Non-asymptotic superlinear convergence of Nesterov accelerated BFGS

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

[1]Mathematics Department, NIT Rourkela, Sector-2, Rourkela, 769008, Odisha, India.

[2*]Mathematics Department, NIT Rourkela, Sector-2, Rourkela, 769008, Odisha, India.

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

Abstract

In this paper, we derive the explicit finite time local convergence of Nesterov accelerated the Broyden-Fletcher-Goldfarb-Shanno (NA-BFGS) under the assumption that the objective function is strongly convex, its gradient is Lipschitz continuous, and its Hessian is Lipschitz continuous at the optimal point. We have shown that the rate of convergence of the NA-BFGS method is $(\frac{1}{k})^{\frac{k}{2}}$. Further, we show that Nesterov accelerated BFGS gives a faster convergence rate than the classical Broyden-Fletcher-Goldfarb-Shanno (BFGS). This is the first work that theoretically guarantees the superlinear convergence of NA-BFGS non-asymptotically.

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

1 Introduction

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

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

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

With solving convex optimization problems, first-order methods such as gradient descent, ADAM, Stochastic gradient descent, and Nesterov accelerated Gradient descent are frequently used [2]. 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 $\tau \in (0, 1)$. The 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 [[16], [20]] where m is the 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 [[1], [21], [7], [19]] such as Newton methods, etc. Newton's method has a quadratic convergence rate under certain assumptions [2] but has some drawbacks. The computational cost in computing the inverse Hessian is high, and it is quite difficult to calculate the exact inverse Hessian matrix when the objective function involves many variables.

Therefore, to solve these problems, the Quasi-Newton method was proposed [[4], [12]], where the inverse Hessian matrix is approximated. The computational cost per iteration for the Newton method was $O(d^3)$, but for Quasi-newton methods, the cost reduces to $O(d^2)$ [20]. There are different variants in Quasi newton methods such as SR1 [6], DFP [[8], [10]], BFGS [[3], [11]], but BFGS optimize faster among them. BFGS method has the super-linear rate of convergence [[5], [9], [13]]. The sequence x_k converge to the optimal point x_* super linearly if the ratio between the error at $k+1$ time and k time tend to zero as k approaches to infinity, i.e.,

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

There is enough literature on the asymptotic analysis of classical BFGS [[5], [9]]. Still, recently, the non-asymptotic convergence analysis of the classical Quasi-Newton method has been carried out [[14], [18]]. Non-asymptotic convergence analysis gives more information about the complexity of the algorithm. In neural network problems, Nesterov accelerated BFGS works better than BFGS [[15], [17]]. But the non-asymptotic analysis of Nesterov accelerated BFGS is not yet studied properly. Hence, in this paper, we find a non-asymptotic convergence analysis of Nesterov accelerated BFGS under certain assumptions. This analysis gives more idea about the complexity of Nesterov accelerated BFGS. Here, we prove that the non-asymptotic rate of convergence of Nestrov accelerated BFGS is better than the classical BFGS. Also, through an example, we show that the Nestrove accelarated BFGS converges better. Also, in our proof, we have assumed weaker Lipscitz conditions on the gradient, and these assumptions are more justified as Nesterov accelerated BFGS shows super-linearity convergence only locally around the optimal point.

In Sec 2, we describe the newly proposed Nesterov accelerated BFGS (NA-BFGS). Sec 3 discusses some notations required for proving our results. In Sec 4, we are discussing our assumptions, which is helpful to prove our desired result, i.e., superlinear

rate of convergence. We prove some important lemmas in Sec 5. In Sec 6, we prove the linear rate of convergence of Nesterov accelerated BFGS. In Sec 7, we prove our main theoretical result, i.e., superlinear convergence rate for a certain class of functions. In Sec 8, we demonstrate the advantage of Nesterov accelerated BFGS over classical BFGS for certain classes of functions. We validate our theoretical results in Sec 9. We conclude our results in Sec 10.

2 Nesterov accelerated BFGS

Nesterov accelerated BFGS is the accelerated version of BFGS, which works better for a certain function class. But the main drawback is the computation of normal gradient $\nabla f(x_k)$ and Nesterov accelerated gradient $\nabla f(x_k + \mu v_k)$. We must compute two gradient operations that increase our computational cost in each iteration. Hence, to avoid this drawback, S. Mahboubi et al. [15] proposed a modified Nesterov accelerated BFGS, which significantly reduces the computational cost of the optimization algorithm and performs better in Neural networks and other problems.

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

$$\hat{f}(x+p) = f(x) + \nabla f(x)^T p + \frac{p^T H p}{2}. \tag{3}$$

where H is the Hessian approximation of $f(x)$. The quadratic model is very accurate in approximating $f(x)$ when x is near to optimal point x_* [2]. Hence, we must take the initial guess x_0 sufficiently close to x_*. In NA-BFGS, $v_0 = 0$ and momentum coefficient $\mu \in (0, 1)$. As f is approximated quadratically in the neighborhood of $x_k + \mu v_k$, so it implies that $\nabla f(x_k + \mu v_k)$ can be approximated linearly in the neighborhood of $x_k + \mu v_k$, i.e.,

$$\nabla f(x_k + \mu v_k) \simeq \nabla f(x_k) + \mu \nabla f(v_k). \tag{4}$$

The newly proposed Nesterov accelerated BFGS is defined as

$$x_{k+1} = x_k + \mu v_k - W_k \nabla f(x_k + \mu v_k) = x_k + \mu v_k - W_k[\nabla f(x_k) + \mu \nabla f(v_k)], \tag{5}$$

where

$$W_{k+1} = \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right)^T W_k \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k}, \tag{6}$$

$s_k = x_{k+1} - (x_k + \mu v_k)$, $y_k = \nabla f(x_{k+1}) - \nabla f(x_k + \mu v_k) = \nabla f(x_{k+1}) - [\nabla f(x_k) + \mu \nabla f(v_k)]$, $\mu \in (0, 1)$. We analyze the above version (5) of Nesterov accelerated BFGS.

3 Notation

This section briefly discusses some notation and definitions used in theorems and their proof. Let us define $[\nabla^2 f(x_*)]^{\frac{1}{2}}$ and $[\nabla^2 f(x_*)]^{-\frac{1}{2}}$ are the real symmetric positive

Algorithm 1 Nesterov accelerated BFGS [15]

1. Take an initial guess $x_0 \in \mathbb{R}^d$, $W_0 \succ 0$ be the initial inverse Hessian approximation and initial momemtum vector $v_0 = 0$ and constant $\mu \in (0, 1)$. Let $k = 0$.
2. In order to get v_{k+1}, we have to solve the equation $v_{k+1} = \mu v_k - W_k \nabla f(x_k + \mu v_k) = \mu v_k - W_k[\nabla f(x_k) + \mu \nabla f(v_k)]$ [15].
3. Then the next iterate $x_{k+1} = x_k + v_{k+1}$.
4. Update W_k using NA-BFGS inverse Hessian formula $W_{k+1} = \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right)^T W_k \left(I - \frac{y_k s_k^T}{s_k^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k}$ where $s_k = x_{k+1} - (x_k + \mu v_k), y_k = \nabla f(x_{k+1}) - \nabla f(x_k + \mu v_k) = \nabla f(x_{k+1}) - [\nabla f(x_k) + \mu \nabla f(v_k)]$.
5. Then $k = k + 1$ and go to step-1.

definite matrix and the square root of $[\nabla^2 f(x_*)]$ and $[\nabla^2 f(x_*)]^{-1}$, respectively. Here, we used a weighted version of the Hessian approximation, i.e.,

$$\hat{B_k} = [\nabla^2 f(x_*)]^{-\frac{1}{2}} B_k [\nabla^2 f(x_*)]^{-\frac{1}{2}}. \tag{7}$$

Hence $\hat{B_k}$ is a real symmetric positive definite matrix as $[\nabla^2 f(x_*)]^{-\frac{1}{2}}$ and B_k are real symmetric positive definite matrix. Here, we used 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

$$\hat{y}_k = [\nabla^2 f(x_*)]^{-\frac{1}{2}} y_k, \quad \hat{s}_k = [\nabla^2 f(x_*)]^{\frac{1}{2}} s_k, \quad \hat{\nabla} f(x_k) = [\nabla^2 f(x_*)]^{-\frac{1}{2}} \nabla f(x_k). \tag{8}$$

$$v_k = x_k - x_{k-1}, \ \hat{v}_k = [\nabla^2 f(x_*)]^{\frac{1}{2}} (x_k - x_{k-1}), \ \hat{\nabla} f(v_k) = [\nabla^2 f(x_*)]^{-\frac{1}{2}} f(v_k). \tag{9}$$

$$J_k = \int_0^1 \left[\nabla^2 f(x_* + \alpha(x_k - x_*))\right] d\alpha, \quad \hat{J}_k = [\nabla^2 f(x_*)]^{-\frac{1}{2}} J_k [\nabla^2 f(x_*)]^{-\frac{1}{2}}. \tag{10}$$

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

$$r_k = [\nabla^2 f(x_*)]^{\frac{1}{2}} (x_k - x_*), \quad \sigma_k = \frac{M}{m^{\frac{3}{2}}} \|r_k\|, \quad \tau_k = max\left(\sigma_k + \frac{M\mu}{m^{\frac{3}{2}}} \|\hat{v}_k\|, \sigma_{k+1}\right). \tag{11}$$

Here, $\|.\|$ and $\|.\|_F$ are denoted as Euclidean and Frobenius norm, respectively.

4 Assumption

We take the following assumptions to prove the superlinear convergence of Nesterov accelerated BFGS.

1. $f(x)$ is C^2 function, i.e., twice continuously differentiable function and $f(x)$ is strongly convex with parameter m , i.e.,

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

2. Gradient of $f(x)$ is Lipschitz continuous with parameter L at the optimal x_* , i.e.,

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

3. Hessian of $f(x)$ is Lipschitz continuous with parameter M at the optimal x_*, i.e.,

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

Remark 1. *In the Nesterov accelerated BFGS, we assume that the gradient and Hessian are Lipschitz continuous at the optimal point only and prove the superlinear convergence of Nesterov accelerated BFGS.*

5 Important Lemmas

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

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

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

Proof. One can refer to [14]. □

Lemma 2. *From the notation section, we define σ_k from (11) and $\hat{J}_k$ from (10). Let us define matrix $H_k = \nabla^2 f(x_* + \alpha_k(x_k - x_*))$, $\hat{H}_k = [\nabla^2 f(x_*)]^{-\frac{1}{2}} H_k \big[\nabla^2 f(x_*)\big]^{-\frac{1}{2}}$ and $\alpha_k \in [0, 1]$. Also, assume that $f(x)$ satisfies Assumption 12 and 14. Then the following inequalities hold for all $k \geq 0$.*

$$\frac{1}{(1 + \frac{\sigma_k}{2})} I \preceq \hat{J}_k \preceq \big(1 + \frac{\sigma_k}{2}\big) I, \tag{16}$$

and

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

Proof. The proof is similar to [14]. □

Lemma 3. *From the notation section, recall the definition of τ_k from (11) 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 $\tau_k \leq 1$ and $\|\hat{B}_k - I\| \leq \delta$. Then B_{k+1} satisfies the following inequalities*

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

we also have

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

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 [14]. □

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

$$\|\hat{y}_t - \hat{s}_t\| \leq \tau_t\|\hat{s}_t\|, \tag{20}$$

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

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

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

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

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

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. By using Cauchy Schwarz inequality, we have

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

We have

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

Therefore, the second claim is proved. Then using reverse triangle inequality and (24), we have

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

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.

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

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

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

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 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

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

3. Let us choose $\epsilon, \delta \in (0, \frac{1}{2})$, and $\rho, \mu \in (0, 1)$ such that they satisfy

$$\frac{3+\epsilon}{1-\epsilon}\left(\frac{\epsilon}{1-\rho}\left(1+\mu+\frac{\mu}{\rho}\right)\right) \leq \delta \quad \text{and} \quad \rho \geq \frac{1}{1-2\delta}\left(\frac{\epsilon}{2}+2\delta+\frac{L\mu N}{m}+\mu(1+2\delta)N\right) \tag{30}$$

where $N \geq \left(1+\frac{1}{\rho}\right)$.

Proposition 1. *Let us assume that $f(x)$ satisfies Assumptions 12-14 and 28-29. Let us choose $\epsilon, \delta \in (0, \frac{1}{2})$ and $\rho, \mu \in (0, 1)$ such that*

$$\frac{3+\epsilon}{1-\epsilon}\left(\frac{\epsilon}{1-\rho}\left(1+\mu+\frac{\mu}{\rho}\right)\right) \leq \delta, \tag{31}$$

$$\rho \geq \frac{1}{1-2\delta}\left(\frac{\epsilon}{2}+2\delta\right). \tag{32}$$

Then

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

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

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 (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 $\hat{B}_0$, respectively. 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

$$\begin{aligned}
\sigma_1 &= \frac{M}{m^{\frac{3}{2}}}\|r_1\| = \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_1 - x_*)\right\| \\
&= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 + \mu v_0 - B_0^{-1}(\nabla f(x_0 + \mu v_0)) - x_*)\right\| \\
&= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 - B_0^{-1}\nabla f(x_0) - x_*)\right\| \quad (\text{As } v_0 = 0) \\
&= \frac{M}{m^{\frac{3}{2}}}\Big\|[\nabla^2 f(x_*)]^{\frac{1}{2}} B_0^{-1}\big[\nabla f(x_0) - \nabla^2 f(x_*)(x_0 - x_*) \\
&\qquad\qquad - (B_0 - \nabla^2 f(x_*))(x_0 - x_*)\big]\Big\| \\
&= \frac{M}{m^{\frac{3}{2}}}\Big\|[\nabla^2 f(x_*)]^{\frac{1}{2}} B_0^{-1}[\nabla^2 f(x_*)]^{\frac{1}{2}}\big([\nabla^2 f(x_*)]^{\frac{-1}{2}}\nabla f(x_0) \\
&- [\nabla^2 f(x_*)]^{\frac{-1}{2}}\nabla^2 f(x_*)(x_0 - x_*) - [\nabla^2 f(x_*)]^{\frac{-1}{2}}(B_0 - \nabla^2 f(x_*)) \\
&\qquad\qquad [\nabla^2 f(x_*)]^{\frac{-1}{2}}[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_0 - x_*)\big)\Big\| \\
&\leq \frac{M}{m^{\frac{3}{2}}}\left\|\hat{B_0}^{-1}\right\|\left\|\left[\hat{\nabla} f(x_0) - r_0 - \left(\hat{B}_0 - I\right) r_0\right]\right\| \\
&\leq \frac{M}{m^{\frac{3}{2}}}\left\|\hat{B_0}^{-1}\right\|\left[\|\hat{\nabla} f(x_0) - r_0\| + \|\hat{B}_0 - I\|\|r_0\|\right] \\
&\leq \frac{M}{m^{\frac{3}{2}}}\frac{1}{1-2\delta}\left[\frac{\sigma_0}{2}\|r_0\| + 2\delta\|r_0\|\right].
\end{aligned}$$

Using (38), (35) and Lemma 4, we get

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

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 28-30. Then*

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

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

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

$$\begin{aligned} Z_0\tau_0 &= \frac{3+\sigma_0}{1-\sigma_0}\max\left(\sigma_0 + \frac{M\mu}{m^{\frac{3}{2}}}\|\hat{v}_0\|, \sigma_1\right) \\ &= \frac{3+\sigma_0}{1-\sigma_0}\sigma_0 \leq \frac{3+\epsilon}{1-\epsilon}\epsilon \leq \delta. \end{aligned}$$

Hence,

$$\begin{aligned} \|\hat{B}_1 - I\|_F &\leq \|(\hat{B}_0 - I)\|_F + Z_0\tau_0 \\ &\leq \delta + \delta = 2\delta. \end{aligned}$$

Similarly, all the eigenvalues of $\hat{B}_1$ are in $[1-2\delta, 1+2\delta]$. Let $\lambda_{\max}(\hat{B}_1)$ and $\lambda_{\min}(\hat{B}_1)$ are the largest eigenvalue and the smallest eigenvalue of $\hat{B}_1$, respectively. Then we have

$$\left\|\hat{B}_1\right\| = \lambda_{\max}(\hat{B}_1) \leq 1 + 2\delta, \tag{41}$$

and

$$\left\|\hat{B}_1^{-1}\right\| = \frac{1}{\lambda_{min}(\hat{B}_1)} \leq \frac{1}{1-2\delta}. \tag{42}$$

$$\begin{aligned} \sigma_2 &= \frac{M}{m^{\frac{3}{2}}}\|r_2\| = \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_2 - x_*)\right\| \\ &= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_1 + \mu v_1 - B_1^{-1}(\nabla f(x_1 + \mu v_1)) - x_*)\right\| \\ &= \frac{M}{m^{\frac{3}{2}}}\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_1 + \mu v_1 - B_1^{-1}(\nabla f(x_1) + \mu\nabla f(v_1)) - x_*)\right\| \\ &= \frac{M}{m^{\frac{3}{2}}}\Big\|[\nabla^2 f(x_*)]^{\frac{1}{2}}B_1^{-1}[\nabla^2 f(x_*)]^{\frac{1}{2}}[\nabla^2 f(x_*)]^{\frac{-1}{2}}\Big[\nabla f(x_1) + \mu\nabla f(v_1) \\ &\qquad - \nabla^2 f(x_*)(x_1 - x_*) - (B_1 - \nabla^2 f(x_*))(x_1 - x_*) - \mu B_1 v_1\Big]\Big\| \\ &\leq \frac{M}{m^{\frac{3}{2}}}\|\hat{B_1}^{-1}\|\Big\|[\nabla f(x_*)]^{\frac{-1}{2}}\Big[\nabla f(x_1) + \mu\nabla f(v_1) - \nabla^2 f(x_*)(x_1 - x_*) \\ &\qquad\qquad - (B_1 - \nabla^2 f(x_*))(x_1 - x^*) - \mu B_1 v_1\Big]\| \\ &= \frac{M}{m^{\frac{3}{2}}}\|\hat{B_1}^{-1}\|\Big\|\Big[\hat{\nabla} f(x_1) + \mu\hat{\nabla} f(v_1) - r_1 - [\nabla f(x_*)]^{\frac{-1}{2}}\big(B_1 - I\big)[\nabla f(x_*)]^{\frac{-1}{2}} \\ &\qquad [\nabla f(x_*)]^{\frac{1}{2}}(x_1 - x_*) - \mu\nabla f(x_*)]^{\frac{-1}{2}}B_1\nabla f(x_*)]^{\frac{-1}{2}}[\nabla^2 f(x_*)]^{\frac{1}{2}}v_1\Big]\Big\| \\ &= \frac{M}{m^{\frac{3}{2}}}\left\|\hat{B_1}^{-1}\right\|\Big\|\Big[\hat{\nabla} f(x_1) + \mu\hat{\nabla} f(v_1) - r_1 - \Big(\hat{B}_1 - I\Big)r_1 - \mu\hat{B}_1[\nabla^2 f(x_*)]^{\frac{1}{2}}v_1\Big]\Big\| \\ &\leq \frac{M}{m^{\frac{3}{2}}}\left\|\hat{B_1}^{-1}\right\|\Big[\Big(\|\hat{\nabla} f(x_1) - r_1\| + \|\hat{B}_1 - I\|\|r_1\|\Big) + \mu\big(\|\hat{\nabla} f(v_1)\| \end{aligned}$$

$$+ \|\hat{B}_1[\nabla^2 f(x_*)]^{\frac{1}{2}} v_1\|\Big)\Big].$$

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

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

As $\sigma_1 \leq \rho\sigma_0$, it implies $\|r_1\| \leq \rho\|r_0\|$. Also,

$$\begin{aligned}
\|\hat{\nabla} f(v_1)\| &\leq \frac{L}{m}\Big(\rho\|r_0\| + \|r_o\|\Big) \\
&= \frac{L}{m}\rho\|r_0\|\Big(1 + \frac{1}{\rho}\Big).
\end{aligned}$$

Hence, their exist $N \in \mathbb{R}$ such that $\rho\|r_0\|\big(1 + \frac{1}{\rho}\big) \leq N\|r_1\|$, i.e., $\rho\|r_0\|\Big(1 + \frac{1}{\rho}\Big) \leq N\|r_1\| \leq N\rho\|r_0\|$.
Hence,

$$N \geq \Big(1 + \frac{1}{\rho}\Big).$$

Therefore,

$$\|\hat{\nabla} f(v_1)\| \leq \frac{L}{m}N\|r_1\|, \tag{44}$$

where $N \geq \Big(1 + \frac{1}{\rho}\Big)$.

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

As $\sigma_1 \leq \rho\sigma_0$, then $\|r_1\| \leq \rho\|r_0\|$.

$$\left\|[\nabla^2 f(x_*)]^{\frac{1}{2}} v_1\right\| \leq \rho\|r_0\|\Big(1+\frac{1}{\rho}\Big).$$

Hence, as above

$$\|[\nabla^2 f(x_*)]^{\frac{1}{2}} v_1\| \leq N\|r_1\|, \tag{45}$$

where $N \geq (1+\frac{1}{\rho})$. Now, using Lemma 4, (45), (44), (42), (41) and put the bounds in (43), we have

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

As $\sigma_1 \leq \rho\sigma_0 \leq \epsilon$, then

$$\sigma_2 \leq \rho\sigma_1,$$

where $\rho \geq \frac{1}{1-2\delta}\left[\frac{\epsilon}{2} + 2\delta + \frac{L\mu N}{m} + \mu(1+2\delta)N\right]$. □

Remark 2. *Proposition 1 and Preposition 2 are the first two cases of induction method used in proving the linear convergence of Nesterov accelerated BFGS.*

Theorem 1. *Let us assume that $f(x)$ satisfies Assumptions 12-14 and 28-30. Then the sequence of iterate x_k generated by the Nesterov accelerated BFGS algorithm (1) converges to an optimal solution x_* with*

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

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 inequalities and the 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 $\sigma_2 \leq \rho\sigma_1$, $\left\|\hat{B}_1 - I\right\|_F \leq 2\delta$, $\left\|\hat{B}_1\right\| \leq 1 + 2\delta$, $\left\|\hat{B}_1^{-1}\right\| \leq \frac{1}{1-2\delta}$, where $\rho \geq \frac{1}{1-2\delta}\left(\frac{\epsilon}{2} + 2\delta + \frac{L\mu N}{m} + \mu(1+2\delta)N\right)$ from Proposition 2. This show that all the condition 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 $\rho \geq \frac{1}{1-2\delta}\left[\frac{\epsilon}{2} + 2\delta + \frac{L\mu N}{m} + \mu(1+2\delta)N\right]$. Now we have to prove for $k = t+1$. Since the condition from (47) is satisfied for $0 \leq k \leq t$, i.e., $\|\hat{B}_k - I\|_F \leq 2\delta$ for $0 \leq k \leq t$, 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 $Z_k = \frac{3+\sigma_k}{1-\sigma_k}$, $\tau_k = \max\left(\sigma_k + \frac{M\mu}{m^{\frac{3}{2}}}\left\|\hat{v_k}\right\|, \sigma_{k+1}\right)$.

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

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

As $\sigma_k \leq \rho\sigma_{k-1} \leq \cdots \leq \rho^k\sigma_0$, then $\|r_k\| \leq \rho^k\|r_0\|$ and $\|r_{k-1}\| \leq \rho^{k-1}\|r_0\|$.

$$\|\hat{v}_k\| \leq \rho^k(1 + \frac{1}{\rho})\|r_0\|. \tag{51}$$

Hence,

$$\begin{aligned}
\tau_k &\leq \epsilon + \frac{M\mu}{m^{3/2}}\|\hat{v_k}\| \\
&\leq \epsilon + \mu\rho^k(1 + \frac{1}{\rho})\sigma_0
\end{aligned}$$

$$\leq \epsilon + \mu\rho^k(1+\frac{1}{\rho})\epsilon = \epsilon\Big(1+\mu\rho^k+\frac{\mu\rho^k}{\rho}\Big)$$
$$\leq \epsilon\Big(1+\mu+\frac{\mu}{\rho}\Big)$$

From (30), we get

$$\tau_k \leq \epsilon\big(\frac{\delta(1-\rho)(1-\epsilon)}{\epsilon(3+\epsilon)}\big) \leq 1,$$

for $0 \leq k \leq t$. As the condition from (47) is satisfied for $0 \leq k \leq t$, i.e., $\|\hat{B}_k - I\|_F \leq 2\delta$ for $0 \leq k \leq t$, we have to show for $k = t+1$. Here, $Z_k = \frac{3+\sigma_k}{1-\sigma_k}$. Also, for $0 \leq k \leq t$, $\sigma_k \leq \epsilon$. Hence,

$$\begin{aligned}\sum_{k=0}^{k=t}\tau_k &\leq \sum_{k=0}^{k=t}\left(\sigma_k + \mu\rho^k(1+\frac{1}{\rho})\sigma_0\right)\\ &\leq \sum_{k=0}^{k=t}\rho^k\sigma_0 + \frac{\mu}{1-\rho}(1+\frac{1}{\rho})\epsilon\\ &\leq \frac{\epsilon}{1-\rho} + \frac{\mu}{1-\rho}(1+\frac{1}{\rho})\epsilon\\ &= \frac{\epsilon}{1-\rho}\left(1+\mu+\frac{\mu}{\rho}\right).\end{aligned}$$

Taking sum from $k=0$ to $k=t$ on both sides from (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

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

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

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

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 $\hat{B}_{t+1}$, respectively. Then we have

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

and

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

Hence,

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

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

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

As $\sigma_t \leq \rho\sigma_{t-1}$, it implies $\|r_t\| \leq \rho\|r_{t-1}\|$

$$\begin{aligned}\big\|\hat{\nabla} f(v_t)\big\| &\leq \frac{L}{m}\left(\rho^t\|r_0\| + \rho^{t-1}\|r_0\|\right)\\ &= \frac{L}{m}\rho^t\|r_0\|\big(1+\frac{1}{\rho}\big).\end{aligned}$$

Hence, there exists $N \in \mathbb{R}$ such that $\rho^t\|r_0\|\big(1+\frac{1}{\rho}\big) \leq N\|r_t\|$, i.e., $\rho^t\|r_0\|\big(1+\frac{1}{\rho}\big) \leq N\|r_t\| \leq N\rho^t\|r_0\|$. Hence,

$$\big\|\hat{\nabla} f(v_t)\big\| \leq \frac{L}{m}N\|r_t\|. \tag{55}$$

where $N \geq (1+\frac{1}{\rho})$.

$$\begin{aligned}\big\|\hat{B}_t[\nabla^2 f(x_*)]^{\frac{1}{2}}v_t\big\| &= \big\|\hat{B}_t[\nabla^2 f(x_*)]^{\frac{1}{2}}(x_t - x_* + x_* - x_{t-1})\big\|\\ &\leq (1+2\delta)\big(\|r_t\| + \|r_{t-1}\|\big)\\ &\leq (1+2\delta)\big(\rho^t\|r_0\| + \rho^{t-1}\|r_0\|\big) \leq \rho^t\|r_0\|(1+2\delta)\big(1+\frac{1}{\rho}\big).\end{aligned}$$

Therefore, as above

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

Similarly, using the same argument from (54-56) and Lemma 4, we have $\|\hat{B}_t - I\|_F \leq 2\delta$, $\left\|\hat{B}_t^{-1}\right\| \leq \frac{1}{1-2\delta}$, $\|\hat{\nabla} f(x_t) - r_t\| \leq \frac{\sigma_t}{2}$, $\big\|\hat{\nabla} f(v_t)\big\| \leq \kappa N\big\|r_t\big\|$, $\big\|\hat{B}_t[\nabla^2 f(x_*)]^{\frac{1}{2}}v_t\big\| \leq N\big(1+2\delta\big)\big\|r_t\big\|$.

$$\begin{aligned}\sigma_{t+1} &\leq \frac{M}{m^{\frac{3}{2}}}\frac{1}{1-2\delta}\left[\frac{\sigma_t}{2}\|r_t\| + 2\delta\|r_t\| + \frac{L\mu N}{m}\|r_t\| + \mu(1+2\delta)N\|r_t\|\right]\\ &= \frac{M\|r_t\|}{m^{\frac{3}{2}}}\frac{1}{1-2\delta}\left[\frac{\sigma_t}{2} + 2\delta + \frac{L\mu N}{m} + \mu(1+2\delta)N\right]\\ &\leq \frac{\sigma_t}{1-2\delta}\left[\frac{\epsilon}{2} + 2\delta + \frac{L\mu N}{m} + \mu(1+2\delta)N\right] \leq \rho\sigma_t.\end{aligned}$$

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

$$\sigma_{k+1} \leq \rho\sigma_k, \text{ where } \rho \geq \frac{1}{1-2\delta}\left[\frac{\epsilon}{2} + 2\delta + \frac{L\mu N}{m} + \mu(1+2\delta)N\right]. \tag{57}$$

Therefore, all the above inequality is true for $k = t+1$. □

Remark 3. *We use the linear convergence results of Nesterov accelerated BFGS in proving the superlinear convergence of Nesterov accelerated BFGS.*

7 Superlinear Convergence of Nesterov Accelerated BFGS

Lemma 5. *Let us assume that $f(x)$ satisfy Assumption 12-14, and 28-30. Then the following inequality holds for all $t \geq 0$,*

$$\left\|\hat{J}_t^{-1}(\hat{J}_t - \hat{B}_t)\hat{s}_t\right\| \leq (1 + \frac{\sigma_t}{2})P\left(\frac{\sigma_t}{2} + \left\|\frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t}\right\|\right)\|r_t\|. \tag{58}$$

$$\left\|\hat{J}_t^{-1}\hat{\nabla}f(v_t)\right\| \leq \frac{L}{m}N\left\|\hat{J}_t^{-1}\right\|\|r_t\|. \tag{59}$$

$$\left\|[\nabla^2 f(x_*)]^{1/2}v_t\right\| \leq N\|r_t\|, \tag{60}$$

where $N \geq (1 + 1/\rho)$, $P \geq \left(\frac{\delta}{M} + \rho\right)$ and $M = \frac{3+\epsilon}{1-\epsilon}\left(\frac{\epsilon}{1-\rho}\right)$.

Proof.

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

$$\|\hat{J}_t^{-1}(\hat{J}_t - \hat{B}_t)\hat{s}_t\| \leq \|\hat{J}_t^{-1}\|\left((\|\hat{J}_t - I)\| + \left\|\frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t}\right\|\right)\|\hat{s}_t\|. \tag{61}$$

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

$$\begin{aligned}
\|\hat{s}_t\| &= \left\|[\nabla^2 f(x_*)]^{1/2}\left(x_{t+1} - x_t - \mu v_t\right)\right\| \\
&= \left\|\left[\nabla^2 f(x_*)\right]^{1/2}\left(x_{t+1} - x_* + x_* - x_t - \mu v_t\right)\right\| \\
&\leq \left\|[\nabla^2 f(x_*)]^{1/2}(x_{t+1} - x_*)\right\| + \left\|[\nabla^2 f(x_*)]^{1/2}(x_t - x_*)\right\| + \mu\left\|[\nabla^2 f(x_*)]^{1/2}v_t\right\| \\
&\leq \left\|r_{t+1}\right\| + \left\|r_t\right\| + \mu\left\|\left[\nabla^2 f(x_*)\right]^{1/2}(x_t - x_{t-1})\right\| \\
&\leq \rho\|r_t\| + \|r_t\| + \mu\big(\|r_t\| + \|r_{t-1}\|\big) = (1 + \mu + \rho)\|r_t\| + \mu\|r_{t-1}\| \\
&\leq (1 + \mu + \rho)\rho^t\|r_0\| + \mu\rho^{t-1}\|r_0\| \\
&= \rho^t(1 + \mu + \rho + \frac{\mu}{\rho})\|r_0\|.
\end{aligned}$$

From (30), we get

$$\|\hat{s}_t\| \leq \rho^t(\frac{\delta}{M} + \rho)\|r_0\|,$$

where $M = \frac{3+\epsilon}{1-\epsilon}\left(\frac{\epsilon}{1-\rho}\right)$ as $\frac{3+\epsilon}{1-\epsilon}\left(\frac{\epsilon}{1-\rho}\left(1+\mu+\frac{\mu}{\rho}\right)\right) \leq \delta$. Hence, their exist $P \in \mathbb{R}$ such that $\rho^t\|r_0\|\left(\frac{\delta}{M}+\rho\right) \leq P\|r_t\|$, i.e., $\rho^t\|r_0\|\left(\frac{\delta}{M}+\rho\right) \leq P\|r_t\| \leq P\rho^t\|r_0\|$.
Hence,

$$P \geq (\frac{\delta}{M} + \rho).$$

Hence, we have

$$\left\|\hat{s}_t\right\| \leq P\left\|r_t\right\|, \tag{62}$$

where $P \geq \frac{\delta}{M} + \rho$. 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 $\|\hat{B}_k{}^{-1}\| \leq \frac{1}{1-2\delta}$. Hence, for any $t \geq 0$, we have $\tau_t = \max(\sigma_t + \frac{M\mu}{m^{3/2}}\|\hat{v_k}\|, \sigma_{t+1}) = \sigma_t + \frac{M\mu}{m^{3/2}}\|\hat{v_k}\|$. Using Lemma 3, we have

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

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

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

Then rearranging the term, we have

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

We have

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

From (51), we have

$$\sum_{t=0}^{k-1} \frac{M\mu}{m^{3/2}} \|\hat{v}_t\| \leq \sum_{t=0}^{k-1} \frac{M\mu}{m^{3/2}} \rho^t (1 + \frac{1}{\rho}) \|r_0\| \leq \mu\sigma_0 \frac{(1 + \frac{1}{\rho})}{1 - \rho} \leq \mu\epsilon \frac{(1 + \frac{1}{\rho})}{1 - \rho}.$$

Hence, we have

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

Therefore,

$$\left[\sum_{t=0}^{k-1} \frac{\hat{s}_t^T (\hat{B}_t - I)\hat{B}_t(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t^T \hat{B}_t \hat{s}_t}\right] \leq 8\delta^2. \tag{63}$$

Using the bounds from (48), we have

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

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

Hence, we have

$$\frac{\hat{s}_t^T (\hat{B}_t - I)\hat{B}_t(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t{}^T \hat{B}_t \hat{s}_t} \geq \frac{1 - 2\delta}{1 + 2\delta} \left\|\frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t}\right\|^2. \tag{64}$$

By combining bounds from (63) and (64), we have

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

then

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

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

$$\sum_{t=0}^{k-1} \left\| \frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t} \right\| \leq 2\sqrt{2}\delta q\sqrt{k}. \tag{65}$$

By combining (65), (62) and Lemma 2 and putting the values in (61), we get

$$\begin{aligned} \left\|\hat{J}_t{}^{-1}(\hat{J}_t - \hat{B}_t)\hat{s}_t\right\| &\leq \|\hat{J}_t{}^{-1}\| \left((\|\hat{J}_t - I)\| + \left\| \frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t} \right\| \right) \|\hat{s}_t\| \\ &\leq (1 + \frac{\sigma_t}{2})P \left(\frac{\sigma_t}{2} + \left\| \frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t} \right\| \right) \|r_t\| \end{aligned}$$

$$\left\|\hat{J}_t{}^{-1}(\hat{J}_t - \hat{B}_t)\hat{s}_t\right\| \leq (1 + \frac{\sigma_t}{2})P \left(\frac{\sigma_t}{2} + \left\| \frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t} \right\| \right) \|r_t\|. \tag{66}$$

$$\begin{aligned} \left\|\hat{J}_t{}^{-1}\hat{\nabla} f(v_t)\right\| \leq \left\|\hat{J}_t{}^{-1}\right\| \left\|\hat{\nabla} f(v_t)\right\| &= \left\|\hat{J}_t{}^{-1}\right\| \left\| [\nabla^2 f(x_*)]^{-1/2} (\nabla f(x_t - x_{t-1})) \right\| \\ &= \left\|\hat{J}_t{}^{-1}\right\| \left\| [\nabla^2 f(x_*)]^{-1/2} (\nabla f(x_t) - \nabla f(x_{t-1})) \right\| \\ &= \left\|\hat{J}_t{}^{-1}\right\| \left\| [\nabla^2 f(x_*)]^{-1/2} \big(\nabla f(x_t) - \nabla f(x_*) \right. \\ &\quad \left. + \nabla f(x_*) - \nabla f(x_{t-1})\big) \right\| \\ &= \frac{L}{m} \|\hat{J}_t{}^{-1}\| \Big(\|r_t\| + \|r_{t-1}\| \Big) \\ &\leq \frac{L}{m} \rho^t \|\hat{J}_t{}^{-1}\| \Big\| r_0 \Big\| (1 + \frac{1}{\rho}). \end{aligned}$$

Hence, their exist $N \in \mathbb{R}$ such that $\rho^t \|r_0\| \big(1 + \frac{1}{\rho}\big) \leq N\|r_t\|$, i.e., $\rho^t \|r_0\| \big(1 + \frac{1}{\rho}\big) \leq N\|r_t\| \leq N\rho^t \|r_0\|$.
Hence,

$$N \geq (1 + \frac{1}{\rho}).$$

$$\left\|\hat{J}_t^{-1}\hat{\nabla} f(v_t)\right\| \leq \frac{L}{m} N \left\|\hat{J}_t^{-1}\right\| \|r_t\|. \tag{67}$$

where $N \geq (1 + \frac{1}{\rho})$.

$$\begin{aligned} \left\| [\nabla^2 f(x_*)]^{1/2} v_t \right\| &= \left\| [\nabla^2 f(x_*)]^{1/2} (x_t - x_{t-1}) \right\| \\ &= \left\| [\nabla^2 f(x_*)]^{1/2} \right\| \left\| (x_t - x_* + x_* - x_{t-1}) \right\| \end{aligned}$$

$$\leq \Big\|r_t\Big\| + \Big\|r_{t-1}\Big\| \leq \rho^t \Big\|r_0\Big\| + \rho^{t-1}\Big\|r_0\Big\|$$
$$\leq \rho^t(1+\frac{1}{\rho})\Big\|r_0\Big\|.$$

Hence, we have

$$\Big\|[\nabla^2 f(x_*)]^{1/2} v_t\Big\| \leq N\Big\|r_t\Big\|, \tag{68}$$

where $N \geq (1+1/\rho)$. □

Theorem 2. *Let us assume that $f(x)$ satisfies Assumptions 12-14 and 28-30. Then x_n generated by Nesterov accelerated BFGS (1) converges to $x*$ superlinearly with a rate of*

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

and

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

where $M_1 = 2\sqrt{2}\delta P(1+\frac{\epsilon}{2})$, $M_2 = \mu N + \frac{L\mu N}{m}(1+\frac{\epsilon}{2}) + P(1+\frac{\epsilon}{2})\frac{\epsilon}{2(1-\rho)}$, $N \geq (1+1/\rho)$, $P \geq \left(\frac{\delta}{M} + \rho\right)$ and $M = \frac{3+\epsilon}{1-\epsilon}\left(\frac{\epsilon}{1-\rho}\right)$.

Proof. As $f(x)$ is twice continuously differentiable function. Therefore we use Taylor theorem around x_*, i.e.,

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

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 (71) and (72), we have

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

Since $\nabla f(x_*) = 0$, we have $J_t(x_t - x_*) = \nabla f(x_t)$ from (10). Hence, $x_t - x_* = {J_t}^{-1}\nabla f(x_t)$. We get from (5) that

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

Then $\nabla f(x_t) = -B_t s_t - \mu \nabla f(v_t)$.

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

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

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

Therefore,

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

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

Now we have to use the bounds of $\| {\hat{J}_t}^{-1}(\hat{J}_t - \hat{B}_t)\hat{s}_t \|$, $\| [\nabla^2 f(x_*)]^{1/2} v_t \|$, and $\| {\hat{J}_t}^{-1} [\nabla^2 f(x_*)]^{-1/2} f(v_t) \|$ from Lemma 5 and put the values in (74). Therefore,

$$\begin{aligned} \frac{\|r_{t+1}\|}{\|r_t\|} &\leq (1 + \frac{\sigma_t}{2}) P \left(\frac{\sigma_t}{2} + \left\| \frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t} \right\| \right) + \mu N + \frac{L\mu N}{m} \| {\hat{J}_t}^{-1} \| \\ &\leq (1 + \frac{\sigma_t}{2}) P \left(\frac{\sigma_t}{2} + \left\| \frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t} \right\| \right) + \mu N + \frac{L\mu N}{m} (1 + \frac{\sigma_t}{2}). \end{aligned}$$

Taking sum both sides from $t = 0$ to $t = k-1$ and use $\sigma_t < \epsilon$, we get

$$\sum_{t=0}^{k-1} \frac{\|r_{t+1}\|}{\|r_t\|} \leq P(1+\frac{\epsilon}{2}) \left(\sum_{t=0}^{k-1} \frac{\sigma_t}{2} + \sum_{t=0}^{k-1} \left\| \frac{(\hat{B}_t - I)\hat{s}_t}{\hat{s}_t} \right\| \right) + \mu N + \frac{L\mu N}{m}(1+\frac{\epsilon}{2})$$
$$\leq P(1+\frac{\epsilon}{2}) \left(\frac{\epsilon}{2(1-\rho)} + 2\sqrt{2}\delta q\sqrt{k} \right) + \mu N + \frac{L\mu N}{m}(1+\frac{\epsilon}{2}).$$

As the Arithmetic mean is greater than equal to the geometric mean, then

$$\frac{\|r_k\|}{\|r_0\|} = \Pi_{t=0}^{k-1} \frac{\|r_{t+1}\|}{\|r_t\|} \leq \left(\frac{\sum_{t=0}^{k-1} \frac{\|r_{t+1}\|}{\|r_t\|}}{k} \right)^k$$
$$\leq \left(\frac{P(1+\frac{\epsilon}{2}) \left(\frac{\epsilon}{2(1-\rho)} + 2\sqrt{2}\delta q\sqrt{k} \right) + \mu N + \frac{L\mu N}{m}(1+\frac{\epsilon}{2})}{k} \right)^k.$$

From (73),

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

where $P \geq \left(\frac{\delta}{M} + \rho\right)$ and $M = \frac{3+\epsilon}{1-\epsilon}\left(\frac{\epsilon}{1-\rho}\right)$.

□

Corollary 1. *Let us assume that $f(x)$ satisfies Assumptions 12-14 and 28. Let us consider the initial Hessian approximation B_0 satisfy*

$$\left\| [\nabla^2 f(x_*)]^{-1/2} (B_0 - \nabla^2 f(x_*)) \nabla^2 f(x_*)]^{-1/2} \right\| = \|\hat{B}_0 - I\| \leq \delta, \tag{75}$$

where $\epsilon, \delta \in (0, 1/2)$ such that for $\rho \in (0,1)$ and $v = 0$ in Theorem 2, they satisfy $\frac{3+\epsilon}{1-\epsilon}\left(\frac{\epsilon}{1-\rho}\right) \leq \delta$,

$$\left[\frac{\epsilon}{2} + 2\delta\right] \leq \rho(1-2\delta), \tag{76}$$

and Nesterov Accelerated BFGS perform like a classical BFGS, and the μ term vanishes as $v = 0$. Then x_n generated by classical BFGS converge to x_* superlinearly with a rate of

$$\frac{\left\| [\nabla^2 f(x_*)]^{\frac{1}{2}} (x_k - x_*) \right\|}{\left\| [\nabla^2 f(x_*)]^{\frac{1}{2}} (x_0 - x_*) \right\|} \leq \left(\frac{M_1 q\sqrt{k} + M_2}{k} \right)^k, \tag{77}$$

and

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq (1+\epsilon)^2 \left(\frac{M_1 q\sqrt{k} + M_2}{k} \right)^{2k}, \tag{78}$$

where $M_1 = 2\sqrt{2}\delta P(1+\frac{\epsilon}{2})$, $M_2 = P(1+\frac{\epsilon}{2})\frac{\epsilon}{2(1-\rho)}$ *and* $P = 1+\rho$.

Proof. The proof of this Theorem is similar to Theorem 2. □

The above result from Corollary 1 can also be found from [14].

8 Comparison of Bounds between BFGS and Nesterov Accelerated BFGS

We take the following assumption to compare with the theoretical bounds of classical BFGS.

1. Gradient of $f(x)$ is Lipschitz continuous everywhere with parameter L, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \qquad \forall x, y \in R^d. \tag{79}$$

This Assumption 12 and (79) assures us that the eigenvalue of the Hessian matrix lies between m and L i.e. $\sqrt{m} \leq [\nabla^2 f(x_*)]^{\frac{1}{2}} \leq \sqrt{L}$. Then the bounds in classical BFGS from (77) and [14] are defined as

$$\frac{\|x_k - x_*\|}{\|x_0 - x_*\|} \leq \sqrt{\frac{L}{m}} \left(\frac{N_1 q\sqrt{k} + N_2}{k} \right)^{k}, \tag{80}$$

where $N_1 = 2\sqrt{2}\delta(1+\rho)(1+\frac{\epsilon}{2})$, $\rho \in (0,1)$, $N_2 = \frac{(1+\frac{\epsilon}{2})(1+\rho)\epsilon}{2(1-\rho)}$, $q = \sqrt{\frac{1+2\delta}{1-2\delta}}$, L and m are Lipschitz constant and strong convexity constant, respectively. However, the bounds in Nesterov accelerated BFGS from (69) can be derived as

$$\frac{\|x_k - x_*\|}{\|x_0 - x_*\|} \leq \sqrt{\frac{L}{m}} \left(\frac{M_1 q\sqrt{k} + M_2}{k} \right)^{k}, \tag{81}$$

where $M_1 = 2\sqrt{2}\delta P(1+\frac{\epsilon}{2})$, $M_2 = \mu N + \frac{L\mu N}{m}(1+\frac{\epsilon}{2}) + P(1+\frac{\epsilon}{2})\frac{\epsilon}{2(1-\rho)}$ and $P \geq \left(\frac{\delta}{M} + \rho\right)$. Let the super-linear convergence in standard BFGS start after k_1 iteration and let the super-linear convergence in Nesterov accelerated BFGS start after k_2 iteration. Assume that $k = \max(k_1, k_2)$. Let us define $C_k = \sqrt{\frac{L}{m}} \left(\frac{M_1 q\sqrt{k} + M_2}{k} \right)^{k}$ and $D_k =$

$\sqrt{\frac{L}{m}}\left(\frac{N_1 q\sqrt{k}+N_2}{k}\right)^k$. Hence,

$$\frac{C_k}{D_k} = \frac{\sqrt{\frac{L}{m}}\left(\frac{M_1 q\sqrt{k}+M_2}{k}\right)^k}{\sqrt{\frac{L}{m}}\left(\frac{N_1 q\sqrt{k}+N_2}{k}\right)^k} = \frac{\left(\frac{M_1 q\sqrt{k}+M_2}{k}\right)^k}{\left(\frac{N_1 q\sqrt{k}+N_2}{k}\right)^k} = \frac{\left(M_1 q\sqrt{k}+M_2\right)^k}{\left(N_1 q\sqrt{k}+N_2\right)^k}.$$

Now, $M_1 q\sqrt{k}+M_2 = M_2\left[1+\frac{M_1 q\sqrt{k}}{M_2}\right] \leq M_2 e^{\frac{M_1 q\sqrt{k}}{M_2}}$ and
$N_1 q\sqrt{k}+N_2 = N_2\left[1+\frac{N_1 q\sqrt{k}}{N_2}\right] \geq N_2 e^{-\frac{N_1 q\sqrt{k}}{N_2}}$. Hence,

$$\left(\frac{C_k}{D_k}\right)^{\frac{1}{k}} \leq \frac{M_2 e^{\frac{M_1 q\sqrt{k}}{M_2}}}{\frac{1}{N_2} e^{\frac{N_1 q\sqrt{k}}{N_2}}} = M_2 N_2 e^{q\sqrt{k}\left(\frac{M_1}{M_2}-\frac{N_1}{N_2}\right)}. \tag{82}$$

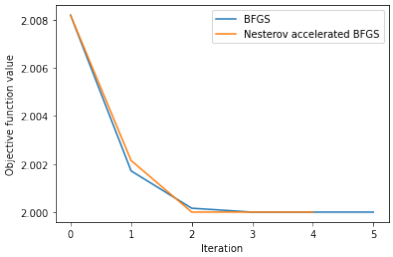
Also, we get

$$\frac{N_1}{N_2} = \frac{2\sqrt{2}\delta(1+\rho)(1+\frac{\epsilon}{2})}{(1+\frac{\epsilon}{2})(1+\rho)\frac{\epsilon}{2(1-\rho)}} = 4\sqrt{2}(1-\rho)\frac{\delta}{\epsilon}$$

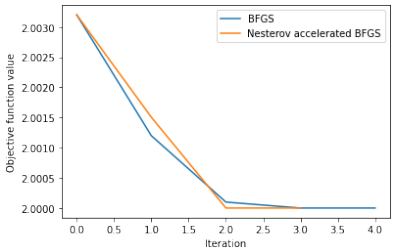
and

$$\begin{aligned}\frac{M_1}{M_2} &= \frac{2\sqrt{2}\delta P(1+\frac{\epsilon}{2})}{(1+\frac{\epsilon}{2})P\frac{\epsilon}{2(1-\rho)}+\mu N+\frac{L\mu N}{m}(1+\frac{\epsilon}{2})} \\ &= \frac{1}{\frac{\epsilon}{4\sqrt{2}(1-\rho)\delta}+\frac{\mu N}{2\sqrt{2}\delta P(1+\frac{\epsilon}{2})}+\frac{L\mu N(1+\frac{\epsilon}{2})}{2\sqrt{2}\delta P(1+\frac{\epsilon}{2})m}} \\ &\leq \frac{N_1}{N_2}.\end{aligned}$$

For large k, we have $\frac{C_k}{D_k} < 1$ from (82). Also, choosing small ϵ and μ, we get $M_2 N_2 < 1$. Hence from (82), we have $\frac{C_k}{D_k} < 1$, always. Hence after K iteration and careful selection of μ and ϵ, Nesterov accelerated BFGS converge to optimal point faster than standard BFGS. Therefore, the convergence of Nesterov accelerated BFGS is very sensitive to the momentum parameter.

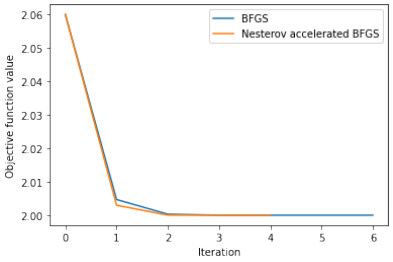


(a) Initial guess (-0.9,-1.3)

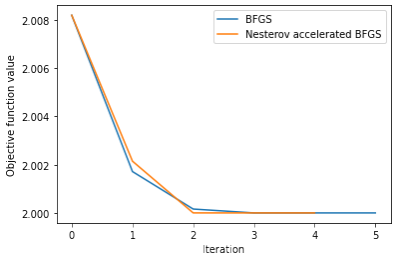


(b) Initial guess (1.1,1.3)

Fig. 1: BFGS VS Nesterov accelerated BFGS



(a) Initial guess (-0.6,-1.1).



(b) Initial guess (0.9,1.3).

Fig. 2: BFGS VS Nesterov accelerated BFGS

9 Numerical Experiment

9.1 Test Function

$f(x)$ is defined as

$$f(x) = f(x, y) = x^4 + y^4 - 2x^2 - 2y^2 + 4. \tag{83}$$

The following function satisfy the Assumption 12,13, and 14. The minimum value is 2 and the minimum value is achieved at $(1, 1)$, $(-1, -1)$. we compare two popular optimization algorithms, BFGS and Nesterov accelerated BFGS, while taking initial guesses very close to optimal. We are interested in the local convergence of Nesterov accelerated BFGS. To verify our theoretical result, we do this numerical experiment. From Figure 1a, 1b, 2a, and 2b, we observe that Nesterov accelerated BFGS outperforms classical BFGS.

10 Discussion

Hence, the Nesterov accelerated BFGS works well in the local neighbourhood of optimal solution. So it is better to use Nesterov accelerated BFGS rather than standard

BFGS when we already reach the local neighbourhood of optimal solution. We know that the rate of convergence of Nesterov accelerated gradient descent is optimal among all first-order methods for higher dimensional problems. Therefore, it is better to use the first Nesterov accelerated gradient descent till we reach the local neighbourhood of optimal, and further, we use Nesterov accelerated BFGS.

11 Conclusion

Here, we show the non-asymptotic superliner convergence of Nestrov accelerated BFGS and prove that its convergence rate is better than the classical BFGS. One may extend further in the following directions. While finding the rate of convergence of NA-BFGS, we assume that the gradient of $f(x)$ is Lipschitz continuous at the optimal point, and L is known to us. One could develop an adapting algorithm that starts from any initial guess L_0 and adjust inverse Hessian approximation so that the original estimate remains the same. All the convergence analysis is valid when we choose an initial guess sufficiently close to the minimizer. We could extend the analysis for global convergence under certain assumptions with an inexact line search.

References

[1] Bennett, A. A.: Newton's method in general analysis. Proc. Natl. Acad. Sci. U. S. A. 2(10), 592-598 (1916)

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

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

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

[5] Broyden, C.G., Broyden, J.E.D., Jr., More, J.J.: On the local and superlinear convergence of quasi-Newton methods. IMA J. Appl. Math. 12, 223-245 (1973)

[6] 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)

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

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

[9] 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)

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

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

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

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

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

[15] 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)

[16] Nesterov, Y.: A method for solving the convex programming problem with convergence rate o(1/k2) Dokl. Akad. Nauk SSSR. 269, 543-547 (1983)

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

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

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

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

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