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 trunction 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 $\phi''^{(1)}_\alpha(x) = -\frac{1/\alpha}{(1+x/\alpha)^2}$, 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 $\phi'^{(2)}_\alpha(x) = \frac{1+\frac{x}{\alpha}}{1+\frac{x}{\alpha}+\frac{x^2}{2\alpha^2}} = 1 - \frac{1}{1+2\alpha/x+2\alpha^2/x^2}$. 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 $\phi''^{(2)}_\alpha(x) = -\frac{1}{\alpha}\frac{\frac{x}{\alpha}+\frac{x^2}{2\alpha^2}}{(1+\frac{x}{\alpha}+\frac{x^2}{2\alpha^2})^2}$, then $|\phi''^{(2)}_\alpha(x)| \leq 1/\alpha$, indicating that it satisfies condition (i).

Example 3.

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

Then we have

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

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

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

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

Next, we will verify the conditions $|x - \phi^{(1)}_\alpha(x)| \leq \frac{Mx^2}{\alpha}$, $|x - \phi^{(2)}_\alpha(x)| \leq \frac{Mx^2}{\alpha}$, and $|x - \phi^h_\alpha(x)| \leq \frac{Mx^2}{\alpha}$.

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

$$|x - \phi^{(1)}_\alpha(x)| \leq \frac{x^2}{2\alpha} \textit{ and } |x - \phi^{(2)}_\alpha(x)| \leq \frac{x^2}{2\alpha}. \tag{1}$$

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

$$\exp(y) \geq 1 + y + \frac{y^2}{2} \text{ for all } y \geq 0. \tag{2}$$

Let's first condsider $\phi^{(1)}_\alpha(x)$, to prove $|x - \alpha\log(1+x/\alpha)| \leq \frac{1}{2\alpha}x^2$, we have to show $|x/\alpha - \log(1+x/\alpha)| \leq \frac{1}{2\alpha^2}x^2$. 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 $f'(y) = -\frac{y^2}{1+y} \leq 0$, then we know $f(y)$ is a decreasing function on $y \geq 0$ thus $f(y) \leq f(0) = 0$, which give the first inequality in (3).

Next let's consider $\phi^{(2)}_\alpha(x)$. 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). □

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

$$|x - \phi^h_\alpha(x)| \leq \frac{x^2}{\alpha}, \tag{3}$$

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

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

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

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*

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

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

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

$$\begin{aligned}
P(\widehat{f}) - P(f^*) =& [P(\widehat{f}) - P(\phi_\alpha(\widehat{f}))] + [P(\phi_\alpha(\widehat{f})) - P_n(\phi_\alpha(\widehat{f}))] + [P_n(\phi_\alpha(\widehat{f})) - P_n(\phi_\alpha(f^*))] \\
&+ [P_n(\phi_\alpha(f^*)) - P(\phi_\alpha(f^*))] + [P(\phi_\alpha(f^*)) - P(f^*)] \\
\leq& [P(\widehat{f}) - P(\phi_\alpha(\widehat{f}))] + [P(\phi_\alpha(\widehat{f})) - P_n(\phi_\alpha(\widehat{f}))] + [P_n(\phi_\alpha(f^*)) - P(\phi_\alpha(f^*))] \\
&+ [P(\phi_\alpha(f^*)) - P(f^*)] \\
\leq& [P(\phi_\alpha(\widehat{f})) - P_n(\phi_\alpha(\widehat{f}))] + [P_n(\phi_\alpha(f^*)) - P(\phi_\alpha(f^*))] + \frac{2M\sigma^2}{\alpha}.
\end{aligned}$$

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

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

□

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 $f, f' \in \mathcal{F}$, we first know that $n(\Lambda(f) - \Lambda(f'))$ is the summation of the following independent random variables with zero mean:

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

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

$$\sum_{i=1}^n \mathrm{E}[C_i(f,f')^2] \leq \sum_{i=1}^n \mathrm{E}[(\phi_\alpha(f(Z_i)) - \phi_\alpha(f'(Z_i)))^2] \leq n\beta^2(\mathcal{F},\alpha)d_e^2(f,f').$$

Then by using Bernstein's inequality we have for any $f, f' \in \mathcal{F}$ and $\theta > 0$,

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

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. □

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 ℓ 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 $d_m(f,f') \leq Ld_m(h,h')$, where $f = \ell(h(\cdot),\cdot)$ and $f' = \ell(h'(\cdot),\cdot)$. Hence an ϵ/L-cover of $\mathcal{H}$ under the metric d_m induces an ϵ-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. □

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*

$$\sup_{f\in\mathcal{F}} |\Lambda(f) - \Lambda(f^*)| \leq C\beta(\mathcal{F},\alpha)\max(\Gamma_\delta, \Delta(\mathcal{F},d_e))\sqrt{\frac{\log\left(\frac{8}{\delta}\right)}{n}},$$

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

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

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

$$P(\widehat{f}) - P(f^*) \leq C\beta(\mathcal{F},\alpha)\max(\Gamma_\delta, \Delta(\mathcal{F},d_e))\sqrt{\frac{\log\left(\frac{8}{\delta}\right)}{n}} + \frac{2M\sigma^2}{\alpha}.$$

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

$$P(\widehat{f}) - P(f^*) \leq O\left(\max(\Gamma_\delta, \Delta(\mathcal{F}, d_e))\sqrt{\frac{\log(8/\delta)}{n}}\right).$$

□

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 $\Gamma_\delta \geq \Delta(\mathcal{F}, d_e)$. Let $(Z'_1, \ldots, Z'_n)$ be an independent copies of $(Z_1, \ldots, Z_n)$, and we define

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

For any $f \in \mathcal{F}$, we define

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

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher random variables. Based on Hoeffding's inequality, we have for all $f, g \in \mathcal{F}$ 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 C

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

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

$$\begin{aligned}
d_{s,s'}(f,g) &= \sqrt{\sum_{i=1}^{n}(W_i(f) - W_i(g))^2} \\
&= \left(\frac{1}{n^2}\sum_{i=1}^{n}[\phi_\alpha(f(Z_i)) - \phi_\alpha(f(Z'_i)) - (\phi_\alpha(g(Z_i)) - \phi_\alpha(g(Z'_i)))]^2\right)^{\frac{1}{2}} \\
&\leq \frac{1}{n}\left(\sum_{i=1}^{n}[\phi_\alpha(f(Z_i)) - \phi_\alpha(g(Z_i))]^2\right)^{\frac{1}{2}} + \frac{1}{n}\left(\sum_{i=1}^{n}[\phi_\alpha(f(Z'_i)) - \phi_\alpha(g(Z'_i))]^2\right)^{\frac{1}{2}} \\
&\leq \frac{1}{\sqrt{n}}\beta(\mathcal{F}, \alpha)\left(\frac{1}{n}\sum_{i=1}^{n}[f(Z_i) - g(Z_i)]^2\right)^{\frac{1}{2}} + \frac{1}{\sqrt{n}}\beta(\mathcal{F}, \alpha)\left(\frac{1}{n}\sum_{i=1}^{n}[f(Z'_i) - g(Z'_i)]^2\right)^{\frac{1}{2}},
\end{aligned}$$

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

$$\gamma(\mathcal{F}, d_{s,s'}(f,g)) \leq \frac{1}{\sqrt{n}}\beta(\mathcal{F}, \alpha)\gamma(\mathcal{F}, d_s(f,g)) + \frac{1}{\sqrt{n}}\beta(\mathcal{F}, \alpha)\gamma(\mathcal{F}, d_{s'}(f,g)). \tag{5}$$

Then we have

$$
\begin{aligned}
&\Pr\left(\sup_{f\in\mathcal{F}}|W(f)-W(f^*)|\geq\theta\right)\\
\leq&\Pr\left(\sup_{f\in\mathcal{F}}|W(f)-W(f^*)|\geq\theta\mid\gamma(\mathcal{F},d_s)\leq\Gamma_\delta\ \text{ and}\ \gamma(\mathcal{F},d_{s'})\leq\Gamma_\delta\right)+2\Pr\left(\gamma(\mathcal{F},d_s)>\Gamma_\delta\right)\\
\leq&\mathrm{E}\left[\exp\left(\lambda\sup_{f\in\mathcal{F}}|W(f)-W(f^*)|\right)\mid\gamma(\mathcal{F},d_s)\leq\Gamma_\delta\ \text{and}\ \gamma(\mathcal{F},d_{s'})\leq\Gamma_\delta\right]\exp(-\lambda\theta)+2\Pr\left(\gamma(\mathcal{F},d_s)>\Gamma_\delta\right)\\
\leq&2\exp\left(\lambda^2C^2\Gamma_\delta^2/n\right)\exp(-\lambda\theta)+2\Pr\left(\gamma(\mathcal{F},d_s)>\Gamma_\delta\right)\\
\leq&2\exp\left(\frac{\lambda^2C^2\Gamma_\delta^2}{n}-\lambda\theta\right)+\frac{\delta}{4}
\end{aligned}
$$

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 Γ_δ which satisfies $\Pr(\gamma(\mathcal{F},d_s)>\Gamma_\delta)\leq\delta/8$.

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

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

Therefore,

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

By Lemma 3.3 from [4], we get

$$
\Pr\left(\sup_{f\in\mathcal{F}}|\Lambda(f)-\Lambda(f^*)|\geq2\theta\right)\leq2\Pr\left(\sup_{f\in\mathcal{F}}|W(f)-W(f^*)|\geq\theta\right)\leq\delta
$$

and for any $f\in\mathcal{F}$, $\Pr\left(\sup_{f\in\mathcal{F}}|\Lambda(f)-\Lambda(f^*)|\geq\theta\right)\leq\frac{1}{2}$. On the other hand, by using $\mathrm{E}[\Lambda(f)-\Lambda(f^*)]=0$ and Lipschitz continuous of $\phi_\alpha(x)$, we have

$$
\frac{\mathrm{Var}(\Lambda(f)-\Lambda(f^*))}{\theta^2}\leq\beta^2(\mathcal{F},\alpha)\frac{\mathrm{E}[f(Z)-f^*(Z)]^2}{n\theta^2}\leq\beta^2(\mathcal{F},\alpha)\frac{\Delta^2(\mathcal{F},d_e)}{n\theta^2}.
$$

By applying Chebyshev's inequality, it suffices to get

$$
\theta\geq\sqrt{2/n}\beta(\mathcal{F},\alpha)\Delta(\mathcal{F},d_e).
$$

If we assume $C>1$ and choose $\delta<1/3$, then $C\beta(\mathcal{F},\alpha)\Gamma_\delta\sqrt{\frac{\log(8/\delta)}{n}}\geq\sqrt{2/n}\beta(\mathcal{F},\alpha)\Delta^2(\mathcal{F},d_e)$. Therefore, we get

$$
\Pr\left(\sup_{f\in\mathcal{F}}|\Lambda(f)-\Lambda(f^*)|\geq2C\beta(\mathcal{F},\alpha)\Gamma_\delta\sqrt{\frac{\log(8/\delta)}{n}}\right)\leq\delta
$$

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

E Proof of Proposition 1

Proof. Let define $z_i=\mathbf{w}^\top\mathbf{x}_i-y_i$, then $\nabla F_\alpha(\mathbf{w})=\frac{1}{n}\sum_{i=1}^n\nabla_{\mathbf{w}}(\phi_{(\alpha)}(z_i^2/2))=\frac{1}{n}\sum_{i=1}^n\phi'_{(\alpha)}(z_i^2/2)z_i\mathbf{x}_i$ and $\nabla^2F_\alpha(\mathbf{w})=\frac{1}{n}\sum_{i=1}^n\nabla_{\mathbf{w}}(\phi'_{(\alpha)}(z_i^2/2)z_i\mathbf{x}_i)=\frac{1}{n}\sum_{i=1}^n\phi''_{(\alpha)}(z_i^2/2)z_i^2\mathbf{x}_i\mathbf{x}_i^\top+\phi'_{(\alpha)}(z_i^2/2)\mathbf{x}_i\mathbf{x}_i^\top$. By the assumptions, there exists a constant $\kappa > 0$, such that $\|\nabla^2 F_\alpha(\mathbf{w})\| \le (\kappa+1)R^2$, indecating that $F_\alpha(\mathbf{w})$ has a $(\kappa+1)R^2$-Lipschitz continous gradient. Then we have

$$\begin{aligned}
F_\alpha(\mathbf{w}_{t+1}) \le& F_\alpha(\mathbf{w}_t) + \nabla F_\alpha(\mathbf{w}_t)^\top(\mathbf{w}_{t+1}-\mathbf{w}_t) + \frac{(\kappa+1)R^2}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2 \\
=& F_\alpha(\mathbf{w}_t) - \eta_t \nabla F_\alpha(\mathbf{w}_t)^\top \phi_\alpha((\mathbf{w}_t^\top \mathbf{x}_i - y_i)^2) + \frac{(\kappa+1)R^2\eta_t^2}{2}\|\phi_\alpha((\mathbf{w}_t^\top \mathbf{x}_i - y_i)^2)\|^2 \\
=& F_\alpha(\mathbf{w}_t) - \eta_t \nabla F_\alpha(\mathbf{w}_t)^\top \phi_\alpha((\mathbf{w}_t^\top \mathbf{x}_i - y_i)^2) \\
&+ \frac{(\kappa+1)R^2\eta_t^2}{2}\|\nabla\phi_\alpha((\mathbf{w}_t^\top \mathbf{x}_i - y_i)^2) - \nabla F_\alpha(\mathbf{w}_t) + \nabla F_\alpha(\mathbf{w}_t)\|^2
\end{aligned}$$

Taking expectation on both sides we have

$$\begin{aligned}
\mathrm{E}[F_\alpha(\mathbf{w}_{t+1}) - F_\alpha(\mathbf{w}_t)] \le& \frac{(\kappa+1)R^2\eta_t^2 - 2\eta_t}{2}\mathrm{E}[\|\nabla F_\alpha(\mathbf{w}_t)\|^2] + \frac{(\kappa+1)R^2\eta_t^2\sigma_\alpha}{2} \\
\le& -\frac{\eta_t}{2}\mathrm{E}[\|\nabla F_\alpha(\mathbf{w}_t)\|^2] + \frac{(\kappa+1)R^2\eta_t^2\sigma_\alpha}{2},
\end{aligned}$$

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

$$\sum_{t=1}^{T}\eta_t \mathrm{E}[\|\nabla F_\alpha(\mathbf{w}_t)\|^2] \le 2(F_\alpha(\mathbf{w}_1) - F_\alpha(\mathbf{w}_*)) + \sum_{t=1}^{T}(\kappa+1)R^2\eta_t^2\sigma_\alpha. \tag{6}$$

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

$$\mathrm{E}_R[\mathrm{E}[\|\nabla F_\alpha(\mathbf{w}_t)\|^2]] \le \frac{2(\kappa+1)R^2(F_\alpha(\mathbf{w}_1) - F_\alpha(\mathbf{w}_*))}{\sqrt{T}} + \frac{\sigma_\alpha}{\sqrt{T}}, \tag{7}$$

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

Remark. The condition of $\left|x^2\phi_\alpha''(x^2/2)\right| \le \kappa$ for three different truncation functions presented in Preliminaries subsection can be easily checked. Example 1: $\left|x^2\phi_\alpha''^{(1)}(x^2/2)\right| = \left|-\frac{x^2/\alpha}{(1+x^2/(2\alpha))^2}\right| = \frac{x^2/\alpha}{1+x^2/\alpha+x^4/(2\alpha)^4} \le 1$; Example 2: $\left|x^2\phi_\alpha''^{(2)}(x^2/2)\right| = \left|\frac{x^4/(2\alpha^2)+x^6/(8\alpha^3)}{(1+x^2/(2\alpha)+x^4/(8\alpha^2))^2}\right| = \frac{x^4/(2\alpha^2)+x^6/(8\alpha^3)}{1+x^2/\alpha+x^4/(2\alpha^2)+x^6/(8\alpha^3)+x^8/(64\alpha^4)} \le 1$; Example 3: $\left|x^2\phi_\alpha''^{h}(x^2/2)\right| = \left|\frac{2x^2(1-x^2/(2\alpha))}{\alpha}\right| = \frac{(2\alpha-x^2)x^2}{\alpha^2} \le 1$ when $0 \le x^2/2 \le \alpha$, otherwise $\left|x^2\phi_\alpha''^{h}(x^2/2)\right| = 0$.

F Proof of Theorem 5

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

Lemma 3. *[5] Under the assumption of Theorem 5, the following inequality holds for any* $\mathbf{w}_1, \mathbf{w}_2 \in \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}_*\|_2 \le r\}$ *with probability* $1 - c\exp(c'\log d)$*,*

$$(\nabla F_\alpha(\mathbf{w}_1) - \nabla F_\alpha(\mathbf{w}_2))^\top(\mathbf{w}_1 - \mathbf{w}_2) \ge \frac{\alpha_T\lambda_{min}(\Sigma_x)}{16}\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2 - \tau\frac{\log(d)}{n}\|\mathbf{w}_1 - \mathbf{w}_2\|_1^2, \tag{8}$$

where $\alpha_T := \min_{|u|\le T}\ell''(u) > 0$, $\tau = \frac{C(\alpha_T+\kappa_2)^2\sigma_x^2T^2}{r^2}$*, and* κ_2 *satisfies* $\ell''(u) \ge -\kappa_2$ *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 $|\ell'(u)| = |\frac{u}{1+u^2/(2\alpha)}| \le \frac{\sqrt{2\alpha}}{2}$ and $\phi_\alpha''(u) = \frac{1-u^2/(2\alpha)}{(1+u^2/(2\alpha))^2} \ge -\frac{1}{8}$, then $\kappa_2 = \frac{1}{8}$. Let $T \le \sqrt{2\alpha}/2$, then $\alpha_T = \frac{12}{25}$. Then

$$(\nabla F_\alpha(\mathbf{w}_\alpha) - \nabla F_\alpha(\mathbf{w}_*))^\top(\mathbf{w}_\alpha - \mathbf{w}_*) \ge a\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2^2 - \tau\frac{\log(d)}{n}\|\mathbf{w}_\alpha - \mathbf{w}_*\|_1^2, \tag{9}$$

where $a = \frac{3\lambda_{\min}(\Sigma_x)}{100}$ and $\tau = \frac{C\sigma_x^2 T^2}{r^2}$ and C is a constant. Suppose SGD returns an approximate stationary point $\mathbf{w}_\alpha$ such that $\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2 \leq r$ and $\|\nabla F_\alpha(\mathbf{w}_\alpha)\|_2 \leq \epsilon$. Since $\mathbf{w}_\alpha$ is a stationary point and $\mathbf{w}_*$ is feasible, we have

$$\nabla F_\alpha(\mathbf{w}_\alpha)^\top(\mathbf{w}_* - \mathbf{w}_\alpha) \geq -\epsilon\|\mathbf{w}_* - \mathbf{w}_\alpha\|_2 \tag{10}$$

By Proposition 1 of [5], we have

$$\nabla F_\alpha(\mathbf{w}_*)^\top(\mathbf{w}_\alpha - \mathbf{w}_*) \geq -c\frac{\sqrt{2\alpha}}{2}\sigma_x\sqrt{\log(d)/n}\|\mathbf{w}_\alpha - \mathbf{w}_*\|_1 \tag{11}$$

Combining inequalities (9) (10) and (11), we have

$$\begin{aligned}
a\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2^2 \leq& \epsilon\|\mathbf{w}_* - \mathbf{w}_\alpha\|_2 + c\frac{\sqrt{2\alpha}}{2}\sigma_x\sqrt{\log(d)/n}\|\mathbf{w}_\alpha - \mathbf{w}_*\|_1 + \tau\frac{\log(d)}{n}\|\mathbf{w}_\alpha - \mathbf{w}_*\|_1^2 \\
\leq& \epsilon\|\mathbf{w}_* - \mathbf{w}_\alpha\|_2 + c\frac{\sqrt{2\alpha}}{2}\sigma_x\sqrt{d\log(d)/n}\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2 + \tau\frac{d\log(d)}{n}\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2^2 \\
\leq& \epsilon\|\mathbf{w}_* - \mathbf{w}_\alpha\|_2 + c\frac{\sqrt{2\alpha}}{2}\sigma_x\sqrt{d\log(d)/n}\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2 + \tau r\frac{d\log(d)}{n}\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2
\end{aligned}$$

Then we get

$$\|\mathbf{w}_\alpha - \mathbf{w}_*\|_2 \leq O\left(\sqrt{\frac{\alpha d\log d}{n}} + \frac{T^2 d\log d}{rn} + \epsilon\right)$$

□

G Proof of Proposition 2

Proof. For similicity, let $\ell(\mathbf{w}) = \ell(\mathbf{w}; \mathbf{x}, \mathbf{y})$. By the defination of truncation function, we know that $\phi_\alpha(x)$ is smooth, i.e., for any $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, there exists a constant L_α such that $\phi_\alpha(\ell(\mathbf{v})) + \phi'_\alpha(\ell(\mathbf{v}))(\ell(\mathbf{w}) - \ell(\mathbf{v})) - \frac{L_\alpha}{2}|\ell(\mathbf{w}) - \ell(\mathbf{v})|^2 \leq \phi_\alpha(\ell(\mathbf{w}))$. Since ℓ is convex, i.e. for any $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, $\ell(\mathbf{w}) \geq \ell(\mathbf{v}) + \partial\ell(\mathbf{v})^\top(\mathbf{w} - \mathbf{v})$, then

$$\begin{aligned}
\phi_\alpha(\ell(\mathbf{w})) - \phi_\alpha(\ell(\mathbf{v})) \geq& \phi'_\alpha(\ell(\mathbf{v}))\partial\ell(\mathbf{v})^\top(\mathbf{w} - \mathbf{v}) - \frac{L_\alpha}{2}|\ell(\mathbf{w}) - \ell(\mathbf{v})|^2 \\
\geq& \phi'_\alpha(\ell(\mathbf{v}))\partial\ell(\mathbf{v})^\top(\mathbf{w} - \mathbf{v}) - \frac{G^2 L_\alpha}{2}\|\mathbf{w} - \mathbf{v}\|^2
\end{aligned}$$

where the first inequality uses $\phi'_\alpha(\ell(\mathbf{v})) \geq 0$; the second inequality uses the fact that $\|\partial\ell(\mathbf{w}; \mathbf{x}_i, y_i)\| \leq G$. That is, $F_\alpha(\mathbf{w})$ is $G^2 L_\alpha$-weakly convex. Finally, by employing the result of Theorem 2.1 from [3], we can complete the proof. □

References

[1] C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015.

[2] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

[3] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[4] S. A. Geer. *Applications of empirical process theory*. Cambridge University Press, 2000.

[5] P.-L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 2017.