Sharper Generalization Bounds for Pairwise Learning: Supplementary Material

A Proof of Theorem 1

To prove Theorem 1, we need to introduce some lemmas. The following lemma is attributed to [7], which provides far-reaching moment bounds for a summation of weakly dependent and mean-zero random functions with bounded increments under a change of any single coordinate. We denote $S\backslash\{z_i\}$ the set $\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$. The L_p-norm of a random variable Z is denoted by $\|Z\|_p := \big(\mathbb{E}[|Z|^p]\big)^{\frac{1}{p}}, p \geq 1$.

Lemma A.1 ([4]). *Let $S = \{z_1, \ldots, z_n\}$ be a set of independent random variables each taking values in $\mathcal{Z}$ and $M > 0$. Let $g_1, \ldots, g_n$ be some functions $g_i : \mathcal{Z}^n \mapsto \mathbb{R}$ such that the following holds for any $i \in [n]$*

- $\big|\mathbb{E}_{S\backslash\{z_i\}}[g_i(S)]\big| \leq M$ *almost surely (a.s.),*
- $\mathbb{E}_{z_i}\big[g_i(S)\big] = 0$ *a.s.,*
- *for any $j \in [n]$ with $j \neq i$, and $z_j'' \in \mathcal{Z}$*

$$\big|g_i(S) - g_i(z_1, \ldots, z_{j-1}, z_j'', z_{j+1}, \ldots, z_n)\big| \leq \beta. \tag{A.1}$$

Then, for any $p \geq 2$

$$\Big\| \sum_{i=1}^{n} g_i(S) \Big\|_p \leq 12\sqrt{6}pn\beta\lceil \log_2 n \rceil + 3\sqrt{2}M\sqrt{pn}.$$

The bounds on moments of random variables can be used to establish concentration inequalities, as shown in the following lemma [4, 16].

Lemma A.2. *Let $a, b \in \mathbb{R}_+$ and $\delta \in (0, 1/e)$. Let Z be a random variable with $\|Z\|_p \leq \sqrt{p}a + pb$ for any $p \geq 2$. Then with probability at least $1 - \delta$*

$$|Z| \leq e\Big(a\sqrt{\log(e/\delta)} + b\log(e/\delta)\Big).$$

The following lemma controls the change on the output of stable algorithms if we perturb a training dataset by two examples.

Lemma A.3. *Let $A : \mathcal{Z}^n \mapsto \mathcal{W}$ be γ-uniformly stable. Then for any $S' = \{z_1', \ldots, z_n'\}$ and $i \neq j$, we have*

$$\sup_{z,\tilde{z}\in\mathcal{Z}} \big|\ell(A(S); z, \tilde{z}) - \ell(A(S_{i,j}); z, \tilde{z})\big| \leq 2\gamma,$$

where $S_{i,j}$ is defined in (3.4).

Proof. For any $i \in [n]$, introduce

$$S_i = \{z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n\}. \tag{A.2}$$

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Note that S, S_i differ only by a single example, and $S_i, S_{i,j}$ differ only by a single example. It then follows from the definition of uniform stability that

$$\begin{aligned}
&\sup_{z,\tilde{z}\in\mathcal{Z}} \big|\ell(A(S);z,\tilde{z}) - \ell(A(S_{i,j});z,\tilde{z})\big| \\
&\leq \sup_{z,\tilde{z}\in\mathcal{Z}} \big|\ell(A(S);z,\tilde{z}) - \ell(A(S_i);z,\tilde{z})\big| + \sup_{z,\tilde{z}\in\mathcal{Z}} \big|\ell(A(S_i);z,\tilde{z}) - \ell(A(S_{i,j});z,\tilde{z})\big| \\
&\leq 2\gamma.
\end{aligned}$$

The proof is complete. □

With these lemmas, we can give the proof of Theorem 1 on high-probability bounds of the generalization gap. The concentration inequality established in Lemma A.1 applies to a summation of n random functions involving n independent random variables, which does not apply to the objective function in pairwise learning since it is a U-statistic. We introduce a novel decomposition to exploit the structure of pairwise learning problems. We abbreviate $\sum_{i,j\in[n]:i\neq j}$ as $\sum_{i\neq j}$.

Proof of Theorem 1. Let $p \geq 2$ be any number. We can decompose the generalization gap associated to $A(S)$ as follows

$$\begin{aligned}
&n(n-1)\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S);Z,\tilde{Z})\big] - \sum_{i\neq j}\ell(A(S);z_i,z_j) = \sum_{i\neq j}\mathbb{E}_{Z,\tilde{Z}}\Big[\ell(A(S);Z,\tilde{Z}) - \mathbb{E}_{z_i',z_j'}[\ell(A(S_{i,j});Z,\tilde{Z})]\Big] \\
&+ \sum_{i\neq j}\mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \ell(A(S_{i,j});z_i,z_j)\Big] + \sum_{i\neq j}\mathbb{E}_{z_i',z_j'}\Big[\ell(A(S_{i,j});z_i,z_j) - \ell(A(S);z_i,z_j)\Big],
\end{aligned}$$

where $S_{i,j}$ is defined in (3.4). According to Lemma A.3, we know

$$\Big|\ell(A(S);Z,\tilde{Z}) - \mathbb{E}_{z_i',z_j'}[\ell(A(S_{i,j});Z,\tilde{Z})]\Big| \leq 2\gamma$$

and

$$\big|\ell(A(S_{i,j});z_i,z_j) - \ell(A(S);z_i,z_j)\big| \leq 2\gamma.$$

Therefore,

$$\Big|n(n-1)\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S);Z,\tilde{Z})\big] - \sum_{i\neq j}\ell(A(S);z_i,z_j)\Big| \leq 4n(n-1)\gamma + \big|\sum_{i\neq j} g_{i,j}(S)\big|, \quad \text{(A.3)}$$

where we introduce

$$g_{i,j}(S) = \mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \ell(A(S_{i,j});z_i,z_j)\Big], \quad \forall i,j\in[n].$$

For any $i,j\in[n]$, we can further decompose $g_{i,j}$ as $g_{i,j} = g_j^{(i)} + \tilde{g}_i^{(j)}$, where (we omit the argument S for brevity)

$$\begin{aligned}
g_j^{(i)} &= \mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \mathbb{E}_Z[\ell(A(S_{i,j});Z,z_j)]\Big] \\
\tilde{g}_i^{(j)} &= \mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_Z\big[\ell(A(S_{i,j});Z,z_j\big] - \ell(A(S_{i,j});z_i,z_j)\Big].
\end{aligned}$$

Let us temporarily fix i, and consider $n-1$ random functions $g_1^{(i)},\ldots,g_{i-1}^{(i)},g_{i+1}^{(i)},\ldots,g_n^{(i)}$. According to the assumption $\big|\mathbb{E}_S[\ell(A(S);z,\tilde{z})]\big| \leq M$ for all $z,\tilde{z}$, we know

$$\big|\mathbb{E}_{S\setminus\{z_j\}}[g_j^{(i)}(S)]\big| \leq 2M, \quad \forall j\in[n].$$

For any $j\neq i$, since z_j is independent of $S_{i,j}$ we know

$$\mathbb{E}_{z_j}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \mathbb{E}_Z[\ell(A(S_{i,j});Z,z_j)]\Big] = 0.$$

Therefore, $\mathbb{E}_{z_j}[g_j^{(i)}] = 0$. For any $k\neq j$ and any $z_k''\in\mathcal{Z}$, it is clear from the uniform stability of A that

$$\Big|\mathbb{E}_{z_i',z_j'}\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \mathbb{E}_{z_i',z_j'}\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j}^{(k)});Z,\tilde{Z})\big]\Big| \leq \gamma,$$

where $S_{i,j}^{(k)}$ is the set derived by replacing the k-th element of $S_{i,j}$ with z_k''. Similarly, one have

$$\Big|\mathbb{E}_{z_i',z_j'}\mathbb{E}_Z[\ell(A(S_{i,j});Z,z_j)] - \mathbb{E}_{z_i',z_j'}\mathbb{E}_Z[\ell(A(S_{i,j}^{(k)});Z,z_j)]\Big| \leq \gamma.$$

It then follows from the above two inequalities that $g_j^{(i)}$ satisfies the bounded increment condition (A.1) with $\beta = 2\gamma$ for all $k \neq j$, i.e.,

$$\Big|\mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j});Z,\tilde{Z})\big] - \mathbb{E}_Z[\ell(A(S_{i,j});Z,z_j)]\Big] \\ - \mathbb{E}_{z_i',z_j'}\Big[\mathbb{E}_{Z,\tilde{Z}}\big[\ell(A(S_{i,j}^{(k)});Z,\tilde{Z})\big] - \mathbb{E}_Z[\ell(A(S_{i,j}^{(k)});Z,z_j)]\Big]\Big| \leq 2\gamma.$$

Therefore, all the assumptions of Lemma A.1 hold for the random functions $g_1^{(i)},\ldots,g_{i-1}^{(i)},g_{i+1}^{(i)},\ldots,g_n^{(i)}$ with n there replaced by $n-1$ and $\beta = 2\gamma$. We can apply Lemma A.1 to derive

$$\Big\|\sum_{j\in[n],j\neq i} g_j^{(i)}\Big\|_p \leq 24\sqrt{6}p(n-1)\gamma\lceil\log_2(n-1)\rceil + 6\sqrt{2}M\sqrt{p(n-1)}, \quad \forall i\in[n].$$

Similarly, we can also show that

$$\Big\|\sum_{i\in[n],i\neq j} \tilde{g}_i^{(j)}\Big\|_p \leq 24\sqrt{6}p(n-1)\gamma\lceil\log_2(n-1)\rceil + 6\sqrt{2}M\sqrt{p(n-1)}, \quad \forall j\in[n].$$

It then follows from the subadditivity of $\|\cdot\|_p$ and the above two inequalities that

$$\begin{aligned}\|\sum_{i\neq j} g_{i,j}\|_p &\leq \Big\|\sum_{i\neq j} g_j^{(i)}\Big\|_p + \Big\|\sum_{i\neq j}\tilde{g}_i^{(j)}\Big\|_p \\ &\leq \sum_{i\in[n]}\Big\|\sum_{j\in[n],j\neq i} g_j^{(i)}\Big\|_p + \sum_{j\in[n]}\Big\|\sum_{i\in[n],i\neq j}\tilde{g}_i^{(j)}\Big\|_p \\ &\leq 48\sqrt{6}p(n-1)n\gamma\lceil\log_2(n-1)\rceil + 12\sqrt{2}M\sqrt{p(n-1)}n.\end{aligned}$$

We can combine the above p-norm and Lemma A.2 to derive the following inequality with probability at least $1-\delta$

$$\Big|\sum_{i\neq j} g_{i,j}\Big| \leq e\Big(12\sqrt{2}M\sqrt{(n-1)}n\sqrt{\log(e/\delta)} + 48\sqrt{6}(n-1)n\gamma\lceil\log_2(n-1)\rceil\log(e/\delta)\Big).$$

Plugging the above inequality back into (A.3) and using the definition of R_S, R, we derive the following inequality with probability at least $1-\delta$

$$\begin{aligned}|R_S(A(S)) - R(A(S))| &\leq 4\gamma + \frac{1}{n(n-1)}\Big|\sum_{i\neq j} g_{i,j}\Big| \\ &\leq 4\gamma + e\Big(12\sqrt{2}M(n-1)^{-\frac{1}{2}}\sqrt{\log(e/\delta)} + 48\sqrt{6}\gamma\lceil\log_2(n-1)\rceil\log(e/\delta)\Big).\end{aligned}$$

The proof is complete. □

B Proof of Theorem 3

In this section, we prove Theorem 3 on high-probability bounds for learning with strongly convex objective functions. We first prove Lemma 2 on the norm of output model.

Proof of Lemma 2. Since $A(S)$ is the minimizer of F_S, we know there is a $F_S'(A(S)) = 0$ (F_S' is a subgradient of F_S at $A(S)$). This together with the definition of strong convexity implies

$$R_S(\mathbf{w}^*) + r(\mathbf{w}^*) - R_S(A(S)) - r(A(S)) \geq \frac{\sigma}{2}\|A(S) - \mathbf{w}^*\|^2. \tag{B.1}$$

Analogous to (A.3), we know

$$n(n-1)\Big(R(A(S)) - R_S(A(S))\Big) \le 4n(n-1)\gamma + \sum_{i,j\in[n]:i\neq j} g_{i,j},$$

where $g_{i,j}$ is defined in the proof of Theorem 1. In the proof of Theorem 1, we have shown $\mathbb{E}[g_{i,j}] = 0$. It then follows that

$$\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \le 4\gamma.$$

We can plug the above inequality back into (B.1) to derive

$$\begin{aligned}\frac{\sigma}{2}\mathbb{E}\big[\|A(S)-\mathbf{w}^*\|^2\big] &\le \mathbb{E}\big[R_S(\mathbf{w}^*) + r(\mathbf{w}^*) - R_S(A(S)) - r(A(S))\big] \\ &\le \mathbb{E}\big[R_S(\mathbf{w}^*) + r(\mathbf{w}^*) - R(A(S)) - r(A(S))\big] + 4\gamma \\ &= \mathbb{E}\big[R(\mathbf{w}^*) + r(\mathbf{w}^*) - R(A(S)) - r(A(S))\big] + 4\gamma \le 4\gamma,\end{aligned}$$

where the last inequality holds since $\mathbf{w}^*$ minimizes $F = R + r$. The stated inequality then follows and finishes the proof. □

To prove Theorem 3, we introduce some lemmas.

Lemma B.1. *For any $S \in \mathcal{Z}^n$, define A as $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$. For any $k \in [n]$, let S_k be defined by* (A.2)*. Then*

$$\begin{aligned}&F_S(A(S_k)) - F_S(A(S)) \le \\ &\frac{1}{n(n-1)}\sum_{i\in[n]:i\neq k}\Big(\big(\ell(A(S_k);z_i,z_k) - \ell(A(S);z_i,z_k)\big) + \big(\ell(A(S_k);z_k,z_i) - \ell(A(S);z_k,z_i)\big) \\ &\qquad + \big(\ell(A(S);z_i,z_k') - \ell(A(S_k);z_i,z_k')\big) + \big(\ell(A(S);z_k',z_i) - \ell(A(S_k);z_k',z_i)\big)\Big).\end{aligned}$$

Proof. Without loss of generality, we can assume $k = n$. Since $A(S_n)$ is a minimizer of F_{S_n}, we know

$$\begin{aligned}&F_S(A(S_n)) - F_S(A(S)) \\ &= F_S(A(S_n)) - F_{S_n}(A(S_n)) + F_{S_n}(A(S_n)) - F_{S_n}(A(S)) + F_{S_n}(A(S)) - F_S(A(S)) \\ &\le F_S(A(S_n)) - F_{S_n}(A(S_n)) + F_{S_n}(A(S)) - F_S(A(S)).\end{aligned} \tag{B.2}$$

By the definition of F_S and F_{S_n}, we know

$$\begin{aligned}&n(n-1)\big(F_S(A(S_n)) - F_{S_n}(A(S_n))\big) = \sum_{i,j\in[n]:i\neq j} f(A(S_n);z_i,z_j) \\ &- \Big(\sum_{i,j\in[n-1]:i\neq j} f(A(S_n);z_i,z_j) + \sum_{i\in[n-1]} f(A(S_n);z_i,z_n') + \sum_{i\in[n-1]} f(A(S_n);z_n',z_i)\Big) \\ &= \sum_{i\in[n-1]}\Big(f(A(S_n);z_i,z_n) + f(A(S_n);z_n,z_i) - f(A(S_n);z_i,z_n') - f(A(S_n);z_n',z_i)\Big).\end{aligned}$$

Similarly, we know

$$\begin{aligned}&n(n-1)\big(F_{S_n}(A(S)) - F_S(A(S))\big) = \\ &\qquad\sum_{i\in[n-1]}\Big(f(A(S);z_i,z_n') + f(A(S);z_n',z_i) - f(A(S);z_i,z_n) - f(A(S);z_n,z_i)\Big).\end{aligned}$$

Therefore, we can combine the above two identities to derive

$$\begin{aligned}&n(n-1)\big(F_S(A(S_n)) - F_{S_n}(A(S_n)) + F_{S_n}(A(S)) - F_S(A(S))\big) = \\ &\sum_{i\in[n-1]}\Big(\big(f(A(S_n);z_i,z_n) - f(A(S);z_i,z_n)\big) + \big(f(A(S_n);z_n,z_i) - f(A(S);z_n,z_i)\big) + \\ &\qquad\big(f(A(S);z_i,z_n') - f(A(S_n);z_i,z_n')\big) + \big(f(A(S);z_n',z_i) - f(A(S_n);z_n',z_i)\big)\Big).\end{aligned}$$

This together with the structure of f ($f = \ell + r$ with r depending only on $\mathbf{w}$) implies

$$
\begin{aligned}
&n(n-1)\big(F_S(A(S_n)) - F_{S_n}(A(S_n)) + F_{S_n}(A(S)) - F_S(A(S))\big) = \\
&\quad \sum_{i\in[n-1]} \Big(\big(\ell(A(S_n); z_i, z_n) - \ell(A(S); z_i, z_n)\big) + \big(\ell(A(S_n); z_n, z_i) - \ell(A(S); z_n, z_i)\big) + \\
&\qquad\qquad \big(\ell(A(S); z_i, z_n') - \ell(A(S_n); z_i, z_n')\big) + \big(\ell(A(S); z_n', z_i) - \ell(A(S_n); z_n', z_i)\big)\Big).
\end{aligned}
$$

Plugging the above identity back into (B.2), we derive

$$
\begin{aligned}
&F_S(A(S_n)) - F_S(A(S)) \\
&\leq \frac{1}{n(n-1)} \sum_{i\in[n-1]} \Big(\big(\ell(A(S_n); z_i, z_n) - \ell(A(S); z_i, z_n)\big) + \big(\ell(A(S_n); z_n, z_i) - \ell(A(S); z_n, z_i)\big) \\
&\qquad + \big(\ell(A(S); z_i, z_n') - \ell(A(S_n); z_i, z_n')\big) + \big(\ell(A(S); z_n', z_i) - \ell(A(S_n); z_n', z_i)\big)\Big).
\end{aligned}
$$

The proof is complete. □

The following lemma establishes the uniform stability of pairwise learning with strongly convex objectives.

Lemma B.2. *Define A as $A(S) = \arg\min_{\mathbf{w}\in\mathcal{W}} F_S(\mathbf{w})$. Suppose F_S is σ-strongly convex w.r.t. $\|\cdot\|$. Assume for all $z, \tilde{z}$ we have* (4.3)*. Then A is $\frac{8L^2}{n\sigma}$-uniformly stable.*

Proof. Let S, S' be two sets that differ by a single example and let $\mathbf{w}_S = A(S)$ and $\mathbf{w}_{S'} = A(S')$. Without loss of generality, we can assume $S' = \{z_1, \ldots, z_{n-1}, z_n'\}$, i.e., S and S' differ by the last example.

Since $\mathbf{w}_S$ is a minimizer of F_S we know there is a subgradient $F_S'(\mathbf{w}_S) = 0$, which together with the σ-strong convexity of F_S, implies

$$
F_S(\mathbf{w}_{S'}) - F_S(\mathbf{w}_S) \geq \frac{\sigma}{2}\|\mathbf{w}_{S'} - \mathbf{w}_S\|^2. \tag{B.3}
$$

According to (4.3) and Lemma B.1, we know

$$
F_S(\mathbf{w}_{S'}) - F_S(\mathbf{w}_S) \leq \frac{4(n-1)L\|\mathbf{w}_S - \mathbf{w}_{S'}\|}{n(n-1)}.
$$

which, together with (B.3), implies

$$
\|\mathbf{w}_S - \mathbf{w}_{S'}\| \leq \frac{8L}{n\sigma}.
$$

This further together with (4.3) implies the $\frac{8L^2}{n\sigma}$-uniform stability of A. The proof is complete. □

To obtain tight control on the term $R(\mathbf{w}^*) - R_S(\mathbf{w}^*)$, we will need a version of Bernstein's inequality for U-statistics. The following theorem is attributed to [10], and can be found in [5] (inequality A.1 on page 868), and in [12] (Theorem 2). A complete proof is provided in [13] (page 4).

Lemma B.3 (Bernstein's inequality for U-Statistic [10, 13])**.** *Let $Z_1, \ldots, \ldots, Z_n$ be independent variables taking values in $\mathcal{Z}$ and $q : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$. Let $b = \sup_{z,\tilde{z}} |q(z, \tilde{z})|$ and σ_0^2 be the variance of $q(Z, \tilde{Z})$. Then for any $\delta \in (0, 1)$ with probability at least $1 - \delta$*

$$
\Big|\frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} q(Z_i, Z_j) - \mathbb{E}_{Z,\tilde{Z}}[q(Z, \tilde{Z})]\Big| \leq \frac{2b\log(1/\delta)}{3\lfloor n/2\rfloor} + \sqrt{\frac{2\sigma_0^2\log(1/\delta)}{\lfloor n/2\rfloor}}. \tag{B.4}
$$

We now give the proof of Theorem 3.

Proof of Theorem 3. According to Lemma B.2, we know that A is $\frac{8L^2}{n\sigma}$-uniformly stable. Using this together with Lemma 2 we derive $\mathbb{E}_S[\|\mathbf{w}^* - A(S)\|^2] \leq \frac{64L^2}{n\sigma^2}$ and therefore

$$\mathbb{E}_S[\|\mathbf{w}^* - A(S)\|] \leq \Big(\mathbb{E}_S[\|\mathbf{w}^* - A(S)\|^2]\Big)^{\frac{1}{2}} \leq \frac{8L}{\sqrt{n}\sigma}. \tag{B.5}$$

For any $\mathbf{w} \in \mathcal{W}$ and $z, \tilde{z}$, define

$$\tilde{\ell}(\mathbf{w}; z, \tilde{z}) = \ell(\mathbf{w}; z, \tilde{z}) - \ell(\mathbf{w}^*; z, \tilde{z}).$$

Then it is clear from Lemma B.2 that A is also $\frac{8L^2}{n\sigma}$-uniformly stable when measured by the "loss" $\tilde{\ell}$, i.e., for any S, S' differing by one example

$$\begin{aligned}
&\sup_{z,\tilde{z}} \big|\tilde{\ell}(A(S); z, \tilde{z}) - \tilde{\ell}(A(S'); z, \tilde{z})\big| \\
&= \sup_{z,\tilde{z}} \big|\ell(A(S); z, \tilde{z}) - \ell(A(S'); z, \tilde{z}) - \ell(\mathbf{w}^*; z, \tilde{z}) + \ell(\mathbf{w}^*; z, \tilde{z})\big| \\
&= \sup_{z,\tilde{z}} \big|\ell(A(S); z, \tilde{z}) - \ell(A(S'); z, \tilde{z})\big| \leq 8L^2/(n\sigma).
\end{aligned}$$

Furthermore, by the Lipschitz continuity (4.3) and (B.5), we know the following inequality for all $z, \tilde{z} \in \mathcal{Z}$

$$\begin{aligned}
\Big|\mathbb{E}_S\big[\tilde{\ell}(A(S); z, \tilde{z})\big]\Big| &= \Big|\mathbb{E}_S\big[\ell(A(S); z, \tilde{z}) - \ell(\mathbf{w}^*; z, \tilde{z})\big]\Big| \\
&\leq L\mathbb{E}_S\big[\|\mathbf{w}^* - A(S)\|\big] \leq \frac{8L^2}{\sqrt{n}\sigma}.
\end{aligned}$$

We can now apply Theorem 1, with $\gamma = \frac{8L^2}{n\sigma}$, $M = 8L^2/(\sqrt{n}\sigma)$ and ℓ replaced by $\tilde{\ell}$, and show the following inequality with probability $1 - \delta/2$

$$\begin{aligned}
\Big|\frac{1}{n(n-1)}\sum_{i\neq j}\tilde{\ell}(A(S); z_i, z_j) - \mathbb{E}_{z,\tilde{z}}\big[\tilde{\ell}(A(S); z, \tilde{z})\big]\Big| &\leq \frac{32L^2}{n\sigma} \\
&+ e\Big(\frac{96\sqrt{2}L^2\sqrt{\log(2e/\delta)}}{\sqrt{n(n-1)}\sigma}\Big) + \frac{384\sqrt{6}L^2\lceil\log_2 n\rceil\log(2e/\delta)}{n\sigma},
\end{aligned}$$

from which we derive the following inequality with probability $1 - \delta/2$

$$\begin{aligned}
\big|R_S(A(S)) - R(A(S))\big| &\leq \Big|\frac{1}{n(n-1)}\sum_{i\neq j}\ell(\mathbf{w}^*; z_i, z_j) - \mathbb{E}_{z,\tilde{z}}\big[\ell(\mathbf{w}^*; z, \tilde{z})\big]\Big| \\
&+ \frac{32L^2}{n\sigma}\Big(1 + 3\sqrt{\frac{2n\log(2e/\delta)}{n-1}} + 12\sqrt{6}\lceil\log_2 n\rceil\log(2e/\delta)\Big).
\end{aligned} \tag{B.6}$$

By to the definition of $\mathbf{w}^*$ ($R'(\mathbf{w}^*) + r'(\mathbf{w}^*) = 0$), the σ-strong convexity and Assumption 1 ($0 \leq \ell(0; z, \tilde{z})$), we know

$$\frac{\sigma\|\mathbf{w}^*\|^2}{2} \leq R(0) + r(0) - R(\mathbf{w}^*) - r(\mathbf{w}^*) \Longrightarrow \|\mathbf{w}^*\| \leq \sqrt{\frac{2(R(0) + r(0))}{\sigma}}.$$

It then follows from the Lipschitz continuity (4.3) that

$$\begin{aligned}
|\ell(\mathbf{w}^*; z, \tilde{z})| &= \big|\ell(\mathbf{w}^*; z, \tilde{z}) - \ell(0; z, \tilde{z}) + \ell(0; z, \tilde{z})\big| \leq L\|\mathbf{w}^*\| + \sup_{z,\tilde{z}}\ell(0; z, \tilde{z}) \\
&\leq L\sqrt{\frac{2(R(0) + r(0))}{\sigma}} + \sup_{z,\tilde{z}}\ell(0; z, \tilde{z}).
\end{aligned}$$

According to Bernstein's inequality (B.4), we derive the following inequality with probability $1 - \delta/2$ that

$$\begin{aligned}
\Big|\frac{1}{n(n-1)}\sum_{i\neq j}\ell(\mathbf{w}^*; z_i, z_j) - \mathbb{E}_{z,\tilde{z}}[\ell(\mathbf{w}^*; z, \tilde{z})]\Big| &\leq \\
\frac{2\big(L\sqrt{2(R(0) + r(0))/\sigma} + b\big)\log(2/\delta)}{3\lfloor n/2\rfloor} &+ \sqrt{\frac{2\sigma_0^2\log(2/\delta)}{\lfloor n/2\rfloor}}.
\end{aligned} \tag{B.7}$$

Plugging the above inequality back into (B.6), we derive the following inequality with probability at least $1-\delta$

$$\big|R_S(A(S)) - R(A(S))\big| \leq \frac{2\big(L\sqrt{2(R(0)+r(0))/\sigma}+b\big)\log(2/\delta)}{3\lfloor n/2\rfloor} + \sqrt{\frac{2\sigma_0^2\log(2/\delta)}{\lfloor n/2\rfloor}} + \frac{32L^2}{n\sigma}\Big(1+3\sqrt{\frac{2n\log(2e/\delta)}{n-1}}+12\sqrt{6}\lceil\log_2 n\rceil\log(2e/\delta)\Big).$$

The above inequality can be written as the stated bound (4.4).

We now turn to (4.5). According to the definition of R and F, we can decompose the excess risk $R(A(S)) - R(\mathbf{w}_R^*)$ as follows

$$\begin{aligned}
&R(A(S)) - R(\mathbf{w}_R^*)\\
&= R(A(S)) - R_S(A(S)) + R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) + R_S(A(S)) - R_S(\mathbf{w}_R^*)\\
&= R(A(S)) - R_S(A(S)) + R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) + F_S(A(S)) - F_S(\mathbf{w}_R^*) + r(\mathbf{w}_R^*) - r(A(S))\\
&\leq R(A(S)) - R_S(A(S)) + R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) + O(\sigma\|\mathbf{w}_R^*\|^2) - r(A(S)), \qquad \text{(B.8)}
\end{aligned}$$

where we have used the inequality $F_S(A(S)) \leq F_S(\mathbf{w}_R^*)$ due to the definition of $A(S)$ and the assumption $r(\mathbf{w}) = O(\sigma\|\mathbf{w}\|^2)$ in the last step. Analogous to (B.7), one can use Bernstein's inequality (Lemma B.3) to show with probability at least $1-\delta/2$ that (under a very mild assumption $\sup_{z,z'} \ell(\mathbf{w}_R^*; z, z') = O(\sqrt{n})$)

$$R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) = O\Big(\frac{\log(1/\delta)}{\sqrt{n}} + \sqrt{\frac{\sigma_0^2\log(1/\delta)}{n}}\Big). \qquad \text{(B.9)}$$

Plugging the above inequality and (4.4) back into (B.8) shows the following inequality with probability at least $1-\delta$

$$R(A(S)) - R(\mathbf{w}_R^*) = O\Big((n\sigma)^{-1}\log n\log(1/\delta) + n^{-\frac{1}{2}}\log(1/\delta)\Big) + O(\sigma\|\mathbf{w}_R^*\|^2).$$

The stated bound (4.5) follows with $\sigma \asymp n^{-1/2}$. The proof is complete. □

Remark B.1. We show here that the existing stability bound (eq. (4.2) with $\gamma = O(1/(n\sigma))$) [1, 6, 11, 17]

$$|R_S(A(S)) - R(A(S))| = O(\sigma^{-1}n^{-\frac{1}{2}}) \qquad \text{(B.10)}$$

yields at best the excess risk bound $R(A(S)) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{4}})$. Indeed, plugging (B.10) and (B.9) back into (B.8), we derive the following inequality with high probability

$$R(A(S)) - R(\mathbf{w}_R^*) = O(\sigma^{-1}n^{-\frac{1}{2}}) + O(\sigma).$$

We can balance the above two terms by taking $\sigma \asymp n^{-\frac{1}{4}}$ and get

$$R(A(S)) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{4}}).$$

C Proof of Theorem 4

To prove Theorem 4, we first introduce some lemmas. Lemma C.1 shows the non-expansiveness of the gradient-update operator, which plays a key role in establishing the stability of SGD. Lemma C.2 is a Chernoff's bound for a summation of independent Bernoulli random variables [2]. In this section, we let $\|\cdot\|_2$ be the Euclidean norm.

Lemma C.1 ([8]). *Assume for all $z \in \mathcal{Z}$, the function $\mathbf{w} \mapsto \ell(\mathbf{w}; z, z')$ is convex and α-smooth. Then for all $\eta \leq 2/\alpha$ and $z, z' \in \mathcal{Z}$ there holds*

$$\|\mathbf{w} - \eta\ell'(\mathbf{w}; z, z') - \mathbf{w}' + \eta\ell'(\mathbf{w}'; z, z')\|_2 \leq \|\mathbf{w} - \mathbf{w}'\|_2.$$

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

We now establish the uniform stability of SGD. The randomness of SGD can be characterized by $\{\{(i_t, j_t)_t\} : i_t, j_t \in [n], i_t \neq j_t\}$. Therefore, SGD can be considered as a deterministic algorithm if $\{\{(i_t, j_t)_t\} : i_t, j_t \in [n], i_t \neq j_t\}$ is fixed. For simplicity, we consider two datasets that differ by the last example. However, our discussion directly extends to the general case where two datasets differ by a single example. Notice that the Lipschitz continuity (4.3) implies $\|\ell'(\mathbf{w}; z, z')\|_2 \leq L$.

Lemma C.3. *Consider fixed $\{\{(i_t, j_t)_t\} : i_t, j_t \in [n], i_t \neq j_t\}$. Let $S = \{z_1, \ldots, z_n\}$ and $S' = \{z'_1, \ldots, z'_n\}$ be two datasets that differ only by the last example, i.e., $z_i = z'_i$ if $i \in [n-1]$. Suppose for all $z, z' \in \mathcal{Z}$ the function $\mathbf{w} \mapsto \ell(\mathbf{w}; z, z')$ is convex, α-smooth and L-Lipschitz w.r.t. $\|\cdot\|_2$. Let $\{\mathbf{w}_t\}, \{\mathbf{w}'_t\}$ be produced by SGD on S and S' respectively with $\eta_t \leq 2/\alpha$, i.e.,* (3.3) *with $r(\mathbf{w}) = 0$. Then SGD with t iterations is γ-uniformly stable with*

$$\gamma \leq 2L^2 \sum_{k=1}^{t} \eta_k \mathbb{I}[i_k = n \text{ or } j_k = n].$$

Proof. Let us consider two cases. We first consider the case $i_t \in [n-1]$ and $j_t \in [n-1]$. In this case, according to the SGD update (3.3) with $r(\mathbf{w}) = 0$ we know

$$\begin{aligned}\mathbf{w}_{t+1} - \mathbf{w}'_{t+1} &= \mathbf{w}_t - \eta_t \ell'(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \mathbf{w}'_t + \eta_t \ell'(\mathbf{w}'_t; z'_{i_t}, z'_{j_t}) \\ &= \mathbf{w}_t - \eta_t \ell'(\mathbf{w}_t; z_{i_t}, z_{j_t}) - \mathbf{w}'_t + \eta_t \ell'(\mathbf{w}'_t; z_{i_t}, z_{j_t}).\end{aligned}$$

It then follows from Lemma C.1 that

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2.$$

We now consider the case that either $i_t = n$ or $j_t = n$. In this case, we know

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 &= \big\|\mathbf{w}_t - \eta_t \ell'(\mathbf{w}_t; z_{i_t}; z_{j_t}) - \mathbf{w}'_t + \eta_t \ell'(\mathbf{w}'_t; z'_{i_t}, z'_{j_t})\big\|_2 \\ &\leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + \big\|\eta_t \ell'(\mathbf{w}'_t; z'_{i_t}, z'_{j_t}) - \eta_t \ell'(\mathbf{w}_t; z_{i_t}, z_{j_t})\big\|_2 \\ &\leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2\eta_t L,\end{aligned}$$

where we have used $\|\ell'(\mathbf{w}; z, z')\|_2 \leq L$ due to the L-Lipschitzness. As a combination of the above two cases, we derive

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + 2\eta_t L \mathbb{I}[i_t = n \text{ or } j_t = n],$$

where $\mathbb{I}[\cdot]$ is the indicator function taking 1 if the argument holds and 0 otherwise. Taking a summation of the above inequality gives ($\mathbf{w}_1 = \mathbf{w}'_1$)

$$\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \leq 2L \sum_{k=1}^{t} \eta_k \mathbb{I}[i_k = n \text{ or } j_k = n].$$

This together with the Lipschitz continuity of ℓ implies the following inequality for all $z, z' \in \mathcal{Z}$

$$\begin{aligned}\big|\ell(\mathbf{w}_{t+1}; z, z') - \ell(\mathbf{w}'_{t+1}; z, z')\big| &\leq L\|\mathbf{w}_{t+1} - \mathbf{w}'_{t+1}\|_2 \\ &\leq 2L^2 \sum_{k=1}^{t} \eta_k \mathbb{I}[i_k = n \text{ or } j_k = n].\end{aligned}$$

The proof is complete. □

We now apply the above uniform stability bounds and Theorem 1 to prove Theorem 4.

Proof of Theorem 4. We can apply Theorem 1 with $A(S) = \mathbf{w}_T$ and the uniform stability bounds in Lemma C.3 to show with probability at least $1 - \delta/2$ that

$$\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O\Big((\log n \log(1/\delta)) \sum_{t=1}^{T} \eta \mathbb{I}[i_t = n \text{ or } j_t = n]\Big) + O(n^{-\frac{1}{2}}\sqrt{\log(1/\delta)}), \tag{C.1}$$

where $\eta = c/\sqrt{T}$. Let $X_t = \mathbb{I}[i_t = n \text{ or } j_t = n]$. It is clear that

$$\mathbb{E}[X_t] = \Pr\{i_t = n \text{ or } j_t = n\} \leq \Pr\{i_t = n\} + \Pr\{j_t = n\} = 2/n.$$

Applying Lemma C.2 with $X_t = \mathbb{I}[i_t = n \text{ or } j_t = n]$ then gives with probability $1 - \delta/2$ that

$$\sum_{t=1}^{T} X_t \leq \Big(1 + \sqrt{3\mu^{-1}\log(1/\delta)}\Big)\mu,$$

where $\mu = \sum_{t=1}^{T} \mathbb{E}[X_t] \leq 2T/n$. It then follows with probability $1 - \delta/2$ that

$$\sum_{t=1}^{T} X_t \leq \frac{2T}{n}\Big(1 + \sqrt{2nT^{-1}\log(1/\delta)}\Big). \tag{C.2}$$

Combining (C.1) and (C.2) together, we derive the following inequality with probability $1 - \delta$

$$\begin{aligned}\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O(n^{-\frac{1}{2}}\sqrt{\log(1/\delta)} + T\eta(\log n \log(1/\delta))n^{-1} \\ + \eta \log n \log(1/\delta)\sqrt{n^{-1}T\log(1/\delta)}).\end{aligned}$$

The proof is complete with $\eta = c/\sqrt{T}$. □

Remark C.1. We now give details on deriving excess risk bounds based on the estimation error bounds in Theorem 4. We can decompose the excess risk into optimization errors and estimation errors as follows (we omit $\log(1/\delta)$) [3]

$$\begin{aligned}R(\mathbf{w}_T) - R(\mathbf{w}_R^*) &= R(\mathbf{w}_T) - R_S(\mathbf{w}_T) + R_S(\mathbf{w}_T) - R_S(\mathbf{w}_R^*) + R_S(\mathbf{w}_R^*) - R(\mathbf{w}_R^*) \\ &= \big(R(\mathbf{w}_T) - R_S(\mathbf{w}_T)\big) + \big(R_S(\mathbf{w}_T) - R_S(\mathbf{w}_R^*)\big) + O(n^{-\frac{1}{2}}),\end{aligned} \tag{C.3}$$

where we have used (B.9). The first term is the estimation error and comes from the approximation of testing errors by training errors. The second term is the optimization error which comes since the optimization algorithm may not output the exact minimizer. Then Theorem 4 actually presents estimation error bounds. If we further assume $\|\mathbf{w}_t\| \leq B$ for some $B > 0$ and all t, then it was shown with high probability that [9]

$$R_S(\mathbf{w}_T) - R_S(\mathbf{w}_R^*) = O(T^{-\frac{1}{2}}\log T). \tag{C.4}$$

We can plug the above optimization error bounds and the estimation error bounds in Theorem 4 into (C.3), and get with high probability

$$R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O\Big(\log n\sqrt{T}/n + n^{-\frac{1}{2}}\log n\Big) + O(T^{-\frac{1}{2}}\log T).$$

One can take an optimal $T \asymp n$ to trade-off the optimization and estimation errors, and get

$$R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O(n^{-\frac{1}{2}}\log n).$$

Remark C.2. If we plug the uniform stability bounds in Lemma C.3 into the existing connection between stability and generalization established in (4.2), we get with high probability that

$$\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O\Big(\sqrt{n}\sum_{t=1}^{T}\eta_t\mathbb{I}[i_t = n \text{ or } j_t = n] + n^{-\frac{1}{2}}\Big).$$

This together with (C.2) shows the following inequality with high probability ($\eta_t = \eta = O(1/\sqrt{T})$)

$$\big|R_S(\mathbf{w}_T) - R(\mathbf{w}_T)\big| = O\Big(\frac{T\eta}{\sqrt{n}}\big(1 + \sqrt{n/T}\big) + n^{-\frac{1}{2}}\Big) = O\Big(\frac{\sqrt{T}}{\sqrt{n}}\big(1 + \sqrt{n/T}\big) + n^{-\frac{1}{2}}\Big).$$

We can plug the above estimation error bound, the optimization error bound (C.4) back into (C.3), and derive the following excess risk bound with high probability

$$R(\mathbf{w}_T) - R(\mathbf{w}_R^*) = O\Big(\frac{\log T}{\sqrt{T}} + \frac{\sqrt{T}}{\sqrt{n}}\big(1 + \sqrt{n/T}\big) + n^{-\frac{1}{2}}\Big) = O(1).$$

D Proofs on Optimistic Bounds

In this section, we prove optimistic bounds in Theorem 6 by using the smoothness of loss functions. We first prove Theorem 5 on the connection between generalization and on-average stability.

Proof of Theorem 5. For all $i, j \in [n]$, let $S_{i,j}$ be defined by (3.4). Due to the symmetry, we know $\mathbb{E}[R(A(S))] = \mathbb{E}[R(A(S_{i,j}))]$ for all $i, j \in [n]$ with $i \neq j$ and therefore

$$\begin{aligned}\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] &= \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}\big[R(A(S_{i,j})) - R_S(A(S))\big] \\ &= \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} \mathbb{E}\Big[\ell\big(A(S_{i,j}); z_i, z_j\big) - \ell\big(A(S); z_i, z_j\big)\Big] \leq \gamma,\end{aligned}$$

where the second identity holds since $A(S_{i,j})$ is independent of z_i and z_j. The proof is complete. □

We then introduce some basic properties of smooth functions. For a α-smooth and non-negative function g, we have the following self-bounding property [14]

$$\|g'(\mathbf{w})\|^2 \leq 2\alpha g(\mathbf{w}), \quad \forall \mathbf{w} \in \mathcal{W} \tag{D.1}$$

and the following elementary inequality

$$g(\mathbf{w}) \leq g(\mathbf{w}') + \langle g'(\mathbf{w}'), \mathbf{w} - \mathbf{w}'\rangle + \frac{\alpha\|\mathbf{w} - \mathbf{w}'\|^2}{2}, \quad \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}. \tag{D.2}$$

We then present a useful lemma.

Lemma D.1. *Let S, S' be defined in Definition 2. Assume for all z, z', $\ell(\cdot, z, z')$ is α-smooth w.r.t. a norm. For all $i \in [n]$, let S_i be defined as* (A.2) *and $\epsilon > 0$. Then*

$$\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \leq \frac{\alpha\mathbb{E}[R_S(A(S))]}{\epsilon} + \frac{2(\epsilon+\alpha)}{n} \sum_{i\in[n]} \mathbb{E}\big[\|A(S_i) - A(S)\|^2\big].$$

Proof. For all $i, j \in [n]$, let $S_{i,j}$ be defined by (3.4). According to (D.2), the Cauchy-Schwartz inequality and (D.1), for all $i, j \in [n]$ we know

$$\begin{aligned}\ell(A(S_{i,j}); z_i, z_j) - \ell(A(S); z_i; z_j) &\leq \langle \ell'(A(S); z_i, z_j), A(S_{i,j}) - A(S)\rangle + \frac{\alpha}{2}\|A(S_{i,j}) - A(S)\|^2 \\ &\leq \big\|\ell'(A(S); z_i, z_j)\big\|\big\|A(S_{i,j}) - A(S)\big\| + \frac{\alpha}{2}\|A(S_{i,j}) - A(S)\|^2 \\ &\leq \frac{\|\ell'(A(S); z_i, z_j)\|^2}{2\epsilon} + \frac{\epsilon+\alpha}{2}\|A(S_{i,j}) - A(S)\|^2 \\ &\leq \frac{\alpha\ell(A(S); z_i, z_j)}{\epsilon} + \frac{\epsilon+\alpha}{2}\|A(S_{i,j}) - A(S)\|^2.\end{aligned}$$

We can plug the above inequality into Theorem 5 to derive

$$\begin{aligned}&\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \\ &\leq \frac{\alpha}{\epsilon n(n-1)} \sum_{i\neq j} \mathbb{E}\big[\ell(A(S); z_i, z_j)\big] + \frac{\epsilon+\alpha}{2n(n-1)} \sum_{i\neq j} \mathbb{E}\big[\|A(S_{i,j}) - A(S)\|^2\big] \\ &= \frac{\alpha\mathbb{E}[R_S(A(S))]}{\epsilon} + \frac{\epsilon+\alpha}{2n(n-1)} \sum_{i\neq j} \mathbb{E}\big[\|A(S_{i,j}) - A(S)\|^2\big].\end{aligned} \tag{D.3}$$

By the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we get the following inequality for all $i \neq j$

$$\begin{aligned}\mathbb{E}\big[\|A(S_{i,j}) - A(S)\|^2\big] &\leq 2\mathbb{E}\big[\|A(S_{i,j}) - A(S_i)\|^2\big] + 2\mathbb{E}\big[\|A(S_i) - A(S)\|^2\big] \\ &= 2\mathbb{E}\big[\|A(S_i) - A(S)\|^2\big] + 2\mathbb{E}\big[\|A(S_j) - A(S)\|^2\big],\end{aligned}$$

where we have used the following identity due to the symmetry between z_i and z_i'

$$\mathbb{E}\big[\|A(S_{i,j}) - A(S_i)\|^2\big] = \mathbb{E}\big[\|A(S_j) - A(S)\|^2\big].$$

Plugging the above inequality back into (D.3), we know

$$\begin{aligned}\mathbb{E}\big[R(A(S)) - R_S(A(S))\big] \leq \frac{\alpha\mathbb{E}[R_S(A(S))]}{\epsilon} \\ + \frac{\epsilon+\alpha}{n(n-1)}\sum_{i\neq j}\Big(\mathbb{E}\big[\|A(S_i) - A(S)\|^2\big] + \mathbb{E}\big[\|A(S_j) - A(S)\|^2\big]\Big).\end{aligned}$$

This yields the stated inequality and finishes the proof. □

Proof of Theorem 6. According to Lemma B.1 and the α-smoothness of ℓ, we know the following inequality for any k

$$\begin{aligned}n(n-1)\big(F_S(A(S_k)) - F_S(A(S))\big) \leq \sum_{i\in[n]:i\neq k}\bigg(\Big\langle \ell'(A(S);z_i,z_k) + \ell'(A(S);z_k,z_i) \\ - \ell'(A(S_k);z_i,z_k') - \ell'(A(S_k);z_k',z_i); A(S_k) - A(S)\Big\rangle + \frac{4\alpha\|A(S_k) - A(S)\|^2}{2}\bigg).\end{aligned}$$

It then follows from the Cauchy-Schwartz inequality that

$$\begin{aligned}n(n-1)\big(F_S(A(S_k)) - F_S(A(S))\big) \leq \sum_{i\in[n]:i\neq k}\Big(\big\|\ell'(A(S);z_i,z_k)\big\| + \big\|\ell'(A(S);z_k,z_i)\big\| \\ + \big\|\ell'(A(S_k);z_i,z_k')\big\| + \big\|\ell'(A(S_k);z_k',z_i)\big\|\Big)\|A(S_k) - A(S)\| + 2\alpha(n-1)\|A(S_k) - A(S)\|^2.\end{aligned}$$

This, together with (D.1) and (B.3), implies

$$\begin{aligned}\frac{\sigma n(n-1)\|A(S_k) - A(S)\|^2}{2} \leq \sqrt{2\alpha}\sum_{i\in[n]:i\neq k}\Big(\sqrt{\ell(A(S);z_i,z_k)} + \sqrt{\ell(A(S);z_k,z_i)} \\ + \sqrt{\ell(A(S_k);z_i,z_k')} + \sqrt{\ell(A(S_k);z_k',z_i)}\Big)\|A(S_k) - A(S)\| + 2\alpha(n-1)\|A(S_k) - A(S)\|^2\end{aligned}$$

and further

$$\begin{aligned}\frac{\sigma n(n-1)\|A(S_k) - A(S)\|}{2} \leq \sqrt{2\alpha}\sum_{i\in[n]:i\neq k}\Big(\sqrt{\ell(A(S);z_i,z_k)} + \sqrt{\ell(A(S);z_k,z_i)} \\ + \sqrt{\ell(A(S_k);z_i,z_k')} + \sqrt{\ell(A(S_k);z_k',z_i)}\Big) + 2\alpha(n-1)\|A(S_k) - A(S)\|.\end{aligned}$$

Since $2\alpha \leq \sigma n/4$, we further get

$$\begin{aligned}\frac{\sigma n(n-1)\|A(S_k) - A(S)\|}{4} \leq \sqrt{2\alpha}\sum_{i\in[n]:i\neq k}\Big(\sqrt{\ell(A(S);z_i,z_k)} + \sqrt{\ell(A(S);z_k,z_i)} \\ + \sqrt{\ell(A(S_k);z_i,z_k')} + \sqrt{\ell(A(S_k);z_k',z_i)}\Big).\end{aligned}$$

Taking a square over both sides and using the standard inequality $\big(\sum_{i=1}^{n-1} a_i\big)^2 \leq (n-1)\sum_{i=1}^{n-1} a_i^2$, we derive

$$\begin{aligned}\frac{\sigma^2 n^2(n-1)^2\|A(S_k) - A(S)\|^2}{16} \leq 2\alpha(n-1)\sum_{i\in[n]:i\neq k}\Big(\sqrt{\ell(A(S);z_i,z_k)} + \sqrt{\ell(A(S);z_k,z_i)} \\ + \sqrt{\ell(A(S_k);z_i,z_k')} + \sqrt{\ell(A(S_k);z_k',z_i)}\Big)^2.\end{aligned}$$

This, further together with the inequality $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$, implies

$$\sigma^2 n^2(n-1)\|A(S_k)-A(S)\|^2 \leq 128\alpha \sum_{i\in[n]:i\neq k} \Big(\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i) + \ell(A(S_k);z_i,z_k')+\ell(A(S_k);z_k',z_i)\Big).$$

Taking a summation of the above inequality from $k=1$ to n, we get

$$\sigma^2 n^2(n-1)\sum_{k=1}^n \|A(S_k)-A(S)\|^2 \leq 128\alpha \sum_{i,k\in[n]:i\neq k} \Big(\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i) + \ell(A(S_k);z_i,z_k')+\ell(A(S_k);z_k',z_i)\Big). \quad \text{(D.4)}$$

Due to the symmetry, we know

$$\mathbb{E}\big[\ell(A(S_k);z_i,z_k')\big] = \mathbb{E}\big[\ell(A(S);z_i,z_k)\big], \quad \forall i\neq k.$$

It then follows that

$$\begin{aligned}
&\sum_{i,k\in[n]:i\neq k} \mathbb{E}\Big[\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i)+\ell(A(S_k);z_i,z_k')+\ell(A(S_k);z_k',z_i)\Big] \\
&= \sum_{i,k\in[n]:i\neq k} \mathbb{E}\Big[\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i)+\ell(A(S);z_i,z_k)+\ell(A(S);z_k,z_i)\Big] \\
&= 4n(n-1)\mathbb{E}\big[R_S(A(S))\big].
\end{aligned}$$

We can plug the above inequality back into (D.4) and derive that

$$\sigma^2 n\sum_{k=1}^n \mathbb{E}\big[\|A(S_k)-A(S)\|^2\big] \leq 512\alpha\mathbb{E}\big[R_S(A(S))\big].$$

We now plug the above inequality back into Lemma D.1 and derive that the following inequality for all $\epsilon>0$

$$\mathbb{E}\big[R(A(S))-R_S(A(S))\big] \leq \frac{\alpha\mathbb{E}[R_S(A(S))]}{\epsilon} + \frac{1024(\epsilon+\alpha)\alpha}{n^2\sigma^2}\mathbb{E}\big[R_S(A(S))\big].$$

We can take $\epsilon = \frac{n\sigma}{32}$ to derive

$$\mathbb{E}\big[R(A(S))-R_S(A(S))\big] \leq \Big(\frac{1024\alpha^2}{n^2\sigma^2}+\frac{64\alpha}{n\sigma}\Big)\mathbb{E}\big[R_S(A(S))\big].$$

Furthermore, according to the definition of $A(S)$ we know ($\mathbf{w}^*$ is independent of S)

$$\begin{aligned}
\mathbb{E}[F(A(S))]-F(\mathbf{w}^*) &= \mathbb{E}\big[F(A(S))-F_S(A(S))\big]+\mathbb{E}\big[F_S(A(S))-F_S(\mathbf{w}^*)\big] \\
&\leq \mathbb{E}\big[F(A(S))-F_S(A(S))\big] = \mathbb{E}\big[R(A(S))-R_S(A(S))\big].
\end{aligned}$$

This finishes the proof of (4.8).

We now turn to the bound of $\mathbb{E}[R(A(S))]-R(\mathbf{w}_R^*)$. Analogously to (B.8), we know

$$\begin{aligned}
\mathbb{E}\big[R(A(S))-R(\mathbf{w}_R^*)\big] &\leq \mathbb{E}\big[R(A(S))-R_S(A(S))\big]+O(\sigma\|\mathbf{w}_R^*\|^2) \\
&= O\Big(\frac{1}{n\sigma}\Big)\mathbb{E}\big[R_S(A(S))\big]+O(\sigma\|\mathbf{w}_R^*\|^2), \quad \text{(D.5)}
\end{aligned}$$

where we have used (4.8) in the last step. According to the definition of $A(S)$, we further know

$$R_S(A(S))+r(A(S)) \leq R_S(\mathbf{w}_R^*)+r(\mathbf{w}_R^*) = R_S(\mathbf{w}_R^*)+O(\sigma\|\mathbf{w}_R^*\|^2).$$

Since $\mathbf{w}_R^*$ is independent of S, we can take expectation to derive

$$\mathbb{E}[R_S(A(S))] = R(\mathbf{w}_R^*)+O(\sigma\|\mathbf{w}_R^*\|^2).$$

We can plug the above inequality back into (D.5), and derive

$$\mathbb{E}\big[R(A(S)) - R(\mathbf{w}_R^*)\big] = O\Big(\frac{R(\mathbf{w}_R^*)}{n\sigma} + O\big(n^{-1} + \sigma\big)\|\mathbf{w}_R^*\|^2\Big).$$

We can take

$$\sigma = \max\Big\{\frac{8\alpha}{n}, \sqrt{\frac{R(\mathbf{w}_R^*)}{n\|\mathbf{w}_R^*\|^2}}\Big\}$$

and derive

$$\mathbb{E}\big[R(A(S)) - R(\mathbf{w}_R^*)\big] = O\Big(\frac{\sqrt{R(\mathbf{w}_R^*)}\|\mathbf{w}_R^*\|}{\sqrt{n}} + \frac{\|\mathbf{w}_R^*\|^2}{n}\Big).$$

This establishes (4.9) and finishes the proof.

□

E Proofs on Applications

In this section, we present proofs for applications of our general results to metric learning.

Proof of Corollary 10. It is well known that F_S is 2λ-strongly convex w.r.t. $\|\cdot\|$. To apply Theorem 3, we require to check (4.3). For all $\mathbf{w}, \mathbf{w}', z, z'$, we know

$$\begin{aligned}
&\big|\ell^\psi(\mathbf{w}; z, z') - \ell^\psi(\mathbf{w}'; z, z')\big| \\
&= \Big|\max\big\{0, 1 - \tau(y, y')(1 - h_{\mathbf{w}}(x, x'))\big\} - \max\big\{0, 1 - \tau(y, y')(1 - h_{\mathbf{w}'}(x, x'))\big\}\Big| \\
&\le |\tau(y, y')|\big|h_{\mathbf{w}}(x, x') - h_{\mathbf{w}'}(x, x')\big| \le \big|\langle \mathbf{w} - \mathbf{w}', (x - x')(x - x')\rangle\big| \\
&\le 4B^2\|\mathbf{w} - \mathbf{w}'\|.
\end{aligned}$$

Therefore, (4.3) holds with $L = 4B^2$. The proof then completes by applying Theorem 3. □

Proof of Corollary 11. To apply Theorem 4, it suffices to show the smoothness of the loss function. The gradient of ℓ^ψ w.r.t. $\mathbf{w}$ can be calculated by

$$\nabla\ell^\psi(\mathbf{w}; z, z') = -\psi'\big(\tau(y, y')(1 - h_{\mathbf{w}}(x, x'))\big)\tau(y, y')(x - x')(x - x')^\top.$$

Then, for any $\mathbf{w}$ and $\mathbf{w}' \in \mathcal{W}$ we have

$$\begin{aligned}
&\big\|\nabla\ell^\psi(\mathbf{w}; z, z') - \nabla\ell^\psi(\mathbf{w}'; z, z')\big\|_K \\
&\le \big\|\tau(y, y')(x - x')(x - x')^\top\big\|\big|\psi'\big(\tau(y, y')(1 - h_{\mathbf{w}}(x, x'))\big) - \psi'\big(\tau(y, y')(1 - h_{\mathbf{w}'}(x, x'))\big)\big| \\
&\le 4B^2\big|\psi'\big(\tau(y, y')(1 - h_{\mathbf{w}}(x, x'))\big) - \psi'\big(\tau(y, y')(1 - h_{\mathbf{w}'}(x, x'))\big)\big| \\
&\le 4B^2|\tau(y, y')|\big|(1 - h_{\mathbf{w}}(x, x')) - (1 - h_{\mathbf{w}'}(x, x'))\big| \\
&= 4B^2\big|\langle \mathbf{w} - \mathbf{w}', (x - x')(x - x')^\top\rangle\big| \\
&\le 16B^4\|\mathbf{w} - \mathbf{w}'\|,
\end{aligned}$$

where we have used the 1-smoothness of the logistic loss in the third step. That is, ℓ^ψ is $(16B^4)$-smooth w.r.t. the Frobenius norm. The stated bound then follows from Theorem 4. □

F Minimax Optimal Excess Risk Bounds for Pairwise Learning

Here we explain that the bound $O(n^{-\frac{1}{2}})$ is minimax optimal for the excess risks in pairwise learning. To see this, we consider pairwise loss functions which do not depend on the second example, i.e., $\ell(\mathbf{w}; z, z') = \ell(\mathbf{w}; z, \tilde{z}')$ for all $z', \tilde{z}' \in \mathcal{Z}$. Then it is clear that R_S defined in (3.1) becomes

$$R_S(\mathbf{w}) = \frac{1}{n}\sum_{i\in[n]}\frac{1}{n-1}\sum_{j\in[n]:j\neq i}\ell(\mathbf{w}; z_i, z_j) = \frac{1}{n}\sum_{i\in[n]}\ell(\mathbf{w}; z_i, z_0) := \widetilde{R}_S(\mathbf{w}),$$

where z_0 is any fixed point in $\mathcal{Z}$. This is actually an objective function for pointwise learning. We know that for any estimator we can find a pointwise learning problem such that this estimator has the excess risk bound $O(n^{-\frac{1}{2}})$ [15]. Then, for any estimator we can build a pairwise learning problem such that this estimator has at best the excess risk bound $O(n^{-\frac{1}{2}})$. Furthermore, we can construct such a pairwise learning problem with the loss function independent of the second example.

References

[1] S. Agarwal and P. Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(Feb):441–474, 2009.

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

[3] O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

[4] O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

[5] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, pages 844–874, 2008.

[6] C. Cortes, M. Mohri, and A. Rastogi. Magnitude-preserving ranking algorithms. In *Proceedings of the 24th international conference on Machine learning*, pages 169–176, 2007.

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

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

[9] N. J. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613, 2019.

[10] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[11] R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems*, pages 862–870, 2009.

[12] T. Peel, S. Anthoine, and L. Ralaivola. Empirical bernstein inequalities for u-statistics. In *Advances in Neural Information Processing Systems*, pages 1903–1911. 2010.

[13] Y. Pitcan. A note on concentration inequalities for u-statistics, 2017.

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

[15] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[16] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[17] B. Wang, H. Zhang, P. Liu, Z. Shen, and J. Pineau. Multitask metric learning: Theory and algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 3362–3371, 2019.