Non-asymptotic superlinear convergence of Nesterov accelerated BFGS

Manish Kumar Sahu[1] and Suvendu Ranjan Pattanaik[2*]

Mathematics Department, National Institute of Technology Rourkela, Rourkela, 769008, Odisha, India.

*Corresponding author(s). E-mail(s): suvendu.pattanaik@gmail.com;
Contributing authors: manishkumarsahu132@gmail.com;

Abstract

This paper studies the convergence of a Nesterov accelerated variant of the Broyden-Fletcher-Goldfarb-Shanno (NA-BFGS) quasi-Newton method in the setting where the objective function is strongly convex, its gradient is Lipschitz continuous, and its Hessian is Lipschitz continuous at the optimal point. We demonstrate that similar to the classic BFGS method, the Nesterov accelerated BFGS method also achieves a nonasymptotic superlinear convergence rate of the form $(\frac{1}{k})^{\frac{k}{4}}$ within a local neighbourhood of the optimal point. The work provides a theoretical explanation of the superlinear convergence of NA-BFGS non-asymptotically and explicitly.

Keywords: Nesterov accelerated Broyden Fletcher Goldfarb Shanno (NA-BFGS), Broyden Fletcher Goldfarb Shanno (BFGS), Superlinear rate of convergence, Local convergence, Non-asymptotic

1 Introduction

Here, we focus on the rate of convergence of the Nesterov accelerated BFGS algorithm non-asymptotically or explicitly after a finite iteration. We minimize a function which is twice continuously differentiable and strongly convex.

$$f : R^d \longrightarrow R. \tag{1}$$

We assume that its gradient $\nabla f(x)$ is Lipschitz continuous, and its Hessian $\nabla^2 f(x)$ is Lipschitz continuous at the optimal point.

When solving convex optimization problems, first-order methods such as gradient descent, ADAM, Stochastic gradient descent, and Nesterov accelerated Gradient descent are frequently used ([1]). The sequence x_k converges to the optimal point x_* linearly if $\|x_k - x_*\| \leq N\tau^k \|x_0 - x_*\|$, where N is a constant whose value depends upon problem parameters and $\tau \in (0, 1)$. Generally, linear convergence rate is achievable if we use these first-order methods. Nesterov accelerated gradient descent achieves a fast linear rate of $(1 - \sqrt{\frac{m}{L}})^{\frac{k}{2}}$ when the dimension of the problem is higher ([2], [3]), where m is strong convexity constant and L is the Lipschitz constant of the gradient of $f(x)$.

To avoid these drawbacks, many researchers were interested in second-order methods ([4], [5], [6], [7]) such as Newton methods, etc. Newton's method has a quadratic convergence rate under certain assumptions ([1]) but has some drawbacks. The computational cost in computing the inverse Hessian is high, and it is pretty demanding to calculate the exact inverse Hessian matrix when the objective function involves many variables.

Therefore, to solve these problems, the quasi-Newton methods are proposed ([8], [9]), where they approximate the inverse Hessian matrix. The computational cost per iteration for the Newton method is $O(d^3)$, but for Quasi-Newton methods, the cost reduces to $O(d^2)$ ([2]). There are different variants in Quasi-Newton methods such as SR1 ([10]), DFP ([11], [12]), BFGS ([13], [14]), but BFGS optimizes faster. BFGS method has a super-linear rate of convergence([15], [16], [17]). The sequence x_k converges to the optimal point x_* super-linearly if the ratio between the errors at $k+1$ time and k time tends to zero as k approaches infinity, i.e.,

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = 0. \tag{2}$$

There have been several works on the asymptotic analysis of classical BFGS ([15],[17]). Recently, the nonasymptotic convergence analysis of the classical Quasi-Newton method has been carried out ([18]),([19]). Nonasymptotic convergence analysis gives more information about the complexity of the algorithm. In neural network problems, Nesterov accelerated BFGS works better than BFGS ([20] [21]). However, the nonasymptotic analysis of Nesterov accelerated BFGS has not yet been studied properly. Hence, in this paper, we are trying to find a nonasymptotic convergence analysis of Nesterov accelerated BFGS under certain assumptions. This analysis gives more insight into the complexity of Nesterov accelerated BFGS. We characterize an explicit upper bound on the error of NA-BFGS methods after a finite number of iterations. As a result, the overall complexity of NA-BFGS methods for achieving an ϵ-accurate solution, i.e., $\|x_k - x_*\| \leq \epsilon$, can be explicitly characterized.

Sec-(2) discusses some required notations used in our results. Sec-(3) describes the newly proposed Nesterov accelerated BFGS (NA-BFGS). In Sec-(4), we discuss our assumptions, which helps to prove our desired results, i.e., superlinear rate of convergence. We prove some necessary lemmas in Sec-(5). In Sec-(6), we demonstrate the linear rate of convergence of Nesterov accelerated BFGS. In Sec-(7), we prove our main theoretical result, i.e., superlinear convergence rate for NA-BFGS.

2 Notation

This section briefly discusses some notation and definitions used in theorems and their proofs. $[\nabla^2 f(x_*)]^{\frac{1}{2}}$ and $[\nabla^2 f(x_*)]^{-\frac{1}{2}}$ are the square root of $[\nabla^2 f(x_*)]$ and $[\nabla^2 f(x_*)]^{-1}$, respectively. Here, we use a weighted version of the Hessian approximation, i.e.,

$$\hat{B_k} = [\nabla^2 f(x_*)]^{-\frac{1}{2}} B_k [\nabla^2 f(x_*)]^{-\frac{1}{2}}. \tag{3}$$

Hence, $\hat{B_k}$ is real symmetric positive definite matrix as $[\nabla^2 f(x_*)]^{-\frac{1}{2}}$ and B_k are real symmetric positive definite matrix. Here, we use the weighted gradient difference $\hat{y}_k$, the weighted variable difference $\hat{s}_k$, the weighted gradient $\hat{\nabla} f(x_k)$, the weighted momentum vector $\hat{v}_k$ and the weighted average Hessian $\hat{J}_k$ such as

$$\hat{y}_k = [\nabla^2 f(x_*)]^{-\frac{1}{2}} y_k, \quad \hat{s}_k = [\nabla^2 f(x_*)]^{\frac{1}{2}} s_k, \quad \hat{\nabla} f(x_k) = [\nabla^2 f(x_*)]^{-\frac{1}{2}} \nabla f(x_k). \tag{4}$$

$$v_k = x_k - x_{k-1}, \ \hat{v}_k = [\nabla^2 f(x_*)]^{\frac{1}{2}} v_k. \tag{5}$$

$$J_k = \int_0^1 \left[\nabla^2 f(x_* + \alpha(x_k - x_*))\right] d\alpha, \quad \hat{J}_k = [\nabla^2 f(x_*)]^{-\frac{1}{2}} J_k [\nabla^2 f(x_*)]^{-\frac{1}{2}}. \tag{6}$$

In order to measure the closeness between the iterate x_k and the minima point x_*, we assume r_k, σ_k and τ_k in such a way that

$$r_k = [\nabla^2 f(x_*)]^{\frac{1}{2}} (x_k - x_*), \quad \sigma_k = \frac{M}{m^{\frac{3}{2}}} \|r_k\|, \quad \tau_k = max\left(\sigma_k + \frac{M\mu}{m^{\frac{3}{2}}} \|\hat{v}_k\|, \sigma_{k+1}\right). \tag{7}$$

3 Nesterov accelerated BFGS

Nesterov accelerated BFGS is the accelerated version of BFGS, which works better in many problems ([20], [21]). The momentum term in NA-BFGS is effective in reducing number of iteration.

Let us assume that the function f is C^2, i.e., the function f can be approximated quadratically in the neighborhood of $x_k + \mu v_k$. Hence, the quadratic model is given by

$$\hat{f}(x + p) = f(x) + \nabla f(x)^T p + \frac{p^T B p}{2}. \tag{8}$$

where B is the Hessian approximation of $f(x)$. The quadratic model is very accurate in approximating $f(x)$ when x is near the optimal point x_* ([1]). In NA-BFGS, the momentum is $v_k = x_{k+1} - x_k$, and the initial momentum is $v_0 = 0$. The momentum coefficient is $\mu \in (0, 1)$. Therefore, the first iteration behaves like classical BFGS. After the second iteration, the momentum term plays its part. In NA-BFGS, it considers the previous two iterations. The Nesterov accelerated BFGS ([20], [21]) is defined as

$$x_{k+1} = x_k + \mu v_k - W_k \nabla f(x_k + \mu v_k), \tag{9}$$

where W_k is the inverse Hessian approximation and updated as follows

$$W_{k+1} = \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right)^T W_k \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k}, \tag{10}$$

$s_k = x_{k+1} - (x_k + \mu v_k)$, $y_k = \nabla f(x_{k+1}) - \nabla f(x_k + \mu v_k)$, $\mu \in (0, 1)$. From the Taylor's expansion, we use the following equality in order to prove the superlinear convergence of Nesterov BFGS, i.e.,

$$\nabla f(x_k + \mu v_k) = \nabla f(x_k) + \mu \nabla^2 f(z_k) v_k. \tag{11}$$

for some z_k on the line joining between x_k and $x_k + \mu v_k$.

Algorithm 1 Nesterov accelerated BFGS ([20], [21])

Require: an initial guess $x_0 \in \mathbb{R}^d$, $W_0 \succ 0$ be the initial inverse Hessian approximation, initial momentum vector $v_0 = 0$ and $\mu \in (0, 1)$.

1: Let $k = 0$.
2: Compute $\nabla f(x_k)$;
3: **while** $(\|\nabla f(x_k)\| > \epsilon)$ **do**
4: Compute $v_{k+1} = \mu v_k - W_k \nabla f(x_k + \mu v_k)$;
5: Update $x_{k+1} = x_k + v_{k+1}$;
6: Compute $\nabla f(x_{k+1})$;
7: Update W_{k+1} using Equation-(10);
8: k=k+1;
9: **end while**

Here, we proposed an adaptive momentum parameter μ_k instead of μ to prove the local superlinear convergence of Nesterov accelerated BFGS.

4 Assumptions

We take the following assumptions on the function f.

1. f is a C^2 function, i.e., twice continuously differentiable function and f is strongly convex with the parameter m, i.e.,

$$\big\|\nabla f(x) - \nabla f(y)\big\| \geq m\big\|x - y\big\| \qquad \forall x, y \in R^d. \tag{12}$$

2. The gradient of f is Lipschitz continuous with parameter L, i.e.,

$$\big\|\nabla f(x) - \nabla f(y)\big\| \leq L\big\|x - y\big\| \qquad \forall x, y \in R^d. \tag{13}$$

3. The Hessian of f is Lipschitz continuous with parameter M at the optimal point x_*, i.e.,

$$\big\|\nabla^2 f(x) - \nabla^2 f(x_*)\big\| \leq M\big\|x - x_*\big\| \qquad \forall x \in R^d. \tag{14}$$

5 Important Lemmas

Lemma 1. *Let us assume that $f(x)$ satisfies Assumption-14, then*

$$\left\|\nabla f(x_1)-\nabla f(x_2)-\nabla^2 f(x_*)(x_1-x_2)\right\| \leq \frac{M}{2}\|x_1-x_2\|\left(\|x_1-x_*\|+\|x_2-x_*\|\right), \quad (15)$$

holds for all $x_1, x_2 \in \mathbb{R}^d$.

Proof. One can refer to [19]. □

Lemma 2. *From the notation section, we define σ_k from (7) and $\hat{J}_k$ from Equation (6). Let us define matrix $H_k = \nabla^2 f(x_* + \alpha_k(x_k - x_*))$, $\hat{H}_k = [\nabla^2 f(x_*)]^{-\frac{1}{2}} H_k \left[\nabla^2 f(x_*)\right]^{-\frac{1}{2}}$ and $\alpha_k \in [0,1]$. Also, assume that $f(x)$ satisfy Assumption-(12) and Assumption-(14). Then the following inequalities hold for all $k \geq 0$*

$$\frac{1}{(1+\frac{\sigma_k}{2})} I \preceq \hat{J}_k \preceq \left(1+\frac{\sigma_k}{2}\right) I, \quad (16)$$

and

$$\frac{1}{(1+\sigma_k)} I \preceq \hat{H}_k \preceq (1+\sigma_k) I. \quad (17)$$

Proof. The proof is similar to [19]. □

Lemma 3. *From the notation section, recall the definition of τ_k from Equation-(7) and let B_{k+1} be the inverse Hessian matrix generated by Nesterov accelerated BFGS. Let us assume that for some $k \geq 0$ and $\delta \geq 0$, we have that $\tau_k < 1$ and $\|\hat{B}_k - I\| \leq \delta$. Then B_{k+1} satisfies the following inequalities*

$$\left\|\hat{B}_{k+1} - I\right\|_F \leq \left\|\hat{B}_k - I\right\|_F - \frac{\hat{s}_k(\hat{B}_k - I)\hat{B}_k(\hat{B}_k - I)\hat{s}_k}{2\delta \hat{s}_k^T \hat{B}_k \hat{s}_k} + \frac{3+\sigma_k}{1-\sigma_k}\tau_k, \quad (18)$$

and

$$\|\hat{B}_{k+1} - I\|_F \leq \|(\hat{B}_k - I)\|_F + Z_k \tau_k, \quad (19)$$

where $Z_k = \frac{3+\sigma_k}{1-\sigma_k}$, $\tau_k = \max\left(\sigma_k + \frac{M\mu}{m^{\frac{3}{2}}}\|\hat{v}_k\|, \sigma_{k+1}\right)$.

Proof. One can refer to [19]. □

Lemma 4. *Let us assume that $f(x)$ satisfies Assumptions-(12-14). Then, the following inequalities hold for all $t \geq 0$.*

$$\|\hat{y}_t - \hat{s}_t\| \leq \tau_t \|\hat{s}_t\|, \quad (20)$$

$$(1-\tau_t)\|\hat{s}_t\|^2 \leq \hat{s}_t^T \hat{y}_t \leq (1+\tau_t)\|\hat{s}_t\|^2, \quad (21)$$

$$(1-\tau_t)\|\hat{s}_t\| \leq \|\hat{y}_t\| \leq (1+\tau_t)\|\hat{s}_t\|, \quad (22)$$

$$\|\hat{\nabla} f(x_t) - r_t\| \leq \frac{\sigma_t}{2}\|r_t\|. \tag{23}$$

Proof.

$$\begin{aligned}
\|\hat{y}_t - \hat{s}_t\| &= \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}} y_t - [\nabla^2 f(x_*)]^{\frac{1}{2}} s_t\right\| \\
&= \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \left\|y_t - \nabla^2 f(x_*) s_t\right\| \\
&= \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \Big\|\nabla f(x_{t+1}) - \nabla f(x_t + \mu v_t) \\
&\quad - \nabla^2 f(x_*)\big(x_{t+1} - (x_t + \mu v_t)\big)\Big\|.
\end{aligned}$$

Now using Lemma-(1) and strong convexity assumption, we get

$$\begin{aligned}
\|\hat{y}_t - \hat{s}_t\| &\leq \frac{M}{2m^{\frac{1}{2}}}\|s_t\|\big(\,\|x_{t+1} - x_*\| + \|(x_t + \mu v_t) - x_*\|\,\big) \\
&\leq \frac{M}{m^{\frac{1}{2}}}\|s_t\| \max\big(\|x_{t+1} - x_*\|, \|(x_t + \mu v_t) - x_*\|\big).
\end{aligned}$$

From the notation, $(x_k - x_*) = r_k[\nabla^2 f(x_*)]^{-\frac{1}{2}}$ and $\hat{v}_k = [\nabla^2 f(x_*)]^{\frac{1}{2}} v_k$, we have

$$\begin{aligned}
\max\left(\|x_{t+1} - x_*\|, \|(x_t + \mu v_t) - x_*\|\right) \leq \|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\| \\
\max\left(\|r_{t+1}\|, \|r_t\| + \mu\|\hat{v}_t\|\right).
\end{aligned}$$

and

$$\begin{aligned}
\|s_t\| &= \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}[\nabla^2 f(x_*)]^{\frac{1}{2}} s_t\right\| \\
&\leq \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \|\hat{s}_t\| \leq \frac{1}{m^{\frac{1}{2}}}\|\hat{s}_t\|.
\end{aligned}$$

As $\tau_t = \max\left(\sigma_{t+1}, \sigma_t + \frac{M\mu}{m^{\frac{3}{2}}}\|\hat{v}_t\|\right)$, therefore,

$$\|\hat{y}_t - \hat{s}_t\| \leq \frac{M}{m^{\frac{3}{2}}} \max\left(\|r_{t+1}\|, \|r_t\| + \mu\|\hat{v}_t\|\right)\|\hat{s}_t\| = \tau_t\|\hat{s}_t\|. \tag{24}$$

Hence, the first claim is proved. Then by using Cauchy Schwarz inequality, we have

$$|(\hat{y}_t - \hat{s}_t)^T \hat{s}_t| \leq \|\hat{y}_t - \hat{s}_t\|\|\hat{s}_t\| \leq \tau_t\|\hat{s}_t\|^2. \tag{25}$$

It implies that

$$(1 - \tau_t)\|\hat{s}_t\|^2 \leq \hat{s}_t{}^T \hat{y}_t \leq (1 + \tau_t)\|\hat{s}_t\|^2. \tag{26}$$

Therefore, the second claim is proved. Applying reverse triangle inequality and Equation-(24), we have

$$\left|\|\hat{y}_t\| - \|s_t\|\right| \leq \|\hat{y}_t - \hat{s}_t\| \leq \tau_t \|\hat{s}_t\|. \tag{27}$$

It implies $(1-\tau_t)\|\hat{s}_t\| \leq \|\hat{y}_t\| \leq (1+\tau_t)\|\hat{s}_t\|$. Hence, the third claim is proved.

$$\begin{aligned}\|\hat{\nabla} f(x_t) - r_t\| &= \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}} \nabla f(x_t) - [\nabla^2 f(x_*)]^{\frac{1}{2}}(x_t - x_*)\right\| \\ &\leq \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \left\|\nabla f(x_t) - \nabla f(x_*) - [\nabla^2 f(x_*)](x_t - x_*)\right\|.\end{aligned}$$

Putting $x_2 = x_*$ in Lemma-(1), we get

$$\begin{aligned}\|\hat{\nabla} f(x_t) - r_t\| &\leq \frac{M}{2m^{\frac{1}{2}}} \|x_t - x_*\|^2 \\ &= \frac{M}{2m^{\frac{1}{2}}} \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}} [\nabla^2 f(x_*)]^{\frac{1}{2}}(x_t - x_*)\right\|^2 \\ &\leq \frac{M}{2m^{\frac{3}{2}}} \|r_t\|^2 = \frac{\sigma_t}{2}\|r_t\|.\end{aligned}$$

Hence, the fourth claim is proved. □

6 Linear Convergence of Nesterov Accelerated BFGS

We take the following assumptions to prove the linear convergence of Nesterov Accelerated BFGS.

1. Let us consider an initial point x_0 such that

$$\sigma_0 = \frac{M}{m^{\frac{3}{2}}} \left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 - x_*)\right\| \leq \epsilon. \tag{28}$$

2. Let us choose the initial Hessian approximation B_0 satisfies

$$\left\|\hat{B}_0 - I\right\|_F = \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}} \left(B_0 - \nabla^2 f(x_*)\right) [\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \leq \delta, \tag{29}$$

where $\epsilon \in (0, \frac{1}{3}), \delta \in (0, \frac{1}{2})$ such that for some $\rho, \mu_k \in (0,1)$, they satisfy

$$\frac{3+\epsilon}{1-\epsilon}\left(\frac{3\epsilon}{1-\rho}\right) \leq \delta, \tag{30}$$

$$\rho \geq \frac{1}{1-2\delta}\left[\frac{\epsilon}{2} + 2\delta\right], \tag{31}$$

$$\mu_k \leq \frac{\sigma_k(1-\rho)}{2g(1-2\delta)(1+\rho)} \tag{32}$$

where $g = \frac{\frac{L}{m}+(1+2\delta)}{1-2\delta}$.

Proposition 1. *Let us assume that $f(x)$ satisfies Assumptions-(12–14) and (28–32). Then*

$$\sigma_1 \leq \rho\sigma_0, \qquad \left\|\hat{B}_0 - I\right\|_F \leq 2\delta, \tag{33}$$

$$\left\|\hat{B}_0\right\| \leq 1+2\delta, \qquad \left\|\hat{B}_0^{-1}\right\| \leq \frac{1}{1-2\delta}, \tag{34}$$

where $\rho \geq \frac{1}{1-2\delta}\left(\frac{\epsilon}{2}+2\delta\right)$.

Proof. As per our Assumption-(29),

$$\|\hat{B}_0 - I\|_F = \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}(B_0 - \nabla^2 f(x_*))[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \leq \delta. \tag{35}$$

Therefore, $\|\hat{B}_0 - I\|_F \leq 2\delta$. From Equation-(35), all the eigenvalues of $\hat{B}_0$ are in $[1-2\delta, 1+2\delta]$. Let $\lambda_{\max}(\hat{B}_0)$ and $\lambda_{\min}(\hat{B}_0)$ are the largest eigenvalue and the smallest eigenvalue of $\hat{B}_0$. Then we have

$$\left\|\hat{B}_0\right\| = \lambda_{\max}(\hat{B}_0) \leq 1+2\delta, \tag{36}$$

and

$$\left\|\hat{B}_0^{-1}\right\| = \frac{1}{\lambda_{\min}(\hat{B}_0)} \leq \frac{1}{1-2\delta}. \tag{37}$$

Hence,

$$\left\|\hat{B}_0\right\| \leq 1+2\delta \quad \text{and} \quad \left\|\hat{B}_0^{-1}\right\| \leq \frac{1}{1-2\delta}. \tag{38}$$

We must show that $\sigma_1 \leq \rho\sigma_0$. As $\sigma_k = \frac{M}{m^{\frac{3}{2}}}\|r_k\|$, we have

$$\begin{aligned}\sigma_1 &= \frac{M}{m^{\frac{3}{2}}}\|r_1\| = \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_1 - x_*)\right\| \\ &= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 + \mu_0 v_0 - B_0^{-1}(\nabla f(x_0+\mu_0 v_0)) - x_*)\right\|.\end{aligned}$$

As $v_0 = 0$, we have

$$= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 - B_0^{-1}\nabla f(x_0) - x_*)\right\|$$

$$
\begin{aligned}
&= \frac{M}{m^{\frac{3}{2}}} \Big\| [\nabla^2 f(x_*)]^{\frac{1}{2}} B_0^{-1} \big[\nabla f(x_0) - \nabla^2 f(x_*)(x_0 - x_*) \\
&\qquad\qquad - (B_0 - \nabla^2 f(x_*))(x_0 - x_*)\big] \Big\| \\
&= \frac{M}{m^{\frac{3}{2}}} \Big\| [\nabla^2 f(x_*)]^{\frac{1}{2}} B_0^{-1} [\nabla^2 f(x_*)]^{\frac{1}{2}} \big([\nabla^2 f(x_*)]^{\frac{-1}{2}} \nabla f(x_0) \\
&- [\nabla^2 f(x_*)]^{\frac{-1}{2}} \nabla^2 f(x_*)(x_0 - x_*) - [\nabla^2 f(x_*)]^{\frac{-1}{2}} (B_0 - \nabla^2 f(x_*)) \\
&\qquad\qquad [\nabla^2 f(x_*)]^{\frac{-1}{2}} [\nabla^2 f(x_*)]^{\frac{1}{2}} (x_0 - x_*) \big) \Big\| \\
&\leq \frac{M}{m^{\frac{3}{2}}} \left\| \hat{B}_0^{-1} \right\| \left\| \left[\hat{\nabla} f(x_0) - r_0 - \left(\hat{B}_0 - I \right) r_0 \right] \right\| \\
&\leq \frac{M}{m^{\frac{3}{2}}} \left\| \hat{B}_0^{-1} \right\| \left[\|\hat{\nabla} f(x_0) - r_0\| + \|\hat{B}_0 - I\| \|r_0\| \right].
\end{aligned}
$$

Using Equation-(38), Lemma-(4),

$$
\begin{aligned}
\sigma_1 &\leq \frac{M}{m^{\frac{3}{2}}} \frac{1}{1 - 2\delta} \left[\frac{\sigma_0}{2} \|r_0\| + 2\delta \|r_0\| \right] \\
&\leq \frac{M \|r_0\|}{m^{\frac{3}{2}}} \left(\frac{1}{1 - 2\delta} \left[\frac{\epsilon}{2} + 2\delta \right] \right) \leq \rho \sigma_0,
\end{aligned}
$$

where $\rho \geq \frac{1}{1-2\delta} \left[\frac{\epsilon}{2} + 2\delta \right]$. □

Proposition 2. *Let us assume that $f(x)$ satisfies Assumptions-(12–14) and (28–32). Then*

$$
\sigma_2 \leq \rho \sigma_1, \qquad \left\| \hat{B}_1 - I \right\|_F \leq 2\delta, \tag{39}
$$

$$
\left\| \hat{B}_1 \right\| \leq 1 + 2\delta, \qquad \left\| \hat{B}_1^{-1} \right\| \leq \frac{1}{1 - 2\delta}, \tag{40}
$$

where $\rho \geq \frac{1}{1-2\delta} \left[\frac{\epsilon}{2} + 2\delta \right]$.

Proof. From Lemma-(3), $\|\hat{B}_1 - I\|_F \leq \|(\hat{B}_0 - I)\|_F + Z_0 \tau_0$ and from our assumption (29), we have $\|(\hat{B}_0 - I)\|_F \leq \delta$ and

$$
Z_0 = \frac{3 + \sigma_0}{1 - \sigma_0} \leq \frac{3 + \epsilon}{1 - \epsilon}.
$$

$$
\tau_0 = \max \left(\sigma_0 + \frac{M \mu_0}{m^{\frac{3}{2}}} \|\hat{v}_0\|, \sigma_1 \right) = \sigma_0 \leq \epsilon.
$$

Hence

$$
Z_0 \tau_0 \leq \frac{3 + \epsilon}{1 - \epsilon} \epsilon \leq \delta.
$$

Therefore,

$$\|\hat{B}_1 - I\|_F \leq \|(\hat{B}_0 - I)\|_F + Z_0\tau_0 \leq \delta + \delta = 2\delta. \tag{41}$$

Similarly, we can easily show that

$$\left\|\hat{B}_1\right\| \leq 1 + 2\delta, \qquad \left\|\hat{B}_1^{-1}\right\| \leq \frac{1}{1-2\delta}. \tag{42}$$

$$\begin{aligned}
\sigma_2 &= \frac{M}{m^{\frac{3}{2}}}\|r_2\| = \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_2 - x_*)\right\| \\
&= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_1 + \mu_1 v_1 - B_1^{-1}(\nabla f(x_1 + \mu_1 v_1)) - x_*)\right\|.
\end{aligned}$$

Using Equation-(11), we have

$$\begin{aligned}
\sigma_2 &= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_1 + \mu_1 v_1 - B_1^{-1}(\nabla f(x_1) + \mu_1 \nabla^2 f(z_1)v_1) - x_*)\right\| \\
&= \frac{M}{m^{\frac{3}{2}}}\Big\|[\nabla^2 f(x_*)]^{\frac{1}{2}} B_1^{-1}[\nabla^2 f(x_*)]^{\frac{1}{2}}[\nabla^2 f(x_*)]^{\frac{-1}{2}}\big[\nabla f(x_1) + \mu_1\nabla^2 f(z_1)v_1 \\
&\qquad - \nabla^2 f(x_*)(x_1 - x_*) - (B_1 - \nabla^2 f(x_*))(x_1 - x_*) - \mu_1 B_1 v_1\big]\Big\| \\
&= \frac{M}{m^{\frac{3}{2}}}\big\|\hat{B_1}^{-1}\big\|\Big\|[\nabla^2 f(x_*)]^{\frac{-1}{2}}\big[\nabla f(x_1) + \mu_1\nabla^2 f(z_1)v_1 - \nabla^2 f(x_*)(x_1 - x_*) \\
&\qquad - (B_1 - \nabla^2 f(x_*))(x_1 - x_*) - \mu_1 B_1 v_1\big]\Big\| \\
&= \frac{M}{m^{\frac{3}{2}}}\big\|\hat{B_1}^{-1}\big\|\Big\|\big[\hat{\nabla} f(x_1) + [\nabla^2 f(x_*)]^{\frac{-1}{2}}\mu_1\nabla^2 f(z_1)v_1 - r_1 - [\nabla^2 f(x_*)]^{\frac{-1}{2}} \\
&\quad \big(B_1 - \nabla^2 f(x_*)\big)[\nabla^2 f(x_*)]^{\frac{-1}{2}}[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_1 - x_*) - \mu_1\nabla^2 f(x_*)]^{\frac{-1}{2}} B_1 \\
&\qquad \nabla^2 f(x_*)]^{\frac{-1}{2}}[\nabla^2 f(x_*)]^{\frac{1}{2}} v_1\big]\Big\| \\
&= \frac{M}{m^{\frac{3}{2}}}\left\|\hat{B_1}^{-1}\right\|\Big\|\big[\hat{\nabla} f(x_1) + [\nabla^2 f(x_*)]^{\frac{-1}{2}}\mu_1\nabla^2 f(z_1)v_1 - r_1 - \left(\hat{B}_1 - I\right) r_1 \\
&\qquad - \mu_1\hat{B}_1[\nabla^2 f(x_*)]^{\frac{1}{2}} v_1\big]\Big\| \\
&\leq \frac{M}{m^{\frac{3}{2}}}\left\|\hat{B_1}^{-1}\right\|\Big[\left(\|\hat{\nabla} f(x_1) - r_1\| + \|\hat{B}_1 - I\|\|r_1\|\right) + \mu_1\big(\|[\nabla^2 f(x_*)]^{\frac{-1}{2}} \\
&\qquad \nabla^2 f(z_1)v_1\| + \|\hat{B}_1[\nabla^2 f(x_*)]^{\frac{1}{2}} v_1\|\big)\Big].
\end{aligned}$$

Hence, we get

$$\begin{aligned}
\sigma_2 \leq \frac{M}{m^{\frac{3}{2}}}\big\|\hat{B_1}^{-1}\big\|\big[&\big(\|\hat{\nabla} f(x_1) - r_1\| + \|\hat{B}_1 - I\|\|r_1\|\big) \\
&+ \mu_1\big(\|[\nabla^2 f(x_*)]^{\frac{-1}{2}}\nabla^2 f(z_1)v_1\| + \|\hat{B}_1[\nabla^2 f(x_*)]^{\frac{1}{2}} v_1\|\big)\big].
\end{aligned} \tag{43}$$

$$\left\|[\nabla^2 f(x_*)]^{\frac{-1}{2}}\nabla^2 f(z_1)v_1\right\| = \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\nabla^2 f(z_1)(x_1 - x_0)\right\|$$

$$
\begin{aligned}
&= \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \left\|\nabla^2 f(z_1)\right\| \|(x_1 - x_*) + (x_* - x_0)\| \\
&= \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \left\|\nabla^2 f(z_1)\right\| \|[\nabla^2 f(x_*)]^{-\frac{1}{2}}[\nabla^2 f(x_*)]^{\frac{1}{2}} \\
&\qquad\qquad \big((x_1 - x_*) + (x_* - x_0)\big)\| \\
&\leq \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \left\|\nabla^2 f(z_1)\right\| \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \|[\nabla^2 f(x_*)]^{\frac{1}{2}} \\
&\qquad\qquad \big((x_1 - x_*) + (x_* - x_0)\big)\| \\
&\leq \frac{L}{m^{1/2}} \left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}} (r_1 + r_0)\right\| \\
&\leq \frac{L}{m} \left(\|r_1\| + \|r_0\|\right).
\end{aligned}
$$

$$
\|[\nabla^2 f(x_*)]^{\frac{-1}{2}} \nabla^2 f(z_1) v_1\| \leq \frac{L}{m} \left(\|r_1\| + \|r_0\|\right). \tag{44}
$$

$$
\begin{aligned}
\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}} v_1\right\| &= \left\|[\nabla^2 f(x_*)]^{\frac{1}{2}} (x_1 - x_0)\right\| \\
&= \left\|[\nabla^2 f(x_*)]^{\frac{1}{2}} (x_1 - x_* + x_* - x_0)\right\| \\
&\leq \|r_1\| + \|r_0\|.
\end{aligned}
$$

Hence,

$$
\|\hat{B}_1 [\nabla^2 f(x_*)]^{\frac{1}{2}} v_1\| \leq (1 + 2\delta)(\|r_1\| + \|r_0\|). \tag{45}
$$

Now using Lemma-(4), Equation-(45), Equation-(44), Equation-(41) and put the bounds in Equation-(43), we have

$$
\begin{aligned}
\sigma_2 &\leq \frac{M}{m^{\frac{3}{2}}} \frac{1}{1 - 2\delta} \big[\frac{\sigma_1}{2} \|r_1\| + 2\delta \|r_1\| + \frac{\mu_1 L}{m} (\|r_1\| + \|r_0\|) + (1 + 2\delta)\mu_1 \big(\|r_1\| \\
&\qquad\qquad + \|r_0\|\big)\big] \\
&= \frac{M\|r_1\|}{m^{\frac{3}{2}}} \frac{1}{1 - 2\delta} \left[\frac{\sigma_1}{2} + 2\delta + \big(\frac{\mu_1 L}{m} + \mu_1(1 + 2\delta)\big)\right] + \frac{M\|r_0\|}{m^{\frac{3}{2}}} \frac{1}{1 - 2\delta} \\
&\qquad\qquad \big(\frac{\mu_1 L}{m} + \mu_1(1 + 2\delta)\big) \\
&\leq \frac{\sigma_1}{1 - 2\delta} \left[\frac{\sigma_1}{2} + 2\delta + \big(\frac{\mu_1 L}{m} + \mu_1(1 + 2\delta)\big)\right] + \frac{\sigma_0}{1 - 2\delta} \left[\frac{\mu_1 L}{m} + \mu_1(1 + 2\delta)\right].
\end{aligned}
$$

As $\mu_1 \leq \frac{(1-\rho)\sigma_1}{2g(1-2\delta)(1+\rho)}$ and take $g = \left[\frac{\frac{L}{m} + (1+2\delta)}{(1-2\delta)}\right]$

$$
\begin{aligned}
\sigma_2 &\leq \frac{\sigma_1^2}{2(1 - 2\delta)} + \frac{2\delta\sigma_1}{(1 - 2\delta)} + \mu_1(\sigma_1 + \sigma_0)g \\
&\leq \frac{\sigma_1^2}{2(1 - 2\delta)} + \frac{2\delta\sigma_1}{(1 - 2\delta)} + (\sigma_1 + \sigma_0)g \frac{(1 - \rho)\sigma_1}{2g(1 - 2\delta)(1 + \rho)}
\end{aligned}
$$

$$
\begin{aligned}
&= \sigma_1 \left[\frac{\sigma_1}{2(1-2\delta)} + \frac{2\delta}{1-2\delta} + (\sigma_1+\sigma_0)\frac{(1-\rho)}{2(1-2\delta)(1+\rho)} \right] \\
&\leq \sigma_1 \left[\frac{\rho\sigma_0}{2(1-2\delta)} + \frac{2\delta}{1-2\delta} + \sigma_0(1+\rho)\frac{(1-\rho)}{2(1-2\delta)(1+\rho)} \right] \\
&= \sigma_1 \left[\frac{2\delta}{1-2\delta} + \frac{\sigma_0}{2(1-2\delta)} \right] = \sigma_1 \left[\frac{\left(\frac{\sigma_0}{2}+2\delta\right)}{(1-2\delta)} \right] \\
&\leq \sigma_1 \left[\frac{\left(\frac{\epsilon}{2}+2\delta\right)}{(1-2\delta)} \right] \\
&\leq \rho\sigma_1.
\end{aligned}
$$

□

Theorem 1. *Let us assume that $f(x)$ satisfies Assumptions-$(12-14)$ and $(28-32)$. Then the sequence of iterate x_k generated by the Nesterov accelerated BFGS algorithm (2) converges to an optimal solution x_* with*

$$\sigma_{k+1} \leq \rho\sigma_k, \qquad \forall k \geq 0. \tag{46}$$

where $\rho \geq \frac{1}{1-2\delta}\left[\frac{\epsilon}{2}+2\delta\right]$. Further, $\left(\|\hat{B}_k\|\right)_{k=0}^{k=\infty}$ lie in a neighbourhood $\nabla^2 f(x_*)$ defined as

$$\left\|\hat{B}_k - I\right\|_F \leq 2\delta, \qquad \forall k \geq 0. \tag{47}$$

Besides, $\left(\left\|\hat{B}_k\right\|\right)_{k=0}^{k=\infty}$ and $\left(\left\|\hat{B}_k^{-1}\right\|\right)_{k=0}^{k=\infty}$ are uniformly bounded by

$$\left\|\hat{B}_k\right\| \leq 1+2\delta, \qquad \left\|\hat{B}_k^{-1}\right\| \leq \frac{1}{1-2\delta}. \tag{48}$$

Proof. We use the induction method to prove all the inequality and linear convergence of Nesterov accelerated BFGS. From Preposition-(1), we have $\sigma_1 \leq \rho\sigma_0$, $\left\|\hat{B}_0 - I\right\|_F \leq 2\delta$, $\left\|\hat{B}_0\right\| \leq 1+2\delta$, $\left\|\hat{B}_0^{-1}\right\| \leq \frac{1}{1-2\delta}$, where $\rho \geq \frac{1}{1-2\delta}\left(\frac{\epsilon}{2}+2\delta\right)$. Nesterov accelerated BFGS behaves like a classical BFGS in the first iteration. After the first iteration, acceleration is added. Then we have $\sigma_2 \leq \rho\sigma_1$, $\left\|\hat{B}_1 - I\right\|_F \leq 2\delta$, $\left\|\hat{B}_1\right\| \leq 1+2\delta$, $\left\|\hat{B}_1^{-1}\right\| \leq \frac{1}{1-2\delta}$. This shows that all the conditions are satisfied for $k = 0, 1$. Then, assume all the conditions are true for $0 \leq k \leq t$. Hence,

$$\|\hat{B}_t - I\|_F \leq 2\delta, \qquad \left\|\hat{B}_t\right\| \leq 1+2\delta, \qquad \left\|\hat{B}_t^{-1}\right\| \leq \frac{1}{1-2\delta}, \qquad \sigma_t \leq \rho\sigma_{t-1}, \tag{49}$$

where $\rho \geq \frac{1}{1-2\delta}\left[\frac{\epsilon}{2}+2\delta\right]$. Now, we have to prove for $k = t+1$. Since the condition from Equation-(47) is satisfied for $0 \leq k \leq t$, i.e., $\|\hat{B}_k - I\|_F \leq 2\delta$ for $0 \leq k \leq t$, now

we have to show for $k = t+1$, i.e., $\|\hat{B}_{t+1} - I\|_F \leq 2\delta$. From Lemma-(3), we have

$$\left\|\hat{B}_{k+1} - I\right\|_F \leq \left\|(\hat{B}_k - I)\right\|_F + Z_k\tau_k, \tag{50}$$

where $Z_k = \frac{3+\sigma_k}{1-\sigma_k}$, $\tau_k = \max\left(\sigma_k + \frac{M\mu}{m^{\frac{3}{2}}}\left\|\hat{v_k}\right\|, \sigma_{k+1}\right)$.

$$\begin{aligned}
\tau_k &= \max\left(\sigma_k + \frac{M\mu_k}{m^{\frac{3}{2}}}\left\|\hat{v}_k\right\|, \sigma_{k+1}\right) \\
&= \sigma_k + \frac{M\mu_k}{m^{\frac{3}{2}}}\left\|\hat{v}_k\right\| \\
&\leq \epsilon + \frac{M\mu_k}{m^{\frac{3}{2}}}\left\|\hat{v}_k\right\|.
\end{aligned}$$

$$\begin{aligned}
\left\|\hat{v}_k\right\| &= \left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_k - x_* + x_* - x_{k-1})\right\| \\
&\leq \left\|r_k\right\| + \left\|r_{k-1}\right\|.
\end{aligned}$$

Hence,

$$\left\|\hat{v}_k\right\| \leq \left\|r_k\right\| + \left\|r_{k-1}\right\|. \tag{51}$$

Hence,

$$\begin{aligned}
\tau_k &\leq \epsilon + \frac{M\mu_k}{m^{3/2}}\|\hat{v_k}\| \\
&\leq \epsilon + \frac{M\mu_k}{m^{3/2}}\left(\left\|r_k\right\| + \left\|r_{k-1}\right\|\right) \\
&= \epsilon + \mu_k\sigma_k + \mu_k\sigma_{k-1} \leq 3\epsilon < 1,
\end{aligned}$$

for $0 \leq k \leq t$. From the induction, we have $\|\hat{B}_k - I\|_F \leq 2\delta$ for $0 \leq k \leq t$, Next, we have to show for $k = t+1$. Hence,

$$\sigma_k \leq \rho\sigma_{k-1} \leq \rho^2\sigma_{k-2} \cdots \leq \rho^k\sigma_0.$$

Now, for $0 \leq k \leq t$, $\sigma_k \leq \epsilon$. We have

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

As $v_0 = 0$, we have

$$\sum_{k=0}^{t}\frac{M\mu_k}{m^{3/2}}\|\hat{v_k}\| = \sum_{k=1}^{t}\frac{M\mu_k}{m^{3/2}}\|\hat{v_k}\| \leq \sum_{k=1}^{t}\frac{M\mu_k}{m^{3/2}}\left(\|r_k\| + \|r_{k-1}\|\right)$$

$$= \sum_{k=1}^{t} \mu_k \big(\sigma_k + \sigma_{k-1}\big) \leq \sum_{k=1}^{t} \mu_k \big(\rho^k \sigma_0 + \rho^{k-1}\sigma_0\big)$$
$$\leq \sum_{k=1}^{t} \mu_k \sigma_0 \big(\rho^k + \rho^{k-1}\big).$$

As $\mu_k < 1$ and $\sigma_0 \leq \epsilon$ from our assumption, we have

$$\leq \sigma_0 \sum_{k=1}^{t} \big(\rho^k + \rho^{k-1}\big)$$
$$\leq \sigma_0 \sum_{k=1}^{t} \rho^{k-1}(1+\rho)$$
$$\leq \epsilon\big(\frac{1+\rho}{1-\rho}\big) \leq \frac{2\epsilon}{1-\rho}.$$

$$\sum_{k=0}^{k=t} \tau_k \leq \sum_{k=0}^{k=t} \left(\sigma_k + \frac{M\mu_k}{m^{\frac{3}{2}}} \big\|\hat{v}_k\big\|\right) \leq \sum_{k=0}^{k=t} \sigma_k + \sum_{k=0}^{k=t} \frac{M\mu_k}{m^{3/2}} \|\hat{v_k}\| \leq \frac{3\epsilon}{1-\rho}.$$

Taking sum from $k = 0$ to $k = t$ on both sides from Equation-(50) and $Z_k = \frac{3+\sigma_k}{1-\sigma_k} \leq \frac{3+\epsilon}{1-\epsilon}$ for $0 \leq k \leq t$, we have

$$\big\|\hat{B}_{t+1} - I\big\|_F \leq \big\|(\hat{B}_0 - I)\big\|_F + \sum_{k=0}^{k=t} Z_k \tau_k$$
$$\leq \delta + \frac{3+\epsilon}{1-\epsilon}\left(\frac{3\epsilon}{1-\rho}\right).$$

From our assumption, $\frac{3+\epsilon}{1-\epsilon}\left(\frac{3\epsilon}{1-\rho}\right) \leq \delta$, then

$$\big\|\hat{B}_{t+1} - I\big\|_F \leq 2\delta.$$

It implies that the above inequality holds for $k = t+1$. Since $\|\hat{B}_{t+1} - I\|_F \leq 2\delta$, therefore all the eigenvalues of $\hat{B}_{t+1}$ lies in $[1-2\delta, 1+2\delta]$. Using the same argument, let us assume $\lambda_{max}(\hat{B}_{t+1})$ and $\lambda_{min}(\hat{B}_{t+1})$ be the largest eigenvalue and the smallest eigenvalue of $\hat{B}_{t+1}$. Then we have

$$\Big\|\hat{B}_{t+1}\Big\| = \lambda_{max}(\hat{B}_{t+1}) \leq 1 + 2\delta, \tag{52}$$

and

$$\Big\|\hat{B}_{t+1}^{-1}\Big\| = \frac{1}{\lambda_{min}(\hat{B}_{t+1})} \leq \frac{1}{1-2\delta}. \tag{53}$$

Hence,

$$\|\hat{B}_{t+1} - I\|_F \leq 2\delta, \qquad \left\|\hat{B}_{t+1}\right\| \leq 1 + 2\delta, \qquad \left\|\hat{B}_{t+1}^{-1}\right\| \leq \frac{1}{1-2\delta}. \tag{54}$$

$$\begin{aligned}
\sigma_{t+1} &= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_t + \mu_t v_t - B_t^{-1}\big(\nabla f(x_t + \mu_t v_t)) - x_*\big)\right\| \\
&= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_t + \mu_t v_t - B_t^{-1}\big(\nabla f(x_t) + \mu_t \nabla^2 f(z_t) v_t - x_*\big)\right\|,
\end{aligned}$$

where $z_t \in (x_t, x_t + \mu_t v_t)$.

$$\begin{aligned}
\sigma_{t+1} &= \frac{M}{m^{\frac{3}{2}}}\Big\|[\nabla^2 f(x_*)]^{\frac{1}{2}} B_t^{-1}\big[\nabla f(x_t) + \mu_t \nabla^2 f(z_t) v_t - \nabla^2 f(x_*)(x_t - x_*) \\
&\qquad - (B_t - \nabla^2 f(x_*))(x_t - x_*) - \mu_t B_t v_t\big]\Big\| \\
&= \frac{M}{m^{\frac{3}{2}}}\Big\|[\nabla^2 f(x_*)]^{\frac{1}{2}} B_t^{-1}[\nabla^2 f(x_*)]^{\frac{1}{2}}[\nabla^2 f(x_*)]^{\frac{-1}{2}}\big[\nabla f(x_t) + \mu_t \nabla^2 f(z_t) v_t \\
&\qquad - \nabla^2 f(x_*)(x_t - x_*) - (B_t - \nabla^2 f(x_*))(x_t - x_*) - \mu_t B_t v_t\big]\Big\| \\
&= \frac{M}{m^{\frac{3}{2}}}\big\|\hat{B}_t^{-1}\big\|\Big\|\big[\hat{\nabla} f(x_t) + [\nabla^2 f(x_*)]^{\frac{-1}{2}}\mu_t \nabla^2 f(z_t) v_t - r_t - [\nabla^2 f(x_*)]^{\frac{-1}{2}} \\
&\qquad \big(B_t - \nabla^2 f(x_*)\big)[\nabla^2 f(x_*)]^{\frac{-1}{2}}[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_t - x_*) - \mu_t[\nabla^2 f(x_*)]^{\frac{-1}{2}} \\
&\qquad\qquad B_t[\nabla^2 f(x_*)]^{\frac{-1}{2}}[\nabla^2 f(x_*)]^{\frac{1}{2}} v_t\big]\Big\| \\
&\leq \frac{M}{m^{\frac{3}{2}}}\big\|\hat{B}_t^{-1}\big\|\Big\|\big[\hat{\nabla} f(x_t) + [\nabla^2 f(x_*)]^{\frac{-1}{2}}\mu_t \nabla^2 f(z_t) v_t - r_t - \big(\hat{B}_t - I\big) r_t \\
&\qquad - \mu_t \hat{B}_t[\nabla^2 f(x_*)]^{\frac{1}{2}} v_t\big]\Big\| \\
&\leq \frac{M}{m^{\frac{3}{2}}}\big\|\hat{B}_t^{-1}\big\|\big[\big(\|\hat{\nabla} f(x_t) - r_t\| + \|\hat{B}_t - I\|\|r_t\|\big) + \mu_t\big(\|[\nabla^2 f(x_*)]^{\frac{-1}{2}} \\
&\qquad\qquad \nabla^2 f(z_t) v_t\| + \|\hat{B}_t[\nabla^2 f(x_*)]^{\frac{1}{2}} v_t\|\big)\big].
\end{aligned}$$

From Equation-(12) and Equation-(13), we have

$$\|[\nabla^2 f(x_*)]^{-1/2}\nabla^2 f(z_t) v_t\| \leq \frac{L}{\sqrt{m}}\|v_t\|.$$

From Equation-(51), we have

$$\begin{aligned}
&\leq \frac{L}{\sqrt{m}}\|[\nabla^2 f(x_*)]^{-1/2}\|\left(\|r_t\| + \|r_{t-1}\|\right) \\
&\leq \frac{L}{m}\left(\|r_t\| + \|r_{t-1}\|\right).
\end{aligned}$$

Therefore,

$$\|[\nabla^2 f(x_*)]^{\frac{-1}{2}} \nabla^2 f(z_t) v_t\| \leq \frac{L}{m} \left(\|r_t\| + \|r_{t-1}\|\right). \tag{55}$$

$$\begin{aligned}\left\|\hat{B}_t[\nabla^2 f(x_*)]^{\frac{1}{2}} v_t\right\| &= \left\|\hat{B}_t[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_t - x_* + x_* - x_{t-1})\right\| \\ &\leq (1+2\delta)\big(\|r_t\| + \|r_{t-1}\|\big).\end{aligned}$$

Therefore,

$$\left\|\hat{B}_t[\nabla^2 f(x_*)]^{\frac{1}{2}} v_t\right\| \leq (1+2\delta)\big(\|r_t\| + \|r_{t-1}\|\big). \tag{56}$$

Using Equations-[54-56] and Lemma-(4), We have,

$$\begin{aligned}\sigma_{t+1} &\leq \frac{M}{m^{\frac{3}{2}}} \frac{1}{1-2\delta}\Big[\frac{\sigma_t}{2}\|r_t\| + 2\delta\|r_t\| + \frac{L\mu_t}{m}(\|r_t\| + \|r_{t-1}\|) + \mu_t(1+2\delta)(\|r_t\| \\ &\quad + \|r_{t-1}\|)\Big] \\ &= \frac{M\|r_t\|}{m^{\frac{3}{2}}} \frac{1}{1-2\delta}\left[\frac{\sigma_t}{2} + 2\delta + \frac{L\mu_t}{m} + \mu_t(1+2\delta)\right] + \frac{M\|r_{t-1}\|}{m^{\frac{3}{2}}} \frac{1}{1-2\delta} \\ &\qquad \left[\frac{L\mu_t}{m} + \mu_t(1+2\delta)\right] \\ &= \frac{\sigma_t}{1-2\delta}\left[\frac{\sigma_t}{2} + 2\delta + \big(\frac{\mu_t L}{m} + \mu_t(1+2\delta)\big)\right] + \frac{\sigma_{t-1}}{1-2\delta}\big(\frac{\mu_t L}{m} + \mu_t(1+2\delta)\big).\end{aligned}$$

As $\mu_t \leq \frac{(1-\rho)\sigma_t}{2g(1-2\delta)(1+\rho)}$ and taking $g := \left[\frac{\frac{L}{m}+(1+2\delta)}{(1-2\delta)}\right]$, we get

$$\begin{aligned}\sigma_{t+1} &\leq \frac{\sigma_t^2}{2(1-2\delta)} + \frac{2\delta\sigma_t}{(1-2\delta)} + \mu_t(\sigma_t + \sigma_{t-1})g \\ &\leq \frac{\sigma_t^2}{2(1-2\delta)} + \frac{2\delta\sigma_t}{(1-2\delta)} + (\sigma_t + \sigma_{t-1})g\frac{(1-\rho)\sigma_t}{2g(1-2\delta)(1+\rho)} \\ &= \sigma_t\left[\frac{\sigma_t}{2(1-2\delta)} + \frac{2\delta}{1-2\delta} + (\sigma_t + \sigma_{t-1})\frac{(1-\rho)}{2(1-2\delta)(1+\rho)}\right] \\ &\leq \sigma_t\left[\frac{\rho\sigma_{t-1}}{2(1-2\delta)} + \frac{2\delta}{1-2\delta} + \sigma_{t-1}(1+\rho)\frac{(1-\rho)}{2(1-2\delta)(1+\rho)}\right] \\ &= \sigma_t\left[\frac{2\delta}{1-2\delta} + \frac{\sigma_{t-1}}{2(1-2\delta)}\right] = \sigma_t\left[\frac{\left(\frac{\sigma_{t-1}}{2} + 2\delta\right)}{(1-2\delta)}\right] \\ &\leq \sigma_t\left[\frac{\left(\frac{\epsilon}{2} + 2\delta\right)}{(1-2\delta)}\right] \\ &\leq \rho\sigma_t.\end{aligned}$$

Hence, it is true for $k = t+1$. Therefore,

$$\sigma_{k+1} \leq \rho\sigma_k. \tag{57}$$

Therefore, all the above inequality is true for $k = t + 1$, and our induction step is complete. □

Using Theorem-1, we assume

$$\mu_k \leq \frac{\rho^k \sigma_0 (1-\rho)}{2g(1-2\delta)(1+\rho)} \tag{58}$$

where $g = \frac{\frac{L}{m}+(1+2\delta)}{1-2\delta}$.

Remark 1. *Here, we choose ϵ and δ and ρ satisfying Equation-30 and Equation-31. The assumptions on the triplet (ϵ, δ, ρ) are the same as in ([19]). The only difference is that $\epsilon \in (0, \frac{1}{3})$ instead of $(0, \frac{1}{2})$. Equations-(30, 31) give an upper and a lower bounds of ρ. Also, one may notice that for the smooth movement of the convergence analysis, ρ should be chosen near 0 and avoided near 1. For the ρ We have to choose μ_0 such that $0 < \mu_0 \leq \frac{\sigma_0(1-\rho)}{2g(1-2\delta)(1+\rho)} < 1$. For implementation purpose, from (32), we choose $\mu_1 = \rho\mu_0$, $\mu_2 = \rho^2\mu_0$, $\ldots$, $\mu_k = \rho^k\mu_0$. Hence, $\mu_k \leq \mu_{k-1} \leq \cdots \leq \mu_0$, i.e., μ_k decreases in every iteration as it nears the optimal point. It behaves like a feedback system. Indeed, as σ_0 measures the distance of the initial point from the solution x_*, $\rho^k\sigma_0$ signifies the distance of the k-th iteration from the solution or the optimal point.*

Algorithm 2 Nesterov accelerated BFGS with adaptive momentum parameter

Require: an initial guess $x_0 \in \mathbb{R}^n$, $L, M, m, \sigma_0, \epsilon, \delta, \rho, W_0 \succ 0$ be the initial inverse Hessian approximation and initial momentum vector $v_0 = 0$, and $\mu_0 \in (0, \frac{\sigma_0(1-\rho)}{2g(1-2\delta)(1+\rho)})$;

1: k=0;
2: Compute $\nabla f(x_k)$;
3: Compute $v_{k+1} = \mu_k v_k - W_k \nabla f(x_k + \mu_k v_k)$;
4: Update $x_{k+1} = x_k + v_{k+1}$;
5: Compute $\nabla f(x_{k+1})$;
6: Update W_{k+1} using Equation-(10);
7: Update $\mu_{k+1} = \rho\mu_k$;
8: k=k+1;

In the next section, we prove the superlinear convergence of Nesterov accelerated BFGS using the above linear rate of convergence of NA-BFGS.

7 Superlinear Convergence of Nesterov Accelerated BFGS

Lemma 5. *Let us assume that $f(x)$ satisfies Assumptions-$\big(12-14\big)$ and $\big(28-32\big)$. Then the following inequalities hold for all $t \geq 0$,*

$$\left\|\hat{J}_t^{-1}(\hat{J}_t - \hat{B}_t)\hat{s}_t\right\| \leq \big(1+\frac{\sigma_t}{2}\big)\big[\rho(1+\|\mu_t\|+\rho)+\|\mu_t\|\big]\left(\frac{\sigma_t}{2}+\left\|\frac{(\hat{B}_t-I)\hat{s}_t}{\hat{s}_t}\right\|\right)\|r_{t-1}\|. \tag{59}$$

$$\left\|\hat{J}_t^{-1}[\nabla^2 f(x_*)]^{-1/2}\nabla^2 f(z_t)v_t\right\| \leq (1+\frac{\sigma_t}{2})\frac{L}{m}(1+\rho)\|r_{t-1}\|. \tag{60}$$

$$\left\|[\nabla^2 f(x_*)]^{1/2}v_t\right\| \leq (1+\rho)\|r_{t-1}\|. \tag{61}$$

Proof.

$$\begin{aligned}
\left\|\hat{J}_t^{-1}(\hat{J}_t-\hat{B}_t)\hat{s}_t\right\| &= \|\hat{J}_t^{-1}[(\hat{J}_t-I)\hat{s}_t-(\hat{B}_t-I)\hat{s}_t]\| \\
&\leq \|\hat{J}_t^{-1}\|\left(\|(\hat{J}_t-I)\hat{s}_t\|+\|(\hat{B}_t-I)\hat{s}_t\|\right) \\
&= \|\hat{J}_t^{-1}\|\left(\|\hat{J}_t-I\|+\left\|\frac{(\hat{B}_t-I)\hat{s}_t}{\hat{s}_t}\right\|\right)\|\hat{s}_t\|.
\end{aligned}$$

Hence,

$$\|\hat{J}_t^{-1}(\hat{J}_t-\hat{B}_t)\hat{s}_t\| \leq \|\hat{J}_t^{-1}\|\left(\|\hat{J}_t-I\|+\left\|\frac{(\hat{B}_t-I)\hat{s}_t}{\hat{s}_t}\right\|\right)\|\hat{s}_t\|. \tag{62}$$

From Lemma-(2), we get $\|\hat{J}_t^{-1}\| \leq 1+\frac{\sigma_t}{2}$ and $\|\hat{J}_t - I\| \leq \frac{\sigma_t}{2}$. From Theorem-(1), we get $\sigma_{t+1} \leq \rho\sigma_t$, where $\rho \in (0,1)$. As $\sigma_t = \frac{M\mu_t}{m^{3/2}}\|r_t\|$, we have $\|r_{t+1}\| \leq \rho\|r_t\|$.

$$\begin{aligned}
\|\hat{s}_t\| &= \left\|[\nabla^2 f(x_*)]^{1/2}(x_{t+1}-x_t-\mu_t v_t)\right\| \\
&= \left\|\left[\nabla^2 f(x_*)\right]^{1/2}(x_{t+1}-x_*+x_*-x_t-\mu_t v_t)\right\| \\
&\leq \left\|[\nabla^2 f(x_*)]^{1/2}(x_{t+1}-x_*)\right\|+\left\|[\nabla^2 f(x_*)]^{1/2}(x_t-x_*)\right\| \\
&+ \|\mu_t\|\left\|[\nabla^2 f(x_*)]^{1/2}v_t\right\| \\
&\leq \left\|r_{t+1}\right\|+\left\|r_t\right\|+\|\mu_t\|\left\|\left[\nabla^2 f(x_*)\right]^{1/2}(x_t-x_{t-1})\right\| \\
&\leq \rho\|r_t\|+\|r_t\|+\|\mu_t\|\big(\|r_t\|+\|r_{t-1}\|\big) \\
&= (1+\|\mu_t\|+\rho)\|r_t\|+\|\mu_t\|\|r_{t-1}\| \\
&\leq \rho(1+\|\mu_t\|+\rho)\|r_{t-1}\|+\|\mu_t\|\|r_{t-1}\| \\
&\leq \big[\rho(1+\|\mu_t\|+\rho)+\|\mu_t\|\big]\|r_{t-1}\|.
\end{aligned}$$

Hence, we have

$$\left\|\hat{s_t}\right\| \leq \left[\rho(1+\|\mu_t\|+\rho)+\|\mu_t\|\right]\|r_{t-1}\|. \tag{63}$$

From Theorem-(1), we have $\sigma_{k+1} \leq \rho\sigma_k$, $\|(\hat{B_k}-I)\|_F \leq 2\delta$, $\|\hat{B_k}\| \leq 1+2\delta$ and $\|\hat{B_k}^{-1}\| \leq \frac{1}{1-2\delta}$. Since, for any $t \geq 0$, we have $\tau_t = \max(\sigma_t + \frac{M\mu_t}{m^{3/2}}\|\hat{v_t}\|, \sigma_{t+1}) = \sigma_t + \frac{M\mu_t}{m^{3/2}}\|\hat{v_t}\|$. Putting $\delta = 2\delta$ in Lemma-(3), we have

$$\left\|(\hat{B}_{t+1}-I)\right\|_F \leq \left\|(\hat{B}_t-I)\right\|_F - \frac{\hat{s}_t^T(\hat{B}_t-I)\hat{B}_t(\hat{B}_t-I)\hat{s}_t}{4\delta\hat{s_t}^T\hat{B}_t\hat{s}_t} + \frac{3+\sigma_t}{1-\sigma_t}\left[\sigma_t + \frac{M\mu_t}{m^{3/2}}\|\hat{v_t}\|\right].$$

Now, taking summation both sides from $t=0$ to $t=k-1$, we get

$$\|(\hat{B_k}-I)\|_F \leq \|(\hat{B_0}-I)\|_F - \sum_{t=0}^{k-1}\frac{\hat{s}_t^T(\hat{B}_t-I)\hat{B}_t(\hat{B}_t-I)\hat{s}_t}{4\delta\hat{s_t}^T\hat{B}_t\hat{s}_t} + \sum_{t=0}^{k-1}\frac{3+\sigma_t}{1-\sigma_t}\left[\sigma_t + \frac{M\mu_t}{m^{3/2}}\|\hat{v_t}\|\right].$$

Then, rearranging the term, we have

$$\begin{aligned}\left[\sum_{t=0}^{k-1}\frac{\hat{s}_t^T(\hat{B}_t-I)\hat{B}_t(\hat{B}_t-I)\hat{s}_t}{4\delta\hat{s}_t^T\hat{B}_t\hat{s}_t}\right] &\leq \|(\hat{B_0}-I)\|_F - \|(\hat{B_k}-I)\|_F \\ &\quad + \sum_{t=0}^{k-1}\frac{3+\sigma_t}{1-\sigma_t}\left[\sigma_t + \frac{M\mu_t}{m^{3/2}}\|\hat{v_t}\|\right] \\ &\leq \|(\hat{B_0}-I)\|_F + \sum_{t=0}^{k-1}\frac{3+\sigma_t}{1-\sigma_t}\left[\sigma_t + \frac{M\mu_t}{m^{3/2}}\|\hat{v_t}\|\right].\end{aligned}$$

We have

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

As $v_0 = 0$, we have

$$\sum_{t=0}^{k-1}\frac{M\mu_t}{m^{3/2}}\|\hat{v_t}\| = \sum_{t=1}^{k-1}\frac{M\mu_t}{m^{3/2}}\|\hat{v_t}\|.$$

From Equation-(51), we have

$$\sum_{t=1}^{k-1} \frac{M\mu_t}{m^{3/2}} \|\hat{v}_t\| \leq \sum_{t=1}^{k-1} \frac{M\mu_t}{m^{3/2}} \big(\|r_t\| + \|r_{t-1}\|\big) = \sum_{t=1}^{k-1} \mu_t \big(\sigma_t + \sigma_{t-1}\big)$$
$$\leq \sum_{t=1}^{k-1} \mu_t \big(\rho^t \sigma_0 + \rho^{t-1}\sigma_0\big) \leq \sum_{t=1}^{k-1} \mu_t \sigma_0 \big(\rho^t + \rho^{t-1}\big).$$

As $\mu_k < 1$ and $\sigma_0 \leq \epsilon$ from our assumption, we have

$$\leq \epsilon \sum_{t=1}^{k-1} \big(\rho^t + \rho^{t-1}\big)$$
$$\leq \epsilon \sum_{t=1}^{k-1} \rho^{t-1}(1+\rho)$$
$$\leq \epsilon \big(\frac{1+\rho}{1-\rho}\big) \leq \frac{2\epsilon}{1-\rho}.$$

Hence, we have

$$\left[\sum_{t=0}^{k-1} \frac{\hat{s}_t^T (\hat{B}_t - I)\hat{B}_t (\hat{B}_t - I)\hat{s}_t}{4\delta \hat{s}_t{}^T \hat{B}_t \hat{s}_t}\right] \leq \|(\hat{B}_0 - I)\|_F + \sum_{t=0}^{k-1} \frac{3+\sigma_t}{1-\sigma_t} \left[\sigma_t + \frac{M\mu_t}{m^{3/2}} \|\hat{v}_t\|\right]$$
$$\leq \delta + \frac{3+\epsilon}{1-\epsilon} \big[\frac{\epsilon}{1-\rho} + \frac{2\epsilon}{1-\rho}\big]$$
$$\leq \delta + \frac{3+\epsilon}{1-\epsilon} \big(\frac{3\epsilon}{1-\rho}\big) \leq \delta + \delta = 2\delta.$$

Therefore,

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

Using the bounds of the Equation-(48), we have

$$\hat{s}_t^T (\hat{B}_t - I)\hat{B}_t (\hat{B}_t - I)\hat{s}_t = x^T \hat{B}_t x \geq \frac{1}{\|\hat{B}_t^{-1}\|} \|(\hat{B}_t - I)\hat{s}_t\|^2$$
$$\geq (1-2\delta)\|(\hat{B}_t - I)\hat{s}_t\|^2,$$

$$\hat{s}_t^T \hat{B}_t \hat{s}_t \leq \left\|\hat{B}_t\right\| \|\hat{s}_t\|^2 \leq (1+2\delta)\|\hat{s}_t\|^2.$$

Hence, we have

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

By combining bounds from Equation-(64) and Equation-(65), we have

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

It implies that

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

By using Cauchy-Schwarz inequality and $q^2 = \frac{1+2\delta}{1-2\delta}$, we have

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

By combining Equation-(66), Equation-(63) and Lemma-(2) and putting the values in Equation-(62), we get

$$\begin{aligned}\left\|\hat{J}_t{}^{-1}(\hat{J}_t - \hat{B}_t)\hat{s}_t\right\| &\leq \|\hat{J}_t{}^{-1}\|\left((\|\hat{J}_t - I)\| + \left\|\frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t}\right\|\right)\|\hat{s}_t\| \\ &\leq (1+\frac{\sigma_t}{2})\big(\rho(1+\|\mu_t\|+\rho) + \|\mu_t\|\big)\left(\frac{\sigma_t}{2} + \left\|\frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t}\right\|\right) \\ &\qquad \|r_{t-1}\|.\end{aligned}$$

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

Using equation-(55) and Lemma-2, we have

$$\begin{aligned}\|\hat{J}_t{}^{-1}[\nabla^2 f(x_*)]^{-1/2}\nabla^2 f(z_t)v_t\| &\leq (1+\frac{\sigma_t}{2})\frac{L}{m}\left(\|r_t\| + \|r_{t-1}\|\right) \\ &\leq (1+\frac{\sigma_t}{2})\frac{L}{m}\left(1+\rho\right)\|r_{t-1}\|.\end{aligned}$$

Using Equation-(51), we get

$$\left\|\hat{v_k}\right\| = \left\|[\nabla^2 f(x_*)]^{1/2}v_t\right\| \leq \left\|r_t\right\| + \left\|r_{t-1}\right\| \leq (1+\rho)\|r_{t-1}\|.$$

Hence, we have

$$\left\|[\nabla^2 f(x_*)]^{1/2} v_t\right\| \leq (1+\rho)\|r_{t-1}\|. \tag{68}$$

□

Theorem 2. *Let us assume that $f(x)$ satisfies Assumptions-$\big(12-14\big)$ and $\big(28-32\big)$. Then x_n generated by Nesterov accelerated BFGS (2) converges to x_* superlinearly with a rate of*

$$\frac{\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_k - x_*)\right\|}{\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 - x_*)\right\|} \leq \left(\frac{M_1 q\sqrt{k} + M_2}{k}\right)^{k/2} \qquad \forall k \geq 1, \tag{69}$$

and

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq (1+\epsilon)^2 \left(\frac{M_1 q\sqrt{k} + M_2}{k}\right)^{k} \qquad \forall k \geq 1, \tag{70}$$

where $M_1 = 4\delta P(1+\frac{\epsilon}{2})$, $M_2 = \big((1+\frac{\epsilon}{2})\frac{L}{m} + 1\big)\frac{\epsilon}{g(1-2\delta)} + P(1+\frac{\epsilon}{2})\frac{\epsilon}{(1-\rho)}$, $P = \big[\rho(1 + \|\mu_0\| + \rho) + \|\mu_0\|\big]$, $g = \frac{\frac{L}{m}+(1+2\delta)}{1-2\delta}$ and $q = \sqrt{\frac{1+2\delta}{1-2\delta}}$.

Proof. As $f(x)$ is twice continuously differentiable function, applying Taylor theorem around x_*, we get

$$\begin{aligned} f(x_t) - f(x_*) &= f(x_*) + \nabla f(x_*)(x_t - x_*) + (1/2)(x_t - x_*)^T H_t (x_t - x_*) - f(x_*) \\ &= (1/2)(x_t - x_*)^T [\nabla^2 f(x_*)]^{1/2} [\nabla^2 f(x_*)]^{-1/2} H_t [\nabla^2 f(x_*)]^{-1/2} \\ &\quad [\nabla^2 f(x_*)]^{1/2}(x_t - x_*) \\ &= \frac{r_t^T \hat{H}_t r_t}{2}, \end{aligned}$$

where $\nabla f(x_*) = 0$, $H_t = \nabla^2 f(x_* + \alpha(x_t - x_*))$, $\alpha \in [0,1]$ and $\hat{H}_t = [\nabla^2 f(x_*)]^{-1/2} H_t [\nabla^2 f(x_*)]^{-1/2}$. Using Lemma-(2) and $\sigma_t \leq \rho\sigma_0 \leq \epsilon$, we get

$$f(x_k) - f(x_*) = \frac{r_k^T \hat{H}_k r_k}{2} \leq \frac{1+\sigma_k}{2}\|r_k\|^2 \leq \frac{1+\epsilon}{2}\|r_k\|^2. \tag{71}$$

Similarly, from Lemma-(2), we have

$$f(x_0) - f(x_*) = \frac{r_0^T \hat{H}_0 r_0}{2} \geq \frac{\|r_0\|^2}{2(1+\sigma_0)} \geq \frac{\|r_0\|^2}{2(1+\epsilon)}. \tag{72}$$

From Equation-(71) and Equation -(72), we have

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq \frac{\frac{1+\epsilon}{2}\|r_k\|^2}{\frac{\|r_0\|^2}{2(1+\epsilon)}} = (1+\epsilon)^2 \frac{\|r_k\|^2}{\|r_0\|^2}. \tag{73}$$

Here, we require to find an upper bound of $\frac{\|r_k\|^2}{\|r_0\|^2}$. We have $J_t(x_t - x_*) = \nabla f(x_t)$ from Equation-(6). Hence, $x_t - x_* = {J_t}^{-1}\nabla f(x_t)$. We get from Equation-(9) that

$$\begin{aligned} s_t = x_{t+1} - (x_t + \mu_t v_t) &= -B_t^{-1}\nabla f(x_t + \mu_t v_t) \\ &= -B_t^{-1}[\nabla f(x_t) + \mu_t \nabla^2 f(z_t) v_t], \end{aligned}$$

where $z_t = \alpha x_t + (1-\alpha)(x_t + \mu_t v_t)$ and $\alpha \in (0,1)$. Hence, $\nabla f(x_t) = -B_t s_t - \mu_t \nabla^2 f(z_t) v_t$. From that above equation, we get

$$\begin{aligned} x_{t+1} - x_* &= x_t - x_* + \mu_t v_t + s_t \\ &= J_t^{-1}\nabla f(x_t) + s_t + \mu_t v_t \\ &= -J_t^{-1} B_t s_t - \mu_t J_t^{-1}\nabla^2 f(z_t) v_t + s_t + \mu_t v_t. \end{aligned}$$

As $\nabla f(x_t) = -B_t s_t - \mu_t \nabla^2 f(z_t) v_t$, multiplying $[\nabla^2 f(x_*)]^{1/2}$ both the sides, we have

$$\begin{aligned} [\nabla^2 f(x_*)]^{1/2}(x_{t+1} - x_*) = -[\nabla^2 f(x_*)]^{1/2} J_t^{-1} [\nabla^2 f(x_*)]^{1/2} [\nabla^2 f(x_*)]^{-1/2} B_t \\ [\nabla^2 f(x_*)]^{-1/2} [\nabla^2 f(x_*)]^{1/2} s_t - \mu_t [\nabla^2 f(x_*)]^{1/2} J_t^{-1} \\ [\nabla^2 f(x_*)]^{1/2} [\nabla^2 f(x_*)]^{-1/2} \nabla^2 f(z_t) v_t + [\nabla^2 f(x_*)]^{1/2} \\ s_t + \mu_t [\nabla^2 f(x_*)]^{1/2} v_t. \end{aligned}$$

Therefore,

$$\begin{aligned} \|r_{t+1}\| &= \| - {\hat{J}_t}^{-1} \hat{B}_t \hat{s}_t - \mu_t {\hat{J}_t}^{-1} [\nabla^2 f(x_*)]^{-1/2} \nabla^2 f(z_t) v_t + \hat{s}_t + \mu_t [\nabla^2 f(x_*)]^{1/2} v_t \| \\ &= \| {\hat{J}_t}^{-1} (\hat{J}_t - \hat{B}_t)\hat{s}_t + \mu_t \left([\nabla^2 f(x_*)]^{1/2} v_t - {\hat{J}_t}^{-1} [\nabla^2 f(x_*)]^{-1/2} \nabla^2 f(z_t) v_t \right) \|. \end{aligned}$$

Hence,

$$\|r_{t+1}\| \leq \| {\hat{J}_t}^{-1} (\hat{J}_t - \hat{B}_t)\hat{s}_t \| + \|\mu_t\| \left(\| [\nabla^2 f(x_*)]^{1/2} v_t \| + \| {\hat{J}_t}^{-1} [\nabla^2 f(x_*)]^{-1/2} \nabla^2 f(z_t) v_t \| \right). \tag{74}$$

Using the bounds of $\| {\hat{J}_t}^{-1} (\hat{J}_t - \hat{B}_t)\hat{s}_t \|$, $\| [\nabla^2 f(x_*)]^{1/2} v_t \|$, $\| {\hat{J}_t}^{-1} [\nabla^2 f(x_*)]^{-1/2} \nabla^2 f(z_t) v_t \|$ from Lemma-(5) and putting the values in Equation-(74), we get

$$\begin{aligned} \|r_{t+1}\| \leq (1 + \frac{\sigma_t}{2}) [\rho(1 + \|\mu_t\| + \rho) + \|\mu_t\|] \left(\frac{\sigma_t}{2} + \left\| \frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t} \right\| \right) \|r_{t-1}\| \\ + \|\mu_t\| \left((1 + \frac{\sigma_t}{2}) \frac{L}{m} (1 + \rho) \|r_{t-1}\| + (1 + \rho) \|r_{t-1}\| \right). \end{aligned}$$

As $\mu_t \leq \rho^t \mu_0$, we have $\|\mu_t\| \leq \|\mu_0\|$. Hence,

$$\begin{aligned}\left\|r_{t+1}\right\| \leq (1+\frac{\sigma_t}{2})\left[\rho(1+\|\mu_0\|+\rho)+\|\mu_0\|\right]\left(\frac{\sigma_t}{2}+\left\|\frac{(\hat{B}_t-I)\hat{s}_t}{\hat{s}_t}\right\|\right)\|r_{t-1}\| \\ +\|\mu_t\|\left((1+\frac{\sigma_t}{2})\frac{L}{m}(1+\rho)\|r_{t-1}\|+(1+\rho)\|r_{t-1}\|\right).\end{aligned}$$

Therefore,

$$\frac{\|r_{t+1}\|}{\|r_{t-1}\|} \leq (1+\frac{\sigma_t}{2})P\left(\frac{\sigma_t}{2}+\left\|\frac{(\hat{B}_t-I)\hat{s}_t}{\hat{s}_t}\right\|\right)+\|\mu_t\|\left((1+\frac{\sigma_t}{2})\frac{L}{m}(1+\rho)+(1+\rho)\right).$$

where $P = \left[\rho(1+\|\mu_0\|+\rho)+\|\mu_0\|\right]$. Taking sum both the sides from $t=0$ to $t=k-1$ and using $\sigma_t \leq \epsilon$, we have

$$\begin{aligned}\sum_{t=0}^{k-1}\frac{\|r_{t+1}\|}{\|r_{t-1}\|} \leq P(1+\frac{\epsilon}{2})\left(\sum_{t=0}^{k-1}\frac{\sigma_t}{2}+\sum_{t=0}^{k-1}\left\|\frac{(\hat{B}_t-I)\hat{s}_t}{\hat{s}_t}\right\|\right)+\sum_{t=0}^{k-1}\|\mu_t\| \\ \left((1+\frac{\epsilon}{2})\frac{L}{m}(1+\rho)+(1+\rho)\right).\end{aligned}$$

Case-1 (Let $k=2n$)
As the arithmetic mean is greater than equal to the geometric mean, we get

$$\begin{aligned}&\frac{\|r_k\|}{\|r_0\|} = \frac{\|r_{2n}\|}{\|r_0\|} = \Pi_{t=0}^{n-1}\frac{\|r_{2t+2}\|}{\|r_{2t}\|} \\ &\leq \left(\frac{\sum_{t=0}^{n-1}\frac{\|r_{2t+2}\|}{\|r_{2t}\|}}{n}\right)^n \\ &\leq \left(\frac{P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)}+2\sqrt{2}\delta q\sqrt{n}\right)+(1+\rho)\left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\sum_{t=0}^{n-1}\|\mu_{2t+1}\|}{n}\right)^n \\ &\leq \left(\frac{P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)}+2\sqrt{2}\delta q\sqrt{n}\right)+(1+\rho)\left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\sum_{t=0}^{n-1}\frac{\rho^{2t+1}\sigma_0(1-\rho)}{2g(1+\rho)(1-2\delta)}}{n}\right)^n \\ &\leq \left(\frac{P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)}+2\sqrt{2}\delta q\sqrt{n}\right)+\left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\frac{\epsilon}{2g(1-2\delta)}}{n}\right)^n \\ &= \left(\frac{P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)}+2\sqrt{2}\delta q\sqrt{\frac{k}{2}}\right)+\left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\frac{\epsilon}{2g(1-2\delta)}}{\frac{k}{2}}\right)^{\frac{k}{2}}\end{aligned}$$

$$= \left(\frac{2P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)} + 2\delta q\sqrt{k}\right) + \left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\frac{\epsilon}{g(1-2\delta)}}{k} \right)^{\frac{k}{2}}.$$

where $P = \left[\rho(1+\|\mu_0\|+\rho) + \|\mu_0\|\right]$.

$$\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)}$$
$$\leq (1+\epsilon)^2 \left(\frac{2P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)} + 2\delta q\sqrt{k}\right) + \left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\frac{\epsilon}{g(1-2\delta)}}{k} \right)^{k}.$$

Case-2 (Let $k = 2n+1$)

$$\frac{\|r_k\|}{\|r_0\|} = \frac{\|r_{2n+1}\|}{\|r_0\|} = \Pi_{t=0}^{n-1} \frac{\|r_{2t+3}\|}{\|r_{2t+1}\|} \frac{\|r_1\|}{\|r_0\|}$$
$$\leq \left(\frac{\sum_{t=0}^{n-1} \frac{\|r_{2t+3}\|}{\|r_{2t+1}\|}}{n} \right)^{n} \frac{\|r_1\|}{\|r_0\|}$$
$$\leq \rho \left(\frac{P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)} + 2\sqrt{2}\delta q\sqrt{n}\right) + (1+\rho)\left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\sum_{t=0}^{n-1}\|\mu_{2t+2}\|}{n} \right)^{n}$$
$$= \rho \left(\frac{P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)} + 2\sqrt{2}\delta q\sqrt{\frac{k-1}{2}}\right) + (1+\rho)\left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\sum_{t=0}^{\frac{k-1}{2}-1}\frac{\rho^{2t+2}\sigma_0(1-\rho)}{2g(1+\rho)(1-2\delta)}}{\frac{k-1}{2}} \right)^{\frac{k-1}{2}}$$
$$= \rho \left(\frac{2P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)} + 2\delta q\sqrt{k-1}\right) + \left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\frac{\epsilon}{g(1-2\delta)}}{k-1} \right)^{\frac{k-1}{2}}.$$

$$\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)}$$
$$\leq (1+\epsilon)^2\rho^2 \left(\frac{2P(1+\frac{\epsilon}{2})\left(\frac{\epsilon}{2(1-\rho)} + 2\delta q\sqrt{k-1}\right) + \left((1+\frac{\epsilon}{2})\frac{L}{m}+1\right)\frac{\epsilon}{g(1-2\delta)}}{k-1} \right)^{k-1}.$$

Hence, the results are established. □

Remark 2. *The superlinear convergence rate of NA-BFGS doesn't depend on the problem dimension d but on the condition number. In our case, the superlinear behaviour*

starts from the first iteration onwards. The superlinear convergence behaviour of NA-BFGS begins when we reach the local neighbourhood of the optimal point, and to reach the local neighbourhood, we can use the Nesterov accelerated gradient descent. The non-asymptotic error bound in NA-BFGS will help predict the maximum number of iterations to achieve the desired accuracy.

Corollary 7.1. *Let us assume that f satisfies Assumptions-(12 − 14). Moreover, suppose the initial point and initial Hessian approximation matrix B_0 satisfy*

$$\frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 - x_*)\right\| \leq \frac{1}{240}. \tag{75}$$

$$\left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\left(B_0 - \nabla^2 f(x_*)\right)[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\|_F \leq \frac{1}{10}. \tag{76}$$

Then x_k generated by Nesterov accelerated BFGS (2) converges to x_* superlinearly with a rate of

$$\frac{\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_k - x_*)\right\|}{\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 - x_*)\right\|} \leq \left(\frac{1}{\sqrt{k}}\right)^{k/2} \qquad \forall k \geq 1, \tag{77}$$

and

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq 1.1\left(\frac{1}{\sqrt{k}}\right)^{k} \qquad \forall k \geq 1. \tag{78}$$

Proof. Comparing Equations (75-76) to Equations (28-29), we get the value of $\epsilon = \frac{1}{240}$, $\delta = \frac{1}{10}$. From Theorem-2, we can take $\rho = \frac{1}{2}$ and $\mu_0 = 0.0004$. Putting all the values in Equations (77-81), we get the desired inequalities. □

Remark 3. *In corollary-7.1, we validate Theorem-2 by taking suitable ϵ, δ, ρ and μ_0. We can also take different values of $\epsilon, \delta, \rho, \mu_0$ to validate our theoretical results.*

To initiate the NA-BFGS, one requires the initial Hessian approximation. In the following Corollary-7.2, we choose the initial Hessian approximation as $\nabla^2 f(x_0)$ and show that the above results hold.

Corollary 7.2. *Let us assume that f satisfies Assumptions-(12 − 14). Moreover, suppose the initial point and initial Hessian approximation matrix B_0 satisfy*

$$\frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 - x_*)\right\| \leq \min\left(\frac{1}{240}, \frac{1}{10\sqrt{d}}\right),\ B_0 = \nabla^2 f(x_0). \tag{79}$$

Then x_k generated by Nesterov accelerated BFGS (2) converges to x_* superlinearly with a rate of

$$\frac{\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_k - x_*)\right\|}{\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 - x_*)\right\|} \leq \left(\frac{1}{\sqrt{k}}\right)^{k/2} \qquad \forall k \geq 1, \tag{80}$$

and

$$\frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)} \leq 1.1\left(\frac{1}{\sqrt{k}}\right)^k \qquad \forall k \geq 1. \tag{81}$$

Proof. From assumption, we have $\frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0-x_*)\right\| \leq \frac{1}{240}$. We know that $\|A\|_F \leq \sqrt{d}\|A\|$ for any matrix of order $\mathbb{R}^{d\times d}$. Hence, we have

$$\begin{aligned}
&\left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\left(\nabla^2 f(x_0)-\nabla^2 f(x_*)\right)[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\|_F \\
&\leq \sqrt{d}\left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\left(\nabla^2 f(x_0)-\nabla^2 f(x_*)\right)[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\| \\
&\leq \sqrt{d}\left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\|^2\left\|\nabla^2 f(x_0)-\nabla^2 f(x_*)\right\| \\
&\leq \sqrt{d}\frac{M}{m}\|x_0-x_*\| \\
&= \sqrt{d}\frac{M}{m}\left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0-x_*)\right\| \\
&\leq \sqrt{d}\frac{M}{m}\left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\|\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0-x_*)\right\| \\
&\leq \sqrt{d}\left(\frac{M}{m^{3/2}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0-x_*)\right\|\right).
\end{aligned}$$

Using Equation-(79), we have

$$\begin{aligned}
&\left\|[\nabla^2 f(x_*)]^{-\frac{1}{2}}\left(\nabla^2 f(x_0)-\nabla^2 f(x_*)\right)[\nabla^2 f(x_*)]^{-\frac{1}{2}}\right\|_F \\
&\leq \sqrt{d}\left(\frac{1}{10\sqrt{d}}\right) = \frac{1}{10}.
\end{aligned}$$

Hence, using Corollary-7.1, we get the desired result. □

Corollary 7.3 (Estimating the norm of the gradient)**.** *Let us assume that $f(x)$ satisfy Assumption-$\big(12-14\big)$, $\big(28-32\big)$. Then, the number of iterations of the NA-BFGS to reach the small norm of the gradient $\|\nabla f(x_k)\| \leq \alpha$ satisfies the following bound*

$$k \geq \frac{M_2 V^2}{\alpha^2}, \tag{82}$$

where $V = \frac{L\epsilon m}{M}$ and $q = \sqrt{\frac{1+2\delta}{1-2\delta}}$.

Proof. From Assumption-(13), putting $x = x_k$ and $y = x_*$, we get

$$\big\|\nabla f(x_k)-\nabla f(x_*)\big\| \leq L\big\|x_k-x_*\big\|. \tag{83}$$

From Equation-(77), we have

$$\|r_k\| \leq \left(\frac{M_1 q\sqrt{k} + M_2}{k}\right)^{k/2} \|r_0\|.$$

we know that $m^{1/2}\|x_k - x_*\| \leq \|r_k\|$ and $\|r_0\| \leq \frac{\epsilon m^{3/2}}{M}$. Therefore,

$$\begin{aligned}
\|x_k - x_*\| &\leq \left(\frac{M_1 q\sqrt{k} + M_2}{k}\right)^{k/2} \frac{\|r_0\|}{m^{1/2}} \\
&\leq \left(\frac{M_1 q\sqrt{k} + M_2}{k}\right)^{k/2} \frac{\epsilon m^{3/2}}{M m^{1/2}} \\
&\leq \left(\frac{M_1 q\sqrt{k} + M_2}{k}\right)^{k/2} \frac{\epsilon m}{M}.
\end{aligned}$$

As $\nabla f(x_*) = 0$, and using Equation-(83), we get

$$\|\nabla f(x_k)\| \leq L\|x_k - x_*\| \leq \left(\frac{M_1 q\sqrt{k} + M_2}{k}\right)^{k/2} \frac{L\epsilon m}{M}.$$

Define $V := \frac{L\epsilon m}{M}$. Hence, $\|\nabla f(x_k)\| \leq \left(\frac{M_1 q\sqrt{k}+M_2}{k}\right)^{k/2} V$. We require to prove the following inequality

$$\begin{aligned}
&\left(\frac{M_1 q\sqrt{k} + M_2}{k}\right)^{k/2} V \leq \alpha \\
&\Rightarrow (M_1 q\sqrt{k} + M_2)^{k/2} \leq \frac{\alpha}{V} (k)^{k/2} \\
&\Rightarrow M_1 q\sqrt{k} + M_2 \leq \left(\frac{\alpha}{V}\right)^{2/k} k \\
&\Rightarrow M_2 \leq \left(\frac{\alpha}{V}\right)^{2/k} k - M_1 q\sqrt{k} \leq k\left(\frac{\alpha}{V}\right)^{2/k} \\
&\Rightarrow lnM_2 \leq lnk + \frac{2}{k} ln\left(\frac{\alpha}{V}\right) \leq lnk + 2ln\left(\frac{\alpha}{V}\right) \\
&\Rightarrow ln\frac{M_2 V^2}{\alpha^2} \leq lnk \\
&\Rightarrow k \geq \frac{M_2 V^2}{\alpha^2}.
\end{aligned}$$

□

Remark 4. *From the above results, we can predict the least iteration number k for which $\nabla f(x_k) \leq \alpha$ is satisfied. Here, iteration number k depends on V, i.e., the initial distance from the optimal point, and M_2. M_2 depends on the condition number and inverse of the $1-\rho$. It shows that if ρ is chosen near to 1, the iteration number k will be higher.*

8 Conclusion

Here, from Equation-(77), we get the rate of convergence of Nesterov accelerated BFGS is $(\frac{1}{k})^{\frac{k}{4}}$, and Nesterov accelerated BFGS works well in the local neighbourhood of the optimal point. We know that the rate of convergence of Nesterov accelerated gradient descent is optimal among all the first-order methods for higher dimensional problems [22]. Therefore, we suggest that one use the Nesterov accelerated gradient descent till one reaches the local neighbourhood of the optimal point and then use Nesterov accelerated BFGS for getting the superliner rate of convergence.

Here, we suggest some further development of the above results. While finding the rate of convergence of NA-BFGS, we assume that the gradient of f is Lipschitz continuous, and L is known to us. One could develop an adapting algorithm that starts from any initial guess L_0 and adjusts inverse Hessian approximation each iteration so that the original estimate remains the same. Also, one can study its rate of convergence in the above analysis. One could extend the analysis to the global convergence.

References

[1] Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)

[2] Nemirovsky, A., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. SIAM, New Delhi (1983)

[3] Nesterov, Y.: A method for solving the convex programming problem with convergence rate o($1/k^2$) Dokl. Akad. Nauk SSSR. **269**(3), 543-547(1983)

[4] Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton method and its global performance. Math. Program. **108**(1), 177-205 (2006)

[5] Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables, vol.30. SIAM, New Delhi (1970)

[6] Bennett, A.A.: Newton's method in general analysis. Proc. Natl. Acad. Sci. U. S. A. **2**(10), 592 (1916)

[7] Conn, A.R., Gould, N.I., Toint, P.L.: Trust Region Methods. SIAM, New Delhi (2000)

[8] Broyden, C.G.: The convergence of single-rank quasi-Newton methods.Math. Comput. **24**(110), 365-382 (1970)

[9] Goldfarb, D.: A family of variable-metric methods derived by variational means. Math. Comput. **24**(109), 23-26 (1970)

[10] 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)

[11] Davidon, W.: Variable metric method for minimization. SIAM J. Optim. (1991)

[12] Fletcher, R., Powell, M.J.: A rapidly convergent descent method for minimization. Comput. J. **6**(2),163-168 (1963)

[13] Fletcher, R.: A new approach to variable metric algorithms. Comput. J. **13**(3), 317-322 (1970)

[14] Broyden, C.G.: A class of methods for solving nonlinear simultaneous equations. Math. Comput. **19**(92), 577-593 (1965)

[15] 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**(3), 223-245 (1973)

[16] Gao, W., Goldfarb, D.: Quasi-newton methods: superlinear convergence without line searches for self-concordant functions. Opt. Methods Softw. **34**(1), 194-217 (2019)

[17] 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)

[18] Rodomanov, A., Nesterov, Y.: Rates of superlinear convergence for classical quasi-Newton methods. Math. Program. 1-32(2021)

[19] Jin, Q., Mokhtari, A.: Non-asymptotic superlinear convergence of standard quasi-Newton methods. Math. Program. **200**(1), 1-49(2022)

[20] Mahboubi, S., Ninomiya, H., Asai, H.: Momentum acceleration of quasi-Newton based optimization technique for neural network training. Nonlinear Theory and Its Applications, IEICE. **12**(3), 554-574 (2021)

[21] Ninomiya, H.: Neural network training based on quasi-Newton method using Nesterov's accelerated gradient. 2016 IEEE Region 10 Conference (TENCON), 51-54 (2016)

[22] Nesterov, Y.: Lectures on convex optimization. Springer, **200**,(2018)