Appendix for "Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent"

Yunwen Lei [1 2] **Yiming Ying** [3]

A. Optimization Error Bounds

For a full picture of generalization errors, we need to address the optimization errors. This is achieved by the following lemma. Parts (a) and (b) consider the convex and strongly convex empirical objectives, respectively (we make no assumptions on the convexity of each $f(\cdot, z)$). Note in Parts (c) and (d), we do not make a bounded gradient assumption. As an alternative, we require convexity of $f(\cdot; z)$ for all z. An appealing property of Parts (c) and (d) is that it involves $O(\sum_{j=1}^t \eta_j^2 F_S(\mathbf{w}))$ instead of $O(\sum_{j=1}^t \eta_j^2)$, which is a requirement for developing fast rates in the case with low noises.

Our discussion on optimization errors requires to use a self-bounding property for functions with Hölder continuous (sub)gradients, which means that gradients can be controlled by function values. The case $\alpha = 1$ was established in Srebro et al. (2010). The case $\alpha \in (0, 1)$ was established in Ying & Zhou (2017) and slightly refined in Lei et al. (2019). The case $\alpha = 0$ follows directly from Definition 3. Define

$$c_{\alpha,1} = \begin{cases} (1 + 1/\alpha)^{\frac{\alpha}{1+\alpha}} L^{\frac{1}{1+\alpha}}, & \text{if } \alpha > 0 \\ \sup_z \|\partial f(0; z)\|_2 + L, & \text{if } \alpha = 0. \end{cases} \tag{A.1}$$

Lemma A.1. *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, and $\mathbf{w} \mapsto \partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1]$. Then for $c_{\alpha,1}$ defined as (A.1) we have*

$$\|\partial f(\mathbf{w}, z)\|_2 \leq c_{\alpha,1} f^{\frac{\alpha}{1+\alpha}}(\mathbf{w}, z), \quad \forall \mathbf{w} \in \mathbb{R}^d, z \in \mathcal{Z}.$$

Lemma A.2. *(a) Let $\{\mathbf{w}_t\}_t$ be produced by (3.3) and Assumption 1 hold. If F_S is convex, then for all $t \in \mathbb{N}$ and $\mathbf{w} \in \Omega$*

$$\mathbb{E}_A[F_S(\mathbf{w}_t^{(1)})] - F_S(\mathbf{w}) \leq \frac{G^2 \sum_{j=1}^t \eta_j^2 + \|\mathbf{w}\|_2^2}{2 \sum_{j=1}^t \eta_j},$$

where $\mathbf{w}_t^{(1)} = \big(\sum_{j=1}^t \eta_j \mathbf{w}_j\big) / \sum_{j=1}^t \eta_j$.

(b) Let F_S be σ_S-strongly convex and Assumption 1 hold. Let $t_0 \geq 0$ and $\{\mathbf{w}_t\}_t$ be produced by (3.3) with $\eta_t = 2/(\sigma_S(t + t_0))$. Then for all $t \in \mathbb{N}$ and $\mathbf{w} \in \Omega$

$$\mathbb{E}_A[F_S(\mathbf{w}_t^{(2)})] - F_S(\mathbf{w}) = O\big(1/(t\sigma_S) + \|\mathbf{w}\|_2^2/t^2\big),$$

where $\mathbf{w}_t^{(2)} = \big(\sum_{j=1}^t (j + t_0 - 1)\mathbf{w}_j\big) / \sum_{j=1}^t (j + t_0 - 1)$.

(c) Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and L-smooth. Let $\{\mathbf{w}_t\}_t$ be produced by (3.3) with $\eta_t \leq 1/(2L)$. If the step size is nonincreasing, then for all $t \in \mathbb{N}$ and $\mathbf{w} \in \Omega$ independent of the SGD algorithm A

$$\sum_{j=1}^t \eta_j \mathbb{E}_A[F_S(\mathbf{w}_j) - F_S(\mathbf{w})] \leq (1/2 + L\eta_1)\|\mathbf{w}\|_2^2 + 2L \sum_{j=1}^t \eta_j^2 F_S(\mathbf{w}).$$

[1]Department of Computer Science, University of Kaiserslautern, Germany [2]School of Computer Science, University of Birmingham, United Kingdom [3]Department of Mathematics and Statistics, State University of New York at Albany, USA. Correspondence to: Yiming Ying <yying@albany.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

(d) Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex, and $\partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1)$. Let $\{\mathbf{w}_t\}_t$ be produced by (3.3) with nonincreasing step sizes. Then for all $t \in \mathbb{N}$ and $\mathbf{w} \in \Omega$ independent of the SGD algorithm A we have

$$2\sum_{j=1}^{t} \eta_j \mathbb{E}_A[F_S(\mathbf{w}_j) - F_S(\mathbf{w})] \leq \|\mathbf{w}\|_2^2 + c_{\alpha,1}^2 \Big(\sum_{j=1}^{t} \eta_j^2\Big)^{\frac{1-\alpha}{1+\alpha}} \Big(\eta_1 \|\mathbf{w}\|_2^2 + 2\sum_{j=1}^{t} \eta_j^2 F_S(\mathbf{w}) + c_{\alpha,2} \sum_{j=1}^{t} \eta_j^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}},$$

where

$$c_{\alpha,2} = \begin{cases} \frac{1-\alpha}{1+\alpha}(2\alpha/(1+\alpha))^{\frac{2\alpha}{1-\alpha}} c_{\alpha,1}^{\frac{2+2\alpha}{1-\alpha}}, & \text{if } \alpha > 0 \\ c_{\alpha,1}^2, & \text{if } \alpha = 0. \end{cases} \tag{A.2}$$

Proof. Parts (a) and (b) can be found in the literature (Lacoste-Julien et al., 2012; Nemirovski et al., 2009). We only prove Parts (c) and (d). We first prove Part (c). The projection operator Π_Ω is non-expansive, i.e.,

$$\big\|\Pi_\Omega(\mathbf{w}) - \Pi_\Omega(\tilde{\mathbf{w}})\big\|_2 \leq \|\mathbf{w} - \tilde{\mathbf{w}}\|_2. \tag{A.3}$$

By the SGD update (3.3), (A.3), convexity and Lemma A.1, we know

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 &\leq \|\mathbf{w}_t - \eta_t \partial f(\mathbf{w}_t; z_{i_t}) - \mathbf{w}\|_2^2 \\ &= \|\mathbf{w}_t - \mathbf{w}\|_2^2 + \eta_t^2 \|\partial f(\mathbf{w}_t; z_{i_t})\|_2^2 + 2\eta_t \langle \mathbf{w} - \mathbf{w}_t, \partial f(\mathbf{w}_t; z_{i_t}) \\ &\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + 2\eta_t^2 L f(\mathbf{w}_t; z_{i_t}) + 2\eta_t (f(\mathbf{w}; z_{i_t}) - f(\mathbf{w}_t; z_{i_t})) \\ &\leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 + 2\eta_t f(\mathbf{w}; z_{i_t}) - \eta_t f(\mathbf{w}_t; z_{i_t}), \end{aligned} \tag{A.4}$$

where the last inequality is due to $\eta_t \leq 1/(2L)$. It then follows that

$$\eta_t f(\mathbf{w}_t; z_{i_t}) \leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 + 2\eta_t f(\mathbf{w}; z_{i_t}).$$

Multiplying both sides by η_t and using the assumption $\eta_{t+1} \leq \eta_t$, we know

$$\begin{aligned} \eta_t^2 f(\mathbf{w}_t; z_{i_t}) &\leq \eta_t \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \eta_t \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 + 2\eta_t^2 f(\mathbf{w}; z_{i_t}) \\ &\leq \eta_t \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \eta_{t+1} \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 + 2\eta_t^2 f(\mathbf{w}; z_{i_t}). \end{aligned}$$

Taking a summation of the above inequality gives ($\mathbf{w}_1 = 0$)

$$\sum_{j=1}^{t} \eta_j^2 f(\mathbf{w}_j; z_{i_j}) \leq \eta_1 \|\mathbf{w}\|_2^2 + 2\sum_{j=1}^{t} \eta_j^2 f(\mathbf{w}; z_{i_j}).$$

Taking an expectation w.r.t. A gives (note $\mathbf{w}_j$ is independent of i_j)

$$\sum_{j=1}^{t} \eta_j^2 \mathbb{E}_A[F_S(\mathbf{w}_j)] = \sum_{j=1}^{t} \eta_j^2 \mathbb{E}_A\big[f(\mathbf{w}_j; z_{i_j})\big] \leq \eta_1 \|\mathbf{w}\|_2^2 + 2\sum_{j=1}^{t} \eta_j^2 \mathbb{E}_A[F_S(\mathbf{w})]. \tag{A.5}$$

On the other hand, taking an expectation w.r.t. i_t over both sides of (A.4) shows

$$2\eta_t \big[F_S(\mathbf{w}_t) - F_S(\mathbf{w})\big] \leq \|\mathbf{w}_t - \mathbf{w}\|_2^2 - \mathbb{E}_{i_t}\big[\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\big] + 2\eta_t^2 L F_S(\mathbf{w}_t).$$

Taking an expectation on both sides followed with a summation, we get

$$\begin{aligned} 2\sum_{j=1}^{t} \eta_j \mathbb{E}_A[F_S(\mathbf{w}_j) - F_S(\mathbf{w})] &\leq \|\mathbf{w}\|_2^2 + 2L\sum_{j=1}^{t} \eta_j^2 \mathbb{E}_A[F_S(\mathbf{w}_j)] \\ &\leq (1 + 2L\eta_1)\|\mathbf{w}\|_2^2 + 4L\sum_{j=1}^{t} \eta_j^2 \mathbb{E}_A[F_S(\mathbf{w})], \end{aligned}$$

where the last step is due to (A.5). The proof is complete since $\mathbf{w}$ is independent of A.

We now prove Part (d). Analogous to (A.4), one can show for loss functions with Hölder continuous (sub)gradients (Lemma A.1)

$$\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2 \le \|\mathbf{w}_t-\mathbf{w}\|_2^2 + c_{\alpha,1}^2\eta_t^2 f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_{i_t}) + 2\eta_t(f(\mathbf{w};z_{i_t})-f(\mathbf{w}_t;z_{i_t})). \tag{A.6}$$

By the Young's inequality

$$ab \le p^{-1}|a|^p + q^{-1}|b|^q, \quad a,b\in\mathbb{R}, p,q>0 \text{ with } p^{-1}+q^{-1}=1, \tag{A.7}$$

we know (notice the following inequality holds trivially if $\alpha=0$)

$$\begin{aligned}
\eta_t c_{\alpha,1}^2 f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_{i_t}) &= \Big(\frac{1+\alpha}{2\alpha}f(\mathbf{w}_t;z_{i_t})\Big)^{\frac{2\alpha}{1+\alpha}}\Big(\frac{2\alpha}{1+\alpha}\Big)^{\frac{2\alpha}{1+\alpha}} c_{\alpha,1}^2\eta_t \\
&\le \frac{2\alpha}{1+\alpha}\Big(\frac{1+\alpha}{2\alpha}f(\mathbf{w}_t;z_{i_t})\Big)^{\frac{2\alpha}{1+\alpha}\frac{1+\alpha}{2\alpha}} + \frac{1-\alpha}{1+\alpha}\Big(\Big(\frac{2\alpha}{1+\alpha}\Big)^{\frac{2\alpha}{1+\alpha}} c_{\alpha,1}^2\eta_t\Big)^{\frac{1+\alpha}{1-\alpha}} \\
&= f(\mathbf{w}_t;z_{i_t}) + c_{\alpha,2}\eta_t^{\frac{1+\alpha}{1-\alpha}}.
\end{aligned}$$

Combining the above two inequalities together, we get

$$\eta_t f(\mathbf{w}_t;z_{i_t}) \le \|\mathbf{w}_t-\mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2 + 2\eta_t f(\mathbf{w};z_{i_t}) + c_{\alpha,2}\eta_t^{\frac{2}{1-\alpha}}.$$

Multiplying both sides by η_t and using $\eta_{t+1}\le\eta_t$, we derive

$$\eta_t^2 f(\mathbf{w}_t;z_{i_t}) \le \eta_t\|\mathbf{w}_t-\mathbf{w}\|_2^2 - \eta_{t+1}\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2 + 2\eta_t^2 f(\mathbf{w};z_{i_t}) + c_{\alpha,2}\eta_t^{\frac{3-\alpha}{1-\alpha}}.$$

Taking a summation of the above inequality gives

$$\sum_{j=1}^t \eta_j^2 f(\mathbf{w}_j;z_{i_j}) \le \eta_1\|\mathbf{w}\|_2^2 + 2\sum_{j=1}^t \eta_j^2 f(\mathbf{w};z_{i_j}) + c_{\alpha,2}\sum_{j=1}^t \eta_j^{\frac{3-\alpha}{1-\alpha}}. \tag{A.8}$$

According to the Jensen's inequality and the concavity of $x\mapsto x^{\frac{2\alpha}{1+\alpha}}$, we know

$$\begin{aligned}
\sum_{j=1}^t \eta_j^2 f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j;z_{i_j}) &\le \sum_{j=1}^t \eta_j^2\Big(\frac{\sum_{j=1}^t \eta_j^2 f(\mathbf{w}_j;z_{i_j})}{\sum_{j=1}^t\eta_j^2}\Big)^{\frac{2\alpha}{1+\alpha}} \\
&\le \Big(\sum_{j=1}^t\eta_j^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big(\eta_1\|\mathbf{w}\|_2^2 + 2\sum_{j=1}^t\eta_j^2 f(\mathbf{w};z_{i_j}) + c_{\alpha,2}\sum_{j=1}^t\eta_j^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}},
\end{aligned} \tag{A.9}$$

where in the last step we have used (A.8). Taking an expectation on both sides of (A.6), we know

$$2\eta_t\mathbb{E}_A\big[F_S(\mathbf{w}_t)-F_S(\mathbf{w})\big] \le \mathbb{E}_A[\|\mathbf{w}_t-\mathbf{w}\|_2^2] - \mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2\big] + c_{\alpha,1}^2\eta_t^2\mathbb{E}_A\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t;z_{i_t})\big].$$

Taking a summation of the above inequality gives

$$\begin{aligned}
2\sum_{j=1}^t \eta_j\mathbb{E}_A[F_S(\mathbf{w}_j)-F_S(\mathbf{w})] &\le \|\mathbf{w}\|_2^2 + c_{\alpha,1}^2\sum_{j=1}^t\eta_j^2\mathbb{E}_A\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j;z_{i_j})\big] \\
&\le \|\mathbf{w}\|_2^2 + c_{\alpha,1}^2\Big(\sum_{j=1}^t\eta_j^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big(\eta_1\|\mathbf{w}\|_2^2 + 2\sum_{j=1}^t\eta_j^2\mathbb{E}_A[F_S(\mathbf{w})] + c_{\alpha,2}\sum_{j=1}^t\eta_j^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}},
\end{aligned}$$

where we have used (A.9) and the concavity of $x\mapsto x^{\frac{2\alpha}{1+\alpha}}$ in the last step. The proof is complete by noting the independence between $\mathbf{w}$ and A. □

B. Proofs on Generalization by On-average Model Stability

To prove Theorem 2, we introduce an useful inequality for L-smooth functions $\mathbf{w} \mapsto f(\mathbf{w}; z)$ (Nesterov, 2013)

$$f(\mathbf{w}; z) \leq f(\tilde{\mathbf{w}}; z) + \langle \mathbf{w} - \tilde{\mathbf{w}}, \partial f(\tilde{\mathbf{w}}; z) \rangle + \frac{L\|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2}{2}. \tag{B.1}$$

Proof of Theorem 2. Due to the symmetry, we know

$$\begin{aligned}
\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] &= \mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^{n}\big(F(A(S^{(i)})) - F_S(A(S))\big)\Big] \\
&= \mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^{n}\big(f(A(S^{(i)}); z_i) - f(A(S); z_i)\big)\Big],
\end{aligned} \tag{B.2}$$

where the last identity holds since $A(S^{(i)})$ is independent of z_i. Under Assumption 1, it is then clear that

$$\big|\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big]\big| \leq \mathbb{E}_{S,\widetilde{S},A}\Big[\frac{G}{n}\sum_{i=1}^{n}\|A(S) - A(S^{(i)})\|_2\Big].$$

This proves Part (a).

We now prove Part (b). According to (B.1) due to the L-smoothness of f and (B.2), we know

$$\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},A}\Big[\langle A(S^{(i)}) - A(S), \partial f(A(S); z_i)\rangle + \frac{L}{2}\|A(S^{(i)}) - A(S)\|_2^2\Big].$$

According to the Schwartz's inequality we know

$$\begin{aligned}
\langle A(S^{(i)}) - A(S), \partial f(A(S); z_i)\rangle &\leq \|A(S^{(i)}) - A(S)\|_2\|\partial f(A(S); z_i)\|_2 \\
&\leq \frac{\gamma}{2}\|A(S^{(i)}) - A(S)\|_2^2 + \frac{1}{2\gamma}\|\partial f(A(S); z_i)\|_2^2 \\
&\leq \frac{\gamma}{2}\|A(S^{(i)}) - A(S)\|_2^2 + \frac{L}{\gamma}f(A(S); z_i),
\end{aligned}$$

where the last inequality is due to the self-bounding property of smooth functions (Lemma A.1). Combining the above two inequalities together, we derive

$$\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq \frac{L+\gamma}{2n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},A}\big[\|A(S^{(i)}) - A(S)\|_2^2\big] + \frac{L}{n\gamma}\sum_{i=1}^{n}\mathbb{E}_{S,A}[f(A(S); z_i)].$$

The stated inequality in Part (b) then follows directly by noting $\frac{1}{n}\sum_{i=1}^{n} f(A(S); z_i) = F_S(A(S))$.

Finally, we consider Part (c). By (B.2) and the convexity of f, we know

$$\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},A}\Big[\langle A(S^{(i)}) - A(S), \partial f(A(S^{(i)}); z_i)\rangle\Big].$$

By the Schwartz's inequality and Lemma A.1 we know

$$\begin{aligned}
\langle A(S^{(i)}) - A(S), \partial f(A(S^{(i)}); z_i)\rangle &\leq \frac{\gamma}{2}\|A(S^{(i)}) - A(S)\|_2^2 + \frac{1}{2\gamma}\|\partial f(A(S^{(i)}); z_i)\|_2^2 \\
&\leq \frac{\gamma}{2}\|A(S^{(i)}) - A(S)\|_2^2 + \frac{c_{\alpha,1}^2}{2\gamma}f^{\frac{2\alpha}{1+\alpha}}(A(S^{(i)}); z_i).
\end{aligned}$$

Combining the above two inequalities together, we get

$$\mathbb{E}_{S,A}\big[F(A(S)) - F_S(A(S))\big] \leq \frac{\gamma}{2n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},A}\big[\|A(S^{(i)}) - A(S)\|_2^2\big] + \frac{c_{\alpha,1}^2}{2\gamma}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S},A}\big[f^{\frac{2\alpha}{1+\alpha}}(A(S^{(i)}); z_i)\big].$$

Since $x \mapsto x^{\frac{2\alpha}{1+\alpha}}$ is concave and z_i is independent of $A(S^{(i)})$, we know

$$\mathbb{E}_{S,\widetilde{S},A}\big[f^{\frac{2\alpha}{1+\alpha}}(A(S^{(i)});z_i)\big] \leq \mathbb{E}_{S,\widetilde{S},A}\Big[\Big(\mathbb{E}_{z_i}\Big[f(A(S^{(i)});z_i)\Big]\Big)^{\frac{2\alpha}{1+\alpha}}\Big] = \mathbb{E}_{S,\widetilde{S},A}\Big[F^{\frac{2\alpha}{1+\alpha}}(A(S^{(i)}))\Big] = \mathbb{E}_{S,A}\Big[F^{\frac{2\alpha}{1+\alpha}}(A(S))\Big].$$

A combination of the above two inequalities then gives the stated bound in Part (c). The proof is complete. □

C. Proof on Learning without Bounded Gradients: Strongly Smooth Case

C.1. Stability bounds

A key property on establishing the stability of SGD is the non-expansiveness of the gradient-update operator established in the following lemma.

Lemma C.1 (Hardt et al. 2016)**.** *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w};z)$ is convex and L-smooth. Then for $\eta \leq 2/L$ we know*

$$\|\mathbf{w} - \eta\partial f(\mathbf{w};z) - \tilde{\mathbf{w}} + \eta\partial f(\tilde{\mathbf{w}};z)\|_2 \leq \|\mathbf{w} - \tilde{\mathbf{w}}\|_2.$$

Based on Lemma C.1, we establish stability bounds of models for SGD applied to two sets differing by a single example.

Lemma C.2. *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w};z)$ is nonnegative, convex and L-smooth. Let $S, \widetilde{S}$ and $S^{(i)}$ be constructed as Definition 4. Let $\{\mathbf{w}_t\}$ and $\{\mathbf{w}_t^{(i)}\}$ be produced by (3.3) with $\eta_t \leq 2/L$ based on S and $S^{(i)}$, respectively. Then for any $p > 0$ we have*

$$\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2\big] \leq \frac{2\sqrt{2L}}{n}\sum_{j=1}^{t}\eta_j\mathbb{E}_{S,A}\big[\sqrt{f(\mathbf{w}_j;z_i)}\big] \tag{C.1}$$

and

$$\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq \frac{8(1+1/p)L}{n}\sum_{j=1}^{t}(1+p/n)^{t-j}\eta_j^2\mathbb{E}_{S,A}\big[f(\mathbf{w}_j;z_i)\big]. \tag{C.2}$$

Proof. If $i_t \neq i$, we know the updates of $\mathbf{w}_{t+1}$ and $\mathbf{w}_{t+1}^{(i)}$ are based on stochastic gradients calculated with the same example z_{i_t}. By Lemma C.1 we then get

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2 \leq \big\|\mathbf{w}_t - \eta_t\partial f(\mathbf{w}_t;z_{i_t}) - \mathbf{w}_t^{(i)} + \eta_t\partial f(\mathbf{w}_t^{(i)},z_{i_t})\big\|_2 \leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2. \tag{C.3}$$

If $i_t = i$, we know

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2 &\leq \big\|\mathbf{w}_t - \eta_t\partial f(\mathbf{w}_t;z_i) - \mathbf{w}_t^{(i)} + \eta_t\partial f(\mathbf{w}_t^{(i)},\tilde{z}_i)\big\|_2 \\
&\leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 + \eta_t\|\partial f(\mathbf{w}_t;z_i) - \partial f(\mathbf{w}_t^{(i)};\tilde{z}_i)\|_2 \qquad \text{(C.4)} \\
&\leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 + \eta_t\|\partial f(\mathbf{w}_t;z_i)\|_2 + \eta_t\|\partial f(\mathbf{w}_t^{(i)};\tilde{z}_i)\|_2 \\
&\leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2 + \sqrt{2L}\eta_t\Big(\sqrt{f(\mathbf{w}_t;z_i)} + \sqrt{f(\mathbf{w}_t^{(i)};\tilde{z}_i)}\Big), \qquad \text{(C.5)}
\end{aligned}$$

where the second inequality follows from the sub-additivity of $\|\cdot\|_2$ and the last inequality is due to Lemma A.1 on the self-bounding property of smooth functions.

We first prove Eq. (C.1). Since i_t is drawn from the uniform distribution over $\{1,\ldots,n\}$, we can combine Eqs. (C.3) and (C.5) to derive

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2\big] \leq \mathbb{E}_A\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|\big] + \frac{\sqrt{2L}\eta_t}{n}\mathbb{E}_A\Big[\sqrt{f(\mathbf{w}_t;z_i)} + \sqrt{f(\mathbf{w}_t^{(i)};\tilde{z}_i)}\Big].$$

Since z_i and $\tilde{z}_i$ follow from the same distribution, we know

$$\mathbb{E}_{S,\widetilde{S},A}\big[\sqrt{f(\mathbf{w}_t^{(i)};\tilde{z}_i)}\big] = \mathbb{E}_{S,A}\big[\sqrt{f(\mathbf{w}_t;z_i)}\big]. \tag{C.6}$$

It then follows that

$$\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\big]\leq \mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|\big]+\frac{2\sqrt{2L}\eta_t}{n}\mathbb{E}_{S,A}\big[\sqrt{f(\mathbf{w}_t;z_i)}\big].$$

Taking a summation of the above inequality and using $\mathbf{w}_1=\mathbf{w}_1^{(i)}$ then give (C.1).

We now turn to (C.2). For the case $i_t=i$, it follows from (C.4) and the standard inequality $(a+b)^2\leq(1+p)a^2+(1+1/p)b^2$ that

$$\begin{aligned}\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2&\leq(1+p)\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+(1+p^{-1})\eta_t^2\|\partial f(\mathbf{w}_t;z_i)-\partial f(\mathbf{w}_t^{(i)};\tilde{z}_i)\|_2^2\\&\leq(1+p)\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+2(1+p^{-1})\eta_t^2\|\partial f(\mathbf{w}_t;z_i)\|_2^2+2(1+p^{-1})\eta_t^2\|\partial f(\mathbf{w}_t^{(i)};\tilde{z}_i)\|_2^2\\&\leq(1+p)\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+4(1+p^{-1})L\eta_t^2 f(\mathbf{w}_t;z_i)+4(1+p^{-1})L\eta_t^2 f(\mathbf{w}_t^{(i)};\tilde{z}_i),\end{aligned}\tag{C.7}$$

where the last inequality is due to Lemma A.1. Combining (C.3), the above inequality together and noticing the distribution of i_t, we derive

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]\leq(1+p/n)\mathbb{E}_A\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\big]+\frac{4(1+p^{-1})L\eta_t^2}{n}\mathbb{E}_A\big[f(\mathbf{w}_t;z_i)+f(\mathbf{w}_t^{(i)};\tilde{z}_i)\big].$$

Analogous to (C.6), we have

$$\mathbb{E}_{S,\widetilde{S},A}\big[f(\mathbf{w}_t^{(i)};\tilde{z}_i)\big]=\mathbb{E}_{S,A}\big[f(\mathbf{w}_t;z_i)\big]\tag{C.8}$$

and get

$$\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]\leq(1+p/n)\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\big]+\frac{8(1+p^{-1})L\eta_t^2}{n}\mathbb{E}_{S,A}\big[f(\mathbf{w}_t;z_i)\big].$$

Multiplying both sides by $(1+p/n)^{-(t+1)}$ yields that

$$\begin{aligned}(1+p/n)^{-(t+1)}\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]&\leq(1+p/n)^{-t}\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\big]\\&\quad+\frac{8(1+p^{-1})L(1+p/n)^{-(t+1)}\eta_t^2}{n}\mathbb{E}_{S,A}\big[f(\mathbf{w}_t;z_i)\big].\end{aligned}$$

Taking a summation of the above inequality and using $\mathbf{w}_1=\mathbf{w}_1^{(i)}$, we get

$$(1+p/n)^{-(t+1)}\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big]\leq\sum_{j=1}^t\frac{8(1+p^{-1})L(1+p/n)^{-(j+1)}\eta_j^2}{n}\mathbb{E}_{S,A}\big[f(\mathbf{w}_j;z_i)\big].$$

The stated bound then follows. □

Proof of Theorem 3. We first prove (4.3). According to Lemma C.2 (Eq. (C.1)), we know

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\Big]\leq\frac{2\sqrt{2L}}{n^2}\sum_{i=1}^n\sum_{j=1}^t\eta_j\mathbb{E}_{S,A}\big[\sqrt{f(\mathbf{w}_j;z_i)}\big].$$

It then follows from the concavity of the square-root function and the Jensen's inequality that

$$\begin{aligned}\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2\Big]&\leq\frac{2\sqrt{2L}}{n}\sum_{j=1}^t\eta_j\mathbb{E}_{S,A}\Big[\sqrt{n^{-1}\sum_{i=1}^n f(\mathbf{w}_j;z_i)}\Big]\\&=\frac{2\sqrt{2L}}{n}\sum_{j=1}^t\eta_j\mathbb{E}_{S,A}\Big[\sqrt{F_S(\mathbf{w}_j)}\Big].\end{aligned}$$

This proves (4.3).

We now turn to (4.4). It follows from Lemma C.2 (Eq. (C.2)) that

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\Big] \le \frac{8(1+p^{-1})L}{n^2}\sum_{i=1}^n\sum_{j=1}^t(1+p/n)^{t-j}\eta_j^2\mathbb{E}_{S,A}\big[f(\mathbf{w}_j;z_i)\big]$$
$$= \frac{8(1+p^{-1})L}{n}\sum_{j=1}^t(1+p/n)^{t-j}\eta_j^2\mathbb{E}_{S,A}\big[F_S(\mathbf{w}_j)\big].$$

The proof is complete. □

Proposition C.3 (Stability bounds for non-convex learning)**.** *Let Assumptions of Theorem 3 hold except that we do not require the convexity of $\mathbf{w}\mapsto f(\mathbf{w};z)$. Then for any $p>0$ we have*

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\Big] \le (1+p/n)(1+\eta_tL)^2\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\Big]+\frac{8(1+p^{-1})L\eta_t^2}{n}\mathbb{E}_{S,A}\big[F_S(\mathbf{w}_t)\big].$$

Proof. If $i_t\neq i$, then by the L-smoothness of f we know

$$\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2 \le \|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2+\eta_t\big\|\partial f(\mathbf{w}_t;z_{i_t})-\partial f(\mathbf{w}_t^{(i)},z_{i_t})\big\|_2 \le (1+\eta_tL)\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2. \tag{C.9}$$

If $i_t=i$, then analogous to (C.7), one can get

$$\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2 \le (1+p)\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2+4(1+p^{-1})L\eta_t^2f(\mathbf{w}_t;z_i)+4(1+p^{-1})L\eta_t^2f(\mathbf{w}_t^{(i)};\tilde{z}_i).$$

By the uniform distribution of $i_t\in\{1,2,\ldots,n\}$ we can combine the above inequality and (C.9) to derive

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big] \le (1+p/n)(1+\eta_tL)^2\mathbb{E}_A\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\big]+\frac{4(1+p^{-1})L\eta_t^2}{n}\mathbb{E}_A\big[f(\mathbf{w}_t;z_i)+f(\mathbf{w}_t^{(i)};\tilde{z}_i)\big].$$

This together with (C.8) implies

$$\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\big] \le (1+p/n)(1+\eta_tL)^2\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\big]+\frac{8(1+p^{-1})L\eta_t^2}{n}\mathbb{E}_{S,A}\big[f(\mathbf{w}_t;z_i)\big].$$

It then follows that

$$\begin{aligned}&\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_{t+1}-\mathbf{w}_{t+1}^{(i)}\|_2^2\Big]\\ &\le (1+p/n)(1+\eta_tL)^2\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\Big]+\frac{8(1+p^{-1})L\eta_t^2}{n^2}\sum_{i=1}^n\mathbb{E}_{S,A}\big[f(\mathbf{w}_t;z_i)\big]\\ &= (1+p/n)(1+\eta_tL)^2\mathbb{E}_A\Big[\frac{1}{n}\sum_{i=1}^n\|\mathbf{w}_t-\mathbf{w}_t^{(i)}\|_2^2\Big]+\frac{8(1+p^{-1})L\eta_t^2}{n}\mathbb{E}_{S,A}\big[F_S(\mathbf{w}_t)\big].\end{aligned}$$

The proof is complete. □

C.2. Generalization bounds

We now prove generalization bounds for SGD.

Proof of Theorem 4. According to Part (c) of Lemma A.2 with $\mathbf{w}=\mathbf{w}^*$, we know the following inequality

$$\sum_{t=1}^T\eta_t\mathbb{E}_A[F_S(\mathbf{w}_t)-F_S(\mathbf{w}^*)] \le (1/2+L\eta_1)\|\mathbf{w}^*\|_2^2+2L\sum_{t=1}^T\eta_t^2F_S(\mathbf{w}^*). \tag{C.10}$$

Let $A(S)$ be the $(t+1)$-th iterate of SGD applied to the dataset S. We plug (4.4) into Part (b) of Theorem 2, and derive

$$\mathbb{E}_{S,A}\big[F(\mathbf{w}_{t+1})\big] \leq (1+L\gamma^{-1})\mathbb{E}_{S,A}\big[F_S(\mathbf{w}_{t+1})\big] + \frac{4(1+p^{-1})(L+\gamma)L\big(1+p/n\big)^{t-1}}{n}\sum_{j=1}^{t}\eta_j^2\mathbb{E}_{S,A}\big[F_S(\mathbf{w}_j)\big].$$

We can plug (A.5) with $\mathbf{w}=\mathbf{w}^*$ into the above inequality, and derive

$$\mathbb{E}_{S,A}\big[F(\mathbf{w}_{t+1})\big] \leq (1+L\gamma^{-1})\mathbb{E}_{S,A}\big[F_S(\mathbf{w}_{t+1})\big] + \frac{4(1+p^{-1})(L+\gamma)L\big(1+p/n\big)^{t-1}}{n}\big(\eta_1\|\mathbf{w}^*\|_2^2 + 2\sum_{j=1}^{t}\eta_j^2\mathbb{E}_{S,A}[F_S(\mathbf{w}^*)]\big).$$

We choose $p=n/T$, then $(1+p/n)^{T-1}=(1+1/T)^{T-1}\leq e$ and therefore the following inequality holds for all $t=1,\ldots,T$ (note $\mathbb{E}_{S,A}[F_S(\mathbf{w}^*)]=F(\mathbf{w}^*)$)

$$\mathbb{E}_{S,A}\big[F(\mathbf{w}_{t+1})\big] \leq (1+L\gamma^{-1})\mathbb{E}_{S,A}\big[F_S(\mathbf{w}_{t+1})\big] + \frac{4(1+T/n)(L+\gamma)Le}{n}\big(\eta_1\|\mathbf{w}^*\|_2^2 + 2\sum_{j=1}^{t}\eta_j^2 F(\mathbf{w}^*)\big).$$

Multiplying both sides by η_{t+1} followed with a summation gives

$$\sum_{t=1}^{T}\eta_t\mathbb{E}_{S,A}[F(\mathbf{w}_t)] \leq \big(1+L/\gamma\big)\sum_{t=1}^{T}\eta_t\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)] + \frac{4(1+T/n)(L+\gamma)Le}{n}\sum_{t=1}^{T}\eta_t\big(\eta_1\|\mathbf{w}^*\|_2^2 + 2\sum_{j=1}^{t-1}\eta_j^2 F(\mathbf{w}^*)\big).$$

Putting (C.10) into the above inequality then gives

$$\begin{aligned}\sum_{t=1}^{T}\eta_t\mathbb{E}_{S,A}[F(\mathbf{w}_t)] \leq \big(1+L/\gamma\big)\Big(\sum_{t=1}^{T}\eta_t\mathbb{E}_{S,A}[F_S(\mathbf{w}^*)] + (1/2+L\eta_1)\|\mathbf{w}^*\|_2^2 + 2L\sum_{t=1}^{T}\eta_t^2\mathbb{E}_{S,A}[F_S(\mathbf{w}^*)]\Big)\\ + \frac{4(1+T/n)(L+\gamma)Le}{n}\sum_{t=1}^{T}\eta_t\big(\eta_1\|\mathbf{w}^*\|_2^2 + 2\sum_{j=1}^{t-1}\eta_j^2 F(\mathbf{w}^*)\big).\end{aligned}$$

Since $\mathbb{E}_S[F_S(\mathbf{w}^*)]=F(\mathbf{w}^*)$, it follows that

$$\begin{aligned}\sum_{t=1}^{T}\eta_t\mathbb{E}_{S,A}[F(\mathbf{w}_t)-F(\mathbf{w}^*)] \leq \frac{L}{\gamma}\sum_{t=1}^{T}\eta_t F(\mathbf{w}^*) + \big(1+L/\gamma\big)\Big((1/2+L\eta_1)\|\mathbf{w}^*\|_2^2 + 2L\sum_{t=1}^{T}\eta_t^2 F(\mathbf{w}^*)]\Big)\\ + \frac{4(1+T/n)(L+\gamma)Le}{n}\sum_{t=1}^{T}\eta_t\big(\eta_1\|\mathbf{w}^*\|_2^2 + 2\sum_{j=1}^{t-1}\eta_j^2 F(\mathbf{w}^*)\big).\end{aligned}$$

The stated inequality then follows from Jensen's inequality. The proof is complete. □

Proof of Corollary 5. We first prove Part (a). For the chosen step size, we know

$$\sum_{t=1}^{T}\eta_t^2 = c^2\sum_{t=1}^{T}\frac{1}{T} = c^2 \quad \text{and} \quad \sum_{t=1}^{T}\eta_t = c\sqrt{T}. \tag{C.11}$$

The stated bound (4.5) then follows from Theorem 4, $\gamma=\sqrt{n}$ and (C.11).

We now prove Part (b). The stated bound (4.6) then follows from Theorem 4, $F(\mathbf{w}^*)=0$ and $\gamma=1$. The proof is complete. □

D. Proof on Learning without Bounded Gradients: Non-smooth Case

D.1. Stability bounds

Theorem 7 is a direct application of the following general stability bounds with $p = n/t$. Therefore, it suffices to prove Theorem D.1.

Theorem D.1. *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and $\partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1)$. Let $S, \widetilde{S}$ and $S^{(i)}$ be constructed as Definition 4 and $c_{\alpha,3} = \frac{\sqrt{1-\alpha}}{\sqrt{1+\alpha}}(2^{-\alpha}L)^{\frac{1}{1-\alpha}}$. Let $\mathbf{w}_t$ and $\mathbf{w}_t^{(i)}$ be the t-th iterate produced by (3.3) based on S and $S^{(i)}$, respectively. Then for any $p > 0$ we have*

$$\mathbb{E}_{S,\widetilde{S},A}\Big[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\Big] \leq c_{\alpha,3}^2 \sum_{j=1}^{t}(1 + p/n)^{t+1-j}\eta_j^{\frac{2}{1-\alpha}} + 4(1 + p^{-1})c_{\alpha,1}^2 \sum_{j=1}^{t}\frac{(1 + p/n)^{t-j}\eta_j^2}{n}\mathbb{E}_{S,A}\Big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j)\Big]. \quad \text{(D.1)}$$

We require several lemmas to prove Theorem D.1. The following lemma establishes the co-coercivity of gradients for convex functions with Hölder continuous (sub)gradients. The case $\alpha = 1$ can be found in Nesterov (2013). The case $\alpha \in (0, 1)$ can be found in Ying & Zhou (2017). The case $\alpha = 0$ follows directly from the convexity of f (in this case, the right-hand side of (D.2) becomes 0).

Lemma D.2. *Assume for all $z \in \mathcal{Z}$, the map $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is convex, and $\mathbf{w} \mapsto \partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1]$. Then for all $\mathbf{w}, \tilde{\mathbf{w}}$ we have*

$$\big\langle \mathbf{w} - \tilde{\mathbf{w}}, \partial f(\mathbf{w}; z) - \partial f(\tilde{\mathbf{w}}; z)\big\rangle \geq \frac{2L^{-\frac{1}{\alpha}}\alpha}{1 + \alpha}\|\partial f(\mathbf{w}; z) - \partial f(\tilde{\mathbf{w}}; z)\|_2^{\frac{1+\alpha}{\alpha}}. \quad \text{(D.2)}$$

Based on Lemma D.2, we can prove Lemma 6 on the approximate contractive behavior of the gradient update associated to a non-smooth function.

Proof of Lemma 6. The following equality holds

$$\|\mathbf{w} - \eta\partial f(\mathbf{w}; z) - \tilde{\mathbf{w}} + \eta\partial f(\tilde{\mathbf{w}}; z)\|_2^2 = \|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 + \eta^2\|\partial f(\mathbf{w}; z) - \partial f(\tilde{\mathbf{w}}; z)\|_2^2 - 2\eta\langle \mathbf{w} - \tilde{\mathbf{w}}, \partial f(\mathbf{w}; z) - \partial f(\tilde{\mathbf{w}}; z)\rangle. \quad \text{(D.3)}$$

We first consider the case $\alpha = 0$. In this case, it follows from Definition 3 and Lemma D.2 with $\alpha = 0$ that

$$\|\mathbf{w} - \eta\partial f(\mathbf{w}; z) - \tilde{\mathbf{w}} + \eta\partial f(\tilde{\mathbf{w}}; z)\|_2^2 \leq \|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 + \eta^2 L^2.$$

We now consider the case $\alpha > 0$. According to Lemma D.2, we know

$$\begin{aligned}
\|\partial f(\mathbf{w}; z) - \partial f(\tilde{\mathbf{w}}; z)\|_2^2 &\leq \Big(\frac{L^{\frac{1}{\alpha}}(1 + \alpha)}{2\alpha}\big\langle \mathbf{w} - \tilde{\mathbf{w}}, \partial f(\mathbf{w}; z) - \partial f(\tilde{\mathbf{w}}; z)\big\rangle\Big)^{\frac{2\alpha}{1+\alpha}} \\
&= \Big(\frac{1 + \alpha}{\eta\alpha}\big\langle \mathbf{w} - \tilde{\mathbf{w}}, \partial f(\mathbf{w}; z) - \partial f(\tilde{\mathbf{w}}; z)\big\rangle\Big)^{\frac{2\alpha}{1+\alpha}}\Big(\eta^{\frac{2\alpha}{1+\alpha}}L^{\frac{2}{1+\alpha}}2^{-\frac{2\alpha}{1+\alpha}}\Big) \\
&\leq \frac{2\alpha}{1 + \alpha}\Big(((1 + \alpha)/(\eta\alpha))^{\frac{2\alpha}{1+\alpha}}\big\langle \mathbf{w} - \tilde{\mathbf{w}}, \partial f(\mathbf{w}; z) - \partial f(\tilde{\mathbf{w}}; z)\big\rangle^{\frac{2\alpha}{1+\alpha}}\Big)^{\frac{1+\alpha}{2\alpha}} + \frac{1 - \alpha}{1 + \alpha}\Big(\eta^{\frac{2\alpha}{1+\alpha}}L^{\frac{2}{1+\alpha}}2^{-\frac{2\alpha}{1+\alpha}}\Big)^{\frac{1+\alpha}{1-\alpha}} \\
&= 2\eta^{-1}\big\langle \mathbf{w} - \tilde{\mathbf{w}}, \partial f(\mathbf{w}; z) - \partial f(\tilde{\mathbf{w}}; z)\big\rangle + \frac{1 - \alpha}{1 + \alpha}\eta^{\frac{2\alpha}{1-\alpha}}(2^{-\alpha}L)^{\frac{2}{1-\alpha}},
\end{aligned}$$

where we have used Young's inequality (A.7). Plugging the above inequality back into (D.3), we derive

$$\|\mathbf{w} - \eta\partial f(\mathbf{w}; z) - \tilde{\mathbf{w}} + \eta\partial f(\tilde{\mathbf{w}}; z)\|_2^2 \leq \|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 + \frac{1 - \alpha}{1 + \alpha}\eta^{2+\frac{2\alpha}{1-\alpha}}(2^{-\alpha}L)^{\frac{2}{1-\alpha}}.$$

Combining the above two cases together, we get

$$\|\mathbf{w} - \eta\partial f(\mathbf{w}; z) - \tilde{\mathbf{w}} + \eta\partial f(\tilde{\mathbf{w}}; z)\|_2^2 \leq \|\mathbf{w} - \tilde{\mathbf{w}}\|_2^2 + c_{\alpha,3}^2\eta^{\frac{2}{1-\alpha}}, \quad \text{(D.4)}$$

where $c_{\alpha,3}$ is defined in Theorem D.1. The proof is complete. □

Proof of Theorem D.1. For the case $i_t \neq i$, it follows from Lemma 6 that

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 \leq \|\mathbf{w}_t - \eta_t \partial f(\mathbf{w}_t; z_{i_t}) - \mathbf{w}_t^{(i)} + \eta_t \partial f(\mathbf{w}_t^{(i)}, z_{i_t})\|_2^2 \leq \|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + c_{\alpha,3}^2 \eta_t^{\frac{2}{1-\alpha}}.$$

If $i_t = i$, by (C.4) and the standard inequality $(a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2$, we get

$$\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2 \leq (1+p)\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2 + 2(1+p^{-1})\eta_t^2\big(\|\partial f(\mathbf{w}_t; z_i)\|_2^2 + \|\partial f(\mathbf{w}_t^{(i)}; \tilde{z}_i)\|_2^2\big).$$

Combining the above two inequalities together, using the self-bounding property (Lemma A.1) and noticing the distribution of i_t, we derive

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq (1+p/n)\Big(\mathbb{E}_A\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big] + c_{\alpha,3}^2 \eta_t^{\frac{2}{1-\alpha}}\Big) + \frac{2(1+p^{-1})c_{\alpha,1}^2\eta_t^2}{n}\mathbb{E}_A\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_i) + f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; \tilde{z}_i)\big].$$

Analogous to (C.6), we know

$$\mathbb{E}_{S,\widetilde{S},A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t^{(i)}; \tilde{z}_i)\big] = \mathbb{E}_{S,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_i)\big]$$

and therefore

$$\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq (1+p/n)\Big(\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big] + c_{\alpha,3}^2 \eta_t^{\frac{2}{1-\alpha}}\Big) + \frac{4(1+p^{-1})c_{\alpha,1}^2\eta_t^2}{n}\mathbb{E}_{S,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_i)\big].$$

Multiplying both sides by $(1+p/n)^{-(t+1)}$ gives

$$\begin{aligned}(1+p/n)^{-(t+1)}\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq{}& (1+p/n)^{-t}\Big(\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_t - \mathbf{w}_t^{(i)}\|_2^2\big] + c_{\alpha,3}^2\eta_t^{\frac{2}{1-\alpha}}\Big) + \\ &\frac{4(1+p^{-1})c_{\alpha,1}^2(1+p/n)^{-(t+1)}\eta_t^2}{n}\mathbb{E}_{S,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_t; z_i)\big].\end{aligned}$$

Taking a summation of the above inequality and using $\mathbf{w}_1 = \mathbf{w}_1^{(i)}$, we derive

$$\begin{aligned}(1+p/n)^{-(t+1)}\mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq{}& c_{\alpha,3}^2\sum_{j=1}^t (1+p/n)^{-j}\eta_j^{\frac{2}{1-\alpha}} \\ &+ \frac{4(1+p^{-1})c_{\alpha,1}^2}{n}\sum_{j=1}^t (1+p/n)^{-(j+1)}\eta_j^2\mathbb{E}_{S,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j; z_i)\big].\end{aligned}$$

We can take an average over i and get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq{}& c_{\alpha,3}^2\sum_{j=1}^t (1+p/n)^{t+1-j}\eta_j^{\frac{2}{1-\alpha}} \\ &+ \frac{4(1+p^{-1})c_{\alpha,1}^2}{n^2}\sum_{i=1}^n\sum_{j=1}^t (1+p/n)^{t-j}\eta_j^2\mathbb{E}_{S,A}\big[f^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j; z_i)\big].\end{aligned}$$

It then follows from the concavity of the function $x \mapsto x^{\frac{2\alpha}{1+\alpha}}$ and the Jensen's inequality that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S,\widetilde{S},A}\big[\|\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{(i)}\|_2^2\big] \leq{}& c_{\alpha,3}^2\sum_{j=1}^t (1+p/n)^{t+1-j}\eta_j^{\frac{2}{1-\alpha}} + \\ &4(1+p^{-1})c_{\alpha,1}^2\sum_{j=1}^t \frac{(1+p/n)^{t-j}\eta_j^2}{n}\mathbb{E}_{S,A}\Big[\Big(\frac{1}{n}\sum_{i=1}^n f(\mathbf{w}_j; z_i)\Big)^{\frac{2\alpha}{1+\alpha}}\Big].\end{aligned}$$

The stated inequality then follows from the definition of F_S. The proof is complete. □

D.2. Generalization errors

Theorem 8 can be considered as an instantiation of the following proposition on generalization error bounds with specific choices of γ, T and θ. In this subsection, we first give the proof of Theorem 8 based on Proposition D.3, and then turn to the proof of Proposition D.3.

Proposition D.3. *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex, and $\partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1)$. Let $\{\mathbf{w}_t\}_t$ be produced by (3.3) with step sizes $\eta_t = cT^{-\theta}, \theta \in [0, 1]$ satisfying $\theta \geq (1-\alpha)/2$. Then for all T satisfying $n = O(T)$ and any $\gamma > 0$ we have*

$$\begin{aligned}\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O\Big(\gamma^{\frac{1+\alpha}{\alpha-1}}\Big) + O(\gamma)\Big(T^{1-\frac{2\theta}{1-\alpha}} + n^{-2}T^{\frac{2-2\theta}{1+\alpha}} + n^{-2}T^{2-2\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)\Big) \\ + O(1/\gamma)\Big(T^{-\frac{2\alpha(1-\theta)}{1+\alpha}} + F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)\Big) + O(T^{\theta-1}) + O(T^{\frac{\alpha\theta-\theta-2\alpha}{1+\alpha}}) + O(T^{-\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)).\end{aligned}$$

Proof of Theorem 8. It can be checked that θ considered in Parts (a)-(c) satisfy $\theta \geq (1-\alpha)/2$. Therefore Proposition D.3 holds. If $\gamma = \sqrt{n}$, then by Proposition D.3 we know

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O(n^{\frac{1}{2}}T^{1-\frac{2\theta}{1-\alpha}}) + O(n^{-\frac{3}{2}}T^{2-2\theta}) + O(n^{-\frac{1}{2}}) + O(T^{\theta-1}) + O(T^{\frac{\alpha\theta-\theta-2\alpha}{1+\alpha}}) + O(T^{-\theta}). \quad \text{(D.5)}$$

We first prove Part (a). Since $\alpha \geq 1/2$, $\theta = 1/2$ and $T \asymp n$, it follows from (D.5) that

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}}) + O(n^{\frac{3}{2}-\frac{1}{1-\alpha}}) + O(n^{-\frac{3\alpha+1}{2(1+\alpha)}}) = O(n^{-\frac{1}{2}}),$$

where we have used $\frac{3}{2} - \frac{1}{1-\alpha} \leq -\frac{1}{2}$ due to $\alpha \geq 1/2$. This shows Part (a).

We now prove Part (b). Since $\alpha < 1/2$, $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$ and $\theta = \frac{3-3\alpha}{2(2-\alpha)} \geq 1/2$ in (D.5), the following inequalities hold

$$\begin{aligned}&\sqrt{n}T^{1-\frac{2\theta}{1-\alpha}} \asymp \sqrt{n}T^{1-\frac{3-3\alpha}{(1-\alpha)(2-\alpha)}} \asymp \sqrt{n}n^{-\frac{(1+\alpha)(2-\alpha)}{(2-\alpha)(1+\alpha)}} \asymp n^{-\frac{1}{2}} \\ &n^{-\frac{3}{2}}T^{2-2\theta} \asymp n^{-\frac{3}{2}}T^{\frac{4-2\alpha-3+3\alpha}{2-\alpha}} \asymp n^{-\frac{3}{2}}T^{\frac{1+\alpha}{2-\alpha}} \asymp n^{-\frac{1}{2}} \\ &T^{\theta-1} \asymp T^{\frac{3-3\alpha-4+2\alpha}{2(2-\alpha)}} \asymp T^{-\frac{1+\alpha}{2(2-\alpha)}} \asymp n^{-\frac{1}{2}} \\ &T^{-\theta} \asymp n^{\frac{2-\alpha}{1+\alpha}\frac{3\alpha-3}{2(2-\alpha)}} \asymp n^{\frac{3\alpha-3}{2(1+\alpha)}} = O(n^{-\frac{1}{2}}),\end{aligned}$$

where we have used $(3\alpha-3)/(1+\alpha) \leq -1$ due to $\alpha < 1/2$. Furthermore, since $\theta \geq 1/2$ and $\alpha < 1/2$ we know $\frac{\alpha\theta-\theta-2\alpha}{1+\alpha} \leq -\frac{1}{2}$ and $T \asymp n^{\frac{2-\alpha}{1+\alpha}} \geq n$. Therefore $T^{\frac{\alpha\theta-\theta-2\alpha}{1+\alpha}} = O(T^{-\frac{1}{2}}) = O(n^{-\frac{1}{2}})$. Plugging the above inequalities into (D.5) gives the stated bound in Part (b).

We now turn to Part (c). Since $F(\mathbf{w}^*) = 0$, Proposition D.3 reduces to

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O\Big(\gamma^{\frac{1+\alpha}{\alpha-1}}\Big) + O(\gamma)\Big(T^{1-\frac{2\theta}{1-\alpha}} + n^{-2}T^{\frac{2-2\theta}{1+\alpha}}\Big) + O\Big(\gamma^{-1}T^{-\frac{2\alpha(1-\theta)}{1+\alpha}}\Big) + O(T^{\theta-1}) + O(T^{\frac{\alpha\theta-\theta-2\alpha}{1+\alpha}}).$$

With $\gamma = nT^{\theta-1}$, we further get

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O\Big(\big(n^{-1}T^{1-\theta}\big)^{\frac{1+\alpha}{1-\alpha}}\Big) + O(nT^{-\frac{(1+\alpha)\theta}{1-\alpha}}) + O(n^{-1}T^{\frac{(\theta-1)(\alpha-1)}{1+\alpha}}) + O(T^{\theta-1}) + O(T^{\frac{\alpha\theta-\theta-2\alpha}{1+\alpha}}). \quad \text{(D.6)}$$

For the choice $T = n^{\frac{2}{1+\alpha}}$ and $\theta = \frac{3-\alpha^2-2\alpha}{4}$, we know $1-\theta = (1+\alpha)^2/4$ and therefore

$$\begin{aligned}&\Big(n^{-1}T^{1-\theta}\Big)^{\frac{1+\alpha}{1-\alpha}} \asymp \Big(n^{-1}n^{\frac{2}{1+\alpha}\frac{(1+\alpha)^2}{4}}\Big)^{\frac{1+\alpha}{1-\alpha}} = n^{-\frac{1+\alpha}{2}} \\ &nT^{-\frac{(1+\alpha)\theta}{1-\alpha}} \asymp n^{1-\frac{2\theta}{1-\alpha}} = n^{1+\frac{2\alpha+\alpha^2-3}{2(1-\alpha)}} = n^{\frac{\alpha^2-1}{2-2\alpha}} = n^{-\frac{1+\alpha}{2}} \\ &n^{-1}T^{\frac{(\theta-1)(\alpha-1)}{1+\alpha}} \asymp n^{-1}T^{\frac{(1-\alpha)(1+\alpha)^2}{4(1+\alpha)}} \asymp n^{-1}T^{\frac{(1-\alpha)(1+\alpha)}{4}} \asymp n^{-\frac{1+\alpha}{2}} \\ &T^{\theta-1} \asymp n^{-\frac{2}{1+\alpha}\frac{(1+\alpha)^2}{4}} \asymp n^{-\frac{1+\alpha}{2}}.\end{aligned}$$

Furthermore,

$$(\theta-1)(1+\alpha)-(\alpha\theta-\theta-2\alpha)=2\theta+\alpha-1=2^{-1}\big(3-\alpha^2-2\alpha+2\alpha-2\big)\geq 0$$

and therefore

$$T^{\theta-1}\geq T^{\frac{\alpha\theta-\theta-2\alpha}{1+\alpha}}.$$

Plugging the above inequalities into (D.6) gives the stated bound in Part (c). The proof is complete. □

To prove Proposition D.3, we first introduce an useful lemma to address some involved series.

Lemma D.4. *Assume for all $z\in\mathcal{Z}$, the function $\mathbf{w}\mapsto f(\mathbf{w};z)$ is nonnegative, convex, and $\partial f(\mathbf{w};z)$ is (α,L)-Hölder continuous with $\alpha\in[0,1)$. Let $\{\mathbf{w}_t\}_t$ be produced by (3.3) with step sizes $\eta_t=cT^{-\theta},\theta\in[0,1]$ satisfying $\theta\geq\frac{1-\alpha}{2}$. Then*

$$\Big(\sum_{t=1}^T\eta_t^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big(\eta_1\|\mathbf{w}^*\|_2^2+2\sum_{t=1}^T\eta_t^2F(\mathbf{w}^*)+c_{\alpha,2}\sum_{t=1}^T\eta_t^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}}=O(T^{\frac{1-\alpha-2\theta}{1+\alpha}})+O(T^{1-2\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)),\tag{D.7}$$

$$\sum_{t=1}^T\eta_t^2\big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}}=O(T^{\frac{1-\alpha-2\theta}{1+\alpha}})+O(T^{1-2\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)),\tag{D.8}$$

$$\sum_{t=1}^T\eta_t\big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}}=O(T^{\frac{(1-\alpha)(1-\theta)}{1+\alpha}})+O(T^{1-\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)).\tag{D.9}$$

Proof. We first prove (D.7). For the step size sequence $\eta_t=cT^{-\theta}$, we have

$$\begin{aligned}
&\Big(\sum_{t=1}^T\eta_t^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big(\eta_1\|\mathbf{w}^*\|_2^2+2\sum_{t=1}^T\eta_t^2F(\mathbf{w}^*)+c_{\alpha,2}\sum_{t=1}^T\eta_t^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}}\\
&=O(T^{\frac{(1-2\theta)(1-\alpha)}{1+\alpha}})\Big(T^{-\theta}+T^{1-2\theta}F(\mathbf{w}^*)+T^{1-\frac{(3-\alpha)\theta}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}}\\
&=O(T^{\frac{1-\alpha-2\theta}{1+\alpha}})+O(T^{1-2\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*))+O(T^{\frac{1-\alpha-2\theta}{1-\alpha}})\\
&=O(T^{\frac{1-\alpha-2\theta}{1+\alpha}})+O(T^{1-2\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)),
\end{aligned}$$

where we have used the subadditivity of $x\mapsto x^{\frac{2\alpha}{1+\alpha}}$, the identity

$$\frac{(1-2\theta)(1-\alpha)}{1+\alpha}+\frac{2\alpha}{1+\alpha}\frac{1-\alpha-3\theta+\alpha\theta}{1-\alpha}=\frac{1-\alpha-2\theta}{1-\alpha}$$

in the second step and $\theta\geq(1-\alpha)/2$ in the third step (the third term is dominated by the first term). This shows (D.7).

We now consider (D.8). Taking an expectation over both sides of (A.8) with $\mathbf{w}=\mathbf{w}^*$, we get

$$\sum_{t=1}^T\eta_t^2\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\leq\eta_1\|\mathbf{w}^*\|_2^2+2\sum_{t=1}^T\eta_t^2\mathbb{E}_S[F_S(\mathbf{w}^*)]+c_{\alpha,2}\sum_{t=1}^T\eta_t^{\frac{3-\alpha}{1-\alpha}}.$$

According to the Jensen's inequality and the concavity of $x\mapsto x^{\frac{2\alpha}{1+\alpha}}$, we know

$$\begin{aligned}
\sum_{t=1}^T\eta_t^2\big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}}&\leq\sum_{t=1}^T\eta_t^2\Big(\frac{\sum_{t=1}^T\eta_t^2\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]}{\sum_{t=1}^T\eta_t^2}\Big)^{\frac{2\alpha}{1+\alpha}}\\
&\leq\Big(\sum_{t=1}^T\eta_t^2\Big)^{\frac{1-\alpha}{1+\alpha}}\Big(\eta_1\|\mathbf{w}^*\|_2^2+2\sum_{t=1}^T\eta_t^2F(\mathbf{w}^*)+c_{\alpha,2}\sum_{t=1}^T\eta_t^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}}\\
&=O(T^{\frac{1-\alpha-2\theta}{1+\alpha}})+O(T^{1-2\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)),
\end{aligned}$$

where we have used (D.7) in the last step. This shows (D.8).

Finally, we show (D.9). Since we consider step sizes $\eta_t = cT^{-\theta}$, it follows from (D.8) that

$$\sum_{t=1}^T \eta_t \big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}} = \big(cT^{-\theta}\big)^{-1} \sum_{t=1}^T \eta_t^2 \big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}} = O(T^{\frac{1-\alpha-2\theta}{1+\alpha}+\theta}) + O(T^{1-\theta} F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)).$$

This proves (D.9) and finishes the proof. □

Proof of Proposition D.3. Since $\mathbb{E}_S[F_S(\mathbf{w}^*)] = F(\mathbf{w}^*)$, we can decompose the excess generalization error into an estimation error and an optimization error as follows

$$\Big(\sum_{t=1}^T \eta_t\Big)^{-1} \sum_{t=1}^T \eta_t \big(\mathbb{E}_{S,A}[F(\mathbf{w}_t)] - F(\mathbf{w}^*)\big) = \Big(\sum_{t=1}^T \eta_t\Big)^{-1} \sum_{t=1}^T \eta_t \mathbb{E}_{S,A}[F(\mathbf{w}_t) - F_S(\mathbf{w}_t)] + \Big(\sum_{t=1}^T \eta_t\Big)^{-1} \sum_{t=1}^T \eta_t \mathbb{E}_{S,A}[F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*)]. \quad \text{(D.10)}$$

Our idea is to address separately the above estimation error and optimization error.

We first address **estimation errors**. Plugging (D.1) back into Theorem 2 (Part (c)) with $A(S) = \mathbf{w}_{t+1}$, we derive

$$\mathbb{E}_{S,A}\big[F(\mathbf{w}_{t+1}) - F_S(\mathbf{w}_{t+1})\big] \leq \frac{c_{\alpha,1}^2}{2\gamma} \mathbb{E}_{S,A}\Big[F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_{t+1})\Big] + 2^{-1}\gamma c_{\alpha,3}^2 \sum_{j=1}^t (1+p/n)^{t+1-j} \eta_j^{\frac{2}{1-\alpha}} + 2\gamma(1+p^{-1}) c_{\alpha,1}^2 \sum_{j=1}^t \frac{(1+p/n)^{t-j}\eta_j^2}{n} \mathbb{E}_{S,A}\Big[F_S^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_j)\Big].$$

By the concavity and sub-additivity of $x \mapsto x^{\frac{2\alpha}{1+\alpha}}$, we know

$$\mathbb{E}_{S,A}\big[F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}_{t+1})\big] \leq \big(\mathbb{E}_{S,A}[F(\mathbf{w}_{t+1})] - \mathbb{E}_{S,A}[F_S(\mathbf{w}_{t+1})] + \mathbb{E}_{S,A}[F_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}} \leq \delta_{t+1}^{\frac{2\alpha}{1+\alpha}} + \big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}},$$

where we denote $\delta_j = \max\{\mathbb{E}_{S,A}[F(\mathbf{w}_j)] - \mathbb{E}_{S,A}[F_S(\mathbf{w}_j)], 0\}$ for all $j \in \mathbb{N}$. It then follows from $p = n/T$ that

$$\delta_{t+1} \leq \frac{c_{\alpha,1}^2}{2\gamma}\Big(\delta_{t+1}^{\frac{2\alpha}{1+\alpha}} + \big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}}\Big) + 2^{-1}\gamma e c_{\alpha,3}^2 \sum_{j=1}^t \eta_j^{\frac{2}{1-\alpha}} + \frac{2e\gamma(1+T/n)c_{\alpha,1}^2}{n} \sum_{j=1}^t \eta_j^2 \big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_j)]\big)^{\frac{2\alpha}{1+\alpha}}.$$

Solving the above inequality of δ_{t+1} gives the following inequality for all $t \leq T$

$$\delta_{t+1} = O\Big(\gamma^{\frac{1+\alpha}{\alpha-1}}\Big) + O\Big(\gamma^{-1}\big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_{t+1})]\big)^{\frac{2\alpha}{1+\alpha}}\Big) + O\Big(\gamma \sum_{j=1}^t \eta_j^{\frac{2}{1-\alpha}}\Big) + O\Big(\gamma(n^{-1}+Tn^{-2}) \sum_{j=1}^t \eta_j^2 \big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_j)]\big)^{\frac{2\alpha}{1+\alpha}}\Big).$$

It then follows from the definition of δ_t that (take a summation of $\eta_t\delta_t$ and note $n = O(T)$)

$$\Big(\sum_{t=1}^T \eta_t\Big)^{-1} \sum_{t=1}^T \eta_t \big(\mathbb{E}_{S,A}[F(\mathbf{w}_t)] - \mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\big) = O\Big(\gamma^{\frac{1+\alpha}{\alpha-1}}\Big) + O\Big(\gamma^{-1}\Big(\sum_{t=1}^T \eta_t\Big)^{-1} \sum_{t=1}^T \eta_t \big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}}\Big) + O\Big(\gamma \sum_{t=1}^T \eta_t^{\frac{2}{1-\alpha}}\Big) + O\Big(\gamma T n^{-2} \sum_{t=1}^T \eta_t^2 \big(\mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\big)^{\frac{2\alpha}{1+\alpha}}\Big).$$

By $\eta_t = cT^{-\theta}$, (D.8) and (D.9), we further get

$$\Big(\sum_{t=1}^T \eta_t\Big)^{-1} \sum_{t=1}^T \eta_t \big(\mathbb{E}_{S,A}[F(\mathbf{w}_t)] - \mathbb{E}_{S,A}[F_S(\mathbf{w}_t)]\big) = O\Big(\gamma^{\frac{1+\alpha}{\alpha-1}}\Big) + O\Big(\gamma T^{1-\frac{2\theta}{1-\alpha}}\Big) + O\Big(\gamma^{-1}T^{\theta-1}\big(T^{\frac{(1-\alpha)(1-\theta)}{1+\alpha}} + T^{1-\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)\big)\Big) + O\Big(\gamma T n^{-2}\big(T^{\frac{1-\alpha-2\theta}{1+\alpha}} + T^{1-2\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)\big)\Big). \quad \text{(D.11)}$$

We now consider **optimization errors**. By Lemma A.2 (Part (d) with $\mathbf{w} = \mathbf{w}^*$) and the concavity of $x \mapsto x^{\frac{2\alpha}{1+\alpha}}$, we know

$$\begin{aligned}
&\Big(\sum_{t=1}^T \eta_t\Big)^{-1} \sum_{t=1}^T \eta_t \mathbb{E}_{S,A}[F_S(\mathbf{w}_t) - F_S(\mathbf{w}^*)] \\
&\leq \Big(2\sum_{t=1}^T \eta_t\Big)^{-1} \|\mathbf{w}^*\|_2^2 + c_{\alpha,1}^2 \Big(2\sum_{t=1}^T \eta_t\Big)^{-1} \Big(\sum_{t=1}^T \eta_t^2\Big)^{\frac{1-\alpha}{1+\alpha}} \Big(\eta_1 \|\mathbf{w}^*\|_2^2 + 2\sum_{t=1}^T \eta_t^2 F(\mathbf{w}^*) + c_{\alpha,2} \sum_{t=1}^T \eta_t^{\frac{3-\alpha}{1-\alpha}}\Big)^{\frac{2\alpha}{1+\alpha}} \\
&= O(T^{\theta-1}) + O(T^{\frac{1-\alpha-2\theta}{1+\alpha}+\theta-1}) + O(T^{-\theta} F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)),
\end{aligned}$$

where we have used (D.7) in the last step.

Plugging the above optimization error bound and the estimation error bound (D.11) back into the error decomposition (D.10), we finally derive the following generalization error bounds

$$\begin{aligned}
\Big(\sum_{t=1}^T \eta_t\Big)^{-1} \sum_{t=1}^T \eta_t \big(\mathbb{E}_{S,A}[F(\mathbf{w}_t)] - F(\mathbf{w}^*)\big) = O\Big(\gamma^{\frac{1+\alpha}{\alpha-1}}\Big) + O(\gamma)\Big(T^{1-\frac{2\theta}{1-\alpha}} + n^{-2}T^{\frac{2-2\theta}{1+\alpha}} + n^{-2}T^{2-2\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)\Big) + \\
O(1/\gamma)\Big(T^{-\frac{2\alpha(1-\theta)}{1+\alpha}} + F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)\Big) + O(T^{\theta-1}) + O(T^{\frac{\alpha\theta-\theta-2\alpha}{1+\alpha}}) + O(T^{-\theta}F^{\frac{2\alpha}{1+\alpha}}(\mathbf{w}^*)).
\end{aligned}$$

The stated inequality then follows from the convexity of F. The proof is complete. □

D.3. Empirical Risk Minimization with Strongly Convex Objectives

In this section, we present an optimistic bound for ERM with strongly convex objectives based on the ℓ_2 on-average model stability. We consider nonnegative and convex loss functions with Hölder continuous (sub)gradients.

Proposition D.5. *Assume for any z, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is nonnegative, convex and $\mathbf{w} \mapsto \partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1]$. Let A be the ERM algorithm, i.e., $A(S) = \arg\min_{\mathbf{w}\in\mathbb{R}^d} F_S(\mathbf{w})$. If for all S, F_S is σ-strongly convex, then*

$$\mathbb{E}_S\big[F(A(S)) - F_S(A(S))\big] \leq \frac{2c_{\alpha,1}^2}{n\sigma} \mathbb{E}_S\Big[F^{\frac{2\alpha}{1+\alpha}}(A(S))\Big].$$

Proof. Let $\widetilde{S}$ and $S^{(i)}, i = 1, \ldots, n$, be constructed as Definition 4. Due to the σ-strong convexity of $F_{S^{(i)}}$ and $\partial F_{S^{(i)}}(A(S^{(i)})) = 0$ (necessity condition for the optimality of $A(S^{(i)})$), we know

$$F_{S^{(i)}}(A(S)) - F_{S^{(i)}}(A(S^{(i)})) \geq 2^{-1}\sigma \big\|A(S) - A(S^{(i)})\big\|_2^2.$$

Taking a summation of the above inequality yields

$$\frac{1}{n}\sum_{i=1}^n \Big(F_{S^{(i)}}(A(S)) - F_{S^{(i)}}(A(S^{(i)}))\Big) \geq \frac{\sigma}{2n}\sum_{i=1}^n \big\|A(S) - A(S^{(i)})\big\|_2^2. \tag{D.12}$$

According to the definition of $S^{(i)}$, we know

$$\begin{aligned}
n\sum_{i=1}^n F_{S^{(i)}}(A(S)) &= \sum_{i=1}^n \Big(\sum_{j\neq i} f(A(S); z_j) + f(A(S); \tilde{z}_i)\Big) \\
&= (n-1)\sum_{j=1}^n f(A(S); z_j) + \sum_{i=1}^n f(A(S); \tilde{z}_i) = (n-1)nF_S(A(S)) + nF_{\widetilde{S}}(A(S)).
\end{aligned}$$

Taking an expectation and dividing both sides by n^2 give ($A(S)$ is independent of $\widetilde{S}$)

$$\frac{1}{n}\mathbb{E}_{S,\widetilde{S}}\Big[\sum_{i=1}^n F_{S^{(i)}}(A(S))\Big] = \frac{n-1}{n}\mathbb{E}_S\big[F_S(A(S))\big] + \frac{1}{n}\mathbb{E}_S\big[F(A(S))\big]. \tag{D.13}$$

Furthermore, by symmetry we know

$$\frac{1}{n}\mathbb{E}_{S,\widetilde{S}}\Big[\sum_{i=1}^{n} F_{S^{(i)}}(A(S^{(i)}))\Big] = \mathbb{E}_S\big[F_S(A(S))\big].$$

Plugging the above identity and (D.13) back into (D.12) gives

$$\frac{\sigma}{2n}\sum_{i=1}^{n}\mathbb{E}_{S,\widetilde{S}}\big[\|A(S^{(i)}) - A(S)\|_2^2\big] \leq \frac{1}{n}\mathbb{E}_{S,\widetilde{S}}\Big[F(A(S)) - F_S(A(S))\Big]. \tag{D.14}$$

We can now apply Part (c) of Theorem 2 to show the following inequality for all $\gamma > 0$ (notice A is a deterministic algorithm)

$$\mathbb{E}_S\big[F(A(S)) - F_S(A(S))\big] \leq \frac{c_{\alpha,1}^2}{2\gamma}\mathbb{E}_S\Big[F^{\frac{2\alpha}{1+\alpha}}(A(S))\Big] + \frac{\gamma}{n\sigma}\mathbb{E}_S\Big[F(A(S)) - F_S(A(S))\Big].$$

Taking $\gamma = n\sigma/2$, we derive

$$\mathbb{E}_S\big[F(A(S)) - F_S(A(S))\big] \leq \frac{c_{\alpha,1}^2}{n\sigma}\mathbb{E}_S\Big[F^{\frac{2\alpha}{1+\alpha}}(A(S))\Big] + \frac{1}{2}\mathbb{E}_S\Big[F(A(S)) - F_S(A(S))\Big],$$

from which we can derive the stated inequality. The proof is complete. □

E. Proofs on Stability with Relaxed Convexity

E.1. Stability and generalization errors

For any convex g, we have (Nesterov, 2013)

$$\langle \mathbf{w} - \tilde{\mathbf{w}}, \partial g(\mathbf{w}) - \partial g(\tilde{\mathbf{w}})\rangle \geq 0, \quad \mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d. \tag{E.1}$$

Proof of Theorem 9. Without loss of generality, we can assume that S and $\widetilde{S}$ differ by the first example, i.e., $z_1 \neq \tilde{z}_1$ and $z_i = \tilde{z}_i, i \neq 1$. According to the update rule (3.3) and (A.3), we know

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2^2 \leq \|\mathbf{w}_t - \eta_t\partial f(\mathbf{w}_t; z_{i_t}) - \tilde{\mathbf{w}}_t + \eta_t\partial f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t})\|_2^2 \\
&= \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2 + \eta_t^2\|\partial f(\mathbf{w}_t; z_{i_t}) - \partial f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t})\|_2^2 + 2\eta_t\langle \mathbf{w}_t - \tilde{\mathbf{w}}_t, \partial f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t}) - \partial f(\mathbf{w}_t; z_{i_t})\rangle.
\end{aligned} \tag{E.2}$$

We first study the term $\|\partial f(\mathbf{w}_t; z_{i_t}) - \partial f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t})\|_2$. The event $i_t \neq 1$ happens with probability $1 - 1/n$, and in this case it follows from the smoothness of f that ($z_{i_t} = \tilde{z}_{i_t}$)

$$\|\partial f(\mathbf{w}_t; z_{i_t}) - \partial f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t})\|_2 \leq L\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2.$$

The event $i_t = 1$ happens with probability $1/n$, and in this case

$$\|\partial f(\mathbf{w}_t; z_{i_t}) - \partial f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t})\|_2 \leq \|\partial f(\mathbf{w}_t; z_{i_t})\|_2 + \|\partial f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t})\|_2 \leq 2G.$$

Therefore, we get

$$\mathbb{E}_{i_t}\big[\|\partial f(\mathbf{w}_t; z_{i_t}) - \partial f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t})\|_2^2\big] \leq \frac{(n-1)L^2}{n}\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2 + \frac{4G^2}{n}. \tag{E.3}$$

It is clear

$$\mathbb{E}_{i_t}\big[f(\mathbf{w}_t; z_{i_t})\big] = F_S(\mathbf{w}_t) \quad \text{and} \quad \mathbb{E}_{i_t}\big[f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t})\big] = F_{\widetilde{S}}(\tilde{\mathbf{w}}_t).$$

Therefore, by (E.1) we derive

$$\begin{aligned}
\mathbb{E}_{i_t}\Big[\langle \mathbf{w}_t - \tilde{\mathbf{w}}_t, \partial f(\tilde{\mathbf{w}}_t; \tilde{z}_{i_t}) - \partial f(\mathbf{w}_t; z_{i_t})\rangle\Big] &= \langle \mathbf{w}_t - \tilde{\mathbf{w}}_t, \partial F_{\widetilde{S}}(\tilde{\mathbf{w}}_t) - \partial F_S(\mathbf{w}_t)\rangle \\
&= \langle \mathbf{w}_t - \tilde{\mathbf{w}}_t, \partial F_{\widetilde{S}}(\tilde{\mathbf{w}}_t) - \partial F_S(\tilde{\mathbf{w}}_t)\rangle + \langle \mathbf{w}_t - \tilde{\mathbf{w}}_t, \partial F_S(\tilde{\mathbf{w}}_t) - \partial F_S(\mathbf{w}_t)\rangle \\
&= \frac{1}{n}\langle \mathbf{w}_t - \tilde{\mathbf{w}}_t, \partial f(\tilde{\mathbf{w}}_t; \tilde{z}_1) - \partial f(\tilde{\mathbf{w}}_t; z_1)\rangle + \langle \mathbf{w}_t - \tilde{\mathbf{w}}_t, \partial F_S(\tilde{\mathbf{w}}_t) - \partial F_S(\mathbf{w}_t)\rangle \\
&\leq \frac{2G\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2}{n}.
\end{aligned} \tag{E.4}$$

Plugging (E.3) and the above inequality back into (E.2), we derive

$$\mathbb{E}_{i_t}\big[\|\mathbf{w}_{t+1}-\tilde{\mathbf{w}}_{t+1}\|_2^2\big] \leq \|\mathbf{w}_t-\tilde{\mathbf{w}}_t\|_2^2 + \frac{4G\eta_t\|\mathbf{w}_t-\tilde{\mathbf{w}}_t\|_2}{n} + \eta_t^2\Big(\frac{(n-1)L^2}{n}\|\mathbf{w}_t-\tilde{\mathbf{w}}_t\|_2^2 + \frac{4G^2}{n}\Big)$$

and therefore

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\tilde{\mathbf{w}}_{t+1}\|_2^2\big] \leq (1+L^2\eta_t^2)\mathbb{E}_A\big[\|\mathbf{w}_t-\tilde{\mathbf{w}}_t\|_2^2\big] + 4G\Big(\frac{\eta_t\mathbb{E}_A[\|\mathbf{w}_t-\tilde{\mathbf{w}}_t\|_2]}{n} + \frac{G\eta_t^2}{n}\Big). \tag{E.5}$$

By the above recurrence relationship and $\mathbf{w}_1=\tilde{\mathbf{w}}_1$, we derive

$$\begin{aligned}\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\tilde{\mathbf{w}}_{t+1}\|_2^2\big] &\leq 4G\sum_{j=1}^t\prod_{\tilde{j}=j+1}^t\Big(1+L^2\eta_{\tilde{j}}^2\Big)\Big(\frac{\eta_j\mathbb{E}_A[\|\mathbf{w}_j-\tilde{\mathbf{w}}_j\|_2]}{n}+\frac{G\eta_j^2}{n}\Big)\\ &\leq 4G\prod_{\tilde{j}=1}^t\Big(1+L^2\eta_{\tilde{j}}^2\Big)\sum_{j=1}^t\Big(\frac{\eta_j\max_{1\leq\tilde{j}\leq t}\mathbb{E}_A[\|\mathbf{w}_{\tilde{j}}-\tilde{\mathbf{w}}_{\tilde{j}}\|_2]}{n}+\frac{G\eta_j^2}{n}\Big).\end{aligned}$$

Since the above inequality holds for all $t\in\mathbb{N}$ and the right-hand side is an increasing function of t, we get

$$\max_{1\leq\tilde{j}\leq t+1}\mathbb{E}_A\big[\|\mathbf{w}_{\tilde{j}}-\tilde{\mathbf{w}}_{\tilde{j}}\|_2^2\big] \leq 4GC_t\sum_{j=1}^t\Big(\frac{\eta_j\max_{1\leq\tilde{j}\leq t+1}\mathbb{E}_A[\|\mathbf{w}_{\tilde{j}}-\tilde{\mathbf{w}}_{\tilde{j}}\|_2]}{n}+\frac{G\eta_j^2}{n}\Big).$$

It then follows that (note $\mathbb{E}_A[\|\mathbf{w}_{\tilde{j}}-\tilde{\mathbf{w}}_{\tilde{j}}\|_2] \leq \big(\mathbb{E}_A\big[\|\mathbf{w}_{\tilde{j}}-\tilde{\mathbf{w}}_{\tilde{j}}\|_2^2\big]\big)^{\frac{1}{2}}$)

$$\max_{1\leq\tilde{j}\leq t+1}\mathbb{E}_A\big[\|\mathbf{w}_{\tilde{j}}-\tilde{\mathbf{w}}_{\tilde{j}}\|_2^2\big] \leq 4GC_t\sum_{j=1}^t\frac{\eta_j}{n}\max_{1\leq\tilde{j}\leq t+1}\Big(\mathbb{E}_A\big[\|\mathbf{w}_{\tilde{j}}-\tilde{\mathbf{w}}_{\tilde{j}}\|_2^2\big]\Big)^{\frac{1}{2}} + 4G^2C_t\sum_{j=1}^t\frac{\eta_j^2}{n}.$$

Solving the above quadratic function of $\max_{1\leq\tilde{j}\leq t+1}\Big(\mathbb{E}_A\big[\|\mathbf{w}_{\tilde{j}}-\tilde{\mathbf{w}}_{\tilde{j}}\|_2^2\big]\Big)^{\frac{1}{2}}$ then shows

$$\max_{1\leq\tilde{j}\leq t+1}\Big(\mathbb{E}_A\big[\|\mathbf{w}_{\tilde{j}+1}-\tilde{\mathbf{w}}_{\tilde{j}+1}\|_2^2\big]\Big)^{\frac{1}{2}} \leq 4GC_t\sum_{j=1}^t\frac{\eta_j}{n} + 2G\Big(C_t\sum_{j=1}^t\frac{\eta_j^2}{n}\Big)^{\frac{1}{2}}.$$

The proof is complete. □

To prove Theorem 10, we require a basic result on series.

Lemma E.1. *We have the following elementary inequalities.*

(a) If $\theta\in(0,1)$, then $(t^{1-\theta}-1)/(1-\theta) \leq \sum_{k=1}^t k^{-\theta} \leq t^{1-\theta}/(1-\theta)$;

(b) If $\theta>1$, then $\sum_{k=1}^t k^{-\theta} \leq \frac{\theta}{\theta-1}$.

We denote by $\epsilon_{\mathrm{stab}}(A,n)$ the infimum over all ϵ for which (3.2) holds, and omit the tuple (A,n) when it is clear from the context.

Proof of Theorem 10. For the step sizes considered in both Part (a) and Part (b), one can check that $\sum_{t=1}^T\eta_t^2$ can be upper bounded by a constant independent of T. Therefore, $C_t<C$ for all $t=1,\ldots,T$ and a universal constant C. We can apply Lemma A.2 (Part (a)) on optimization errors to get

$$\mathbb{E}_A[F_S(\mathbf{w}_T^{(1)})] - F_S(\mathbf{w}^*) = O\Big(\frac{\sum_{t=1}^T\eta_t^2+\|\mathbf{w}^*\|_2^2}{\sum_{t=1}^T\eta_t}\Big). \tag{E.6}$$

By the convexity of norm, we know

$$\mathbb{E}_A\big[\|\mathbf{w}_T^{(1)} - \tilde{\mathbf{w}}_T^{(1)}\|_2\big] \leq \big(\sum_{t=1}^T \eta_t\big)^{-1} \sum_{t=1}^T \eta_t \mathbb{E}_A[\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2] = O\Big(\sum_{t=1}^T \frac{\eta_t}{n} + n^{-\frac{1}{2}}\big(\sum_{t=1}^T \eta_t^2\big)^{\frac{1}{2}}\Big),$$

where we have applied Theorem 9 (the upper bound in Theorem 9 is an increasing function of t). It then follows from the Lipschitz continuity that $\epsilon_{\text{stab}} = O\Big(\sum_{t=1}^T \frac{\eta_t}{n} + n^{-\frac{1}{2}}\big(\sum_{t=1}^T \eta_t^2\big)^{\frac{1}{2}}\Big)$. This together with the error decomposition (3.1), Lemma 1 and the optimization error bound (E.6) shows

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) = O\Big(\sum_{t=1}^T \frac{\eta_t}{n} + n^{-\frac{1}{2}}\big(\sum_{t=1}^T \eta_t^2\big)^{\frac{1}{2}}\Big) + O\Big(\frac{\sum_{t=1}^T \eta_t^2 + \|\mathbf{w}^*\|_2^2}{\sum_{t=1}^T \eta_t}\Big). \tag{E.7}$$

For the step sizes $\eta_t = \eta_1 t^{-\theta}$ with $\theta \in (1/2, 1)$, we can apply Lemma E.1 to show

$$\begin{aligned}\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(1)})] - F(\mathbf{w}^*) &= O\Big(\sum_{t=1}^T \frac{t^{-\theta}}{n} + n^{-\frac{1}{2}}\big(\sum_{t=1}^T t^{-2\theta}\big)^{\frac{1}{2}}\Big) + O\Big(\frac{\sum_{t=1}^T t^{-2\theta} + \|\mathbf{w}^*\|_2^2}{\sum_{t=1}^T t^{-\theta}}\Big) \\ &= O\Big(n^{-1}T^{1-\theta} + n^{-\frac{1}{2}} + T^{\theta-1}\Big).\end{aligned}$$

This proves the first part.

Part (b) follows by plugging (C.11) into (E.7). The proof is complete. □

F. Proofs on Stability with Relaxed Strong Convexity

Proof of Theorem 11. Due to the σ_S-strong convexity of F_S, we can analyze analogously to (E.4) to derive

$$\mathbb{E}_{i_t}\Big[\langle \mathbf{w}_t - \tilde{\mathbf{w}}_t, \partial f(\tilde{\mathbf{w}}_t; z_{i_t}) - \partial f(\mathbf{w}_t; z_{i_t})\rangle\Big] \leq \frac{2G\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2}{n} - \sigma_S\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2.$$

Therefore, analogous to the derivation of (E.5) we can derive

$$\begin{aligned}\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2^2\big] &\leq (1 + L^2\eta_t^2 - 2\sigma_S\eta_t)\mathbb{E}_A\big[\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2\big] + 4G\Big(\frac{G\eta_t^2}{n} + \frac{\eta_t\mathbb{E}_A[\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2]}{n}\Big) \\ &\leq (1 + L^2\eta_t^2 - \frac{3}{2}\sigma_S\eta_t)\mathbb{E}_A\big[\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2\big] + \frac{4G^2\eta_t^2}{n} + \frac{8G^2\eta_t}{n^2\sigma_S},\end{aligned}$$

where we have used

$$\frac{4G}{n}\mathbb{E}_A[\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2] \leq \frac{8G^2}{n^2\sigma_S} + \frac{\sigma_S\mathbb{E}_A[\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2]}{2}.$$

We find $t_0 \geq 4L^2/\sigma_S^2$. Then $\eta_t \leq \sigma_S/(2L^2)$ and it follows that

$$\begin{aligned}\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2^2\big] &\leq (1 - \sigma_S\eta_t)\mathbb{E}_A\big[\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2\big] + \frac{4G^2\eta_t^2}{n} + \frac{8G^2\eta_t}{n^2\sigma_S} \\ &= \Big(1 - \frac{2}{t + t_0}\Big)\mathbb{E}_A\big[\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2\big] + \frac{4G^2}{n}\Big(\eta_t^2 + \frac{2\eta_t}{n\sigma_S}\Big).\end{aligned}$$

Multiplying both sides by $(t + t_0)(t + t_0 - 1)$ yields

$$(t + t_0)(t + t_0 - 1)\mathbb{E}_A\big[\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2^2\big] \leq (t + t_0 - 1)(t + t_0 - 2)\mathbb{E}_A\big[\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2^2\big] + \frac{8G^2(t + t_0 - 1)}{n\sigma_S}\Big(\eta_t + \frac{2}{n\sigma_S}\Big).$$

Taking a summation of the above inequality and using $\mathbf{w}_1 = \tilde{\mathbf{w}}_1$ then give

$$\begin{aligned}(t+t_0)(t+t_0-1)\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\tilde{\mathbf{w}}_{t+1}\|_2^2\big] &\leq \frac{8G^2}{n\sigma_S}\sum_{j=1}^{t}(j+t_0-1)\Big(\eta_j+\frac{2}{n\sigma_S}\Big)\\ &= \frac{8G^2}{n\sigma_S}\Big(\sum_{j=1}^{t}(j+t_0-1)\eta_j+\frac{2}{n\sigma_S}\sum_{j=1}^{t}(j+t_0-1)\Big)\\ &\leq \frac{8G^2}{n\sigma_S}\Big(\frac{2t}{\sigma_S}+\frac{t(t+2t_0-1)}{n\sigma_S}\Big).\end{aligned}$$

It then follows

$$\mathbb{E}_A\big[\|\mathbf{w}_{t+1}-\tilde{\mathbf{w}}_{t+1}\|_2^2\big] \leq \frac{16G^2}{n\sigma_S^2}\Big(\frac{1}{t+t_0}+\frac{1}{n}\Big).$$

The stated bound then follows from the elementary inequality $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ for $a, b \geq 0$. The proof is complete. □

Proof of Theorem 12. By the convexity of norm, we know

$$\begin{aligned}\mathbb{E}_A\big[\|\mathbf{w}_T^{(2)}-\tilde{\mathbf{w}}_T^{(2)}\|_2\big] &\leq \big(\sum_{t=1}^{T}(t+t_0-1)\big)^{-1}\sum_{t=1}^{T}(t+t_0-1)\mathbb{E}_A[\|\mathbf{w}_t-\tilde{\mathbf{w}}_t\|_2]\\ &\leq \frac{4G}{\sigma_S}\big(\sum_{t=1}^{T}(t+t_0-1)\big)^{-1}\sum_{t=1}^{T}(t+t_0-1)\Big(\frac{1}{\sqrt{n(t+t_0)}}+\frac{1}{n}\Big)\\ &= O(\sigma_S^{-1}\big((nT)^{-\frac{1}{2}}+n^{-1}\big)),\end{aligned}$$

where we have used Lemma E.1 in the last step. Since the above bound holds for all $S, \widetilde{S}$ differing by a single example, it follows that ℓ_1 on-average model stability is bounded by $O(\mathbb{E}_S[\sigma_S^{-1}]\big((nT)^{-\frac{1}{2}}+n^{-1}\big))$. By Part (b) of Lemma A.2 we know

$$\mathbb{E}_A[F_S(\mathbf{w}_T^{(2)})]-F_S(\mathbf{w}^*) = O\big(1/(T\sigma_S)+\|\mathbf{w}^*\|_2^2/T^2\big).$$

It then follows from (3.1) and Part (a) of Theorem 2 that

$$\mathbb{E}_{S,A}[F(\mathbf{w}_T^{(2)})]-F(\mathbf{w}^*) = O(\mathbb{E}_S\big[\sigma_S^{-1}\big((nT)^{-\frac{1}{2}}+n^{-1}\big)\big])+O\big(\mathbb{E}_S\big[1/(T\sigma_S)\big]+1/T^2\big).$$

The stated bound holds since $T \asymp n$. The proof is complete. □

Proposition F.1. *Let $S=\{z_1,\ldots,z_n\}$ and $C_S=\frac{1}{n}\sum_{i=1}^{n}x_ix_i^\top$. Then the range of C_S is the linear span of $\{x_1,\ldots,x_n\}$.*

Proof. It suffices to show that the kernel of C_S is the orthogonal complement of $V = \text{span}\{x_1,\ldots,x_n\}$ (we denote span$\{x_1,\ldots,x_n\}$ the linear span of $x_1,\ldots,x_n$). Indeed, for any x in the kernel of C_S, we know $C_Sx=0$ and therefore $x^\top C_Sx = \frac{1}{n}\sum_{i=1}^{n}(x_i^\top x)^2 = 0$, from which we know that x must be orthogonal to V. Furthermore, for any x orthogonal to V, it is clear that $C_Sx=0$, i.e., x belongs to the kernel of C_S. The proof is complete. □

G. Extensions

In this section, we present some extensions of our analyses. We consider three extensions: extension to stochastic proximal gradient descent, extension to high probability analysis and extension to SGD without replacement.

G.1. Stochastic proximal gradient descent

Our discussions can be directly extended to study the performance of stochastic proximal gradient descent (SPGD). Let $r:\mathbb{R}^d\to\mathbb{R}^+$ be a convex regularizer. SPGD updates the models by

$$\mathbf{w}_{t+1} = \text{Prox}_{\eta_t r}\big(\mathbf{w}_t-\eta_t\partial f(\mathbf{w}_t;z_{i_t})\big),$$

where $\text{Prox}_g(\mathbf{w}) = \arg\min_{\tilde{\mathbf{w}}\in\mathbb{R}^d}\big[g(\tilde{\mathbf{w}})+\frac{1}{2}\|\mathbf{w}-\tilde{\mathbf{w}}\|_2^2\big]$ is the proximal operator. SPGD has found wide applications in solving optimization problems with a composite structure (Parikh & Boyd, 2014). It recovers the projected SGD as a specific case by taking an appropriate r. Our stability bounds for SGD can be trivially extend to SPGD due to the non-expansiveness of proximal operators: $\|\text{Prox}_g(\mathbf{w})-\text{Prox}_g(\tilde{\mathbf{w}})\|_2 \leq \|\mathbf{w}-\tilde{\mathbf{w}}\|_2, \forall \mathbf{w}, \tilde{\mathbf{w}}$ if g is convex.

G.2. Stability bounds with high probabilities

We can also extend our stability bounds stated in expectation to high-probability bounds, which would be helpful to understand the fluctuation of SGD w.r.t. different realization of random indices.

Proposition G.1. *Let Assumption 1 hold. Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is convex and $\mathbf{w} \mapsto \partial f(\mathbf{w}; z)$ is (α, L)-Hölder continuous with $\alpha \in [0, 1]$. Let $S = \{z_1, \ldots, z_n\}$ and $\widetilde{S} = \{\tilde{z}_1, \ldots, \tilde{z}_n\}$ be two sets of training examples that differ by a single example. Let $\{\mathbf{w}_t\}_t$ and $\{\tilde{\mathbf{w}}_t\}_t$ be produced by (3.3) based on S and $\widetilde{S}$, respectively, and $\delta \in (0, 1)$. If we take step size $\eta_j = ct^{-\theta}$ for $j = 1, \ldots, t$ and $c > 0$, then with probability at least $1 - \delta$*

$$\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2 = O\Big(t^{1-\frac{\theta}{1-\alpha}} + n^{-1}t^{1-\theta}\Big(1 + \sqrt{nt^{-1}\log(1/\delta)}\Big)\Big).$$

High-probability generalization bounds can be derived by combining the above stability bounds and the recent result on relating generalization and stability in a high-probability analysis (Bousquet et al., 2019; Feldman & Vondrak, 2019).

To prove Proposition G.1, we need to introduce a special concentration inequality called Chernoff's bound for a summation of independent Bernoulli random variables (Boucheron et al., 2013).

Lemma G.2 (Chernoff's Bound)**.** *Let $X_1, \ldots, X_t$ be independent random variables taking values in $\{0, 1\}$. Let $X = \sum_{j=1}^t X_j$ and $\mu = \mathbb{E}[X]$. Then for any $\tilde{\delta} \in (0, 1)$ with probability at least $1 - \exp\big(-\mu\tilde{\delta}^2/3\big)$ we have $X \leq (1 + \tilde{\delta})\mu$.*

Proof of Proposition G.1. Without loss of generality, we can assume that S and $\widetilde{S}$ differ by the first example, i.e., $z_1 \neq \tilde{z}_1$ and $z_i = \tilde{z}_i$ for $i \neq 1$. If $i_t \neq 1$, we can apply Lemma 6 and (A.3) to derive

$$\begin{aligned}\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2 &\leq \|\mathbf{w}_t - \eta_t\partial f(\mathbf{w}_t; z_{i_t}) - \tilde{\mathbf{w}}_t + \eta_t\partial f(\tilde{\mathbf{w}}_t; z_{i_t})\|_2 \\ &\leq \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2 + c_{\alpha,3}\eta_t^{\frac{1}{1-\alpha}}.\end{aligned}$$

If $i_t = 1$, we know

$$\begin{aligned}\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2 &\leq \|\mathbf{w}_t - \eta_t\partial f(\mathbf{w}_t; z_1) - \tilde{\mathbf{w}}_t + \eta_t\partial f(\tilde{\mathbf{w}}_t; \tilde{z}_1)\|_2 \\ &\leq \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2 + 2\eta_t G.\end{aligned}$$

Combining the above two cases together, we derive

$$\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2 \leq \|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2 + c_{\alpha,3}\eta_t^{\frac{1}{1-\alpha}} + 2\eta_t G\mathbb{I}_{[i_t=1]}.$$

Taking a summation of the above inequality then yields

$$\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2 \leq c_{\alpha,3}\sum_{j=1}^t \eta_j^{\frac{1}{1-\alpha}} + 2G\sum_{j=1}^t \eta_j\mathbb{I}_{[i_j=1]}.$$

Applying Lemma G.2 with $X_j = \mathbb{I}_{[i_j=1]}$ and $\mu = t/n$ (note $\mathbb{E}_A[X_j] = 1/n$), with probability $1 - \delta$ there holds

$$\sum_{j=1}^t \mathbb{I}_{[i_j=1]} \leq \frac{t}{n}\big(1 + \sqrt{3nt^{-1}\log(1/\delta)}\big).$$

Therefore, for the step size $\eta_j = ct^{-\theta}, j = 1, \ldots, t$, we know

$$\|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_{t+1}\|_2 \leq c_{\alpha,3}c^{\frac{1}{1-\alpha}}t^{1-\frac{\theta}{1-\alpha}} + 2Gcn^{-1}\big(1 + \sqrt{3nt^{-1}\log(1/\delta)}\big)t^{1-\theta}.$$

The proof is complete. □

G.3. SGD without replacement

Our stability bounds can be further extended to SGD without replacement. In this case, we run SGD in epochs. For the k-th epoch, we start with a model $\mathbf{w}_1^k \in \mathbb{R}^d$, and draw an index sequence $(i_1^k, \ldots, i_n^k)$ from the uniform distribution over all permutations of $\{1, \ldots, n\}$. Then we update the model by

$$\mathbf{w}_{t+1}^k = \mathbf{w}_t^k - \eta_t^k \partial f(\mathbf{w}_t^k; z_{i_t^k}), \quad t = 1, \ldots, n, \tag{G.1}$$

where $\{\eta_t^k\}$ is the step size sequence. We set $\mathbf{w}_1^{k+1} = \mathbf{w}_{n+1}^k$, i.e., each epoch starts with the last iterate of the previous epoch. The following proposition establishes stability bounds for SGD without replacement when applied to loss functions with Hölder continuous (sub)gradients.

Proposition G.3. *Suppose assumptions of Proposition G.1 hold. Let $\{\mathbf{w}_t\}_t$ and $\{\tilde{\mathbf{w}}_t\}_t$ be produced by (G.1) based on S and $\widetilde{S}$, respectively. Then*

$$\mathbb{E}_A[\|\mathbf{w}_1^{K+1} - \tilde{\mathbf{w}}_1^{K+1}\|_2] \leq \frac{2G}{n} \sum_{k=1}^K \sum_{t=1}^n \eta_t^k + c_{\alpha,3} \sum_{k=1}^K \sum_{t=1}^n (\eta_t^k)^{\frac{1}{1-\alpha}}.$$

Proof. Without loss of generality, we can assume that S and $\widetilde{S}$ differ by the first example, i.e., $z_1 \neq \tilde{z}_1$ and $z_i = \tilde{z}_i$ for $i \neq 1$. Analogous to the proof of Proposition G.1, we derive the following inequality for all $k \in \mathbb{N}$ and $t = 1, \ldots, n$

$$\|\mathbf{w}_{t+1}^k - \tilde{\mathbf{w}}_{t+1}^k\|_2 \leq \|\mathbf{w}_t^k - \tilde{\mathbf{w}}_t^k\|_2 + c_{\alpha,3}(\eta_t^k)^{\frac{1}{1-\alpha}} \mathbb{I}_{[i_t^k \neq 1]} + 2\eta_t^k G \mathbb{I}_{[i_t^k = 1]}.$$

Taking a summation of the above inequality from $t = 1$ to n gives

$$\|\mathbf{w}_{n+1}^k - \tilde{\mathbf{w}}_{n+1}^k\|_2 \leq \|\mathbf{w}_1^k - \tilde{\mathbf{w}}_1^k\|_2 + c_{\alpha,3} \sum_{t=1}^n (\eta_t^k)^{\frac{1}{1-\alpha}} \mathbb{I}_{[i_t^k \neq 1]} + 2G \sum_{t=1}^n \eta_t^k \mathbb{I}_{[i_t^k = 1]}.$$

Let i^k be the unique $t \in \{1, \ldots, n\}$ such that $i_t^k = 1$. Since $\mathbf{w}_1^{k+1} = \mathbf{w}_{n+1}^k$, we derive

$$\|\mathbf{w}_1^{k+1} - \tilde{\mathbf{w}}_1^{k+1}\|_2 \leq \|\mathbf{w}_1^k - \tilde{\mathbf{w}}_1^k\|_2 + c_{\alpha,3} \sum_{t=1}^n (\eta_t^k)^{\frac{1}{1-\alpha}} + 2G\eta_{i^k}^k.$$

Since we draw $(i_1^k, \ldots, i_n^k)$ from the uniform distribution of all permutations, i^k takes an equal probability to each $1, \ldots, n$. Therefore, we can take expectations over A to derive

$$\mathbb{E}_A\big[\|\mathbf{w}_1^{k+1} - \tilde{\mathbf{w}}_1^{k+1}\|_2\big] \leq \mathbb{E}_A\big[\|\mathbf{w}_1^k - \tilde{\mathbf{w}}_1^k\|_2\big] + c_{\alpha,3} \sum_{t=1}^n (\eta_t^k)^{\frac{1}{1-\alpha}} + \frac{2G \sum_{t=1}^n \eta_t^k}{n}.$$

We can take a summation of the above inequality from $k = 1$ to K to derive the stated bound. The proof is complete. □

References

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. *arXiv preprint arXiv:1910.07833*, 2019.

Feldman, V. and Vondrak, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pp. 1270–1279, 2019.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234, 2016.

Lacoste-Julien, S., Schmidt, M., and Bach, F. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Lei, Y., Hu, T., Li, G., and Tang, K. Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Transactions on Neural Networks and Learning Systems*, 2019. doi: 10.1109/TNNLS.2019.2952219.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

Parikh, N. and Boyd, S. P. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pp. 2199–2207, 2010.

Ying, Y. and Zhou, D.-X. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224—-244, 2017.