

Appendix

A Notations

Given weight matrix $P \in \mathbb{S}_{++}^d$, recall the weighted vectors defined in (9). We define the quantity $\hat{\theta}_t$,

$$\cos(\hat{\theta}_t) = \frac{-\hat{g}_t^\top \hat{s}_t}{\|\hat{g}_t\| \|\hat{s}_t\|}. \quad (26)$$

We also define the following terms which play important roles in all the convergence analysis,

$$\hat{p}_t := \frac{f(x_t) - f(x_{t+1})}{-\hat{g}_t^\top \hat{s}_t}, \quad \hat{q}_t := \frac{\|\hat{g}_t\|^2}{f(x_t) - f(x_*)}, \quad \hat{m}_t := \frac{\hat{y}_t^\top \hat{s}_t}{\|\hat{s}_t\|^2}, \quad \hat{n}_t = \frac{\hat{y}_t^\top \hat{s}_t}{-\hat{g}_t^\top \hat{s}_t}. \quad (27)$$

Moreover, we define that

$$\tilde{g}_t := \nabla^2 f(x_*)^{-\frac{1}{2}} g_t. \quad (28)$$

For $\tau_1, \tau_2 \in [0, 1]$, we define the Hessian matrices J_t and G_t as

$$J_t := \nabla^2 f(x_t + \tau_1(x_{t+1} - x_t)), \quad (29)$$

$$G_t := \nabla^2 f(x_t + \tau_2(x_* - x_t)). \quad (30)$$

We have that $y_t = \nabla f(x_{t+1}) - \nabla f(x_t) = J_t(x_{t+1} - x_t) = J_t s_t$ and $\nabla f(x_t) = \nabla f(x_t) - \nabla f(x_*) = G_t(x_t - x_*)$ for some $\tau_1, \tau_2 \in [0, 1]$ by mean value theorem.

B Lemmas and Propositions

Lemma B.1. *Consider the BFGS method with Armijo-Wolfe inexact line search, where the step size satisfies the conditions in (5) and (6). If B_t is symmetric positive definite, we have $f(x_{t+1}) \leq f(x_t)$ and the following results hold:*

$$\hat{p}_t = \frac{f(x_t) - f(x_{t+1})}{-g_t^\top s_t} \geq \alpha, \quad \hat{n}_t = \frac{y_t^\top s_t}{-g_t^\top s_t} \geq 1 - \beta. \quad (31)$$

Proof. Please check Lemma 2.1 in [42]. \square

Proposition B.2. *Let $\{B_t\}_{t \geq 0}$ be the Hessian approximation matrices generated by the BFGS update in (4). For a given weight matrix $P \in \mathbb{S}_{++}^d$, recall the weighted vectors and the weighted matrix in (9). Then, we have that for any $t \geq 0$,*

$$\Psi(\hat{B}_{t+1}) \leq \Psi(\hat{B}_t) + \frac{\|\hat{y}_t\|^2}{\hat{y}_t^\top \hat{s}_t} - 1 + \log \frac{\cos^2 \hat{\theta}_t}{\hat{m}_t}, \quad (32)$$

where \hat{m}_t is defined in (27) and $\cos(\hat{\theta}_t)$ is defined in (26). As a corollary, we have that for any $t \geq 1$,

$$\sum_{i=0}^{t-1} \log \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq -\Psi(\hat{B}_0) + \sum_{i=0}^{t-1} \left(1 - \frac{\|\hat{y}_i\|^2}{\hat{y}_i^\top \hat{s}_i}\right). \quad (33)$$

Proof. Please check Proposition 2 in [41]. \square

Lemma B.3. *Recall the definition of function $\omega(x)$ in (10) and define the function $\omega_*(x)$ as*

$$\omega_*(x) := -x - \log(1 - x), \quad \forall x < 1. \quad (34)$$

We have that

(a) $\omega(x)$ is increasing function for $x > 0$ and decreasing function for $-1 < x < 0$. Moreover, $\omega(x) \geq 0$ for all $x > -1$.

(b) When $x \geq 0$, we have that $\omega(x) \geq \frac{x^2}{2(1+x)}$.

(c) When $-1 < x \leq 0$, we have that $\omega(x) \geq \frac{x^2}{2+x}$.

(d) When $0 < x < 1$, we have that $\omega_*(x) \leq \frac{x^2}{2(1-x)}$.

(e) We have $\sqrt{2x} + \frac{2x}{3} \leq \omega^{-1}(x) \leq \sqrt{2x} + x$, where ω^{-1} is the inverse function of $\omega(x)$ when $x > 0$.

Proof. Please check Lemma G.1 in [42] for the proof of (a), (b), and (c). Please check Lemma 5.1.5 from [46] for the proof of (d) and Lemma A.1 from [40] for the proof of (e). \square

Lemma B.4. Recall the definition of function $\omega(x)$ and $\omega_*(x)$ in (10) and (34). If Assumption 2.2 holds, we have that

$$f(y) \geq f(x) + g(x)^\top (y - x) + \frac{4}{M^2} \omega\left(\frac{M}{2} \|y - x\|_x\right). \quad (35)$$

Moreover, when $\|y - x\|_x < \frac{2}{M}$, we have that

$$f(y) \leq f(x) + g(x)^\top (y - x) + \frac{4}{M^2} \omega_*\left(\frac{M}{2} \|y - x\|_x\right). \quad (36)$$

Proof. Check Theorem 5.1.8 and Theorem 5.1.9 of [46]. \square

Lemma B.5. Suppose Assumptions 2.2 holds, and recall the definitions of the matrices J_t and G_t in (29) and (30), and the quantity D_t in (11). The following statements hold:

(a) For any $t \geq 0$, we have that

$$\frac{1}{1 + D_t} \nabla^2 f(x_*) \preceq \nabla^2 f(x_t) \preceq (1 + D_t) \nabla^2 f(x_*). \quad (37)$$

(b) For any $t \geq 0$, we have that

$$\frac{1}{1 + D_t} \nabla^2 f(x_*) \preceq G_t \preceq (1 + D_t) \nabla^2 f(x_*). \quad (38)$$

(c) For any $t \geq 0$ and $\tau \in [0, 1]$, we have that

$$\frac{1}{1 + D_t} G_t \preceq \nabla^2 f(x_t + \tau(x_* - x_t)) \preceq (1 + D_t) G_t. \quad (39)$$

(d) Suppose that $f(x_{t+1}) \leq f(x_t)$ for $t \geq 0$, we have that

$$\frac{1}{1 + D_t} \nabla^2 f(x_*) \preceq J_t \preceq (1 + D_t) \nabla^2 f(x_*). \quad (40)$$

Proof. Proof for (a): In Lemma B.4, take $y = x_t$ and $x = x_*$, we have that $\omega(\frac{M}{2} \|x_t - x_*\|_{x_*}) \leq \frac{M^2}{4} (f(x_t) - f(x_*))$. Hence, we have that

$$\|x_t - x_*\|_{x_*} \leq \frac{2}{M} \omega^{-1}\left(\frac{M^2}{4} C_t\right).$$

In Assumptions 2.2, take $x = x_t$, $y = x_*$ and $z = x_*$, we prove that

$$\nabla^2 f(x_t) \preceq (1 + M \|x_t - x_*\|_{x_*}) \nabla^2 f(x_*) \preceq (1 + 2\omega^{-1}(\frac{M^2}{4} C_t)) \nabla^2 f(x_*) = (1 + D_t) \nabla^2 f(x_*). \quad (41)$$

Similarly, take $x = x_*$, $y = x_t$, $z = x_*$ and $w = x_t$, we prove that

$$\nabla^2 f(x_t) \succeq \frac{1}{1 + M \|x_t - x_*\|_{x_*}} \nabla^2 f(x_*) \succeq \frac{1}{1 + 2\omega^{-1}(\frac{M^2}{4} C_t)} \nabla^2 f(x_*) = \frac{1}{1 + D_t} \nabla^2 f(x_*). \quad (42)$$

Proof for (b): Recall the definition of G_t in (30). Notice that $\|x_t + \tau_2(x_* - x_t) - x_*\|_{x_*} = (1 - \tau_2)\|x_t - x_*\|_{x_*} \leq \|x_t - x_*\|_{x_*} \leq \frac{2}{M}\omega^{-1}(\frac{M^2}{4}C_t)$. Similar to the arguments in (a), in Assumptions 2.2, take $x = x_t + \tau_2(x_* - x_t)$, $y = x_*$ and $z = x_*$, we prove that

$$\begin{aligned} G_t &\preceq (1 + M\|x_t + \tau(x_* - x_t) - x_*\|_{x_*})\nabla^2 f(x_*) \\ &\preceq (1 + 2\omega^{-1}(\frac{M^2}{4}C_t))\nabla^2 f(x_*) = (1 + D_t)\nabla^2 f(x_*). \end{aligned} \quad (43)$$

Similarly, take $x = x_*$, $y = x_t + \tau_2(x_* - x_t)$ and $z = x_*$, we prove that

$$\begin{aligned} G_t &\succeq \frac{1}{1 + M\|x_t + \tau(x_* - x_t) - x_*\|_{x_*}}\nabla^2 f(x_*) \\ &\succeq \frac{1}{1 + 2\omega^{-1}(\frac{M^2}{4}C_t)}\nabla^2 f(x_*) = \frac{1}{1 + D_t}\nabla^2 f(x_*). \end{aligned} \quad (44)$$

Proof for (c): Notice that $\|x_t + \tau(x_* - x_t) - (x_t + \tau_2(x_* - x_t))\|_{x_*} = |\tau - \tau_2|\|x_t - x_*\|_{x_*} \leq \|x_t - x_*\|_{x_*} \leq \frac{2}{M}\omega^{-1}(\frac{M^2}{4}C_t)$. Similar to the arguments in (a), in Assumptions 2.2, take $x = x_t + \tau(x_* - x_t)$, $y = x_t + \tau_2(x_* - x_t)$ and $z = x_*$, we prove that

$$\begin{aligned} \nabla^2 f(x_t + \tau(x_* - x_t)) &\preceq (1 + M\|x_t + \tau(x_* - x_t) - x_*\|_{x_*})G_t \\ &\preceq (1 + 2\omega^{-1}(\frac{M^2}{4}C_t))G_t = (1 + D_t)G_t. \end{aligned} \quad (45)$$

Similarly, take $x = x_t + \tau_2(x_* - x_t)$, $y = x_t + \tau(x_* - x_t)$ and $z = x_*$, we prove that

$$\begin{aligned} \nabla^2 f(x_t + \tau(x_* - x_t)) &\succeq \frac{1}{1 + M\|x_t + \tau(x_* - x_t) - x_*\|_{x_*}}G_t \\ &\succeq \frac{1}{1 + 2\omega^{-1}(\frac{M^2}{4}C_t)}G_t = \frac{1}{1 + D_t}G_t. \end{aligned} \quad (46)$$

Proof for (d): Recall the definition of J_t in (29). Notice that $\|x_t + \tau(x_{t+1} - x_t) - x_*\|_{x_*} = (1 - \tau)\|x_t - x_*\|_{x_*} + \tau\|x_{t+1} - x_*\|_{x_*} \leq (1 - \tau)\frac{2}{M}\omega^{-1}(\frac{M^2}{4}C_t) + \tau\frac{2}{M}\omega^{-1}(\frac{M^2}{4}C_t) \leq \frac{2}{M}\omega^{-1}(\frac{M^2}{4}C_t)$ where the last inequality holds since $f(x_{t+1}) \leq f(x_t)$ and $\omega^{-1}(x)$ is increasing function. Hence, similar to the arguments in (a), in Assumptions 2.2, take $x = x_t + \tau(x_{t+1} - x_t)$, $y = x_*$ and $z = x_*$, we prove that

$$\begin{aligned} J_t &\preceq (1 + M\|x_t + \tau(x_{t+1} - x_t) - x_*\|_{x_*})\nabla^2 f(x_*) \\ &\preceq (1 + 2\omega^{-1}(\frac{M^2}{4}C_t))\nabla^2 f(x_*) = (1 + D_t)\nabla^2 f(x_*). \end{aligned} \quad (47)$$

Similarly, take $x = x_*$, $y = x_t + \tau(x_{t+1} - x_t)$ and $z = x_*$, we prove that

$$\begin{aligned} J_t &\succeq \frac{1}{1 + M\|x_t + \tau(x_{t+1} - x_t) - x_*\|_{x_*}}\nabla^2 f(x_*) \\ &\succeq \frac{1}{1 + 2\omega^{-1}(\frac{M^2}{4}C_t)}\nabla^2 f(x_*) = \frac{1}{1 + D_t}\nabla^2 f(x_*). \end{aligned} \quad (48)$$

□

Proposition B.6. Let $\{x_t\}_{t \geq 0}$ be the iterates generated by BFGS. Recall the definitions of weighted vectors in (9) and notations in (27). Then, for any weight matrix $P \in \mathbb{S}_{++}^d$ and any $t \geq 1$, we have

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - \left(\prod_{i=0}^{t-1} \hat{p}_i \hat{q}_i \hat{n}_i \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i}\right)^{\frac{1}{t}}\right)^t. \quad (49)$$

Proof. Please check Proposition 1 in [41] □

Lemma B.7. Suppose Assumption 2.2 holds for the convex objective function $f(x)$ and recall the definition \tilde{g}_t in (28) and D_t in (11). We have the following condition,

$$\frac{\|\tilde{g}_t\|^2}{f(x_t) - f(x_*)} \geq \frac{1}{1 + D_t}, \quad \forall t \geq 0. \quad (50)$$

Proof. Since f is convex, we know that for any $x, y \in \mathbb{R}^d$, we have $f(y) \geq f(x) + g(x)^\top (y - x)$. Take $x = x_t$ and $y = x_*$, we obtain that

$$f(x_t) - f(x_*) \leq g_t^\top (x_t - x_*).$$

Using mean value theorem and the fact that $\nabla f(x_*) = 0$, we have that

$$g_t = \nabla f(x_t) = \nabla f(x_t) - \nabla f(x_*) = G_t(x_t - x_*),$$

where G_t is defined in (30) for some $\tau_2 \in [0, 1]$. Hence, we prove that

$$\begin{aligned} f(x_t) - f(x_*) &\leq g_t^\top (x_t - x_*) = g_t^\top G_t^{-1} g_t \\ &\leq \left(1 + 2\omega^{-1}\left(\frac{M^2}{4}C_t\right)\right) g_t^\top \nabla^2 f(x_*)^{-1} g_t = (1 + D_t) \|\tilde{g}_t\|^2, \end{aligned}$$

where we use the result in (38) from Lemma B.5. \square

Lemma B.8. Suppose Assumption 2.2 holds for the convex objective function $f(x)$ and recall the definition \tilde{g}_t in (28) and D_t in (11). We have the following condition,

$$\frac{2}{(1 + D_t)^2} \leq \frac{\|\tilde{g}_t\|^2}{f(x_t) - f(x_*)} \leq 2(1 + D_t)^2, \quad \forall t \geq 0. \quad (51)$$

Proof. By applying Taylor's theorem with Lagrange remainder, there exists $\tilde{\tau}_t \in [0, 1]$ such that

$$\begin{aligned} f(x_t) &= f(x_*) + \nabla f(x_*)^\top (x_t - x_*) + \frac{1}{2}(x_t - x_*)^\top \nabla^2 f(x_t + \tilde{\tau}_t(x_* - x_t))(x_t - x_*) \\ &= f(x_*) + \frac{1}{2}(x_t - x_*)^\top \nabla^2 f(x_t + \tilde{\tau}_t(x_* - x_t))(x_t - x_*), \end{aligned} \quad (52)$$

where we used the fact that $\nabla f(x_*) = 0$ in the last equality. Moreover, by the fundamental theorem of calculus, we have

$$\nabla f(x_t) - \nabla f(x_*) = \nabla^2 f(x_t + \tau(x_* - x_t))(x_t - x_*) = G_t(x_t - x_*),$$

where we use the definition of G_t in (30). Since $\nabla f(x_*) = 0$ and we denote $g_t = \nabla f(x_t)$, this further implies that

$$x_t - x_* = G_t^{-1}(\nabla f(x_t) - \nabla f(x_*)) = G_t^{-1} g_t. \quad (53)$$

Combining (52) and (53) leads to

$$f(x_t) - f(x_*) = \frac{1}{2} g_t^\top G_t^{-1} \nabla^2 f(x_t + \tilde{\tau}_t(x_* - x_t)) G_t^{-1} g_t. \quad (54)$$

Based on (39) in Lemma B.5, we have $\nabla^2 f(x_t + \tilde{\tau}_t(x_* - x_t)) \preceq (1 + D_t) G_t$, which implies

$$G_t^{-1} \nabla^2 f(x_t + \tilde{\tau}_t(x_* - x_t)) G_t^{-1} \preceq (1 + D_t)^2 G_t^{-1}. \quad (55)$$

Moreover, it follows from (38) in Lemma B.5 that $\frac{1}{1+D_t} \nabla^2 f(x_*) \preceq G_t$, which implies that

$$G_t^{-1} \preceq (1 + D_t)(\nabla^2 f(x_*))^{-1}. \quad (56)$$

Combining (55) and (56), we obtain that

$$G_t^{-1} \nabla^2 f(x_t + \tilde{\tau}_t(x_* - x_t)) G_t^{-1} \preceq (1 + D_t)^2 (\nabla^2 f(x_*))^{-1},$$

and hence

$$g_t^\top G_t^{-1} \nabla^2 f(x_t + \tilde{\tau}_t(x_* - x_t)) G_t^{-1} g_t \leq (1 + D_t)^2 g_t^\top (\nabla^2 f(x_*))^{-1} g_t.$$

By using (54) and the fact that $g_t^\top (\nabla^2 f(x_*))^{-1} g_t = \|\tilde{g}_t\|^2$, we obtain that

$$\frac{\|\tilde{g}_t\|^2}{f(x_t) - f(x_*)} \geq \frac{2}{(1 + D_t)^2},$$

and the left claim follows. Using the similar method, we can prove the right claim. \square

Lemma B.9. Suppose Assumption 2.2 holds and $C_t \leq \min\{\frac{1}{16}, \frac{4}{M^2}\omega(\frac{1}{32})\}$ and $\rho_t \geq \frac{15}{16}$ at iteration t , then we have that

$$f(x_t + d_t) \leq f(x_t). \quad (57)$$

Proof. Notice that using (37) from Lemma B.5, we have that

$$d_t^\top \nabla^2 f(x_t) d_t \leq (1 + D_t) d_t^\top \nabla^2 f(x_*) d_t = (1 + D_t) \|\tilde{d}_t\|^2. \quad (58)$$

Since $-\tilde{g}_t^\top \tilde{d}_t \leq \|\tilde{g}_t\| \|\tilde{d}_t\|$ by Cauchy-Schwarz inequality where $\tilde{g}_t = \nabla^2 f(x_*)^{-\frac{1}{2}} g_t$, we obtain

$$\|\tilde{d}_t\| = \|\tilde{g}_t\| \frac{\|\tilde{d}_t\|}{\|\tilde{g}_t\|} \leq \|\tilde{g}_t\| \frac{\|\tilde{d}_t\|^2}{-\tilde{g}_t^\top \tilde{d}_t} = \frac{1}{\rho_t} \|\tilde{g}_t\|. \quad (59)$$

Using the right inequality in Lemma B.8, we have that

$$\|\tilde{g}_t\|^2 \leq 2(1 + D_t)^2 (f(x_t) - f(x_*)) = 2(1 + D_t)^2 C_t. \quad (60)$$

Leveraging (58), (59) and (60), we obtain that

$$d_t^\top \nabla^2 f(x_t) d_t \leq (1 + D_t) \|\tilde{d}_t\|^2 \leq \frac{1 + D_t}{\rho_t^2} \|\tilde{g}_t\|^2 \leq \frac{2(1 + D_t)^3}{\rho_t^2} C_t.$$

Since $C_t \leq \frac{1}{16}$, $D_t = 2\omega^{-1}(\frac{M^2}{4} C_t) \leq \frac{1}{16}$ and $\rho_t \geq \frac{15}{16}$, we have that

$$\sqrt{d_t^\top \nabla^2 f(x_t) d_t} \leq \sqrt{\frac{2(1 + D_t)^3}{\rho_t^2} C_t} \leq \frac{13}{30} < 1.$$

Applying the second inequality from Lemma B.4 with $x = x_t$ and $y = x_t + d_t$, we have that

$$f(x_t + d_t) \leq f(x_t) + g_t^\top d_t + \omega_*(\|d_t\|_{x_t}),$$

since $\|d_t\|_{x_t} = \sqrt{d_t^\top \nabla^2 f(x_t) d_t} < 1$. Using (d) from Lemma B.3, we have that

$$f(x_t + d_t) - f(x_t) \leq g_t^\top d_t + \omega_*(\|d_t\|_{x_t}) \leq g_t^\top d_t + \frac{\|d_t\|_{x_t}^2}{2(1 - \|d_t\|_{x_t})}.$$

Applying the fact that $\|d_t\|_{x_t} = \sqrt{d_t^\top \nabla^2 f(x_t) d_t} \leq \frac{13}{30}$ and (58), we have that

$$\begin{aligned} f(x_t + d_t) - f(x_t) &\leq g_t^\top d_t + \frac{\|d_t\|_{x_t}^2}{2(1 - \|d_t\|_{x_t})} \leq g_t^\top d_t + \frac{15}{17} \|d_t\|_{x_t}^2 \leq g_t^\top d_t + \frac{15}{17} (1 + D_t) \|\tilde{d}_t\|^2 \\ &= g_t^\top d_t + \frac{15}{17} (1 + D_t) \frac{\|\tilde{d}_t\|^2}{-\tilde{g}_t^\top \tilde{d}_t} (-g_t^\top d_t) = -g_t^\top d_t \left(\frac{15}{17} (1 + D_t) \frac{\|\tilde{d}_t\|^2}{-\tilde{g}_t^\top \tilde{d}_t} - 1 \right) \\ &= -g_t^\top d_t \left(\frac{15}{17} \frac{1 + D_t}{\rho_t} - 1 \right) \end{aligned} \quad (61)$$

Notice that $-g_t^\top d_t = -g_t^\top B_t^{-1} g_t > 0$ and when $D_t \leq \frac{1}{16}$ and $\rho_t \geq \frac{15}{16}$, we can verify that

$$\frac{15}{17} \frac{1 + D_t}{\rho_t} \leq 1.$$

Therefore, (61) implies the conclusion that

$$f(x_t + d_t) - f(x_t) \leq 0.$$

□

Lemma B.10. Recall $\hat{p}_t = \frac{f(x_t) - f(x_{t+1})}{-\tilde{g}_t^\top \tilde{s}_t}$ and $\hat{n}_t = \frac{\tilde{g}_t^\top \tilde{s}_t}{-\tilde{g}_t^\top \tilde{s}_t}$ defined in (27). If the unit step size $\eta_t = 1$ satisfies the Armijo-Wolfe conditions (5) and (6), then we have that

$$\hat{p}_t \geq 1 - \frac{1 + D_t}{2\rho_t}, \quad \hat{n}_t \geq \frac{1}{(1 + D_t)\rho_t}. \quad (62)$$

Proof. Please check Lemma 6.1 in [42]. The only difference is that C_t is replaced by D_t defined in (11). \square

Proposition B.11. *Let $\{B_t\}_{t \geq 0}$ be the Hessian approximation matrices generated by the BFGS update in (4). Suppose Assumptions 2.2 holds and $f(x_{t+1}) \leq f(x_t)$ for any $t \geq 0$. Recall the definition of $\Psi(\cdot)$ in (8) and D_t in (11), we have*

$$\sum_{i=0}^{t-1} \omega(\rho_i - 1) \leq \Psi(\bar{B}_0) + 2 \sum_{i=0}^{t-1} D_i. \quad (63)$$

Proof. Please check Proposition G.2 in [42]. The only difference is C_t is replaced by D_t defined in (11). \square

C Proof of Lemmas, Propositions and Theorems

C.1 Proof of Proposition 2.1

We use induction to prove that $x_t = A\dot{x}_t$ and $B_t = (A^{-1})^\top \dot{B}_t A^{-1}$ for any $t \geq 0$. Notice that when $t = 0$, we already have that $x_0 = A\dot{x}_0$ and $B_0 = (A^{-1})^\top \dot{B}_0 A^{-1}$ since $\dot{x}_0 = A^{-1}x_0$ and $\dot{B}_0 = A^\top B_0 A$ where A is non-singular. Suppose that the conditions hold for $t = k$ with $k \geq 0$, i.e., $x_k = A\dot{x}_k$ and $B_k = (A^{-1})^\top \dot{B}_k A^{-1}$. We consider the case $t = k + 1$. We have

$$\begin{aligned} x_{k+1} &= x_k - \eta_k B_k^{-1} \nabla f(x_k) = A\dot{x}_k - \eta_k A \dot{B}_k^{-1} A^\top \nabla f(A\dot{x}_k) \\ &= A\dot{x}_k - \eta_k A \dot{B}_k^{-1} A^\top (A^\top)^{-1} \nabla \phi(\dot{x}_k) = A(\dot{x}_k - \eta_k \dot{B}_k^{-1} \nabla \phi(\dot{x}_k)) = A\dot{x}_{k+1}. \end{aligned}$$

Suppose that $\dot{s}_k = \dot{x}_{k+1} - \dot{x}_k$ and $\dot{y}_k = \nabla \phi(\dot{x}_{k+1}) - \nabla \phi(\dot{x}_k)$, we have that $s_k = x_{k+1} - x_k = A\dot{s}_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k) = \nabla f(A\dot{x}_{k+1}) - \nabla f(A\dot{x}_k) = (A^\top)^{-1}(\nabla \phi(\dot{x}_{k+1}) - \nabla \phi(\dot{x}_k)) = (A^\top)^{-1} \dot{y}_k$. Hence, we have that

$$\begin{aligned} B_{k+1} &= B_k - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{s_k^\top y_k} \\ &= (A^{-1})^\top \dot{B}_k A^{-1} - \frac{(A^{-1})^\top \dot{B}_k A^{-1} A \dot{s}_k \dot{s}_k^\top A^\top (A^{-1})^\top \dot{B}_k A^{-1}}{\dot{s}_k A^\top (A^{-1})^\top \dot{B}_k A^{-1} A \dot{s}_k} + \frac{(A^\top)^{-1} \dot{y}_k \dot{y}_k^\top A^{-1}}{\dot{s}_k A^\top (A^\top)^{-1} \dot{y}_k} \\ &= (A^{-1})^\top \left(\dot{B}_k - \frac{\dot{B}_k \dot{s}_k \dot{s}_k^\top \dot{B}_k}{\dot{s}_k^\top \dot{B}_k \dot{s}_k} + \frac{\dot{y}_k \dot{y}_k^\top}{\dot{s}_k^\top \dot{y}_k} \right) A^{-1} \\ &= (A^{-1})^\top \dot{B}_{k+1} A^{-1}. \end{aligned} \quad (64)$$

We prove that $x_{k+1} = A\dot{x}_{k+1}$ and $B_{k+1} = (A^{-1})^\top \dot{B}_{k+1} A^{-1}$. Therefore, we prove that $\dot{x}_t = A^{-1}x_t$ and $\dot{B}_t = A^\top B_t A$ for any $t \geq 0$ using induction. It is obvious that $\phi(\dot{x}_t) = f(A\dot{x}_t) = f(AA^{-1}x_t) = f(x_t)$ for any $t \geq 0$. Therefore, The BFGS quasi-Newton method is affine invariant.

C.2 Proof of Theorem 3.1

We choose $P = (1 + D_0) \nabla^2 f(x_*)$ throughout the proof. Note that given this weight matrix P , it can be easily verified that for any $t \geq 0$,

$$\frac{\|\hat{y}_t\|^2}{\hat{s}_t^\top \hat{y}_t} = \frac{\hat{s}_t^\top \hat{J}_t^2 \hat{s}_t}{\hat{s}_t^\top \hat{J}_t \hat{s}_t} \leq \|\hat{J}_t\| = \frac{\|\nabla^2 f(x_*)^{-\frac{1}{2}} J_t \nabla^2 f(x_*)^{-\frac{1}{2}}\|}{1 + D_0} \leq \frac{1 + D_t}{1 + D_0} \leq 1, \quad (65)$$

where J_t is defined in (29) and we use (40) in Lemma B.5 as well as the fact that $f(x_{t+1}) \leq f(x_t)$, $D_t \leq D_0$ and ω^{-1} is increasing function. Hence, we use (33) in Proposition B.2 with $\bar{B}_0 = \bar{B}_0$ defined in (12) to obtain

$$\sum_{i=0}^{t-1} \log \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq -\Psi(\bar{B}_0) + \sum_{i=0}^{t-1} \left(1 - \frac{\|\hat{y}_i\|^2}{\hat{s}_i^\top \hat{y}_i} \right) \geq -\Psi(\bar{B}_0),$$

which further implies that

$$\prod_{i=0}^{t-1} \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq e^{-\Psi(\bar{B}_0)}.$$

Moreover, for the choice $P = (1 + D_0) \nabla^2 f(x_*)$, it can be shown that

$$\hat{q}_t = \frac{\|\tilde{g}_t\|^2}{(1 + D_0)(f(x_t) - f(x_*))} \geq \frac{1}{(1 + D_0)(1 + D_t)} \geq \frac{1}{(1 + D_0)^2}. \quad (66)$$

by using result in Lemma B.7. From Lemma B.1, we know $\hat{p}_t \geq \alpha$ and $\hat{n}_t \geq 1 - \beta$, which lead to

$$\prod_{i=0}^{t-1} \frac{\hat{p}_i \hat{n}_i \hat{q}_i}{\hat{m}_i} \cos^2(\hat{\theta}_i) \geq \prod_{i=0}^{t-1} \hat{p}_i \prod_{i=0}^{t-1} \hat{q}_i \prod_{i=0}^{t-1} \hat{n}_i \prod_{i=0}^{t-1} \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq \left(\frac{\alpha(1 - \beta)}{(1 + D_0)^2} \right)^t e^{-\Psi(\bar{B}_0)}.$$

Thus, it follows from Proposition B.6 that

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left[1 - \left(\prod_{i=0}^{t-1} \frac{\hat{p}_i \hat{q}_i \hat{n}_i}{\hat{m}_i} \cos^2(\hat{\theta}_i) \right)^{\frac{1}{t}} \right]^t \leq \left(1 - \frac{\alpha(1 - \beta) e^{-\frac{\Psi(\bar{B}_0)}{t}}}{(1 + D_0)^2} \right)^t.$$

This completes the proof. (14) can be easily verified since when $t \geq \Psi(\bar{B}_0)$, we have $e^{-\frac{\Psi(\bar{B}_0)}{t}} \geq \frac{1}{3}$.

C.3 Proof of Theorem 3.3

First, we prove the following result holds:

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - 2\alpha(1 - \beta) e^{-\frac{\Psi(\bar{B}_0) + 3 \sum_{i=0}^{t-1} D_i}{t}} \right)^t. \quad (67)$$

We choose the weight matrix as $P = \nabla^2 f(x_*)$ throughout the proof. Similar to the proof of Theorem 3.1, we start from the key inequality in (49), but we apply different bounds on the \hat{q}_t and $\frac{\cos^2(\hat{\theta}_t)}{\hat{m}_t}$. Specifically, we have that

$$\frac{\|\hat{y}_t\|^2}{\hat{s}_t^\top \hat{y}_t} = \frac{\hat{s}_t^\top \hat{J}_t^2 \hat{s}_t}{\hat{s}_t^\top \hat{J}_t \hat{s}_t} \leq \|\hat{J}_t\| = \|\nabla^2 f(x_*)^{-\frac{1}{2}} J_t \nabla^2 f(x_*)^{-\frac{1}{2}}\| \leq 1 + D_t. \quad (68)$$

where J_t is defined in (29) and we use (40) in Lemma B.5 as well as the fact that $f(x_{t+1}) \leq f(x_t)$. Hence, we use (33) in Proposition B.2 with $\tilde{B}_0 = \bar{B}_0$ defined in (12) to obtain

$$\sum_{i=0}^{t-1} \log \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq -\Psi(\bar{B}_0) + \sum_{i=0}^{t-1} \left(1 - \frac{\|\hat{y}_i\|^2}{\hat{s}_i^\top \hat{y}_i} \right) \geq -\Psi(\bar{B}_0) - \sum_{i=0}^{t-1} D_i,$$

which further implies that

$$\prod_{i=0}^{t-1} \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq e^{-\Psi(\bar{B}_0) - \sum_{i=0}^{t-1} D_i}. \quad (69)$$

Moreover, since $\hat{q}_t = \frac{\|\tilde{g}_t\|^2}{f(x_t) - f(x_*)} \geq \frac{2}{(1 + 2\omega^{-1}(\frac{M^2}{4} C_t))^2}$ for any $t \geq 0$ by using Lemma B.8, we obtain that

$$\prod_{i=0}^{t-1} \hat{q}_i \geq \prod_{i=0}^{t-1} \frac{2}{(1 + D_i)^2} \geq 2^t \prod_{i=0}^{t-1} e^{-2D_i} = 2^t e^{-2 \sum_{i=0}^{t-1} D_i}, \quad (70)$$

where we use the inequality $1 + x \leq e^x$ for any $x \in \mathbb{R}$. From Lemma B.1, we know $\hat{p}_t \geq \alpha$ and $\hat{n}_t \geq 1 - \beta$, which lead to

$$\prod_{i=0}^{t-1} \hat{p}_i \hat{n}_i \geq \alpha^t (1 - \beta)^t. \quad (71)$$

Combining (69), (70), (71) and (49) from Proposition B.6, we prove that

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left[1 - \left(\prod_{i=0}^{t-1} \frac{\hat{p}_i \hat{q}_i \hat{n}_i}{\hat{m}_i} \cos^2(\hat{\theta}_i) \right)^{\frac{1}{t}} \right]^t \leq \left(1 - 2\alpha(1 - \beta) e^{-\frac{\Psi(\bar{B}_0) + 3 \sum_{i=0}^{t-1} D_i}{t}} \right)^t.$$

This completes the proof. Notice that when

$$t \geq \Psi(\bar{B}_0) + 3 \sum_{i=0}^{t-1} D_i, \quad (72)$$

(67) implies the condition that

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - \frac{2\alpha(1 - \beta)}{e} \right)^t \leq \left(1 - \frac{2\alpha(1 - \beta)}{3} \right)^t, \quad (73)$$

which leads to the linear rate in (18).

Hence, it is sufficient to establish an upper bound on $\sum_{i=0}^{t-1} D_i$. We decompose the sum into two parts: $\sum_{i=0}^{\lceil \Psi(\bar{B}_0) \rceil - 1} D_i$ and $\sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t D_i$. For the first part, note that since $f(x_{i+1}) \leq f(x_i)$ by Lemma B.1, we also have $D_{i+1} \leq D_i$ for $i \geq 0$ using the fact that ω^{-1} is increasing. Hence, we have $\sum_{i=0}^{\lceil \Psi(\bar{B}_0) \rceil - 1} D_i \leq D_0 \lceil \Psi(\bar{B}_0) \rceil \leq D_0(\Psi(\bar{B}_0) + 1)$. Moreover, by Theorem 3.1, when $t \geq \Psi(\bar{B}_0)$ we have

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - e^{-\frac{\Psi(\bar{B}_0)}{t}} \frac{\alpha(1 - \beta)}{(1 + D_0)^2} \right)^t \leq \left(1 - \frac{\alpha(1 - \beta)}{3(1 + D_0)^2} \right)^t.$$

Hence, using the fact that $\omega^{-1}(x) \leq \sqrt{2x} + x$ and the definition of D_t in (11), we obtain that

$$\begin{aligned} \sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t D_i &\leq \sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t \left(\frac{M}{2} \sqrt{2C_i} + \frac{M^2}{4} C_i \right) = \frac{\sqrt{2}M}{2} \sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t \sqrt{C_i} + \frac{M^2}{4} \sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t C_i \\ &= \frac{\sqrt{2}M}{2} \sqrt{C_0} \sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t \sqrt{\frac{f(x_i) - f(x_*)}{f(x_0) - f(x_*)}} + \frac{M^2}{4} C_0 \sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t \frac{f(x_i) - f(x_*)}{f(x_0) - f(x_*)} \\ &\leq \frac{\sqrt{2}M}{2} \sqrt{C_0} \sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t \left(1 - \frac{\alpha(1 - \beta)}{3(1 + D_0)^2} \right)^{\frac{i}{2}} + \frac{M^2}{4} C_0 \sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t \left(1 - \frac{\alpha(1 - \beta)}{3(1 + D_0)^2} \right)^i \\ &\leq \frac{\sqrt{2}M}{2} \sqrt{C_0} \sum_{i=1}^{\infty} \left(1 - \frac{\alpha(1 - \beta)}{3(1 + D_0)^2} \right)^{\frac{i}{2}} + \frac{M^2}{4} C_0 \sum_{i=1}^{\infty} \left(1 - \frac{\alpha(1 - \beta)}{3(1 + D_0)^2} \right)^i \\ &\leq \frac{\sqrt{2}M}{2} \sqrt{C_0} \left(\frac{6(1 + D_0)^2}{\alpha(1 - \beta)} - 1 \right) + \frac{M^2}{4} C_0 \left(\frac{3(1 + D_0)^2}{\alpha(1 - \beta)} - 1 \right) \end{aligned}$$

where we used the fact that $\sum_{i=1}^{\infty} (1 - \rho)^{\frac{i}{2}} = \frac{\sqrt{1 - \rho}}{1 - \sqrt{1 - \rho}} = \frac{\sqrt{1 - \rho} + 1 - \rho}{\rho} \leq \frac{2}{\rho} - 1$ and $\sum_{i=1}^{\infty} (1 - \rho)^i = \frac{1 - \rho}{1 - (1 - \rho)} = \frac{1}{\rho} - 1$ for any $\rho \in (0, 1)$. Hence, by combining both inequalities, we have

$$\begin{aligned} \sum_{i=0}^{t-1} D_i &= \sum_{i=0}^{\lceil \Psi(\bar{B}_0) \rceil - 1} D_i + \sum_{i=\lceil \Psi(\bar{B}_0) \rceil}^t D_i \\ &\leq D_0 \Psi(\bar{B}_0) + \frac{\sqrt{2}M}{2} \sqrt{C_0} \frac{6(1 + D_0)^2}{\alpha(1 - \beta)} + \frac{M^2}{4} C_0 \frac{3(1 + D_0)^2}{\alpha(1 - \beta)} \\ &= D_0 \Psi(\bar{B}_0) + (M\sqrt{2C_0} + \frac{M^2}{4} C_0) \frac{3(1 + D_0)^2}{\alpha(1 - \beta)} \leq D_0 \left(\Psi(\bar{B}_0) + \frac{3(1 + D_0)^2}{\alpha(1 - \beta)} \right), \end{aligned} \quad (74)$$

where the last inequality is due to (d) from Lemma B.3 and the definition of D_t in (11). Hence, when

$$t \geq \Psi(\bar{B}_0) + 3D_0 \left(\Psi(\bar{B}_0) + \frac{3(1 + D_0)^2}{\alpha(1 - \beta)} \right) \geq \Psi(\bar{B}_0) + 3 \sum_{i=0}^{t-1} D_i,$$

using result from (67), we have the linear convergence rate in (18).

C.4 Proof of Lemma 4.1

Denote $\bar{x}_{t+1} = x_t + d_t$ and $\bar{s}_t = \bar{x}_{t+1} - x_t = d_t$. Since $\delta_1 \leq \min\{\frac{1}{16}, \frac{4}{M^2}\omega(\frac{1}{32})\}$ and $\delta_2 \geq \frac{15}{16}$, we have $f(\bar{x}_{t+1}) \leq f(x_t)$ from Lemma B.9. Using Taylor's expansion, we have that $f(\bar{x}_{t+1}) = f(x_t) + g_t^\top d_t + \frac{1}{2}d_t^\top \nabla^2 f(x_t + \hat{\tau}(\bar{x}_{t+1} - x_t))d_t$, where $\hat{\tau} \in [0, 1]$. Hence, we have

$$\begin{aligned} \frac{f(x_t) - f(\bar{x}_{k+1})}{-g_t^\top d_t} &= \frac{-g_t^\top d_t - \frac{1}{2}d_t^\top \nabla^2 f(x_t + \hat{\tau}(\bar{x}_{t+1} - x_t))d_t}{-g_t^\top d_t} \\ &= 1 - \frac{1}{2} \frac{d_t^\top \nabla^2 f(x_t + \hat{\tau}(\bar{x}_{t+1} - x_t))d_t}{-g_t^\top d_t} \geq 1 - \frac{1 + D_t}{2} \frac{d_t^\top \nabla^2 f(x_*)d_t}{-g_t^\top d_t} = 1 - \frac{1 + D_t}{2\rho_t}, \end{aligned}$$

where we apply the (40) from Lemma B.5 since $f(\bar{x}_{t+1}) \leq f(x_t)$. Therefore, when $C_t \leq \delta_1 \leq \frac{4}{M^2}\omega(\frac{\sqrt{2(1-\alpha)}-1}{2})$ and $\rho_t \geq \delta_2 \geq \frac{1}{\sqrt{2(1-\alpha)}}$, we obtain that $\frac{f(x_t) - f(\bar{x}_{k+1})}{-g_t^\top d_t} \geq 1 - \frac{1+D_t}{2\rho_t} = 1 - \frac{1+2\omega^{-1}(\frac{M}{4}C_t)}{2\rho_t} \geq \alpha$ and unit step size $\eta_t = 1$ satisfies the sufficient condition (5).

Similarly, using (40) from Lemma B.5 since $f(\bar{x}_{t+1}) \leq f(x_t)$ and denote $\bar{g}_{k+1} = \nabla f(\bar{x}_{t+1})$, $\bar{y}_t = \bar{g}_{k+1} - g_t$, we have that

$$\frac{\bar{y}_t^\top \bar{s}_t}{-g_t^\top \bar{s}_t} = \frac{\bar{s}_t^\top J_t \bar{s}_t}{-g_t^\top \bar{s}_t} = \frac{d_t^\top J_t d_t}{-g_t^\top d_t} \geq \frac{1}{1 + D_t} \frac{d_t^\top \nabla^2 f(x_*)d_t}{-g_t^\top d_t} = \frac{1}{(1 + D_t)\rho_t}.$$

Therefore, when $C_t \leq \delta_1 \leq \frac{4}{M^2}\omega(\frac{1}{2}(\frac{1}{\sqrt{1-\beta}} - 1))$ and $\rho_t \leq \delta_3 = \frac{1}{\sqrt{1-\beta}}$, we obtain that $\frac{\bar{y}_t^\top \bar{s}_t}{-g_t^\top \bar{s}_t} \geq \frac{1}{(1+D_t)\rho_t} = \frac{1}{(1+2\omega^{-1}(\frac{M}{4}C_t))\rho_t} \geq 1 - \beta$, which indicates that $\bar{g}_{t+1}^\top d_t = \bar{g}_{t+1}^\top \bar{s}_t = \bar{y}_t^\top \bar{s}_t + g_t^\top \bar{s}_t \geq -g_t^\top \bar{s}_t(1 - \beta) + g_t^\top \bar{s}_t = \beta g_t^\top \bar{s}_t = \beta g_t^\top d_t$. Hence, unit step size $\eta_t = 1$ satisfies the curvature condition (6). Therefore, we prove that when $C_t \leq \delta_1$ and $\delta_2 \leq \rho_t \leq \delta_3$, step size $\eta_t = 1$ satisfies the Armijo-Wolfe conditions (5) and (6).

C.5 Proof of Lemma 4.2

Since in Theorem 3.1, we already prove that

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - \frac{\alpha(1-\beta)e^{-\frac{\Psi(\bar{B}_0)}{t}}}{(1+D_0)^2}\right)^t.$$

This implies that

$$C_t \leq \left(1 - \frac{\alpha(1-\beta)e^{-\frac{\Psi(\bar{B}_0)}{t}}}{(1+D_0)^2}\right)^t C_0.$$

When $t \geq \Psi(\bar{B}_0)$, we obtain that

$$C_t \leq \left(1 - \frac{\alpha(1-\beta)}{3(1+D_0)^2}\right)^t C_0.$$

When $t \geq \frac{3(1+D_0)^2}{\alpha(1-\beta)} \log \frac{C_0}{\delta_1}$, we obtain that

$$C_t \leq \left(1 - \frac{\alpha(1-\beta)}{3(1+D_0)^2}\right)^t C_0 \leq \delta_1.$$

Therefore, the first claim in (23) follows.

Define $I_1 = \{t_0 \leq i \leq t-1 : \rho_i < \delta_2\}$ and $I_2 = \{t_0 \leq i \leq t-1 : \rho_i > \delta_3\}$, we know that $|I| = |I_1| + |I_2|$. Notice that for $t \in I_1$, we have that $\rho_t - 1 < \delta_2 - 1 < 0$ since $\delta_2 < 1$ and the function $\omega(x)$ defined in (10) is decreasing for $-1 < x < 0$ from (a) in Lemma B.3. Hence, we have that $\sum_{i \in I_1} \omega(\rho_i - 1) \geq \sum_{i \in I_1} \omega(\delta_2 - 1) = \omega(\delta_2 - 1)|I_1|$. Similarly, we have that for $t \in I_2$, we have that $\rho_i - 1 > \delta_3 - 1 > 0$ since $\delta_3 > 1$ and the function $\omega(x)$ is increasing for $x > 0$ from (a) in Lemma B.3. Hence, we have that $\sum_{i \in I_2} \omega(\rho_i - 1) \geq \sum_{i \in I_2} \omega(\delta_3 - 1) = \omega(\delta_3 - 1)|I_2|$. Using

(63) from Proposition B.11, we have that $\sum_{i=0}^{t-1} \omega(\rho_i - 1) \leq \Psi(\tilde{B}_0) + 2 \sum_{i=0}^{t-1} D_i$ for any $t \geq 1$. Therefore, we obtain that

$$\begin{aligned} \Psi(\tilde{B}_0) + 2 \sum_{i=0}^{t-1} D_i &\geq \sum_{i=0}^{t-1} \omega(\rho_i - 1) \geq \sum_{i \in I_1} \omega(\beta_i - 1) + \sum_{i \in I_2} \omega(\beta_i - 1) \\ &\geq \omega(\delta_2 - 1)|I_1| + \omega(\delta_3 - 1)|I_2| \geq \min\{\omega(\delta_2 - 1), \omega(\delta_3 - 1)\}(|I_1| + |I_2|), \end{aligned}$$

which leads to the result

$$\begin{aligned} |I| = |I_1| + |I_2| &\leq \frac{\Psi(\tilde{B}_0) + 2 \sum_{i=0}^{t-1} D_i}{\min\{\omega(\delta_2 - 1), \omega(\delta_3 - 1)\}} \\ &= \delta_4 \left(\Psi(\tilde{B}_0) + 2 \sum_{i=0}^{t-1} D_i \right) \leq \delta_4 \left(\Psi(\tilde{B}_0) + 2D_0(\Psi(\tilde{B}_0) + \frac{3(1+D_0)^2}{\alpha(1-\beta)}) \right), \end{aligned} \quad (75)$$

where $\delta_4 = \frac{1}{\min\{\omega(\delta_2-1), \omega(\delta_3-1)\}}$ and the last inequality is due to (74).

C.6 Proof of Theorem 4.3

First, we prove the following result:

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(\frac{\delta_6 t_0 + \delta_7 \Psi(\tilde{B}_0) + \delta_8 \sum_{i=0}^{t-1} D_i}{t} \right)^t. \quad (76)$$

We choose the weight matrix as $P = \nabla^2 f(x_*)$ throughout the proof. Taking the sum from 0 to $t-1$ in inequality (32) of the Proposition B.2, we obtain that

$$\sum_{i=0}^{t-1} \log \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq -\Psi(\tilde{B}_0) + \sum_{i=0}^{t-1} \left(1 - \frac{\|\hat{y}_i\|^2}{\hat{y}_i^\top \hat{s}_i} \right), \quad \forall t \geq 1. \quad (77)$$

Notice that $\frac{\|\hat{y}_i\|^2}{\hat{y}_i^\top \hat{s}_i} = \frac{\|\hat{J}_i \hat{s}_i\|^2}{\hat{s}_i^\top \hat{J}_i \hat{s}_i} \leq \|\hat{J}_i\| \leq 1 + D_i$ where \hat{J}_i is defined in (9) with $P = \nabla^2 f(x_*)$ and we use (40) from Lemma B.5. Therefore, we have that

$$\prod_{i=0}^{t-1} \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq e^{-\Psi(\tilde{B}_0) + \sum_{i=0}^{t-1} \left(1 - \frac{\|\hat{y}_i\|^2}{\hat{y}_i^\top \hat{s}_i} \right)} \geq e^{-\Psi(\tilde{B}_0) - \sum_{i=0}^{t-1} D_i}. \quad (78)$$

where $\cos(\hat{\theta}_i)$ is defined in (26). Recall the definitions in (27) and the results in Lemma B.8, we have

$$\prod_{i=0}^{t-1} \hat{q}_i \geq \prod_{i=0}^{t-1} \frac{2}{(1+D_i)^2} \geq 2^t e^{-2 \sum_{i=0}^{t-1} D_i}. \quad (79)$$

Recall the definition of the set $I = \{t_0 \leq i \leq t-1 : \rho_i \notin [\delta_2, \delta_3]\}$ and define the set $\bar{I} = \{t_0 \leq i \leq t-1 : \rho_i \in [\delta_2, \delta_3]\}$ for any $t > t_0$. Then, we have that

$$\prod_{i=0}^{t-1} \hat{p}_i \hat{n}_i = \prod_{i=0}^{t_0-1} \hat{p}_i \hat{n}_i \prod_{i \in I} \hat{p}_i \hat{n}_i \prod_{i \in \bar{I}} \hat{p}_i \hat{n}_i. \quad (80)$$

From Lemma B.1, we know $\hat{p}_t \geq \alpha$ and $\hat{n}_t \geq 1 - \beta$ for any $t \geq 0$, which lead to

$$\prod_{i=0}^{t_0-1} \hat{p}_i \hat{n}_i \geq \alpha^{t_0} (1 - \beta)^{t_0} = \frac{1}{2^{t_0}} e^{-t_0 \log \frac{1}{2\alpha(1-\beta)}}. \quad (81)$$

$$\begin{aligned} \prod_{i \in I} \hat{p}_i \hat{n}_i &\geq \prod_{i \in I} \alpha(1 - \beta) = \frac{1}{2^{|I|}} e^{-|I| \log \frac{1}{2\alpha(1-\beta)}} \\ &\geq \frac{1}{2^{|I|}} e^{-|I| \log \frac{1}{2\alpha(1-\beta)}} \geq \frac{1}{2^{|I|}} e^{-\delta_4 \left(\Psi(\tilde{B}_0) + 2 \sum_{i=0}^{t-1} D_i \right) \log \frac{1}{2\alpha(1-\beta)}}, \end{aligned} \quad (82)$$

where the second inequality holds since $\log \frac{1}{2\alpha(1-\beta)} > 0$ and the last inequality holds since (75) from the proof of Lemma 4.2. Notice that when index $i \in \bar{I}$, we have $C_i \leq \delta_1$ from Lemma 4.2 and $\rho_i \in [\delta_2, \delta_3]$. Applying Lemma B.10 and Lemma 4.1, we know that for $i \in \bar{I}$, $\eta_i = 1$ satisfies the Armijo-Wolfe conditions (5), (6) and we have $\hat{p}_i \geq 1 - \frac{1+D_i}{2\rho_i} > 0$ and $\hat{n}_i \geq \frac{1}{(1+D_i)\rho_i}$ from (62). Hence, we obtain that

$$\prod_{i \in \bar{I}} \hat{p}_i \hat{n}_i \geq \frac{1}{2^{\bar{I}}} \prod_{i \in \bar{I}} \left(2 - \frac{1+D_i}{\rho_i}\right) \frac{1}{(1+D_i)\rho_i} \geq \frac{1}{2^{|\bar{I}|}} e^{-\sum_{i \in \bar{I}} D_i} \prod_{i \in \bar{I}} \left(2 - \frac{1+D_i}{\rho_i}\right) \frac{1}{\rho_i}, \quad (83)$$

where the last inequality holds since $\frac{1}{1+D_i} \geq e^{-D_i}$. Using the fact that $\log x \geq 1 - \frac{1}{x}$, we obtain

$$\begin{aligned} \prod_{i \in \bar{I}} \left(2 - \frac{1+D_i}{\rho_i}\right) \frac{1}{\rho_i} &= \prod_{i \in \bar{I}} e^{\log(2 - \frac{1+D_i}{\rho_i}) - \log \rho_i} \geq \prod_{i \in \bar{I}} e^{1 - \frac{1}{2 - \frac{1+D_i}{\rho_i}} - \log \rho_i} \\ &= \prod_{i \in \bar{I}} e^{\frac{\rho_i - 1 - D_i}{2\rho_i - 1 - D_i} - \log \rho_i} = \prod_{i \in \bar{I}} e^{\frac{\rho_i - 1 - \log \rho_i + 2(1 - \rho_i) \log \rho_i - (1 - \log \rho_i) D_i}{2\rho_i - 1 - D_i}} \\ &= \prod_{i \in \bar{I}} e^{\frac{\omega(\rho_i - 1) + 2(1 - \rho_i) \log \rho_i - (1 - \log \rho_i) D_i}{2\rho_i - 1 - D_i}} \geq \prod_{i \in \bar{I}} e^{\frac{-2(\rho_i - 1) \log \rho_i - (1 - \log \rho_i) D_i}{2\rho_i - 1 - D_i}} \\ &= \prod_{i \in \bar{I}} e^{-\frac{2(\rho_i - 1) \log \rho_i + (1 - \log \rho_i) D_i}{2\rho_i - 1 - D_i}} \geq \prod_{i \in \bar{I}} e^{-\frac{2(\rho_i - 1) \log \rho_i + (1 - \log \delta_2) D_i}{2\delta_2 - 1 - 1/16}} = \prod_{i \in \bar{I}} e^{-\frac{2(\rho_i - 1) \log \rho_i + (1 - \log \delta_2) D_i}{2\delta_2 - 17/16}}, \end{aligned} \quad (84)$$

where the second inequality holds since $\omega(\rho_i - 1) \geq 0$ and the third inequality holds since $\rho_i \geq \delta_2$ due to $i \in \bar{I}$ and $C_i \leq \delta_1 \leq \frac{4}{M^2} \omega(\frac{1}{32})$, $D_i = 2\omega^{-1}(\frac{M^2}{4} C_i) \leq \frac{1}{16}$ due to $i \geq t_0$ and Lemma 4.2. Notice that $2\rho_i - 1 - D_i \geq 2\delta_2 - 1 - \frac{1}{16} > 0$ for all $i \in \bar{I}$ since $\rho_i \geq \delta_2 \geq \frac{15}{16}$.

When $\rho_i \geq 1$, using $\log \rho_i \leq \rho_i - 1$, (b) in Lemma B.3 and $\rho_i \leq \delta_3$ due to $i \in \bar{I}$, we have that

$$(\rho_i - 1) \log \rho_i \leq (\rho_i - 1)^2 \leq 2\rho_i \omega(\rho_i - 1) \leq 2\delta_3 \omega(\rho_i - 1). \quad (85)$$

Similarly, when $\rho_i < 1$, using $\log \rho_i \geq 1 - \frac{1}{\rho_i}$, (c) in Lemma B.3 and $\rho_i \geq \delta_2$ due to $i \in \bar{I}$, we have

$$(\rho_i - 1) \log \rho_i \leq \frac{(\rho_i - 1)^2}{\rho_i} \leq \frac{\rho_i + 1}{\rho_i} \omega(\rho_i - 1) \leq (1 + \frac{1}{\delta_2}) \omega(\rho_i - 1). \quad (86)$$

Combining (84), (85) and (86), we obtain that

$$\begin{aligned} \prod_{i \in \bar{I}} \left(2 - \frac{1+D_i}{\rho_i}\right) \frac{1}{\rho_i} &\geq \prod_{i \in \bar{I}} e^{-\frac{2(\rho_i - 1) \log \rho_i + (1 - \log \delta_2) D_i}{2\delta_2 - 17/16}} = \prod_{i \in \bar{I}} e^{-\frac{2(\rho_i - 1) \log \rho_i}{2\delta_2 - 17/16}} \prod_{i \in \bar{I}} e^{-\frac{(1 - \log \delta_2) D_i}{2\delta_2 - 17/16}} \\ &= \prod_{i \in \bar{I}, \rho_i < 1} e^{-\frac{2(\rho_i - 1) \log \rho_i}{2\delta_2 - 17/16}} \prod_{i \in \bar{I}, \rho_i \geq 1} e^{-\frac{2(\rho_i - 1) \log \rho_i}{2\delta_2 - 17/16}} \prod_{i \in \bar{I}} e^{-\frac{(1 - \log \delta_2) D_i}{2\delta_2 - 17/16}} \\ &\geq \prod_{i \in \bar{I}, \rho_i < 1} e^{-\frac{2(1 + \frac{1}{\delta_2}) \omega(\rho_i - 1)}{2\delta_2 - 17/16}} \prod_{i \in \bar{I}, \rho_i \geq 1} e^{-\frac{4\delta_3 \omega(\rho_i - 1)}{2\delta_2 - 17/16}} \prod_{i \in \bar{I}} e^{-\frac{(1 - \log \delta_2) D_i}{2\delta_2 - 17/16}} \\ &= e^{-\frac{2 + \frac{2}{\delta_2}}{2\delta_2 - 17/16} \sum_{i \in \bar{I}, \rho_i < 1} \omega(\rho_i - 1) - \frac{4\delta_3}{2\delta_2 - 17/16} \sum_{i \in \bar{I}, \rho_i \geq 1} \omega(\rho_i - 1) - \frac{(1 - \log \delta_2)}{2\delta_2 - 17/16} \sum_{i \in \bar{I}} D_i} \\ &\geq e^{-\delta_5 \left(\sum_{i \in \bar{I}, \rho_i < 1} \omega(\rho_i - 1) + \sum_{i \in \bar{I}, \rho_i \geq 1} \omega(\rho_i - 1) \right) - \frac{(1 - \log \delta_2)}{2\delta_2 - 17/16} \sum_{i \in \bar{I}} D_i} \\ &= e^{-\delta_5 \sum_{i \in \bar{I}} \omega(\rho_i - 1) - \frac{(1 - \log \delta_2)}{2\delta_2 - 17/16} \sum_{i \in \bar{I}} D_i} \end{aligned} \quad (87)$$

where $\delta_5 = \max\{\frac{2+\frac{2}{\delta_2}}{2\delta_2-17/16}, \frac{4\delta_3}{2\delta_2-17/16}\}$. Combining (83) and (87), we obtain that

$$\begin{aligned}
\prod_{i \in \bar{I}} \hat{p}_i \hat{n}_i &\geq \frac{1}{2^{|\bar{I}|}} e^{-\sum_{i \in \bar{I}} D_i} \prod_{i \in \bar{I}} \left(2 - \frac{1 + D_i}{\rho_i}\right) \frac{1}{\rho_i} \\
&\geq \frac{1}{2^{|\bar{I}|}} e^{-\delta_5 \sum_{i \in \bar{I}} \omega(\rho_i - 1) - (1 + \frac{1 - \log \delta_2}{2\delta_2 - 17/16}) \sum_{i \in \bar{I}} D_i} \geq \frac{1}{2^{|\bar{I}|}} e^{-\delta_5 \sum_{i=0}^{t-1} \omega(\rho_i - 1) - \frac{2\delta_2 - \delta_1 - \log \delta_2}{2\delta_2 - 17/16} \sum_{i=0}^{t-1} D_i} \\
&\geq \frac{1}{2^{|\bar{I}|}} e^{-\delta_5 \left(\Psi(\tilde{B}_0) + 2 \sum_{i=0}^{t-1} D_i\right) - \frac{2\delta_2 - 1/16 - \log \delta_2}{2\delta_2 - 17/16} \sum_{i=0}^{t-1} D_i},
\end{aligned} \tag{88}$$

where the last inequality is due to (63) from Lemma B.3. Combining (80), (81), (82) and (88), we obtain that

$$\begin{aligned}
\prod_{i=0}^{t-1} \hat{p}_i \hat{n}_i &= \prod_{i=0}^{t_0-1} \hat{p}_i \hat{n}_i \prod_{i \in I} \hat{p}_i \hat{n}_i \prod_{i \in \bar{I}} \hat{p}_i \hat{n}_i \\
&\geq \frac{1}{2^{t_0}} e^{-t_0 \log \frac{1}{2\alpha(1-\beta)}} \frac{1}{2^{|\bar{I}|}} e^{-\delta_4 \left(\Psi(\tilde{B}_{t_0}) + 2 \sum_{i=0}^{t-1} D_i\right) \log \frac{1}{2\alpha(1-\beta)}} \\
&\quad \frac{1}{2^{|\bar{I}|}} e^{-\delta_5 \left(\Psi(\tilde{B}_0) + 2 \sum_{i=0}^{t-1} D_i\right) - \frac{2\delta_2 - 1/16 - \log \delta_2}{2\delta_2 - 17/16} \sum_{i=0}^{t-1} D_i} \\
&= \frac{1}{2^t} e^{-\left(t_0 \log \frac{1}{2\alpha(1-\beta)} + (\delta_4 \log \frac{1}{2\alpha(1-\beta)} + \delta_5) \Psi(\tilde{B}_0) + (2\delta_4 \log \frac{1}{2\alpha(1-\beta)} + 2\delta_5 + \frac{2\delta_2 - 1/16 - \log \delta_2}{2\delta_2 - 17/16}) \sum_{i=0}^{t-1} D_i\right)}.
\end{aligned} \tag{89}$$

Leveraging (78), (79), (89) with (49) from Proposition B.6, we prove that

$$\begin{aligned}
\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} &\leq \left[1 - \left(\prod_{i=0}^{t-1} \frac{\hat{p}_i \hat{q}_i \hat{n}_i \cos^2(\hat{\theta}_i)}{\hat{m}_i}\right)^{\frac{1}{t}}\right]^t = \left[1 - \left(\prod_{i=0}^{t-1} \hat{p}_i \hat{n}_i \prod_{i=0}^{t-1} \hat{q}_i \prod_{i=0}^{t-1} \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i}\right)^{\frac{1}{t}}\right]^t \\
&\leq \left(1 - e^{-\frac{t_0 \log \frac{1}{2\alpha(1-\beta)} + (1 + \delta_4 \log \frac{1}{2\alpha(1-\beta)} + \delta_5) \Psi(\tilde{B}_0) + (2 + 2\delta_4 \log \frac{1}{2\alpha(1-\beta)} + 2\delta_5 + \frac{2\delta_2 - 1/16 - \log \delta_2}{2\delta_2 - 17/16}) \sum_{i=0}^{t-1} D_i}{t}}\right)^t \\
&= \left(1 - e^{-\frac{\delta_6 t_0 + \delta_7 \Psi(\tilde{B}_0) + \delta_8 \sum_{i=0}^{t-1} D_i}{t}}\right)^t \leq \left(\frac{\delta_6 t_0 + \delta_7 \Psi(\tilde{B}_0) + \delta_8 \sum_{i=0}^{t-1} D_i}{t}\right)^t,
\end{aligned}$$

where the inequality is due to the fact that $1 - e^{-x} \leq x$ for any $x \in \mathbb{R}$ and $\delta_6, \delta_7, \delta_8$ are defined in (25). Hence, we prove that for any $t > t_0$,

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - e^{-\frac{\delta_6 t_0 + \delta_7 \Psi(\tilde{B}_0) + \delta_8 \sum_{i=0}^{t-1} D_i}{t}}\right)^t \leq \left(\frac{\delta_6 t_0 + \delta_7 \Psi(\tilde{B}_0) + \delta_8 \sum_{i=0}^{t-1} D_i}{t}\right)^t. \tag{90}$$

From (74) in Theorem 3.3, we have that

$$\sum_{i=0}^{t-1} D_i \leq D_0 \left(\Psi(\tilde{B}_0) + \frac{3(1 + D_0)^2}{\alpha(1 - \beta)}\right). \tag{91}$$

Therefore, combining the above inequality with (90), we prove that

$$\begin{aligned}
\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} &\leq \left(\frac{\delta_6 t_0 + \delta_7 \Psi(\tilde{B}_0) + \delta_8 \sum_{i=0}^{t-1} D_i}{t}\right)^t \\
&\leq \left(\frac{\delta_6 t_0 + \delta_7 \Psi(\tilde{B}_0) + \delta_8 D_0 \left(\Psi(\tilde{B}_0) + \frac{3(1 + D_0)^2}{\alpha(1 - \beta)}\right)}{t}\right)^t.
\end{aligned}$$

D Proof of Iteration Complexity

We treat the line search parameters α and β as absolute constants. The first linear rate from Theorem 3.1 leads to the global complexity of

$$\mathcal{O}(\Psi(\bar{B}_0) + (1 + D_0)^2 \log \frac{1}{\epsilon}) \quad (92)$$

The second linear rate from Theorem 3.3 leads to the global complexity of

$$\mathcal{O}(\Psi(\tilde{B}_0) + (\Psi(\bar{B}_0) + (1 + D_0)^2 D_0 + \log \frac{1}{\epsilon})) \quad (93)$$

where the first term is the number of iterations required to reach the linear rate in (18). For the analysis of the superlinear convergence rate, we denote that

$$\Omega = \Psi(\tilde{B}_0) + (\Psi(\bar{B}_0) + (1 + D_0)^2 D_0)$$

From Theorem 4.3, we have that

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(\frac{\Omega}{t}\right)^t$$

Let T_* be the number such that the inequality $(\frac{\Omega}{t})^t \leq \epsilon$ above becomes equality. we have

$$\log \frac{1}{\epsilon} = T_* \log \frac{T_*}{\Omega} \leq T_* \left(\frac{T_*}{\Omega} - 1\right),$$

which leads to

$$T_* \geq \frac{\Omega + \sqrt{\Omega^2 + 4\Omega \log \frac{1}{\epsilon}}}{2}.$$

Hence, we have that

$$\log \frac{1}{\epsilon} = T_* \log \frac{T_*}{\Omega} \geq T_* \log \frac{\Omega + \sqrt{\Omega^2 + 4\Omega \log \frac{1}{\epsilon}}}{2\Omega} \geq T_* \log \left(\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\log \frac{1}{\epsilon}}{\Omega}} \right),$$

which implies that

$$T_* \leq \frac{\log \frac{1}{\epsilon}}{\log \left(\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{\log \frac{1}{\epsilon}}{\Omega}} \right)}.$$

Hence, to reach the accuracy of ϵ , we need the number of iterations t to be at least

$$\mathcal{O}\left(\frac{\log \frac{1}{\epsilon}}{\log \left(\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{\Omega} \log \frac{1}{\epsilon}} \right)}\right). \quad (94)$$

Therefore, we prove the iteration complexity by choosing the minimal from (92), (93), and (94). For the special case of $B_0 = aI$ for $a > 0$, just replace $\Psi(\bar{B}_0)$ and $\Psi(\tilde{B}_0)$ by Δ_1 and Δ_2 defined in (16), (20), respectively.

E Proof of Line Search Complexity

Proposition E.1. *Suppose that Assumption 2.2 holds. Consider the BFGS method with inexact line search defined in (5) and (6) and we choose the step size η_t according to Algorithm 1. At iteration t , denote λ_t as the number of loops in Algorithm 1 to terminate and return the η_t satisfying the Wolfe conditions (5) and (6). Then λ_t is finite and upper bounded by*

$$\begin{aligned} \lambda_t \leq & 2 + \log_2 \left(1 + \frac{(1 - \beta)(1 + 2D_t)}{\beta - \alpha} \right) \\ & + 2 \log_2 \left(1 + \log_2 (2(1 - \alpha)(1 + D_t)) + \max\{\log_2 \rho_t, \log_2 \frac{1}{\rho_t}\} \right). \end{aligned} \quad (95)$$

Algorithm 1 Log Bisection Algorithm for Weak Wolfe Conditions

Require: Initial step size $\eta^{(0)} = 1$, $\eta_{min}^{(0)} = 0$, $\eta_{max}^{(0)} = +\infty$

```

1: for  $i = 0, 1, 2, \dots$  do
2:   if  $f(x_t + \eta^{(i)} d_t) > f(x_t) + \alpha \eta^{(i)} \nabla f(x_t)^\top d_t$  then
3:     Set  $\eta_{max}^{(i+1)} = \eta^{(i)}$  and  $\eta_{min}^{(i+1)} = \eta_{min}^{(i)}$ 
4:     if  $\eta_{min}^{(i)} = 0$  then
5:        $\eta^{(i+1)} = (\frac{1}{2})^{2^{i+1}-1}$ 
6:     else
7:        $\eta^{(i+1)} = \sqrt{\eta_{max}^{(i+1)} \eta_{min}^{(i+1)}}$ 
8:     end if
9:   else if  $\nabla f(x_t + \eta^{(i)} d_t)^\top d_t < \beta \nabla f(x_t)^\top d_t$  then
10:    Set  $\eta_{max}^{(i+1)} = \eta_{max}^{(i)}$  and  $\eta_{min}^{(i+1)} = \eta^{(i)}$ 
11:    if  $\eta_{max}^{(i)} = +\infty$  then
12:       $\eta^{(i+1)} = 2^{2^{i+1}-1}$ 
13:    else
14:       $\eta^{(i+1)} = \sqrt{\eta_{max}^{(i+1)} \eta_{min}^{(i+1)}}$ 
15:    end if
16:   else
17:     Return  $\eta^{(i)}$ 
18:   end if
19: end for

```

Proof. Please check Proposition K.2 in [42]. The only difference is C_t is replaced by D_t defined in (11). \square

We can prove the line search complexity in Proposition 5.1 using result from Proposition E.1. We have that

$$\begin{aligned}
\Lambda_t &= \frac{1}{t} \sum_{i=0}^{t-1} \lambda_i \leq 2 + \frac{1}{t} \sum_{i=0}^{t-1} \log_2 \left(1 + \frac{(1-\beta)(1+2D_i)}{\beta-\alpha} \right) \\
&\quad + \frac{2}{t} \sum_{i=0}^{t-1} \log_2 \left(1 + \log_2 (2(1-\alpha)(1+D_i)) + \max\{\log_2 \rho_i, \log_2 \frac{1}{\rho_i}\} \right).
\end{aligned} \tag{96}$$

Using Jensen's inequality, we have that

$$\frac{1}{t} \sum_{i=0}^{t-1} \log_2 \left(1 + \frac{(1-\beta)(1+2D_i)}{\beta-\alpha} \right) \leq \log_2 \left(1 + \frac{1-\beta}{\beta-\alpha} + \frac{2(1-\beta)}{\beta-\alpha} \frac{\sum_{i=0}^{t-1} D_i}{t} \right). \tag{97}$$

$$\begin{aligned}
&\frac{1}{t} \sum_{i=0}^{t-1} \log_2 \left(1 + \log_2 (2(1-\alpha)(1+D_i)) + \max\{\log_2 \rho_i, \log_2 \frac{1}{\rho_i}\} \right) \\
&\leq \log_2 \left(1 + \log_2 2(1-\alpha) + \frac{1}{t} \sum_{i=0}^{t-1} \log_2 (1+D_i) + \frac{1}{t} \sum_{i=0}^{t-1} \max\{\log_2 \rho_i, \log_2 \frac{1}{\rho_i}\} \right) \\
&\leq \log_2 \left(1 + \log_2 2(1-\alpha) + \log_2 \left(1 + \frac{\sum_{i=0}^{t-1} D_i}{t} \right) + \frac{1}{t} \sum_{i=0}^{t-1} \max\{\log_2 \rho_i, \log_2 \frac{1}{\rho_i}\} \right).
\end{aligned} \tag{98}$$

We also have that

$$\begin{aligned}
& \frac{1}{t} \sum_{i=0}^{t-1} \max\{\log_2 \rho_i, \log_2 \frac{1}{\rho_i}\} = \frac{1}{t} \sum_{i=0, \rho_i \geq 1}^{t-1} \log_2 \rho_i + \frac{1}{t} \sum_{i=0, 0 \leq \rho_i < 1}^{t-1} \log_2 \frac{1}{\rho_i} \\
&= \frac{1}{t} \sum_{i=0, \rho_i \geq 2}^{t-1} \log_2 \rho_i + \frac{1}{t} \sum_{i=0, 1 \leq \rho_i < 2}^{t-1} \log_2 \rho_i + \frac{1}{t} \sum_{i=0, \frac{1}{2} < \rho_i < 1}^{t-1} \log_2 \frac{1}{\rho_i} + \frac{1}{t} \sum_{i=0, \rho_i \leq \frac{1}{2}}^{t-1} \log_2 \frac{1}{\rho_i} \quad (99) \\
&\leq 2 + \frac{1}{t} \sum_{i=0, \rho_i \geq 2}^{t-1} \log_2 \rho_i + \frac{1}{t} \sum_{i=0, \rho_i \leq \frac{1}{2}}^{t-1} \log_2 \frac{1}{\rho_i},
\end{aligned}$$

where the inequality is due to $\log_2 \rho_i \leq 1$ for $\rho_i < 2$ and $\log_2 \frac{1}{\rho_i} \leq 1$ for $\rho_i > \frac{1}{2}$. Using the definition of ω and (b) in Lemma B.3, we obtain that

$$\begin{aligned}
& \frac{1}{t} \sum_{i=0, \rho_i \geq 2}^{t-1} \log_2 \rho_i = \frac{\log_2 e}{t} \sum_{i=0, \rho_i \geq 2}^{t-1} \log \rho_i = \frac{\log_2 e}{t} \sum_{i=0, \rho_i \geq 2}^{t-1} (\rho_i - 1 - \omega(\rho_i - 1)) \\
&\leq \frac{\log_2 e}{t} \sum_{i=0, \rho_i \geq 2}^{t-1} \left(\frac{2\rho_i}{\rho_i - 1} \omega(\rho_i - 1) - \omega(\rho_i - 1) \right) = \frac{\log_2 e}{t} \sum_{i=0, \rho_i \geq 2}^{t-1} \frac{\rho_i + 1}{\rho_i - 1} \omega(\rho_i - 1) \quad (100) \\
&\leq \frac{3 \log_2 e}{t} \sum_{i=0, \rho_i \geq 2}^{t-1} \omega(\rho_i - 1).
\end{aligned}$$

Similarly, using (c) in Lemma B.3, we obtain that

$$\begin{aligned}
& \frac{1}{t} \sum_{i=0, \rho_i \leq \frac{1}{2}}^{t-1} \log_2 \frac{1}{\rho_i} = \frac{\log_2 e}{t} \sum_{i=0, \rho_i \leq \frac{1}{2}}^{t-1} \log \frac{1}{\rho_i} = \frac{\log_2 e}{t} \sum_{i=0, \rho_i \leq \frac{1}{2}}^{t-1} (\omega(\rho_i - 1) + 1 - \rho_i) \\
&\leq \frac{\log_2 e}{t} \sum_{i=0, \rho_i \leq \frac{1}{2}}^{t-1} (\omega(\rho_i - 1) + \frac{1 + \rho_i}{1 - \rho_i} \omega(\rho_i - 1)) = \frac{\log_2 e}{t} \sum_{i=0, \rho_i \leq \frac{1}{2}}^{t-1} \frac{2}{1 - \rho_i} \omega(\rho_i - 1) \quad (101) \\
&\leq \frac{4 \log_2 e}{t} \sum_{i=0, \rho_i \leq \frac{1}{2}}^{t-1} \omega(\rho_i - 1).
\end{aligned}$$

Combining (99), (100) and (101), we prove that

$$\begin{aligned}
& \frac{1}{t} \sum_{i=0}^{t-1} \max\{\log_2 \rho_i, \log_2 \frac{1}{\rho_i}\} \leq 2 + \frac{1}{t} \sum_{i=0, \rho_i \geq 2}^{t-1} \log_2 \rho_i + \frac{1}{t} \sum_{i=0, \rho_i \leq \frac{1}{2}}^{t-1} \log_2 \frac{1}{\rho_i} \\
&\leq 2 + \frac{4 \log_2 e}{t} \sum_{i=0}^{t-1} \omega(\rho_i - 1) \leq 2 + \frac{6}{t} \left(\Psi(\tilde{B}_0) + 2 \sum_{i=0}^{t-1} D_i \right). \quad (102)
\end{aligned}$$

where we use the fact that $\omega(\rho_i - 1) \geq 0$ for any $i \geq 0$ and the last inequality is due to (63) in Proposition B.11. Leveraging (96), (97), (98) and (102), we have that

$$\begin{aligned}
\Lambda_t &\leq 2 + \log_2 \left(1 + \frac{1-\beta}{\beta-\alpha} + \frac{2(1-\beta)}{\beta-\alpha} \frac{\sum_{i=0}^{t-1} D_i}{t} \right) \\
&\quad + 2 \log_2 \left(3 + \log_2 2(1-\alpha) + \log_2 \left(1 + \frac{\sum_{i=0}^{t-1} D_i}{t} \right) + \frac{6}{t} (\Psi(\tilde{B}_0) + 2 \sum_{i=0}^{t-1} D_i) \right) \\
&\leq 2 + \log_2 \left(1 + \frac{1-\beta}{\beta-\alpha} + \frac{2(1-\beta)}{\beta-\alpha} \frac{\sum_{i=0}^{t-1} D_i}{t} \right) \\
&\quad + 2 \log_2 \left(\log_2 16(1-\alpha) + \log_2 \left(1 + \frac{\sum_{i=0}^{t-1} D_i}{t} \right) + \frac{6\Psi(\tilde{B}_0) + 12 \sum_{i=0}^{t-1} D_i}{t} \right) \\
&\leq 2 + \log_2 \left(1 + \frac{1-\beta}{\beta-\alpha} + \frac{2(1-\beta)}{\beta-\alpha} \frac{\sum_{i=0}^{t-1} D_i}{t} \right) \\
&\quad + 2 \log_2 \left(\log_2 16(1-\alpha) + \log_2 \left(1 + \frac{6\Psi(\tilde{B}_0) + 14 \sum_{i=0}^{t-1} D_i}{t} \right) \right).
\end{aligned}$$

Using (74) from the proof of Theorem 3.3, i.e.,

$$\sum_{i=0}^{t-1} D_i \leq D_0 \left(\Psi(\bar{B}_0) + \frac{3(1+D_0)^2}{\alpha(1-\beta)} \right).$$

We prove the line search complexity as

$$\Lambda_t = \mathcal{O} \left(\log \left(1 + \frac{\Gamma}{t} \right) + \log \log \left(1 + \frac{\Psi(\tilde{B}_0) + \Gamma}{t} \right) \right)$$

where

$$\Gamma = \mathcal{O} (D_0(\Psi(\bar{B}_0) + (1+D_0)^2))$$

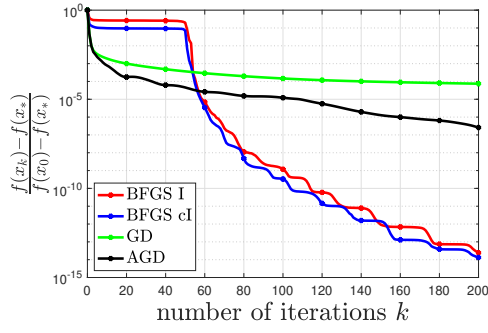
For the special case of $B_0 = aI$ for $a > 0$, just replace $\Psi(\bar{B}_0)$ and $\Psi(\tilde{B}_0)$ by Δ_1 and Δ_2 defined in (16), (20), respectively.

F Proof of Strong Self-Concordance

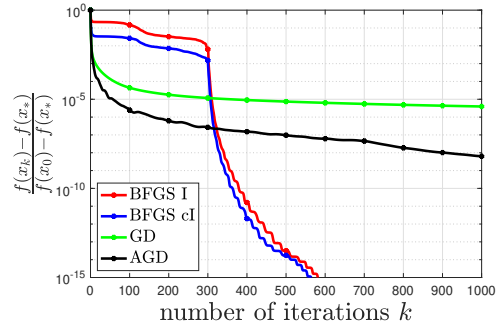
Consider the log-sum-exp function defined as $f(x) = \log(\sum_{i=1}^n \exp(c_i^\top x - b_i)) + \frac{1}{2} \sum_{i=1}^n (c_i^\top x)^2$, we have that $\nabla f(x) = \sum_{i=1}^n \pi_i c_i + \sum_{i=1}^n (c_i^\top x) c_i$ where $\pi_i = \frac{\exp(c_i^\top x - b_i)}{\sum_{j=1}^n \exp(c_j^\top x - b_j)}$ and $\nabla^2 f(x) = \sum_{i=1}^n (\pi_i + 1) c_i c_i^\top - (\sum_{i=1}^n \pi_i c_i)(\sum_{i=1}^n \pi_i c_i)^\top$. From proof in [48], this log-sum-exp function is strictly convex. Moreover, we also need to prove that this function is strongly self-concordant. Notice that, with respect to the operator $B = \sum_{i=1}^n c_i c_i^\top$, this function f is strongly convex with parameter 1 and its Hessian is Lipschitz continuous with parameter 2 (check example 1 of [49]¹). Hence, using results from Example 4.1 in [17], the log-sum-exp function is strongly self-concordant.

The proof of that the logistic regression function $f(x) = \frac{1}{N} \sum_{i=1}^N \ln(1 + e^{-y_i z_i^\top x})$ without l_2 regularization is strongly self-concordant is almost the same. It has the similar structure with the log-sum-exp function $f(x) = \log(\sum_{i=1}^n e^{c_i^\top x - b_i}) + \frac{1}{2} \sum_{i=1}^n (c_i^\top x)^2$. Notice that in our BFGS method, we use the line search scheme such that we always have $f(x_t) \leq f(x_0)$ for any $t \geq 0$. Hence, the iterations generated by BFGS method with weak-Wolfe line search conditions always stay in the bounded set $\{x | f(x) \leq f(x_0)\}$ where x_0 is the initial point. In this bounded set, the logistic regression function is strongly convex and its Hessian is smooth with respect to the operator matrix $B = \sum_{i=1}^n z_i z_i^\top$. According to the Example 4.1 from that greedy quasi-Newton paper, if a function is strongly convex and its Hessian is smooth with respect to some matrix B , then the function is strongly self-concordant. Hence, the logistic regression function is strongly self-concordant. Similarly, for the hard cubic function, we can show that it is strongly convex and its Hessian is Lipschitz smooth with respect to some operator matrix B and therefore the hard cubic function is also strongly self-concordant.

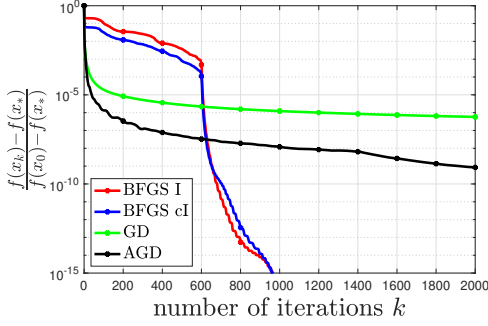
¹N. Doikov and Y. Nesterov. Minimizing uniformly convex functions by cubic regularization of newton method. arXiv, 1905.02671, 2019



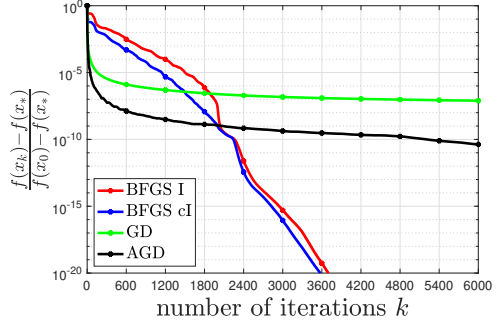
(a) $d = 50$.



(b) $d = 300$.



(c) $d = 600$.

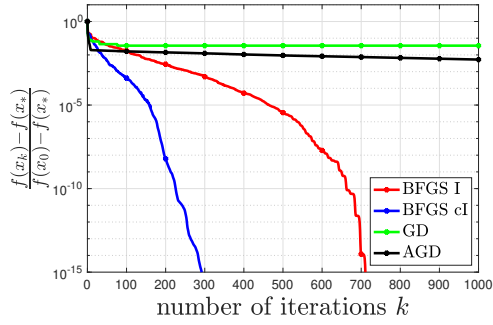


(d) $d = 2000$.

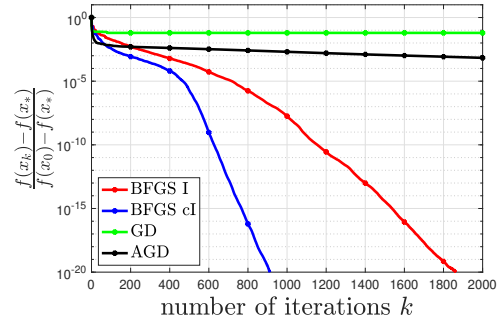
Figure 5: Convergence rates of BFGS with different B_0 , gradient descent and accelerated gradient descent for solving the hard cubic function with different dimensions.

G Additional Numerical Experiments

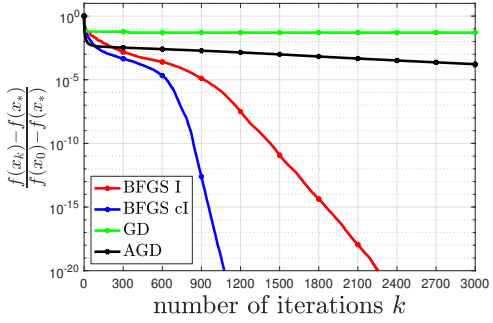
Additional numerical experiments on the hard cubic function and the logistic regression for different dimensions are presented in figures 5 and 6. The empirical results of the performance of different optimization methods for the hard cubic function with respect to the number of gradient evaluations and the time in seconds are in Figures 7 and 8. Additional numerical results of the values of the step sizes of BFGS method are in Figure 9. Additional results of the performance of different optimization methods with transformation matrix are in Figure 10. The convergence performance of BFGS method is similar to the empirical results from Figures 1, 2, 3, and 4 in section 6.



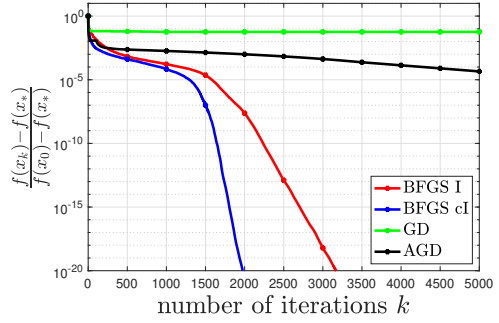
(a) $d = 50$.



(b) $d = 300$.

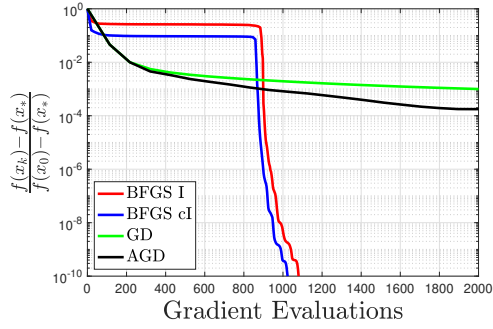


(c) $d = 600$.

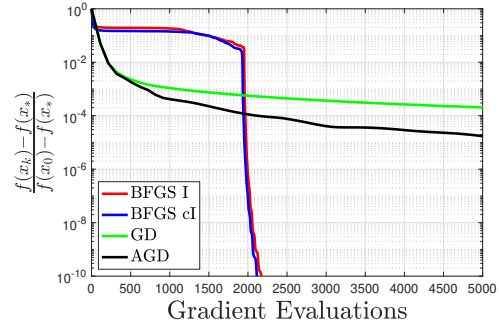


(d) $d = 2000$.

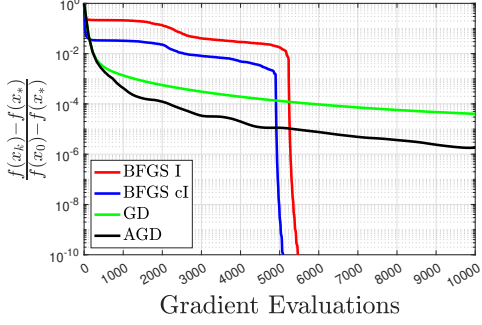
Figure 6: Convergence rates of BFGS with different B_0 , gradient descent and accelerated gradient descent for solving the logistic regression function with different dimensions.



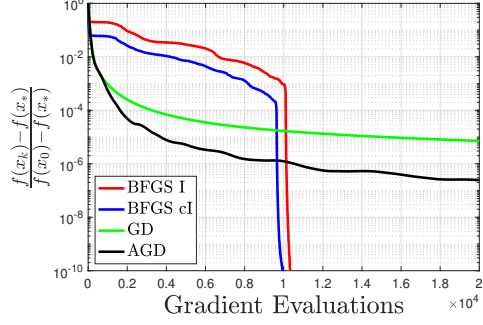
(a) $d = 100$.



(b) $d = 200$.

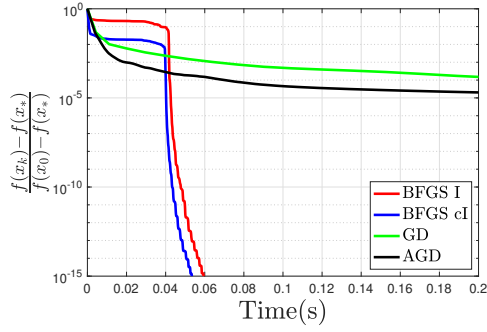


(c) $d = 500$.

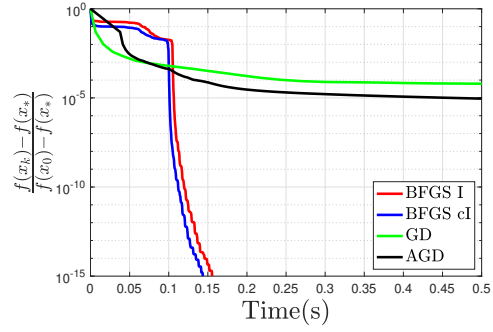


(d) $d = 1000$.

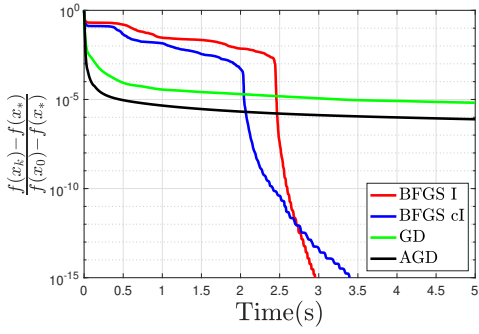
Figure 7: Convergence rates of BFGS with different B_0 , gradient descent and accelerated gradient descent for solving the hard cubic function with respect to the number of gradient evaluations.



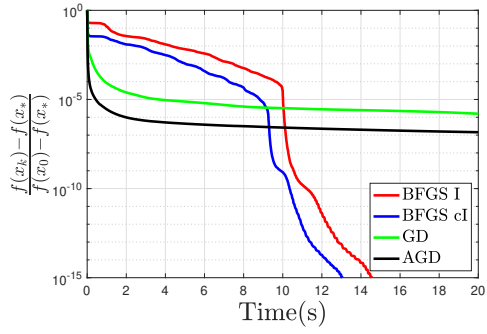
(a) $d = 100$.



(b) $d = 200$.

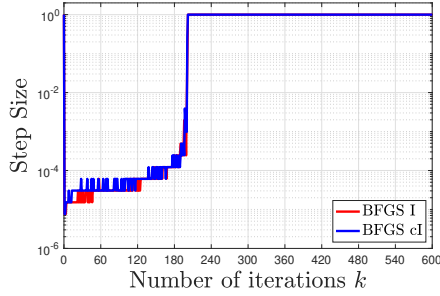


(c) $d = 500$.

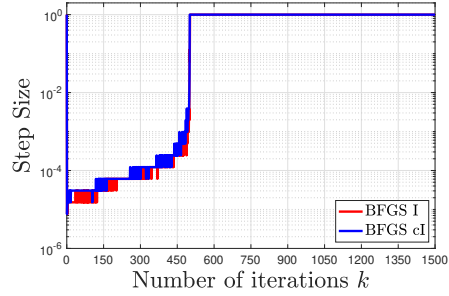


(d) $d = 1000$.

Figure 8: Convergence rates of BFGS with different B_0 , gradient descent and accelerated gradient descent for solving the hard cubic function with respect to the time in seconds.



(a) $d = 100$.



(b) $d = 1000$.

Figure 9: Step size of BFGS with different B_0 using inexact line search for solving the hard cubic function with different dimensions.

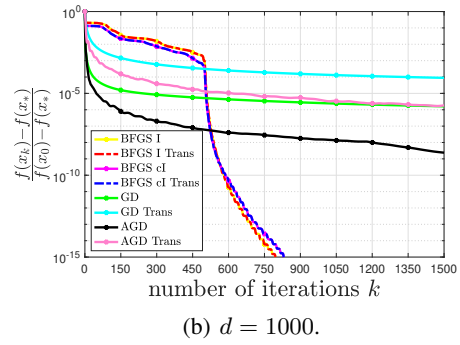
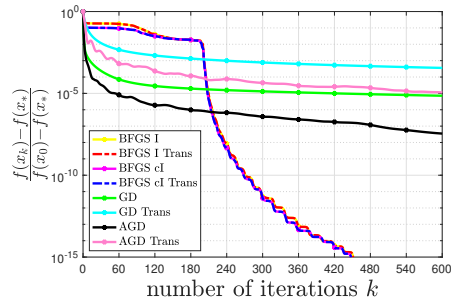


Figure 10: Convergence rates of BFGS with different B_0 , gradient descent and accelerated gradient descent for solving the hard cubic function with transformation matrix A .