New Results on Superlinear Convergence of Classical Quasi-Newton Methods *

Anton Rodomanov†　　　Yurii Nesterov‡

April 29, 2020

Abstract

We present a new theoretical analysis of local superlinear convergence of the classical quasi-Newton methods from the convex Broyden class. Our analysis is based on the potential function involving the logarithm of determinant of Hessian approximation and the trace of inverse Hessian approximation. For the well-known DFP and BFGS methods, we obtain the rates of the form $\left[\frac{L}{\mu}\left(\exp\left\{\frac{n}{k}\ln\frac{L}{\mu}\right\}-1\right)\right]^{k/2}$ and $\left[\exp\left\{\frac{n}{k}\ln\frac{L}{\mu}\right\}-1\right]^{k/2}$ respectively, where k is the iteration counter, n is the dimension of the problem, μ is the strong convexity parameter, and L is the Lipschitz constant of the gradient. Currently, these are the best known superlinear convergence rates for these methods. In particular, our results show that the starting moment of superlinear convergence of BFGS method depends on the *logarithm* of the condition number $\frac{L}{\mu}$ in the worst case.

Keywords: quasi-Newton methods, convex Broyden class, DFP, BFGS, superlinear convergence, local convergence, rate of convergence.

*The research results of this paper were obtained with support of ERC Advanced Grant 788368.

†Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Catholic University of Louvain (UCL). E-mail: anton.rodomanov@uclouvain.be.

‡Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL). E-mail: yurii.nesterov@uclouvain.be.

1 Introduction

Motivation. We study local superlinear convergence of classical quasi-Newton methods for smooth unconstrained optimization. These algorithms can be seen as an approximation of the standard Newton method, in which the exact Hessian is replaced by some operator, that is updated in iterations by using the gradients of the objective function. The two most famous examples of quasi-Newton algorithms are the *Davidon–Fletcher–Powell (DFP)* method [7,13] and the *Broyden–Fletcher–Goldfarb–Shanno (BFGS)* method [2,3,14,16, 25], which together belong to a more general Broyden family [1] of quasi-Newton algorithms. For an introduction into the topic, see [9] and [21, Chapter 6]. See also [19] for the discussion of quasi-Newton algorithms in the context of nonsmooth optimization.

The superlinear convergence of quasi-Newton methods was established as early as in 1970s, firstly by Powell [22] and Dixon [10,11] for the methods with exact line search, and then by Broyden, Dennis and Moré [4] and Dennis and Moré [8] for the methods without line search. The latter two approaches have been extended onto more general methods under various settings (see e.g. [6,12,15,17,20,26–29]).

However, explicit *rates* of superlinear convergence for quasi-Newton algorithms have been obtained only recently. The first results were presented in [23] for the *greedy* quasi-Newton methods. These algorithms are based on the updating formulas from the Broyden family, and use greedily chosen basis vectors as the updating directions. The superlinear convergence rate of the greedy quasi-Newton methods has the form $\left(1 - \frac{\mu}{nL}\right)^{k^2/2} \left(\frac{nL}{\mu}\right)^k$, where k is the iteration counter, n is the dimension of the problem, μ is the strong convexity parameter, and L is the Lipschitz constant of the gradient.

After that, in [24], the *classical* quasi-Newton methods were considered, for which the authors established the superlinear convergence rates of the form $\left(\frac{nL^2}{\mu^2 k}\right)^{k/2}$ and $\left(\frac{nL}{\mu k}\right)^{k/2}$ for DFP and BFGS respectively. The analysis was based on the trace potential function, which was then augmented by the logarithm of determinant of the *inverse Hessian* approximation to extend the proof onto the general nonlinear case.

In this paper, we provide a further development of the results of [24]. In particular, for DFP and BFGS methods, we establish new superlinear convergence rates of the form $\left[\frac{L}{\mu}\left(\exp\left\{\frac{n}{k}\ln\frac{L}{\mu}\right\} - 1\right)\right]^{k/2}$ and $\left[\exp\left\{\frac{n}{k}\ln\frac{L}{\mu}\right\} - 1\right]^{k/2}$ respectively. Interestingly, according to our results, the starting moment of superlinear convergence of BFGS method has a *logarithmic* dependency on the condition number $\frac{L}{\mu}$. As compared to the previous work, the main difference in the analysis is the choice of the potential function: now the main part is formed by the logarithm of determinant of Hessian approximation, which is then augmented by the trace of *inverse Hessian* approximation to extend the proof onto the general nonlinear case.

It is worth noting that recently, in [18], another independent analysis of superlinear convergence of the classical DFP and BFGS methods was presented with the resulting rate $\left(\frac{1}{k}\right)^{k/2}$ for both methods. Note that this rate does not depend on any of the constants n, μ and L. However, to obtain it, the authors had to make an additional assumption that the methods start from a sufficiently good initial Hessian approximation. Without this assumption, to our knowledge, their proof technique, based on the Frobenius-norm potential function, leads only to the rate $\left(\frac{nL^2}{\mu^2 k}\right)^{k/2}$ for both DFP and BFGS, which is weaker than the corresponding rates in [24].

Contents. This paper is organized as follows. In Section 2, we study the convex Broyden class of quasi-Newton updates for approximating a self-adjoint positive definite operator. We introduce a certain measure of closeness of quasi-Newton approximations to the target operator along the updating directions, and relate this measure to the improvement in two potential functions: the log-det barrier and the augmented log-det barrier. We also show that the introduced measure is an upper bound for another measure, where the metrics are taken with respect to the successive quasi-Newton approximations.

In Section 3, we analyze the rate of convergence of the classical quasi-Newton methods from the convex Broyden class as applied to minimizing a quadratic function. On this simple example, where the Hessian is constant, we illustrate the main ideas of our analysis, using the both potential functions.

In Section 4, we consider the general unconstrained optimization problem. Assuming that the initial point is sufficiently close to the solution, we establish the same convergence rates as in the quadratic case, up to some absolute constants.

Finally, in Section 5, we explain why the new superlinear convergence rates, that we have obtained in this paper, are better than the previously known ones, and discuss some open questions.

Notation. In what follows, $\mathbb{E}$ denotes an arbitrary n-dimensional real vector space. Its dual space, composed of all linear functionals on $\mathbb{E}$, is denoted by $\mathbb{E}^*$. The value of a linear function $s \in \mathbb{E}^*$, evaluated at a point $x \in \mathbb{E}$, is denoted by $\langle s, x \rangle$.

For a smooth function $f : \mathbb{E} \to \mathbb{R}$, we denote by $\nabla f(x)$ and $\nabla^2 f(x)$ its gradient and Hessian respectively, evaluated at a point $x \in \mathbb{E}$. Note that $\nabla f(x) \in \mathbb{E}^*$, and $\nabla^2 f(x)$ is a self-adjoint linear operator from $\mathbb{E}$ to $\mathbb{E}^*$.

The partial ordering of self-adjoint linear operators is defined in the standard way. We write $A \preceq A_1$ for $A, A_1 : \mathbb{E} \to \mathbb{E}^*$ if $\langle (A_1 - A)x, x \rangle \geq 0$ for all $x \in \mathbb{E}$, and $W \preceq W_1$ for $W, W_1 : \mathbb{E}^* \to \mathbb{E}$ if $\langle s, (W_1 - W)s \rangle \geq 0$ for all $s \in \mathbb{E}^*$.

Any self-adjoint positive definite linear operator $A : \mathbb{E} \to \mathbb{E}^*$ induces in the spaces $\mathbb{E}$ and $\mathbb{E}^*$ the following pair of conjugate Euclidean norms:

$$\|h\|_A \stackrel{\mathrm{def}}{=} \langle Ah, h \rangle^{1/2}, \quad h \in \mathbb{E}, \qquad \|s\|_A^* \stackrel{\mathrm{def}}{=} \langle s, A^{-1} s \rangle^{1/2}, \quad s \in \mathbb{E}^*. \tag{1.1}$$

When $A = \nabla^2 f(x)$, where $f : \mathbb{E} \to \mathbb{R}$ is a smooth function with positive definite Hessian, and $x \in \mathbb{E}$, we prefer to use notation $\|\cdot\|_x$ and $\|\cdot\|_x^*$, provided that there is no ambiguity with the reference function f.

Sometimes, in the formulas, involving products of linear operators, it is convenient to treat $x \in \mathbb{E}$ as a linear operator from $\mathbb{R}$ to $\mathbb{E}$, defined by $x\alpha = \alpha x$, and x^* as a linear operator from $\mathbb{E}^*$ to $\mathbb{R}$, defined by $x^* s = \langle s, x \rangle$. Likewise, any $s \in \mathbb{E}^*$ can be treated as a linear operator from $\mathbb{R}$ to $\mathbb{E}^*$, defined by $s\alpha = \alpha s$, and s^* as a linear operator from $\mathbb{E}$ to $\mathbb{R}$, defined by $s^* x = \langle s, x \rangle$. In this case, xx^* and ss^* are rank-one self-adjoint linear operators from $\mathbb{E}^*$ to $\mathbb{E}$ and from $\mathbb{E}^*$ to $\mathbb{E}$ respectively, acting as follows:

$$(xx^*)s = \langle s, x \rangle x, \qquad (ss^*)x = \langle s, x \rangle s, \qquad x \in \mathbb{E},\ s \in \mathbb{E}^*.$$

Given two self-adjoint linear operators $A : \mathbb{E} \to \mathbb{E}^*$ and $W : \mathbb{E}^* \to \mathbb{E}$, we define the trace and the determinant of A with respect to W as follows:

$$\langle W, A \rangle \stackrel{\mathrm{def}}{=} \mathrm{Tr}(WA), \qquad \mathrm{Det}(W, A) \stackrel{\mathrm{def}}{=} \mathrm{Det}(WA).$$

Note that WA is a linear operator from $\mathbb{E}$ to itself, and hence its trace and determinant are well-defined by the eigenvalues (they coincide with the trace and determinant of the matrix representation of WA with respect to an arbitrary chosen basis in the space $\mathbb{E}$, and the result is independent of the particular choice of the basis). In particular, if W is positive definite, then $\langle W, A\rangle$ and $\mathrm{Det}(W, A)$ are respectively the sum and the product of the eigenvalues of A relative to W^{-1}. Observe that $\langle\cdot,\cdot\rangle$ is a bilinear form, and for any $x \in \mathbb{E}$, we have

$$\langle Ax, x\rangle \;=\; \langle xx^*, A\rangle.$$

When A is invertible, we also have

$$\langle A^{-1}, A\rangle \;=\; n, \qquad \mathrm{Det}(A^{-1}, \delta A) \;=\; \delta^n. \tag{1.2}$$

for any $\delta \in \mathbb{R}$. Also recall the following multiplicative formula for the determinant:

$$\mathrm{Det}(W, A) \;=\; \mathrm{Det}(W, G) \cdot \mathrm{Det}(G^{-1}, A), \tag{1.3}$$

which is valid for any invertible linear operator $G : \mathbb{E} \to \mathbb{E}^*$. If the operator W is positive semidefinite, and $A \preceq A_1$ for some self-adjoint linear operator $A_1 : \mathbb{E} \to \mathbb{E}^*$, then $\langle W, A\rangle \le \langle W, A_1\rangle$ and $\mathrm{Det}(W, A) \le \mathrm{Det}(W, A_1)$. Similarly, if A is positive semidefinite and $W \preceq W_1$ for some self-adjoint linear operator $W_1 : \mathbb{E}^* \to \mathbb{E}$, then $\langle W, A\rangle \le \langle W_1, A\rangle$ and $\mathrm{Det}(W, A) \le \mathrm{Det}(W_1, A)$.

2 Convex Broyden Class

Let A and G be two self-adjoint positive definite linear operators from $\mathbb{E}$ to $\mathbb{E}^*$, where A is the target operator, that we want to approximate, and G is its current approximation. The *Broyden class* of quasi-Newton updates of G with respect to A along a direction $u \in \mathbb{E} \setminus \{0\}$ is the following family of updating formulas, parameterized by a scalar $\tau \in \mathbb{R}$:

$$\begin{aligned} \mathrm{Broyd}_\tau(A, G, u) \;&=\; \phi_\tau \left[G - \tfrac{Auu^*G + Guu^*A}{\langle Au, u\rangle} + \left(\tfrac{\langle Gu, u\rangle}{\langle Au, u\rangle} + 1 \right) \tfrac{Auu^*A}{\langle Au, u\rangle} \right] \\ &\quad + (1 - \phi_\tau) \left[G - \tfrac{Guu^*G}{\langle Gu, u\rangle} + \tfrac{Auu^*A}{\langle Au, u\rangle} \right], \end{aligned} \tag{2.1}$$

where

$$\phi_\tau \;\stackrel{\text{def}}{=}\; \phi_\tau(A, G, u) \;\stackrel{\text{def}}{=}\; \frac{\tau \frac{\langle Au, u\rangle}{\langle AG^{-1}Au, u\rangle}}{\tau \frac{\langle Au, u\rangle}{\langle AG^{-1}Au, u\rangle} + (1-\tau) \frac{\langle Gu, u\rangle}{\langle Au, u\rangle}}. \tag{2.2}$$

If the denominator in (2.2) equals zero, we left both ϕ_τ and $\mathrm{Broyd}_\tau(A, G, u)$ undefined. For the sake of convenience, we also set $\mathrm{Broyd}_\tau(A, G, u) = G$ for $u = 0$.

In this paper, we will be interested in the *convex* Broyden class, which is described by the values of $\tau \in [0, 1]$. Note that for all such τ the denominator in (2.2) is always strictly positive for any $u \neq 0$, so both ϕ_τ and $\mathrm{Broyd}_\tau(A, G, u)$ are well-defined; moreover, $\phi_\tau \in [0, 1]$. For $\tau = 1$, we have $\phi_\tau = 1$, and (2.1) becomes the DFP update; for $\tau = 0$, we have $\phi_\tau = 0$, and (2.1) becomes the BFGS update.

Remark 2.1 *Usually the Broyden class is defined directly in terms of the parameter ϕ. However, in the context of this paper, it turns out to be more convenient to work in terms of τ instead of ϕ. As can be seen from* (A.1)*, the parameter τ is exactly the weight of the DFP component in the updating formula for the inverse operator.*

One of the basic properties of the convex Broyden class is that each update from this class preserves the bounds on the relative eigenvalues with respect to the target operator.

Lemma 2.1 (see [24, Lemma 2.1]) *If $\frac{1}{\xi}A \preceq G \preceq \eta A$ for some $\xi, \eta \geq 1$, then, for any $u \in \mathbb{E}$, and any $\tau \in [0,1]$, we have $\frac{1}{\xi}A \preceq \mathrm{Broyd}_\tau(A,G,u) \preceq \eta A$.*

Define the following measure of closeness of G to A along direction $u \in \mathbb{E} \setminus \{0\}$:

$$\nu(A,G,u) \overset{\text{def}}{=} \frac{\langle (G-A)G^{-1}(G-A)u,u\rangle^{1/2}}{\langle Au,u\rangle^{1/2}} \overset{(1.1)}{=} \frac{\|(G-A)u\|_G^*}{\|u\|_A}. \tag{2.3}$$

Let us present two potential functions, whose improvement after one update from the convex Broyden class can be bounded from below by a certain non-negative monotonically increasing function of ν, vanishing at zero.

First, consider the *log-det barrier*:

$$V(A,G) = \ln \mathrm{Det}(A^{-1},G). \tag{2.4}$$

We will use this potential function only when $A \preceq G$. Note that in this case $V(A,G) \geq 0$.

Lemma 2.2 *Let $A, G : \mathbb{E} \to \mathbb{E}^*$ be self-adjoint positive definite linear operators such that*

$$A \preceq G \preceq \eta A \tag{2.5}$$

for some $\eta \geq 1$. Then, for any $\tau \in [0,1]$ and any $u \in \mathbb{E} \setminus \{0\}$, we have

$$V(A,G) - V(A,\mathrm{Broyd}_\tau(A,G,u)) \geq \ln\Big(1 + (\tau\tfrac{1}{\eta} + 1 - \tau)\nu^2(A,G,u)\Big).$$

Proof: Indeed, denoting $G_+ \overset{\text{def}}{=} \mathrm{Broyd}_\tau(A,G,u)$, we obtain

$$\begin{aligned} V(A,G) - V(A,G_+) &\overset{(2.4)}{=} \ln \mathrm{Det}(A^{-1},G) - \ln \mathrm{Det}(A^{-1},G_+) \\ &\overset{(1.3)}{=} \ln \mathrm{Det}(G_+^{-1},G) \\ &\overset{(A.2)}{=} \ln\Big(\tau\tfrac{\langle Au,u\rangle}{\langle AG^{-1}Au,u\rangle} + (1-\tau)\tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle}\Big) \\ &= \ln\Big(1 + \tau\tfrac{\langle A(A^{-1}-G^{-1})Au,u\rangle}{\langle AG^{-1}Au,u\rangle} + (1-\tau)\tfrac{\langle (G-A)u,u\rangle}{\langle Au,u\rangle}\Big). \end{aligned} \tag{2.6}$$

In view of (2.5), we have $0 \preceq G - A \preceq (1-\frac{1}{\eta})G$. Hence[1],

$$(G-A)G^{-1}(G-A) \preceq \Big(1-\tfrac{1}{\eta}\Big)(G-A) \preceq \tfrac{1}{1+\frac{1}{\eta}}(G-A) \preceq G - A. \tag{2.7}$$

[1]This is obvious when $G - A$ is non-degenerate. The general case then follows by continuity.

At the same time, we have the following identity:

$$A(A^{-1} - G^{-1})A \quad = \quad G - A - (G-A)G^{-1}(G-A). \tag{2.8}$$

Therefore, denoting $\nu \stackrel{\mathrm{def}}{=} \nu(A, G, u)$, we can write

$$\frac{\langle (G-A)u,u\rangle}{\langle Au,u\rangle} \quad \stackrel{(2.7)}{\geq} \quad \frac{\langle (G-A)G^{-1}(G-A)u,u\rangle}{\langle Au,u\rangle} \quad \stackrel{(2.3)}{=} \quad \nu^2,$$

and

$$\begin{aligned} \frac{\langle A(A^{-1}-G^{-1})Au,u\rangle}{\langle AG^{-1}Au,u\rangle} \quad &\stackrel{(2.8)}{=} \quad \frac{\langle (G-A-(G-A)G^{-1}(G-A))u,u\rangle}{\langle AG^{-1}Au,u\rangle} \quad \stackrel{(2.7)}{\geq} \quad \frac{1}{\eta} \frac{\langle (G-A)G^{-1}(G-A)u,u\rangle}{\langle AG^{-1}Au,u\rangle} \\ &\stackrel{(2.5)}{\geq} \quad \frac{1}{\eta} \frac{\langle (G-A)G^{-1}(G-A)u,u\rangle}{\langle Au,u\rangle} \quad \stackrel{(2.3)}{=} \quad \frac{1}{\eta}\nu^2. \end{aligned}$$

Substituting the above two inequalities into (2.6), we obtain the claim. □

Now consider another potential function, the *augmented log-det barrier*:

$$\psi(G, A) \quad \stackrel{\mathrm{def}}{=} \quad \ln \mathrm{Det}(A^{-1}, G) - \langle G^{-1}, G - A\rangle. \tag{2.9}$$

As compared to the log-det barrier, this potential function is more universal since it works even if the condition $A \preceq G$ is violated. Note that the augmented log-det barrier is in fact the Bregman divergence, generated by the strictly convex function $d(A) \stackrel{\mathrm{def}}{=} -\ln \mathrm{Det}(B^{-1}, A)$, defined on the set of self-adjoint positive definite linear operators from $\mathbb{E}$ to $\mathbb{E}^*$, where $B : \mathbb{E} \to \mathbb{E}^*$ is an arbitrary fixed self-adjoint positive definite linear operator. Indeed,

$$\begin{aligned} \psi(G, A) \quad &\stackrel{(1.3)}{=} \quad -\ln \mathrm{Det}(B^{-1}, A) + \ln \mathrm{Det}(B^{-1}, G) - \langle -G^{-1}, A - G\rangle \\ &= \quad d(A) - d(G) - \langle \nabla d(G), A - G\rangle \;\geq\; 0. \end{aligned} \tag{2.10}$$

Remark 2.2 *The idea of combining together the trace and the logarithm of the determinant to form a potential function for the analysis of quasi-Newton methods, can be traced back to the work [5]. Note also that in [24], the authors studied the evolution of* $\psi(A, G)$*, i.e. the Bregman divergence was centered at* A *instead of* G*.*

Let us establish an auxiliary inequality.

Lemma 2.3 *For any real* $\alpha \geq \beta > 0$*, we have* $\alpha + \frac{1}{\beta} - 1 \geq 1$*, and*

$$\alpha - \ln \beta - 1 \quad \geq \quad \frac{\sqrt{3}}{2+\sqrt{3}} \ln\left(\alpha + \frac{1}{\beta} - 1\right) \quad \geq \quad \frac{6}{13} \ln\left(\alpha + \frac{1}{\beta} - 1\right). \tag{2.11}$$

Proof: We only need to prove the first inequality in (2.11) since the second one follows from it and the fact that $\frac{\sqrt{3}+2}{\sqrt{3}} = 1 + \frac{2}{\sqrt{3}} \leq 1 + \frac{7}{6} = \frac{13}{6}$ (since $2 \leq \frac{7}{2\sqrt{3}}$).

Let $\beta > 0$ be fixed, and let $\zeta_1 : (1 - \frac{1}{\beta}, +\infty) \to \mathbb{R}$ be the function

$$\zeta_1(\alpha) \stackrel{\text{def}}{=} \alpha - \tfrac{\sqrt{3}}{2+\sqrt{3}} \ln\left(\alpha + \tfrac{1}{\beta} - 1\right). \tag{2.12}$$

Note that the domain of ζ_1 includes the point $\alpha = \beta$ since $\beta \geq 2 - \frac{1}{\beta} > 1 - \frac{1}{\beta}$. Let us show that ζ_1 is an increasing function on the interval $[\beta, +\infty)$. Indeed, for any $\alpha \geq \beta$, we have

$$\zeta_1'(\alpha) \stackrel{(2.12)}{=} 1 - \tfrac{\sqrt{3}}{2+\sqrt{3}} \tfrac{1}{\alpha + \frac{1}{\beta} - 1} \;\geq\; 1 - \tfrac{1}{\alpha + \frac{1}{\beta} - 1} \;=\; \tfrac{\alpha + \frac{1}{\beta} - 2}{\alpha + \frac{1}{\beta} - 1} \;\geq\; \tfrac{\beta + \frac{1}{\beta} - 2}{\alpha + \frac{1}{\beta} - 1} \;\geq\; 0.$$

Thus, it suffices to prove (2.11) only in the case when $\alpha = \beta$, or, equivalently, to show that the function $\zeta_2 : (0, +\infty) \to \mathbb{R}$, defined by

$$\zeta_2(\alpha) \stackrel{\text{def}}{=} \alpha - \ln\alpha - 1 - \tfrac{\sqrt{3}}{2+\sqrt{3}} \ln\left(\alpha + \tfrac{1}{\alpha} - 1\right), \tag{2.13}$$

is non-negative. Differentiating, we find that, for all $\alpha > 0$, we have

$$\begin{aligned}
\zeta_2'(\alpha) &\stackrel{(2.13)}{=} 1 - \tfrac{1}{\alpha} - \tfrac{\sqrt{3}}{2+\sqrt{3}} \tfrac{1 - \frac{1}{\alpha^2}}{\alpha + \frac{1}{\alpha} - 1} \;=\; \left(1 - \tfrac{1}{\alpha}\right)\left(1 - \tfrac{\sqrt{3}}{2+\sqrt{3}} \tfrac{1 + \frac{1}{\alpha}}{\alpha + \frac{1}{\alpha} - 1}\right) \\
&= \left(1 - \tfrac{1}{\alpha}\right) \tfrac{\alpha + \frac{1}{\alpha} - 1 - (2\sqrt{3} - 3)(1 + \frac{1}{\alpha})}{\alpha + \frac{1}{\alpha} - 1} \;=\; \left(1 - \tfrac{1}{\alpha}\right) \tfrac{\alpha - 2(\sqrt{3}-1) + (\sqrt{3}-1)^2 \frac{1}{\alpha}}{1 + \frac{1}{\alpha} - 1} \\
&= \left(1 - \tfrac{1}{\alpha}\right) \tfrac{(\sqrt{\alpha} - (\sqrt{3}-1)\frac{1}{\sqrt{\alpha}})^2}{\alpha + \frac{1}{\alpha} - 1}.
\end{aligned}$$

Hence, $\zeta_2'(\alpha) \leq 0$ for $0 < \alpha \leq 1$, and $\zeta_2'(\alpha) \geq 0$ for $\alpha \geq 1$. Thus, the minimum of ζ_2 is attained at $\alpha = 1$. Consequently, $\zeta_2(\alpha) \geq \zeta_2(1) = 0$ for all $\alpha > 0$. □

It turns out that, up to some constants, the improvement in the augmented log-det barrier can be lower bounded exactly by the same logarithmic function of ν, that we used for the simple log-det barrier.

Lemma 2.4 *Let $A, G : \mathbb{E} \to \mathbb{E}^*$ be self-adjoint positive definite linear operators such that*

$$\tfrac{1}{\xi} A \;\preceq\; G \;\preceq\; \eta A \tag{2.14}$$

for some $\xi, \eta \geq 1$. Then, for any $\tau \in [0, 1]$ and any $u \in \mathbb{E} \setminus \{0\}$, we have

$$\psi(G, A) - \psi(\mathrm{Broyd}_\tau(A, G, u), A) \;\geq\; \tfrac{6}{13} \ln\left(1 + (\tau \tfrac{1}{\xi\eta} + 1 - \tau)\nu^2(A, G, u)\right).$$

Proof: Indeed, denoting $G_+ \stackrel{\text{def}}{=} \mathrm{Broyd}_\tau(A, G, u)$, we obtain

$$\langle G^{-1} - G_+^{-1}, A\rangle \stackrel{(A.1)}{=} \tau\left[\tfrac{\langle AG^{-1}AG^{-1}Au, u\rangle}{\langle AG^{-1}Au, u\rangle} - 1\right] + (1 - \tau)\left[\tfrac{\langle AG^{-1}Au, u\rangle}{\langle Au, u\rangle} - 1\right],$$

and

$$\mathrm{Det}(G_+^{-1}, G) \stackrel{(A.2)}{=} \tau \tfrac{\langle Au, u\rangle}{\langle AG^{-1}Au, u\rangle} + (1 - \tau) \tfrac{\langle Au, u\rangle}{\langle Gu, u\rangle}.$$

Thus,

$$
\begin{aligned}
\psi(G,A) - \psi(G_+,A) &\overset{(2.9)}{=} \langle G^{-1} - G_+^{-1}, A\rangle + \ln \mathrm{Det}(G_+^{-1}, G) \\
&= \tau\alpha_1 + (1-\tau)\alpha_0 + \ln(\tau\beta_1^{-1} + (1-\tau)\beta_0^{-1}) - 1 \qquad (2.15) \\
&= \alpha - \ln\beta - 1,
\end{aligned}
$$

where

$$
\begin{array}{rclrcl}
\alpha_1 & \overset{\mathrm{def}}{=} & \frac{\langle AG^{-1}AG^{-1}Au,u\rangle}{\langle AG^{-1}Au,u\rangle}, & \beta_1 & \overset{\mathrm{def}}{=} & \frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle}, \\
\alpha_0 & \overset{\mathrm{def}}{=} & \frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle}, & \beta_0 & \overset{\mathrm{def}}{=} & \frac{\langle Au,u\rangle}{\langle Gu,u\rangle}, \\
\alpha & \overset{\mathrm{def}}{=} & \tau\alpha_1 + (1-\tau)\alpha_0, & \beta & \overset{\mathrm{def}}{=} & (\tau\beta_1^{-1} + (1-\tau)\beta_0^{-1})^{-1}.
\end{array} \qquad (2.16)
$$

Note that $\alpha_1 \geq \beta_1$ and $\alpha_0 \geq \beta_0$ in view of the Cauchy-Schwartz inequality. At the same time, $\tau\beta_1 + (1-\tau)\beta_2 \geq \beta$ by the convexity of the inverse function $t \mapsto t^{-1}$. Hence, we can apply Lemma 2.3 to estimate (2.15) from below. Note that

$$
\begin{aligned}
\alpha + \tfrac{1}{\beta} - 1 &\overset{(2.16)}{=} \tau\tfrac{\langle (A+AG^{-1}AG^{-1}A)u,u\rangle}{\langle AG^{-1}Au,u\rangle} + (1-\tau)\tfrac{\langle (G+A)u,u\rangle}{\langle Au,u\rangle} - 1 \\
&= 1 + \tau\tfrac{\langle (G-A)G^{-1}AG^{-1}(G-A)\rangle}{\langle AG^{-1}Au,u\rangle} + (1-\tau)\tfrac{\langle (G-A)G^{-1}(G-A)u,u\rangle}{\langle Au,u\rangle} \\
&\overset{(2.14)}{\geq} 1 + (\tau\tfrac{1}{\xi\eta} + 1 - \tau)\tfrac{\langle (G-A)G^{-1}(G-A)u,u\rangle}{\langle Au,u\rangle} \\
&\overset{(2.3)}{=} 1 + (\tau\tfrac{1}{\xi\eta} + 1 - \tau)\nu^2(A,G,u). \qquad \square
\end{aligned}
$$

The measure $\nu(A,G,u)$, defined in (2.3), is the ratio of the norm of $(G-A)u$, measured with respect to G, and the norm of u, measured with respect to A. It is important that we can change the corresponding metrics to G_+ and G respectively by paying only with the minimal eigenvalue of G relative to A.

Lemma 2.5 *Let $A, G : \mathbb{E} \to \mathbb{E}^*$ be self-adjoint positive definite linear operators such that*

$$
\tfrac{1}{\xi}A \preceq G \qquad (2.17)
$$

for some $\xi > 0$. Then, for any $\tau \in [0,1]$, any $u \in \mathbb{E} \setminus \{0\}$, and $G_+ \overset{\mathrm{def}}{=} \mathrm{Broyd}_\tau(A,G,u)$,

$$
\nu^2(A,G,u) \geq \tfrac{1}{1+\xi}\tfrac{\langle (G-A)G_+^{-1}(G-A)u,u\rangle}{\langle Gu,u\rangle}.
$$

Proof: From (A.1), it is easy to see that $G_+^{-1}Au = u$. Hence,

$$
\begin{aligned}
\tfrac{\langle (G-A)G_+^{-1}(G-A)u,u\rangle}{\langle Gu,u\rangle} &= \tfrac{\langle GG_+^{-1}Gu,u\rangle}{\langle Gu,u\rangle} + \tfrac{\langle Au, G_+^{-1}Au\rangle}{\langle Gu,u\rangle} - 2\tfrac{\langle Gu, G_+^{-1}Au\rangle}{\langle Gu,u\rangle} \\
&= \tfrac{\langle GG_+^{-1}Gu,u\rangle}{\langle Gu,u\rangle} + \tfrac{\langle Au,u\rangle}{\langle Gu,u\rangle} - 2.
\end{aligned} \qquad (2.18)
$$

Since $1 - t \leq \frac{1}{t} - 1$ for all $t > 0$, we further have

$$
\begin{aligned}
\tfrac{\langle GG_+^{-1}Gu,u\rangle}{\langle Gu,u\rangle} &\overset{(A.1)}{=} \tau\left[1 - \tfrac{\langle Au,u\rangle^2}{\langle Gu,u\rangle\langle AG^{-1}Au,u\rangle} + \tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle}\right] \\
&\quad + (1-\tau)\left[\left(\tfrac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} + 1\right)\tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle} - 1\right] \qquad (2.19) \\
&\leq \left(\tfrac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} + 1\right)\tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle} - 1.
\end{aligned}
$$

Denote $\nu \overset{\text{def}}{=} \nu(A, G, u)$. Then,

$$\nu^2 \overset{(2.3)}{=} \frac{\langle (G-A)G^{-1}(G-A)u,u\rangle}{\langle Au,u\rangle} = \frac{\langle Gu,u\rangle}{\langle Au,u\rangle} + \frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} - 2. \tag{2.20}$$

Consequently,

$$\begin{aligned}(1+\xi)\nu^2 &\overset{(2.17)}{\geq} \left(\frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} + 1\right)\nu^2 \\ &\overset{(2.20)}{=} \left(\frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} + 1\right)\frac{\langle Gu,u\rangle}{\langle Au,u\rangle} + \frac{\langle AG^{-1}Au,u\rangle^2}{\langle Au,u\rangle^2} - \frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} - 2 \\ &\overset{(2.19)}{\geq} \frac{\langle GG_+^{-1}Gu,u\rangle}{\langle Au,u\rangle} + \frac{\langle AG^{-1}Au,u\rangle^2}{\langle Au,u\rangle} - \frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} - 1.\end{aligned} \tag{2.21}$$

Thus,

$$\begin{aligned}(1+\xi)\nu^2 - \frac{\langle (G-A)G_+^{-1}(G-A)u,u\rangle}{\langle Gu,u\rangle} &\overset{(2.18)}{=} (1+\xi)\nu^2 - \frac{\langle GG_+^{-1}Gu,u\rangle}{Gu,u\rangle} - \frac{\langle Au,u\rangle}{\langle Gu,u\rangle} + 2 \\ &\overset{(2.21)}{\geq} \frac{\langle AG^{-1}Au,u\rangle^2}{\langle Au,u\rangle^2} - \frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} - \frac{\langle Au,u\rangle}{\langle Gu,u\rangle} + 1 \\ &\geq \frac{\langle AG^{-1}Au,u\rangle^2}{\langle Au,u\rangle^2} - 2\frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} + 1 \;\geq\; 0,\end{aligned}$$

where we have applied the Cauchy–Schwartz inequality $\frac{\langle Au,u\rangle}{\langle Gu,u\rangle} \leq \frac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle}$. □

3 Unconstrained Quadratic Minimization

Let us study the convergence properties of the classical quasi-Newton methods from the convex Broyden class, as applied to minimizing the quadratic function

$$f(x) \overset{\text{def}}{=} \tfrac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle, \tag{3.1}$$

where $A : \mathbb{E} \to \mathbb{E}^*$ is a self-adjoint positive definite linear operator, and $b \in \mathbb{E}^*$.

Let $B : \mathbb{E} \to \mathbb{E}^*$ be a fixed self-adjoint positive definite linear operator, and let $\mu, L > 0$ be such that

$$\mu B \preceq A \preceq LB. \tag{3.2}$$

Thus, μ is the *strong convexity* parameter of f, and L is the constant of *Lipschitz continuity* of the gradient of f, both measured relative to B.

Consider the following standard quasi-Newton process for minimizing (3.1):

(3.3)
Initialization: Choose $x_0 \in \mathbb{E}$. Set $G_0 = LB$.
For $k \geq 0$ **iterate:** 1. Update $x_{k+1} = x_k - G_k^{-1}\nabla f(x_k)$. 2. Set $u_k = x_{k+1} - x_k$ and choose $\tau_k \in [0, 1]$. 3. Compute $G_{k+1} = \text{Broyd}_{\tau_k}(A, G_k, u_k)$.

For measuring its rate of convergence, we use the norm of the gradient, taken with respect to the Hessian:

$$\lambda_k \stackrel{\mathrm{def}}{=} \|\nabla f(x_k)\|_A^* \stackrel{(1.1)}{=} \langle \nabla f(x_k), A^{-1}\nabla f(x_k)\rangle^{1/2}.$$

It is known that the process (3.3) has at least a linear convergence rate of the standard gradient method:

Theorem 3.1 (see [24, Theorem 3.1]) *In scheme* (3.3), *for all* $k \geq 0$, *we have*

$$A \preceq G_k \preceq \tfrac{L}{\mu} A, \tag{3.4}$$

$$\lambda_k \leq \left(1 - \tfrac{\mu}{L}\right)^k \lambda_0. \tag{3.5}$$

Let us establish the superlinear convergence. According to (3.4), for the quadratic function, we have $A \preceq G_k$ for all $k \geq 0$. Therefore, in our analysis, we can use both potential functions: the log-det barrier and the augmented log-det barrier. Let us study both variants. We start with the first one.

Theorem 3.2 *In scheme* (3.3), *for all* $k \geq 1$, *we have*

$$\lambda_k \leq \left[\frac{2}{\prod_{i=0}^{k-1}(\tau_i \frac{\mu}{L} + 1 - \tau_i)^{1/k}} \left(e^{\frac{n}{k}\ln\frac{L}{\mu}} - 1\right)\right]^{k/2} \sqrt{\tfrac{L}{\mu}} \cdot \lambda_k. \tag{3.6}$$

Proof: Let $k \geq 1$ be arbitrary. Without loss of generality, we can assume that $u_i \neq 0$ for all $0 \leq i \leq k$. Denote $V_i \stackrel{\mathrm{def}}{=} V(A, G_i)$, $\nu_i \stackrel{\mathrm{def}}{=} \nu(A, G_i, u_i)$, $p_i \stackrel{\mathrm{def}}{=} \tau_i \frac{\mu}{L} + 1 - \tau_i$, and $g_i \stackrel{\mathrm{def}}{=} \|\nabla f(x_i)\|_{G_i}^*$ for any $0 \leq i \leq k$. By Lemma 2.2 and (3.4), we have

$$\ln(1 + p_i \nu_i^2) \leq V_i - V_{i+1}$$

for all $0 \leq i \leq k-1$. Summing up these inequalities, we obtain

$$\begin{aligned} \sum_{i=0}^{k-1} \ln(1 + p_k \nu_k^2) &\leq V_0 - V_k \stackrel{(3.4)}{\leq} V_0 \stackrel{(3.3)}{=} V(A, LB) \stackrel{(2.4)}{=} \ln \mathrm{Det}(A^{-1}, LB) \\ &\stackrel{(3.2)}{\leq} \ln \mathrm{Det}(\tfrac{1}{\mu} B^{-1}, LB) = n \ln \tfrac{L}{\mu}. \end{aligned} \tag{3.7}$$

Hence, by convexity of function $t \mapsto \ln(1 + e^t)$, we get

$$\begin{aligned} \tfrac{n}{k} \ln \tfrac{L}{\mu} &\stackrel{(3.7)}{\geq} \tfrac{1}{k} \sum_{i=0}^{k-1} \ln(1 + p_i \nu_i^2) = \tfrac{1}{k} \sum_{i=0}^{k-1} \ln(1 + e^{\ln(p_i \nu_i^2)}) \\ &\geq \ln\left(1 + e^{\frac{1}{k}\sum_{i=0}^{k-1} \ln(p_i \nu_i^2)}\right) = \ln\left(1 + \left[\prod_{i=0}^{k-1} p_i \nu_i^2\right]^{1/k}\right). \end{aligned} \tag{3.8}$$

At the same time, by Lemma 2.5 and (3.4), for all $0 \leq i \leq k-1$, we have

$$\nu_i^2 \geq \tfrac{1}{2} \frac{\langle (G_i - A) G_{i+1}^{-1} (G_i - A) u_i, u_i \rangle}{\langle G_i u_i, u_i \rangle} = \tfrac{1}{2} \frac{g_{i+1}^2}{g_i^2}$$

since $G_i u_i = -\nabla f(x_i)$ and $A u_i = \nabla f(x_{i+1}) - \nabla f(x_i)$. Consequently, $\prod_{i=0}^{k-1} \nu_i^2 \geq \frac{1}{2^k} \frac{g_k^2}{g_0^2}$. Thus,

$$\frac{n}{k} \ln \frac{L}{\mu} \quad \stackrel{(3.8)}{\geq} \quad \ln \left(1 + \frac{1}{2} \left[\prod_{i=0}^{k-1} p_i \right]^{1/k} \left[\frac{g_k}{g_0} \right]^{2/k} \right).$$

Rearranging, we obtain

$$g_k \quad \leq \quad \left[\frac{2}{\prod_{i=0}^{k-1} p_i^{1/k}} \left(e^{\frac{n}{k} \ln \frac{L}{\mu}} - 1 \right) \right]^{k/2} g_0.$$

It remains to note that $\lambda_k \leq \sqrt{\frac{L}{\mu}} \cdot g_k$ and $g_0 \leq \lambda_0$ in view of (3.4). $\square$

Remark 3.1 *As can be seen from* (3.7)*, the factor* $n \ln \frac{L}{\mu}$ *in the efficiency estimate* (3.6) *can be improved up to* $\ln \mathrm{Det}(A^{-1}, LB) = \sum_{i=1}^{n} \ln \frac{L}{\lambda_i}$*, where* $\lambda_1, \ldots, \lambda_n$ *are the eigenvalues of* A *relative to* B*. This improved factor can be significantly smaller than the original one if the majority of the eigenvalues* λ_i *are much larger than* μ*.*

Now let us briefly present another approach, that is based on the analysis of the *augmented* log-det barrier. The resulting efficiency estimate will be the same as in Theorem 3.2 up to a slightly worse absolute constant under the exponent. However, in contrast to the previous one, this proof can be generalized onto general nonlinear functions.

Theorem 3.3 *In scheme* (3.3)*, for all* $k \geq 0$*, we have*

$$\lambda_k \quad \leq \quad \left[\frac{2}{\prod_{i=0}^{k-1} (\tau_i \frac{\mu}{L} + 1 - \tau_i)^{1/k}} \left(e^{\frac{13}{6} \frac{n}{k} \ln \frac{L}{\mu}} - 1 \right) \right]^{k/2} \sqrt{\frac{L}{\mu}} \cdot \lambda_0.$$

Proof: Let $k \geq 1$ be arbitrary. Without loss of generality, we can assume that $u_i \neq 0$ for all $0 \leq i \leq k$. Denote $\psi_i \stackrel{\mathrm{def}}{=} \psi(G_i, A)$, $\nu_i \stackrel{\mathrm{def}}{=} \nu(A, G_i, u_i)$, $p_i = \tau_i \frac{\mu}{L} + 1 - \tau_i$, and $g_i \stackrel{\mathrm{def}}{=} \|\nabla f(x_i)\|_{G_i}^*$ for all $0 \leq i \leq k$. By Lemma 2.4 and (3.4), we have

$$\frac{6}{13} \ln(1 + p_i \nu_i^2) \quad \leq \quad \psi_i - \psi_{i+1}$$

for all $0 \leq i \leq k-1$. Hence,

$$\begin{aligned} \frac{6}{13} \sum_{i=0}^{k-1} \ln(1 + p_i \nu_i^2) \quad &\leq \quad \psi_0 - \psi_k \stackrel{(2.10)}{\leq} \psi_0 \stackrel{(3.3)}{=} \psi(LB, A) \\ &\stackrel{(2.9)}{=} \ln \mathrm{Det}(A^{-1}, LB) - \langle \tfrac{1}{L} B^{-1}, LB - A \rangle \qquad (3.9) \\ &\stackrel{(3.2)}{\leq} \ln \mathrm{Det}(\tfrac{1}{\mu} B^{-1}, LB) \; = \; n \ln \tfrac{L}{\mu}, \end{aligned}$$

and we can continue exactly in the same manner as in the proof of Theorem 3.2. $\square$

4 Minimization of General Functions

In this section, we consider the general unconstrained minimization problem:

$$\min_{x \in \mathbb{E}} f(x), \tag{4.1}$$

where $f : \mathbb{E} \to \mathbb{R}$ is a twice continuously differentiable function with positive definite second derivative. Our goal is to study the convergence properties of the following standard quasi-Newton scheme for (4.1):

$$\boxed{\begin{array}{l} \textbf{Initialization: } \text{Choose } x_0 \in \mathbb{E}. \text{ Set } G_0 = LB. \\[1em] \textbf{For } k \geq 0 \textbf{ iterate:} \\[1em] 1. \text{ Update } x_{k+1} = x_k - G_k^{-1} \nabla f(x_k). \\[1em] 2. \text{ Set } u_k = x_{k+1} - x_k \text{ and choose } \tau_k \in [0,1]. \\[1em] 3. \text{ Denote } J_k = \int_0^1 \nabla^2 f(x_k + t u_k) dt. \\[1em] 4. \text{ Set } G_{k+1} = \mathrm{Broyd}_{\tau_k}(J_k, G_k, u_k). \end{array}} \tag{4.2}$$

Here $B : \mathbb{E} \to \mathbb{E}^*$ is a fixed self-adjoint positive definite linear operator, and L is a fixed positive constant, which together define the initial Hessian approximation G_0.

In what follows, we assume that there exist constants $\mu > 0$ and $M \geq 0$, such that

$$\mu B \;\preceq\; \nabla^2 f(x) \;\preceq\; LB, \tag{4.3}$$

$$\nabla^2 f(y) - \nabla^2 f(x) \;\preceq\; M \|y - x\|_z \nabla^2 f(w) \tag{4.4}$$

for all $x, y, z, w \in \mathbb{E}$. The first assumption (4.3) specifies that, relative to the operator B, the objective function f is *μ-strongly convex* and its gradient is *L-Lipschitz continuous*. The second assumption (4.4) means that f is *M-strongly self-concordant*. This assumption was recently introduced in [23] as a convenient affine-invariant alternative to the standard assumption of the Lipschitz second derivative, and is satisfied at least for any strongly convex function with Lipschitz continuous Hessian (see [23, Example 4.1]). The main facts, that we will use about strongly self-concordant functions, are summarized in the following lemma (see [23, Lemma 4.1]):

Lemma 4.1 *For any $x, y \in \mathbb{E}$, $J \overset{\text{def}}{=} \int_0^1 \nabla^2 f(x + t(y - x)) dt$, $r \overset{\text{def}}{=} \|y - x\|_x$, we have*

$$\left(1 + \tfrac{Mr}{2}\right)^{-1} \nabla^2 f(x) \;\preceq\; J \;\preceq\; \left(1 + \tfrac{Mr}{2}\right) \nabla^2 f(x), \tag{4.5}$$

$$\left(1 + \tfrac{Mr}{2}\right)^{-1} \nabla^2 f(y) \;\preceq\; J \;\preceq\; \left(1 + \tfrac{Mr}{2}\right) \nabla^2 f(y). \tag{4.6}$$

Note that for a quadratic function, we have $M = 0$, and (4.5), (4.6) become equalities.

Let us analyze the process (4.2). For measuring its rate of convergence, we use the local norm of the gradient:

$$\lambda_k \overset{\text{def}}{=} \|\nabla f(x_k)\|^*_{x_k} \overset{(1.1)}{=} \langle \nabla f(x_k), \nabla^2 f(x_k)^{-1} \nabla f(x_k) \rangle^{1/2}. \tag{4.7}$$

It will also be convenient to introduce the following quantities[2] for all $k \geq 0$:

$$r_k \overset{\text{def}}{=} \|u_k\|_{x_k}, \qquad \xi_k \overset{\text{def}}{=} e^{M \sum_{i=0}^{k-1} r_i} \quad (\geq 1). \tag{4.8}$$

We analyze the process (4.2) in several steps. The first step is to establish the bounds on the relative eigenvalues of the Hessian approximations with respect to the corresponding Hessians.

Lemma 4.2 *For all $k \geq 0$, we have*

$$\tfrac{1}{\xi_k} \nabla^2 f(x_k) \preceq G_k \preceq \xi_k \tfrac{L}{\mu} \nabla^2 f(x_k), \tag{4.9}$$

$$\tfrac{1}{\xi_{k+1}} J_k \preceq G_k \preceq \xi_{k+1} \tfrac{L}{\mu} J_k. \tag{4.10}$$

Proof: For $k = 0$, (4.9) follows from (4.3) and the fact that $G_0 = LB$ while $\xi_0 = 1$.

Now suppose that $k \geq 0$, and that (4.9) has already been proved for all indices up to k. Then, applying Lemma 4.1 to (4.9), we obtain

$$\frac{1}{\xi_k \left(1 + \frac{M r_k}{2}\right)} J_k \preceq G_k \preceq \left(1 + \tfrac{M r_k}{2}\right) \xi_k \tfrac{L}{\mu} J_k. \tag{4.11}$$

This gives us (4.10) since $(1 + \frac{M r_k}{2}) \xi_k \leq \xi_{k+1}$ by the definition of ξ.

Further, applying Lemma 2.1 to (4.11), we obtain

$$\frac{1}{\xi_k \left(1 + \frac{M r_k}{2}\right)} J_k \preceq G_{k+1} \preceq \left(1 + \tfrac{M r_k}{2}\right) \xi_k \tfrac{L}{\mu} J_k.$$

Consequently,

$$\begin{aligned} G_{k+1} &\overset{(4.6)}{\preceq} \left(1 + \tfrac{M r_k}{2}\right)^2 \xi_k \tfrac{L}{\mu} \nabla^2 f(x_{k+1}) \overset{(4.8)}{\preceq} \xi_{k+1} \tfrac{L}{\mu} \nabla^2 f(x_{k+1}), \\ G_{k+1} &\overset{(4.6)}{\succeq} \frac{1}{\left(1 + \frac{M r_k}{2}\right)^2 \xi_k} \nabla^2 f(x_{k+1}) \overset{(4.8)}{\succeq} \tfrac{1}{\xi_{k+1}} \nabla^2 f(x_{k+1}). \end{aligned}$$

Thus, (4.9) is now proved for the next index, $k+1$, so we can continue by induction. □

Lemma 4.2 has the following useful corollary:

[2]We follow the standard convention that the sum over the empty set is defined as 0, so $\xi_0 = 1$. Similarly, the product over the empty set is defined as 1.

Corollary 4.1 *For all $k \geq 0$, we have*

$$r_k \;\leq\; \xi_k \lambda_k. \tag{4.12}$$

Proof: Indeed,

$$\begin{aligned} r_k &\overset{(4.8)}{=} \|u_k\|_{x_k} \overset{(4.2)}{=} \|G_k^{-1}\nabla f(x_k)\|_{x_k} \overset{(1.1)}{=} \langle \nabla f(x_k), G_k^{-1}\nabla^2 f(x_k) G_k^{-1}\nabla f(x_k)\rangle^{1/2} \\ &\overset{(4.9)}{\leq} \xi_k \langle \nabla f(x_k), \nabla^2 f(x_k)^{-1}\nabla f(x_k)\rangle^{1/2} \overset{(4.7)}{=} \xi_k \lambda_k. \qquad \Box \end{aligned}$$

The second step in our analysis is to establish a preliminary version of the linear convergence theorem for the scheme (4.2).

Lemma 4.3 *For all $k \geq 0$, we have*

$$\lambda_k \;\leq\; \sqrt{\xi_k}\lambda_0 \prod_{i=0}^{k-1} q_i, \tag{4.13}$$

where

$$q_i \;\overset{\text{def}}{=}\; \max\left\{1 - \tfrac{\mu}{\xi_{i+1}L}, \xi_{i+1} - 1\right\}. \tag{4.14}$$

Proof: Let $k, i \geq 0$ be arbitrary. By Taylor's formula, we have

$$\nabla f(x_{i+1}) \;\overset{(4.2)}{=}\; \nabla f(x_i) + J_i u_i \;\overset{(4.2)}{=}\; J_i(J_i^{-1} - G_i^{-1})\nabla f(x_i). \tag{4.15}$$

Hence,

$$\begin{aligned} \lambda_{i+1} &\overset{(4.7)}{=} \langle \nabla f(x_{i+1}), \nabla^2 f(x_{i+1})^{-1}\nabla f(x_{i+1})\rangle^{1/2} \\ &\overset{(4.6)}{\leq} \sqrt{1 + \tfrac{Mr_i}{2}}\langle \nabla f(x_{i+1}), J_i^{-1}\nabla f(x_{i+1})\rangle^{1/2} \\ &\overset{(4.15)}{=} \sqrt{1 + \tfrac{Mr_i}{2}}\langle \nabla f(x_i), (J_i^{-1} - G_i^{-1})J_i(J_i^{-1} - G_i^{-1})\nabla f(x_i)\rangle^{1/2}. \end{aligned} \tag{4.16}$$

Note that

$$-(\xi_{i+1} - 1)J_i^{-1} \;\overset{(4.10)}{\preceq}\; J_i^{-1} - G_i^{-1} \;\overset{(4.10)}{\preceq}\; \left(1 - \tfrac{\mu}{\xi_{i+1}L}\right) J_i^{-1}.$$

Therefore,

$$(J_i^{-1} - G_i^{-1})J_i(J_i^{-1} - G_i^{-1}) \;\overset{(4.14)}{\preceq}\; q_i^2 J_i^{-1} \;\overset{(4.5)}{\preceq}\; q_i^2 \left(1 + \tfrac{Mr_i}{2}\right) \nabla^2 f(x_i)^{-1}.$$

Thus,

$$\lambda_{i+1} \;\overset{(4.16)}{\leq}\; \left(1 + \tfrac{Mr_i}{2}\right) q_i \langle \nabla f(x_i), \nabla^2 f(x_i)^{-1}\nabla f(x_i)\rangle^{1/2} \;\overset{(4.7)}{=}\; \left(1 + \tfrac{Mr_i}{2}\right) q_i \lambda_i.$$

Consequently,

$$\lambda_k \;\leq\; \lambda_0 \prod_{i=0}^{k-1} \left(1 + \tfrac{Mr_i}{2}\right) q_i \;\leq\; \lambda_0 \prod_{i=0}^{k-1} e^{\frac{Mr_i}{2}} q_i \;\overset{(4.8)}{=}\; \sqrt{\xi_k}\lambda_0 \prod_{i=0}^{k-1} q_i. \Box$$

Next, we establish a preliminary version of the theorem on superlinear convergence of the scheme (4.2).

Lemma 4.4 *For all $k \geq 1$, we have*

$$\lambda_k \quad \leq \quad \left[\frac{1+\xi_k}{\prod_{i=0}^{k-1}(\tau_i \frac{\mu}{\xi_{i+1}^2 L}+1-\tau_i)^{1/k}} \left(e^{\frac{13}{6}\frac{n}{k}\left(\ln\frac{L}{\mu}+\xi_{k+1}\ln\xi_{k+1}\right)}-1\right)\right]^{k/2} \sqrt{\xi_k \frac{L}{\mu}} \cdot \lambda_0. \tag{4.17}$$

Proof: Let $k \geq 1$ be arbitrary. Without loss of generality, we can assume that $u_i \neq 0$ for all $0 \leq i \leq k$. Denote $\psi_i \stackrel{\text{def}}{=} \psi(G_i, J_i)$, $\tilde{\psi}_{i+1} \stackrel{\text{def}}{=} \psi(G_{i+1}, J_i)$, $\nu_i \stackrel{\text{def}}{=} \nu(J_i, G_i, u_i)$, $p_i \stackrel{\text{def}}{=} \tau_i \frac{\mu}{\xi_{i+1}^2 L} + 1 - \tau_i$, and $g_i \stackrel{\text{def}}{=} \|\nabla f(x_i)\|_{G_i}^*$ for any $0 \leq i \leq k$.

Let $0 \leq i \leq k-1$ be arbitrary. By Lemma 2.4 and (4.10), we have

$$\tfrac{6}{13}\ln\left(1+p_i\nu_i^2\right) \quad \leq \quad \psi_i - \tilde{\psi}_{i+1} \;=\; \psi_i - \psi_{i+1} + \Delta_i, \tag{4.18}$$

where

$$\Delta_i \quad \stackrel{\text{def}}{=} \quad \psi_{i+1} - \tilde{\psi}_{i+1} \quad \stackrel{(2.9)}{=} \quad \langle G_{i+1}^{-1}, J_{i+1} - J_i\rangle + \ln \operatorname{Det}(J_{i+1}^{-1}, J_i). \tag{4.19}$$

Denote

$$\delta_i \quad \stackrel{\text{def}}{=} \quad \left(1+\tfrac{Mr_i}{2}\right)\left(1+\tfrac{Mr_{i+1}}{2}\right) \;\geq\; 1. \tag{4.20}$$

Clearly,

$$J_i \quad \stackrel{(4.6)}{\succeq} \quad \left(1+\tfrac{Mr_i}{2}\right)^{-1}\nabla^2 f(x_{i+1}) \quad \stackrel{(4.5)}{\succeq} \quad \delta_i^{-1} J_{i+1}.$$

Therefore,

$$\langle G_{i+1}^{-1}, J_{i+1} - J_i\rangle \quad \leq \quad (1-\delta_i^{-1})\langle G_{i+1}^{-1}, J_{i+1}\rangle \quad \stackrel{(4.10)}{\leq} \quad n\xi_{i+2}(1-\delta_i^{-1}).$$

Hence,

$$\begin{aligned}
\sum_{i=0}^{k-1}\langle G_{i+1}^{-1}, J_{i+1} - J_i\rangle \quad & \leq \quad n\sum_{i=0}^{k-1}\xi_{i+2}(1-\delta_i^{-1}) \;\stackrel{(4.8)}{\leq}\; n\xi_{k+1}\sum_{i=0}^{k-1}(1-\delta_i^{-1}) \\
& \stackrel{(4.20)}{\leq} \quad n\xi_{k+1}\sum_{i=0}^{k-1}\left(1-e^{-\frac{M}{2}(r_i+r_{i+1})}\right) \\
& \leq \quad n\xi_{k+1}\tfrac{M}{2}\sum_{i=0}^{k-1}(r_i+r_{i+1}) \\
& \leq \quad n\xi_{k+1}M\sum_{i=0}^{k} r_i \;\stackrel{(4.8)}{=}\; n\xi_{k+1}\ln\xi_{k+1}.
\end{aligned}$$

Consequently,

$$\sum_{i=0}^{k-1}\Delta_i \quad \stackrel{(4.19)}{\leq} \quad n\xi_{k+1}\ln\xi_{k+1} + \ln\operatorname{Det}(J_k^{-1}, J_0). \tag{4.21}$$

Summing up (4.18), we obtain

$$
\begin{array}{rcl}
\frac{6}{13}\sum\limits_{i=0}^{k-1}\ln(1+p_i\nu_i^2) & \leq & \psi_0-\psi_k+\sum\limits_{i=0}^{k-1}\Delta_i \overset{(2.10)}{\leq} \psi_0+\sum\limits_{i=0}^{k-1}\Delta_i \\
& \overset{(4.2)}{=} & \psi(LB,J_0)+\sum\limits_{i=0}^{k-1}\Delta_i \\
& \overset{(2.9)}{=} & \ln\mathrm{Det}(J_0^{-1},LB)-\langle \frac{1}{L}B^{-1}, LB-J_0\rangle+\sum\limits_{i=0}^{k-1}\Delta_i \\
& \overset{(4.21)}{\leq} & \ln\mathrm{Det}(J_k^{-1},LB)-\langle \frac{1}{L}B^{-1}, LB-J_0\rangle+n\xi_{k+1}\ln\xi_{k+1} \\
& \overset{(4.3)}{\leq} & \ln\mathrm{Det}(\frac{1}{\mu}B^{-1},LB)+n\xi_{k+1}\ln\xi_{k+1} \\
& = & n\ln\frac{L}{\mu}+n\xi_{k+1}\ln\xi_{k+1} \;=\; n\left(\ln\frac{L}{\mu}+\xi_{k+1}\ln\xi_{k+1}\right).
\end{array}
\tag{4.22}
$$

Hence, by convexity of function $t\mapsto \ln(1+e^t)$,

$$
\begin{array}{l}
\frac{13}{6}\frac{n}{k}\left(\ln\frac{L}{\mu}+\xi_{k+1}\ln\xi_{k+1}\right) \overset{(4.22)}{\geq} \frac{1}{k}\sum\limits_{i=0}^{k-1}\ln(1+p_i\nu_i^2) \;=\; \frac{1}{k}\sum\limits_{i=0}^{k-1}\ln(1+e^{\ln(p_i\nu_i^2)}) \\
\geq \ln\left(1+e^{\frac{1}{k}\sum_{i=0}^{k-1}\ln(p_i\nu_i^2)}\right) \;=\; \ln\left(1+\left[\prod\limits_{i=0}^{k-1}p_i\nu_i^2\right]^{1/k}\right).
\end{array}
\tag{4.23}
$$

At the same time, by Lemma 2.5 and (4.10), we have

$$
\nu_i^2 \;\geq\; \frac{1}{1+\xi_{i+1}}\frac{\langle (G_i-J_i)G_{i+1}^{-1}(G_i-J_i)u_i,u_i\rangle}{\langle G_iu_i,u_i\rangle} \;=\; \frac{1}{1+\xi_{i+1}}\frac{g_{i+1}^2}{g_i^2}
$$

since $G_iu_i=-\nabla f(x_i)$ and $J_iu_i=\nabla f(x_{i+1})-\nabla f(x_i)$. Consequently,

$$
\prod_{i=0}^{k-1}\nu_i^2 \;\geq\; \frac{g_k^2}{g_0^2}\prod_{i=0}^{k-1}\frac{1}{1+\xi_{i+1}} \overset{(4.8)}{\geq} \frac{1}{(1+\xi_k)^k}\frac{g_k^2}{g_0^2}.
$$

Thus,

$$
\frac{13}{6}\frac{n}{k}\left(\ln\frac{L}{\mu}+\xi_{k+1}\ln\xi_{k+1}\right) \overset{(4.23)}{\geq} \ln\left(1+\frac{\prod_{i=0}^{k-1}p_i^{1/k}}{1+\xi_k}\left[\frac{g_k}{g_0}\right]^{2/k}\right),
$$

Consequently,

$$
g_k \;\leq\; \left[\frac{1+\xi_k}{\prod_{i=0}^{k-1}p_i^{1/k}}\left(e^{\frac{13}{6}\frac{n}{k}\left(\ln\frac{L}{\mu}+\xi_{k+1}\ln\xi_{k+1}\right)}-1\right)\right]^{k/2} g_0.
$$

It remains to note that $\lambda_k\leq\sqrt{\xi_k\frac{L}{\mu}}\cdot g_k$ in view of (4.9), and $g_0\leq\lambda_0$ since in view of (4.3) and the fact that $G_0=LB$. □

Note that, in the quadratic case ($M=0$), we have $\xi_k\equiv 1$ (see (4.8)), and so Lemmas 4.2, 4.3 reduce to the already known Theorem 3.1, and Lemma 4.4 reduces to the already known Theorem 3.2. In the general case, the quantities ξ_k can grow with iterations. However, as we will see in a moment, by requiring the initial point x_0 in the scheme

(4.2) to be sufficiently close to the solution, we can still ensure that ξ_k stay *uniformly bounded* by a sufficiently small absolute constant. This allows us to recover all the main results of the quadratic case.

To write down the region of local convergence of the scheme (4.2), we need to introduce one more quantity, which is related to the starting moment of superlinear convergence[3]:

$$K_0 \stackrel{\text{def}}{=} \left\lceil \frac{1}{\tau \frac{4\mu}{9L} + 1 - \tau} 8n \left(\ln \frac{L}{\mu} + 1 \right) \right\rceil, \qquad \tau \stackrel{\text{def}}{=} \sup_{k \ge 0} \tau_k \quad (\le 1). \tag{4.24}$$

For DFP ($\tau_k \equiv 1$) and BFGS ($\tau_k \equiv 0$), we have respectively

$$K_0^{\text{DFP}} = \left\lceil \frac{18nL}{\mu} \left(\ln \frac{L}{\mu} + 1 \right) \right\rceil, \qquad K_0^{\text{BFGS}} = \left\lceil 8n \left(\ln \frac{L}{\mu} + 1 \right) \right\rceil. \tag{4.25}$$

Now we are ready to prove the main result of this section.

Theorem 4.1 *Suppose that, in scheme* (4.2)*, we have*

$$M\lambda_0 \le \frac{\ln \frac{3}{2}}{\left(\frac{3}{2}\right)^{3/2}} \cdot \max\left\{ \frac{\mu}{2L}, \frac{1}{K_0 + 9} \right\}. \tag{4.26}$$

Then, for all $k \ge 0$*,*

$$\tfrac{2}{3} \nabla^2 f(x_k) \preceq G_k \preceq \tfrac{3L}{2\mu} \nabla^2 f(x_k), \tag{4.27}$$

$$\lambda_k \le \left(1 - \tfrac{\mu}{2L}\right)^k \sqrt{\tfrac{3}{2}} \cdot \lambda_0, \tag{4.28}$$

and, for all $k \ge 1$*,*

$$\lambda_k \le \left[\frac{5}{2 \prod_{i=0}^{k-1} (\tau_i \frac{4\mu}{9L} + 1 - \tau_i)^{1/k}} \left(e^{\frac{13}{6} \frac{n}{k} \left(\ln \frac{L}{\mu} + 1 \right)} - 1 \right) \right]^{k/2} \sqrt{\tfrac{3L}{2\mu}} \cdot \lambda_0. \tag{4.29}$$

Proof: Let us prove by induction that, for all $k \ge 0$,

$$\xi_k \le \tfrac{3}{2}. \tag{4.30}$$

Clearly, for $k = 0$, (4.30) is satisfied since $\xi_0 = 1$. It is also satisfied for $k = 1$ since

$$\xi_1 \stackrel{(4.8)}{=} e^{Mr_0} \stackrel{(4.12)}{\le} e^{\xi_0 M \lambda_0} \stackrel{(4.8)}{=} e^{M\lambda_0} \stackrel{(4.26)}{\le} \tfrac{3}{2}.$$

Now let $k \ge 0$, and suppose that (4.30) has already been proved for all indices up to $k+1$. Then, applying Lemma 4.2, we obtain (4.27) for all indices up to $k+1$. Applying now Lemma 4.3 and using for all $0 \le i \le k$ the relation

$$q_i \stackrel{(4.14)}{=} \max\left\{ 1 - \frac{\mu}{\xi_{i+1} L}, \xi_{i+1} - 1 \right\} \stackrel{(4.30)}{\le} \max\left\{ 1 - \frac{2\mu}{3L}, \frac{1}{2} \right\} \le 1 - \frac{\mu}{2L},$$

[3] Hereinafter, $\lceil t \rceil$ for $t > 0$ denotes the smallest positive integer greater or equal to t.

we obtain (4.28) for all indices up to $k+1$. Finally, if $k \geq 1$, then, applying Lemma 4.4 and using for all $0 \leq i \leq k$ the bound

$$\xi_{i+1} \ln \xi_{i+1} \overset{(4.30)}{\leq} \tfrac{3}{2} \ln \tfrac{3}{2} \;\leq\; \tfrac{3}{2} \cdot \tfrac{1}{2} \;=\; \tfrac{3}{4} \;\leq\; 1,$$

we obtain (4.29) for all indices up to k. Thus, at this moment, the inequalities (4.27) and (4.28) are proved for all indices up to $k+1$, while (4.29) is proved only up to the index k.

To conclude the inductive step, it remains to prove that (4.30) is also satisfied for the next index, $k+2$, or, equivalently, in view of (4.8), that

$$M \sum_{i=0}^{k+1} r_i \;\leq\; \ln \tfrac{3}{2}.$$

Since

$$M \sum_{i=0}^{k+1} r_i \overset{(4.12)}{\leq} M \sum_{i=0}^{k+1} \xi_i \lambda_i \overset{(4.30)}{\leq} \tfrac{3}{2} M \sum_{i=0}^{k+1} \lambda_i,$$

it suffices to show that

$$\tfrac{3}{2} M \sum_{i=0}^{k+1} \lambda_i \;\leq\; \ln \tfrac{3}{2}. \tag{4.31}$$

Note that

$$\tfrac{3}{2} M \sum_{i=0}^{k+1} \lambda_i \overset{(4.28)}{\leq} \left(\tfrac{3}{2}\right)^{3/2} M \lambda_0 \sum_{i=0}^{k+1} \left(1 - \tfrac{\mu}{2L}\right)^i \;\leq\; \left(\tfrac{3}{2}\right)^{3/2} \tfrac{2L}{\mu} M \lambda_0. \tag{4.32}$$

Therefore, if we could prove that

$$\tfrac{3}{2} M \sum_{i=0}^{k+1} \lambda_i \;\leq\; \left(\tfrac{3}{2}\right)^{3/2} (K_0 + 9) M \lambda_0, \tag{4.33}$$

then, combining (4.32) and (4.33), we would obtain

$$\tfrac{3}{2} M \sum_{i=0}^{k+1} \lambda_i \;\leq\; \left(\tfrac{3}{2}\right)^{3/2} \min\left\{ \tfrac{2L}{\mu}, K_0 + 9 \right\} M \lambda_0 \overset{(4.26)}{\leq} \ln \tfrac{3}{2},$$

which is exactly (4.31).

Let us prove (4.33). If $k \leq K_0$, we have

$$\tfrac{3}{2} M \sum_{i=0}^{k+1} \lambda_i \overset{(4.28)}{\leq} \left(\tfrac{3}{2}\right)^{3/2} (k+2) M \lambda_0 \;\leq\; \left(\tfrac{3}{2}\right)^{3/2} (K_0 + 2) M \lambda_0,$$

and (4.33) follows. Therefore, from now on, we can assume that $k \geq K_0$. Then[4],

$$\begin{aligned} \tfrac{3}{2} M \sum_{i=0}^{k+1} \lambda_i \;&=\; \tfrac{3}{2} M \left(\sum_{i=0}^{K_0 - 1} \lambda_i + \lambda_{k+1} \right) + \tfrac{3}{2} M \sum_{i=K_0}^{k} \lambda_i \\ &\overset{(4.28)}{\leq} \left(\tfrac{3}{2}\right)^{3/2} (K_0 + 1) M \lambda_0 + \tfrac{3}{2} M \sum_{i=K_0}^{k} \lambda_i. \end{aligned}$$

[4]We will estimate the second sum using (4.29). However, as was noted previously, at this moment, (4.29) is proved only up to the index k. This is the reason why we move the term λ_{k+1} into the first sum.

It remains to show that

$$\tfrac{3}{2} M \sum_{i=K_0}^{k} \lambda_i \;\; \leq \;\; \left(\tfrac{3}{2}\right)^{3/2} \cdot 8 \cdot M \lambda_0. \tag{4.34}$$

We can do this using (4.29).

First, let us make some estimations. Clearly, for all $0 < t < 1$,

$$e^t \;=\; \sum_{j=0}^{\infty} \tfrac{t^j}{j!} \;\leq\; 1 + t + \tfrac{t^2}{2} \sum_{j=0}^{\infty} t^j \;=\; 1 + t + \tfrac{t^2}{2(1-t)} \;=\; 1 + t\left(1 + \tfrac{t}{2(1-t)}\right). \tag{4.35}$$

Hence, using that $\frac{83}{70} \leq \frac{6}{5}$, for all $0 < t \leq 1$, we obtain

$$e^{\frac{13t}{48}} - 1 \;\stackrel{(4.35)}{\leq}\; \tfrac{13t}{48}\left(1 + \tfrac{\frac{13}{48}}{2(1-\frac{13}{48})}\right) \;=\; \tfrac{13t}{48} \cdot \tfrac{83}{70} \;\leq\; \tfrac{13t}{48} \cdot \tfrac{6}{5} \;=\; \tfrac{13t}{40}. \tag{4.36}$$

Since $\frac{13}{16} \leq \frac{121}{144}$, it follows that

$$\left[\tfrac{5}{2t}\left(e^{\frac{13t}{48}} - 1\right)\right]^{1/2} \;\stackrel{(4.36)}{\leq}\; \sqrt{\tfrac{5}{2t} \cdot \tfrac{13t}{40}} \;=\; \sqrt{\tfrac{13}{16}} \;\leq\; \tfrac{11}{12}. \tag{4.37}$$

At the same time, $\frac{11}{12} = 1 - \frac{1}{12} \leq e^{-1/12}$. Hence,

$$\left(\tfrac{11}{12}\right)^{8 \ln \frac{L}{\mu}} \cdot \sqrt{\tfrac{L}{\mu}} \;\;\leq\;\; e^{-\frac{2}{3} \ln \frac{L}{\mu}} \cdot \sqrt{\tfrac{L}{\mu}} \;=\; \left(\tfrac{L}{\mu}\right)^{-2/3} \cdot \sqrt{\tfrac{L}{\mu}} \;=\; \left(\tfrac{L}{\mu}\right)^{-1/6} \;\leq\; 1, \tag{4.38}$$

and

$$\left(\tfrac{11}{12}\right)^{K_0} \sqrt{\tfrac{L}{\mu}} \;\stackrel{(4.24)}{\leq}\; \left(\tfrac{11}{12}\right)^{8\left(\ln \frac{L}{\mu} + 1\right)} \sqrt{\tfrac{L}{\mu}} \;\stackrel{(4.38)}{\leq}\; \left(\tfrac{11}{12}\right)^{8} \;\leq\; e^{-\frac{2}{3}} \;\leq\; \tfrac{1}{1+\frac{2}{3}} \;=\; \tfrac{3}{5} \;\leq\; \tfrac{2}{3}. \tag{4.39}$$

Thus, for all $K_0 \leq i \leq k$, denoting

$$p \;\stackrel{\mathrm{def}}{=}\; \tau \tfrac{4\mu}{9L} + 1 - \tau \;\stackrel{(4.24)}{\leq}\; \left[\prod_{j=0}^{i-1} \left(\tau_i \tfrac{4\mu}{9L} + 1 - \tau_i\right)\right]^{1/i}, \tag{4.40}$$

we obtain

$$\begin{aligned} \lambda_i \;&\stackrel{(4.29)}{\leq}\; \left[\tfrac{5}{2p}\left(e^{\frac{13}{6}\frac{n}{i}\left(\ln \frac{L}{\mu} + 1\right)} - 1\right)\right]^{i/2} \sqrt{\tfrac{3L}{2\mu}} \cdot \lambda_0 \\ &\stackrel{(4.24)}{\leq}\; \left[\tfrac{5}{2p}\left(e^{\frac{13p}{48}} - 1\right)\right]^{i/2} \sqrt{\tfrac{3L}{2\mu}} \cdot \lambda_0 \;\stackrel{(4.37)}{\leq}\; \left(\tfrac{11}{12}\right)^{i} \sqrt{\tfrac{3L}{2\mu}} \cdot \lambda_0 \\ &=\; \left(\tfrac{11}{12}\right)^{i-K_0} \left(\tfrac{11}{12}\right)^{K_0} \sqrt{\tfrac{3L}{2\mu}} \cdot \lambda_0 \;\stackrel{(4.39)}{\leq}\; \left(\tfrac{11}{12}\right)^{i-K_0} \tfrac{2}{3} \cdot \sqrt{\tfrac{3}{2}} \cdot \lambda_0. \end{aligned} \tag{4.41}$$

Consequently,

$$\begin{aligned} \tfrac{3}{2} M \sum_{i=K_0}^{k} \lambda_i \;&\stackrel{(4.41)}{\leq}\; \left(\tfrac{3}{2}\right)^{3/2} M \lambda_0 \cdot \tfrac{2}{3} \sum_{i=K_0}^{k} \left(\tfrac{11}{12}\right)^{i-K_0} \\ &\leq\; \left(\tfrac{3}{2}\right)^{3/2} M \lambda_0 \cdot \tfrac{2}{3} \cdot 12 \;=\; \left(\tfrac{3}{2}\right)^{3/2} \cdot 8 \cdot M \lambda_0, \end{aligned}$$

and (4.34) follows. □

Remark 4.1 *According to Theorem 4.1, the parameter of strong self-concordancy M only affects the size of the region of local convergence of the process* (4.2)*, and not its rate of convergence. For a quadratic function, we have $M = 0$, and so the scheme* (4.2) *is globally convergent.*

The region of local convergence, specified by (4.26), depends on the *maximum* of two quantities: $\frac{\mu}{L}$ and $\frac{1}{K_0}$. For DFP, the $\frac{1}{K_0}$ part in this maximum is in fact redundant, and its region of local convergence is simply inversely proportional to the condition number:

$$O\left(\tfrac{\mu}{L}\right).$$

However, for BFGS, the $\frac{1}{K_0}$ part does not disappear, and we obtain the following region of local convergence:

$$\max\left\{O\left(\tfrac{\mu}{L}\right),\ O\left(\tfrac{1}{n(\ln\frac{L}{\mu}+1)}\right)\right\}.$$

Clearly, this second region can be much bigger than the first one when the condition number $\frac{L}{\mu}$ is significantly larger than the dimension n.

Remark 4.2 *The previous estimate of the size of the region of local convergence, established in [24], was $O(\frac{\mu}{L})$ for both DFP and BFGS.*

5 Discussion

Let us compare the new superlinear convergence rates, obtained in this paper for the classical DFP and BFGS methods, with the previously known results from [24]. Since the efficiency estimates in the general nonlinear case differ from those for the quadratic one just in absolute constants, we only discuss the quadratic case.

In what follows, we use our standard notation: n is the dimension of the space, μ is the strong convexity parameter, L is the Lipschitz constant of the gradient, and λ_k is the local norm of the gradient, taken at the k-th iteration of the method.

For BFGS method, the previously known estimate of the superlinear convergence rate (see [24, Theorem 3.2]) is

$$\lambda_k \quad \leq \quad \left(\tfrac{nL}{\mu k}\right)^{k/2}\lambda_0. \tag{5.1}$$

Although (5.1) is formally valid for all $k \geq 1$, it becomes useful[5] only after

$$\widehat{K}_0^{\mathrm{BFGS}} \quad \stackrel{\mathrm{def}}{=} \quad \tfrac{nL}{\mu} \tag{5.2}$$

iterations. Thus, $\widehat{K}_0^{\mathrm{BFGS}}$ can be thought of as the *starting moment* of the superlinear convergence, according to the estimate (5.1).

In this paper, we have obtained a new estimate (Theorem 3.2):

$$\lambda_k \quad \leq \quad \left[2\left(e^{\frac{n}{k}\ln\frac{L}{\mu}} - 1\right)\right]^{k/2}\sqrt{\tfrac{L}{\mu}}\cdot\lambda_0. \tag{5.3}$$

[5]Indeed, according to Theorem 3.1, we always have at least that $\lambda_k \leq (1-\frac{\mu}{L})^k\lambda_0$ for all $k \geq 0$.

Its starting moment of superlinear convergence can be described as follows:

$$K_0^{\rm BFGS} \overset{\rm def}{=} 4n \ln \tfrac{L}{\mu}. \tag{5.4}$$

Indeed, since $e^t \leq \frac{1}{1-t} = 1 + \frac{t}{1-t}$ for any $t < 1$, we have, for all $k \geq K_0^{\rm BFGS}$,

$$e^{\frac{n}{k}\ln\frac{L}{\mu}} - 1 \;\leq\; \frac{\frac{n}{k}\ln\frac{L}{\mu}}{1-\frac{n}{k}\ln\frac{L}{\mu}} \overset{(5.4)}{\leq} \frac{\frac{n}{k}\ln\frac{L}{\mu}}{1-\frac{1}{4}} \;=\; \frac{4n}{3k}\ln\frac{L}{\mu}. \tag{5.5}$$

At the same time, for all $k \geq K_0^{\rm BFGS}$,

$$\sqrt{\tfrac{L}{\mu}} \;=\; e^{\frac{1}{2}\ln\frac{L}{\mu}} \overset{(5.4)}{\leq} e^{\frac{k}{8}} \;=\; (e^{\frac{1}{4}})^{k/2} \;\leq\; \left(1 + \tfrac{\frac{1}{4}}{1-\frac{1}{4}}\right)^{k/2} \;\leq\; \left(\tfrac{4}{3}\right)^{k/2} \;\leq\; \left(\tfrac{3}{2}\right)^{k/2}. \tag{5.6}$$

Hence, according the new estimate (5.3), for all $k \geq K_0^{\rm BFGS}$, we have

$$\lambda_k \overset{(5.5)}{\leq} \left(\tfrac{8n}{3k}\ln\tfrac{L}{\mu}\right)^{k/2}\sqrt{\tfrac{L}{\mu}}\cdot\lambda_0 \overset{(5.6)}{\leq} \left(\tfrac{4n}{k}\ln\tfrac{L}{\mu}\right)^{k/2}\lambda_0 \qquad (\overset{(5.4)}{\leq} \lambda_0). \tag{5.7}$$

Comparing the previously known efficiency estimate (5.1) and its starting moment of superlinear convergence (5.2) with the new ones (5.7), (5.4), we thus conclude that we manage to put the condition number $\frac{L}{\mu}$ *under the logarithm*.

For DFP, the previously known rate (see [24, Theorem 3.2]) is

$$\lambda_k \;\leq\; \left(\tfrac{nL^2}{\mu^2 k}\right)^{k/2}\lambda_0$$

with the following starting moment of the superlinear convergence:

$$\widehat{K}_0^{\rm DFP} \overset{\rm def}{=} \tfrac{nL^2}{\mu^2}. \tag{5.8}$$

The new rate, that we have obtained in this paper (Theorem 3.2), is

$$\lambda_k \;\leq\; \left[\tfrac{2L}{\mu}\left(e^{\frac{n}{k}\ln\frac{L}{\mu}} - 1\right)\right]^{k/2}\sqrt{\tfrac{L}{\mu}}\cdot\lambda_0. \tag{5.9}$$

Repeating the same reasoning as above, we can easily obtain that the new starting moment of the superlinear convergence can be described as follows:

$$K_0^{\rm DFP} \overset{\rm def}{=} \tfrac{4nL}{\mu}\ln\tfrac{L}{\mu}, \tag{5.10}$$

and, for all $k \geq K_0^{\rm DFP}$, the new estimate (5.9) takes the following form:

$$\lambda_k \;\leq\; \left(\tfrac{4nL}{\mu k}\ln\tfrac{L}{\mu}\right)^{k/2}\lambda_0 \quad (\overset{(5.10)}{\leq} \lambda_0).$$

Thus, compared to the old result, we have improved the factor $\frac{L^2}{\mu^2}$ up to $\frac{L}{\mu}\ln\frac{L}{\mu}$. Interestingly enough, we see that the ratio between the old starting moments (5.8), (5.2) of the superlinear convergence of DFP and BFGS and the new ones (5.10), (5.4) have remained the same, $\frac{L}{\mu}$, although the both estimates have been improved.

To conclude, let us mention several open questions. First, looking at the starting moment of superlinear convergence of the BFGS method, $n \ln \frac{L}{\mu}$, in addition to the dimension n, we see the presence of the logarithm of the condition number $\ln \frac{L}{\mu}$. Although typically such logarithmic factors are considered small, it is still interesting to understand whether this factor can be completely removed.

Second, note that all the estimates of superlinear convergence, that we have obtained in this paper for the convex Broyden class are expressed in terms of the parameter τ, which controls the weight of the DFP component in the updating formula for the *inverse* operator (see (A.1)). At the same time, in [24], the corresponding estimates were presented in terms of the parameter ϕ, which controls the weight of the DFP component in the updating formula for the *primal* operator (see (2.1)). Of course, for the extreme members of the convex Broyden class, DFP and BFGS, both parameters ϕ and τ coincide. However, in general, they could be quite different. We do not know if it is possible to express the results of this paper in terms of ϕ instead of τ.

Finally, recall that, in all the quasi-Newton methods, which we considered, the initial Hessian approximation G_0 was set to LB, where L is the Lipschitz constant of the gradient, measured relative to the operator B. We always assume that this Lipschitz constant is available to the methods. Of course, it is interesting to develop some *adaptive* algorithms, which could start from any initial guess L_0 for the constant L, and then somehow dynamically adjust the Hessian approximations in iterations, yet retaining all the original efficiency estimates.

References

[1] C. Broyden. Quasi-Newton methods and their application to function minimization. *Mathematics of Computation*, **21**(99), 368-381 (1967).

[2] C. Broyden. The convergence of a class of double-rank minimization algorithms: 1. General considerations. *IMA Journal of Applied Mathematics*, **6**(1), 76-90 (1970).

[3] C. Broyden. The convergence of a class of double-rank minimization algorithms: 2. The new algorithm. *IMA Journal of Applied Mathematics*, **6**(3), 222-231 (1970).

[4] C. Broyden, J. Dennis, and J. Moré. On the local and superlinear convergence of quasi-Newton methods. *IMA Journal of Applied Mathematics*, **12**(3), 223-245 (1973).

[5] R. Byrd and J. Nocedal. A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. *SIAM Journal on Numerical Analysis*, **26**(3), 727-739 (1989).

[6] R. Byrd, D. Liu, and J. Nocedal. On the behavior of Broyden's class of quasi-Newton methods. *SIAM Journal on Optimization*, **2**(4), 533-557 (1992).

[7] W. Davidon. Variable metric method for minimization. Argonne National Laboratory Research and Development Report 5990 (1959).

[8] J. Dennis and J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Mathematics of Computation*, **28**(126), 549-560 (1974).

[9] J. Dennis and J. Moré. Quasi-Newton methods, motivation and theory. *SIAM Review*, **19**(1), 46-89 (1977).

[10] L. Dixon. Quasi-Newton algorithms generate identical points. *Mathematical Programming*, **2**(1), 383-387 (1972).

[11] L. Dixon. Quasi Newton techniques generate identical points II: The proofs of four new theorems. *Mathematical Programming*, **3**(1), 345-358 (1972).

[12] J. Engels and H. Martínez. Local and superlinear convergence for partially known quasi-Newton methods. SIAM Journal on Optimization, **1**(1), 42-56 (1991).

[13] R. Fletcher and M. Powell. A rapidly convergent descent method for minimization. *Computer Journal*, **6**(2), 163-168 (1963).

[14] R. Fletcher. A new approach to variable metric algorithms. *Computer Journal*, **13**(3), 317-322 (1970).

[15] W. Gao and D. Goldfarb. Quasi-Newton methods: superlinear convergence without line searches for self-concordant functions. *Optimization Methods and Software*, **34**(1), 194-217 (2019).

[16] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, **24**(109), 23-26 (1970).

[17] A. Griewank and P. Toint. Local convergence analysis for partitioned quasi-Newton updates. *Numerische Mathematik*, **39**(3), 429-448 (1982).

[18] Q. Jin, A. Mokhtari. Non-asymptotic Superlinear Convergence of Standard Quasi-Newton Methods. *arXiv preprint*, arXiv:2003.13607 (2020).

[19] A. Lewis and M. Overton. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, **141**(1-2), 135-163 (2013).

[20] A. Mokhtari, M. Eisen, and A. Ribeiro. IQN: An incremental quasi-Newton method with local superlinear convergence rate. *SIAM Journal on Optimization*, **28**(2), 1670-1698 (2018).

[21] J. Nocedal and S. Wright. Numerical optimization. *Springer Science & Business Media* (2006).

[22] M. Powell. On the convergence of the variable metric algorithm. *IMA Journal of Applied Mathematics*, **7**(1), 21-36 (1971).

[23] A. Rodomanov, Y. Nesterov. Greedy quasi-Newton methods with explicit superlinear convergence. *CORE Discussion Papers*, **06** (2020).

[24] A. Rodomanov, Y. Nesterov. Rates of Superlinear Convergence for Classical Quasi-Newton Methods. *CORE Discussion Papers*, **11** (2020).

[25] D. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, **24**(111), 647-656 (1970).

[26] A. Stachurski. Superlinear convergence of Broyden's bounded θ-class of methods. *Mathematical Programming*, **20**(1), 196-212 (1981).

[27] Z. Wei, G. Yu, G. Yuan, and Z. Lian. The superlinear convergence of a modified BFGS-type method for unconstrained optimization. *Computational Optimization and Applications*, **29**(3), 315-332 (2004).

[28] H. Yabe and N. Yamaki. Local and superlinear convergence of structured quasi-Newton methods for nonlinear optimization. *Journal of the Operations Research Society of Japan*, **39**(4), 541-557 (1996).

[29] H. Yabe, H. Ogasawara, and M. Yoshino. Local and superlinear convergence of quasi-Newton methods based on modified secant conditions. *Journal of Computational and Applied Mathematics*, **205**(1), 617-632 (2007).

A Appendix

Lemma A.1 *Let $A, G : \mathbb{E} \to \mathbb{E}^*$ be self-adjoint positive definite linear operators, let $u \in \mathbb{E}$ be non-zero, and let $\tau \in \mathbb{R}$ be such that $G_+ \stackrel{\mathrm{def}}{=} \mathrm{Broyd}_\tau(A, G, u)$ is well-defined. Then,*

$$\begin{aligned} G_+^{-1} &= \tau\left[G^{-1} - \tfrac{G^{-1}Auu^*AG^{-1}}{\langle AG^{-1}Au,u\rangle} + \tfrac{uu^*}{\langle Au,u\rangle}\right] \\ &\quad + (1-\tau)\left[G^{-1} - \tfrac{G^{-1}Auu^* + uu^*AG^{-1}}{\langle Au,u\rangle} + \left(\tfrac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} + 1\right)\tfrac{uu^*}{\langle Au,u\rangle}\right], \end{aligned} \tag{A.1}$$

and

$$\mathrm{Det}(G_+^{-1}, G) = \tau\tfrac{\langle Au,u\rangle}{\langle AG^{-1}Au,u\rangle} + (1-\tau)\tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle}. \tag{A.2}$$

Proof: Denote $\phi \stackrel{\mathrm{def}}{=} \phi_\tau(A, G, u)$. According to Lemma 6.2 in [24], we have

$$\mathrm{Det}(G^{-1}, G_+) = \phi\tfrac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} + (1-\phi)\tfrac{\langle Au,u\rangle}{\langle Gu,u\rangle} \stackrel{(2.2)}{=} \left[\tau\tfrac{\langle Au,u\rangle}{\langle AG^{-1}Au,u\rangle} + (1-\tau)\tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle}\right]^{-1}.$$

This proves (A.2) since $\mathrm{Det}(G_+^{-1}, G) = \frac{1}{\mathrm{Det}(G^{-1}, G_+)}$ in view of (1.3) and (1.2).

Let us prove (A.1). Denote

$$G_0 \stackrel{\mathrm{def}}{=} G - \tfrac{Guu^*G}{\langle Gu,u\rangle} + \tfrac{Auu^*A}{\langle Au,u\rangle}, \qquad s \stackrel{\mathrm{def}}{=} \tfrac{Au}{\langle Au,u\rangle} - \tfrac{Gu}{\langle Gu,u\rangle}. \tag{A.3}$$

Note that

$$\begin{aligned} G_+ &\stackrel{(2.1)}{=} G_0 + \phi\left[\tfrac{\langle Gu,u\rangle Auu^*A}{\langle Au,u\rangle^2} + \tfrac{Guu^*G}{\langle Gu,u\rangle} - \tfrac{\langle Auu^*G + Guu^*A}{\langle Au,u\rangle}\right] \\ &= G_0 + \phi\langle Gu,u\rangle ss^*. \end{aligned} \tag{A.4}$$

Let $I_{\mathbb{E}}$ and $I_{\mathbb{E}^*}$ be the identity operators in $\mathbb{E}$, $\mathbb{E}^*$. Since $G_0 u = Au$, we have

$$\begin{aligned} &\left[\left(I_{\mathbb{E}} - \tfrac{uu^*A}{\langle Au,u\rangle}\right)G^{-1}\left(I_{\mathbb{E}^*} - \tfrac{Auu^*}{\langle Au,u\rangle}\right) + \tfrac{uu^*}{\langle Au,u\rangle}\right]G_0 \\ = &\left(I_{\mathbb{E}} - \tfrac{uu^*A}{\langle Au,u\rangle}\right)G^{-1}\left(G_0 - \tfrac{Auu^*A}{\langle Au,u\rangle}\right) + \tfrac{uu^*A}{\langle Au,u\rangle} \\ \stackrel{(A.3)}{=} &\left(I_{\mathbb{E}} - \tfrac{uu^*A}{\langle Au,u\rangle}\right)G^{-1}\left(G - \tfrac{Guu^*G}{\langle Gu,u\rangle}\right) + \tfrac{uu^*A}{\langle Au,u\rangle} = I_{\mathbb{E}}. \end{aligned}$$

Hence, we can conclude that

$$\begin{aligned} G_0^{-1} &= \left(I_{\mathbb{E}} - \tfrac{uu^*A}{\langle Au,u\rangle}\right)G^{-1}\left(I_{\mathbb{E}^*} - \tfrac{Auu^*}{\langle Au,u\rangle}\right) + \tfrac{uu^*}{\langle Au,u\rangle} \\ &= G^{-1} - \tfrac{G^{-1}Auu^* + uu^*AG^{-1}}{\langle Au,u\rangle} + \left(\tfrac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} + 1\right)\tfrac{uu^*}{\langle Au,u\rangle}. \end{aligned}$$

Thus, we see that the right-hand side of (A.1) equals

$$\begin{aligned} H_+ &\stackrel{\text{def}}{=} G_0^{-1} - \tau \left[\tfrac{\langle AG^{-1}Au,u\rangle uu^*}{\langle Au,u\rangle^2} + \tfrac{G^{-1}Auu^*AG^{-1}}{\langle AG^{-1}Au,u\rangle} - \tfrac{G^{-1}Auu^*+uu^*AG^{-1}}{\langle Au,u\rangle} \right] \\ &= G_0^{-1} - \tau \langle AG^{-1}Au,u\rangle ww^*, \end{aligned} \tag{A.5}$$

where

$$w \stackrel{\text{def}}{=} \tfrac{G^{-1}Au}{\langle AG^{-1}Au,u\rangle} - \tfrac{u}{\langle Au,u\rangle}. \tag{A.6}$$

It remains to verify that $H_+G_+ = I_{\mathbb{E}}$. Clearly,

$$\begin{aligned} \langle AG^{-1}Au,u\rangle G_0 w &\stackrel{(A.6)}{=} G_0 G^{-1}Au - \tfrac{\langle AG^{-1}Au,u\rangle G_0 u}{\langle Au,u\rangle} \\ &\stackrel{(A.3)}{=} Au - \tfrac{\langle Au,u\rangle Gu}{\langle Gu,u\rangle} \stackrel{(A.3)}{=} \langle Au,u\rangle s. \end{aligned} \tag{A.7}$$

Hence,

$$\begin{aligned} \langle AG^{-1}Au,u\rangle \langle G_0 w,w\rangle &\stackrel{(A.7)}{=} \langle Au,u\rangle\langle s,w\rangle \stackrel{(A.6)}{=} \tfrac{\langle Au,u\rangle\langle s,G^{-1}Au\rangle}{\langle AG^{-1}Au,u\rangle} - \langle s,u\rangle \\ &\stackrel{(A.3)}{=} \tfrac{\langle Au,u\rangle}{\langle AG^{-1}Au,u\rangle}\left(\tfrac{\langle AG^{-1}Au,u\rangle}{\langle Au,u\rangle} - \tfrac{\langle Au,u\rangle}{\langle Gu,u\rangle} \right) \\ &= 1 - \tfrac{\langle Au,u\rangle^2}{\langle AG^{-1}Au,u\rangle\langle Gu,u\rangle}. \end{aligned} \tag{A.8}$$

Consequently,

$$\begin{aligned} \tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle} H_+ G_0 ww^* G_0 &\stackrel{(A.5)}{=} \tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle}(G_0^{-1} - \tau\langle AG^{-1}Au,u\rangle ww^*)G_0 ww^* G_0 \\ &= \tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle}(1 - \tau\langle AG^{-1}Au,u\rangle\langle G_0 w,w\rangle) ww^* G_0 \\ &\stackrel{(A.8)}{=} \tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle}\left(1 - \tau + \tau\tfrac{\langle Au,u\rangle^2}{\langle AG^{-1}Au,u\rangle\langle Gu,u\rangle}\right) ww^* G_0 \\ &= \left[\tau\tfrac{\langle Au,u\rangle}{\langle AG^{-1}Au,u\rangle} + (1-\tau)\tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle}\right] ww^* G_0. \end{aligned} \tag{A.9}$$

Thus,

$$\begin{aligned} H_+G_+ &\stackrel{(A.4)}{=} H_+(G_0 + \phi\langle Gu,u\rangle ss^*) \stackrel{(A.7)}{=} H_+\left(G_0 + \phi\tfrac{\langle AG^{-1}Au,u\rangle^2}{\langle Au,u\rangle}\tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle} G_0 ww^* G_0\right) \\ &\stackrel{(A.9)}{=} H_+G_0 + \phi\tfrac{\langle AG^{-1}Au,u\rangle^2}{\langle Au,u\rangle}\left[\tau\tfrac{\langle Au,u\rangle}{\langle AG^{-1}Au,u\rangle} + (1-\tau)\tfrac{\langle Gu,u\rangle}{\langle Au,u\rangle}\right] \\ &\stackrel{(2.2)}{=} H_+G_0 + \tau\langle AG^{-1}Au,u\rangle ww^* G_0 \stackrel{(A.5)}{=} I_{\mathbb{E}}. \end{aligned}$$

$\square$