Appendix A. Some Important Lemmas

In this section, we give several important lemmas which will be used in the proof of the theorems of this paper.

Lemma 9 *If A and B are $d \times d$ symmetric positive matrices, and $(1-\epsilon_0)B \preceq A \preceq (1+\epsilon_0)B$ where $0 < \epsilon_0 < 1$, then we have*

$$\|A^{1/2}B^{-1}A^{1/2} - I\| \leq \epsilon_0,$$

where I is the identity matrix.

Proof Because $A \preceq (1+\epsilon_0)B$, we have $z^T[A - (1+\epsilon_0)B]z \leq 0$ for any nonzero $z \in \mathbb{R}^d$. This implies $\frac{z^TAz}{z^TBz} \leq 1 + \epsilon_0$ for any $z \neq 0$. Subsequently,

$$\begin{aligned}\lambda_{\max}(B^{-1}A) =& \lambda_{\max}(B^{-1/2}AB^{-1/2}) \\ =& \max_{u \neq 0} \frac{u^TB^{-1/2}AB^{-1/2}u}{u^Tu} \\ =& \max_{z \neq 0} \frac{z^TAz}{z^TBz} \\ \leq& 1 + \epsilon_0,\end{aligned}$$

where the last equality is obtained by setting $z = B^{-1/2}u$. Similarly, we have $\lambda_{\min}(B^{-1}A) \geq 1 - \epsilon_0$. Since $B^{-1}A$ and $A^{1/2}B^{-1}A^{1/2}$ are similar, the eigenvalues of $A^{1/2}B^{-1}A^{1/2}$ are all between $1 - \epsilon_0$ and $1 + \epsilon_0$. Therefore, we have

$$\|A^{1/2}B^{-1}A^{1/2} - I\| \leq \epsilon_0.$$

■

Lemma 10 ([3]) *Let $X_1, X_2, \ldots, X_k$ be independent, random, symmetric, real matrices of size $d \times d$ with $0 \preceq X_i \preceq LI$, where I is the $d \times d$ identity matrix. Let $Y = \sum_{i=1}^k X_i$, $\mu_{\min} = \lambda_{\min}(\mathbb{E}[Y])$ and $\mu_{\max} = \lambda_{\max}(\mathbb{E}[Y])$. Then,*

$$\mathbb{P}\left(\lambda_{\min}(Y) \leq (1-\epsilon)\mu_{\min}\right) \leq d \cdot e^{-\epsilon^2\mu_{\min}/L}$$

Lemma 11 ([3]) *Given a matrix $A \in \mathbb{R}^{m \times n}$, construct an $m \times n$ random matrix R such that*

$$\mathbb{E}[R] = A \quad \textit{and} \quad \|R\| \leq L.$$

Compute the per-sample second moment:

$$M = \max\{\|\mathbb{E}[RR^T]\|, \mathbb{E}[R^TR]\|\}.$$

Form the matrix sampling estimator

$$\bar{R} = \frac{1}{s}\sum_{i=1}^{s} R_i, \textit{ where each } R_i \textit{ is an independent copy of } R.$$

Then, for all $t \geq 0$

$$\mathbb{P}\left[\|\bar{R} - A\| \geq t\right] \leq (m+n)\exp\left(\frac{-st^2/2}{M + 2Lt/3}\right).$$

Lemma 12 *Assume* (9) *and* (10) *hold. Let* $0 < \delta < 1$, $0 < \epsilon < 1$ *and* $0 < c$ *be given. If we sample* f_i*'s uniformly with the sample size* $|\mathcal{S}|$ *and construct* $H^{(t)} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)})$, *then we have the following results:*
(a) If $|\mathcal{S}| \geq \frac{16K^2\log(2d/\delta)}{c^2\epsilon^2}$, *it holds that*

$$\|H^{(t)} - \nabla^2 F(x^{(t)})\| \leq \epsilon c.$$

(b) If $|\mathcal{S}| \geq \frac{K\log(2d/\delta)}{\sigma\epsilon^2}$, *it holds that*

$$\lambda_{\min}(H^{(t)}) \geq (1-\epsilon)\sigma.$$

Proof Consider $|\mathcal{S}|$ i.i.d random matrces $H_j^{(t)}, j = 1, \ldots, |\mathcal{S}|$ such that $\mathbb{P}\left(H_j^{(t)} = \nabla^2 f_i(x^{(t)})\right) = 1/n$ for all $i = 1, \ldots, n$. Then, we have $\mathbb{E}(H_j^{(t)}) = \nabla^2 F(x^{(t)})$ for all $j = 1, \ldots, |\mathcal{S}|$. By (9) and the positive semi-definite property of $H_j^{(t)}$, we have $\lambda_{\max}(H_j^{(t)}) \leq K$ and $\lambda_{\min}(H_j^{(t)}) \geq 0$. By Lemma 10, we have that if $|\mathcal{S}| \geq \frac{K\log(d/\delta)}{\sigma\epsilon^2}$, $\lambda_{\min}(H^{(t)}) \geq (1-\epsilon)\sigma$ holds with probability at least $1-\delta$.

We define random maxtrices $X_j = H_j^{(t)} - \nabla^2 F(x^{(t)})$ for all $j = 1, \ldots, |\mathcal{S}|$. We have $\mathbb{E}[X_j] = 0$, $\|X_j\| \leq 2K$ and $\|X_j\|^2 \leq 4K^2$. By Lemma 11, we have

$$\mathbb{P}(\|H^{(t)} - \nabla^2 F(x^{(t)})\| \geq \epsilon c) \leq 2d\exp^{-\frac{c^2\epsilon^2|\mathcal{S}|}{16K^2}}.$$

When $|\mathcal{S}| \geq \frac{16K^2\log(2d/\delta)}{c^2\epsilon^2}$, $\|H^{(t)} - \nabla^2 F(x^{(t)})\| \leq \epsilon c$ holds with probability at least $1-\delta$. ■

Appendix B. Proofs of theorems of Section 3

Proof of Theorem 3 By Assumption 1 and 2, we have that $F(x)$ is μ-strongly convex and $\nabla F(x)$ is L-Lipschitz continuous. Hence, we have

$$\mu \leq \lambda_{\min}(\nabla^2 F(x)) \leq \lambda_{\max}(\nabla^2 F(x)) \leq L.$$

Hence, for any x in domain, it holds that

$$\kappa = \frac{L}{\mu} \geq \kappa(\nabla^2 F(x)).$$

By Taylor's theorem, we obtain

$$
\begin{aligned}
&\nabla F(x^{(t+1)})\\
=&\nabla F(x^{(t)}) + \nabla^2 F(x^{(t)})(-p^{(t)}) + \int_0^1 [\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})](-p^{(t)})ds\\
=&\nabla F(x^{(t)}) - \nabla^2 F(x^{(t)})[H^{(t)}]^{-1}\nabla F(x^{(t)}) + \nabla^2 F(x^{(t)})[H^{(t)}]^{-1}\nabla F(x^{(t)}) - \nabla^2 F(x^{(t)})p^{(t)}\\
&+ \int_0^1 [\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})](-p^{(t)})ds\\
=&\left[\nabla^2 F(x^{(t)})\right]^{\frac{1}{2}} \left(I - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\right) \left[\nabla^2 F(x^{(t)})\right]^{-\frac{1}{2}} \nabla F(x^{(t)})\\
&+ \nabla^2 F(x^{(t)})([H^{(t)}]^{-1}\nabla F(x^{(t)}) - p^{(t)}) + \int_0^1 [\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})](-p^{(t)})ds.
\end{aligned}
$$

Hence, we have the following identity

$$
\begin{aligned}
\left[\nabla^2 F(x^{(t)})\right]^{-\frac{1}{2}} \nabla F(x^{(t+1)}) =& \left(I - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\right) \left[\nabla^2 F(x^{(t)})\right]^{-\frac{1}{2}} \nabla F(x^{(t)})\\
&+ [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}([H^{(t)}]^{-1}\nabla F(x^{(t)}) - p^{(t)})\\
&+ [\nabla^2 F(x^{(t)})]^{-\frac{1}{2}} \int_0^1 [\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})](-p^{(t)})ds.
\end{aligned}
$$

For notational simplicity, we denote $M = \left[\nabla^2 F(x^{(t)})\right]^{-1}$ and $M^* = \left[\nabla^2 F(x^*)\right]^{-1}$. Then we can obtain

$$
\begin{aligned}
\|\nabla F(x^{(t+1)})\|_M \le& \left\|I - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\right\| \|\nabla F(x^{(t)})\|_M\\
&+ \|[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\|\|[H^{(t)}]^{-1}\nabla F(x^{(t)}) - p^{(t)}\|\\
&+ \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 \|\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})\|\|p^{(t)}\|ds.
\end{aligned}
$$

We bound the three terms on the right-hand side of the above equation respectively.

For the first term, using Lemma 9, we have

$$
\left\|I - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\right\| \|\nabla F(x^{(t)})\|_M \le \epsilon_0 \|\nabla F(x^{(t)})\|_M.
$$

For the second term, by the fact that $\|AB\| \ge \|A\|\sigma_{\min}(B)$ and condition

$$
\|\nabla F(x^{(t)}) - H^{(t)}p^{(t)}\| \le \frac{\epsilon_1}{\kappa}\|\nabla F(x^{(t)})\| \le \frac{\epsilon_1}{\kappa(\nabla^2 F(x^{(t)}))}\|\nabla F(x^{(t)})\|,
$$

we obtain

$$\begin{aligned}
&\|[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\|\|[H^{(t)}]^{-1}\nabla F(x^{(t)}) - p^{(t)}\|\\
=&\frac{\|[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\|}{\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})}\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})\|[H^{(t)}]^{-1}\|\|\nabla F(x^{(t)}) - H^{(t)}p^{(t)}\|\\
\le&\frac{\epsilon_1}{\kappa(\nabla^2 F(x^{(t)}))}\frac{\|[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\|}{\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})}\|[H^{(t)}]^{-1}\|(\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})\|\nabla F(x^{(t)})\|)\\
\le&\frac{\epsilon_1}{\kappa(\nabla^2 F(x^{(t)}))}\|\nabla^2 F(x^{(t)})\|\|[H^{(t)}]^{-1}\|\|\nabla F(x^{(t)})\|_M\\
\le&\frac{\epsilon_1}{1-\epsilon_0}\|\nabla F(x^{(t)})\|_M.
\end{aligned}$$

For the third term, we bound it for the case that $\nabla^2 F(x)$ is not Lipschitz continuous and the case $\nabla^2 F(x)$ is Lipschitz continuous respectively.

First, we consider the case that $\nabla^2 F(x)$ is not Lipschitz continuous but is continuous close to the optimal point x^*. Because $\nabla^2 F(x)$ is continuous near x^*, there exists a sufficient small value γ such that it holds that

$$\|[\nabla^2 F(x^*)]^{-1} - [\nabla^2 F(x^{(t)})]^{-1}\| < \frac{\nu(t)}{L}, \tag{14}$$

and

$$\|\nabla^2 F(x^*) - \nabla^2 F(x^{(t)})\| < \frac{\eta(t)\mu}{\sqrt{\kappa}}, \tag{15}$$

when $\|x^{(t)} - x^*\| \le \gamma$. Therefore, $\nu(t)$ and $\eta(t)$ will go to 0 as $x^{(t)}$ goes to x^*.

By μ-strong convexity, we have $\|[\nabla^2 F(x^t)]^{-1}\| \le \frac{1}{\mu}$ for all $x^{(t)}$ sufficiently close to x^*. Because of Eqn. (2), we have

$$\|[H^{(t)}]^{-1}\| \le (1+\epsilon_0)\|[\nabla^2 F(x^t)]^{-1}\| \le \frac{1}{(1-\epsilon_0)\mu}.$$

We define $r^{(t)} = \nabla F(x^{(t)}) - H^{(t)}p^{(t)}$. Then we have that the direction vector satisfies

$$\|p^{(t)}\| = \|[H^{(t)}]^{-1}\|(\|r^{(t)}\| + \|\nabla F(x^{(t)})\|) \le \frac{2}{(1-\epsilon_0)\mu}\|\nabla F(x^{(t)})\|, \tag{16}$$

where the second inequality is because

$$\|r^{(t)}\| = \|\nabla F(x^{(t)}) - H^{(t)}p^{(t)}\| \le \frac{\epsilon_1}{\kappa}\|\nabla F(x^{(t)})\| \le \|\nabla F(x^{(t)})\|.$$

Hence, with $\|x - x^*\| \le \gamma$, combining condition (15), we have

$$\begin{aligned}
&\|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 \|\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})\| \|p^{(t)}\| ds \\
\le& \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 \frac{\mu\eta(t)}{\sqrt{\kappa}} \|p^{(t)}\| ds \\
\le& \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \frac{2}{(1-\epsilon_0)\mu} \frac{\mu\eta(t)}{\sqrt{\kappa}} \|\nabla F(x^{(t)})\| \\
\le& \frac{2\eta(t)}{1-\epsilon_0} \frac{\|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\|}{\sqrt{\kappa}\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})} \lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}) \|\nabla F(x^{(t)})\| \\
\le& \frac{2\eta(t)}{1-\epsilon_0} \|\nabla F(x^{(t)})\|_M.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|\nabla F(x^{(t+1)})\|_M \le& \epsilon_0 \|\nabla F(x^{(t)})\|_M + \frac{\epsilon_1}{1-\epsilon_0} \|\nabla F(x^{(t)})\|_M + \frac{2\eta(t)}{1-\epsilon_0} \|\nabla F(x^{(t)})\|_M \\
=& \left(\epsilon_0 + \frac{\epsilon_1}{1-\epsilon_0} + \frac{2\eta(t)}{1-\epsilon_0}\right) \|\nabla F(x^{(t)})\|_M.
\end{aligned}$$

Now, we show the relationship between $\|\cdot\|_M$ and $\|\cdot\|_{M^*}$. By Eqn. (14), we have

$$-\frac{\nu(t)}{\lambda_{\max}(\nabla^2 F(x^*))} u^T u \le u^T([\nabla^2 F(x^*)]^{-1} - [\nabla^2 F(x^{(t)})]^{-1}) u \le \frac{\nu(t)}{\lambda_{\max}(\nabla^2 F(x^*))} u^T u,$$

for any nonzero $u \in \mathbb{R}^d$, which implies that

$$(1 - \nu(t)) u^T [\nabla^2 F(x^{(t)})]^{-1} u \le u^T [\nabla^2 F(x^*)]^{-1} u \le (1 + \nu(t)) u^T [\nabla^2 F(x^{(t)})]^{-1} u.$$

That is,

$$(1 - \nu(t)) \|u\|_M \le \|u\|_{M^*} \le (1 + \nu(t)) \|u\|_M.$$

By this relationship between $\|\cdot\|_M$ and $\|\cdot\|_{M^*}$, we get

$$\|\nabla F(x^{(t+1)})\|_{M^*} \le \left(\epsilon_0 + \frac{\epsilon_1}{1-\epsilon_0} + \frac{2\eta(t)}{1-\epsilon_0}\right) \frac{1+\nu(t)}{1-\nu(t)} \|\nabla F(x^{(t)})\|_{M^*}$$

Second, we consider the case that $\nabla^2 F(x)$ is Lipschitz continuous with parameter $\hat{L}$. We have that the direction vector satisfies

$$\|p^{(t)}\| \le \frac{2}{(1-\epsilon_0)\lambda_{\min}(\nabla^2 F(x^{(t)}))} \|\nabla F(x^{(t)})\|.$$

Because $\nabla^2 F(x)$ is Lipschitz continuous with parameter $\hat{L}$, we have

$$\begin{aligned}
&\|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 \|\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})\| \|p^{(t)}\| ds \\
&\leq \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 s\hat{L}\|p^{(t)}\|^2 ds \\
&= \frac{\hat{L}}{2}\|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\|\lambda_{\min}^{-2}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})\lambda_{\min}^{2}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})\|p^{(t)}\|^2 \\
&\leq \frac{\hat{L}}{2}\|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\|\lambda_{\min}^{-2}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}) \left(\frac{2}{(1-\epsilon_0)\lambda_{\min}(\nabla^2 F(x^{(t)}))}\right)^2 \|\nabla F(x^{(t)})\|_M^2 \\
&= \frac{2\hat{L}\lambda_{\max}(\nabla^2 F(x^{(t)}))}{(1-\epsilon_0)^2\lambda_{\min}^2(\nabla^2 F(x^{(t)}))\sqrt{\lambda_{\min}(\nabla^2 F(x^{(t)}))}}\|\nabla F(x^{(t)})\|_M^2 \\
&\leq \frac{2}{(1-\epsilon_0)^2} \cdot \frac{\hat{L}\kappa}{\mu\sqrt{\mu}}\|\nabla F(x^{(t)})\|_M^2.
\end{aligned}$$

Thus, we have

$$\|\nabla F(x^{(t+1)})\|_M \leq \left(\epsilon_0 + \frac{\epsilon_1}{1-\epsilon_0}\right)\|\nabla F(x^{(t)})\|_M + \frac{2}{(1-\epsilon_0)^2} \cdot \frac{\hat{L}\kappa}{\mu\sqrt{\mu}}\|\nabla F(x^{(t)})\|_M^2.$$

By the Lipschitz continuity of $\nabla^2 F(x)$ and the condition

$$\|x^{(t)} - x^*\| \leq \frac{\mu}{\hat{L}\kappa} \leq \frac{\lambda_{\min}(\nabla^2 F(x^*))}{\hat{L}\kappa(\nabla^2 F(x^{(t)}))},$$

we obtain

$$\begin{aligned}
\|[\nabla^2 F(x^*)]^{-1} - [\nabla^2 F(x^{(t)})]^{-1}\| \leq& \|[\nabla^2 F(x^*)]^{-1}\|\|[\nabla^2 F(x^{(t)})]^{-1}\|\|\nabla^2 F(x^*) - \nabla^2 F(x^{(t)})\| \\
\leq& \hat{L}\|[\nabla^2 F(x^*)]^{-1}\|\|[\nabla^2 F(x^{(t)})]^{-1}\|\|x^{(t)} - x^*\| \\
\leq& \nu(t)\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-1}).
\end{aligned}$$

Hence, we can obtain that for any $u \in \mathbb{R}^d$,

$$-\nu(t)\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-1})u^T y \leq y^T([\nabla^2 F(x^*)]^{-1} - [\nabla^2 F(x^{(t)})]^{-1})y \leq \nu(t)\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-1})y^T y,$$

which yields

$$(1-\nu(t))u^T[\nabla^2 F(x^{(t)})]^{-1}u \leq u^T[\nabla^2 F(x^*)]^{-1}u \leq (1+\nu(t))u^T[\nabla^2 F(x^{(t)})]^{-1}u.$$

That is,

$$(1-\nu(t))\|u\|_M \leq \|u\|_{M^*} \leq (1+\nu(t))\|u\|_M.$$

Accordingly, we have

$$\|\nabla F(x^{(t+1)})\|_{M^*} \leq \left(\epsilon_0 + \frac{\epsilon_1}{1-\epsilon_0}\right)\frac{1+\nu(t)}{1-\nu(t)}\|\nabla F(x^{(t)})\|_{M^*} + \frac{2}{(1-\epsilon_0)^2} \cdot \frac{\hat{L}\kappa}{\mu\sqrt{\mu}}\frac{(1+\nu(t))^2}{1-\nu(t)}\|\nabla F(x^{(t)})\|_{M^*}^2$$

■

Appendix C. Proofs of theorems of Section 4

Proof of Theorem 4 If S is an ϵ_0-subspace embedding matrix w.r.t. $B(x^{(t)})$, then we have

$$(1-\epsilon_0)\nabla^2 F(x^{(t)}) \preceq [B(x^{(t)})]^T S^T S B(x^{(t)}) \preceq (1+\epsilon_0)\nabla^2 F(x^{(t)}). \tag{17}$$

By simple transformation and omitting ϵ_0^2, (17) can be transformed into

$$(1-\epsilon_0)[B(x^{(t)})]^T S^T S \nabla^2 B(x^{(t)}) \preceq \nabla^2 F(x^{(t)}) \preceq (1+\epsilon_0)[B(x^{(t)})]^T S^T S B(x^{(t)}).$$

The convergence rate can be derived directly from Theorem 3. ■

Appendix D. Proofs of theorems of Section 5

Proof of Theorem 5 By Lemma 12, when $|\mathcal{S}| \geq \frac{16K^2\log(2d/\delta)}{\sigma^2\epsilon_0^2}$, $H^{(t)}$ has the following property:

$$\|H^{(t)} - \nabla^2 F(x^{(t)})\| \leq \epsilon_0\sigma.$$

The above property implies the following:

$$\begin{aligned}
&|y^T(H^{(t)} - \nabla^2 F(x^{(t)}))y| \leq \epsilon_0\sigma y^T y,\\
\Rightarrow & -\epsilon_0\sigma y^T y \leq y^T(H^{(t)} - \nabla^2 F(x^{(t)}))y \leq \epsilon_0\sigma y^T y\\
\Rightarrow & \, H^{(t)} - \epsilon_0\sigma I \preceq \nabla^2 F(x^{(t)}) \preceq H^{(t)} + \epsilon_0\sigma I\\
\Rightarrow & \, (1-\epsilon_0)H^{(t)} \preceq \nabla^2 F(x^{(t)}) \preceq (1+\epsilon_0)H^{(t)}.
\end{aligned}$$

The convergence rate can be derived directly from Theorem 3. ■

Proof of Theorem 6
By Lemma 12, when $|\mathcal{S}| \geq \frac{16K^2\log(2d/\delta)}{\beta^2}$, we have

$$\|\nabla^2 F(x^{(t)}) - H^{(t)}_{|\mathcal{S}|}\| \leq \beta,$$

with probability at least $1-\delta$. Hence, we can derive

$$\begin{aligned}
&|y^T(\nabla^2 F(x^{(t)}) - H^{(t)}_{|\mathcal{S}|})y| \leq \beta y^T y\\
\Rightarrow & y^T H^{(t)}_{|\mathcal{S}|} y - \beta y^T y \leq y^T \nabla^2 F(x^{(t)}) y \leq y^T H^{(t)}_{|\mathcal{S}|} y + \beta y^T y\\
\Rightarrow & y^T H^{(t)} y - \alpha y^T y - \beta y^T y \leq y^T \nabla^2 F(x^{(t)}) y \leq y^T H^{(t)} y - \alpha y^T y + \beta y^T y\\
\Rightarrow & y^T H^{(t)} y - (\alpha+\beta) y^T y \overset{(1)}{\leq} y^T \nabla^2 F(x^{(t)}) y \overset{(2)}{\leq} y^T H^{(t)} y + (\beta-\alpha) y^T y.
\end{aligned}$$

Now we first consider $\overset{(1)}{\leq}$ case, we have

$$\begin{aligned}
&y^T H^{(t)} y - (\alpha+\beta) y^T y \leq y^T \nabla^2 F(x^{(t)}) y \\
\Rightarrow & y^T H^{(t)} y \leq y^T \nabla^2 F(x^{(t)}) y + (\alpha+\beta) y^T y \\
\Rightarrow & y^T H^{(t)} y \leq y^T \nabla^2 F(x^{(t)}) y + \frac{\alpha+\beta}{\sigma} y^T \nabla^2 F(x^{(t)}) y \\
\Rightarrow & y^T H^{(t)} y \leq \left(1 + \frac{\alpha+\beta}{\sigma}\right) y^T \nabla^2 F(x^{(t)}) y \\
\Rightarrow & \left(1 - \frac{\alpha+\beta}{\sigma+\alpha+\beta}\right) y^T H^{(t)} y \leq y^T \nabla^2 F(x^{(t)}) y \\
\Rightarrow & \left(1 - \frac{\alpha+\beta}{\sigma+\alpha+\beta}\right) H^{(t)} \preceq F(x^{(t)}).
\end{aligned}$$

For $\overset{(2)}{\leq}$ case, we consider two cases respectively. The first case is $\beta - \sigma/2 \leq \alpha \leq \beta$, and we have

$$\begin{aligned}
&y^T \nabla^2 F(x^{(t)}) y \leq y^T H^{(t)} y + (\beta-\alpha) y^T y \\
\Rightarrow & y^T \nabla^2 F(x^{(t)}) y - (\beta-\alpha) y^T y \leq y^T H^{(t)} y \\
\Rightarrow & y^T \nabla^2 F(x^{(t)}) y - \frac{\beta-\alpha}{\sigma} y^T \nabla^2 F(x^{(t)}) y \leq y^T H^{(t)} y \\
\Rightarrow & \left(1 - \frac{\beta-\alpha}{\sigma}\right) y^T \nabla^2 F(x^{(t)}) y \leq y^T H^{(t)} y \\
\Rightarrow & y^T \nabla^2 F(x^{(t)}) y \leq \left(1 + \frac{\beta-\alpha}{\sigma-(\beta-\alpha)}\right) y^T H^{(t)} y \\
\Rightarrow & \nabla^2 F(x^{(t)}) \preceq \left(1 + \frac{\beta-\alpha}{\sigma+\alpha-\beta}\right) H^{(t)}.
\end{aligned}$$

For the case $\beta < \alpha$, we can derive

$$\begin{aligned}
&y^T \nabla^2 F(x^{(t)}) y \leq y^T H^{(t)} y + (\beta-\alpha) y^T y \leq y^T H^{(t)} y \\
\Rightarrow & \nabla^2 F(x^{(t)}) \preceq (1+0) H^{(t)}.
\end{aligned}$$

Hence, for $\beta - \sigma \leq \alpha$, we have

$$\left(1 - \frac{\alpha+\beta}{\sigma+\alpha+\beta}\right) H^{(t)} \preceq F(x^{(t)}) \preceq \left(1 + \frac{\beta-\alpha}{\sigma+\alpha-\beta}\right) H^{(t)}.$$

Therefore, ϵ_0 in Theorem 3 can be set as follows:

$$\epsilon_0 = \max\left(\frac{\beta-\alpha}{\sigma+\alpha-\beta}, \frac{\alpha+\beta}{\sigma+\alpha+\beta}\right).$$

The convergence properties can derived from Theorem 3 directly. ■

Proof of Theorem 7
We denote the SVD of $H_{\mathcal{S}}^{(t)}$ as follows

$$H_{\mathcal{S}}^{(t)} = U\hat{\Lambda}U^T = U_r\hat{\Lambda}_r U_r^T + U_{\backslash r}\hat{\Lambda}_{\backslash r}U_{\backslash r}^T.$$

By Lemma 12, when $|\mathcal{S}| \geq \frac{16K^2\log(2d/\delta)}{\beta^2}$, we have

$$\|\nabla^2 F(x^{(t)}) - H_{|\mathcal{S}|}^{(t)}\| \leq \beta,$$

with probability at least $1-\delta$. Hence, we can derive

$$\begin{aligned}
&|y^T(\nabla^2 F(x^{(t)}) - H_{|\mathcal{S}|}^{(t)})y| \leq \beta y^T y \\
\Rightarrow & y^T H_{|\mathcal{S}|}^{(t)} y - \beta y^T y \leq y^T \nabla^2 F(x^{(t)}) y \leq y^T H_{|\mathcal{S}|}^{(t)} y + \beta y^T y \\
\Rightarrow & y^T H^{(t)} y + y^T U_{\backslash r}(\hat{\Lambda}_{\backslash r} - \hat{\lambda}_{r+1}^{(t)} I)U_{\backslash r}^T y - \beta y^T y \leq y^T \nabla^2 F(x^{(t)}) y \\
& \leq y^T H^{(t)} y + y^T U_{\backslash r}(\hat{\Lambda}_{\backslash r} - \hat{\lambda}_{r+1}^{(t)} I)U_{\backslash r}^T y + \beta y^T y \\
\Rightarrow & y^T H^{(t)} y - y^T U \begin{bmatrix} \beta I_r & \\ & (\beta + \hat{\lambda}_{r+1}^{(t)}) I_{\backslash r} - \hat{\Lambda}_{\backslash r} \end{bmatrix} U^T y \overset{(1)}{\leq} y^T \nabla^2 F(x^{(t)}) y \\
& \overset{(2)}{\leq} y^T H^{(t)} y + y^T U \begin{bmatrix} \beta I_r & \\ & (\beta - \hat{\lambda}_{r+1}^{(t)}) I_{\backslash r} + \hat{\Lambda}_{\backslash r} \end{bmatrix} U^T y
\end{aligned}$$

Now we first consider $\overset{(1)}{\leq}$ case, we have

$$\begin{aligned}
& y^T H^{(t)} y - y^T U \begin{bmatrix} \beta I_r & \\ & (\beta + \hat{\lambda}_{r+1}^{(t)}) I_{\backslash r} - \hat{\Lambda}_{\backslash r} \end{bmatrix} U^T y \overset{(1)}{\leq} y^T \nabla^2 F(x^{(t)}) y \\
\Rightarrow & y^T H^{(t)} y \leq y^T \nabla^2 F(x^{(t)}) y + (\beta + \hat{\lambda}_{r+1}^{(t)}) y^t y \\
\Rightarrow & y^T H^{(t)} y \leq y^T \nabla^2 F(x^{(t)}) y + \frac{\beta + \hat{\lambda}_{r+1}^{(t)}}{\sigma} y^T \nabla^2 F(x^{(t)}) y \\
\Rightarrow & y^T H^{(t)} y \leq y^T \nabla^2 F(x^{(t)}) y + \frac{2\beta + \hat{\lambda}_{r+1}}{\sigma} y^T \nabla^2 F(x^{(t)}) y \\
\Rightarrow & \left(1 - \frac{2\beta + \lambda_{r+1}^{(t)}}{\sigma + 2\beta + \lambda_{r+1}^{(t)}}\right) y^T H^{(t)} y \leq y^T \nabla^2 F(x^{(t)}) y.
\end{aligned}$$

Hence we have

$$\left(1 + \frac{\beta}{\lambda_{r+1}^{(t)} - \beta}\right) H^{(t)} \preceq \nabla^2 F(x).$$

Now we first consider $\overset{(2)}{\leq}$ case, we have

$$\begin{aligned}
y^T\nabla^2F(x^{(t)})y \leq & y^TH^{(t)}y + y^TU\begin{bmatrix} \beta I_r & \\ & (\beta - \hat{\lambda}_{r+1}^{(t)})I_{\backslash r} + \hat{\Lambda}_{\backslash r}\end{bmatrix}U^Ty \\
\leq & y^TH^{(t)}y + \frac{\beta}{\hat{\lambda}_{r+1}^{(t)}}y^TH^{(t)}y \\
\leq & \left(1 + \frac{\beta}{\lambda_{r+1}^{(t)} - \beta}\right)y^TH^{(t)}y,
\end{aligned}$$

where the last inequality is because $\lambda_{r+1}^{(t)} - \beta \leq \hat{\lambda}_{r+1}^{(t)}$. Hence, we have

$$\nabla^2F(x) \preceq \left(1 + \frac{\beta}{\lambda_{r+1}^{(t)} - \beta}\right)H^{(t)}.$$

Hence, we have

$$\epsilon_0 = \max\left(\frac{\beta}{\lambda_{r+1}^{(t)} - \beta}, \frac{2\beta + \lambda_{r+1}^{(t)}}{\sigma + 2\beta + \lambda_{r+1}^{(t)}}\right) < 1,$$

because $\beta \leq \frac{\lambda_{r+1}^{(t)}}{2}$.

The convergence properties can be derived directly by Theorem 3. ■

Appendix E. Subsampled Hessian and Gradient

In fact, we can also subsample gradient to accelerate the subsampled Newton method. The detailed procedure is presented in Algorithm 5 [1, 2].

Theorem 13 *Let $F(x)$ satisfy the properties described in Theorem 3. We also assume Eqn.* (9) *and Eqn.* (10) *hold and let $0 < \delta < 1$ and $0 < \epsilon_0 < 1/2$ be given. Let $|\mathcal{S}_H|$ and $|\mathcal{S}_{\mathrm{g}}|$ be set such that Eqn.* (2) *holds and it holds that*

$$\|\mathrm{g}(x^{(t)}) - \nabla F(x^{(t)})\| \leq \frac{\epsilon_2}{\kappa}\|\nabla F(x^{(t)})\|.$$

The direction vector $\mathrm{p}^{(t)}$ is computed as in Algorithm 5. Then for $t = 1, \dots, T$, we have the following convergence properties:

(a) There exists a sufficient small value γ, $0 < \nu(t) < 1$, and $0 < \eta(t) < 1$ such that when $\|x^{(t)} - x^*\| \leq \gamma$, then for each iteration, it holds that

$$\|\nabla F(x^{(t+1)})\|_{M^*} \leq (\epsilon_0 + 2\epsilon_2 + 4\eta(t))\frac{1 + \nu(t)}{1 - \nu(t)}\|\nabla F(x^{(t)})\|_{M^*}$$

with probability at least $1 - \delta$.

Algorithm 5 Subsampled Hessian and Subsampled Gradient.

```
Input: $x^{(0)}$, $0 < \delta < 1$, $0 < \epsilon_0 < 1$;
Set the sample size $|\mathcal{S}_H|$ and $|\mathcal{S}_{\mathrm{g}}|$.
for $t = 0, 1, \ldots$ until termination do
    Select a sample set $\mathcal{S}_H$, of size $|\mathcal{S}|$ and construct $H^{(t)} = \frac{1}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} \nabla^2 f_j(x^{(t)})$;
    Select a sample set $\mathcal{S}_{\mathrm{g}}$ of size $|\mathcal{S}_{\mathrm{g}}|$ and calculate $\mathrm{g}(x^{(t)}) = \frac{1}{|\mathcal{S}_{\mathrm{g}}|}\sum_{i\in\mathcal{S}_{\mathrm{g}}} \nabla f_i(x^{(t)})$.
    Calculate $p^{(t)} = [H^{(t)}]^{-1}\mathrm{g}(x^{(t)})$;
    Update $x^{(t+1)} = x^{(t)} - p^{(t)}$;
end for
```

(b) *If $\nabla^2 F(x^{(t)})$ is also Lipschitz continuous and $\{x^{(t)}\}$ satisfies Eqn.* (6)*, then for each iteration, it holds that*

$$\|\nabla F(x^{(t+1)})\|_{M^*} \le (\epsilon_0 + 2\epsilon_2)\frac{1+\nu(t)}{1-\nu(t)}\|\nabla F(x^{(t)})\|_{M^*} + \frac{8\hat{L}\kappa}{\mu\sqrt{\mu}}\frac{(1+\nu(t))^2}{1-\nu(t)}\|\nabla F(x^{(t)})\|^2_{M^*}.$$

with probability at least $1-\delta$.

In common cases, subsampled gradient $\mathrm{g}(x^{(t)})$ needs to subsample over 80% of samples to guarantee convergence of the algorithm. Roosta-Khorasani and Mahoney [2] showed that it needs $|\mathcal{S}_{\mathrm{g}}| \ge G(x^{(t)})^2\kappa^2/(\nu^2(t)\|\nabla F(x^{(t)})\|^2)$, where $G(x^{(t)}) = \max_i \|\nabla f_i(x^{(t)})\|$ for $i = 1, \ldots, n$. When $x^{(t)}$ is close to x^*, $\|\nabla F(x^{(t)})\|$ is close to 0. Hence $|\mathcal{S}_{\mathrm{g}}|$ will go to n as iteration goes. This is the reason why the Newton method and variants of the subsampled Newton method are very sensitive to the accuracy of subsampled gradient.

The proof of Theorem 13 is almost the same with Theorem 3. For completeness, we give the detailed proof as follows.

Proof By Taylor's theorem, we obtain

$$\begin{aligned}
&\nabla F(x^{(t+1)}) \\
=&\nabla F(x^{(t)}) + \nabla^2 F(x^{(t)})(-p^{(t)}) + \int_0^1 [\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})](-p^{(t)})ds \\
=&\nabla F(x^{(t)}) - \nabla^2 F(x^{(t)})[H^{(t)}]^{-1}\nabla F(x^{(t)}) + \nabla^2 F(x^{(t)})[H^{(t)}]^{-1}\nabla F(x^{(t)}) - \nabla^2 F(x^{(t)})p^{(t)} \\
&+ \int_0^1 [\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})](-p^{(t)})ds \\
=&\left[\nabla^2 F(x^{(t)})\right]^{\frac{1}{2}}\left(I - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\right)\left[\nabla^2 F(x^{(t)})\right]^{-\frac{1}{2}}\nabla F(x^{(t)}) \\
&+ \nabla^2 F(x^{(t)})([H^{(t)}]^{-1}\nabla F(x^{(t)}) - p^{(t)}) + \int_0^1 [\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})](-p^{(t)})ds.
\end{aligned}$$

Hence, we have the following identity

$$\begin{aligned}
\left[\nabla^2 F(x^{(t)})\right]^{-\frac{1}{2}}\nabla F(x^{(t+1)}) =& \left(I - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}[H^{(t)}]^{-1}[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\right)\left[\nabla^2 F(x^{(t)})\right]^{-\frac{1}{2}}\nabla F(x^{(t)}) \\
&+ [\nabla^2 F(x^{(t)})]^{\frac{1}{2}}([H^{(t)}]^{-1}\nabla F(x^{(t)}) - p^{(t)}) \\
&+ [\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\int_0^1 [\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})](-p^{(t)})ds.
\end{aligned}$$

Further more, we define $M = \left[\nabla^2 F(x^{(t)})\right]^{-1}$, we can obtain

$$\begin{aligned}\|\nabla F(x^{(t+1)})\|_M \le & \left\| I - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{\frac{1}{2}} \right\| \|\nabla F(x^{(t)})\|_M \\ & + \|[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\| \|[H^{(t)}]^{-1}(\nabla F(x^{(t)}) - \mathrm{g}(x^{(t)}))\| \\ & + \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 \|\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})\| \|p^{(t)}\| ds.\end{aligned}$$

We will bound the three terms on the right hand of above equation seperately.

For the first term, using Lemma 9, we have

$$\left\| I - [\nabla^2 F(x^{(t)})]^{\frac{1}{2}} [H^{(t)}]^{-1} [\nabla^2 F(x^{(t)})]^{\frac{1}{2}} \right\| \|\nabla F(x^{(t)})\|_M \le \epsilon_0 \|\nabla F(x^{(t)})\|_M.$$

For the second term, by the fact that $\|AB\| \ge \|A\| \sigma_{\min}(B)$ and condition $\|\mathrm{g}(x^{(t)}) - \nabla F(x^{(t)})\| \le \epsilon_2 \|\nabla F(x^{(t)})\|$, we obtain

$$\begin{aligned}& \|[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\| \|[H^{(t)}]^{-1}(\nabla F(x^{(t)}) - \mathrm{g}(x^{(t)}))\| \\ = & \frac{\|[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\|}{\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})} \lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}) \|[H^{(t)}]^{-1}\| \|\nabla F(x^{(t)}) - \mathrm{g}(x^{(t)})\| \\ \le & \epsilon_2 \frac{\|[\nabla^2 F(x^{(t)})]^{\frac{1}{2}}\|}{\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})} \|[H^{(t)}]^{-1}\| \|\nabla F(x^{(t)})\|_M \\ \le & \frac{\epsilon_2}{1-\epsilon_0} \|\nabla F(x^{(t)})\|_M \\ \le & 2\epsilon_2 \|\nabla F(x^{(t)})\|_M\end{aligned}$$

For the third term, we bound it for the case that $\nabla^2 F(x)$ is not Lipschitz continuous and the case $\nabla^2 F(x)$ is Lipschitz continuous respectively.

(a) Now we consider the case that $\nabla^2 F(x)$ is not Lipschitz continuous but is continuous close to the optimal point x^*. Because $\nabla^2 F(x)$ is continuous near x^*, there exists a sufficient small value δ such that Eqn. (14) and Eqn. (15) hold when $\|x^{(t)} - x^*\| \le \delta$.

By μ-strong convexity, we have $\|[\nabla^2 F(x^t)]^{-1}\| \le \frac{1}{\mu}$ for all $x^{(t)}$ sufficiently close to x^*. Then we have

$$\|p^{(t)}\| = \|[H^{(t)}]^{-1}\| \|\mathrm{g}(x^{(t)})\| \le \frac{1 + \epsilon_2/\kappa}{(1-\epsilon_0)\mu} \|\nabla F(x^{(t)})\| \le \frac{2}{(1-\epsilon_0)\mu} \|\nabla F(x^{(t)})\|.$$

Hence, with $\|x - x^*\| \le \delta$, combining condition (15), we have

$$\begin{aligned}& \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 \|\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})\| \|p^{(t)}\| ds \\ \le & \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 \frac{\mu \eta(t)}{\sqrt{\kappa}} \|p^{(t)}\| ds \\ \le & \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \frac{2}{(1-\epsilon_0)\mu} \frac{\mu \eta(t)}{\sqrt{\kappa}} \|\nabla F(x^{(t)})\| \\ \le & \frac{2\eta(t)}{1-\epsilon_0} \frac{\|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\|}{\sqrt{\kappa} \lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})} \lambda_{\min}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}) \|\nabla F(x^{(t)})\| \\ \le & 4\eta(t) \|\nabla F(x^{(t)})\|_M,\end{aligned}$$

Therefore, we have

$$\begin{aligned}\|\nabla F(x^{(t+1)})\|_M \leq & \epsilon_0\|\nabla F(x^{(t)})\|_M + 2\epsilon_2\|\nabla F(x^{(t)})\|_M + 4\eta(t)\|\nabla F(x^{(t)})\|_M \\ = & (\epsilon_0 + 2\epsilon_2 + 4\eta(t))\|\nabla F(x^{(t)})\|_M.\end{aligned}$$

By Eqn. (14), we have

$$\begin{aligned}& -\frac{\nu(t)}{\lambda_{\max}(\nabla^2 F(x^*))}y^Ty \leq y^T([\nabla^2 F(x^*)]^{-1} - [\nabla^2 F(x^{(t)})]^{-1})y \leq \frac{\nu(t)}{\lambda_{\max}(\nabla^2 F(x^*))}y^Ty, \\ \Rightarrow & (1-\nu(t))y^T[\nabla^2 F(x^{(t)})]^{-1}y \leq y^T[\nabla^2 F(x^*)]^{-1}y \leq (1+\nu(t))y^T[\nabla^2 F(x^{(t)})]^{-1}y \\ \Rightarrow & (1-\nu(t))\|y\|_M \leq \|y\|_{M^*} \leq (1+\nu(t))\|y\|_M.\end{aligned}$$

By this relationship between $\|\cdot\|_M$ and $\|\cdot\|_{M^*}$, we get

$$\|\nabla F(x^{(t+1)})\|_{M^*} \leq (\epsilon_0 + 2\epsilon_2 + 4\eta(t))\frac{1+\nu(t)}{1-\nu(t)}\|\nabla F(x^{(t)})\|_{M^*}$$

(b) Now we consider the case that $\nabla^2 F(x)$ is Lipschitz continuous with parameter $\hat{L}$. The same to the previous proof, we have

$$\|p^{(t)}\| = \|[H^{(t)}]^{-1}\|\|\mathrm{g}(x^{(t)})\| \leq \frac{1+\epsilon_2/\kappa}{(1-\epsilon_0)\mu}\|\nabla F(x^{(t)})\| \leq \frac{2}{(1-\epsilon_0)\mu}\|\nabla F(x^{(t)})\|.$$

Because $\nabla^2 F(x)$ is Lipschitz continuous with parameter $\hat{L}$, we have

$$\begin{aligned}& \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 \|\nabla^2 F(x^{(t)} + sp^{(t)}) - \nabla^2 F(x^{(t)})\|\|p^{(t)}\|ds \\ \leq & \|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\| \int_0^1 sL\|p^{(t)}\|^2 ds \\ = & \frac{\hat{L}}{2}\|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\|\lambda_{\min}^{-2}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})\lambda_{\min}^2([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})\|p^{(t)}\|^2 \\ \leq & \frac{\hat{L}}{2}\|[\nabla^2 F(x^{(t)})]^{-\frac{1}{2}}\|\lambda_{\min}^{-2}([\nabla^2 F(x^{(t)})]^{-\frac{1}{2}})\left(\frac{2}{(1-\epsilon_0)\lambda_{\min}(\nabla^2 F(x^{(t)}))}\right)^2\|\nabla F(x^{(t)})\|_M^2 \\ = & \frac{2\hat{L}\lambda_{\max}(\nabla^2 F(x^{(t)}))}{(1-\epsilon_0)^2\lambda_{\min}^2(\nabla^2 F(x^{(t)}))\sqrt{\lambda_{\min}(\nabla^2 F(x^{(t)}))}}\|\nabla F(x^{(t)})\|_M^2 \\ \leq & \frac{8\hat{L}\kappa}{\mu\sqrt{\mu}}\|\nabla F(x^{(t)})\|_M^2,\end{aligned}$$

where the last inequality is because $\epsilon_0 \leq 1/2$. Hence, we have

$$\|\nabla F(x^{(t+1)})\|_M \leq (\epsilon_0 + 2\epsilon_2)\|\nabla F(x^{(t)})\|_M + \frac{8\hat{L}\kappa}{\mu\sqrt{\mu}}\|\nabla F(x^{(t)})\|_M^2.$$

By the Lipschitz continuity of $\nabla^2 F(x)$ and the condition

$$\|x^{(t)} - x^*\| \leq \frac{\mu}{\hat{L}\kappa} \leq \frac{\lambda_{\min}(\nabla^2 F(x^*))}{\hat{L}\kappa(\nabla^2 F(x^{(t)}))},$$

we obtain

$$
\begin{aligned}
\|[\nabla^2 F(x^*)]^{-1} - [\nabla^2 F(x^{(t)})]^{-1}\| \leq & \|[\nabla^2 F(x^*)]^{-1}\| \|[\nabla^2 F(x^{(t)})]^{-1}\| \|\nabla^2 F(x^*) - \nabla^2 F(x^{(t)})\| \\
\leq & \hat{L} \|[\nabla^2 F(x^*)]^{-1}\| \|[\nabla^2 F(x^{(t)})]^{-1}\| \|x^{(t)} - x^*\| \\
\leq & \nu(t) \lambda_{\min}([\nabla^2 F(x^{(t)})]^{-1}).
\end{aligned}
$$

Hence, we can derive

$$
\begin{aligned}
& -\nu(t)\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-1}) y^T y \leq y^T ([\nabla^2 F(x^*)]^{-1} - [\nabla^2 F(x^{(t)})]^{-1}) y \leq \nu(t)\lambda_{\min}([\nabla^2 F(x^{(t)})]^{-1}) y^T y, \\
\Rightarrow & (1-\nu(t)) y^T [\nabla^2 F(x^{(t)})]^{-1} y \leq y^T [\nabla^2 F(x^*)]^{-1} y \leq (1+\nu(t)) y^T [\nabla^2 F(x^{(t)})]^{-1} y \\
\Rightarrow & (1-\nu(t)) \|y\|_M \leq \|y\|_{M^*} \leq (1+\nu(t)) \|y\|_M .
\end{aligned}
$$

Hence, we have

$$
\|\nabla F(x^{(t+1)})\|_{M^*} \leq (\epsilon_0 + 2\epsilon_2) \frac{1+\nu(t)}{1-\nu(t)} \|\nabla F(x^{(t)})\|_{M^*} + \frac{8\hat{L}\kappa}{\mu\sqrt{\mu}} \frac{(1+\nu(t))^2}{1-\nu(t)} \|\nabla F(x^{(t)})\|_{M^*}^2
$$

■

Table 2: Datasets Description

Dataset	n	d	source
mushrooms	8124	112	UCI
a9a	32561	123	UCI
Covertype	581012	54	UCI

Appendix F. Unnecessity of Lipschitz continuity of Hessian

In this section, we validate our theoretical results about unnecessity of the Lipschitz continuity condition of $\nabla^2 F(x)$. We conduct experiment on the primal problem for the linear SVM which can be written as

$$
\min_x F(x) = \frac{1}{2}\|x\|^2 + \frac{C}{2n} \sum_{i=1}^{n} \ell(b_i, \langle x, a_i \rangle)
$$

where (a_i, b_i) denotes the training data, x defines the separating hyperplane, $C > 0$, and $\ell(\cdot)$ is the loss function. In our experiment, we choose Hinge-2 loss as our loss function whose definition is

$$
\ell(b, \langle x, a \rangle) = \max(0, 1 - b\langle x, a \rangle)^2.
$$

Let $SV^{(t)}$ denote the set of indices of all the support vectors at iteration t, i.e.,

$$
SV^{(t)} = \{i : b_i \langle x^{(t)}, a_i \rangle < 1\}.
$$

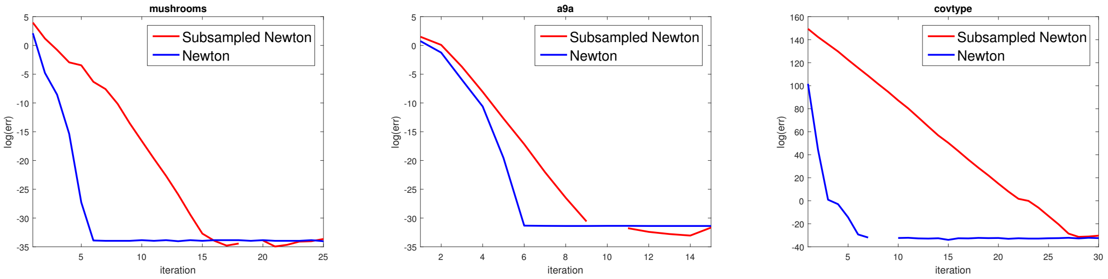
Then the Hessian matrix of $F(x^{(t)})$ can be written as

$$\nabla^2 F(x^{(t)}) = I + \frac{1}{n} \sum_{i \in SV^{(t)}} a_i a_i^T.$$

From the above equation, we can see that $\nabla^2 F(x)$ is not Lipschitz continuous.

Without loss of generality, we use the Subsampled Newton method (Algorithm 2) in our experiment. We sample 5% support vectors in each iteration. Our experiments on three datasets whose detailed description is in Table 2 and report our results in Figure 3.

From Figure 3, we can see that Subsampled Newton converges linearly and the Newton method converges superlinearly. This matches our theory that the Lipschitz continuity of $\nabla^2 F(x)$ is not necessary to achieve a linear or superlinear convergence rate.



(a) mushrooms.

(b) a9a.

(c) covtype.

Figure 3: Convergence properties on different datasets.

References

[1] Richard H Byrd, Gillian M Chin, Will Neveitt, and Jorge Nocedal. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3):977–995, 2011.

[2] Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled newton methods ii: Local convergence rates. *arXiv preprint arXiv:1601.04738*, 2016.

[3] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.