Mathematical Programming
https://doi.org/10.1007/s10107-022-01887-4

FULL LENGTH PAPER

Series A

Non-asymptotic superlinear convergence of standard quasi-Newton methods

Qiujiang Jin[1] · **Aryan Mokhtari**[1]

Received: 16 December 2020 / Accepted: 29 August 2022

Abstract

In this paper, we study and prove the non-asymptotic superlinear convergence rate of the Broyden class of quasi-Newton algorithms which includes the Davidon–Fletcher–Powell (DFP) method and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method. The asymptotic superlinear convergence rate of these quasi-Newton methods has been extensively studied in the literature, but their explicit finite–time local convergence rate is not fully investigated. In this paper, we provide a finite–time (non-asymptotic) convergence analysis for Broyden quasi-Newton algorithms under the assumptions that the objective function is strongly convex, its gradient is Lipschitz continuous, and its Hessian is Lipschitz continuous at the optimal solution. We show that in a local neighborhood of the optimal solution, the iterates generated by both DFP and BFGS converge to the optimal solution at a superlinear rate of $(1/k)^{k/2}$, where k is the number of iterations. We also prove a similar local superlinear convergence result holds for the case that the objective function is self-concordant. Numerical experiments on several datasets confirm our explicit convergence rate bounds. Our theoretical guarantee is one of the first results that provide a non-asymptotic superlinear convergence rate for quasi-Newton methods.

Keywords Quasi-Newton method · Superlinear convergence rate · Non-asymptotic analysis · DFP algorithm · BFGS algorithm

Mathematics Subject Classification 90-08 · 90C25 · 90C53

✉ Qiujiang Jin
qiujiang@austin.utexas.edu

Aryan Mokhtari
mokhtari@austin.utexas.edu

[1] Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA

Published online: 17 September 2022

1 Introduction

In this paper, we focus on the non-asymptotic convergence analysis of quasi-Newton methods for the problem of minimizing a convex function $f : \mathbb{R}^d \to \mathbb{R}$, i.e.,

$$\min_{x \in \mathbb{R}^d} f(x).$$

Specifically, we focus on two different settings. In the first case, we assume that the objective function f is strongly convex, smooth (its gradient is Lipschitz continuous), and its Hessian is Lipschitz continuous at the optimal solution. In the second case, we study the setting where the objective function f is self-concordant. We formally define these settings in the following sections. In both considered cases, the optimal solution is unique and denoted by x_*.

There is an extensive literature on the use of first-order methods for convex optimization, and it is well-known that the best achievable convergence rate for first-order methods, when the objective function is strongly convex and smooth, is a linear convergence rate. Specifically, we say a sequence $\{x_k\}$ converges linearly if $\|x_k - x_*\| \leq C\gamma^k \|x_0 - x_*\|$, where $\gamma \in (0, 1)$ is the constant of linear convergence, and C is a constant possibly depending on problem parameters. Among first-order methods, the accelerated gradient method proposed in [1] achieves a fast linear convergence rate of $(1 - \sqrt{\mu/L})^{k/2}$, where μ is the strong convexity parameter and L is the smoothness parameter (the Lipschitz constant of the gradient) [2]. It is also known that the convergence rate of the accelerated gradient method is optimal for first-order methods in the setting that the problem dimension d is sufficiently larger than the number of iterations [3].

Classical alternatives to improve the convergence rate of first-order methods are second-order methods [4–7] and in particular Newton's method. It has been shown that if in addition to smoothness and strong convexity assumptions, the objective function f has a Lipschitz continuous Hessian, then the iterates generated by Newton's method converge to the optimal solution at a quadratic rate in a local neighborhood of the optimal solution; see [8, Chapter 9]. A similar result has been established for the case that the objective function is self-concordant [9]. Despite the fact that the quadratic convergence rate of Newton's method holds only in a local neighborhood of the optimal solution, it could reduce the overall number of iterations significantly as it is substantially faster than the linear rate of first-order methods. The fast quadratic convergence rate of Newton's method, however, does not come for free. Implementation of Newton's method requires solving a linear system at each iteration with the matrix defined by the objective function Hessian $\nabla^2 f(x)$. As a result, the computational cost of implementing Newton's method in high-dimensional problems is prohibitive, as it could be $\mathcal{O}(d^3)$, unlike first-order methods that have a per iteration cost of $\mathcal{O}(d)$.

Quasi-Newton algorithms are quite popular since they serve as a middle ground between first-order methods and Newton-type algorithms. They improve the linear convergence rate of first-order methods and achieve a local superlinear rate, while their computational cost per iteration is $\mathcal{O}(d^2)$ instead of $\mathcal{O}(d^3)$ of Newton's method. The main idea of quasi-Newton methods is to approximate the step of Newton's method

without computing the objective function Hessian $\nabla^2 f(x)$ or its inverse $\nabla^2 f(x)^{-1}$ at every iteration [10, Chapter 6]. To be more specific, quasi-Newton methods aim at approximating the curvature of the objective function by using only first-order information of the function, i.e., its gradients $\nabla f(x)$; see Sect. 2 for more details. There are several different approaches for approximating the objective function Hessian and its inverse using first-order information, which leads to different quasi-Newton updates, but perhaps the most popular quasi-Newton algorithms are the Symmetric Rank-One (SR1) method [11], the Broyden method [12–14], the Davidon-Fletcher-Powell (DFP) method [15, 16], the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [17–20], and the limited-memory BFGS (L-BFGS) method [21, 22].

As mentioned earlier, a major advantage of quasi-Newton methods is their asymptotic local superlinear convergence rate. More precisely, we state that the sequence $\{x_k\}$ converges to the optimal solution x_* superlinearly when the ratio between the distance to the optimal solution at time $k+1$ and k approaches zero as k approaches infinity, i.e.,

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = 0.$$

For various settings, this superlinear convergence result has been established for a large class of quasi-Newton methods, including the Broyden method [13, 17, 23], the DFP method [13, 24, 25], the BFGS method [13, 25–27], and several other variants of these algorithms [28–34]. Although this result is promising and lies between the linear rate of first-order methods and the quadratic rate of Newton's method, it only holds asymptotically and does not characterize an explicit upper bound on the error of quasi-Newton methods after a finite number of iterations. As a result, the overall complexity of quasi-Newton methods for achieving an ϵ-accurate solution, i.e., $\|x_k - x_*\| \leq \epsilon$, cannot be explicitly characterized. Hence, it is essential to establish a non-asymptotic convergence rate for quasi-Newton methods, which is the main goal of this paper.

In this paper, we show that if the initial iterate is close to the optimal solution and the initial Hessian approximation error is sufficiently small, then the iterates of the convex Broyden class including both the DFP and BFGS methods converge to the optimal solution at a superlinear rate of $(1/k)^{k/2}$. We further show that our theoretical result suggests a trade-off between the size of the superlinear convergence neighborhood and the rate of superlinear convergence. In other words, one can improve the numerical constant in the above rate at the cost of reducing the radius of the neighborhood in which DFP and BFGS converge superlinearly. We believe that our theoretical guarantee provides one of the first non-asymptotic results for the superlinear convergence rate of BFGS and DFP.

Related work In a recent work [35], the authors studied the non-asymptotic analysis of a class of *greedy* quasi-Newton methods that are based on the update rule of the Broyden family and use a greedily selected basis vectors for updating Hessian approximations. In particular, they show a superlinear convergence rate of $(1 - \frac{\mu}{dL})^{k^2/2}(\frac{dL}{\mu})^k$ for this class of algorithms. However, greedy quasi-Newton methods are more computationally costly than standard quasi-Newton methods, as they require computing a greedily selected basis vector at each iteration. It is worth noting that such computation

requires access to additional information beyond the objective function gradient, e.g., the diagonal components of the Hessian. Also, two recent concurrent papers study the non-asymptotic superlinear convergence rate of the DFP and BFGS methods [36, 37]. In [36], the authors show that when the objective function is smooth, strongly convex, and strongly self-concordant, the iterates of BFGS and DFP, in a local neighborhood of the optimal solution, achieve the superlinear convergence rate of $(\frac{dL}{\mu k})^{k/2}$ and $(\frac{dL^2}{\mu^2 k})^{k/2}$, respectively. In their follow-up paper [37], they improve the superlinear convergence results to $[e^{\frac{d}{k}\ln\frac{L}{\mu}}-1]^{k/2}$ and $[\frac{L}{\mu}(e^{\frac{d}{k}\ln\frac{L}{\mu}}-1)]^{k/2}$, respectively. We would like to highlight that the proof techniques, assumptions, and final theoretical results of [36, 37] and our paper are different and derived independently. The major difference in the analysis is that in [36, 37], the authors use a potential function related to the trace and the logarithm of the determinant of the Hessian approximation matrix, while we use a Frobenius norm potential function. In addition, our convergence rates for both DFP and BFGS are independent of the problem dimension d. Nevertheless, in our results, the neighborhood of superlinear convergence depends on d. Moreover, to derive our results we consider two settings where in the first case the objective function is strongly convex, smooth, and has a Lipschitz continuous Hessain at the optimal solution, and in the second setting the function is self-concordant. Both of these settings are more general than the setting in [36, 37], which requires the objective function to be strongly convex, smooth, and strongly self-concordant.

Outline In Sect. 2, we discuss the Broyden class of quasi-Newton methods, DFP and BFGS. In Sect. 3, we mention our assumptions, notations as well as some general technical lemmas. Then, in Sect. 4, we present the main theoretical results of our paper on the non-asymptotic superlinear convergence of DFP and BFGS for the setting that the objective function is strongly convex, smooth, and its Hessian is Lipschitz continuous at the optimal solution. In Sect. 5, we extend our theoretical results to the class of self-concordant functions, by exploiting the proof techniques developed in Sect. 4. In Sect. 6, we provide a detailed discussion on the advantages and drawbacks of our theoretical results and compare them with some concurrent works. In Sect. 7, we numerically evaluate the performance of DFP and BFGS on several datasets and compare their convergence rates with our theoretical bounds. Finally, in Sect. 8, we close the paper with some concluding remarks.

Notation For vector $v \in \mathbb{R}^d$, its Euclidean norm (l-2 norm) is denoted by $\|v\|$. We denote the Frobenius norm of matrix $A \in \mathbb{R}^{d\times d}$ as $\|A\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^d A_{ij}^2}$ and its induced 2-norm is denoted by $\|A\| = \max_{\|v\|=1} \|Av\|$. The trace of matrix A, which is the sum of its diagonal elements, is denoted by $\mathrm{Tr}\,(A)$. For any two symmetric matrices $A, B \in \mathbb{R}^{d\times d}$, we denote that $A \preceq B$ if and only if $B - A$ is a symmetric positive semidefinite matrix.

2 Quasi-Newton methods

In this section, we review standard quasi-Newton methods, and, in particular, we discuss the updates of the DFP and BFGS algorithms. Consider a time index k, a step

size η_k, and a positive-definite matrix B_k to define a generic descent algorithm through the iteration

$$x_{k+1} = x_k - \eta_k B_k^{-1} \nabla f(x_k). \tag{1}$$

Note that if we simply replace B_k by the identity matrix I, we recover the update of gradient descent, and if we replace it by the objective function Hessian $\nabla^2 f(x_k)$, we obtain the update of Newton's method. The main goal of quasi-Newton methods is to find a symmetric positive-definite matrix B_k using only first-order information such that B_k is close to the Hessian $\nabla^2 f(x_k)$. Note that the step size η_k is often computed according to a line search routine for the global convergence of quasi-Newton methods. Our focus in this paper, however, is on the local convergence of quasi-Newton methods, which requires the unit step size $\eta_k = 1$. Hence, in the rest of the paper, we assume that the iterate x_k is sufficiently close to the optimal solution x_* and the step size is $\eta_k = 1$.

In most quasi-Newton methods, the function's curvature is approximated in a way that it satisfies the *secant condition*. To better explain this property, let us first define the variable difference s_k and gradient difference y_k as

$$s_k = x_{k+1} - x_k, \quad \text{and} \quad y_k = \nabla f(x_{k+1}) - \nabla f(x_k). \tag{2}$$

The goal is to find a matrix B_{k+1} that satisfies the secant condition $B_{k+1} s_k = y_k$. The rationale for satisfying the secant condition is that the Hessian $\nabla^2 f(x_k)$ approximately satisfies this condition when x_{k+1} and x_k are close to each other, e.g., they are both close to the optimal solution x_*. However, the secant condition alone is not sufficient to specify B_{k+1}. To resolve this indeterminacy, different quasi-Newton algorithms consider different additional conditions. One common constraint is to enforce the Hessian approximation (or its inverse) at time $k+1$ to be close to the one computed at time k. This is a reasonable extra condition as we expect the Hessian (or its inverse) evaluated at x_{k+1} to be close to the one computed at x_k.

In the DFP method, we enforce the proximity condition on Hessian approximations B_k and B_{k+1}. Basically, we aim to find the closest positive-definite matrix to B_k (in some weighted matrix norm) that satisfies the secant condition; see Chapter 6 of [10] for more details. The update of the Hessian approximation matrices of DFP is given by

$$B_{k+1}^{\text{DFP}} = \left(I - \frac{y_k s_k^\top}{y_k^\top s_k}\right) B_k \left(I - \frac{s_k y_k^\top}{s_k^\top y_k}\right) + \frac{y_k y_k^\top}{y_k^\top s_k}. \tag{3}$$

Since implementation of the update in (1) requires access to the inverse of the Hessian approximation, it is essential to derive an explicit update for the Hessian inverse approximation to avoid the cost of inverting a matrix at each iteration. If we define H_k as the inverse of B_k, i.e., $H_k = B_k^{-1}$, using the Sherman-Morrison-Woodbury formula, one can write

$$H_{k+1}^{\text{DFP}} = H_k - \frac{H_k y_k y_k^\top H_k}{y_k^\top H_k y_k} + \frac{s_k s_k^\top}{s_k^\top y_k}. \tag{4}$$

The BFGS method can be considered as the dual of DFP. In BFGS, we also seek a positive-definite matrix that satisfies the secant condition, but instead of forcing the proximity condition on the Hessian approximation B, we enforce it on the Hessian inverse approximation H. To be more precise, we aim to find a positive-definite matrix H_{k+1} that satisfies the secant condition $s_k = H_{k+1} y_k$ and is the closest matrix (in some weighted norm) to the previous Hessian inverse approximation H_k. The update of the Hessian inverse approximation matrices of BFGS is given by,

$$H_{k+1}^{\text{BFGS}} = \left(I - \frac{s_k y_k^\top}{y_k^\top s_k}\right) H_k \left(I - \frac{y_k s_k^\top}{s_k^\top y_k}\right) + \frac{s_k s_k^\top}{y_k^\top s_k}. \tag{5}$$

Similarly, by the Sherman-Morrison-Woodbury formula, the update of BFGS method for the Hessian approximation matrices is given by,

$$B_{k+1}^{\text{BFGS}} = B_k - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{s_k^\top y_k}. \tag{6}$$

Note that both DFP and BFGS belong to a more general class of quasi-Newton methods called the Broyden class. The Hessian approximation B_{k+1} of the Broyden class is defined as

$$B_{k+1} = \phi_k B_{k+1}^{\text{DFP}} + (1 - \phi_k) B_{k+1}^{\text{BFGS}}, \tag{7}$$

and the Hessian inverse approximation is defined as

$$H_{k+1} = (1 - \psi_k) H_{k+1}^{\text{DFP}} + \psi_k H_{k+1}^{\text{BFGS}}, \tag{8}$$

where $\phi_k, \psi_k \in \mathbb{R}$. In this paper, we only focus on the convex class of Broyden quasi-Newton methods, where $\phi_k, \psi_k \in [0, 1]$. The steps of this class of methods are summarized in Algorithm 1. In fact, in Algorithm 1, if we set $\psi_k = 0$, we recover DFP, and if we set $\psi_k = 1$, we recover BFGS. It is worth noting that the cost of computing the descent direction $H_k \nabla f(x_k)$ for this class of quasi-Newton methods is of $\mathcal{O}(d^2)$, which improves $\mathcal{O}(d^3)$ per iteration cost of Newton's method.

Remark 1 Note that when $s_k = 0$, we have $\nabla f(x_k) = 0$ from (1) and thus $x_k = x_*$. Hence, in our implementation and analysis we assume $s_k \neq 0$. Moreover, in both considered settings, the objective function is at least strictly convex. As a result, if $s_k \neq 0$, then it follows that $y_k \neq 0$ and $s_k^\top y_k > 0$. This observation shows that the updates of BFGS and DFP are well-defined. Finally, it is well-known that for the convex class of Broyden methods if B_k is symmetric positive-definite and $s_k^\top y_k > 0$, then B_{k+1} is also symmetric positive-definite [10]. In Algorithm 1, we assume that the initial Hessian approximation B_0 is symmetric positive-definite, and, hence,

Algorithm 1 The convex Broyden class of quasi-Newton methods

Require: Initial iterate x_0 and initial Hessian inverse approximation H_0.
1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Update the variable: $x_{k+1} = x_k - H_k \nabla f(x_k)$;
3: Compute the variable difference $s_k = x_{k+1} - x_k$;
4: **if** $s_k = 0$ **then**
5: Terminate the algorithm
6: **else**
7: Compute the gradient difference $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$;
8: Update the Hessian inverse approximation $H_{k+1} = (1 - \psi_k) H_{k+1}^{\text{DFP}} + \psi_k H_{k+1}^{\text{BFGS}}$ according to (4) and (5);
9: **end if**
10: **end for**

all Hessian approximation matrices B_k and their inverse matrices H_k are symmetric positive-definite.

3 Preliminaries

In this section, we first specify the required assumptions for our results in Sect. 4 and introduce some notations to simplify our expressions. Moreover, we present some intermediate lemmas that will be use later in Sect. 4 to prove our main theoretical results for the setting that the objective function is strongly convex, smooth, and its Hessian is Lipschitz continuous at the optimal solution. In Sect. 5, we will use a subset of these intermediate results to extend our analysis to the class of self-concordant functions.

3.1 Assumptions

We next state the required assumptions for establishing our theoretical results in Sect. 4.

Assumption 3.1 The objective function $f(x)$ is twice-differentiable. Moreover, the function $f(x)$ is strongly convex with parameter $\mu > 0$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \geq \mu \|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{9}$$

Assumption 3.2 The gradient of the objective function $f(x)$ is Lipschitz continuous with parameter $L > 0$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{10}$$

As f is twice-differentiable, Assumption 3.1 and 3.2 imply that the eigenvalues of the Hessian are larger than μ and smaller than L, i.e., $\mu I \preceq \nabla^2 f(x) \preceq L I, \forall x \in \mathbb{R}^d$. Note that for our main theoretical results, we only require Assumption 3.1, but to compare our results with other theoretical bounds we will use the condition in Assumption 3.2 in our discussions.

Assumption 3.3 The Hessian $\nabla^2 f(x)$ satisfies the following condition for some constant $M \geq 0$,

$$\|\nabla^2 f(x) - \nabla^2 f(x_*)\| \leq M\|x - x_*\|, \quad \forall x \in \mathbb{R}^d. \tag{11}$$

The condition in Assumption 3.3 is common for analyzing second-order methods as we require a regularity condition on the objective function Hessian. In fact, Assumption 3.3 is one of the least strict conditions required for the analysis of second-order type methods as it requires Lipschitz continuity of the Hessian only at (near) the optimal solution. This condition is, indeed, weaker than assuming that the Hessian is Lipschitz continuous everywhere. Note that for the class of strongly convex and smooth functions, the strongly self-concordance assumption required in [36, 37] is equivalent to assuming that the Hessian is Lipschitz continuous everywhere. Hence, the condition in Assumption 3.3 is also weaker than the one in [36, 37]. Assumption 3.3 leads to the following corollary.

Corollary 1 *If the condition in Assumption 3.3 holds, then for all $x, y \in \mathbb{R}^d$, we have*

$$\|\nabla f(x) - \nabla f(y) - \nabla^2 f(x_*)(x - y)\| \leq \frac{M}{2}\|x - y\|(\|x - x_*\| + \|y - x_*\|). \tag{12}$$

Proof Check Appendix A. □

Remark 2 Our analysis can be extended to the case that Assumptions 3.1, 3.2 and 3.3 only hold in a local neighborhood of the optimal solution x_*. Here, we assume they hold in $\mathbb{R}^d$ to simplify our proofs.

3.2 Notations

Next, we briefly mention some of the definitions and notations that will be used in following theorems and proofs. We consider $\nabla^2 f(x_*)^{\frac{1}{2}}$ and $\nabla^2 f(x_*)^{-\frac{1}{2}}$ as the square root of the matrices $\nabla^2 f(x_*)$ and $\nabla^2 f(x_*)^{-1}$, i.e., $\nabla^2 f(x_*) = \nabla^2 f(x_*)^{\frac{1}{2}} \nabla^2 f(x_*)^{\frac{1}{2}}$ and $\nabla^2 f(x_*)^{-1} = \nabla^2 f(x_*)^{-\frac{1}{2}} \nabla^2 f(x_*)^{-\frac{1}{2}}$. By Assumption 3.1, both $\nabla^2 f(x_*)^{\frac{1}{2}}$ and $\nabla^2 f(x_*)^{-\frac{1}{2}}$ are symmetric positive-definite. Throughout the paper, we analyze and study weighted version of the Hessian approximation $\hat{B}_k$ defined as

$$\hat{B}_k = \nabla^2 f(x_*)^{-\frac{1}{2}} B_k \nabla^2 f(x_*)^{-\frac{1}{2}}. \tag{13}$$

$\hat{B}_k$ is symmetric positive-definite, since B_k and $\nabla^2 f(x_*)^{-\frac{1}{2}}$ are both symmetric positive-definite. We also use $\|\hat{B}_k - I\|_F$ as the measure of closeness between B_k and $\nabla^2 f(x_*)$, which can be written as

$$\|\hat{B}_k - I\|_F = \|\nabla^2 f(x_*)^{-\frac{1}{2}} \left(B_k - \nabla^2 f(x_*)\right) \nabla^2 f(x_*)^{-\frac{1}{2}}\|_F. \tag{14}$$

We further define the weighted gradient difference $\hat{y}_k$, the weighted variable difference $\hat{s}_k$, and the weighted gradient $\widehat{\nabla f}(x_k)$ as

$$\hat{y}_k = \nabla^2 f(x_*)^{-\frac{1}{2}} y_k, \qquad \hat{s}_k = \nabla^2 f(x_*)^{\frac{1}{2}} s_k, \qquad \widehat{\nabla f}(x_k) = \nabla^2 f(x_*)^{-\frac{1}{2}} \nabla f(x_k). \tag{15}$$

To measure closeness to the optimal solution for iterate x_k, we use $r_k \in \mathbb{R}^d$, $\sigma_k \in \mathbb{R}$, and $\tau_k \in \mathbb{R}$ which are formally defined as

$$r_k = \nabla^2 f(x_*)^{\frac{1}{2}} (x_k - x_*), \qquad \sigma_k = \frac{M}{\mu^{\frac{3}{2}}} \|r_k\|, \qquad \tau_k = \max\{\sigma_k, \sigma_{k+1}\}. \tag{16}$$

In (16), μ is the strong convexity parameter defined in Assumption 3.1 and M is the Lipschitz continuity parameter of the Hessian at the optimal solution defined in Assumption 3.3. In our analysis, we also use the average Hessian J_k and its weighted version $\hat{J}_k$ that are formally defined as

$$J_k = \int_0^1 \nabla^2 f(x_* + \alpha(x_k - x_*)) d\alpha, \qquad \hat{J}_k = \nabla^2 f(x_*)^{-\frac{1}{2}} J_k \nabla^2 f(x_*)^{-\frac{1}{2}}. \tag{17}$$

3.3 Intermediate Lemmas

Next, we present some lemmas that we will later use to establish the non-asymptotic superlinear convergence of DFP and BFGS. Proofs of these lemmas are relegated to the appendix.

Lemma 1 *For any matrix $A \in \mathbb{R}^{d\times d}$ and vector $u \in \mathbb{R}^d$ with $\|u\| = 1$, we have*

$$\|A\|_F^2 - \|(I - uu^\top)A(I - uu^\top)\|_F^2 \geq \|Au\|^2. \tag{18}$$

Proof Check Appendix B. □

Lemma 2 *For any matrices $A, B \in \mathbb{R}^{d\times d}$, we have*

$$\|AB\|_F \leq \|A\|\|B\|_F, \qquad \|B^\top AB\|_F \leq \|A\|\|B\|\|B\|_F. \tag{19}$$

Proof Check Appendix C. □

The results in Lemma 1 and Lemma 2 hold for arbitrary matrices. The next lemma focuses on some properties of the weighted average Hessian $\hat{J}_k$ under Assumptions 3.1 and 3.3.

Lemma 3 *Recall the definition of σ_k in (16) and $\hat{J}_k$ in (17). Suppose $\alpha_k \in [0, 1]$ and define the matrix $H_k = \nabla^2 f(x_* + \alpha_k(x_k - x_*))$ and $\hat{H}_k = \nabla^2 f(x_*)^{-\frac{1}{2}} H_k \nabla^2 f(x_*)^{-\frac{1}{2}}$. If Assumptions 3.1 and 3.3 hold, then the following inequalities hold for all $k \geq 0$,*

$$\frac{1}{1+\frac{\sigma_k}{2}} I \preceq \hat{J}_k \preceq (1+\frac{\sigma_k}{2}) I, \qquad \frac{1}{1+\sigma_k} I \preceq \hat{H}_k \preceq (1+\sigma_k) I. \tag{20}$$

Proof Check Appendix D. □

In the following lemma, we establish some bounds that depend on the weighted gradient difference $\hat{y}_k$ and the weighted variable difference $\hat{s}_k$.

Lemma 4 *Recall the definitions in (13–16). If Assumptions 3.1 and 3.3 hold, then the following inequalities hold for all $k \geq 0$,*

$$\|\hat{y}_k - \hat{s}_k\| \leq \tau_k \|\hat{s}_k\|, \tag{21}$$

$$(1-\tau_k)\|\hat{s}_k\|^2 \leq \hat{s}_k^\top \hat{y}_k \leq (1+\tau_k)\|\hat{s}_k\|^2, \tag{22}$$

$$(1-\tau_k)\|\hat{s}_k\| \leq \|\hat{y}_k\| \leq (1+\tau_k)\|\hat{s}_k\|, \tag{23}$$

$$\|\widehat{\nabla f(x_k)} - r_k\| \leq \frac{\sigma_k}{2}\|r_k\|. \tag{24}$$

Proof Check Appendix E. □

4 Main theoretical results

In this section, we characterize the non-asymptotic superlinear convergence of the Broyden class of quasi-Newton methods, when Assumptions 3.1, 3.2 and 3.3 hold. In Sect. 4.1, we first establish a crucial proposition which characterizes the error of Hessian approximation for this class of quasi-Newton methods. Then, in Sect. 4.2, we leverage this result to show that the iterates of this class of algorithms converge at least linearly to the optimal solution, if the initial distance to the optimal solution and the initial Hessian approximation error are sufficiently small. Finally, we use these intermediate results in Sect. 4.3 to prove that the iterates of the convex Broyden class, including both DFP and BFGS, converge to the optimal solution at a superlinear rate of $(1/k)^{k/2}$. Note that in Algorithm 1 we use the Hessian inverse approximation matrix H_k to describe the algorithm, but in our analysis we will study the behavior of the Hessian approximation matrix B_k.

4.1 Hessian approximation error: Frobenius norm potential function

Next, we use the Frobenius norm of the Hessian approximation error $\|\hat{B}_k - I\|_F$ as the potential function in our analysis. Specifically, we will use the results of Lemma 1, Lemma 2, and Lemma 4 to study the dynamic of the Hessian approximation error $\|\hat{B}_k - I\|_F$ for both DFP and BFGS. First, start with the DFP method.

Lemma 5 *Consider the update of DFP in (3) and recall the definition of τ_k in (16). Suppose that for some $\delta > 0$ and some $k \geq 0$, we have that $\tau_k < 1$ and $\|\hat{B}_k - I\|_F \leq \delta$. Then, the matrix B_{k+1}^{DFP} generated by the DFP update satisfies the following inequality*

$$\|\hat{B}_{k+1}^{DFP} - I\|_F \leq \|\hat{B}_k - I\|_F - \frac{\|(\hat{B}_k - I)\hat{s}_k\|^2}{2\delta\|\hat{s}_k\|^2} + W_k\tau_k, \tag{25}$$

where $W_k = \|\hat{B}_k\|\frac{4}{(1-\tau_k)^2} + \frac{3+\tau_k}{1-\tau_k}$.

Proof The proof and conclusion of this lemma are similar to the ones in Lemma 3.2 in [33], except the value of parameter W_k. This difference comes from the fact that [33] analyzed the modified DFP update, while we consider the standard DFP method. Recall the DFP update in (3) and multiply both sides of that expression by the matrix $\nabla^2 f(x_*)^{-\frac{1}{2}}$ from left and right to obtain

$$\hat{B}_{k+1}^{\text{DFP}} = \left(I - \frac{\hat{y}_k\hat{s}_k^\top}{\hat{y}_k^\top\hat{s}_k}\right)\hat{B}_k\left(I - \frac{\hat{s}_k\hat{y}_k^\top}{\hat{s}_k^\top\hat{y}_k}\right) + \frac{\hat{y}_k\hat{y}_k^\top}{\hat{y}_k^\top\hat{s}_k}, \tag{26}$$

where we used the fact that $s_k^\top y_k = s_k^\top\nabla^2 f(x_*)^{\frac{1}{2}}\nabla^2 f(x_*)^{-\frac{1}{2}} y_k = \hat{s}_k^\top\hat{y}_k$. To simplify the proof, we use the following notations:

$$\begin{aligned} &B = \hat{B}_k, \quad B_+ = \hat{B}_{k+1}^{\text{DFP}}, \quad s = \hat{s}_k, \quad y = \hat{y}_k, \quad \tau = \tau_k, \quad P = I - \frac{ss^\top}{\|s\|^2}, \\ &Q = \frac{ss^\top}{\|s\|^2} - \frac{sy^\top}{s^\top y}. \end{aligned} \tag{27}$$

Hence, (26) is equivalent to

$$B_+ = \left(I - \frac{ys^\top}{s^\top y}\right)B\left(I - \frac{sy^\top}{s^\top y}\right) + \frac{yy^\top}{s^\top y}.$$

Moreover, we can express $B_+ - I$ as

$$\begin{aligned} B_+ - I &= (P + Q^\top)B(P + Q) - I + \frac{yy^\top}{s^\top y} \\ &= PBP + Q^\top BP + PBQ + Q^\top BQ - I + \frac{yy^\top}{s^\top y} \\ &= P(B - I)P + P^2 - I + \frac{yy^\top}{s^\top y} + Q^\top BP + PBQ + Q^\top BQ. \end{aligned} \tag{28}$$

Notice that $P^2 = P$ and $P = P^\top$. Thus, (28) can be simplified as

$$B_+ - I = D + E + G^\top + G + H,$$

where

$$D = P(B - I)P, \quad E = \frac{yy^\top}{s^\top y} - \frac{ss^\top}{\|s\|^2}, \quad G = PBQ, \quad H = Q^\top BQ.$$

Next, we proceed to upper bound $\|B_+ - I\|_F$. To do so, we derive upper bounds on the Frobenius norm of matrices D, E, G and H. We start by $\|D\|_F$. If we set $u = s/\|s\|$ and $A = B - I$ in Lemma 1, we obtain that

$$\frac{\|(B - I)s\|^2}{\|s\|^2} \leq \|B - I\|_F^2 - \|D\|_F^2, \tag{29}$$

which implies $\|B-I\|_F^2-\|D\|_F^2 \geq 0$. Moreover, using the fact that $a^2-b^2 \leq 2a(a-b)$ we can write

$$\|B - I\|_F^2 - \|D\|_F^2 \leq 2\|B - I\|_F(\|B - I\|_F - \|D\|_F) \leq 2\delta(\|B - I\|_F - \|D\|_F), \tag{30}$$

where the second inequality follows from the fact that $\|B - I\|_F^2 - \|D\|_F^2 \geq 0$ and the assumption that $\|B - I\|_F \leq \delta$. Next, if we replace the right hand side of (29) by its upper bound in (30) and massage the resulted expression, we obtain that

$$\|D\|_F \leq \|B - I\|_F - \frac{\|(B - I)s\|^2}{2\delta\|s\|^2}, \tag{31}$$

which provides an upper bound on $\|D\|_F$. To derive upper bounds for $\|E\|_F$, $\|G\|_F$ and $\|H\|_F$, we first need to find an upper bound for $\|Q\|_F$, where Q is defined in (27). Note that

$$\begin{aligned}\|Q\|_F &= \left\| \frac{ss^\top}{\|s\|^2} - \frac{sy^\top}{s^\top y} \right\|_F = \left\| \frac{ss^\top}{\|s\|^2} - \frac{sy^\top}{\|s\|^2} + \frac{sy^\top}{\|s\|^2} - \frac{sy^\top}{s^\top y} \right\|_F \\ &\leq \left\| \frac{ss^\top}{\|s\|^2} - \frac{sy^\top}{\|s\|^2} \right\|_F + \left\| \frac{sy^\top}{\|s\|^2} - \frac{sy^\top}{s^\top y} \right\|_F,\end{aligned}$$

where the first equality holds by the definition of Q, the second equality is obtained by adding and subtracting $\frac{sy^\top}{\|s\|^2}$, and the inequality holds due to the triangle inequality. We can further simplify the right hand side as

$$\begin{aligned}\|Q\|_F &\leq \frac{\|s(s - y)^\top\|_F}{\|s\|^2} + \frac{|s^T(s - y)|\|sy^\top\|_F}{\|s\|^2 s^\top y} \leq \frac{\|y - s\|}{\|s\|} + \frac{\|y - s\|\|y\|}{s^\top y} \\ &\leq \tau + \frac{\tau(1 + \tau)\|s\|^2}{(1 - \tau)\|s\|^2} = \frac{2\tau}{1 - \tau},\end{aligned} \tag{32}$$

where the second inequality holds using the Cauchy–Schwarz inequality and the fact that $\|ab^\top\|_F = \|a\|\|b\|$ for $a, b \in \mathbb{R}^d$, and the last inequality holds due to the results in (21), (22), and (23).

Next using the upper bound in (32) on $\|Q\|_F$ we derive an upper bound on $\|E\|_F$. Note that

$$\begin{aligned}\|E\|_F &= \left\|\frac{yy^\top}{s^\top y} - \frac{ss^\top}{\|s\|^2}\right\|_F = \left\|\frac{yy^\top}{s^\top y} - \frac{sy^\top}{s^\top y} + \frac{sy^\top}{s^\top y} - \frac{ss^\top}{\|s\|^2}\right\|_F \\ &\le \left\|\frac{yy^\top}{s^\top y} - \frac{sy^\top}{s^\top y}\right\|_F + \left\|\frac{sy^\top}{s^\top y} - \frac{ss^\top}{\|s\|^2}\right\|_F,\end{aligned}$$

where we used the triangle inequality in the last step. Using the definition of Q we can show that

$$\begin{aligned}\|E\|_F &\le \frac{\|(y-s)y^\top\|_F}{s^\top y} + \|Q\|_F \le \frac{\|y-s\|\|y\|}{s^\top y} + \frac{2\tau}{1-\tau} \\ &\le \frac{\tau(1+\tau)\|s\|^2}{(1-\tau)\|s\|^2} + \frac{2\tau}{1-\tau} = \frac{3+\tau}{1-\tau}\tau,\end{aligned} \tag{33}$$

where for the second inequality we use (32) and $\|ab^\top\|_F = \|a\|\|b\|$, and for the third inequality we use the results in (21), (22), and (23).

We proceed to derive an upper bound for $\|G\|_F$. Note that $0 \preceq P \preceq I$ and thus $\|P\| \le 1$. Using this observation, (32) and the first inequality in (19), we can show that $\|G\|_F$ is bounded above by

$$\|G\|_F = \|PBQ\|_F \le \|PB\|\|Q\|_F \le \|P\|\|B\|\|Q\|_F \le \|B\|\|Q\|_F \le \|B\|\frac{2}{1-\tau}\tau. \tag{34}$$

Finally, we provide an upper bound for $\|H\|_F$. By leveraging the second inequality in (19) and the fact that $\|A\| \le \|A\|_F$ for any matrix $A \in \mathbb{R}^{d\times d}$, we can show that

$$\|H\|_F = \|Q^\top BQ\|_F \le \|B\|\|Q\|\|Q\|_F \le \|B\|\|Q\|_F^2 \le \|B\|\frac{4\tau^2}{(1-\tau)^2}, \tag{35}$$

where for the last inequality we used the result in (32).

If we replace $\|D\|_F$, $\|E\|_F$, $\|G\|_F$, and $\|H\|_F$ with their upper bounds in (31), (33), (34) and (35), respectively, we obtain that

$$\begin{aligned}\|B_+ - I\|_F &\le \|D\|_F + \|E\|_F + 2\|G\|_F + \|H\|_F \\ &\le \|B-I\|_F - \frac{\|(B-I)s\|^2}{2\delta\|s\|^2} + \frac{3+\tau}{1-\tau}\tau + \|B\|\frac{4}{1-\tau}\tau + \|B\|\frac{4\tau}{(1-\tau)^2}\tau \\ &= \|B-I\|_F - \frac{\|(B-I)s\|^2}{2\delta\|s\|^2} + W\tau,\end{aligned}$$

where $W = \|B\|\frac{4}{1-\tau} + \|B\|\frac{4\tau}{(1-\tau)^2} + \frac{3+\tau}{1-\tau} = \|B\|\frac{4}{(1-\tau)^2} + \frac{3+\tau}{1-\tau}$. Considering the notations introduced in (27), the result in (25) follows from the above inequality and the proof is complete. □

The result in Lemma 5 shows how the error of Hessian approximation in DFP evolves as we run the updates. Next, we establish a similar result for the BFGS method.

Lemma 6 *Consider the update of BFGS in (6) and recall the definition of τ_k in (16). Suppose that for some $\delta > 0$ and some $k \geq 0$, we have that $\tau_k < 1$ and $\|\hat{B}_k - I\|_F \leq \delta$. Then, the matrix B_{k+1}^{BFGS} generated by the BFGS update satisfies the following inequality*

$$\|\hat{B}_{k+1}^{BFGS} - I\|_F \leq \|\hat{B}_k - I\|_F - \frac{\hat{s}_k^\top (\hat{B}_k - I)\hat{B}_k(\hat{B}_k - I)\hat{s}_k}{2\delta \hat{s}_k^\top \hat{B}_k \hat{s}_k} + V_k \tau_k, \tag{36}$$

where $V_k = \frac{3+\tau_k}{1-\tau_k}$.

Proof The proof of this lemma is adapted from the proof of Lemma 3.6 in [32]. We should also add that our upper bound in (36) improves the bound in [32] as it contains an additional negative term, i.e., $-\frac{\hat{s}_k^\top (\hat{B}_k - I)\hat{B}_k(\hat{B}_k - I)\hat{s}_k}{2\delta \hat{s}_k^\top \hat{B}_k \hat{s}_k}$. Recall the BFGS update in (6) and multiply both sides of that expression with $\nabla^2 f(x_*)^{-\frac{1}{2}}$ from left and right to obtain

$$\hat{B}_{k+1}^{\text{BFGS}} = \hat{B}_k - \frac{\hat{B}_k \hat{s}_k \hat{s}_k^\top \hat{B}_k}{\hat{s}_k^\top \hat{B}_k \hat{s}_k} + \frac{\hat{y}_k \hat{y}_k^\top}{\hat{s}_k^\top \hat{y}_k}, \tag{37}$$

where we used the fact that $s_k^\top B_k s_k = s_k^\top \nabla^2 f(x_*)^{\frac{1}{2}} \nabla^2 f(x_*)^{-\frac{1}{2}} B_k \nabla^2 f(x_*)^{-\frac{1}{2}} \nabla^2 f(x_*)^{\frac{1}{2}} s_k = \hat{s}_k^\top \hat{B}_k \hat{s}_k$. To simplify the proof, we use the following notations:

$$B = \hat{B}_k, \quad B_+ = \hat{B}_{k+1}^{\text{BFGS}}, \quad s = \hat{s}_k, \quad y = \hat{y}_k, \quad \tau = \tau_k. \tag{38}$$

Considering these notations, the expression in (37) can be written as

$$B_+ = B - \frac{Bss^\top B}{s^\top Bs} + \frac{yy^\top}{s^\top y}.$$

Moreover, we can show that $B_+ - I$ is given by

$$B_+ - I = B - I - \frac{Bss^\top B}{s^\top Bs} + \frac{ss^\top}{\|s\|^2} + \frac{yy^\top}{s^\top y} - \frac{ss^\top}{\|s\|^2} = D + E,$$

where

$$D = B - I - \frac{Bss^\top B}{s^\top Bs} + \frac{ss^\top}{\|s\|^2}, \qquad E = \frac{yy^\top}{s^\top y} - \frac{ss^\top}{\|s\|^2}.$$

To establish an upper bound on $\|B_+ - I\|_F$, we find upper bounds on $\|D\|_F^2$ and $\|E\|_F^2$. Note that using the fact that $\|D\|_F^2 = \mathrm{Tr}\left[DD^\top\right]$ and properties of the trace operator we can show that

$$\begin{aligned}
\|D\|_F^2 &= \mathrm{Tr}\left[\left(B - I - \frac{Bss^\top B}{s^\top Bs} + \frac{ss^\top}{\|s\|^2}\right)\left(B - I - \frac{Bss^\top B}{s^\top Bs} + \frac{ss^\top}{\|s\|^2}\right)^\top\right] \\
&= \mathrm{Tr}\left[(B-I)^2 - \frac{Bss^\top B(B-I) + (B-I)Bss^\top B}{s^\top Bs} - \frac{ss^\top Bss^\top B + Bss^\top Bss^\top}{s^\top Bs\|s\|^2}\right] \\
&\quad + \mathrm{Tr}\left[\frac{ss^\top(B-I) + (B-I)ss^\top}{\|s\|^2} + \frac{Bss^\top BBss^\top B}{(s^\top Bs)^2} + \frac{ss^\top ss^\top}{\|s\|^4}\right]. \qquad (39) \\
&= \mathrm{Tr}\left[(B-I)^2 - \frac{Bss^\top B(B-I) + (B-I)Bss^\top B}{s^\top Bs} - \frac{ss^\top B + Bss^\top}{\|s\|^2}\right] \\
&\quad + \mathrm{Tr}\left[\frac{ss^\top(B-I) + (B-I)ss^\top}{\|s\|^2} + \frac{\|Bs\|^2 Bss^\top B}{(s^\top Bs)^2} + \frac{ss^\top}{\|s\|^2}\right].
\end{aligned}$$

Using the fact that $\mathrm{Tr}\left(ab^\top\right) = a^\top b$ for any $a, b \in \mathbb{R}^d$ we can write the following simplifications:

$$\begin{aligned}
&\mathrm{Tr}\left[\frac{Bss^\top B(B-I) + (B-I)Bss^\top B}{s^\top Bs}\right] = 2\frac{s^\top B(B-I)Bs}{s^\top Bs}, \\
&\mathrm{Tr}\left[(B-I)^2\right] = \|B-I\|_F^2, \qquad \mathrm{Tr}\left[\frac{Bss^\top + ss^\top B}{\|s\|^2}\right] = 2\frac{s^\top Bs}{\|s\|^2}, \qquad \mathrm{Tr}\left[\frac{ss^\top}{\|s\|^2}\right] = 1, \\
&\mathrm{Tr}\left[\frac{ss^\top(B-I) + (B-I)ss^\top}{\|s\|^2}\right] = 2\frac{s^\top(B-I)s}{\|s\|^2}, \qquad \mathrm{Tr}\left[\frac{\|Bs\|^2 Bss^\top B}{(s^\top Bs)^2}\right] = \frac{\|Bs\|^4}{(s^\top Bs)^2}.
\end{aligned}$$

Substituting the above simplifications into (39), we obtain that

$$\begin{aligned}
\|D\|_F^2 &= \|B-I\|_F^2 - 2\frac{s^\top B(B-I)Bs}{s^\top Bs} - 2\frac{s^\top Bs}{\|s\|^2} + 2\frac{s^\top(B-I)s}{\|s\|^2} + \frac{\|Bs\|^4}{(s^\top Bs)^2} + 1 \\
&= \|B-I\|_F^2 + \left[\left(\frac{\|Bs\|^2}{s^\top Bs}\right)^2 - \frac{s^\top B^3 s}{s^\top Bs}\right] - \frac{s^\top(B-I)B(B-I)s}{s^\top Bs}. \qquad (40)
\end{aligned}$$

Next, we proceed to show that the second term on the right hand side of (40), i.e., $\left(\frac{\|Bs\|^2}{s^\top Bs}\right)^2 - \frac{s^\top B^3 s}{s^\top Bs}$, is non-positive. Note that by using the Cauchy–Schwarz inequality, we have

$$\|Bs\|^2 = s^\top B^2 s = s^\top B^{\frac{3}{2}} B^{\frac{1}{2}} s \le \|B^{\frac{3}{2}} s\| \|B^{\frac{1}{2}} s\|.$$

Now by computing the squared of both sides we obtain $\|Bs\|^4 \le \|B^{\frac{3}{2}} s\|^2 \|B^{\frac{1}{2}} s\|^2 = s^\top B^3 s s^\top Bs$, which implies that

$$\left(\frac{\|Bs\|^2}{s^\top Bs}\right)^2 - \frac{s^\top B^3 s}{s^\top Bs} \le 0. \qquad (41)$$

By combining (40) and (41), we obtain that

$$\frac{s^\top (B-I)B(B-I)s}{s^\top Bs} \leq \|B-I\|_F^2 - \|D\|_F^2. \tag{42}$$

The above inequality implies that $\|B-I\|_F^2 - \|D\|_F^2 \geq 0$. Moreover, using the fact that $a^2 - b^2 \leq 2a(a-b)$, $\forall a, b \in \mathbb{R}$, we can show that

$$\begin{aligned}\|B-I\|_F^2 - \|D\|_F^2 &\leq 2\|B-I\|_F(\|B-I\|_F - \|D_k\|_F)\\ &\leq 2\delta(\|B-I\|_F - \|D\|_F).\end{aligned} \tag{43}$$

where the second inequality follows from $\|B-I\|_F^2 - \|D\|_F^2 \geq 0$ and the fact that $\|B-I\|_F \leq \delta$. Now if combine the results in (42) and (43), we obtain that

$$\|D\|_F \leq \|B-I\|_F - \frac{s^\top (B-I)B(B-I)s}{2\delta s^\top Bs}, \tag{44}$$

which provides an upper bound on $\|D\|_F$. Moreover, according to (33), $\|E\|_F$ is bounded above by

$$\|E\|_F \leq \frac{3+\tau}{1-\tau}\tau. \tag{45}$$

If we replace $\|D\|_F$ and $\|E\|_F$ with their upper bounds in (44) and (45), we obtain that

$$\|B_+ - I\|_F \leq \|D\|_F + \|E\|_F \leq \|B-I\|_F - \frac{s^\top (B-I)B(B-I)s}{2\delta s^\top Bs} + V\tau,$$

where $V = \frac{3+\tau}{1-\tau}$. Considering the notations in (38), the claim follows from the above inequality. □

Now we can combine Lemma 5 and Lemma 6 to derive a bound on the error of Hessian approximation for the (convex) Broyden class of quasi-Newton methods.

Lemma 7 *Consider the update of the (convex) Broyden family in (7) and recall the definition of τ_k in (16). Suppose that for some $\delta > 0$ and some $k \geq 0$, we have that $\tau_k < 1$ and $\|\hat{B}_k - I\|_F \leq \delta$. Then, the matrix B_{k+1} generated by (7) satisfies the following inequality*

$$\begin{aligned}\|\hat{B}_{k+1} - I\|_F \leq \|\hat{B}_k - I\|_F &- \phi_k \frac{\|(\hat{B}_k - I)\hat{s}_k\|^2}{2\delta\|\hat{s}_k\|^2}\\ &-(1-\phi_k)\frac{\hat{s}_k^\top (\hat{B}_k - I)\hat{B}_k(\hat{B}_k - I)\hat{s}_k}{2\delta \hat{s}_k^\top \hat{B}_k \hat{s}_k} + Z_k\tau_k,\end{aligned} \tag{46}$$

where $Z_k = \phi_k \|\hat{B}_k\| \frac{4}{(1-\tau_k)^2} + \frac{3+\tau_k}{1-\tau_k}$. *We also have that*

$$\|\hat{B}_{k+1} - I\|_F \leq \|\hat{B}_k - I\|_F + Z_k \tau_k. \tag{47}$$

Proof Notice that $B_{k+1} = \phi_k B_{k+1}^{\text{DFP}} + (1-\phi_k) B_{k+1}^{\text{BFGS}}$. Using this expression and the convexity of the norm, we can show that

$$\begin{aligned}\|\hat{B}_{k+1} - I\|_F = \|\phi_k B_{k+1}^{\text{DFP}} + (1-\phi_k) B_{k+1}^{\text{BFGS}} - I\|_F \leq \phi_k \|B_{k+1}^{\text{DFP}} \\ - I\|_F + (1-\phi_k)\|B_{k+1}^{\text{BFGS}} - I\|_F.\end{aligned}$$

By replacing $\|B_{k+1}^{\text{DFP}} - I\|_F$ and $\|B_{k+1}^{\text{BFGS}} - I\|_F$ with their upper bounds in Lemma 5 and Lemma 6, the claim in (46) follows. Moreover, since $\phi_k \in [0,1]$, $\delta > 0$, $\frac{\|(\hat{B}_k - I)\hat{s}_k\|^2}{\|\hat{s}_k\|^2} \geq 0$ and $\frac{\hat{s}_k^\top (\hat{B}_k - I)\hat{B}_k(\hat{B}_k - I)\hat{s}_k}{\hat{s}_k^\top \hat{B}_k \hat{s}_k} \geq 0$, the result in (46) implies (47). □

4.2 Linear convergence

In this section, we leverage the results from the previous section on the error of Hessian approximation to show that if the initial iterate is sufficiently close to the optimal solution and the initial Hessian approximation matrix is close to the Hessian at the optimal solution, the iterates of BFGS and DFP converge at least linearly to the optimal solution. Moreover, the Hessian approximation matrices always stay close to the Hessian at the optimal solution and the norms of Hessian approximation matrix and its inverse are always bounded above. These results are essential in proving our non-asymptotic superlinear convergence results.

Lemma 8 *Consider the convex Broyden class of quasi-Newton methods described in Algorithm 1, and recall the definitions in (13–16). Suppose Assumptions 3.1 and 3.3 hold. Moreover, suppose the initial point* x_0 *and initial Hessian approximation matrix* B_0 *satisfy*

$$\sigma_0 \leq \epsilon, \qquad \|\hat{B}_0 - I\|_F \leq \delta, \tag{48}$$

where $\epsilon, \delta \in (0, \frac{1}{2})$ *such that for some* $\rho \in (0,1)$, *they satisfy*

$$\begin{aligned}&\left(\phi_{max}\frac{4(2\delta+1)}{(1-\epsilon)^2} + \frac{3+\epsilon}{1-\epsilon}\right)\frac{\epsilon}{1-\rho} \leq \delta, \\ &\frac{\epsilon}{2} + 2\delta \leq (1-2\delta)\rho, \qquad \phi_{max} = \sup_{k\geq 0} \phi_k\,.\end{aligned} \tag{49}$$

Then, the sequence of iterates $\{x_k\}_{k=0}^{+\infty}$ *converges to the optimal solution* x_* *with*

$$\sigma_{k+1} \leq \rho \sigma_k, \qquad \forall k \geq 0. \tag{50}$$

Furthermore, the matrices $\{B_k\}_{k=0}^{+\infty}$ *stay in a neighborhood of* $\nabla^2 f(x_*)$ *defined as*

$$\|\hat{B}_k - I\|_F \leq 2\delta, \qquad \forall k \geq 0. \tag{51}$$

Moreover, the norms $\{\|\hat{B}_k\|\}_{k=0}^{+\infty}$ *and* $\{\|\hat{B}_k^{-1}\|\}_{k=0}^{+\infty}$ *are all uniformly bounded above by*

$$\|\hat{B}_k\| \leq 1 + 2\delta, \qquad \|\hat{B}_k^{-1}\| \leq \frac{1}{1-2\delta}, \qquad \forall k \geq 0. \tag{52}$$

Proof The proof of this lemma is adapted from the proof of Theorem 3.1 in [33]. In [33], the authors prove the results for the modified DFP method, while we consider the more general class of Broyden methods. We will use induction to prove (50), (51) and (52). First consider the base case of $k = 0$. By the initial condition (48), it's obvious that (51) holds for $k = 0$. From (51) we know that all the eigenvalues of $\hat{B}_0$ are in the interval $[1 - 2\delta, 1 + 2\delta]$. Suppose that $\lambda_{max}(\hat{B}_0)$ is the largest eigenvalue of $\hat{B}_0$ and $\lambda_{min}(\hat{B}_0)$ is the smallest eigenvalue of $\hat{B}_0$, we have

$$\|\hat{B}_0\| = \lambda_{max}(\hat{B}_0) \leq 1 + 2\delta, \quad \|\hat{B}_0^{-1}\| = \frac{1}{\lambda_{min}(\hat{B}_0)} \leq \frac{1}{1-2\delta}.$$

Hence, (52) holds for $k = 0$. Based on Assumptions 3.1 and 3.3 and the definitions in (13–16), we have

$$\begin{aligned}
\sigma_1 &= \frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_1 - x_*)\| = \frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - B_0^{-1}\nabla f(x_0) - x_*)\| \\
&= \frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}} B_0^{-1}[\nabla f(x_0) - \nabla^2 f(x_*)(x_0 - x_*) - (B_0 - \nabla^2 f(x_*))(x_0 - x_*)]\| \\
&= \frac{M}{\mu^{\frac{3}{2}}}\|\hat{B}_0^{-1}[\widehat{\nabla f}(x_0) - r_0 - (\hat{B}_0 - I)r_0]\| \\
&\leq \frac{M}{\mu^{\frac{3}{2}}}\|\hat{B}_0^{-1}\|\left(\|\widehat{\nabla f}(x_0) - r_0\| + \|\hat{B}_0 - I\|\|r_0\|\right).
\end{aligned} \tag{53}$$

Now using the result in (24), and the bounds in (48), (49), (51) and (52) for $k = 0$, we can write

$$\begin{aligned}
\sigma_1 &\leq \frac{M}{\mu^{\frac{3}{2}}}\|\hat{B}_0^{-1}\|(\frac{\sigma_0}{2}\|r_0\| + \|\hat{B}_0 - I\|\|r_0\|) = \|\hat{B}_0^{-1}\|(\frac{\sigma_0}{2} + \|\hat{B}_0 - I\|)\sigma_0 \\
&\leq \frac{1}{1-2\delta}(\frac{\epsilon}{2} + 2\delta)\sigma_0 \leq \rho\sigma_0.
\end{aligned}$$

This indicates that the condition in (50) holds for $k = 0$. Hence, all the conditions in (50), (51) and (52) hold for $k = 0$, and the base of induction is complete. Now we assume that the conditions in (50), (51) and (52) hold for all $0 \leq k \leq t$, where $t \geq 0$. Our goal is to show that these conditions are also satisfied for the case of $k = t + 1$. Since (50) holds for all $0 \leq k \leq t$, we have $\tau_k = \max\{\sigma_k, \sigma_{k+1}\} = \sigma_k \leq \epsilon < 1$ for

$0 \leq k \leq t$. Moreover, since the condition in (51) holds for $0 \leq k \leq t$, we know that $\|\hat{B}_k - I\|_F \leq 2\delta$ for $0 \leq k \leq t$. Hence, by (47) in Lemma 7, we obtain that

$$\|\hat{B}_{k+1} - I\|_F \leq \|\hat{B}_k - I\|_F + Z_k\sigma_k, \qquad 0 \leq k \leq t, \tag{54}$$

where $Z_k = \phi_k\|\hat{B}_k\|\frac{4}{(1-\sigma_k)^2} + \frac{3+\sigma_k}{1-\sigma_k}$. Using (52) and $\sigma_k \leq \epsilon$ for $0 \leq k \leq t$, we obtain that

$$Z_k \leq \phi_k \frac{4(2\delta+1)}{(1-\epsilon)^2} + \frac{3+\epsilon}{1-\epsilon}, \qquad 0 \leq k \leq t.$$

Further if (48) and (50) hold for $0 \leq k \leq t$, we have that

$$\sum_{k=0}^{t} \sigma_k \leq \sum_{k=0}^{t} \rho^k \sigma_0 \leq \frac{\sigma_0}{1-\rho} \leq \frac{\epsilon}{1-\rho}. \tag{55}$$

Considering these results we can show that

$$\begin{aligned} \sum_{k=0}^{t} Z_k\sigma_k &\leq \sup_{k\geq 0}\left[\phi_k \frac{4(2\delta+1)}{(1-\epsilon)^2} + \frac{3+\epsilon}{1-\epsilon}\right] \sum_{k=0}^{t} \sigma_k \\ &\leq \left(\phi_{\max} \frac{4(2\delta+1)}{(1-\epsilon)^2} + \frac{3+\epsilon}{1-\epsilon}\right) \frac{\epsilon}{1-\rho} \leq \delta, \end{aligned} \tag{56}$$

where we use the definition $\phi_{\max} = \sup_{k\geq 0} \phi_k$ and the last inequality holds due to the first inequality in (49). By leveraging (56) and (48) and computing the sum of the terms in the left and right hand side of (54) from $k = 0$ to t, we obtain

$$\|\hat{B}_{t+1} - I\|_F \leq \|\hat{B}_0 - I\|_F + \sum_{k=0}^{t} Z_k\sigma_k \leq \delta + \delta = 2\delta,$$

which implies that (51) holds for $k = t + 1$. Applying the same techniques we used in the base case, we can prove that (50) and (52) hold for $k = t + 1$. Hence, all the claims in (50), (51) and (52) hold for $k = t + 1$, and our induction step is complete. □

4.3 Explicit non-asymptotic superlinear rate

In the previous section, we established local linear convergence of iterates generated by the convex Broyden class including DFP and BFGS. Indeed, these local linear results are not our ultimate goal, as first-order methods are also linearly convergent under the same assumptions. However, the linear convergence is required to establish a local non-asymptotic superlinear convergence result, which is our main contribution. Next, we state the main results of this paper on the non-asymptotic superlinear convergence rate of the convex Broyden class of quasi-Newton methods. To prove this claim, we use the results in Lemma 7 and Lemma 8.

Theorem 1 *Consider the convex Broyden class of quasi-Newton methods described in Algorithm 1. Suppose the objective function f satisfies the conditions in Assumptions 3.1 and 3.3. Moreover, suppose the initial point x_0 and initial Hessian approximation matrix B_0 satisfy*

$$\frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \epsilon,$$
$$\|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F \leq \delta, \tag{57}$$

where $\epsilon, \delta \in (0, \frac{1}{2})$ such that for some $\rho \in (0, 1)$, they satisfy

$$\left(\phi_{max}(2\delta + 1)\frac{4}{(1-\epsilon)^2} + \frac{3+\epsilon}{1-\epsilon}\right)\frac{\epsilon}{1-\rho} \leq \delta,$$
$$\frac{\epsilon}{2} + 2\delta \leq (1 - 2\delta)\rho, \qquad \phi_{max} = \sup_{k\geq 0}\phi_k\,. \tag{58}$$

Then the iterates $\{x_k\}_{k=0}^{+\infty}$ generated by the convex Broyden class of quasi-Newton methods converge to x_* at a superlinear rate of

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{C_1 q\sqrt{k} + C_2}{k}\right)^k, \qquad \forall k \geq 1, \tag{59}$$

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq (1+\epsilon)^2\left(\frac{C_1 q\sqrt{k} + C_2}{k}\right)^{2k}, \qquad \forall k \geq 1, \tag{60}$$

$$C_1 = 2\sqrt{2}\delta(1+\rho)\left(1 + \frac{\epsilon}{2}\right), \qquad C_2 = \frac{(1+\rho)(1+\frac{\epsilon}{2})\epsilon}{2(1-\rho)}, \tag{61}$$

where $q = \frac{1}{\sqrt{\phi_{min}\frac{4\delta}{1+2\delta} + \frac{1-2\delta}{1+2\delta}}} \in \left[1, \sqrt{\frac{1+2\delta}{1-2\delta}}\right]$ and $\phi_{min} = \inf_{k\geq 0}\phi_k$.

Proof When both conditions (57) and (58) hold, by Lemma 8, the results in (50), (51) and (52) hold. This indicates that for any $t \geq 0$, we have

$$\tau_t = \max\{\sigma_t, \sigma_{t+1}\} = \sigma_t \leq \sigma_0 \leq \epsilon < 1, \qquad \|\hat{B}_t - I\|_F \leq 2\delta.$$

Hence, using Lemma 7 for any $t \geq 0$, we can show that

$$\|\hat{B}_{t+1} - I\|_F \leq \|\hat{B}_t - I\|_F - \phi_t\frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{4\delta\|\hat{s}_t\|^2}$$
$$-(1-\phi_t)\frac{\hat{s}_t^\top(\hat{B}_t - I)\hat{B}_t(\hat{B}_t - I)\hat{s}_t}{4\delta\hat{s}_t^\top\hat{B}_t\hat{s}_t} + Z_t\sigma_t, \tag{62}$$

where $Z_t = \phi_t \|\hat{B}_t\| \frac{4}{(1-\sigma_t)^2} + \frac{3+\sigma_t}{1-\sigma_t}$. Using (55) and (56), for $k \geq 0$ we have

$$\sum_{t=0}^{k} \sigma_t \leq \frac{\epsilon}{1-\rho}, \qquad \sum_{t=0}^{k} Z_t \sigma_t \leq \delta. \tag{63}$$

Now compute the sum of both sides of (62) from $t = 0$ to $k-1$ to obtain

$$\|\hat{B}_k - I\|_F \leq \|\hat{B}_0 - I\|_F - \sum_{t=0}^{k-1} \left[\phi_t \frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{4\delta \|\hat{s}_t\|^2} + (1-\phi_t) \frac{\hat{s}_t^\top (\hat{B}_t - I)\hat{B}_t (\hat{B}_t - I)\hat{s}_t}{4\delta \hat{s}_t^\top \hat{B}_t \hat{s}_t} \right] + \sum_{t=0}^{k-1} Z_t \sigma_t.$$

Regroup the terms and use the results in (57) and (63) to show that

$$\sum_{t=0}^{k-1} \left[\phi_t \frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{4\delta \|\hat{s}_t\|^2} + (1-\phi_t) \frac{\hat{s}_t^\top (\hat{B}_t - I)\hat{B}_t (\hat{B}_t - I)\hat{s}_t}{4\delta \hat{s}_t^\top \hat{B}_t \hat{s}_t} \right]$$
$$\leq \|\hat{B}_0 - I\|_F - \|\hat{B}_k - I\|_F + \sum_{t=0}^{k-1} Z_t \sigma_t \leq \|\hat{B}_0 - I\|_F + \sum_{t=0}^{k-1} Z_t \sigma_t \leq \delta + \delta = 2\delta,$$

which leads to

$$\sum_{t=0}^{k-1} \left[\phi_t \frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{\|\hat{s}_t\|^2} + (1-\phi_t) \frac{\hat{s}_t^\top (\hat{B}_t - I)\hat{B}_t (\hat{B}_t - I)\hat{s}_t}{\hat{s}_t^\top \hat{B}_t \hat{s}_t} \right] \leq 8\delta^2. \tag{64}$$

Moreover, using the bounds in (52) we can show that

$$\hat{s}_t^\top (\hat{B}_t - I)\hat{B}_t (\hat{B}_t - I)\hat{s}_t \geq \frac{1}{\|\hat{B}_t^{-1}\|} \|(\hat{B}_t - I)\hat{s}_t\|^2 \geq (1-2\delta)\|(\hat{B}_t - I)\hat{s}_t\|^2,$$
$$\hat{s}_t^\top \hat{B}_t \hat{s}_t \leq \|\hat{B}_t\| \|\hat{s}_t\|^2 \leq (1+2\delta)\|\hat{s}_t\|^2.$$

Hence, we have

$$\frac{\hat{s}_t^\top (\hat{B}_t - I)\hat{B}_t (\hat{B}_t - I)\hat{s}_t}{\hat{s}_t^\top \hat{B}_t \hat{s}_t} \geq \frac{1-2\delta}{1+2\delta} \frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{\|\hat{s}_t\|^2}. \tag{65}$$

By combining the bounds in (64) and (65), we obtain

$$\sum_{t=0}^{k-1} \left[\phi_t + (1-\phi_t) \frac{1-2\delta}{1+2\delta} \right] \frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{\|\hat{s}_t\|^2} \leq 8\delta^2.$$

Now by computing the minimum value of the term $\phi_t + (1-\phi_t)\frac{1-2\delta}{1+2\delta}$, we can show

$$\inf_{k\geq 0}\left[\phi_k + (1-\phi_k)\frac{1-2\delta}{1+2\delta}\right]\sum_{t=0}^{k-1}\frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{\|\hat{s}_t\|^2} \leq 8\delta^2,$$
$$\left(\phi_{\min}\frac{4\delta}{1+2\delta} + \frac{1-2\delta}{1+2\delta}\right)\sum_{t=0}^{k-1}\frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{\|\hat{s}_t\|^2} \leq 8\delta^2,$$

where $\phi_{\min} = \inf_{k\geq 0}\phi_k$ and by regrouping the terms, we obtain that

$$\sum_{t=0}^{k-1}\frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{\|\hat{s}_t\|^2} \leq \frac{8\delta^2}{\phi_{\min}\frac{4\delta}{1+2\delta} + \frac{1-2\delta}{1+2\delta}}.$$

Considering the definition $q := \frac{1}{\sqrt{\phi_{\min}\frac{4\delta}{1+2\delta} + \frac{1-2\delta}{1+2\delta}}}$, we can simplify our upper bound as

$$\sum_{t=0}^{k-1}\frac{\|(\hat{B}_t - I)\hat{s}_t\|^2}{\|\hat{s}_t\|^2} \leq 8\delta^2 q^2.$$

By using the Cauchy-Schwarz inequality, we obtain that

$$\sum_{t=0}^{k-1}\frac{\|(\hat{B}_t - I)\hat{s}_t\|}{\|\hat{s}_t\|} \leq 2\sqrt{2}\delta q\sqrt{k}. \tag{66}$$

Note that since $\phi_k \in [0,1]$, we have $q \in \left[1, \sqrt{\frac{1+2\delta}{1-2\delta}}\right]$. The result in (66) provides an upper bound on $\sum_{t=0}^{k-1}\frac{\|(\hat{B}_t - I)\hat{s}_t\|}{\|\hat{s}_t\|}$, which is a crucial term in the remaining of our proof.

Now, note that $\nabla f(x_t) = J_t(x_t - x_*)$, where J_t is defined in (17). This implies that $x_t - x_* = J_t^{-1}\nabla f(x_t)$ and hence we have

$$x_{t+1} - x_* = x_t - x_* + s_t = J_t^{-1}\nabla f(x_t) + s_t = -J_t^{-1}B_t s_t + s_t = J_t^{-1}(J_t - B_t)s_t.$$

where the third equality holds since $-B_t s_t = \nabla f(x_t)$. Pre-multiply both sides of the above expression by $\nabla^2 f(x_*)^{\frac{1}{2}}$ to obtain

$$r_{t+1} = \hat{J}_t^{-1}(\hat{J}_t - \hat{B}_t)\hat{s}_t = \hat{J}_t^{-1}[(\hat{J}_t - I)\hat{s}_t - (\hat{B}_t - I)\hat{s}_t].$$

Therefore, we obtain that

$$\|r_{t+1}\| \leq \|\hat{J}_t^{-1}\|\left(\|(\hat{J}_t - I)\hat{s}_t\| + \|(\hat{B}_t - I)\hat{s}_t\|\right) \leq \|\hat{J}_t^{-1}\|\left(\|\hat{J}_t - I\| + \frac{\|(\hat{B}_t - I)\hat{s}_t\|}{\|\hat{s}_t\|}\right)\|\hat{s}_t\|.$$

From Lemma 3 we know that $\|\hat{J}_t^{-1}\| \leq 1 + \frac{\sigma_t}{2}$ and $\|\hat{J}_t - I\| \leq \frac{\sigma_t}{2}$. Therefore, we have

$$\|r_{t+1}\| \leq \left(1 + \frac{\sigma_t}{2}\right)\left(\frac{\sigma_t}{2} + \frac{\|(\hat{B}_t - I)\hat{s}_t\|}{\|\hat{s}_t\|}\right)\|\hat{s}_t\|. \tag{67}$$

Also, since $\sigma_{t+1} \leq \rho\sigma_t$ and $\sigma_t = \frac{M}{\mu^{\frac{3}{2}}}\|r_t\|$, we obtain that $\|r_{t+1}\| \leq \rho\|r_t\|$. Hence, we can write

$$\|\hat{s}_t\| = \|\nabla f(x_*)^{\frac{1}{2}}(x_{t+1} - x_* + x_* - x_t)\| \leq \|r_{t+1}\| + \|r_t\| \leq (1+\rho)\|r_t\|. \tag{68}$$

Using the expressions in (67) and (68), we can show that $\frac{\|r_{t+1}\|}{\|r_t\|}$ is bounded above by

$$\frac{\|r_{t+1}\|}{\|r_t\|} \leq (1+\rho)\left(1 + \frac{\sigma_t}{2}\right)\left(\frac{\sigma_t}{2} + \frac{\|(\hat{B}_t - I)\hat{s}_t\|}{\|\hat{s}_t\|}\right). \tag{69}$$

Compute the sum of both sides of (69) from $t = 0$ to $k-1$ and use $\sigma_t \leq \epsilon$, (63), and (66) to obtain

$$\begin{aligned}\sum_{t=0}^{k-1}\frac{\|r_{t+1}\|}{\|r_t\|} &\leq (1+\rho)\left(1+\frac{\epsilon}{2}\right)\left(\sum_{t=0}^{k-1}\frac{\sigma_t}{2} + \sum_{t=0}^{k-1}\frac{\|(\hat{B}_t - I)\hat{s}_t\|}{\|\hat{s}_t\|}\right)\\ &\leq (1+\rho)\left(1+\frac{\epsilon}{2}\right)\left(\frac{\epsilon}{2(1-\rho)} + 2\sqrt{2}\delta q\sqrt{k}\right).\end{aligned}$$

By leveraging the arithmetic-geometric inequality, we obtain that

$$\begin{aligned}\frac{\|r_k\|}{\|r_0\|} &= \prod_{t=0}^{k-1}\frac{\|r_{t+1}\|}{\|r_t\|}\\ &\leq \left(\frac{\sum_{t=0}^{k-1}\frac{\|r_{t+1}\|}{\|r_t\|}}{k}\right)^k \leq \left(\frac{2\sqrt{2}\delta(1+\rho)(1+\frac{\epsilon}{2})q\sqrt{k} + \frac{(1+\rho)(1+\frac{\epsilon}{2})\epsilon}{2(1-\rho)}}{k}\right)^k.\end{aligned} \tag{70}$$

Using condition (61), the proof of (59) is complete. Next, we proceed to prove (60). Based on the Taylor's Theorem, there exists $\alpha_t \in [0,1]$ and the matrix $H_t = \nabla^2 f(x_* + \alpha_t(x_t - x_*))$ such that

$$f(x_t) - f(x_*) = \nabla f(x_*)(x_t - x_*) + \frac{1}{2}(x_t - x_*)^\top H_t(x_t - x_*) = \frac{1}{2}r_t^\top \hat{H}_t r_t,$$

where we used $\nabla f(x_*) = 0$ and $\hat{H}_t = \nabla^2 f(x_*)^{-\frac{1}{2}} H_t \nabla^2 f(x_*)^{-\frac{1}{2}}$. By Lemma 3 and $\sigma_t \leq \epsilon$, we have

$$f(x_0) - f(x_*) = \frac{1}{2} r_0^\top \hat{H}_0 r_0 \geq \frac{1}{2(1+\sigma_0)} \|r_0\|^2 \geq \frac{1}{2(1+\epsilon)} \|r_0\|^2, \tag{71}$$

and

$$f(x_k) - f(x_*) = \frac{1}{2} r_k^\top \hat{H}_k r_k \leq \frac{1+\sigma_k}{2} \|r_k\|^2 \leq \frac{1+\epsilon}{2} \|r_k\|^2. \tag{72}$$

By combining (70), (71) and (72), we obtain that

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq \frac{\frac{1+\epsilon}{2}\|r_k\|^2}{\frac{1}{2(1+\epsilon)}\|r_0\|^2} \leq (1+\epsilon)^2 \left(\frac{2\sqrt{2}\delta(1+\rho)(1+\frac{\epsilon}{2})q\sqrt{k} + \frac{(1+\rho)(1+\frac{\epsilon}{2})\epsilon}{2(1-\rho)}}{k} \right)^{2k},$$

and the claim in (60) holds. □

The above theorem establishes the non-asymptotic superlinear convergence of the Broyden class of quasi-Newton methods. Notice that we use the weighted norm in (59) to characterize the convergence rate. If in addition to the strong convexity condition in Assumption 3.1, we also assume that the gradient is Lipschitz continuous as in Assumption 3.2, then we have that $\sqrt{\mu}\|x_t - x_*\| \leq \|r_t\| \leq \sqrt{L}\|x_t - x_*\|, \forall t \geq 0$. Hence, the result in (59) implies that

$$\frac{\|x_k - x_*\|}{\|x_0 - x_*\|} \leq \sqrt{\frac{L}{\mu}} \left(\frac{C_1 q\sqrt{k} + C_2}{k} \right)^k, \qquad \forall k \geq 1, \tag{73}$$

where C_1 and C_2 are defined in (61). Next, we use the above theorem to report the results for DFP and BFGS, which are two special cases of the convex Broyden class of quasi-Newton methods.

Corollary 2 *Consider the DFP and BFGS methods. Suppose Assumptions 3.1 and 3.3 hold and for some $\epsilon, \delta \in (0, \frac{1}{2})$ and $\rho \in (0,1)$, the initial point x_0 and initial Hessian approximation B_0 satisfy*

$$\frac{M}{\mu^{\frac{3}{2}}} \|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \epsilon, \quad \|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F \leq \delta. \tag{74}$$

– *For the DFP method, if the tuple (ϵ, δ, ρ) satisfies*

$$\left[\frac{4(2\delta+1)}{(1-\epsilon)^2} + \frac{3+\epsilon}{1-\epsilon} \right] \frac{\epsilon}{1-\rho} \leq \delta, \qquad \frac{\epsilon}{2} + 2\delta \leq (1-2\delta)\rho\,, \tag{75}$$

then the iterates $\{x_k\}_{k=0}^{+\infty}$ *generated by the DFP method converge to* x_* *at a superlinear rate of*

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{C_1\sqrt{k} + C_2}{k}\right)^k, \quad \forall k \geq 1, \tag{76}$$

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq (1+\epsilon)^2 \left(\frac{C_1\sqrt{k} + C_2}{k}\right)^{2k}, \quad \forall k \geq 1. \tag{77}$$

– *For the BFGS method, if the tuple* (ϵ, δ, ρ) *satisfies*

$$\frac{(3+\epsilon)\epsilon}{(1-\epsilon)(1-\rho)} \leq \delta, \qquad \frac{\epsilon}{2} + 2\delta \leq (1-2\delta)\rho\,, \tag{78}$$

then the iterates $\{x_k\}_{k=0}^{+\infty}$ *generated by the BFGS method converge to* x_* *at a superlinear rate of*

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{C_1\sqrt{\frac{1+2\delta}{1-2\delta}}\sqrt{k} + C_2}{k}\right)^k, \quad \forall k \geq 1, \tag{79}$$

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq (1+\epsilon)^2 \left(\frac{C_1\sqrt{\frac{1+2\delta}{1-2\delta}}\sqrt{k} + C_2}{k}\right)^{2k}, \quad \forall k \geq 1, \tag{80}$$

where C_1 *and* C_2 *are defined in (61).*

Proof In Theorem 1, set $\phi_k = 1$ for all $k \geq 0$ to obtain the results for DFP and set $\phi_k = 0$ for all $k \geq 0$ to obtain the results for BFGS. □

The results in Corollary 2 indicate that, in a local neighborhood of the optimal solution, the iterates generated by DFP and BFGS converge to the optimal solution at a superlinear rate of $((C_1\sqrt{k} + C_2)/k)^k$, where the constants C_1 and C_2 are determined by ρ, ϵ and δ. Indeed, as time progresses, the rate behaves as $\mathcal{O}\left((1/\sqrt{k})^k\right)$. The tuple (ρ, ϵ, δ) is independent of the problem parameters (μ, L, M, d), and the only required condition for the tuple (ρ, ϵ, δ) is that they should satisfy (75) or (78). Note that the superlinear rate in (76) and (79) is faster than linear rate of first-order methods as the contraction coefficient approaches zero at a sublinear rate of $\mathcal{O}(1/\sqrt{k})$. Similarly, in terms of the function value, the superlinear rate shown in (77) and (80) behaves as $\mathcal{O}\left((1/k)^k\right)$. The result in Corollary 2 also shows the existence of a trade-off between the *rate of convergence* and the *neighborhood of superlinear convergence*. We highlight this point in the following remark.

Remark 3 There exists a trade-off between the size of the local neighborhood in which DFP or BFGS converges superlinearly and their rate of convergence. To be more precise, by choosing larger values for ϵ and δ (as long as they satisfy (75) or (78)), we

can increase the size of the region in which quasi-Newton method has a fast superlinear convergence rate, but on the other hand, it will lead to a slower superlinear convergence rate according to the bounds in (76), (77), (79) and (80). Conversely, by choosing small values for ϵ and δ, the rate of convergence becomes faster, but the local neighborhood defined in (74) becomes smaller.

The final convergence results of Corollary 2 depend on the choice of parameters (ρ, ϵ, δ), and it may not be easy to quantify the exact convergence rate at first glance. To better quantify the superlinear convergence rate of DFP and BFGS, in the following corollary, we state the results of Corollary 2 for specific choices of ρ, ϵ and δ which simplifies our expressions. Indeed, one can choose another set of values for these parameters to control the neighborhood and rate of superlinear convergence, as long as they satisfy the conditions in (75) for DFP and (78) for BFGS.

Corollary 3 *Consider the DFP and BFGS methods and suppose Assumptions 3.1 and 3.3 hold. Moreover, suppose the initial point x_0 and initial Hessian approximation matrix B_0 of DFP satisfy*

$$\frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \frac{1}{120}, \quad \|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F \leq \frac{1}{7}, \tag{81}$$

and the initial point x_0 and initial Hessian approximation matrix B_0 of BFGS satisfy

$$\frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \frac{1}{50}, \quad \|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F \leq \frac{1}{7}. \tag{82}$$

Then, the iterates $\{x_k\}_{k=0}^{+\infty}$ generated by the DFP and BFGS methods satisfy

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{1}{k}\right)^{\frac{k}{2}}, \qquad \frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq 1.1\left(\frac{1}{k}\right)^k, \qquad \forall k \geq 1. \tag{83}$$

Proof The results for DFP can be shown by setting $\rho = \frac{1}{2}$, $\epsilon = \frac{1}{120}$ and $\delta = \frac{1}{7}$ in Corollary 2. We can check that for those values, the conditions in (75) are all satisfied. Moreover, the expressions in (76) and (77) can be simplified as

$$\frac{2\sqrt{2}\delta(1+\rho)(1+\frac{\epsilon}{2})\sqrt{k} + \frac{(1+\rho)(1+\frac{\epsilon}{2})\epsilon}{2(1-\rho)}}{k} = \frac{\frac{2\sqrt{2}}{7}(1+\frac{1}{2})(1+\frac{1}{240})\sqrt{k} + \frac{(1+\frac{1}{2})(1+\frac{1}{240})\frac{1}{120}}{2(1-\frac{1}{2})}}{k} < \frac{1}{\sqrt{k}},$$

and $(1+\epsilon)^2 = (1+\frac{1}{120})^2 \leq 1.1$. So the claims in (83) follow. The results for BFGS can be shown similarly by setting $\rho = \frac{1}{2}$, $\epsilon = \frac{1}{50}$ and $\delta = \frac{1}{7}$ in (78), (79) and (80). □

The results in Corollary 3 show that for some specific choices of (ϵ, δ, ρ), the convergence rate of DFP and BFGS is $(1/k)^{k/2}$, which is asymptotically faster than any linear convergence rate of first-order methods. Moreover, we observe that the neighborhood in which this fast superlinear rate holds is slightly larger for BFGS compared to DFP, i.e., compare the first conditions in (81) and (82). This is in consistence with the fact that in practice, BFGS often outperforms DFP.

A major shortcoming of the results in Corollary 2 and Corollary 3 is that, in addition to assuming that the initial iterate x_0 is sufficiently close to the optimal solution, we also require the initial Hessian approximation error to be sufficiently small. In the following theorem, we resolve this issue by suggesting a practical choice for B_0 such that the second assumption in (81) and (82) can be satisfied under some conditions. To be more precise, we show that if $\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|$ is sufficiently small (we formally describe this condition), then by setting $B_0 = \nabla^2 f(x_0)$, the second condition in (81) and (82) for Hessian approximation is satisfied, and we can achieve the convergence rate in (83).

Theorem 2 *Consider the DFP and BFGS methods and suppose Assumptions 3.1 and 3.3 hold. Moreover, for DFP, suppose the initial point x_0 and initial Hessian approximation B_0 satisfy*

$$\frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \min\left\{\frac{1}{120}, \frac{1}{7\sqrt{d}}\right\}, \qquad B_0 = \nabla^2 f(x_0), \tag{84}$$

and for BFGS, they satisfy

$$\frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \min\left\{\frac{1}{50}, \frac{1}{7\sqrt{d}}\right\}, \qquad B_0 = \nabla^2 f(x_0). \tag{85}$$

Then, the iterates $\{x_k\}_{k=0}^{+\infty}$ generated by the DFP and BFGS methods satisfy

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{1}{k}\right)^{\frac{k}{2}}, \qquad \frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq 1.1\left(\frac{1}{k}\right)^{k}, \qquad \forall k \geq 1. \tag{86}$$

Proof First we consider the case of the DFP method. Notice that by (84), we obtain

$$\frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \frac{1}{120}.$$

Hence, the first part of (81) is satisfied. Moreover, using Assumptions 3.1 and 3.3, we have

$$\begin{aligned}
&\|\nabla^2 f(x_*)^{-\frac{1}{2}}(\nabla^2 f(x_0)-\nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F \\
&\leq \sqrt{d}\|\nabla^2 f(x_*)^{-\frac{1}{2}}(\nabla^2 f(x_0)-\nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\| \\
&\leq \sqrt{d}\|\nabla^2 f(x_*)^{-\frac{1}{2}}\|^2\|\nabla^2 f(x_0)-\nabla^2 f(x_*)\| \leq \sqrt{d}\frac{M}{\mu}\|x_0-x_*\| \\
&= \sqrt{d}\frac{M}{\mu}\|\nabla^2 f(x_*)^{-\frac{1}{2}}\nabla^2 f(x_*)^{\frac{1}{2}}(x_0-x_*)\| \\
&\leq \sqrt{d}\frac{M}{\mu}\|\nabla^2 f(x_*)^{-\frac{1}{2}}\|\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0-x_*)\| \\
&\leq \sqrt{d}\frac{M}{\mu^{\frac{3}{2}}}\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0-x_*)\| \leq \frac{1}{7}.
\end{aligned}$$

The first inequality holds as $\|A\|_F \leq \sqrt{d}\|A\|$ for any matrix $A \in \mathbb{R}^{d\times d}$, and the last inequality is due to the first part of (84). The above bound shows that the second part of the (81) is also satisfied, and by Corollary 3 the claim follows. The proof for BFGS is similar to the proof for DFP. It can be derived by following the steps of proof of DFP and exploiting the BFGS results in Corollary 3. □

According to Theorem 2, if the initial weighted error $\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0-x_*)\|$ is sufficiently small, then by setting the initial Hessian approximation B_0 as the Hessian at the initial point $\nabla^2 f(x_0)$, the iterates will converge superlinearly at a rate of $(1/k)^{k/2}$. More specifically, based on the result in (24), it suffices to have $\|\nabla^2 f(x^*)^{-\frac{1}{2}}\nabla f(x_0)\| \leq \mathcal{O}(\mu^{\frac{3}{2}}/(M\sqrt{d}))$ to ensure $\|\nabla^2 f(x^*)^{\frac{1}{2}}(x_0-x^*)\| \leq \mathcal{O}(\mu^{\frac{3}{2}}/(M\sqrt{d}))$ as stated in (84) and (85). Hence, this condition is satisfied when $\|\nabla f(x_0)\| \leq \mathcal{O}(\mu^2/(M\sqrt{d}))$. This observation implies that, in practice, we can exploit any optimization algorithm to find an initial point x_0 such that $\|\nabla f(x_0)\| \leq \mathcal{O}(\mu^2/(M\sqrt{d}))$, and once this condition is satisfied, by setting $B_0 = \nabla^2 f(x_0)$ we obtain the guaranteed superlinear convergence result. The suggested procedure requires only one evaluation of the Hessian inverse for the initial iterate, and in the rest of the algorithm, the Hessian inverse approximations are updated according to the convex Broyden update in (8).

5 Analysis of self-concordant functions

The results that we have presented so far require three assumptions: (i) the objective function is strongly convex, (ii) its gradient is Lipschitz continuous (iii) and its Hessian is Lipschitz continuous only at the optimal solution. In this section, we extend our theoretical results to a different setting where the objective function is self-concordant.

Assumption 5.1 The objective function f is standard self-concordant. In other words, it satisfies the following conditions: (i) f is closed with open domain $dom(f)$. (ii) it is three times continuously differentiable, (iii) $\nabla^2 f(x) \succ 0$ for all $x \in dom(f)$, and

(iv) the Hessian satisfies

$$\frac{d}{dt}\nabla^2 f(x+ty)\Big|_{t=0} \preceq 2\left(y^\top \nabla^2 f(x) y\right)^{\frac{1}{2}} \nabla^2 f(x), \quad \forall x \in dom(f), \quad \forall y \in \mathbb{R}^d. \tag{87}$$

Notice that the constant 2 in the above condition corresponds to standard self-concordant functions, but, in principle, there can be any arbitrary constant instead of 2. The analysis of Newton-type methods for self-concordant functions (see, e.g., [38, 39]) expands the theory of second-order algorithms beyond the classic setting considered in the previous section. This family of functions are of interest as it includes a large set of loss functions that are widely used in machine learning, such as linear functions, convex quadratic functions, and negative logarithm functions. In this section, we extend our results to this class of functions.

We should mention that the setup considered in this section is neither more general nor more strict than the setup in the previous section. For instance, the function $f(x) = -\log x$ is self-concordant and satisfies Assumption 5.1, but it does not satisfy Assumption 3.1, 3.2 or 3.3 for any $x > 0$. Conversely, the self-concordance assumption is not a necessary condition for the assumption that the Hessian is Lipschitz continuous only at the optimal solution. For instance, the objective function

$$f(x) = \begin{cases} 7x^2 + 8x + 3 & \text{if} \quad x \in (-\infty, -1) \\ x^4 + x^2 & \text{if} \quad x \in [-1, 1] \\ 7x^2 - 8x + 3 & \text{if} \quad x \in (1, +\infty) \end{cases} \tag{88}$$

satisfies the conditions in Assumptions 3.1, 3.2 and 3.3. However, it is not self-concordant, as its third derivative is not continuous.

Based on these points, the analysis in this section extends our convergence analysis of quasi-Newton methods to a new setting that is not covered by the setup in the previous section.

We should also mention that in [35–37] for the finite-time analysis of quasi-Newton methods, the authors assume that the objective function is *strongly* self-concordant which forms a subclass of self-concordant functions, formally defined in [35]. Note that a function f is strongly self-concordant when there exists a constant $K \geq 0$ such that for any $x, y, z, w \in dom(f)$, we have

$$\nabla^2 f(y) - \nabla^2 f(x) \preceq K\left((y-x)^\top \nabla^2 f(z)(y-x)\right)^{\frac{1}{2}} \nabla^2 f(w). \tag{89}$$

In addition, in [35–37] the authors require the objective function to be strongly convex and smooth. Indeed, our considered setting in this section is more general than the setup in these works as we only require the function to be self-concordant.

Note that the condition $\nabla^2 f(x) \succ 0$ guarantees that the inner product $s_k^\top y_k$ in quasi-Newton updates is always positive in all iterations, as stated in Sect. 2. Also by the definition of self-concordance, the function $f(x)$ is always strictly convex. We

start our analysis by stating the following lemma which plays an important role in our analysis for self-concordant functions.

Lemma 9 *Suppose function f satisfies Assumption 5.1 and $x, y \in dom(f)$. Further, consider the definition $G := \int_0^1 \nabla^2 f(x+\alpha(y-x))d\alpha$. If x and y are such that $r = \|\nabla^2 f(x)^{\frac{1}{2}}(y-x)\| < 1$, then*

$$(1-r)^2\nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \frac{1}{(1-r)^2}\nabla^2 f(x), \tag{90}$$

$$(1-r+\frac{r^2}{3})\nabla^2 f(x) \preceq G \preceq \frac{1}{1-r}\nabla^2 f(x). \tag{91}$$

Proof Check Theorem 4.1.6 and Corollary 4.1.4 of [9]. □

The next two lemmas are based on Lemma 9 and are similar to the results in Lemma 3 and 4, except here we prove them for the case that the conditions in Assumption 5.1 are satisfied.

Lemma 10 *Recall the definition of r_k in (16) and $\hat{J}_k$ in (17). Suppose that there exists $\alpha_k \in [0,1]$ and define the matrix $H_k = \nabla^2 f(x_* + \alpha_k(x_k - x_*))$ and $\hat{H}_k = \nabla^2 f(x_*)^{-\frac{1}{2}} H_k \nabla^2 f(x_*)^{-\frac{1}{2}}$. If Assumption 5.1 holds and $\|r_k\| \leq \frac{1}{2}$, then for all $k \geq 0$ we have*

$$\frac{1}{1+2\|r_k\|}I \preceq \hat{J}_k \preceq (1+2\|r_k\|)I, \qquad (1-\|r_k\|)^2 I \preceq \hat{H}_k \preceq \frac{1}{(1-\|r_k\|)^2}I. \tag{92}$$

Proof Check Appendix F □

Lemma 11 *Recall the definitions in (13) - (16) and consider the definition $\theta_k := \max\{\|r_k\|, \|r_{k+1}\|\}$. Suppose that for some $k \geq 0$, we have $\theta_k \leq \frac{1}{2}$. If Assumption 5.1 holds, we have*

$$\|\hat{y}_k - \hat{s}_k\| \leq 6\theta_k\|\hat{s}_k\|, \tag{93}$$

$$(1-6\theta_k)\|\hat{s}_k\|^2 \leq \hat{s}_k^\top \hat{y}_k \leq (1+6\theta_k)\|\hat{s}_k\|^2, \tag{94}$$

$$(1-6\theta_k)\|\hat{s}_k\| \leq \|\hat{y}_k\| \leq (1+6\theta_k)\|\hat{s}_k\|, \tag{95}$$

$$\|\widehat{\nabla f}(x_k) - r_k\| \leq 2\|r_k\|^2. \tag{96}$$

Proof Check Appendix G. □

By comparing Lemma 10 and Lemma 11 with Lemma 3 and Lemma 4, respectively, we observe that the only difference between these results is that we replaced $\sigma_k/2 = (M/(2\mu^{\frac{3}{2}}))\|r_k\|$ by $2\|r_k\|$ and $\tau_k = \max\{\sigma_k, \sigma_{k+1}\}$ by $6\theta_k = 6\max\{\|r_k\|, \|r_{k+1}\|\}$. Due to this similarity, our results for the self-concordant setting are very similar to the previous case considered in Sect. 4. As a result, the superlinear convergence proof in this section is also similar to the one in Sect. 4. Next, we directly present the final superlinear convergence rate results for self-concordant functions.

Theorem 3 *Consider the convex Broyden class of quasi-Newton methods described in Algorithm 1. Suppose the objective function f satisfies the conditions in Assumption 5.1. Moreover, suppose the initial point x_0 and initial Hessian approximation matrix B_0 satisfy*

$$\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \frac{\epsilon}{6}, \qquad \|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F \leq \delta, \tag{97}$$

where $\epsilon, \delta \in (0, \frac{1}{2})$ such that for some $\rho \in (0, 1)$, they satisfy

$$\left(\phi_{max}(2\delta+1)\frac{4}{(1-\epsilon)^2} + \frac{3+\epsilon}{1-\epsilon}\right)\frac{\epsilon}{1-\rho} \leq \delta, \qquad \frac{\epsilon}{3} + 2\delta \leq (1-2\delta)\rho,$$
$$\phi_{max} = \sup_{k\geq 0}\phi_k\,. \tag{98}$$

Then the iterates $\{x_k\}_{k=0}^{+\infty}$ generated by the convex Broyden class of quasi-Newton methods converge to x_* at a superlinear rate of

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{C_3 q\sqrt{k} + C_4}{k}\right)^k, \qquad \forall k \geq 1, \tag{99}$$

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq \frac{1}{(1-\frac{\epsilon}{6})^4}\left(\frac{C_3 q\sqrt{k} + C_4}{k}\right)^{2k}, \qquad \forall k \geq 1, \tag{100}$$

$$C_3 = 2\sqrt{2}\delta(1+\rho)\left(1+\frac{\epsilon}{3}\right), \qquad C_4 = \frac{(1+\rho)(1+\frac{\epsilon}{3})\epsilon}{3(1-\rho)}, \tag{101}$$

where $q = \frac{1}{\sqrt{\phi_{min}\frac{4\delta}{1+2\delta} + \frac{1-2\delta}{1+2\delta}}} \in \left[1, \sqrt{\frac{1+2\delta}{1-2\delta}}\right]$ and $\phi_{min} = \inf_{k\geq 0}\phi_k$.

Proof Check Appendix H. □

Similarly, we can set $\phi_k = 1$ or $\phi_k = 0$ for all $k \geq 0$ to obtain the results for DFP and BFGS, respectively, as stated in Corollary 2. We can also select specific values for (ϵ, δ, ρ) to simplify our bounds.

Corollary 4 *Consider the DFP and BFGS methods and suppose Assumption 5.1 holds. Moreover, suppose for the DFP method, the initial point x_0 and initial Hessian approximation matrix B_0 satisfy*

$$\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \frac{1}{720}, \qquad \|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F \leq \frac{1}{7}, \tag{102}$$

and for the BFGS method, the initial point x_0 and initial Hessian approximation matrix B_0 satisfy

$$\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \frac{1}{300}, \qquad \|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F \leq \frac{1}{7}. \tag{103}$$

Then, the iterates $\{x_k\}_{k=0}^{+\infty}$ generated by these methods satisfy

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{1}{k}\right)^{\frac{k}{2}}, \qquad \frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq 1.1\left(\frac{1}{k}\right)^k, \qquad \forall k \geq 1. \tag{104}$$

Proof As in the proof of Corollary 3, we set $\phi_k = 1$, $\rho = \frac{1}{2}$, $\epsilon = \frac{1}{120}$, $\delta = \frac{1}{7}$ for the DFP method and $\phi_k = 0$, $\rho = \frac{1}{2}$, $\epsilon = \frac{1}{50}$, $\delta = \frac{1}{7}$ for the BFGS method in Theorem 3. Then, the claims follow. □

We can also set the initial Hessian approximation matrix to be $\nabla^2 f(x_0)$ as in Theorem 2 to achieve the same superlinear convergence rate as long as the distance between the initial point x_0 and the optimal point x_* is sufficiently small.

Theorem 4 *Consider the DFP and BFGS methods and suppose Assumption 5.1 holds. Moreover, suppose for the DFP method, the initial point x_0 and initial Hessian approximation matrix B_0 satisfy*

$$\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \min\left\{\frac{1}{720}, \frac{1}{21\sqrt{d}}\right\}, \qquad B_0 = \nabla^2 f(x_0), \tag{105}$$

and for the BFGS method, they satisfy

$$\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \min\left\{\frac{1}{300}, \frac{1}{21\sqrt{d}}\right\}, \qquad B_0 = \nabla^2 f(x_0). \tag{106}$$

Then, the iterates $\{x_k\}_{k=0}^{+\infty}$ generated by these methods satisfy

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{1}{k}\right)^{\frac{k}{2}}, \qquad \frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq 1.1\left(\frac{1}{k}\right)^k, \quad \forall k \geq 1. \tag{107}$$

Proof Check Appendix I. □

In summary, we established the local convergence rate of the convex Broyden class of quasi-Newton methods for self-concordant functions. We showed that if the initial distance to the optimal solution is $\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| = \mathcal{O}(1)$ and the initial Hessian approximation error is $\|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F = \mathcal{O}(1)$, the

iterations converge to the optimal solution at a superlinear rate of $\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)^k$ and $\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)} = \mathcal{O}\left(\frac{1}{k}\right)^k$. Moreover, we can achieve the same superlinear rate if the initial error is $\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| = \mathcal{O}(\frac{1}{\sqrt{d}})$ and the initial Hessian approximation matrix is $B_0 = \nabla^2 f(x_0)$.

6 Discussion

In this section, we discuss the strengths and shortcomings of our theoretical results and compare them with concurrent papers [36, 37] on the non-asymptotic superlinear convergence of DFP and BFGS.

Initial Hessian approximation condition Note that in our main theoretical results, in addition to the fact that the initial iterate x_0 has to be close to the optimal solution x_*, which is a common condition for local convergence results, we also need the initial Hessian approximation B_0 to be close to the Hessian at the optimal solution $\nabla^2 f(x_*)$. At first glance, this might seem restrictive, but as we have shown in Theorem 2 and Theorem 4, if we set the initial Hessian approximation to the Hessian at the initial point $\nabla^2 f(x_0)$, this condition is automatically satisfied as long as the initial iterate error $\|x_0 - x_*\|$ is sufficiently small. From a complexity point of view, this approach is reasonable as quasi-Newton methods and Newton's method outperform first-order methods in a local neighborhood of the optimal solution, and their global linear convergence rate may not be faster than the linear convergence rate of first-order methods. Hence, as suggested in [2], to optimize the overall iteration complexity according to theoretical bounds, one might use first-order methods such as Nesterov's accelerated gradient method to reach a local neighborhood of the optimal solution, and then switch to locally fast methods such as quasi-Newton methods. If this procedure is used, our theoretical results show that by setting $B_0 = \nabla^2 f(x_0)$ (and equivalently $H_0 = \nabla^2 f(x_0)^{-1}$) for the convex Broyden class of quasi-Newton, the fast superlinear convergence rate of $(1/k)^{k/2}$ can be obtained.

It is worth noting that, however, in practice algorithms that do not require switching between algorithms or knowledge of problem parameters are more favorable. Due to these reasons, quasi-Newton methods with an Armijo-Wolfe line search are more practical, as they offer an adaptive choice of the steplength with global convergence and avoid specifying typically unknown constants such as the Lipschitz constant of the gradient, Lipschitz constant of the Hessian, and strong convexity parameter.

It is worth noting that, in practice, algorithms that do not switch between different methods or require knowledge of the problem parameters are more favorable. Because of these reasons, quasi-Newton methods with an Armijo-Wolfe line search are more practical as they offer an adaptive choice of the stepsize with global convergence guarantees without knowledge of unknown constants such as the Lipschitz constant of the gradient, Lipschitz constant of the Hessian, and strong convexity parameter. Indeed, both this framework and the framework in [36, 37] requires re-initializing the Hessian approximation when one is sufficiently close to the solution. An ideal theoretical

guarantee would follow a line-search approach that guarantees that, once the iterates reach a local neighborhood of the optimal solution, the Hessian approximation for DFP or BFGS automatically satisfies the required conditions for superlinear convergence without any modification to the Hessian approximation.

Convergence rate-neighborhood trade-off As mentioned earlier, we observe a trade-off between the radius of the neighborhood in which BFGS and DFP converge superlinearly to the optimal solution and the rate (speed) of superlinear convergence. One important observation here is that for specific choices of ϵ, δ and ρ, the rate of convergence could be independent of the problem dimension d, while the neighborhood of the convergence would depend on d. Note that by selecting different parameters we could improve the dependency of the neighborhood on d, at the cost of achieving a contraction factor that depends on d. In this case, the contraction factor may not be always smaller than 1, and we can only guarantee that after a few iterations it becomes smaller than 1 and eventually behaves as $1/k$. The results in [36, 37] have a similar structure. For instance, in [36], the authors show that when the initial Newton decrement is smaller than $\frac{\mu^{\frac{5}{2}}}{ML}$, which is independent of the problem dimension, the convergence rate would be of the form $(\frac{dL}{\mu k})^{k/2}$. Hence, to observe the superlinear convergence rate one need to run the BFGS method at least for dL/μ iterations to ensure the contraction factor is smaller than 1. A similar conclusion could be made using our results, if we adjust the neighborhood. In our main result, we only report the case that the neighborhood depends on d and the rate is independent of that, since in this case the contraction factor is always smaller than 1 and the superlinear behavior starts from the first iteration.

7 Numerical experiments

In this section, we present our numerical experiments and compare the non-asymptotic performance of quasi-Newton methods with Newton's method and the gradient descent algorithm. We further investigate if the convergence rates of quasi-Newton methods are consistent with our theoretical guarantees. In particular, we solve the following logistic regression problem with l_2 regularization

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{N}\sum_{i=1}^{N} \ln\left(1+e^{-y_i z_i^\top x}\right) + \frac{\mu}{2}\|x\|^2. \tag{108}$$

We assume that $\{z_i\}_{i=1}^N$ are the data points and $\{y_i\}_{i=1}^N$ are their corresponding labels where $z_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ for $1 \leq i \leq N$. Note that the function $f(x)$ in (108) is strongly convex with parameter $\mu > 0$. We normalize all data points such that $\|z_i\| = 1$ for all $1 \leq i \leq N$. Therefore, the gradient of the function $f(x)$ is Lipschitz continuous with parameter $L = 1/4 + \mu$. It is also known that the logistic regression objective function is self-concordant after a suitable scaling, i.e., it is self-concordant but not standard self-concordant (with constant 2). Moreover,

Table 1 Sample size N, dimension d, initial point parameter c and regularization μ of each dataset

Dataset	N	d	c	μ
Colon-cancer	62	2000	0.1	0.01
Covertype	581, 012	54	1	0.001
GISETTE	6000	5000	0.1	0.01
MNIST	11774	784	0.1	0.01

its Hessian is Lipschitz continuous. In summary, the objective function f in (108) satisfies Assumptions 3.1, 3.2, 3.3 and Assumption 5.1.

We conduct our experiments on four different datasets: (i) colon-cancer dataset [40], (ii) Covertype dataset [41], (iii) GISETTE handwritten digits classification dataset from the NIPS 2003 feature selection challenge [42] and (iv) MNIST dataset of handwritten digits [43].[1] We compare the performance of DFP, BFGS, Newton's method, and gradient descent. We initialize all the algorithms with the same initial point $x_0 = c * \mathbf{1}$ where $c > 0$ is a tuned parameter and $\mathbf{1} \in \mathbb{R}^d$ is the one vector. We set the initial Hessian inverse approximation matrix as $\nabla^2 f(x_0)^{-1}$ for the DFP and BFGS methods. The step size is 1 for DFP, BFGS, and Newton's method. The step size of the gradient descent method is tuned by hand to achieve the best performance on each dataset.

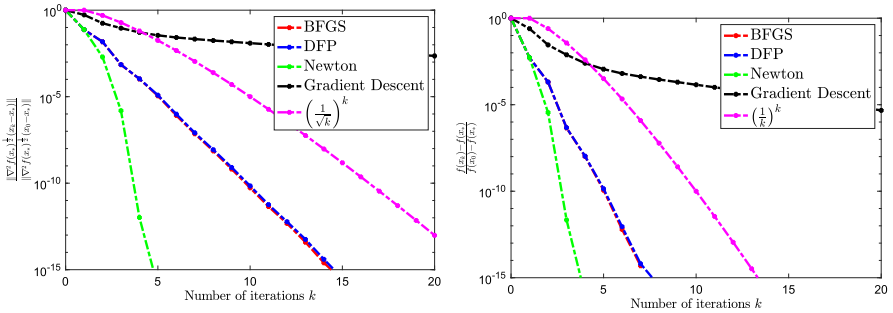
All the parameters (sample size N, dimension d, initial point parameter c and regularization μ) of these different datasets are provided in Table 1. Notice that the initial point parameter c is selected from the set $\mathcal{A} = \{0.001, 0.01, 0.1, 1, 10\}$ to guarantee that the initial point x_0 is close enough to the optimal solution x_* so that we can achieve the superlinear convergence rate of DFP and BFGS on each dataset. The regularization parameter μ is also chosen from the same set $\mathcal{A}$ to obtain the best performance on each dataset.

From the theoretical results of Sects. 4.3 and 5, we expect the iterates $\{x_k\}_{k=0}^{\infty}$ generated by the DFP method and the BFGS method to satisfy the following superlinear convergence rate

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{1}{\sqrt{k}}\right)^k, \qquad \frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq 1.1\left(\frac{1}{k}\right)^k, \qquad \forall k \geq 1.$$

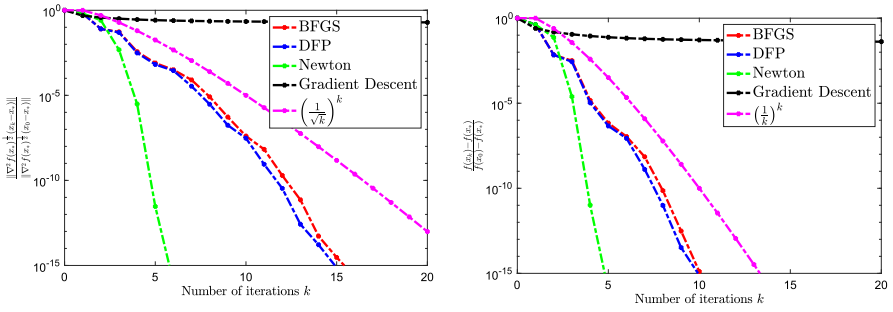
Hence, in our numerical experiments, we compare the convergence rate of $\frac{\|\nabla^2 f(x_*)^{1/2}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{1/2}(x_0 - x_*)\|}$ with $(\frac{1}{\sqrt{k}})^k$ and the convergence rate of $\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)}$ with $(\frac{1}{k})^k$ to check the tightness of our theoretical bounds. Our numerical experiments are shown in Figs. 1, 2, 3 and 4 for different datasets. Note that for each problem, we present two plots. The left plot (plot (a)) showcases $\frac{\|\nabla^2 f(x_*)^{1/2}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{1/2}(x_0 - x_*)\|}$ for different algorithms as well as our theoretical bound which is $(\frac{1}{\sqrt{k}})^k$. In the right plot (plot (b)), we compare $\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)}$ for different methods with our theoretical bound which is $(\frac{1}{k})^k$.

[1] We use LIBSVM [44] with license: https://www.csie.ntu.edu.tw/~cjlin/libsvm/COPYRIGHT.



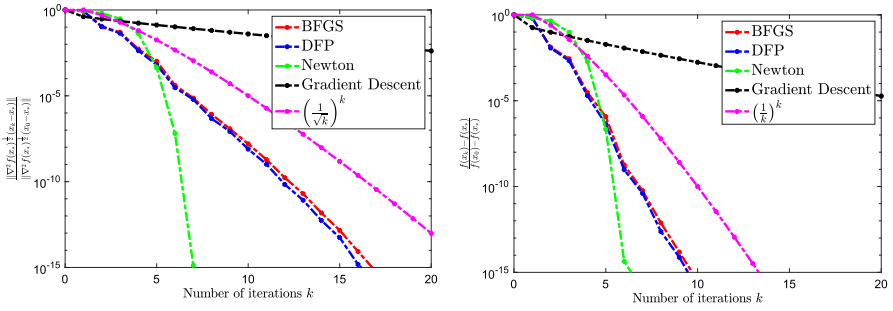
(a) Results of $\frac{\|\nabla^2 f(x_*)^{1/2}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{1/2}(x_0 - x_*)\|}$. **(b)** Results of $\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)}$.

Fig. 1 Convergence rates of logistic regression on the Colon-cancer dataset



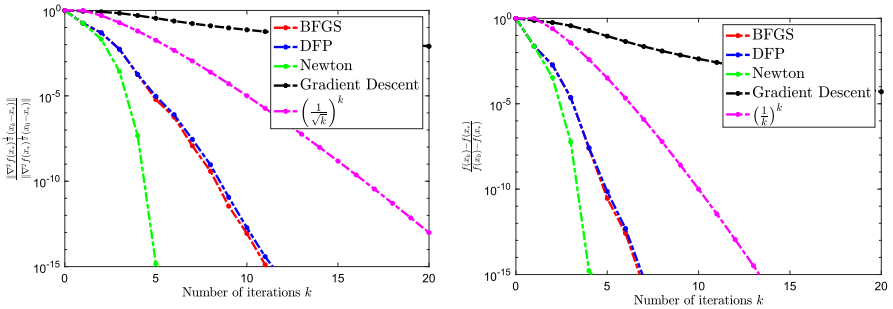
(a) Results of $\frac{\|\nabla^2 f(x_*)^{1/2}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{1/2}(x_0 - x_*)\|}$. **(b)** Results of $\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)}$.

Fig. 2 Convergence rates of logistic regression on the Covertype dataset



(a) Results of $\frac{\|\nabla^2 f(x_*)^{1/2}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{1/2}(x_0 - x_*)\|}$. **(b)** Results of $\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)}$.

Fig. 3 Convergence rates of logistic regression on the GISETTE dataset



(a) Results of $\frac{\|\nabla^2 f(x_*)^{1/2}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{1/2}(x_0 - x_*)\|}$. **(b)** Results of $\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)}$.

Fig. 4 Convergence rates of logistic regression on the MNIST dataset

We observe that $\frac{\|\nabla^2 f(x_*)^{1/2}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{1/2}(x_0 - x_*)\|}$ for the DFP and BFGS methods are bounded above by $(\frac{1}{\sqrt{k}})^k$ and $\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)}$ for the DFP and BFGS methods are bounded above by $(\frac{1}{k})^k$. Therefore, these experimental results confirm our theoretical superlinear convergence rates of quasi-Newton methods.

8 Conclusion

In this paper, we studied the local convergence rate of the convex Broyden class of quasi-Newton methods which includes the DFP and BFGS methods. We focused on two settings: (i) the objective function is μ-strongly convex, its gradient is L-Lipschitz continuous, and its Hessian is Lipschitz continuous at the optimal solution with parameter M, (ii) the objective function is self-concordant. For these two settings we characterized the explicit non-asymptotic superlinear convergence rate of Broyden class of quasi-Newton methods. In particular, for the first setting, we showed that if the initial distance to the optimal solution is $\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| = \mathcal{O}(\frac{\mu^{\frac{3}{2}}}{M})$ and the initial Hessian approximation error is $\|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F = \mathcal{O}(1)$, the iterations generated by the DFP and BFGS methods converge to the optimal solution at a superlinear rate of $\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)^k$ and $\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} = \mathcal{O}\left(\frac{1}{k}\right)^k$. We further showed that we can achieve the same superlinear convergence rate if the initial error is $\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| = \mathcal{O}(\frac{\mu^{\frac{3}{2}}}{M\sqrt{d}})$ and the initial Hessian approximation matrix is $B_0 = \nabla^2 f(x_0)$. We proved similar convergence rate results for the second setting where the objective function is self-concordant.

Acknowledgements Q. Jin acknowledges support from the National Initiative for Modeling and Simulation (NIMS) Graduate Research Fellowship. This research is supported by NSF Award CCF-2007668.

Code Availability The code of the numerical experiments is available at the following link: https://github.com/QiujiangJin/Non-asymptotic-Superlinear-Convergence-of-Standard-Quasi-Newton-Methods

Appendix

A Proof of corollary 1

According to the definition $J = \int_0^1 \nabla^2 f(x + t(y - x))dt$, we have $\nabla f(x) - \nabla f(y) = J(x - y)$. Hence, we can write

$$\begin{aligned} \|\nabla f(x) - \nabla f(y) - \nabla^2 f(x_*)(x - y)\| &= \|(J - \nabla^2 f(x_*))(x - y)\| \\ &\leq \|J - \nabla^2 f(x_*)\|\|x - y\|. \end{aligned} \tag{109}$$

Moreover, we can show that

$$\begin{aligned} \|J - \nabla^2 f(x_*)\| &= \left\| \int_0^1 [\nabla^2 f(x + t(y - x)) - \nabla^2 f(x_*)]dt \right\| \\ &\leq \int_0^1 \|\nabla^2 f(x + t(y - x)) - \nabla^2 f(x_*)\|dt. \end{aligned}$$

By Assumption 3.3, we can replace the upper bound in the above expression by the following

$$\begin{aligned} \|J - \nabla^2 f(x_*)\| &\leq M \int_0^1 \|x + t(y - x) - x_*\|dt \\ &\leq M \left[\int_0^1 (1 - t)\|x - x_*\|dt + \int_0^1 t\|y - x_*\|dt \right] \\ &= \frac{M}{2}(\|x - x_*\| + \|y - x_*\|). \end{aligned} \tag{110}$$

By combining (109) and (110), the result in (12) follows.

B Proof of Lemma 1

Define $P := I - uu^\top$. Since $\|u\| = 1$, we have $P = P^\top$, $P^2 = P$, and $0 \preceq P \preceq I$. These properties imply that

$$\begin{aligned}\|PAP\|_F^2 &= \mathrm{Tr}(PAPPA^\top P) = \mathrm{Tr}(PAPA^\top P)\\ &= \mathrm{Tr}(PPAPA^\top) = \mathrm{Tr}(PAPA^\top).\end{aligned} \tag{111}$$

Moreover, for symmetric matrices X_1 and X_2 that satisfy $X_1 \preceq X_2$ we have $\mathrm{Tr}(X_1 Y) \leq \mathrm{Tr}(X_2 Y)$ when Y is positive-semidefinite. This result and $0 \preceq P \preceq I$ imply that

$$\mathrm{Tr}(PAPA^\top) \leq \mathrm{Tr}(APA^\top) = \mathrm{Tr}(AA^\top - Auu^\top A^\top) = \|A\|_F^2 - \|Au\|^2. \tag{112}$$

By combining the results in (111) and (112), and considering the definition $P := I - uu^\top$, the claim in (18) follows.

C Proof of Lemma 2

Notice that $\mathrm{Tr}(X_1 Y) \leq \mathrm{Tr}(X_2 Y)$ for any symmetric matrices $X_1 \preceq X_2$ and symmetric positive-semidefinite matrix Y. Since $A^\top A \preceq \|A\|^2 I$, we obtain that

$$\|AB\|_F^2 = \mathrm{Tr}(B^\top A^\top AB) = \mathrm{Tr}(A^\top ABB^\top) \leq \|A\|^2 \mathrm{Tr}(BB^\top) = \|A\|^2 \|B\|_F^2,$$

which leads to the first inequality in (19). The second inequality in (19) follows from the first one, since

$$\|B^\top AB\|_F \leq \|B^\top A\| \|B\|_F \leq \|A\| \|B\| \|B\|_F.$$

D Proof of Lemma 3

By Assumption 3.3, we have that

$$\begin{aligned}\|J_k - \nabla^2 f(x_*)\| &= \left\| \int_0^1 \left[\nabla^2 f(x_* + \alpha(x_k - x_*)) - \nabla^2 f(x_*) \right] d\alpha \right\|\\ &\leq \int_0^1 M\alpha \|x_k - x_*\| d\alpha = \frac{M}{2} \|x_k - x_*\|.\end{aligned}$$

Hence, we have $J_k - \nabla^2 f(x_*) \preceq \frac{M}{2}\|x_k - x_*\| I$. Considering this bound and Assumption 3.1, we obtain

$$\begin{aligned}J_k - \nabla^2 f(x_*) &\preceq \frac{M}{2\mu} \|x_k - x_*\| \nabla^2 f(x_*) = \frac{M}{2\mu} \|\nabla^2 f(x_*)^{-\frac{1}{2}} \nabla^2 f(x_*)^{\frac{1}{2}} (x_k - x_*)\| \nabla^2 f(x_*)\\ &\preceq \frac{M}{2\mu} \|\nabla^2 f(x_*)^{-\frac{1}{2}}\| \|r_k\| \nabla^2 f(x_*) \preceq \frac{M}{2\mu^{\frac{3}{2}}} \|r_k\| \nabla^2 f(x_*) = \frac{\sigma_k}{2} \nabla^2 f(x_*).\end{aligned} \tag{113}$$

Similarly, we have that

$$\nabla^2 f(x_*) - J_k \preceq \|J_k - \nabla^2 f(x_*)\| I \preceq \frac{M}{2}\|x_k - x_*\| I \preceq \frac{M}{2\mu}\|x_k - x_*\| J_k \preceq \frac{\sigma_k}{2} J_k. \tag{114}$$

Combining (113) and (114), we obtain that

$$\frac{1}{1+\frac{\sigma_k}{2}}\nabla^2 f(x_*) \preceq J_k \preceq \left(1+\frac{\sigma_k}{2}\right)\nabla^2 f(x_*).$$

Multiplying both side of the above expression by $\nabla^2 f(x_*)^{-\frac{1}{2}}$ from left and right leads to the first result in (20). By Assumption 3.3 and $\alpha_k \in [0, 1]$, we have that

$$\begin{aligned}\|H_k - \nabla^2 f(x_*)\| &= \|\nabla^2 f(x_* + \alpha_k(x_k - x_*)) - \nabla^2 f(x_*)\| \leq M\alpha_k\|x_k - x_*\| \\ &\leq M\|x_k - x_*\|.\end{aligned}$$

Hence, we have $H_k - \nabla^2 f(x_*) \preceq M\|x_k - x_*\| I$. Considering this bound and Assumption 3.1, we obtain

$$\begin{aligned}H_k - \nabla^2 f(x_*) &\preceq \frac{M}{\mu}\|x_k - x_*\|\nabla^2 f(x_*) = \frac{M}{\mu}\|\nabla^2 f(x_*)^{-\frac{1}{2}}\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|\nabla^2 f(x_*) \\ &\preceq \frac{M}{\mu}\|\nabla^2 f(x_*)^{-\frac{1}{2}}\|\|r_k\|\nabla^2 f(x_*) \preceq \frac{M}{\mu^{\frac{3}{2}}}\|r_k\|\nabla^2 f(x_*) = \sigma_k\nabla^2 f(x_*).\end{aligned} \tag{115}$$

Similarly, we have that

$$\nabla^2 f(x_*) - H_k \preceq \|H_k - \nabla^2 f(x_*)\| I \preceq M\|x_k - x_*\| I \preceq \frac{M}{\mu}\|x_k - x_*\| H_k \preceq \sigma_k H_k. \tag{116}$$

Combining (115) and (116), we obtain that

$$\frac{1}{1+\sigma_k}\nabla^2 f(x_*) \preceq H_k \preceq (1+\sigma_k)\nabla^2 f(x_*).$$

Multiplying both side of the above expression by $\nabla^2 f(x_*)^{-\frac{1}{2}}$ from left and right leads to the second result in (20).

E Proof of Lemma 4

By Assumption 3.1 and Corollary 1, we have

$$\begin{aligned}\|\hat{y}_k - \hat{s}_k\| &= \|\nabla^2 f(x_*)^{-\frac{1}{2}} y_k - \nabla^2 f(x_*)^{\frac{1}{2}} s_k\| \le \|\nabla^2 f(x_*)^{-\frac{1}{2}}\| \|y_k - \nabla^2 f(x_*) s_k\| \\ &= \|\nabla^2 f(x_*)^{-\frac{1}{2}}\| \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_*)(x_{k+1} - x_k)\| \\ &\le \frac{M}{\mu^{\frac{1}{2}}} \|s_k\| \frac{1}{2} (\|x_{k+1} - x_*\| + \|x_k - x_*\|) \le \frac{M}{\mu^{\frac{1}{2}}} \|s_k\| \max\{\|x_{k+1} - x_*\|, \|x_k - x_*\|\}.\end{aligned} \tag{117}$$

Notice that

$$\|s_k\| = \|\nabla^2 f(x_*)^{-\frac{1}{2}} \nabla^2 f(x_*)^{\frac{1}{2}} s_k\| \le \|\nabla^2 f(x_*)^{-\frac{1}{2}}\| \|\hat{s}_k\| \le \frac{1}{\mu^{\frac{1}{2}}} \|\hat{s}_k\|. \tag{118}$$

Based on the definition $r_k = \nabla^2 f(x_*)^{\frac{1}{2}} (x_k - x_*)$, we have $x_k - x_* = \nabla^2 f(x_*)^{-\frac{1}{2}} r_k$ and hence

$$\begin{aligned}\max\{\|x_{k+1} - x_*\|, \|x_k - x_*\|\} &\le \|\nabla^2 f(x_*)^{-\frac{1}{2}}\| \max\{\|r_k\|, \|r_{k+1}\|\} \\ &\le \frac{1}{\mu^{\frac{1}{2}}} \max\{\|r_k\|, \|r_{k+1}\|\}.\end{aligned} \tag{119}$$

Substitute (118) and (119) into (117) and recall the definition in (16) to obtain

$$\|\hat{y}_k - \hat{s}_k\| \le \frac{M}{\mu^{\frac{3}{2}}} \max\{\|r_k\|, \|r_{k+1}\|\} \|\hat{s}_k\| = \tau_k \|\hat{s}_k\|.$$

Hence, the proof of the first claim in (21) is complete. By using the Cauchy-Schwarz inequality and (21), we can write

$$|(\hat{y}_k - \hat{s}_k)^\top \hat{s}_k| \le \|\hat{y}_k - \hat{s}_k\| \|\hat{s}_k\| \le \tau_k \|\hat{s}_k\|^2.$$

Therefore, we obtain that

$$(1 - \tau_k)\|\hat{s}_k\|^2 \le \hat{s}_k^\top \hat{y}_k \le (1 + \tau_k)\|\hat{s}_k\|^2,$$

and the second claim in (22) holds. Using the reverse triangle inequality and (21), we have $|\|\hat{y}_k\| - \|\hat{s}_k\|| \le \|\hat{y}_k - \hat{s}_k\| \le \tau_k \|\hat{s}_k\|$. Hence, the third claim in (23) holds.

Finally, to prove the last claim in (24), we use Assumption 3.1 and Corollary 1 to show that

$$\begin{aligned}\|\widehat{\nabla f}(x_k) - r_k\| &= \|\nabla^2 f(x_*)^{-\frac{1}{2}} \nabla f(x_k) - \nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\| \\ &\leq \|\nabla^2 f(x_*)^{-\frac{1}{2}}\| \|\nabla f(x_k) - \nabla f(x_*) - \nabla^2 f(x_*)(x_k - x_*)\| \\ &\leq \frac{M}{2\mu^{\frac{1}{2}}} \|x_k - x_*\|^2 = \frac{M}{2\mu^{\frac{1}{2}}} \|\nabla^2 f(x_*)^{-\frac{1}{2}} \nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|^2 \\ &\leq \frac{M}{2\mu^{\frac{1}{2}}} \|\nabla^2 f(x_*)^{-\frac{1}{2}}\|^2 \|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|^2 \leq \frac{M}{2\mu^{\frac{3}{2}}} \|r_k\|^2 = \frac{\sigma_k}{2}\|r_k\|.\end{aligned}$$

F Proof of Lemma 10

Set $x = x_*$ and $y = x_k$ in (91) of Lemma 9 and note that $\|r_k\| \leq \frac{1}{2} < 1$. By (91), we have

$$\left(1 - \|r_k\| + \frac{\|r_k\|^2}{3}\right) \nabla^2 f(x_*) \preceq J_k \preceq \frac{1}{1 - \|r_k\|} \nabla^2 f(x_*).$$

Multiply the above expressions form left and right by $\nabla^2 f(x_*)^{-\frac{1}{2}}$ to obtain

$$\left(1 - \|r_k\| + \frac{\|r_k\|^2}{3}\right) I \preceq \hat{J}_k \preceq \frac{1}{1 - \|r_k\|} I. \tag{120}$$

Using the fact that $\|r_k\| \leq \frac{1}{2}$, we have

$$1 - \|r_k\| + \frac{\|r_k\|^2}{3} \geq \frac{1}{1 + 2\|r_k\|} \tag{121}$$

$$\frac{1}{1 - \|r_k\|} \leq 1 + 2\|r_k\|. \tag{122}$$

Replace the lower and upper bounds in (120) with the ones in (121) and (122), respectively, to obtain the first result in (92). Set $x = x_*$ and $y = x_* + \alpha_k(x_k - x_*)$ in (90) of Lemma 9 and notice that since $\alpha_k \in [0, 1]$, we obtain that

$$\begin{aligned}r &= \|\nabla^2 f(x)^{\frac{1}{2}}(y - x)\| = \|\nabla^2 f(x_*)^{\frac{1}{2}}(x_* + \alpha_k(x_k - x_*) - x_*)\| \\ &= \alpha_k \|r_k\| \leq \|r_k\| \leq \frac{1}{2} < 1.\end{aligned}$$

By (90), we have that

$$(1 - r)^2 \nabla^2 f(x_*) \preceq H_k \preceq \frac{1}{(1 - r)^2} \nabla^2 f(x_*).$$

Using that $r \leq \|r_k\| < 1$, we get that

$$(1-\|r_k\|)^2\nabla^2 f(x_*) \preceq (1-r)^2\nabla^2 f(x_*) \preceq H_k \preceq \frac{1}{(1-r)^2}\nabla^2 f(x_*) \preceq \frac{1}{(1-\|r_k\|)^2}\nabla^2 f(x_*).$$

Multiply the above expressions form left and right by $\nabla^2 f(x_*)^{-\frac{1}{2}}$ to obtain the second result in (92).

G Proof of Lemma 11

We first show that for $x = x_*$ and $y = x_k + \alpha(x_{k+1} - x_k)$, where $\alpha \in [0, 1]$, the value of $r = \|\nabla^2 f(x)^{\frac{1}{2}}(y - x)\|$ defined in Lemma 9 is less than 1. To do so, note that

$$\begin{aligned} r &= \|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k + \alpha(x_{k+1} - x_k) - x_*)\| \leq \alpha\|r_{k+1}\| + (1-\alpha)\|r_k\| \\ &\leq \max\{\|r_k\|, \|r_{k+1}\|\} = \theta_k \leq \frac{1}{2} < 1, \end{aligned}$$

where $\|r_k\| = \|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|$. Note that in the above simplification we used the assumption that $\theta \leq 1/2$. Now using the result in (90) we have

$$(1-r)^2\nabla^2 f(x_*) \preceq \nabla^2 f(x_k + \alpha(x_{k+1} - x_k)) \preceq \frac{1}{(1-r)^2}\nabla^2 f(x_*).$$

Moreover, since $r \leq \theta_k \in [0, 1)$, we can write

$$(1-\theta_k)^2\nabla^2 f(x_*) \preceq \nabla^2 f(x_k + \alpha(x_{k+1} - x_k)) \preceq \frac{1}{(1-\theta_k)^2}\nabla^2 f(x_*).$$

By computing the integral for α from 0 to 1 in the above inequality, we get that

$$(1-\theta_k)^2\nabla^2 f(x_*) \preceq G_k \preceq \frac{1}{(1-\theta_k)^2}\nabla^2 f(x_*),$$

where we used the definition $G_k := \int_0^1 \nabla^2 f(x_k + \alpha(x_{k+1} - x_k))d\alpha$. Multiplying the above expression from left and right by $\nabla^2 f(x_*)^{-\frac{1}{2}}$ leads to

$$(1-\theta_k)^2 I \preceq \hat{G}_k \preceq \frac{1}{(1-\theta_k)^2} I,$$

where $\hat{G}_k = \nabla^2 f(x_*)^{-\frac{1}{2}} G_k \nabla^2 f(x_*)^{-\frac{1}{2}}$. The above inequality is equivalent to

$$\left((1-\theta_k)^2 - 1\right) I \preceq \hat{G}_k - I \preceq \left(\frac{1}{(1-\theta_k)^2} - 1\right) I,$$

which indicates that

$$\|\hat{G}_k - I\| \leq \max\left\{\frac{1}{(1-\theta_k)^2} - 1, 1 - (1-\theta_k)^2\right\}. \tag{123}$$

Since $\theta_k \in [0, 1)$, we have that

$$\frac{1}{(1-\theta_k)^2} - 1 = \frac{1-(1-\theta_k)^2}{(1-\theta_k)^2} \geq 1 - (1-\theta_k)^2.$$

Hence, (123) can be simplified as

$$\|\hat{G}_k - I\| \leq \frac{1}{(1-\theta_k)^2} - 1 = \frac{(2-\theta_k)}{(1-\theta_k)^2}\theta_k \leq \frac{(2-\frac{1}{2})}{(1-\frac{1}{2})^2}\theta_k = 6\theta_k, \tag{124}$$

where the second inequality holds due to $\theta_k \leq \frac{1}{2}$.

Considering the definition $G_k := \int_0^1 \nabla^2 f(x_k + \alpha(x_{k+1} - x_k))d\alpha$, we have $y_k = G_k s_k$. Using this observation, we have

$$\begin{aligned}
\|\hat{y}_k - \hat{s}_k\| &= \|\nabla^2 f(x_*)^{-\frac{1}{2}} y_k - \nabla^2 f(x_*)^{\frac{1}{2}} s_k\| \\
&= \|\nabla^2 f(x_*)^{-\frac{1}{2}} G_k \nabla^2 f(x_*)^{-\frac{1}{2}} \nabla^2 f(x_*)^{\frac{1}{2}} s_k - \nabla^2 f(x_*)^{\frac{1}{2}} s_k\| \\
&= \|\hat{G}_k \hat{s}_k - \hat{s}_k\| \leq \|\hat{G}_k - I\|\|\hat{s}_k\| \leq 6\theta_k \|\hat{s}_k\|,
\end{aligned}$$

where the last inequality holds due to (124). Hence, the proof of the first claim in (93) is complete.

By using the Cauchy-Schwarz inequality and (93), we can write

$$|(\hat{y}_k - \hat{s}_k)^\top \hat{s}_k| \leq \|\hat{y}_k - \hat{s}_k\|\|\hat{s}_k\| \leq 6\theta_k \|\hat{s}_k\|^2.$$

Therefore, we obtain that

$$(1 - 6\theta_k)\|\hat{s}_k\|^2 \leq \hat{s}_k^\top \hat{y}_k \leq (1 + 6\theta_k)\|\hat{s}_k\|^2,$$

and the second claim in (94) holds. Using the reverse triangle inequality and (93), we have $|\|\hat{y}_k\| - \|\hat{s}_k\|| \leq \|\hat{y}_k - \hat{s}_k\| \leq 6\theta_k\|\hat{s}_k\|$. Hence, the third claim in (95) holds. Finally, using Lemma 10 we know that

$$\|\hat{J}_k - I\| \leq \max\{2\|r_t\|, 1 - \frac{1}{1 + 2\|r_t\|}\} = 2\|r_t\|,$$

and

$$\begin{aligned}\|\widehat{\nabla f}(x_k) - r_k\| &= \|\nabla^2 f(x_*)^{-\frac{1}{2}} \nabla f(x_k) - \nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\| \\ &= \|\nabla^2 f(x_*)^{-\frac{1}{2}} J_k(x_k - x_*) - \nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\| \\ &= \|\hat{J}_k r_k - r_k\| \leq \|\hat{J}_k - I\|\|r_k\| \leq 2\|r_k\|^2.\end{aligned}$$

Thus, the last claim in (96) holds.

H Proof of Theorem 3

From the assumption on the initial point, it follows that $x_k \in \operatorname{dom} f$ for all $k \geq 0$, so that Algorithm 1 is well-defined. The proof of this Theorem 3 is very similar to the proof of Theorem 1. The only difference is that we utilize the Lemma 10 and Lemma 11 instead of Lemma 3 and Lemma 4. Hence, we need to replace all the terms $\frac{\sigma_k}{2} = \frac{M}{2\mu^{\frac{3}{2}}}\|r_k\|$ by $2\|r_k\|$ and $\tau_k = \max\{\sigma_k, \sigma_{k+1}\}$ by $6\theta_k = 6\max\{\|r_k\|, \|r_{k+1}\|\}$. Here, we only stated the outline of the proof and omit the details to avoid redundancy.

First we present the potential function similar to the (46) from Lemma 7. Suppose that for some $\delta > 0$ and some $k \geq 0$, we have $\theta_k = \max\{\|r_k\|, \|r_{k+1}\|\} < \frac{1}{6}$ and $\|\hat{B}_k - I\|_F \leq \delta$. Then, the matrix B_{k+1} generated by the convex Broyden class update (7) satisfies

$$\begin{aligned}\|\hat{B}_{k+1} - I\|_F \leq \|\hat{B}_k - I\|_F - \phi_k \frac{\|(\hat{B}_k - I)\hat{s}_k\|^2}{2\delta\|\hat{s}_k\|^2} \\ -(1-\phi_k)\frac{\hat{s}_k^\top(\hat{B}_k - I)\hat{B}_k(\hat{B}_k - I)\hat{s}_k}{2\delta\hat{s}_k^\top \hat{B}_k \hat{s}_k} + 6Z_k\theta_k,\end{aligned} \tag{125}$$

where $Z_k = \phi_k\|\hat{B}_k\|\frac{4}{(1-6\theta_k)^2} + \frac{3+6\theta_k}{1-6\theta_k}$. We also have that

$$\|\hat{B}_{k+1} - I\|_F \leq \|\hat{B}_k - I\|_F + 6Z_k\theta_k. \tag{126}$$

The proof of the above conclusion is the same as the proof we presented in Lemmas 5, 6, and 7 except that we use the results of Lemma 11 instead of Lemma 4. Then, we present the similar linear convergence results like Lemma 8. Suppose that the objective function f satisfies the conditions in Assumption 5.1. Moreover, suppose the initial point x_0 and initial Hessian approximation matrix B_0 satisfy

$$\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \frac{\epsilon}{6}, \qquad \|\nabla^2 f(x_*)^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F \leq \delta, \tag{127}$$

where $\epsilon, \delta \in (0, \frac{1}{2})$ such that for some $\rho \in (0, 1)$, they satisfy

$$\left(\phi_{\max}(2\delta+1)\frac{4}{(1-\epsilon)^2}+\frac{3+\epsilon}{1-\epsilon}\right)\frac{\epsilon}{1-\rho} \leq \delta,$$
$$\frac{\epsilon}{3}+2\delta \leq (1-2\delta)\rho,$$
$$\phi_{\max} = \sup_{k\geq 0}\phi_k\ . \tag{128}$$

Then, the sequence of iterates $\{x_k\}_{k=0}^{+\infty}$ converges to the optimal solution x_* with

$$\|r_{k+1}\| \leq \rho\|r_k\|, \qquad \forall k \geq 0. \tag{129}$$

Furthermore, the matrices $\{B_k\}_{k=0}^{+\infty}$ stay in a neighborhood of $\nabla^2 f(x_*)$ defined as

$$\|\hat{B}_{k+1} - I\|_F \leq 2\delta, \qquad \forall k \geq 0. \tag{130}$$

Moreover, the norms $\{\|\hat{B}_k\|\}_{k=0}^{+\infty}$ and $\{\|\hat{B}_k^{-1}\|\}_{k=0}^{+\infty}$ are all uniformly bounded above by

$$\|\hat{B}_k\| \leq 1+2\delta, \qquad \|\hat{B}_k^{-1}\| \leq \frac{1}{1-2\delta}, \qquad \forall k \geq 0. \tag{131}$$

We apply the same induction technique used in the proof of Lemma 8 to prove the above linear convergence results and utilize the potential function in (126) and Lemma 11. Finally we can prove the superlinear convergence results of

$$\frac{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_k - x_*)\|}{\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\|} \leq \left(\frac{C_3 q\sqrt{k}+C_4}{k}\right)^k, \qquad \forall k \geq 1, \tag{132}$$

$$\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)} \leq \frac{1}{(1-\frac{\epsilon}{6})^4}\left(\frac{C_3 q\sqrt{k}+C_4}{k}\right)^{2k}, \qquad \forall k \geq 1, \tag{133}$$

$$C_3 = 2\sqrt{2}\delta(1+\rho)\left(1+\frac{\epsilon}{3}\right), \qquad C_4 = \frac{(1+\rho)(1+\frac{\epsilon}{3})\epsilon}{3(1-\rho)}, \tag{134}$$

where $q = \frac{1}{\sqrt{\phi_{\min}\frac{4\delta}{1+2\delta}+\frac{1-2\delta}{1+2\delta}}} \in \left[1, \sqrt{\frac{1+2\delta}{1-2\delta}}\right]$ and $\phi_{\min} = \inf_{k\geq 0}\phi_k$. This proof is based on the linear convergence results of (129), (130), (131) and is the same as the proof in Theorem 1, except that here we replace the results of Lemma 3 by the results of Lemma 10, substitute the results of Lemma 4 with the results of Lemma 11 and utilize the intermediate inequality (125) instead of (46). Notice that all the term $\frac{\epsilon}{2}$ has been replaced with the term $\frac{\epsilon}{3}$ since in this setting, we use the term $2\|r_t\|$ instead of the term $\frac{\sigma_t}{2}$ and $2\|r_t\| \leq 2\|r_0\| \leq 2\frac{\epsilon}{6} = \frac{\epsilon}{3}$.

I Proof of Theorem 4

First we focus on the DFP method. Notice that by (105) we have

$$\|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \frac{1}{720}.$$

Hence, the first condition in (102) is satisfied. Set $x = x_*$ and $y = x_0$ in Lemma 9. Notice that $\|r_0\| = \|\nabla^2 f(x_*)^{\frac{1}{2}}(x_0 - x_*)\| \leq \frac{1}{720} < 1$. Hence, using (90) we obtain that

$$(1 - \|r_0\|)^2 \nabla^2 f(x_*) \preceq \nabla^2 f(x_0) \preceq \frac{1}{(1 - \|r_0\|)^2} \nabla^2 f(x_*).$$

Multiply the above expression by $\nabla^2 f(x_*)^{-\frac{1}{2}}$ from left and right to obtain

$$(1 - \|r_0\|)^2 I \preceq \nabla^2 f(x_*)^{-\frac{1}{2}} \nabla^2 f(x_0) \nabla^2 f(x_*)^{-\frac{1}{2}} \preceq \frac{1}{(1 - \|r_0\|)^2} I,$$

which implies that

$$\begin{aligned}
\nabla^2 f(x_*)^{-\frac{1}{2}} (\nabla^2 f(x_0) - \nabla^2 f(x_*)) \nabla^2 f(x_*)^{-\frac{1}{2}} &\preceq \left(\frac{1}{(1 - \|r_0\|)^2} - 1 \right) I, \\
\nabla^2 f(x_*)^{-\frac{1}{2}} (\nabla^2 f(x_0) - \nabla^2 f(x_*)) \nabla^2 f(x_*)^{-\frac{1}{2}} &\succeq \left((1 - \|r_0\|)^2 - 1 \right) I.
\end{aligned}$$

The above two inequalities indicate that

$$\begin{aligned}
&\|\nabla^2 f(x_*)^{-\frac{1}{2}} (\nabla^2 f(x_0) - \nabla^2 f(x_*)) \nabla^2 f(x_*)^{-\frac{1}{2}}\| \\
&\leq \max \left\{ \frac{1}{(1 - \|r_0\|)^2} - 1, 1 - (1 - \|r_0\|)^2 \right\}. \qquad (135)
\end{aligned}$$

Since $\|r_0\| \in [0, 1)$, we have that

$$\frac{1}{(1 - \|r_0\|)^2} - 1 = \frac{1 - (1 - \|r_0\|)^2}{(1 - \|r_0\|)^2} \geq 1 - (1 - \|r_0\|)^2.$$

Hence, (135) can be simplified as

$$\begin{aligned}
&\|\nabla^2 f(x_*)^{-\frac{1}{2}} (\nabla^2 f(x_0) - \nabla^2 f(x_*)) \nabla^2 f(x_*)^{-\frac{1}{2}}\| \\
&\leq \frac{1}{(1 - \|r_0\|)^2} - 1 = \frac{(2 - \|r_0\|)}{(1 - \|r_0\|)^2} \|r_0\| \leq \frac{(2 - \frac{1}{720})}{(1 - \frac{1}{720})^2} \|r_0\| \leq 3\|r_0\|,
\end{aligned}$$

where the second inequality holds due to $\|r_0\| \leq \frac{1}{720}$. Therefore, we can show that

$$\begin{aligned}&\|\nabla^2 f(x_*)^{-\frac{1}{2}}(\nabla^2 f(x_0) - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\|_F\\ \leq&\sqrt{d}\|\nabla^2 f(x_*)^{-\frac{1}{2}}(\nabla^2 f(x_0) - \nabla^2 f(x_*))\nabla^2 f(x_*)^{-\frac{1}{2}}\| \leq 3\sqrt{d}\|r_0\| \leq \frac{1}{7},\end{aligned}$$

where the first inequality is true since $\|A\|_F \leq \sqrt{d}\|A\|$ for any matrix $A \in \mathbb{R}^{d\times d}$ and the last inequality is due to the first part of (105). Hence, the second condition in (102) is also satisfied. By Corollary 4, we can conclude that (107) holds. The proof for BFGS is similar to the proof for DFP. It can be derived by following the steps of proof of DFP and exploiting the BFGS results in Corollary 4.

References

1. Nesterov, Y.: A method for solving the convex programming problem with convergence rate o(1/k2). Dokl. Akad. Nauk SSSR **269**, 543–547 (1983)
2. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer Science & Business Media, Berlin (2013)
3. Nemirovsky, A., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. SIAM, New Delhi (1983)
4. Bennett, A.A.: Newton's method in general analysis. Proc. Natl. Acad. Sci. U. S. A. **2**(10), 592 (1916)
5. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables, vol. 30. SIAM, New Delhi (1970)
6. Conn, A.R., Gould, N.I., Toint, P.L.: Trust Region Methods. SIAM, New Delhi (2000)
7. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton method and its global performance. Math. Program. **108**(1), 177–205 (2006)
8. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
9. Nesterov, Y.: Introductory Lectures on Convex Optimization. Springer Science & Business Media, Berlin (2004)
10. Nocedal, J., Wright, S.: Numerical Optimization. Springer Science & Business Media, Berlin (2006)
11. Conn, A.R., Gould, N.I.M., Toint, P.L.: Convergence of quasi-Newton matrices generated by the symmetric rank one update. Math. Program. **50**(1–3), 177–195 (1991)
12. Broyden, C.G.: A class of methods for solving nonlinear simultaneous equations. Math. Comput. **19**(92), 577–593 (1965)
13. Broyden, C.G., Broyden, J.E.D., Jr., More, J.J.: On the local and superlinear convergence of quasi-Newton methods. IMA J. Appl. Math. **12**, 223–245 (1973)
14. Gay, D.M.: Some convergence properties of Broyden's method. SIAM J. Numer. Anal. **16**(4), 623–630 (1979)
15. Davidon, W.: Variable metric method for minimization. SIAM J. Optim. (1991). https://epubs.siam.org/doi/10.1137/0801001
16. Fletcher, R., Powell, M.J.: A rapidly convergent descent method for minimization. Comput. J. **6**(2), 163–168 (1963)
17. Broyden, C.G.: The convergence of single-rank quasi-Newton methods. Math. Comput. **24**(110), 365–382 (1970)
18. Fletcher, R.: A new approach to variable metric algorithms. Comput. J. **13**(3), 317–322 (1970)
19. Goldfarb, D.: A family of variable-metric methods derived by variational means. Math. Comput. **24**(109), 23–26 (1970)
20. Shanno, D.F.: Conditioning of quasi-Newton methods for function minimization. Math. Comput. **24**(111), 647–656 (1970)
21. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Math. Comput. **35**(151), 773–782 (1980)

22. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. **45**(1–3), 503–528 (1989)
23. Moré, J.J., Trangenstein, J.A.: On the global convergence of Broyden's method. Math. Comput. **30**(135), 523–540 (1976)
24. Powell, M.: On the convergence of the variable metric algorithm. IMA J. Appl. Math. **7**(1), 21–36 (1971)
25. Dennis, J.E., Moré, J.J.: A characterization of superlinear convergence and its application to quasi-Newton methods. Math. Comput. **28**(126), 549–560 (1974)
26. Byrd, R.H., Nocedal, J., Yuan, Y.-X.: Global convergence of a class of quasi-Newton methods on convex problems. SIAM J. Numer. Anal. **24**(5), 1171–1190 (1987)
27. Gao, W., Goldfarb, D.: Quasi-newton methods: superlinear convergence without line searches for self-concordant functions. Opt. Methods Softw. **34**(1), 194–217 (2019)
28. Griewank, A., Toint, P.L.: Local convergence analysis for partitioned quasi-Newton updates. Numer. Math. **39**(3), 429–448 (1982)
29. Dennis, J., Martinez, H.J., Tapia, R.A.: Convergence theory for the structured BFGS secant method with an application to nonlinear least squares. J. Optim. Theory Appl. **61**(2), 161–178 (1989)
30. Yuan, Y.-X.: A modified BFGS algorithm for unconstrained optimization. IMA J. Numer. Anal. **11**(3), 325–332 (1991)
31. Al-Baali, M.: Global and superlinear convergence of a restricted class of self-scaling methods with inexact line searches, for convex functions. Comput. Optim. Appl. **9**(2), 191–203 (1998)
32. Li, D., Fukushima, M.: A globally and superlinearly convergent Gauss-Newton-based BFGS method for symmetric nonlinear equations. SIAM J. Numer. Anal. **37**(1), 152–172 (1999)
33. Yabe, H., Ogasawara, H., Yoshino, M.: Local and superlinear convergence of quasi-Newton methods based on modified secant conditions. J. Comput. Appl. Math. **205**(1), 617–632 (2007)
34. Mokhtari, A., Eisen, M., Ribeiro, A.: IQN: an incremental quasi-Newton method with local superlinear convergence rate. SIAM J. Optim. **28**(2), 1670–1698 (2018)
35. Rodomanov, A., Nesterov, Y.: Greedy quasi-Newton methods with explicit superlinear convergence. SIAM J. Optim. **31**(1), 785–811 (2021)
36. Rodomanov, A., Nesterov, Y.: Rates of superlinear convergence for classical quasi-Newton methods. Math. Program. **194**, 159–190 (2022)
37. Rodomanov, A., Nesterov, Y.: New results on superlinear convergence of classical quasi-Newton methods. J. Optim. Theory Appl. **188**(3), 744–769 (2021)
38. Nesterov, J.E.: "Self-concordant functions and polynomial-time methods in convex programming," Report, Central Economic and Mathematic Institute. USSR Acad, Sci (1989)
39. Nesterov, Y., Nemirovskii, A.: Interior-Point Polynomial Algorithms in Convex Programming. SIAM, New Delhi (1994)
40. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Cell Biol. **96**, 6745–6750 (1999)
41. Blackard, A. Jock., Dean, D.J.: Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. Comput. Electron. Agric. **24**(3), 131–151 (2000)
42. Isabelle, G., Gunn, S., Ben-Hur, A., Dror, G.: Result analysis of the nips: feature selection challenge. Adv. Neural Inf. Process. Syst. **17**, 2005 (2003)
43. LeCun, Y., Cortes, C., Burges, C. J.: MNIST handwritten digit database. AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist/, 2010
44. Chang, C.-C., Lin, C.-J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. doi (2011). https://doi.org/10.1145/1961189.1961199