Supplement for “Learning with Non-Convex Truncated Losses by SGD”

A Properties of truncation functions

In this section, we first verify that three examples of truncation functions satisfy Definition 1.

Example 1. \( \phi^{(1)}_{\alpha}(x) = \alpha \log(1 + \frac{x}{\alpha}) \). We have \( \phi^{(1)}_{\alpha}(x) = \frac{1}{1+x/\alpha} \). Then it is easy to check it satisfies condition (ii), (iii), and for any \( \alpha_1 \leq \alpha_2 \), we have \( \phi'_{\alpha_1}(x) \leq \phi'_{\alpha_2}(x) \). Since \( \phi^{(1)}_{\alpha}(x) = -\frac{1}{\alpha x} \frac{1}{1+x/\alpha^2} \), then \( |\phi^{(1)}_{\alpha}(x)| \leq 1/\alpha \), indicating that it satisfies condition (i).

Example 2. \( \phi^{(2)}_{\alpha}(x) = \alpha \log(1 + \frac{x}{\alpha} + \frac{x^2}{2\alpha^2}) \). We have \( \phi^{(2)}_{\alpha}(x) = \frac{1}{1+x/\alpha^2} = 1 - \frac{1}{1+2x/\alpha+2\alpha^2/x^2} \). Then it is easy to check it satisfies condition (ii), (iii), and for any \( \alpha_1 \leq \alpha_2 \), we have \( \phi'_{\alpha_1}(x) \leq \phi'_{\alpha_2}(x) \). Since \( \phi^{(2)}_{\alpha}(x) = -\frac{1}{\alpha} \frac{1}{(1+x/\alpha^2)} \), then \( |\phi^{(2)}_{\alpha}(x)| \leq 1/\alpha \), indicating that it satisfies condition (i).

Example 3.

\[
\phi^h_{\alpha}(x) = \begin{cases} \frac{\alpha}{3} [1 - \left(1 - \frac{x}{\alpha}\right)^3] & \text{if } 0 \leq x < \alpha, \\ \frac{\alpha}{3} & \text{otherwise.} \end{cases}
\]

Then we have

\[
\phi'_{\alpha}(x) = \begin{cases} (1 - \frac{x}{\alpha})^2 & \text{if } 0 \leq x < \alpha, \\ 0 & \text{otherwise.} \end{cases}
\]

Then it is easy to check it satisfies condition (ii), (iii), and for any \( \alpha_1 \leq \alpha_2 \), we have \( \phi'_{\alpha_1}(x) \leq \phi'_{\alpha_2}(x) \). Since

\[
\phi^h_{\alpha}(x) = \begin{cases} -\frac{\alpha}{3} (1 - \frac{x}{\alpha}) & \text{if } 0 \leq x < \alpha, \\ 0 & \text{otherwise.} \end{cases}
\]

then \( |\phi^h_{\alpha}(x)| \leq 2/\alpha \), indicating that it satisfies condition (i).

Next, we will verify the conditions \( |x - \phi_{\alpha}^{(1)}(x)| \leq \frac{M x^2}{\alpha} \), \( |x - \phi_{\alpha}^{(2)}(x)| \leq \frac{M x^2}{\alpha} \), and \( |x - \phi_{\alpha}^{h}(x)| \leq \frac{M x^2}{\alpha} \).

**Proposition 1.** For any \( \alpha > 0 \) and \( x \geq 0 \), we have

\[
|x - \phi_{\alpha}^{(1)}(x)| \leq \frac{x^2}{2\alpha} \quad \text{and} \quad |x - \phi_{\alpha}^{(2)}(x)| \leq \frac{x^2}{2\alpha}.
\]

**Proof.** We first need the following result to prove the proposition:

\[
\exp(y) \geq 1 + y + \frac{y^2}{2} \quad \text{for all } y \geq 0.
\]

Let’s first consider \( \phi_{\alpha}^{(1)}(x) \), to prove \( |x - \alpha \log(1 + x/\alpha)| \leq \frac{1}{2\alpha} x^2 \), we have to show \( |x/\alpha - \log(1 + x/\alpha)| \leq \frac{1}{2\alpha} x^2 \). Let \( y = x/\alpha \geq 0 \), we only need to show \( |y - \log(1+y)| \leq \frac{y^2}{2} \). By the inequality (2) we know that \( \log(1 + y) - y \leq 0 \), so we only need to show \( f(y) := y - \log(1+y) - \frac{y^2}{2} \leq 0 \) for all \( y \geq 0 \). Since \( f'(y) = -\frac{y}{1+y} \leq 0 \), then we know \( f(y) \) is a decreasing function on \( y \geq 0 \) thus \( f(y) \leq f(0) = 0 \), which gives the first inequality in (3).

Next let’s consider \( \phi_{\alpha}^{(2)}(x) \). Similarly, we only need to show \( f(y) := y - \log(1 + y + y^2/2) - \frac{y^2}{2} \leq 0 \) for all \( y \geq 0 \). Since \( f'(y) = -\frac{y+y^2/2+y^3/2}{1+y+y^2/2} \leq 0 \), then we know \( f(y) \) is a decreasing function on \( y \geq 0 \) thus \( f(y) \leq f(0) = 0 \), which gives the second inequality in (3).

Next let’s consider \( \phi_{\alpha}^{h}(x) \). Similarly, we only need to show \( f(y) := y - \log(1 + y + y^2/2) - \frac{y^2}{2} \leq 0 \) for all \( y \geq 0 \). Since \( f'(y) = -\frac{y+y^2/2+y^3/2}{1+y+y^2/2} \leq 0 \), then we know \( f(y) \) is a decreasing function on \( y \geq 0 \) thus \( f(y) \leq f(0) = 0 \), which gives the third inequality in (3).

**Proposition 2.** For any \( \alpha > 0 \) and \( x \geq 0 \), we have

\[
|x - \phi_{\alpha}^{h}(x)| \leq \frac{x^2}{\alpha},
\]

(3)
Proof. Let first consider $0 \leq x < \alpha$, then we want to show $|x - \frac{3}{4}[1 - (1 - \frac{3}{4})^3]| \leq \frac{M \alpha^2}{\alpha}$, or equivalently $\left| \frac{3}{4} - \frac{3}{4}[1 - (1 - \frac{3}{4})^3] \right| \leq \frac{M \alpha^2}{\alpha}$. Let $y = \frac{3}{4} \in [0, 1)$, we only need to show $\left| y - \frac{3}{4}[1 - (1 - y)^3] \right| \leq M \alpha^2$.

(i) When $y - \frac{3}{4}[1 - (1 - y)^3] > 0$, then we need to show $f(y) := y - \frac{3}{4}[1 - (1 - y)^3] - M \alpha^2 \leq 0$. In fact, $f'(y) = 1 - (1 - y)^2 - 2My = 2(1 - M)y - y^2$. By setting $M \geq 1$, we know $f'(y) < 0$. Therefore, $f(y) \leq f(0) = 0$ for all $0 \leq y < 1$.

(ii) When $y - \frac{3}{4}[1 - (1 - y)^3] \leq 0$, then we need to show $f(y) := \frac{3}{4}[1 - (1 - y)^3] - y - M \alpha^2 \leq 0$. In fact, $f'(y) = (1 - y)^2 - 1 - 2My = -1 + 2My \leq 0$, then $f(y) \leq f(0) = 0$ for all $0 \leq y < 1$.

Next we consider $x \geq \alpha$, then we want to show $|x - \frac{3}{4}| \leq \frac{M \alpha^2}{\alpha}$, or equivalently $\left| \frac{3}{4} - \frac{3}{4} \right| \leq \frac{M \alpha^2}{\alpha}$. Let $y = \frac{3}{4} \geq 1$, we only need to show $\left| y - \frac{3}{4} \right| \leq M \alpha^2$. Since $y > 1$, we must show $y - \frac{3}{4} \leq M \alpha^2$. By setting $M \geq 1$, this trivially holds. In summary, we can choose $M = 1$, which completes the proof. \qed

B Proof of Theorem 2

We will use the following lemma to prove this theorem. The proof of this lemma can be found in subsection B.1.

Lemma 1. Under the same setting as Theorem 2, with a probability at least $1 - 3\delta$, we have

$$\sup_{f \in F} |\Lambda(f) - \Lambda(f^*)| \leq C \beta(F, \alpha) \log(2/\delta) \left( \frac{\gamma_2(F, d_e)}{\sqrt{n}} + \frac{\gamma_1(F, d_m)}{n} \right),$$

where $\Lambda(f) = P(\phi_\alpha(f)) - P_n(\phi_\alpha(f))$, $C$ is a universal constant.

Proof of Theorem 2. By (6), we know $\tilde{f} = \arg\min_{f \in F} P_n(\phi_\alpha(f))$, and thus $P_n(\phi_\alpha(\tilde{f})) - P_n(\phi_\alpha(f^*)) \leq 0$, where $f^* = \arg\min_{f \in F} P(f)$. Then we have

$$P(\tilde{f}) - P(f^*) \leq P(\tilde{f} - P(\phi_\alpha(\tilde{f}))) + [P(\phi_\alpha(\tilde{f})) - P_n(\phi_\alpha(\tilde{f}))] + [P_n(\phi_\alpha(\tilde{f})) - P_n(\phi_\alpha(f^*))]$$

$$+ [P_n(\phi_\alpha(f^*)) - P(\phi_\alpha(f^*))) + [P(\phi_\alpha(f^*))) - P(\phi_\alpha(f^*)))$$

$$\leq P(\tilde{f}) - P(\phi_\alpha(\tilde{f})) + [P(\phi_\alpha(\tilde{f})) - P_n(\phi_\alpha(f^*)) + [P_n(\phi_\alpha(f^*)) - P(\phi_\alpha(f^*)))$$

$$+ [P(\phi_\alpha(f^*))) - P(\phi_\alpha(f^*))) + 2M \alpha^2.$$

where the last inequality is derived using the fact that $E[X - \phi_\alpha(X)] \leq E \left[ \frac{M \alpha}{\alpha} X^2 \right]$ for a random variable $X$. Then by Lemma 1, with a probability at least $1 - 3\delta$,

$$P(\tilde{f}) - P(f^*) \leq C \beta(F, \alpha) \log(2/\delta) \left( \frac{\gamma_2(F, d_e)}{\sqrt{n}} + \frac{\gamma_1(F, d_m)}{n} \right) + \frac{2M \alpha^2}{\alpha}.$$

\qed

B.1 Proof of Lemma 1

Proof. This proof is similar to the analysis in Proposition 5 and Lemma 6 from [1]. For completeness, we include it here. For any $f, f' \in F$, we first know that $n(\Lambda(f) - \Lambda(f'))$ is the summation of the following independent random variables with zero mean:

$$C_i(f, f') = \phi_\alpha(f(Z_i)) - \phi_\alpha(f'(Z_i)) - [E[\phi_\alpha(f(Z))] - E[\phi_\alpha(f'(Z))] \leq 2\beta(F, \alpha) d_m(f, f'),$$

where the last inequality is due to $\phi_\alpha$ is Lipschitz continuous and $\beta(F, \alpha) = \sup_{f, Z} \phi_\alpha(f(Z))$. On the other hand,

$$\sum_{i=1}^n E(C_i(f, f')^2) \leq n \beta^2(F, \alpha) d_m^2(f, f').$$
Then by using Bernstein’s inequality we have for any \( f, f' \in \mathcal{F} \) and \( \theta > 0 \),

\[
\Pr(|\Lambda(f) - \Lambda(f')| > \theta) \leq 2 \exp \left( -\frac{n\theta^2}{2(\beta^2(\mathcal{F}, \alpha) d_f^2(f, f') + \theta \beta(\mathcal{F}, \alpha) d_m(f, f')/3)} \right).
\]

Then by using Theorem 12 and inequality (14) from [1], let \( f' = f^* \) we get

\[
\sup_{f \in \mathcal{F}} |\Lambda(f) - \Lambda(f^*)| \leq C \beta(\mathcal{F}, \alpha) \log(2/\delta) \left( \frac{\gamma_2(\mathcal{F}, d_e)}{\sqrt{n}} + \frac{\gamma_1(\mathcal{F}, d_m)}{n} \right),
\]

where \( C \) is a constant.

\( \square \)

**C  Proof of Corollary 3**

**Proof.** By assumption we know that there exists a constant \( D > 0 \) such that \( \max_{X \in \mathcal{X}, h, h' \in \mathcal{H}} |h(X) - h'(X)| \leq D \). Then for any \( X \in \mathcal{X} \), by the Lipschitz continuity of \( \ell \) function, we know that

\[
|\ell(h(X), Y) - \ell(h'(X), Y)| \leq L|h(X) - h'(X)| \leq LD.
\]

where \( L \) is the Lipschitz constant of \( \ell() \) with respect to its first argument. By the definition of \( \mathcal{H} \), Since for any \( f, f' \in \mathcal{F} \), we have \( \ell_m(f, f') \leq Ld_m(h, h') \), where \( f = \ell(h(\cdot), \cdot) \) and \( f' = \ell(h'(\cdot), \cdot) \). Hence an \( \epsilon/L \)-cover of \( \mathcal{H} \) under the metric \( d_m \) induces an \( \epsilon \)-cover of \( \mathcal{F} \) under the metric \( d_m \). Therefore, we have

\[
\log N(\mathcal{F}, \epsilon, d_m) \leq \log N(\mathcal{H}, \epsilon/L, d_m).
\]

Since \( \mathcal{H} \) is a compact set under distance measure \( d_m \) by the assumption, its covering number is finite \([2]\). Then

\[
\gamma_1(\mathcal{F}, d_m) \leq \int_0^1 \log N(\mathcal{F}, \epsilon, d_m) d\epsilon \leq \int_0^1 \log N(\mathcal{H}, \epsilon/L, d_m) d\epsilon < \infty.
\]

Similarly,

\[
\gamma_2(\mathcal{F}, d_e) \leq \int_0^1 \log N(\mathcal{F}, \epsilon, d_e)^{1/2} d\epsilon \leq \int_0^1 \log N(\mathcal{F}, \epsilon, d_m)^{1/2} d\epsilon \leq \int_0^1 \log N(\mathcal{H}, \epsilon/L, d_m)^{1/2} d\epsilon \leq \infty
\]

By setting \( \alpha \geq \Omega(\sqrt{n}) \) in Theorem 2, we get the result.  \( \square \)

**D  Proof of Theorem 4**

We will use the following lemma to prove this theorem. The proof of this lemma can be found in subsection D.1.

**Lemma 2.** Under the same setting as Theorem 4, with a probability at least \( 1 - 3\delta \), we have

\[
\sup_{f \in \mathcal{F}} |\Lambda(f) - \Lambda(f^*)| \leq C \beta(\mathcal{F}, \alpha) \max(\Gamma_\delta, \Delta(\mathcal{F}, d_e)) \sqrt{\frac{\log \left( \frac{n}{\delta} \right)}{n}},
\]

where \( \Lambda(f) = P(\phi_\alpha(f)) - P_n(\phi_\alpha(f)) \), \( C \) is a universal constant.

**Proof of Theorem 4.** Similar to the proof of Theorem 2, we have

\[
P(\tilde{f}) - P(f^*) \leq \left[ P(\phi_\alpha(\tilde{f})) - P_n(\phi_\alpha(\tilde{f})) \right] + \left[ P_n(\phi_\alpha(f^*)) - P(\phi_\alpha(f^*)) \right] + \frac{2M\sigma^2}{\alpha}.
\]

Then by Lemma 2, with a probability at least \( 1 - 3\delta \),

\[
P(\tilde{f}) - P(f^*) \leq C \beta(\mathcal{F}, \alpha) \max(\Gamma_\delta, \Delta(\mathcal{F}, d_e)) \sqrt{\frac{\log \left( \frac{n}{\delta} \right)}{n}} + \frac{2M\sigma^2}{\alpha}.
\]
By the definition of $d \phi$, where the second inequality uses the fact that $\sqrt{n} \in F$, and any $i$ in $W_i(f) = \frac{1}{n} \phi_\alpha(f(Z_i)) - \frac{1}{n} \phi_\alpha(f(Z'_i))$.

For any $f \in F$, we define

$$W(f) = \sum_{i=1}^{n} \epsilon_i W_i(f),$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables. Based on Hoeffding’s inequality, we have for all $f, g \in F$ and any $\theta > 0$,

$$\Pr(|W(f) - W(g)| > \theta) \leq 2 \exp \left( -\frac{\theta^2}{2d_{s,s'}(f, g)} \right),$$

where the probability is taken over Rademacher variables conditional on $Z_i$ and $Z'_i$, and $d_{s,s'}(f, g) = \sqrt{\sum_{i=1}^{n} (W_i(f) - W_i(g))^2}$. Then by using Proposition 14 of [1], we have for all $\lambda > 0$, and a universal constant

$$E \left[ \exp \left( \lambda \sup_{f \in F} |W(f) - W(f^*)| \right) \right] \leq 2 \exp \left( \lambda^2 C^2 \gamma(F, d_{s,s'}(f, f^*))^2 / 4 \right),$$

(4)

By the definition of $d_{s,s'}(f, g)$, we have

$$d_{s,s'}(f, g) = \sqrt{\sum_{i=1}^{n} (W_i(f) - W_i(g))^2}$$

$$= \left( \frac{1}{n^2} \sum_{i=1}^{n} [\phi_\alpha(f(Z_i)) - \phi_\alpha(f(Z'_i)) - (\phi_\alpha(g(Z_i)) - \phi_\alpha(g(Z'_i)))]^2 \right)^{1/2}$$

$$\leq \frac{1}{n} \left( \sum_{i=1}^{n} [\phi_\alpha(f(Z_i)) - \phi_\alpha(g(Z_i))]^2 \right)^{1/2} + \frac{1}{n} \left( \sum_{i=1}^{n} [\phi_\alpha(f(Z'_i)) - \phi_\alpha(g(Z'_i))]^2 \right)^{1/2}$$

$$\leq \frac{1}{\sqrt{n}} \beta(F, \alpha) \left( \frac{1}{n} \sum_{i=1}^{n} [f(Z_i) - g(Z_i)]^2 \right)^{1/2} + \frac{1}{\sqrt{n}} \beta(F, \alpha) \left( \frac{1}{n} \sum_{i=1}^{n} [f(Z'_i) - g(Z'_i)]^2 \right)^{1/2},$$

where the second inequality uses the fact that $\phi_\alpha(x)$ is Lipschitz continuous. Thus, we have

$$\gamma(F, d_{s,s'}(f, g)) \leq \frac{1}{\sqrt{n}} \beta(F, \alpha) \gamma(F, d_{s}(f, g)) + \frac{1}{\sqrt{n}} \beta(F, \alpha) \gamma(F, d_{s'}(f, g)).$$

(5)

### D.1 Proof of Lemma 2

Proof. This proof is similar to the analysis in Theorem 7 from [1]. For completeness, we include it here. First, we assume $\Gamma_\delta \geq \Delta(F, d_c)$. Let $(Z'_1, \ldots, Z'_n)$ be an independent copies of $(Z_1, \ldots, Z_n)$, and we define

$$W_i(f) = \frac{1}{n} \phi_\alpha(f(Z_i)) - \frac{1}{n} \phi_\alpha(f(Z'_i)).$$

Then by setting $\alpha \geq \sqrt{n} \sigma^2/(2 \log(1/\delta))$, we get

$$P(f) - P(f^*) \leq O \left( \max(\Gamma_\delta, \Delta(F, d_c)) \sqrt{\frac{\log(8/\delta)}{n}} \right).$$
Then we have
\[
\Pr \left( \sup_{f \in F} |W(f) - W(f^*)| \geq \theta \right) \\
\leq \Pr \left( \sup_{f \in F} |W(f) - W(f^*)| \geq \theta \mid \gamma(F, d_x) \leq \Gamma_\delta \text{ and } \gamma(F, d_\epsilon) \leq \Gamma_\delta \right) + 2\Pr \left( \gamma(F, d_x) > \Gamma_\delta \right)
\]
\[
\leq \mathbb{E} \left[ \exp \left( \lambda \sup_{f \in F} |W(f) - W(f^*)| \right) \mid \gamma(F, d_x) \leq \Gamma_\delta \text{ and } \gamma(F, d_\epsilon) \leq \Gamma_\delta \right] \exp(-\lambda \theta) + 2\Pr \left( \gamma(F, d_x) > \Gamma_\delta \right)
\]
\[
\leq 2 \exp \left( \frac{\lambda^2 C^2 \Gamma_\delta^2}{n} \right) \exp(-\lambda \theta) + 2\Pr \left( \gamma(F, d_x) > \Gamma_\delta \right)
\]
\[
\leq 2 \exp \left( \frac{\lambda^2 C^2 \Gamma_\delta^2}{n} - \lambda \theta \right) + \frac{\delta}{4}
\]
where the second inequality uses Markov inequality, the third inequality uses the results of (4) and (5), where the last inequality is due to the definition of $\Gamma_\delta$ which satisfies $\Pr(\gamma(F, d_x) > \Gamma_\delta) \leq \delta/8$.

Let $\theta = 2C\Gamma_\delta \sqrt{\frac{\log(8/\delta)}{n}}$ and $\lambda = \frac{\sqrt{n \log(8/\delta)}}{C\Gamma_\delta}$ then
\[
\frac{\lambda^2 C^2 \Gamma_\delta^2}{n} - \lambda \theta = \frac{\lambda^2 C^2 \Gamma_\delta^2}{n} - 2C\Gamma_\delta \sqrt{\frac{\log(8/\delta)}{n}} \lambda = -\log(8/\delta).
\]

Therefore,
\[
\Pr \left( \sup_{f \in F} |W(f) - W(f^*)| \geq \theta \right) \leq \frac{\delta}{4} + \frac{\delta}{4} = \frac{\delta}{2}
\]

By Lemma 3.3 from [4], we get
\[
\Pr \left( \sup_{f \in F} |\Lambda(f) - \Lambda(f^*)| \geq 2\theta \right) \leq 2\Pr \left( \sup_{f \in F} |W(f) - W(f^*)| \geq \theta \right) \leq \delta
\]
and for any $f \in F$, $\Pr \left( \sup_{f \in F} |\Lambda(f) - \Lambda(f^*)| \geq \theta \right) \leq \frac{1}{2}$. On the other hand, by using $\mathbb{E}[\Lambda(f) - \Lambda(f^*)] = 0$ and Lipschitz continuous of $\phi_\alpha(x)$, we have
\[
\text{Var}(\Lambda(f) - \Lambda(f^*)) \leq \beta^2(\mathcal{F}, \alpha) \frac{\mathbb{E}[f(Z) - f^*(Z)]^2}{n\theta^2} \leq \beta^2(\mathcal{F}, \alpha) \frac{\Delta^2(\mathcal{F}, d_\epsilon)}{n\theta^2}.
\]

By applying Chebyshev’s inequality, it suffices to get
\[
\theta \geq \sqrt{2/n} \beta(\mathcal{F}, \alpha) \Delta(\mathcal{F}, d_\epsilon).
\]

If we assume $C > 1$ and choose $\delta < 1/3$, then $C\beta(\mathcal{F}, \alpha) \Gamma_\delta \sqrt{\frac{\log(8/\delta)}{n}} \geq \sqrt{2/n} \beta(\mathcal{F}, \alpha) \Delta^2(\mathcal{F}, d_\epsilon)$. Therefore, we get
\[
\Pr \left( \sup_{f \in F} |\Lambda(f) - \Lambda(f^*)| \geq 2C\beta(\mathcal{F}, \alpha) \Gamma_\delta \sqrt{\frac{\log(8/\delta)}{n}} \right) \leq \delta
\]

We can get the similar result for $\Gamma_\delta < \Delta(\mathcal{F}, d_\epsilon)$ instead of $\Gamma_\delta$ by using the similar analysis. We then complete the proof.

\section*{E Proof of Proposition 1}

\textbf{Proof.} Let define $z_i = w^T x_i - y_i$, then $\nabla F_\alpha(w) = \frac{1}{n} \sum_{i=1}^n \nabla w(\phi_\alpha(z_i^2/2)) = \frac{1}{n} \sum_{i=1}^n \phi'_\alpha(z_i^2/2) z_i x_i$ and $\nabla^2 F_\alpha(w) = \frac{1}{n} \sum_{i=1}^n \nabla w(\phi'_\alpha(z_i^2/2) z_i x_i) = \frac{1}{n} \sum_{i=1}^n \phi''_\alpha(z_i^2/2) z_i^2 x_i x_i^T + \phi'_\alpha(z_i^2/2) x_i x_i^T$. By the assumptions,
there exists a constant $\kappa > 0$, such that $\|\nabla^2 F_\alpha(w)\| \leq (\kappa + 1)R^2$, indicating that $F_\alpha(w)$ has a $(\kappa + 1)R^2$-Lipschitz continuous gradient. Then we have

$$F_\alpha(w_{t+1}) \leq F_\alpha(w_t) + \nabla F_\alpha(w_t)\top (w_{t+1} - w_t) + \frac{(\kappa + 1)R^2}{2}\|w_{t+1} - w_t\|^2$$

$$= F_\alpha(w_t) - \eta_t \nabla F_\alpha(w_t)\top \phi_\alpha((w_t\top x_i - y_i)^2) + \frac{(\kappa + 1)R^2\eta_t^2}{2}\|\phi_\alpha((w_t\top x_i - y_i)^2)\|^2$$

$$+ \frac{(\kappa + 1)R^2\eta_t^2}{2}\nabla \phi_\alpha((w_t\top x_i - y_i)^2) - \nabla F_\alpha(w_t) + \nabla F_\alpha(w_t)\parallel^2$$

Taking expectation on both sides we have

$$E[F_\alpha(w_{t+1}) - F_\alpha(w_t)] \leq \frac{(\kappa + 1)R^2\eta_t^2 - 2\eta_t E[\|\nabla F_\alpha(w_t)\|^2] + \frac{(\kappa + 1)R^2\eta_t^2\sigma_\alpha}{2}}{2}$$

where the last inequality uses the fact that $\eta_t \leq \frac{1}{(\kappa + 1)R^2}$. Summing up $t$ over $1, \ldots, T$, we have

$$\sum_{t=1}^{T} \eta_t E[\|\nabla F_\alpha(w_t)\|^2] \leq 2(F_\alpha(w_1) - F_\alpha(w_\epsilon)) + \sum_{t=1}^{T} (\kappa + 1)R^2\eta_t^2\sigma_\alpha.$$  \hspace{1cm} (6)

By setting $\eta_t = \frac{1}{(\kappa + 1)R^2\sqrt{T}}$, we have

$$E[\|\nabla F_\alpha(w_t)\|^2] \leq \frac{2(\kappa + 1)R^2(F_\alpha(w_1) - F_\alpha(w_\epsilon)) + \sigma_\alpha}{\sqrt{T}},$$  \hspace{1cm} (7)

where $R$ is a uniform random variable supported on $\{1, \ldots, T\}$. To achieve an approximate stationary point $E[\|\nabla F_\alpha(w_t)\|^2] \leq \epsilon^2$, the iteration complexity is $T = O(\sigma_\alpha^2/\epsilon^4)$.

**Remark.** The condition of $|x^2 \phi_\alpha''(x^2/2)| \leq \kappa$ for three different truncation functions presented in Preliminaries subsection can be easily checked. Example 1: $|x^2 \phi_\alpha''(1)(x^2/2)| = \frac{-x^7/\alpha}{(1+x^2/(2\alpha))^7} \leq 1$; Example 2: $|x^2 \phi_\alpha''(2)(x^2/2)| = \frac{x^7/(2\alpha)^7 + x^7/(8\alpha)^7 + x^7/(64\alpha)^7}{1 + x^2/(2\alpha)^2 + x^2/(8\alpha)^2 + x^2/(64\alpha)^2} \leq 1$; Example 3: $|x^2 \phi_\alpha''(x^2/2)| = \frac{2x^2(1-x^2/(2\alpha))}{\alpha x^2} \leq 1$ when $0 \leq x^2/2 \leq \alpha$, otherwise $|x^2 \phi_\alpha''(x^2/2)| = 0$.

**F Proof of Theorem 5**

**Proof.** We will use the following lemma in our proof.

**Lemma 3.** \cite{5} Under the assumption of Theorem 5, the following inequality holds for any $w_1, w_2 \in \{w : \|w - w_\epsilon\|_2 \leq \tau\}$ with probability $1 - c\exp(c'\log d)$,

$$\langle \nabla F_\alpha(w_1) - \nabla F_\alpha(w_2) \rangle \top (w_1 - w_2) \geq \frac{\alpha_T \lambda_{\min}(\Sigma_2)}{16}\|w_1 - w_2\|^2_2 - \tau \frac{\log(d)}{n}\|w_1 - w_2\|^2_1,$$  \hspace{1cm} (8)

where $\alpha_T := \min_{|u| \leq T} \ell''(u) > 0$, $\tau = \frac{C(\alpha + \kappa)^2\sigma^2 T^2}{r^2}$, and $\kappa$ satisfies $\ell''(u) \geq -\kappa_2$ for all $u$.

Then let's start our proof by setting $\ell(u) := \phi_\alpha(u^2/2) = \alpha \log(1 + u^2/(2\alpha))$. It is easy to show that $|\ell''(u)| = \frac{|u^2/(2\alpha)|}{(1 + u^2/(2\alpha))^2} \leq \frac{\sqrt{2}\alpha}{2}$ and $\phi_\alpha''(u) = \frac{1 - u^2/(2\alpha)}{(1 + u^2/(2\alpha))^2} \geq -\frac{1}{8}$, then $\kappa_2 \leq \frac{1}{8}$. Let $\tau \leq \sqrt{2}\alpha/2$, then $\alpha_T \geq \frac{12}{25}$. Then

$$\langle \nabla F_\alpha(w_\alpha) - \nabla F_\alpha(w_\epsilon) \rangle \top (w_\alpha - w_\epsilon) \geq a\|w_\alpha - w_\epsilon\|^2_2 - \tau \frac{\log(d)}{n}\|w_\alpha - w_\epsilon\|^2_1,$$  \hspace{1cm} (9)
where \( a = \frac{3\lambda_{\text{min}}(\Sigma)}{100} \) and \( \tau = \frac{C\sigma^2 r^2}{\mu^2} \) and \( C \) is a constant. Suppose SGD returns an approximate stationary point \( w_\alpha \) such that \( \| w_\alpha - w_* \|_2 \leq \tau \) and \( \| \nabla F_\alpha(w_\alpha) \|_2 \leq \epsilon \). Since \( w_\alpha \) is a stationary point and \( w_* \) is feasible, we have
\[
\nabla F_\alpha(w_\alpha)^T(w_* - w_\alpha) \geq -\epsilon \| w_* - w_\alpha \|_2
\] (10)

By Proposition 1 of [5], we have
\[
\nabla F_\alpha(w_\alpha)^T(w_* - w_\alpha) \geq -c \frac{\sqrt{2\alpha}}{2} \sigma_x \sqrt{\log(d)/n} \| w_* - w_\alpha \|_1
\] (11)

Combining inequalities (9) (10) and (11), we have
\[
a \| w_\alpha - w_* \|_2^2 \leq \epsilon \| w_* - w_\alpha \|_2 + c \frac{\sqrt{2\alpha}}{2} \sigma_x \sqrt{\log(d)/n} \| w_* - w_\alpha \|_1 + \tau \frac{\log(d)}{n} \| w_* - w_\alpha \|_2
\]
\[
\leq \epsilon \| w_* - w_\alpha \|_2 + c \frac{\sqrt{2\alpha}}{2} \sigma_x \sqrt{d \log(d)/n} \| w_* - w_\alpha \|_1 + \tau \frac{d \log(d)}{n} \| w_* - w_\alpha \|_2
\]
\[
\leq \epsilon \| w_* - w_\alpha \|_2 + c \frac{\sqrt{2\alpha}}{2} \sigma_x \sqrt{d \log(d)/n} \| w_* - w_\alpha \|_1 + \tau \frac{d \log(d)}{n} \| w_* - w_\alpha \|_2
\]

Then we get
\[
\| w_\alpha - w_* \|_2 \leq O \left( \sqrt{\frac{\alpha d \log d}{n}} + \frac{T^2 d \log d}{rn} + \epsilon \right)
\]

\[ \square \]

**G Proof of Proposition 2**

Proof. For simplicity, let \( \ell(w) = \ell(w; x, y) \). By the definition of truncation function, we know that \( \phi_\alpha(x) \) is smooth, i.e., for any \( w, v \in \mathbb{R}^d \), there exists a constant \( L_\alpha \) such that \( \phi_\alpha(\ell(v)) + \phi'_\alpha(\ell(v))(\ell(w) - \ell(v)) - \frac{L_\alpha}{2} |\ell(w) - \ell(v)|^2 \leq \phi_\alpha(\ell(w)) \). Since \( \ell \) is convex, i.e. for any \( w, v \in \mathbb{R}^d \), \( \ell(w) \geq \ell(v) + \ell(v)^T (w - v) \), then
\[
\phi_\alpha(\ell(w)) - \phi_\alpha(\ell(v)) \geq \phi'_\alpha(\ell(v)) \ell(v)^T (w - v) - \frac{L_\alpha}{2} |\ell(w) - \ell(v)|^2
\]
\[
\geq \phi'_\alpha(\ell(v)) \ell(v)^T (w - v) - \frac{G^2 L_\alpha}{2} \| w - v \|^2
\]
where the first inequality uses \( \phi'_\alpha(\ell(v)) \geq 0 \); the second inequality uses the fact that \( \| \ell(w; x_i, y_i) \| \leq G \). That is, \( F_\alpha(w) \) is \( G^2 L_\alpha \)-weakly convex. Finally, by employing the result of Theorem 2.1 from [3], we can complete the proof. \[ \square \]

**References**


