(\mathbf{I}_d - \frac{k}{d}\mathbf{I}_d\right) = \frac{k+1}{d}\mathbf{I}_d,
\end{aligned}$$

which completes the induction. In the above derivation, the first and the fifth equalities come from the inductive hypothesis; third equality is achieved by applying Lemma A.1 with $\mathbf{P} = \mathbf{Q}\mathbf{Q}^\top$ and $r = d-k$, and the fact $\mathbf{Q}\mathbf{Q}^\top\mathbf{v} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{Q}\mathbf{Q}^\top)$. ■

Appendix B. Auxiliary Lemmas for Non-negative Sequences

The following lemmas on non-negative sequences extend Theorem 23 of Lin et al. (2022) and Theorem 4.7 of Rodomanov and Nesterov (2021a), leading to a general framework for the analysis of (block) quasi-Newton methods.

Lemma B.1 *Let $\{\lambda_t\}$ and $\{\delta_t\}$ be two non-negative random sequences that satisfy*

$$\mathbb{E}_t\left[\delta_{t+1}\right] \le \left(1 - \frac{1}{\alpha}\right)(1+c_1\lambda_t)^2(\delta_t + c_2\lambda_t), \quad \lambda_{t+1} \le (1+c_1\lambda_t)^2(\delta_t + c_3\lambda_t)\lambda_t, \tag{37}$$

$$\delta_0 + c\lambda_0 \le s, \quad \textit{and} \quad \lambda_t \le \left(1 - \frac{1}{\beta}\right)^t \lambda_0, \tag{38}$$

for some $c_1, c_2, c_3 \ge 0$ with $c = \max\{c_2, c_3\}$, $s \ge 0$, $\alpha > 1$, and $\beta > 1$, where we define $\mathbb{E}_t[\,\cdot\,] \triangleq \mathbb{E}[\,\cdot \,|\, \delta_0, \cdots, \delta_t, \lambda_0, \cdots, \lambda_t]$. If $\lambda_0 > 0$ is sufficient small such that

$$\lambda_0 \le \frac{\ln 2}{\beta(2c_1 + c(\alpha/(\alpha-1)))}, \tag{39}$$

we have $\mathbb{E}\left[\lambda_{t+1}/\lambda_t\right] \le 2s\left(1 - 1/\alpha\right)^t$.

Proof We denote $\theta_t \triangleq \delta_t + c\lambda_t$. Noticing that we have $\exp(x) \ge 1 + x$ for all $x \ge 0$. Applying this fact with $x = c_1\lambda_t$ and the definition $c = \max\{c_2, c_3\}$, we have

$$\mathbb{E}_t[\delta_{t+1}] \overset{(37)}{\le} \left(1 - \frac{1}{\alpha}\right)(\delta_t + c_2\lambda_t)(1+c_1\lambda_t)^2 \le \left(1 - \frac{1}{\alpha}\right)\theta_t \exp\left(2c_1\lambda_t\right) \tag{40}$$

$$\text{and} \quad \lambda_{t+1} \overset{(37)}{\le} (\delta_t + c_3\lambda_t)(1+c_1\lambda_t)^2\lambda_t \le \theta_t\lambda_t\exp\left(2c_1\lambda_t\right). \tag{41}$$

We denote $\tilde{c} \triangleq 2c_1 + c\alpha/(\alpha-1)$, then it holds that

$$\begin{aligned}
\mathbb{E}_t[\theta_{t+1}] &= \mathbb{E}_t[\delta_{t+1} + c\lambda_{t+1}] \\
&\overset{(40),(41)}{\leq} \left(1-\frac{1}{\alpha}\right)\theta_t \exp\left(2c_1\lambda_t\right) + c\theta_t\lambda_t \exp\left(2c_1\lambda_t\right) \\
&= \left(1-\frac{1}{\alpha}\right)\left(1+\frac{c\alpha\lambda_t}{\alpha-1}\right)\theta_t \exp\left(2c_1\lambda_t\right) \\
&\leq \left(1-\frac{1}{\alpha}\right)\theta_t \exp\left(\left(2c_1 + \frac{c\alpha}{\alpha-1}\right)\lambda_t\right) \\
&= \left(1-\frac{1}{\alpha}\right)\theta_t \exp\left(\tilde{c}\lambda_t\right) \overset{(38)}{\leq} \left(1-\frac{1}{\alpha}\right)\theta_t \exp\left(\tilde{c}\left(1-\frac{1}{\beta}\right)^t \lambda_0\right),
\end{aligned} \tag{42}$$

where the second inequality comes from $\exp(x) \geq 1 + x$ with $x = c\alpha\lambda_t/(\alpha-1)$. Taking expectation on both sides of equation (42), we have

$$\mathbb{E}[\theta_{t+1}] \leq \left(1-\frac{1}{\alpha}\right)\exp\left(\tilde{c}\left(1-\frac{1}{\beta}\right)^t \lambda_0\right)\mathbb{E}[\theta_t], \tag{43}$$

where we use the fact $\mathbb{E}[\mathbb{E}_t[\delta_{t+1}]] = \mathbb{E}[\delta_{t+1}]$. Therefore, we finish the proof as follows

$$\begin{aligned}
\mathbb{E}\left[\frac{\lambda_{t+1}}{\lambda_t}\right] &\overset{(41)}{\leq} \mathbb{E}[\theta_t \exp\left(2c_1\lambda_t\right)] \overset{(38)}{\leq} \mathbb{E}[\theta_t]\exp\left(\tilde{c}\big(1-\frac{1}{\beta}\big)^t\lambda_0\right) \\
&\overset{(43)}{\leq} \left(1-\frac{1}{\alpha}\right)\mathbb{E}[\theta_{t-1}]\exp\left(\tilde{c}\big(1-\frac{1}{\beta}\big)^{t-1}\lambda_0 + \tilde{c}\big(1-\frac{1}{\beta}\big)^t\lambda_0\right) \\
&\overset{(43)}{\leq} \left(1-\frac{1}{\alpha}\right)^t \mathbb{E}[\theta_0]\exp\left(\tilde{c}\sum_{i=0}^{t}\big(1-\frac{1}{\beta}\big)^i\lambda_0\right) \\
&\leq \left(1-\frac{1}{\alpha}\right)^t \mathbb{E}[\delta_0 + c\lambda_0]\exp\left(\tilde{c}\beta\lambda_0\right) \overset{(38),(39)}{\leq} \left(1-\frac{1}{\alpha}\right)^t 2s.
\end{aligned}$$

■

Lemma B.2 *Let $\{\lambda_t\}$ and $\{\tilde{\eta}_t\}$ be two positive sequences with $\tilde{\eta}_t \geq 1$ and satisfy*

$$\lambda_{t+1} \leq \left(1-\frac{1}{\tilde{\eta}_t}\right)\lambda_t + \frac{m_1\lambda_t^2}{2} + \frac{m_1^2\lambda_t^3}{4\tilde{\eta}_t} \quad \textit{for all } \lambda_t \textit{ such that } m_1\lambda_t \leq 2 \tag{44}$$

$$\textit{and} \quad \tilde{\eta}_{t+1} \leq (1+m_2\lambda_t)^2\tilde{\eta}_t \tag{45}$$

for some $m_1, m_2 > 0$. If

$$m\lambda_0 \leq \frac{\ln(3/2)}{4\tilde{\eta}_0}, \tag{46}$$

where $m \triangleq \max\{m_1, m_2\}$*, then it holds that*

$$\tilde{\eta}_t \leq \tilde{\eta}_0 \exp\left(2m \sum_{i=0}^{t-1} \lambda_i\right) \leq \frac{3\tilde{\eta}_0}{2} \tag{47}$$

$$\textit{and} \quad \lambda_t \leq \left(1 - \frac{1}{2\tilde{\eta}_0}\right)^t \lambda_0. \tag{48}$$

Proof We prove the results of (47) and (48) by induction. In the case of $t = 0$, inequalities (47) and (48) are satisfied naturally by conditions $\tilde{\eta}_t \geq 1$ and condition (45), where we define $\sum_{i=0}^{t} \lambda_i = 0$ if $t < 0$. Now we suppose inequalities (47) and (48) holds for $t = 0, \ldots, \hat{t}$, then we have

$$m \sum_{i=0}^{\hat{t}} \lambda_i \overset{(48)}{\leq} m\lambda_0 \sum_{i=0}^{\hat{t}} \left(1 - \frac{1}{2\tilde{\eta}_0}\right)^i \leq m\lambda_0 \cdot 2\tilde{\eta}_0 \overset{(46)}{\leq} \frac{\ln(3/2)}{2}. \tag{49}$$

In the case of $t = \hat{t} + 1$, we use induction to achieve

$$\begin{aligned} \frac{1 - m_1\lambda_{\hat{t}}/2}{\tilde{\eta}_{\hat{t}}} &\geq \frac{\exp\left(-m_1\lambda_{\hat{t}}\right)}{\tilde{\eta}_{\hat{t}}} \overset{(47)}{\geq} \frac{\exp\left(-m_1\lambda_{\hat{t}}\right) \cdot \exp\left(-2m\sum_{i=0}^{\hat{t}-1}\lambda_i\right)}{\tilde{\eta}_0} \\ &\geq \frac{\exp\left(-2m\sum_{i=0}^{\hat{t}}\lambda_i\right)}{\tilde{\eta}_0} \overset{(49)}{\geq} \frac{2}{3\tilde{\eta}_0}, \end{aligned} \tag{50}$$

where the first step is due to the fact $\exp(-2x) \leq 1 - x$ with $x = m_1\lambda_{\hat{t}}/2 \in [0, 1/2]$ and the last second step is based on $m \geq m_1$. We also have

$$m_1\lambda_{\hat{t}} \leq m\lambda_{\hat{t}} \overset{(48)}{\leq} m\lambda_0 \overset{(46)}{\leq} \frac{1}{8\tilde{\eta}_0} \leq 2, \tag{51}$$

where the last step is due to $\tilde{\eta}_0 \geq 1$. According to the condition (44), we have

$$\begin{aligned} \lambda_{\hat{t}+1} &\overset{(44)}{\leq} \left(1 - \frac{1}{\tilde{\eta}_{\hat{t}}}\right)\lambda_{\hat{t}} + \frac{m_1\lambda_{\hat{t}}^2}{2} + \frac{m_1^2\lambda_{\hat{t}}^3}{4\tilde{\eta}_{\hat{t}}} = \left(1 + \frac{m_1\lambda_{\hat{t}}}{2}\right) \cdot \left(1 - \frac{1 - m_1\lambda_{\hat{t}}/2}{\tilde{\eta}_{\hat{t}}}\right)\lambda_{\hat{t}} \\ &\overset{(50),(51)}{\leq} \left(1 + \frac{1}{16\tilde{\eta}_0}\right) \cdot \left(1 - \frac{2}{3\tilde{\eta}_0}\right)\lambda_{\hat{t}} \leq \left(1 - \frac{1}{2\tilde{\eta}_0}\right)\lambda_{\hat{t}} \overset{(48)}{\leq} \left(1 - \frac{1}{2\tilde{\eta}_0}\right)^{\hat{t}+1}\lambda_0. \end{aligned}$$

The induction also implies

$$\tilde{\eta}_{\hat{t}+1} \overset{(45)}{\leq} (1 + m_2\lambda_{\hat{t}})^2\tilde{\eta}_{\hat{t}} \leq \tilde{\eta}_{\hat{t}} \exp\left(2m\lambda_{\hat{t}}\right) \overset{(47)}{\leq} \tilde{\eta}_0 \exp\left(2m\sum_{i=0}^{\hat{t}}\lambda_{\hat{t}}\right) \overset{(49)}{\leq} \frac{3\tilde{\eta}_0}{2},$$

where the second inequality is due to the fact $\exp(x) \geq 1 + x$ with $x = m_2\lambda_{\hat{t}} \geq 0$. Thus, we finish the proof. ■

Appendix C. The Proof of Theorem 4.2

We first present a lemma from Lin et al. (2022).

Lemma C.1 (Lin et al. (2022, Lemma 25)) *If the twice differentiable function $f(\cdot)$ satisfies Assumptions 2.2 and 2.4, and the positive-definite matrix $\mathbf{G} \in \mathbb{R}^{d\times d}$ and the vector $\mathbf{x} \in \mathbb{R}^d$ satisfy $\nabla^2 f(\mathbf{x}) \preceq \mathbf{G} \preceq \eta \nabla^2 f(\mathbf{x})$ for some $\eta > 1$, then we have*

$$\nabla^2 f(\mathbf{x}_+) \preceq \tilde{\mathbf{G}} \preceq \eta(1+Mr)^2 \nabla^2 f(\mathbf{x}_+) \tag{52}$$

$$\text{and} \quad \tau_{\nabla^2 f(\mathbf{x}_+)}(\tilde{\mathbf{G}}) \leq (1+Mr)^2 \left(\frac{\tau_{\nabla^2 f(\mathbf{x})}(\mathbf{G})}{\operatorname{tr}(\nabla^2 f(\mathbf{x}))} + 2Mr \right) \operatorname{tr}\big(\nabla^2 f(\mathbf{x}_+)\big) \tag{53}$$

for any $\mathbf{x}_+ \in \mathbb{R}^d$, where $\tilde{\mathbf{G}} = (1+Mr)\mathbf{G}$ and $r = \|\mathbf{x} - \mathbf{x}_+\|_{\mathbf{x}}$.

Now, we prove Theorem 4.2 by using Lemma 3.2, 4.1, B.2, and C.1.

Proof Denote $\tilde{\eta}_t \triangleq \min_{\mathbf{G}_t \preceq \eta \nabla^2 f(\mathbf{x}_t)} \eta$. We first prove that for all $t \geq 0$, it holds

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{G}_t \preceq \tilde{\eta}_t \nabla^2 f(\mathbf{x}_t). \tag{54}$$

The result of $\mathbf{G}_t \preceq \tilde{\eta}_t \nabla^2 f(\mathbf{x}_t)$ holds by the definition of $\tilde{\eta}_t$, and we prove $\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{G}_t$ by induction. For the case of $t = 0$, it holds naturally by the initial condition. Suppose it holds for $t = 0, 1, \ldots, \hat{t}$. Then for the case of $t = \hat{t} + 1$, we have

$$\mathbf{G}_{\hat{t}+1} = \text{SR-}k(\tilde{\mathbf{G}}_{\hat{t}}, \nabla^2 f(\mathbf{x}_{\hat{t}+1}), \mathbf{U}_{\hat{t}}) \overset{(7)}{\succeq} \tilde{\mathbf{G}}_{\hat{t}} \overset{(52)}{\succeq} \nabla^2 f(\mathbf{x}_{\hat{t}+1}),$$

which finishes the induction.

Let $\lambda_t \triangleq \lambda(\mathbf{x}_t)$. Applying equation (54) and Lemma 4.1, we achieve

$$\lambda_{t+1} \leq \left(1 - \frac{1}{\tilde{\eta}_t}\right) \lambda_t + \frac{M\lambda_t^2}{2} + \frac{M^2 \lambda_t^3}{4\tilde{\eta}_t} \quad \text{for all } \lambda_t \text{ such that } M\lambda_t \leq 2.$$

We then figure out the relation between $\tilde{\eta}_{t+1}$ and $\tilde{\eta}_t$. Let $r_t = \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_{\mathbf{x}_t}$. Applying equation (54) and Lemma C.1 with $\mathbf{G} = \mathbf{G}_t$, $\mathbf{x} = \mathbf{x}_t$, $\mathbf{x}_+ = \mathbf{x}_{t+1}$, $\eta = \tilde{\eta}_t$, and $r = r_t$, we achieve

$$\nabla^2 f(\mathbf{x}_{t+1}) \preceq \tilde{\mathbf{G}}_t \preceq \tilde{\eta}_t (1+Mr_t)^2 \nabla^2 f(\mathbf{x}_{t+1}). \tag{55}$$

Using Lemma 3.2 with $\mathbf{G} = \tilde{\mathbf{G}}_t$, $\mathbf{A} = \nabla^2 f(\mathbf{x}_{t+1})$, and $\eta = \tilde{\eta}_t(1+Mr_t)^2$, we have

$$\mathbf{G}_{t+1} \overset{(7),(55)}{\preceq} \tilde{\eta}_t (1+Mr_t)^2 \nabla^2 f(\mathbf{x}_{t+1}) \overset{(17)}{\preceq} \tilde{\eta}_t (1+M\lambda_t)^2 \nabla^2 f(\mathbf{x}_{t+1}),$$

which means $\tilde{\eta}_{t+1} = \min_{\mathbf{G}_{t+1} \preceq \eta \nabla^2 f(\mathbf{x}_{t+1})} \eta \leq \tilde{\eta}_t (1+M\lambda_t)^2$. Hence, sequences $\{\tilde{\eta}_t\}$ and $\{\lambda_t\}$ satisfy the conditions of Lemma B.2 with $m_1 = m_2 = M$. We then apply Lemma B.2 to obtain

$$\nabla^2 f(\mathbf{x}_t) \preceq \tilde{\mathbf{G}}_t \preceq \frac{3\tilde{\eta}_0}{2} \nabla^2 f(\mathbf{x}_t) \preceq \frac{3\eta_0}{2} \nabla^2 f(\mathbf{x}_t)$$

and

$$\lambda(\mathbf{x}_t) \leq \left(1 - \frac{1}{2\tilde{\eta}_0}\right)^t \lambda(\mathbf{x}_0) \leq \left(1 - \frac{1}{2\eta_0}\right)^t \lambda(\mathbf{x}_0).$$

■

Appendix D. The Proof of Theorem 4.3

We first provide some auxiliary lemmas that will be used in our later proof.

Lemma D.1 *For any positive definite symmetric matrices $\mathbf{G}, \mathbf{H} \in \mathbb{R}^{d\times d}$ such that $\mathbf{H} \preceq \mathbf{G}$, it holds that*

$$\mathbf{G} \preceq \left(1 + \frac{\hat{\varkappa} d \tau_{\mathbf{H}}(\mathbf{G})}{\mathrm{tr}(\mathbf{H})}\right)\mathbf{H}, \tag{56}$$

where $\hat{\varkappa}$ is the condition number of $\mathbf{H}$ and the notation of $\tau_{\mathbf{H}}(\mathbf{G})$ follows the expression of (8)*. If it further satisfies that $\mathbf{G} \preceq \eta\mathbf{H}$ for some $\eta \geq 1$, then we have*

$$\frac{\tau_{\mathbf{H}}(\mathbf{G})}{\mathrm{tr}(\mathbf{H})} \leq \eta - 1. \tag{57}$$

Proof We obtain inequality (56) by statement 3) of equation (42) from Lin et al. (Lin et al., 2022, Lemma 25). We prove inequality (57) by the definition of $\tau_{\mathbf{H}}(\mathbf{G})$ as follows

$$\frac{\tau_{\mathbf{H}}(\mathbf{G})}{\mathrm{tr}(\mathbf{H})} = \frac{\mathrm{tr}(\mathbf{G}-\mathbf{H})}{\mathrm{tr}(\mathbf{H})} \leq \frac{\mathrm{tr}((\eta-1)\mathbf{H})}{\mathrm{tr}(\mathbf{H})} = \eta - 1.$$

■

Lemma D.2 (Lin et al. (2022, Lemma 26)) *Suppose the non-negative random sequence $\{X_t\}$ satisfies $\mathbb{E}[X_t] \leq c\,(1-1/\alpha)^t$ for all $t \geq 0$ and some constants $c \geq 0$ and $\alpha > 1$. Then for any $\delta \in (0,1)$, we have $X_t \leq c\alpha^2\,(1-1/(1+\alpha))^t/\delta$ for all t with probability at least $1-\delta$.*

Now, we provide the proof of Theorem 4.3.

Proof We denote $\mathbb{E}_t[\,\cdot\,] \triangleq \mathbb{E}[\,\cdot \mid \mathbf{U}_0, \cdots, \mathbf{U}_{t-1}]$, $g_t = \tau_{\nabla^2 f(\mathbf{x}_t)}(\mathbf{G}_t)/\mathrm{tr}(\nabla^2 f(\mathbf{x}_t))$, $\lambda_t = \lambda(\mathbf{x}_t)$, $r_t = \|\mathbf{x}_{t+1}-\mathbf{x}_t\|_{\mathbf{x}_t}$, and $\delta_t = d\varkappa g_t$. The initial condition (19) means we can apply Theorem 4.2 to achieve

$$\nabla^2 f(\mathbf{x}_t) \preceq \mathbf{G}_t \qquad \text{and} \qquad \lambda_t \leq \left(1 - \frac{1}{2\eta_0}\right)^t \lambda_0. \tag{58}$$

According to Theorem 3.6, we have

$$\mathbb{E}_t\left[\tau_{\nabla^2 f(\mathbf{x}_{t+1})}(\mathbf{G}_{t+1})\right] \overset{(8)}{\leq} \left(1 - \frac{k}{d}\right)\tau_{\nabla^2 f(\mathbf{x}_{t+1})}(\tilde{\mathbf{G}}_t). \tag{59}$$

According to Lemma C.1 with $\mathbf{G} = \mathbf{G}_t$, $\mathbf{x} = \mathbf{x}_t$, $\mathbf{x}_+ = \mathbf{x}_{t+1}$, and $r = r_t$, we have

$$\tau_{\nabla^2 f(\mathbf{x}_{t+1})}(\tilde{\mathbf{G}}_t) \overset{(53)}{\leq} (1+Mr_t)^2(g_t + 2Mr_t)\mathrm{tr}(\nabla^2 f(\mathbf{x}_{t+1})), \tag{60}$$

which means

$$\begin{aligned}\mathbb{E}_t[\delta_{t+1}] = \mathbb{E}_t\left[\frac{d\varkappa\tau_{\nabla^2 f(\mathbf{x}_{t+1})}(\mathbf{G}_{t+1})}{\mathrm{tr}(\nabla^2 f(\mathbf{x}_{t+1}))}\right] &\overset{(59)}{\leq} \left(1-\frac{k}{d}\right)\left(\frac{d\varkappa\tau_{\nabla^2 f(\mathbf{x}_{t+1})}(\tilde{\mathbf{G}}_t)}{\mathrm{tr}(\nabla^2 f(\mathbf{x}_{t+1}))}\right)\\ &\overset{(60)}{\leq} \left(1-\frac{k}{d}\right)\left(\frac{d\varkappa(1+Mr_t)^2(g_t+2Mr_t)\mathrm{tr}(\nabla^2 f(\mathbf{x}_{t+1}))}{\mathrm{tr}(\nabla^2 f(\mathbf{x}_{t+1}))}\right)\\ &\overset{(17)}{\leq} \left(1-\frac{k}{d}\right)(1+M\lambda_t)^2(\delta_t+2\varkappa dM\lambda_t).\end{aligned} \tag{61}$$

According to Lemma D.1 with $\mathbf{G}=\mathbf{G}_t$ and $\mathbf{H}=\nabla^2 f(\mathbf{x}_t)$, we have

$$\nabla^2 f(\mathbf{x}_t) \overset{(58)}{\preceq} \mathbf{G}_t \overset{(56)}{\preceq} (1+\delta_t)\nabla^2 f(\mathbf{x}_t).$$

According to Lemma 4.1 with $\eta_t = 1+\delta_t$, we have

$$\begin{aligned}\lambda_{t+1} &\overset{(17)}{\leq} \left(1-\frac{1}{1+\delta_t}\right)\lambda_t + \frac{M\lambda_t^2}{2} + \frac{M^2\lambda_t^3}{4(1+\delta_t)}\\ &= \left(1+\frac{M\lambda_t}{2}\right)\cdot\left(\frac{\delta_t+M\lambda_t/2}{1+\delta_t}\right)\lambda_t \ \leq (1+M\lambda_t)^2\left(\delta_t+\frac{M\lambda_t}{2}\right)\lambda_t.\end{aligned} \tag{62}$$

According to Lemma D.1 with $\mathbf{G}=\mathbf{G}_0$, $\mathbf{H}=\nabla^2 f(\mathbf{x}_0)$, $\eta=\eta_0$, and the initial condition (19), we have

$$\delta_0 = \frac{d\varkappa\tau_{\nabla^2 f(\mathbf{x}_0)}(\mathbf{G}_0)}{\mathrm{tr}(\nabla^2 f(\mathbf{x}_0))} \overset{(57)}{\leq} d\varkappa(\eta_0-1) \tag{63}$$

$$\text{and}\quad \delta_0 + 2d\varkappa M\lambda_0 \overset{(63)}{\leq} d\varkappa(\eta_0-1)+2d\varkappa M\lambda_0 \overset{(19)}{\leq} d\varkappa\eta_0. \tag{64}$$

Then we apply Lemma B.1 on the random sequences $\{\lambda_t\}$ and $\{\delta_t\}$ with $c_1=M$, $c_2=M/2$, $c_3=2\varkappa dM$, $\alpha=d/k$, $\beta=2\eta_0$, and $s=d\varkappa\eta_0$ to achieve inequality (20), where we can verify conditions (37) and (38) by equations (61), (62) and (58), (64) respectively.

Now, we prove the two-stage convergence of SR-k methods as follows.

(a) For SR-k method with randomized strategy such that $[\mathbf{U}_t]_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$, we apply Lemma D.2 with $X_t=\lambda_{t+1}/\lambda_t$, $\alpha=d/k$ and $c=2d\varkappa\eta_0$ to obtain

$$\frac{\lambda_{t+1}}{\lambda_t} \leq \frac{2d^3\varkappa\eta_0}{k^2\delta}\left(1-\frac{k}{d+k}\right)^t \tag{65}$$

holds for all t with probability at least $1-\delta$. We take $t_0=\mathcal{O}(d\ln(d\varkappa\eta_0/\delta)/k)$, which leads to

$$\frac{2d^3\varkappa\eta_0}{k^2\delta}\left(1-\frac{k}{d+k}\right)^{t_0} \leq \frac{1}{2}. \tag{66}$$

Together with the linear convergence (18) provided by Theorem 4.2, we have

$$\begin{aligned}\lambda_{t+t_0} &\overset{(65)}{\leq} \frac{2d^3\varkappa\eta_0}{k^2\delta}\cdot\left(1-\frac{k}{d+k}\right)^{t+t_0-1}\cdot\lambda_{t+t_0-1}\\ &\overset{(66)}{\leq}\left(1-\frac{k}{d+k}\right)^{t-1}\cdot\frac{\lambda_{t+t_0-1}}{2}\leq\cdots\overset{(65),(66)}{\leq}\left(1-\frac{k}{d+k}\right)^{t(t-1)/2}\cdot\left(\frac{1}{2}\right)^t\lambda_{t_0}\\ &\overset{(18)}{\leq}\left(1-\frac{k}{d+k}\right)^{t(t-1)/2}\cdot\left(\frac{1}{2}\right)^t\cdot\left(1-\frac{1}{2\eta_0}\right)^{t_0}\lambda_0\end{aligned}$$

with probability at least $1-\delta$.

(b) For SR-k method with greedy strategy such that $\mathbf{U}_t=\mathbf{E}_k(\tilde{\mathbf{G}}_t-\nabla^2 f(\mathbf{x}_{t+1}))$, we can take $t_0=\mathcal{O}\left(d\ln(\eta_0 d\varkappa)/k\right)$ such that

$$2d\varkappa\eta_0\left(1-\frac{k}{d}\right)^{t_0}\leq\frac{1}{2}. \tag{67}$$

Together with the linear convergence (18) provided by Theorem 4.2, we have

$$\begin{aligned}\lambda_{t+t_0} &\overset{(20)}{\leq} 2d\varkappa\eta_0\left(1-\frac{k}{d}\right)^{t+t_0-1}\lambda_{t+t_0-1}\\ &\overset{(67)}{\leq}\left(1-\frac{k}{d}\right)^{t-1}\cdot\frac{\lambda_{t+t_0-1}}{2}\leq\cdots\overset{(20),(67)}{\leq}\left(1-\frac{k}{d}\right)^{t(t-1)/2}\cdot\left(\frac{1}{2}\right)^t\lambda_{t_0}\\ &\overset{(18)}{\leq}\left(1-\frac{k}{d}\right)^{t(t-1)/2}\cdot\left(\frac{1}{2}\right)^t\cdot\left(1-\frac{1}{2\eta_0}\right)^{t_0}\lambda_0.\end{aligned}$$

■

Appendix E. The Proof of Theorem 5.5

We first provide the linear convergence for the proposed block BFGS and DFP methods (Algorithm 2).

Lemma E.1 *Under the setting of Theorem 5.5, Algorithm 2 holds that*

$$\lambda(\mathbf{x}_t)\leq\left(1-\frac{1}{2\eta_0}\right)^t\lambda(\mathbf{x}_0)\quad\textit{and}\quad\nabla^2 f(\mathbf{x}_t)\preceq\mathbf{G}_t\preceq\frac{3\eta_0}{2}\nabla^2 f(\mathbf{x}_t) \tag{68}$$

for all $t\geq 0$.

Proof We can obtain this result by following the proof of Theorem 4.2 by replacing the steps of using Lemma 3.2 with using Lemma 5.2. ■

We then provide some auxiliary lemmas for later analysis.

Lemma E.2 (Rodomanov and Nesterov, 2021a, Lemma 4.8)) *If the twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is M-strongly self-concordant and μ-strongly convex and the positive definite matrix $\mathbf{G} \in \mathbb{R}^{d\times d}$ and $\mathbf{x} \in \mathbb{R}^d$ satisfy $\nabla^2 f(\mathbf{x}) \preceq \mathbf{G} \preceq \eta \nabla^2 f(\mathbf{x})$ for some $\eta > 1$, then we have*

$$\sigma_{\nabla^2 f(\mathbf{x}_+)}(\tilde{\mathbf{G}}) \leq (1 + Mr)^2(\sigma_{\nabla^2 f(\mathbf{x})}(\mathbf{G}) + 2dMr) \tag{69}$$

for any $\mathbf{x}, \mathbf{x}_+ \in \mathbb{R}^d$ where $\tilde{\mathbf{G}} = (1 + Mr)\mathbf{G}$ and $r = \|\mathbf{x} - \mathbf{x}_+\|_{\mathbf{x}}$.

Lemma E.3 *For any positive definite symmetric matrices $\mathbf{G}, \mathbf{H} \in \mathbb{R}^{d\times d}$ such that $\mathbf{H} \preceq \mathbf{G}$, we have*

$$\mathbf{G} \preceq (1 + \sigma_{\mathbf{H}}(\mathbf{G}))\mathbf{H}. \tag{70}$$

If it further satisfies that $\mathbf{G} \preceq \eta\mathbf{H}$ for some $\eta \geq 1$, then we have

$$\sigma_{\mathbf{H}}(\mathbf{G}) \leq d(\eta - 1). \tag{71}$$

Proof We obtain inequality (70) by statement 1) of Equation (42) by Lin et al. (Lin et al., 2022, Lemma 25). We prove inequality (71) by the definition of $\sigma_{\mathbf{H}}(\mathbf{G})$ as follows

$$\begin{aligned}\sigma_{\mathbf{H}}(\mathbf{G}) &= \operatorname{tr}\big(\mathbf{H}^{-1}(\mathbf{G} - \mathbf{H})\big) = \operatorname{tr}\big(\mathbf{H}^{-1/2}(\mathbf{G} - \mathbf{H})\mathbf{H}^{-1/2}\big)\\ &\leq \operatorname{tr}\big((\eta - 1)\mathbf{H}^{-1/2}\mathbf{H}\mathbf{H}^{-1/2}\big) = d(\eta - 1).\end{aligned}$$

■

Now, we present the proof for Theorem 5.5.

Proof Denote $\delta_t \triangleq \sigma_{\nabla^2 f(\mathbf{x}_t)}(\mathbf{G}_t)$ and $\lambda_t \triangleq \lambda(\mathbf{x}_t)$. According to Lemma E.1 and Lemma E.3 with $\mathbf{G} = \mathbf{G}_t$ and $\mathbf{H} = \nabla^2 f(\mathbf{x}_t)$, we have

$$\nabla^2 f(\mathbf{x}_t) \overset{(68)}{\preceq} \mathbf{G}_t \overset{(70)}{\preceq} (1 + \delta_t)\nabla^2 f(\mathbf{x}_t). \tag{72}$$

According to Theorem 5.3 with $\mathbf{A} = \nabla^2 f(\mathbf{x}_{t+1})$ and $\mathbf{G} = \tilde{\mathbf{G}}_t$, we have

$$\mathbb{E}_t[\delta_{t+1}] = \mathbb{E}_t\big[\sigma_{\nabla^2 f(\mathbf{x}_{t+1})}(\mathbf{G}_{t+1})\big] \overset{(26)}{\leq} \left(1 - \frac{k}{d\varkappa}\right)\sigma_{\nabla^2 f(\mathbf{x}_{t+1})}(\tilde{\mathbf{G}}_t). \tag{73}$$

According to Lemma E.2 with $\mathbf{G} = \mathbf{G}_t$ and $\mathbf{x} = \mathbf{x}_t$, we have

$$\sigma_{\nabla^2 f(\mathbf{x}_{t+1})}(\tilde{\mathbf{G}}_t) \overset{(69)}{\leq} (1 + Mr_t)^2(\delta_t + 2dMr_t). \tag{74}$$

Thus, we can obtain following result

$$\begin{aligned}\mathbb{E}_t[\delta_{t+1}] &\overset{(73)}{\leq} \left(1 - \frac{k}{d\varkappa}\right)\sigma_{\nabla^2 f(\mathbf{x}_{t+1})}(\tilde{\mathbf{G}}_t)\\ &\overset{(74)}{\leq} \left(1 - \frac{k}{d\varkappa}\right)(1 + Mr_t)^2(\delta_t + 2dMr_t)\\ &\overset{(17)}{\leq} \left(1 - \frac{k}{d\varkappa}\right)(1 + M\lambda_t)^2(\delta_t + 2dM\lambda_t).\end{aligned} \tag{75}$$

We follow Lemma 4.1 with $\eta_t = 1 + \delta_t$ and the derivation of equation (62) to achieve

$$\lambda_{t+1} \leq (1 + M\lambda_t)^2 \left(\delta_t + \frac{M\lambda_t}{2}\right) \lambda_t. \tag{76}$$

According to Lemma E.1, we have

$$\lambda_t \leq \left(1 - \frac{1}{2\eta_0}\right)^t \lambda_0. \tag{77}$$

According to Lemma E.3 with $\mathbf{G} = \mathbf{G}_0$, $\mathbf{H} = \nabla^2 f(\mathbf{x}_0)$, $\eta = \eta_0$, and the initial condition (32), we have

$$\delta_0 = \sigma_{\nabla^2 f(\mathbf{x}_0)}(\mathbf{G}_0) \overset{(71)}{\leq} d(\eta_0 - 1) \qquad \text{and} \qquad \delta_0 + 2dM\lambda_0 \overset{(32)}{\leq} d\eta_0. \tag{78}$$

Then we apply Lemma B.1 on the random sequences $\{\lambda_t\}$ and $\{\delta_t\}$ with $c_1 = M$, $c_2 = M/2$, $c_3 = 2dM$, $\alpha = d\varkappa/k$, $\beta = 2\eta_0$, and $s = d\eta_0$ to finish the proof, where we can verify conditions (37) and (38) by equations (75), (76) and (77), (78) respectively. ■

Appendix F. The proof of Theorem 5.8

Proof We define $\delta_t \triangleq \sigma_{\nabla^2 f(\mathbf{x}_t)}(\mathbf{G}_t)$ and $\lambda_t \triangleq \lambda(\mathbf{x}_t)$. The proof of this theorem can follow the counterpart of Theorem 5.5 by using $\tilde{\mathbf{L}}_t^\top \mathbf{U}_t$ to replace $\mathbf{U}_t$ (line 8 in Algorithm 3). This leads to the faster rate for matrix approximation, i.e., applying Theorem 5.6 with $\mathbf{G} = \tilde{\mathbf{G}}_t$, $\mathbf{A} = \nabla^2 f(\mathbf{x}_{t+1})$, and $\mathbf{U} = \mathbf{U}_t$ to achieve

$$\mathbb{E}_t[\delta_{t+1}] = \mathbb{E}_t[\sigma_{\nabla^2 f(\mathbf{x}_{t+1})}(\mathbf{G}_{t+1})] \overset{(34)}{\leq} \left(1 - \frac{k}{d}\right) \sigma_{\nabla^2 f(\mathbf{x}_{t+1})}(\tilde{\mathbf{G}}_t). \tag{79}$$

We follow the proof of Theorem 5.5 by replacing equation (73) with (79) and replacing all the terms of $\varkappa d$ with d to finish the proof of Theorem 5.8. ■

References

Azam Asl, Haihao Lu, and Jinwen Yang. A J-symmetric quasi-Newton method for minimax problems. *Mathematical Programming*, 204(1–2):207–254, 2024.

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

Jaya P. N. Bishwal. *Parameter estimation in stochastic differential equations*, volume 1923 of *Lecture Notes in Mathematics*. Springer Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-74448-1.

Nicolas Boutet, Rob Haelterman, and Joris Degroote. Secant update version of quasi-Newton PSB with weighted multisecant equations. *Computational Optimization and Applications*, 75(2):441–466, 2020.

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

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

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, 1970b.

Charles G. Broyden, J. E. Dennis, and Jorge J. Moré. On the local and superlinear convergence of quasi-Newton methods. *IMA Journal of Applied Mathematics*, 12(3): 223–245, 1973.

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.

Yasuko Chikuse. *Statistics on special manifolds*. Springer, 2003.

Yu-Hong Dai. Convergence properties of the BFGS algoritm. *SIAM Journal on Optimization*, 13(3):693–701, 2002.

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

J. E. Dennis, Jr. and Jorge J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Mathematics of Computation*, 28(126):549–560, 1974.

Roger Fletcher and Michael J. D. Powell. A rapidly convergent descent method for minimization. *The Computer Journal*, 6(2):163–168, 1963.

Wenbo Gao and Donald Goldfarb. Block BFGS methods. *SIAM Journal on Optimization*, 28(2):1205–1231, 2018.

Zhan Gao, Aryan Mokhtari, and Alec Koppel. Limited-memory greedy quasi-Newton method with non-asymptotic superlinear convergence rate. *preprint, arXiv:2306.15444v2 [math.OC]*, 2023.

Donald Goldfarb, Yi Ren, and Achraf Bahamou. Practical quasi-Newton methods for training deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 2386–2396, 2020.

Robert M. Gower and Peter Richtárik. Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1380–1409, 2017.

Robert M. Gower, Donald Goldfarb, and Peter Richtárik. Stochastic block BFGS: Squeezing more curvature out of data. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 1869–1878. PMLR, 2016.

Philipp Hennig and Martin Kiefel. Quasi-Newton methods: A new direction. *Journal of Machine Learning Research*, 14(26):843–865, 2013.

Mortaza Jamshidian and Robert I. Jennrich. Acceleration of the EM algorithm by using quasi-Newton methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 59(3):569–587, 1997.

Zhen-Yuan Ji and Yu-Hong Dai. Greedy PSB methods with explicit superlinear convergence. *Computational Optimization and Applications*, 85(3):753–786, 2023.

Ruichen Jiang and Aryan Mokhtari. Accelerated quasi-Newton proximal extragradient: Faster rate for smooth convex optimization. In *Advances in Neural Information Processing Systems*, volume 36, pages 8114–8151, 2023.

Ruichen Jiang, Qiujiang Jin, and Aryan Mokhtari. Online learning guided curvature approximation: A quasi-Newton method with global non-asymptotic superlinear convergence. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 1962–1992. PMLR, 2023.

Qiujiang Jin and Aryan Mokhtari. Non-asymptotic superlinear convergence of standard quasi-Newton methods. *Mathematical Programming*, 200(1):425–473, 2023.

Qiujiang Jin, Alec Koppel, Ketan Rajawat, and Aryan Mokhtari. Sharpened quasi-Newton methods: Faster superlinear rate and larger local convergence neighborhood. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 10228–10250. PMLR, 2022.

Qiujiang Jin, Ruichen Jiang, and Aryan Mokhtari. Non-asymptotic global convergence analysis of BFGS with the Armijo-Wolfe line search. In *Advances in Neural Information Processing Systems*, volume 37, pages 16810–16851, 2024. doi: 10.52202/079017-0536.

Qiujiang Jin, Ruichen Jiang, and Aryan Mokhtari. Non-asymptotic global convergence rates of BFGS with exact line search. *Mathematical Programming*, 2025. doi: 10.1007/s10107-025-02256-7.

Dmitry Kovalev, Robert M. Gower, Peter Richtárik, and Alexander Rogozin. Fast linear convergence of randomized BFGS. *preprint, arXiv:2002.11337v4 [math.OC]*, 2020.

Ching-pei Lee, Cong Han Lim, and Stephen J. Wright. A distributed quasi-Newton algorithm for empirical risk minimization with nonsmooth regularization. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1646–1655, 2018. doi: 10.1145/3219819.3220075.

Mokhwa Lee and Yifan Sun. Almost multisecant BFGS quasi-Newton method. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*, pages 908–915, 2024. doi: 10.1109/IEEECONF60004.2024.10942893.

Zhigang Li, Wenchuan Wu, Boming Zhang, Hongbin Sun, and Qinglai Guo. Dynamic economic dispatch using Lagrangian relaxation with multiplier updates based on a quasi-Newton method. *IEEE Transactions on Power Systems*, 28(4):4516–4527, 2013.

Dachao Lin, Haishan Ye, and Zhihua Zhang. Explicit convergence rates of greedy and random quasi-Newton methods. *Journal of Machine Learning Research*, 23(162):1–40, 2022.

Chengchang Liu and Luo Luo. Quasi-Newton methods for saddle point problems. In *Advances in Neural Information Processing Systems*, volume 35, pages 3975–3987, 2022.

Chengchang Liu, Shuxian Bi, Luo Luo, and John C. S. Lui. Partial-Quasi-Newton methods: Efficient algorithms for minimax optimization problems with unbalanced dimensionality. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1031–1041, 2022. doi: 10.1145/3534678.3539379.

Chengchang Liu, Cheng Chen, Luo Luo, and John C. S. Lui. Block Broyden's methods for solving nonlinear equations. In *Advances in Neural Information Processing Systems*, volume 36, pages 47487–47499, 2023. doi: 10.52202/075280-2056.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1–3):503–528, 1989.

Alexander Ludwig. The Gauss–Seidel–quasi-Newton method: A hybrid algorithm for solving dynamic economic models. *Journal of Economic Dynamics and Control*, 31(5):1610–1632, 2007.

Aryan Mokhtari, Mark Eisen, and Alejandro Ribeiro. IQN: An incremental quasi-Newton method with local superlinear convergence rate. *SIAM Journal on Optimization*, 28(2): 1670–1698, 2018.

Yurii Nesterov. *Lectures on convex optimization*. Springer, 2018.

Dianne P. O'Leary and A. Yeremin. The linear algebra of block quasi-Newton algorithms. *Linear Algebra and its Applications*, 212–213:153–168, 1994.

M. J. D. Powell. On the convergence of the variable metric algorithm. *IMA Journal of Applied Mathematics*, 7(1):21–36, 1971.

Zicheng Qiu, Jie Jiang, and Xiaojun Chen. A quasi-Newton subspace trust region algorithm for nonmonotone variational inequalities in adversarial learning over box constraints. *Journal of Scientific Computing*, 101(2):45, 2024. doi: 10.1007/s10915-024-02679-y.

Anton Rodomanov. Global complexity analysis of BFGS. *preprint, arXiv:2404.15051v1 [math.OC]*, 2024.

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

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, 2021b.

Anton Rodomanov and Yurii Nesterov. Rates of superlinear convergence for classical quasi-Newton methods. *Mathematical Programming*, 194(1–2):159–190, 2022.

Robert B. Schnabel. Quasi-Newton methods using multiple secant equations. Technical report, University of Colorado Boulder, 1983.

David F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

Xiao Wang, Shiqian Ma, Donald Goldfarb, and Wei Liu. Stochastic quasi-Newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017.

Xiaoyu Wang, Xiao Wang, and Ya-xiang Yuan. Stochastic proximal quasi-Newton methods for non-convex composite optimization. *Optimization Methods and Software*, 34(5):922–948, 2019.

Max A. Woodbury. Inverting modified matrices. Memorandum Report 42, Statistical Research Group, Princeton University, 1950.

Minghan Yang, Andre Milzarek, Zaiwen Wen, and Tong Zhang. A stochastic extra-step quasi-Newton method for nonsmooth nonconvex optimization. *Mathematical Programming*, 194(1–2):257–303, 2022.

Haishan Ye, Dachao Lin, Xiangyu Chang, and Zhihua Zhang. Towards explicit superlinear convergence rate for SR1. *Mathematical Programming*, 199(1–2):1273–1303, 2023.

Ya-Xiang Yuan. A modified BFGS algorithm for unconstrained optimization. *IMA Journal of Numerical Analysis*, 11(3):325–332, 1991.

Yichuan Zhang and Charles Sutton. Quasi-Newton methods for Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems*, volume 24, pages 2393–2401, 2011.