Appendix for

Investigating the Role of Weight Decay in Enhancing Nonconvex SGD

Contents

A. Technical Lemmas

Lemma 3 *Assume $(\boldsymbol{w}^t)_{t\geq 0}$ be generated by SGDW, as $0<\alpha<1$, we have*

$$\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{w}^t\|^2 \leq \frac{\gamma^2(1-\alpha^T)}{(1-\alpha)^2}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|^2 + \frac{\gamma^2 T(1-\alpha^T)}{1-\alpha}\sigma^2,$$

Lemma 4 *Let $(\boldsymbol{w}^t)_{t\geq 0}$ be generated by SignSGDW, it holds that*

$$\|\boldsymbol{w}^t\| \leq \frac{1-\alpha^T}{1-\alpha}\gamma\sqrt{d} \leq \frac{1}{1-\alpha}\gamma\sqrt{d},\ \|\boldsymbol{w}^t\|_\infty \leq \frac{1-\alpha^T}{1-\alpha}\gamma,$$

provided $0 \leq t \leq T$.

Lemma 5 *Let $\boldsymbol{x}^\dagger, \boldsymbol{m} \in \mathbb{R}^d$ be arbitrary vectors such that $\|\boldsymbol{x}^\dagger\| \leq \frac{1}{1-\alpha}\gamma\sqrt{d},\ \|\boldsymbol{x}^\dagger\|_\infty \leq \frac{1-\alpha^T}{1-\alpha}\gamma$, and*

$$\boldsymbol{x}^\ddagger = \alpha\boldsymbol{x}^\dagger - \gamma \text{Sign}(\boldsymbol{m}), \tag{3}$$

and $\boldsymbol{\epsilon} := \boldsymbol{m} - \nabla f_S(\boldsymbol{x}^\dagger)$. If $|\langle \nabla f_S(\boldsymbol{x}^\dagger), \boldsymbol{x}^\dagger\rangle| \leq \delta\|\nabla f_S(\boldsymbol{x}^\dagger)\|_1 \cdot \|\boldsymbol{x}^\dagger\|_\infty$, then we have

$$f_S(\boldsymbol{x}^\ddagger) - f_S(\boldsymbol{x}^\dagger) \leq -\gamma[1-(1-\alpha^T)\delta]\|\nabla f_S(\boldsymbol{x}^\dagger)\|_1 + 2\sqrt{d}\gamma\|\boldsymbol{\epsilon}\| + 2Ld\gamma^2.$$

Lemma 6 *Let Assumption 1, 2, and 3 hold, and let $(\boldsymbol{x}^t)_{t\geq 1}$ and $(\boldsymbol{y}^k)_{t\geq 1}$ be the output of SignSGDW on these two neighbor sets of n samples with the same initial points. Then:*

1) In the general case, it holds that

$$\sup_{\xi\in\Xi}\sqrt{\mathbb{E}_{\mathcal{A}_{\text{SignSGDW}}}\|\nabla f(\mathcal{A}_{\text{SignSGDW}}(S);\xi) - \nabla f(\mathcal{A}_{\text{SignSGDW}}(S');\xi)\|^2} = \mathcal{O}\Big(\frac{T^{3/4}}{\sqrt{n}}\Big).$$

2) If $\delta < 1$ and $\gamma = \frac{1}{L\sqrt{d}T^{3/4}}, \alpha = 1 - \frac{1}{T^{1/4}}$, then we have

$$\sup_{\xi\in\Xi}\sqrt{\mathbb{E}_{\mathcal{A}_{\text{SignSGDW}}}\|\nabla f(\mathcal{A}_{\text{SignSGDW}}(S);\xi) - \nabla f(\mathcal{A}_{\text{SignSGDW}}(S');\xi)\|^2} = \mathcal{O}\Big(\frac{L\gamma\sqrt{d}T^{3/4}}{\sqrt{n}}\Big) = \mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big).$$

B. Proof of Theorem 1

The Lipschitz property gives us

$$\begin{aligned}
f_S(\boldsymbol{w}^{t+1}) &\leq f_S(\boldsymbol{w}^t) + \langle \nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^{t+1} - \boldsymbol{w}^t\rangle + \frac{L^2}{2}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \\
&\leq f_S(\boldsymbol{w}^t) + \langle \nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^{t+1} - \alpha\boldsymbol{w}^t + (\alpha-1)\boldsymbol{w}^t\rangle + \frac{L^2}{2}\|\boldsymbol{w}^{t+1} - \alpha\boldsymbol{w}^t + (\alpha-1)\boldsymbol{w}^t\|^2 \\
&= f_S(\boldsymbol{w}^t) + \langle \nabla f_S(\boldsymbol{w}^t), (\alpha-1)\boldsymbol{w}^t - \gamma\boldsymbol{g}^t\rangle + \frac{L^2}{2}\|(\alpha-1)\boldsymbol{w}^t - \gamma\boldsymbol{g}^t\|^2 \\
&= f_S(\boldsymbol{w}^t) - \gamma\langle \boldsymbol{g}^t, \nabla f_S(\boldsymbol{w}^t)\rangle - (1-\alpha)\langle \nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^t\rangle + \frac{L^2\gamma^2}{2}\|\boldsymbol{g}^t\|^2 \\
&\quad + (1-\alpha)\gamma L^2\langle \boldsymbol{g}^t, \boldsymbol{w}^t\rangle + \frac{(1-\alpha)^2L^2}{2}\|\boldsymbol{w}^t\|^2.
\end{aligned}$$

Taking the expectation of both sides over χ^t, we get

$$\begin{aligned}
\mathbb{E}\Big[f_S(\boldsymbol{w}^{t+1}) \mid \chi^t\Big] &\leq f_S(\boldsymbol{w}^t) - \gamma\|\nabla f_S(\boldsymbol{w}^t)\|^2 - (1-\alpha)(1-\gamma L^2)\langle \nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^t\rangle \\
&\quad + \frac{(1-\alpha)^2L^2}{2}\|\boldsymbol{w}^t\|^2 + \frac{L^2\gamma^2}{2}\mathbb{E}\Big[\|\boldsymbol{g}^t\|^2 \mid \chi^t\Big] \\
&\leq f_S(\boldsymbol{w}^t) - \gamma(1-\frac{L\gamma}{2})\|\nabla f_S(\boldsymbol{w}^t)\|^2 + (1-\alpha)(1-\gamma L^2)\underbrace{\langle -\nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^t\rangle}_{(\dagger)} \\
&\quad + \frac{(1-\alpha)^2L^2}{2}\|\boldsymbol{w}^t\|^2 + \frac{L^2\gamma^2\sigma^2}{2}.
\end{aligned} \tag{4}$$

I. Using $\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \|\boldsymbol{a}\|\|\boldsymbol{b}\| \leq \frac{\|\boldsymbol{a}\|^2}{2\sqrt{1-\alpha}} + \frac{\sqrt{1-\alpha}\|\boldsymbol{b}\|^2}{2}$ for $(\dagger)$ in (4) as $\nabla f_S(\boldsymbol{w}^t) \to \boldsymbol{a}$ and $\boldsymbol{w}^t \to \boldsymbol{b}$, we get

$$\begin{aligned}\mathbb{E}\Big[f_S(\boldsymbol{w}^{t+1}) \mid \chi^t\Big] \leq f_S(\boldsymbol{w}^t) - \Big[\gamma(1-\frac{L\gamma}{2}) - \frac{\sqrt{1-\alpha}(1-\gamma L^2)}{2}\Big]\|\nabla f_S(\boldsymbol{w}^t)\|^2 \\ + \Big[\frac{(1-\alpha)^2L^2}{2} + \frac{(1-\alpha)^{3/2}(1-\gamma L^2)}{2}\Big]\|\boldsymbol{w}^t\|^2 + \frac{L^2\gamma^2\sigma^2}{2},\end{aligned} \tag{5}$$

We use some shorthand notation for constants as follows

$$C_1 := \gamma(1-\frac{L\gamma}{2}) - \frac{\sqrt{1-\alpha}(1-\gamma L^2)}{2},\ C_2 := \frac{(1-\alpha)^2L^2}{2} + \frac{(1-\alpha)^{3/2}(1-\gamma L^2)}{2}.$$

Taking full expectations on both sides of (5) results in

$$\mathbb{E}f_S(\boldsymbol{w}^{t+1}) \leq \mathbb{E}f_S(\boldsymbol{w}^t) - C_1\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|^2 + C_2\mathbb{E}\|\boldsymbol{w}^t\|^2 + \frac{L^2\gamma^2\sigma^2}{2}.$$

Summing the inequality from $t=1$ to T and using Lemma 3, we get

$$\Big[C_1 - C_2\frac{\gamma^2(1-\alpha^T)}{(1-\alpha)^2}\Big]\sum_{t=1}^{T}\|\nabla f_S(\boldsymbol{w}^t)\|^2 \leq f_S(\boldsymbol{w}^1) - \min f_S + C_2\frac{\gamma^2\sigma^2T}{1-\alpha} + \frac{L^2\gamma^2\sigma^2T}{2}.$$

Noticing that γ is very small and $1-\gamma^2 \leq \alpha < 1$, it follows

$$C_1 \geq \frac{\gamma}{2},\ C_2 \approx \frac{(1-\alpha)^{3/2}}{2}, \frac{1-\alpha^T}{1-\alpha} \approx T.$$

and

$$C_2\frac{\gamma^2(1-\alpha^T)}{2(1-\alpha)^2} \approx \frac{\gamma^2(1-\alpha^T)}{2\sqrt{1-\alpha}} = \frac{\gamma^2(1-\alpha^T)\sqrt{1-\alpha}}{2(1-\alpha)} \approx \frac{\gamma^2T\sqrt{1-\alpha}}{2} \leq \gamma/4.$$

Based on these approximations, as γ is small, we can see that

$$C_1 - C_2\frac{\gamma^2(1-\alpha^T)}{(1-\alpha)^2} \geq \frac{\gamma}{4}.$$

Therefore, we obtain the following bound for the rate

$$\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|^2}{T} = \mathcal{O}\Big(\frac{\mathbb{E}f_S(\boldsymbol{w}^1) - \min f_S}{\gamma T} + \frac{L^2\gamma\sigma^2}{2} + \sqrt{1-\alpha}\gamma\sigma^2\Big).$$

II. If $0 \leq \delta < 1$: Using $\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \delta_T\|\boldsymbol{a}\|\|\boldsymbol{b}\| \leq \frac{\delta_T\|\boldsymbol{a}\|^2}{2} + \frac{\delta_T\|\boldsymbol{b}\|^2}{2}$ for $(\dagger)$ in (4) as $\nabla f_S(\boldsymbol{w}^t) \to \boldsymbol{a}$ and $\boldsymbol{w}^t \to \boldsymbol{b}$, and using the fact that $\langle -\nabla f_S(\boldsymbol{w}^t), \boldsymbol{w}^t\rangle \leq \|\nabla f_S(\boldsymbol{w}^t)\| \cdot \|\boldsymbol{w}^t\|$, we get

$$\begin{aligned}\mathbb{E}\Big[f_S(\boldsymbol{w}^{t+1}) \mid \chi^t\Big] \leq f_S(\boldsymbol{w}^t) - \Big[\gamma(1-\frac{L\gamma}{2}) - \frac{\delta_T(1-\alpha)(1-\gamma L^2)}{2}\Big]\|\nabla f_S(\boldsymbol{w}^t)\|^2 \\ + \Big[\frac{(1-\alpha)^2L^2}{2} + \frac{\delta_T(1-\alpha)(1-\gamma L^2)}{2}\Big]\|\boldsymbol{w}^t\|^2 + \frac{L^2\gamma^2\sigma^2}{2},\end{aligned} \tag{6}$$

We use some shorthand notation for constants as follows

$$C_1 := \gamma(1-\frac{L\gamma}{2}) - \frac{\delta_T(1-\alpha)(1-\gamma L^2)}{2},\ C_2 := \frac{(1-\alpha)^2L^2}{2} + \frac{\delta_T(1-\alpha)(1-\gamma L^2)}{2}.$$

Taking full expectations on both sides of (4), we get

$$\mathbb{E}f_S(\boldsymbol{w}^{t+1}) \leq \mathbb{E}f_S(\boldsymbol{w}^t) - C_1\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|^2 + C_2\mathbb{E}\|\boldsymbol{w}^t\|^2 + \frac{L^2\gamma^2\sigma^2}{2}.$$

Summing the inequality from $t=1$ to T and using Lemma 3, we get

$$\Big[C_1 - C_2\frac{\gamma^2(1-\alpha^T)}{(1-\alpha)^2}\Big]\sum_{t=1}^{T}\|\nabla f_S(\boldsymbol{w}^t)\|^2 \le f_S(\boldsymbol{w}^1) - \min f_S + C_2\frac{\gamma^2(1-\alpha^T)\sigma^2 T}{1-\alpha} + \frac{L^2\gamma^2\sigma^2 T}{2}.$$

Noticing that γ is very small and $1-\alpha=\gamma$, it follows

$$C_1 \approx \gamma - \frac{\delta_T(1-\alpha)}{2} \ge \gamma - \frac{\delta_T\gamma}{2},\ C_2 \approx \frac{\delta_T(1-\alpha)}{2} \le \frac{\delta_T\gamma}{2},$$

and

$$C_2\frac{\gamma^2(1-\alpha^T)}{(1-\alpha)^2} \approx \delta_T\frac{\gamma^2(1-\alpha^T)}{2(1-\alpha)} \le \frac{\delta_T\gamma^2}{2(1-\alpha)} = \frac{\delta_T\gamma}{2}.$$

Therefore, we are led to

$$C_1 - C_2\frac{\gamma^2(1-\alpha^T)}{(1-\alpha)^2} \ge (1-\delta_T)\gamma.$$

Hence, we get

$$\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|^2}{T} = \mathcal{O}\Big(\frac{\mathbb{E}f_S(\boldsymbol{w}^1) - \min f_S}{(1-\delta_T)\gamma T} + \frac{(L^2+1)\gamma\sigma^2}{2(1-\delta_T)}\Big).$$

Combining the two cases above, we then complete the proof.

C. Proof of Theorem 2

With the Cauchy's inequality, it holds that

$$\begin{aligned}\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|}{T} &= \frac{\sum_{t=1}^{T}\mathbb{E}(1\cdot\|\nabla f_S(\boldsymbol{w}^t)\|)}{T}\\ &\le \frac{\sqrt{T\sum_{i=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|^2}}{T} = \frac{\sqrt{\sum_{i=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|^2}}{\sqrt{T}}.\end{aligned}$$

I. As $1-\alpha \le \frac{1}{2T}$, from Theorem 1, it holds that $\sum_{i=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|^2 = \mathcal{O}(\sqrt{T})$, indicating the following result

$$\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|}{T} = \mathcal{O}\Big(\frac{1}{T^{1/4}}\Big).$$

From Lemmas 2 and 1, we can see that

$$\mathbb{E}\|\nabla f(\boldsymbol{w}^t) - \nabla f_S(\boldsymbol{w}^t)\| = \mathcal{O}\Big(\sqrt{\frac{t}{n}}\Big) = \mathcal{O}\Big(\sqrt{\frac{T}{n}}\Big).$$

Based on the inequality above, we have

$$\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t) - \nabla f_S(\boldsymbol{w}^t)\|}{T} = \mathcal{O}\Big(\sqrt{\frac{T}{n}}\Big).$$

By the triangle inequality, we have

$$\begin{aligned}\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|}{T} &= \frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|}{T} + \frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t) - \nabla f_S(\boldsymbol{w}^t)\|}{T}\\ &= \mathcal{O}\Big(\frac{1}{T^{1/4}} + \sqrt{\frac{T}{n}}\Big).\end{aligned}$$

II. For the case $1-\alpha=\frac{1}{\sqrt{T}}$, from Theorem 1, we can get

$$\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|}{T}=\mathcal{O}\Big(\frac{1}{\sqrt{1-\delta_T}T^{1/4}}\Big).$$

In this case, we also have

$$\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)-\nabla f_S(\boldsymbol{w}^t)\|}{T}=\mathcal{O}\Big(\sqrt{\frac{\gamma T}{n}}\Big)=\mathcal{O}\Big(\frac{T^{1/4}}{\sqrt{n}}\Big).$$

Therefore, we can get the following generalization bound

$$\begin{aligned}\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|}{T}&=\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|}{T}+\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)-\nabla f_S(\boldsymbol{w}^t)\|}{T}\\&=\mathcal{O}\Big(\frac{1}{\sqrt{1-\delta_T}T^{1/4}}+\frac{T^{1/4}}{\sqrt{n}}\Big).\end{aligned}$$

Combination of the two cases above then completes the proof.

D. Proof of Theorem 3

Let

$$\boldsymbol{g}^t:=\nabla f(\boldsymbol{w}^t;\xi_{i_t}),\boldsymbol{\epsilon}^t:=\boldsymbol{m}^t-\nabla f_S(\boldsymbol{w}^t),\boldsymbol{\delta}^t:=\boldsymbol{g}^t-\nabla f_S(\boldsymbol{w}^t).$$

Through direct computations, we obtain

$$\begin{aligned}&\boldsymbol{m}^t=\theta\boldsymbol{m}^{t-1}+(1-\theta)\boldsymbol{g}^t=\theta(\boldsymbol{\epsilon}^{t-1}+\nabla f_S(\boldsymbol{w}^{t-1}))+(1-\theta)(\boldsymbol{\delta}^t+\nabla f_S(\boldsymbol{w}^t))\\&\Rightarrow\boldsymbol{\epsilon}^t=\boldsymbol{m}^t-\nabla f_S(\boldsymbol{w}^t)=\theta\boldsymbol{\epsilon}^{t-1}+\theta(\nabla f_S(\boldsymbol{w}^{t-1})-\nabla f_S(\boldsymbol{w}^t))+(1-\theta)\boldsymbol{\delta}^t.\end{aligned}$$

Noting the smoothness of the gradient, we have

$$\begin{aligned}\|\boldsymbol{s}^t\|&=\|\nabla f_S(\boldsymbol{w}^{t-1})-\nabla f_S(\boldsymbol{w}^t)\|\le L\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^t\|\\&=L\|\alpha\boldsymbol{w}^{t-1}-\boldsymbol{w}^t+(1-\alpha)\boldsymbol{w}^{t-1}\|\le L\|\alpha\boldsymbol{w}^{t-1}-\boldsymbol{w}^t\|+L\|(\alpha-1)\boldsymbol{w}^{t-1}\|\\&\le L\sqrt{d}\gamma+(1-\alpha)L\|\boldsymbol{w}^{t-1}\|\le 2L\sqrt{d}\gamma.\end{aligned}$$

This leads us to

$$\boldsymbol{\epsilon}^t=\theta\boldsymbol{\epsilon}^{t-1}+\boldsymbol{s}^t+(1-\theta)\boldsymbol{\delta}^t.$$

Using Mathematical Induction, we derive the following result

$$\boldsymbol{\epsilon}^t=\theta^t\boldsymbol{\epsilon}^0+\theta\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{s}^i+(1-\theta)\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i.\tag{7}$$

Taking norms on both sides gives

$$\|\boldsymbol{\epsilon}^t\|\le 2L\sqrt{d}\gamma\sum_{i=1}^{t}\theta^i+(1-\theta)\Big\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\Big\|+\theta^t\|\boldsymbol{\epsilon}^0\|.$$

Furthermore, by taking expectations on both sides, we have

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\|\le\frac{2L\sqrt{d}\gamma\theta}{1-\theta}+(1-\theta)\mathbb{E}\Big\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\Big\|+\theta^t\|\boldsymbol{\epsilon}^0\|.\tag{8}$$

Applying Cauchy's inequality and considering the independence of $(\boldsymbol{s}^i)_{1\leq i\leq t}$, together with the fact that $\mathbb{E}|\boldsymbol{\delta}^i|^2 \leq \sigma^2$, we have

$$\mathbb{E}\Big\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\Big\| \leq \sqrt{\mathbb{E}\Big\|\sum_{i=1}^{t}\theta^{t-i}\boldsymbol{\delta}^i\Big\|^2} = \sqrt{\mathbb{E}\sum_{i=1}^{t}\theta^{2t-2i}\|\boldsymbol{\delta}^i\|^2} \leq \frac{\sigma}{\sqrt{1-\theta^2}}.$$

Consequently, we obtain

$$\mathbb{E}\|\boldsymbol{\epsilon}^t\| \leq \frac{2L\sqrt{d}\gamma\theta}{1-\theta} + \frac{\sqrt{1-\theta}}{\sqrt{1+\theta}}\sigma + \theta^t\|\boldsymbol{\epsilon}^0\| \leq \frac{2L\sqrt{d}\gamma}{1-\theta} + \sqrt{1-\theta}\sigma + \theta^t\|\boldsymbol{\epsilon}^0\|.$$

Leveraging Lemma 5 with $\boldsymbol{x}^\dagger \to \boldsymbol{x}^t$ and $\boldsymbol{m} \to \boldsymbol{m}^t$ and Lemma 4, we derive

$$f_S(\boldsymbol{w}^{t+1}) - f_S(\boldsymbol{w}^t) \leq -\gamma[1-(1-\alpha^T)\delta_T]\|\nabla f_S(\boldsymbol{w}^t)\|_1 + 2\sqrt{d}\gamma\|\boldsymbol{\epsilon}^t\| + 2Ld\gamma^2.$$

Summing the recursion from $t=1$ to T yields

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{x}^t)\|_1 \leq{}& \frac{f_S(\boldsymbol{w}^1)-\min f_S}{\gamma[1-(1-\alpha^T)\delta_T]T} + \frac{4Ld\gamma}{[1-(1-\alpha^T)\delta_T](1-\theta)} + \frac{2\sqrt{d}\sqrt{1-\theta}\sigma}{[1-(1-\alpha^T)\delta_T]}\\ &+ \frac{2L\gamma d}{[1-(1-\alpha^T)\delta_T]} + \frac{2\sqrt{d}\|\boldsymbol{\epsilon}^0\|}{[1-(1-\alpha^T)\delta_T](1-\theta)T}.\end{aligned}$$

Let $1-\theta = \frac{1}{\sqrt{T}}$ and $\gamma = \frac{1}{L\sqrt{d}T^{3/4}}$, we arrive at

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{x}^t)\|_1 = \mathcal{O}\Big(\frac{\sqrt{d}(f_S(\boldsymbol{w}^1)-\min f_S+\sigma+\|\nabla f_S(\boldsymbol{w}^0)\|)}{[1-(1-\alpha^T)\delta_T]T^{1/4}}\Big).$$

Recalling Lemma 4, we observe that the weight decay is approximately $\frac{1}{T^{3/4}}$ when $\gamma = \frac{1}{L\sqrt{d}T^{3/4}}$. Otherwise, the optimization will be confined to a very small region around $\mathbf{0}$.

I. As $1-\frac{1}{T} \leq \alpha < 1$, $\alpha^T \asymp 1$ yielding

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{x}^t)\|_1 = \mathcal{O}\Big(\frac{\sqrt{d}(f_S(\boldsymbol{w}^1)-\min f_S+\sigma+\|\nabla f_S(\boldsymbol{w}^0)\|)}{T^{1/4}}\Big).$$

II. As $1-\frac{1}{T^{3/4}} = \alpha$, $\alpha^T \asymp 0$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{x}^t)\|_1 &= \mathcal{O}\Big(\frac{\sqrt{d}(f_S(\boldsymbol{w}^1)-\min f_S+\sigma+\|\nabla f_S(\boldsymbol{w}^0)\|)}{[(1-\delta_T)+\delta_T\alpha^T]T^{1/4}}\Big)\\ &= \mathcal{O}\Big(\frac{\sqrt{d}(f_S(\boldsymbol{w}^1)-\min f_S+\sigma+\|\nabla f_S(\boldsymbol{w}^0)\|)}{(1-\delta_T)T^{1/4}}\Big).\end{aligned}$$

Note that $\boldsymbol{w}^1 = \boldsymbol{w}^0$ because $\boldsymbol{m}^0 = \mathbf{0}$, we then proved the result.

E. Proof of Theorem 4

I. Using Lemmas 1 and 6, in the general case, we then get

$$\begin{aligned}\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|_1}{T} &\leq \frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)-\nabla f_S(\boldsymbol{w}^t)\|_1}{T} + \frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|_1}{T}\\ &\leq \frac{\sum_{t=1}^{T}\sqrt{d}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)-\nabla f_S(\boldsymbol{w}^t)\|}{T} + \frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|_1}{T}\\ &= \mathcal{O}\Big(\frac{\sqrt{d}}{T^{1/4}} + \frac{\sqrt{d}T^{3/4}}{\sqrt{n}}\Big).\end{aligned}$$

II. If $\delta < 1$, we have

$$\begin{aligned}\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|_1}{T} &\leq \frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)-\nabla f_S(\boldsymbol{w}^t)\|_1}{T}+\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|_1}{T}\\ &\leq \frac{\sum_{t=1}^{T}\sqrt{d}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)-\nabla f_S(\boldsymbol{w}^t)\|}{T}+\frac{\sum_{t=1}^{T}\mathbb{E}\|\nabla f_S(\boldsymbol{w}^t)\|_1}{T}\\ &= \mathcal{O}\Big(\frac{\sqrt{d}}{(1-\delta)T^{1/4}}+\sqrt{\frac{T}{n}}\Big).\end{aligned}$$

Based on these two cases, we then prove the proof.

F. Proofs of Technical Lemmas

F.1. Proof of Lemma 2

I. ($1-\alpha \leq \min\Big\{\gamma^2, \frac{1}{2T}\Big\}$) The general case is almost identical to the SGD results proved by [25]. We present the proof here for completeness. We can assume, without loss of generality, that $S = \{\xi_1, \ldots, \xi_{n-1}, \xi_n\}$ and $S' = \{\xi_1, \ldots, \xi_{n-1}, \xi_n'\}$. Let $I := \{i_1, \ldots, i_T\}$ denote the set of indices selected during the implementation of SignSGDW, and $(\boldsymbol{x}^t)_{t\geq 1}$ and $(\boldsymbol{y}^t)_{t\geq 1}$ be the sequences generated by the two neighbor sets. By direct computation, we have

$$\begin{aligned}&\mathbb{E}_{\text{SGDW}}\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\\ &= \mathbb{E}_{\text{SGDW}}\Big[\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\Big|n\in I\Big]\text{Prob}(n\in I)\\ &\quad+\mathbb{E}_{\text{SGDW}}\Big[\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\Big|n\notin I\Big]\text{Prob}(n\notin I).\end{aligned}$$

Noticing that $\boldsymbol{y}^T = \boldsymbol{x}^T$ as $n \notin I$, we then derive the following bound

$$\begin{aligned}&\mathbb{E}_{\text{SignSGDW}}\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\\ &= \mathbb{E}_{\text{SignSGDW}}\Big[\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\Big|n\in I\Big]\text{Prob}(n\in I)\\ &\leq 4G^2[\sum_{t=1}^{T}\text{Prob}(i_t=n)] \leq \frac{4G^2T}{n} = \mathcal{O}\Big(\frac{T}{n}\Big).\end{aligned}$$

II. ($1-\alpha=\gamma$) Without loss of generality, we assume that $S = \{\xi_1, \ldots, \xi_{n-1}, \xi_n\}$ and $S' = \{\xi_1, \ldots, \xi_{n-1}, \xi_n'\}$. Let $(\boldsymbol{x}^t)_{t\geq 0}$ and $(\boldsymbol{y}^t)_{t\geq 0}$ be the output of SGDW on these two neighbor sets.

1. With probability $1-\frac{1}{n}$, we derive

$$\begin{aligned}&\|\boldsymbol{x}^{t+1}-\boldsymbol{y}^{t+1}\|^2 = \|\alpha\boldsymbol{x}^t-\alpha\boldsymbol{y}^t-\gamma(\nabla f(\boldsymbol{x}^t;\xi_{i_t})-\nabla f(\boldsymbol{y}^t;\xi_{i_t}))\|^2\\ &\leq (1+h)\alpha^2\|\boldsymbol{x}^t-\boldsymbol{y}^t\|^2+(1+\frac{1}{h})\gamma^2\|\nabla f(\boldsymbol{x}^t;\xi_{i_t})-\nabla f(\boldsymbol{y}^t;\xi_{i_t}')\|^2.\end{aligned}$$

Setting $h = \frac{1-\alpha}{\alpha}$, we are led to

$$\|\boldsymbol{x}^{t+1}-\boldsymbol{y}^{t+1}\|^2 \leq \alpha\|\boldsymbol{x}^t-\boldsymbol{y}^t\|^2+\frac{\gamma^2L^2}{1-\alpha}\|\boldsymbol{x}^t-\boldsymbol{y}^t\|^2 \leq \|\boldsymbol{x}^t-\boldsymbol{y}^t\|^2.$$

2. With probability $\frac{1}{n}$, we get

$$\begin{aligned}&\|\boldsymbol{x}^{t+1}-\boldsymbol{y}^{t+1}\|^2 = \|\alpha\boldsymbol{x}^t-\alpha\boldsymbol{y}^t-\gamma(\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{y}^t;\xi_n'))\|^2\\ &\leq \alpha\|\boldsymbol{x}^t-\boldsymbol{y}^t\|^2+\frac{4\gamma^2G^2}{1-\alpha}\\ &\leq \alpha\|\boldsymbol{x}^t-\boldsymbol{y}^t\|^2+4\gamma G^2.\end{aligned}$$

In summary, we then get

$$\mathbb{E}\|\boldsymbol{x}^{t+1}-\boldsymbol{y}^{t+1}\|^2 \leq \mathbb{E}\|\boldsymbol{x}^{t}-\boldsymbol{y}^{t}\|^2+\frac{4\gamma G^2}{n},$$

which then yields

$$\mathbb{E}\|\boldsymbol{x}^{T}-\boldsymbol{y}^{T}\|^2 \leq \frac{4\gamma T G^2}{n}.$$

The Lipschitz property of the gradient then indicates

$$\sup_{\xi\in\Xi}\mathbb{E}\|\nabla f(\boldsymbol{x}^T;\xi)-\nabla f(\boldsymbol{y}^T;\xi)\|^2 \leq L^2\mathbb{E}\|\boldsymbol{x}^T-\boldsymbol{y}^T\|^2 \leq \frac{4\gamma T L^2 G^2}{n}.$$

Hence, we can get the bound as

$$\sup_{\xi\in\Xi}\mathbb{E}\|\nabla f(\boldsymbol{x}^T;\xi)-\nabla f(\boldsymbol{y}^T;\xi)\|^2 \leq \frac{4\gamma T L^2 G^2}{n}.$$

F.2. Proof of Lemma 3

Similarly, we can get

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{w}^{t+1}\|^2 &= \mathbb{E}\|\alpha\boldsymbol{w}^t-\gamma\boldsymbol{g}^t\|^2\\
&= \mathbb{E}\|\alpha\boldsymbol{w}^t\|^2-2\gamma\alpha\langle\boldsymbol{w}^t,\mathbb{E}\boldsymbol{g}^t\rangle+\gamma^2\mathbb{E}\|\boldsymbol{g}^t\|^2\\
&= \mathbb{E}\|\alpha\boldsymbol{w}^t\|^2-2\gamma\alpha\mathbb{E}\langle\boldsymbol{w}^t,\nabla f(\boldsymbol{w}^t)\rangle+\gamma^2\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|^2+\gamma^2\sigma^2\\
&= \mathbb{E}\|\alpha\boldsymbol{w}^t\|^2+2\gamma\alpha\mathbb{E}(\|\boldsymbol{w}^t\|\cdot\|\nabla f(\boldsymbol{w}^t)\|)+\gamma^2\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|^2+\gamma^2\sigma^2\\
&\leq (1+\frac{1}{h})\alpha^2\mathbb{E}\|\boldsymbol{w}^t\|^2+(1+h)\gamma^2\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|^2+\gamma^2\sigma^2.
\end{aligned}$$

Setting $h=\frac{\alpha}{1-\alpha}$, we get

$$\mathbb{E}\|\boldsymbol{w}^{t+1}\|^2 \leq \alpha\mathbb{E}\|\boldsymbol{w}^t\|^2+\frac{\gamma^2}{1-\alpha}\mathbb{E}\|\nabla f(\boldsymbol{w}^t)\|^2+\gamma^2\sigma^2.$$

Thus, we are led to

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{w}^{t+1}\|^2 &\leq \alpha^t\|\boldsymbol{w}^1\|^2+\sum_{j=1}^{t}\alpha^{t-j}(\frac{1}{1-\alpha}\gamma^2\mathbb{E}\|\nabla f(\boldsymbol{w}^j)\|^2+\gamma^2\sigma^2)\\
&\leq \alpha^t\|\boldsymbol{w}^1\|^2+\frac{1}{1-\alpha}\gamma^2\sum_{j=1}^{t}\alpha^{t-j}\mathbb{E}\|\nabla f(\boldsymbol{w}^j)\|^2+\frac{\gamma^2(1-\alpha^T)}{1-\alpha}\sigma^2
\end{aligned}$$

as $t\leq T$. Summation from $t=1$ to T leads to

$$\sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{w}^t\|^2 \leq \sum_{t=1}^{T}\mathbb{E}\|\boldsymbol{w}^{t+1}\|^2 \leq \frac{1}{1-\alpha}\|\boldsymbol{w}^1\|^2+\frac{\gamma^2(1-\alpha^T)}{(1-\alpha)^2}\sum_{j=1}^{t}\mathbb{E}\|\nabla f(\boldsymbol{w}^j)\|^2+\frac{\gamma^2 T(1-\alpha^T)}{1-\alpha}\sigma^2.$$

Noting that $\boldsymbol{w}^1=\boldsymbol{0}$, we then proved the result.

F.3. Proof of Lemma 4

With the scheme of the algorithm, we have

$$\|\boldsymbol{w}^{t+1}\| \leq \alpha\|\boldsymbol{w}^t\|+\gamma\sqrt{d} \leq \gamma\sqrt{d}(\sum_{j=0}^{t}\alpha^j) \leq \gamma\sqrt{d}(\sum_{j=0}^{T-1}\alpha^j) \leq \frac{1-\alpha^T}{1-\alpha}\gamma\sqrt{d}.$$

For any $i\in\{1,2,\ldots,n\}$, it follows that

$$\|\boldsymbol{w}_i^{t+1}\| \leq \alpha\|\boldsymbol{w}_i^t\|+\gamma \leq \gamma(\sum_{j=0}^{t}\alpha^j) \leq \gamma(\sum_{j=0}^{T-1}\alpha^j) \leq \frac{1-\alpha^T}{1-\alpha}\gamma.$$

F.4. Proof of Lemma 5

The proof of this lemma is inspired by the work of [34]. With the smoothness of ∇f_S, we have

$$\begin{aligned}
&f_S(\boldsymbol{x}^{\ddagger}) - f_S(\boldsymbol{x}^{\dagger}) \\
&\leq \langle \nabla f_S(\boldsymbol{x}^{\dagger}), \boldsymbol{x}^{\ddagger} - \boldsymbol{x}^{\dagger} \rangle + \frac{L}{2}\|\boldsymbol{x}^{\ddagger} - \boldsymbol{x}^{\dagger}\|^2 \\
&\leq -\gamma \langle \nabla f(\boldsymbol{x}^{\dagger}), \text{Sign}(\boldsymbol{m}) \rangle + (\alpha - 1)\langle \nabla f_S(\boldsymbol{x}^{\dagger}), \boldsymbol{x}^{\dagger} \rangle + L\|\gamma \text{Sign}(\boldsymbol{m})\|^2 + L\|(\alpha-1)\boldsymbol{x}^{\dagger}\|^2 \\
&\leq -\gamma \langle \nabla f_S(\boldsymbol{x}^{\dagger}), \text{Sign}(\boldsymbol{m}) \rangle + Ld\gamma^2 + L(\alpha-1)^2\|\boldsymbol{x}^{\dagger}\|^2 + (1-\alpha)\delta\|\nabla f_S(\boldsymbol{x}^{\dagger})\|_1 \cdot \|\boldsymbol{x}^{\dagger}\|_{\infty} \\
&= -\gamma \langle \nabla f_S(\boldsymbol{x}^{\dagger}), \text{Sign}[\nabla f_S(\boldsymbol{x}^{\dagger})] \rangle + \gamma \langle \nabla f_S(\boldsymbol{x}^{\dagger}), \text{Sign}(\boldsymbol{m}) - \text{Sign}[\nabla f_S(\boldsymbol{x}^{\dagger})] \rangle \\
&\quad + Ld\gamma^2 + L(\alpha-1)^2\|\boldsymbol{x}^{\dagger}\|^2 + (1-\alpha)\delta\|\nabla f_S(\boldsymbol{x}^{\dagger})\|_1 \cdot \|\boldsymbol{x}^{\dagger}\|_{\infty} \\
&= -\gamma\|\nabla f_S(\boldsymbol{x}^{\dagger})\|_1 + \gamma \langle \nabla f_S(\boldsymbol{x}^{\dagger}), \text{Sign}(\boldsymbol{m}) - \text{Sign}[\nabla f_S(\boldsymbol{x}^{\dagger})] \rangle \\
&\quad + Ld\gamma^2 + L(\alpha-1)^2\|\boldsymbol{x}^{\dagger}\|^2 + (1-\alpha)\delta\|\nabla f_S(\boldsymbol{x}^{\dagger})\|_1 \cdot \|\boldsymbol{x}^{\dagger}\|_{\infty}.
\end{aligned}$$

With direct computation,

$$\langle \nabla f_S(\boldsymbol{x}^{\dagger}), \text{Sign}[\nabla f_S(\boldsymbol{x}^{\dagger})] - \text{Sign}(\boldsymbol{m}) \rangle = \sum_{i=1}^{d} [\nabla f_S(\boldsymbol{x}^{\dagger})]_i \cdot [\text{Sign}([\nabla f_S(\boldsymbol{x}^{\dagger})]_i) - \text{Sign}([\boldsymbol{m}]_i)].$$

I. If $\text{Sign}([\nabla f_S(\boldsymbol{x}^{\dagger})]_i) = \text{Sign}([\boldsymbol{m}]_i)$, $[\nabla f_S(\boldsymbol{x}^{\dagger})]_i \cdot [\text{Sign}([\nabla f_S(\boldsymbol{x}^{\dagger})]_i) - \text{Sign}([\boldsymbol{m}]_i)] = 0$.
II. Otherwise, $[\nabla f_S(\boldsymbol{x}^{\dagger})]_i \cdot [\boldsymbol{m}]_i \leq 0$ gives us

$$|\boldsymbol{\epsilon}_i| = |[\nabla f_S(\boldsymbol{x}^{\dagger})]_i - [\boldsymbol{m}]_i| \geq |[\nabla f_S(\boldsymbol{x}^{\dagger})]_i|.$$

And we can get

$$[\nabla f_S(\boldsymbol{x}^{\dagger})]_i \cdot [\text{Sign}([\nabla f_S(\boldsymbol{x}^{\dagger})]_i) - \text{Sign}([\boldsymbol{m}]_i)] \leq 2|[\nabla f_S(\boldsymbol{x}^{\dagger})]_i| \leq 2|\boldsymbol{\epsilon}_i|. \tag{9}$$

Therefore, for any $i \in [d]$, inequality (9) holds. Based on this, we have

$$\langle \nabla f_S(\boldsymbol{x}^{\dagger}), \text{Sign}[\nabla f_S(\boldsymbol{x}^{\dagger})] - \text{Sign}(\boldsymbol{m}) \rangle \leq 2\sum_{i=1}^{d} |\boldsymbol{\epsilon}_i| \leq 2\sqrt{d}\|\boldsymbol{\epsilon}\|,$$

which further yields

$$\begin{aligned}
f_S(\boldsymbol{x}^{\ddagger}) - f_S(\boldsymbol{x}^{\dagger}) \leq &-\gamma\|\nabla f(\boldsymbol{x}^{\dagger})\|_1 + 2\sqrt{d}\gamma\|\boldsymbol{\epsilon}\| + Ld\gamma^2 \\
&+ L(\alpha-1)^2\|\boldsymbol{x}^{\dagger}\|^2 + (1-\alpha)\delta\|\nabla f_S(\boldsymbol{x}^{\dagger})\|_1 \cdot \|\boldsymbol{x}^{\dagger}\|_{\infty}.
\end{aligned}$$

Recall that $\|\boldsymbol{x}^{\dagger}\| \leq \frac{\gamma\sqrt{d}}{1-\alpha}$, $\|\boldsymbol{x}^{\dagger}\|_{\infty} \leq \frac{1-\alpha^T}{1-\alpha}\gamma$, we then get

$$\begin{aligned}
f_S(\boldsymbol{x}^{\ddagger}) - f_S(\boldsymbol{x}^{\dagger}) &\leq -\gamma\|\nabla f_S(\boldsymbol{x}^{\dagger})\|_1 + 2\sqrt{d}\gamma\|\boldsymbol{\epsilon}\| + 2Ld\gamma^2 + (1-\alpha^T)\delta\gamma\|\nabla f(\boldsymbol{x}^{\dagger})\|_1 \\
&= -\gamma[1 - (1-\alpha^T)\delta]\|\nabla f(\boldsymbol{x}^{\dagger})\|_1 + 2\sqrt{d}\gamma\|\boldsymbol{\epsilon}\| + 2Ld\gamma^2.
\end{aligned}$$

F.5. Proof of Lemma 6

Without loss of generality, we assume that $S = \{\xi_1, \ldots, \xi_{n-1}, \xi_n\}$ and $S' = \{\xi_1, \ldots, \xi_{n-1}, \xi_n'\}$. Let $I := \{i_1, \ldots, i_T\}$ be the set of indices selected during the implementation of SignSGDW, and $(\boldsymbol{x}^t)_{t\geq 1}$ and $(\boldsymbol{y}^t)_{t\geq 1}$ be generated by the two neighbor sets. The scheme of the algorithm gives us

$$\begin{aligned}
&\|\boldsymbol{x}^{t+1} - \boldsymbol{y}^{t+1}\| = \|\alpha\boldsymbol{x}^t - \alpha\boldsymbol{y}^t - \gamma(\text{Sign}(\boldsymbol{m}^t) - \text{Sign}(\hat{\boldsymbol{m}}^t))\| \\
&\leq \alpha\|\boldsymbol{x}^t - \boldsymbol{y}^t\|^2 + \gamma\|\text{Sign}(\boldsymbol{m}^t) - \text{Sign}(\hat{\boldsymbol{m}}^t)\| \\
&\leq \alpha\|\boldsymbol{x}^t - \boldsymbol{y}^t\| + 2\gamma\sqrt{d}.
\end{aligned}$$

Hence, we are led to

$$\sup_{t\geq 1}\|\boldsymbol{x}^t-\boldsymbol{y}^t\|\leq\frac{2\gamma\sqrt{d}(1-\alpha^T)}{(1-\alpha)}.$$

It then follows

$$\begin{aligned}
&\mathbb{E}_{\text{SignSGDW}}\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\\
&=\mathbb{E}_{\text{SignSGDW}}\Big[\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\Big|n\in I\Big]\text{Prob}(n\in I)\\
&\quad+\mathbb{E}_{\text{SignSGDW}}\Big[\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\Big|n\notin I\Big]\text{Prob}(n\notin I).
\end{aligned}$$

Noticing that $\boldsymbol{y}^T=\boldsymbol{x}^T$ as $n\notin I$, we then get

$$\begin{aligned}
&\mathbb{E}_{\text{SignSGDW}}\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\\
&=\mathbb{E}_{\text{SignSGDW}}\Big[\|\nabla f(\boldsymbol{y}^T;\xi)-\nabla f(\boldsymbol{x}^T;\xi)\|^2\Big|n\in I\Big]\text{Prob}(n\in I)\\
&\leq L^2\mathbb{E}\|\boldsymbol{x}^T-\boldsymbol{y}^T\|^2\text{Prob}(n\in I)\leq\frac{4L^2\gamma^2d(1-\alpha^T)^2}{(1-\alpha)^2}\text{Prob}(n\in I)\\
&\leq\frac{4L^2\gamma^2d(1-\alpha^T)^2}{(1-\alpha)^2}[\sum_{t=1}^{T}\text{Prob}(i_t=n)]\leq\frac{4L^2\gamma^2d(1-\alpha^T)^2}{(1-\alpha)^2}\cdot\frac{T}{n}.
\end{aligned}\tag{10}$$

I. In the general case ($\delta=1$), $1-\alpha\leq\frac{1}{T}$, it holds $\frac{1-\alpha^T}{1-\alpha}\asymp T$ yielding

$$\frac{4L^2\gamma^2d(1-\alpha^T)^2}{(1-\alpha)^2}\cdot\frac{T}{n}=\mathcal{O}\Big(\frac{4L^2\gamma^2dT^3}{n}\Big)=\mathcal{O}\Big(\frac{T^{3/2}}{n}\Big).$$

II. If $\delta<1$ and $\alpha=1-\frac{1}{T^{1/4}}$, we have

$$\frac{4L^2\gamma^2d(1-\alpha^T)^2}{(1-\alpha)^2}\cdot\frac{T}{n}\leq\frac{4L^2\gamma^2d}{(1-\alpha)^2}\cdot\frac{T}{n}=\mathcal{O}\Big(\frac{4L^2\gamma^2dT^{5/2}}{n}\Big)=\mathcal{O}\Big(\frac{T}{n}\Big).$$