# A Supplementary

## A.1 Supporting Theorems and Lemmas

Let us recall the excess risk of a randomized algorithm $\mathcal{A}$ defined as $\epsilon_{\text{risk}}(\mathcal{A}(S)) = R(\mathcal{A}(S)) - R(\mathbf{w}_*)$, which can be decomposed by

$$\epsilon_{\text{risk}}(\mathcal{A}(S)) = R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) + R_S(\mathbf{w}_*) - R(\mathbf{w}_*) + R_S(\mathcal{A}(S)) - R_S(\mathbf{w}_*). \tag{A.14}$$

Hence, before introducing the proofs we will give some theorems and lemmas that are repeatedly used to bound each term in Equation (A.14).

Here we simply assume bounds for $\|\mathbf{w}\|$. A simple lemma indicates that if $\mathbf{w}$ is bounded, then $\ell(\mathbf{w}, \cdot, \cdot)$ is also bounded. In the subsequent sections, we will characterize the bounds for the iterates $\{\mathbf{w}_t\}$ whenever the parameter space $\mathcal{W}$ is bounded or unbounded.

**Lemma A.6.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Let $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}')$. For any $\mathbf{w} \in W$ that $\mathbf{w} \leq B$ for some $0 \leq B < \infty$, then $\sup_{\mathbf{z},\mathbf{z}'} \ell(\mathbf{w}, \mathbf{z}, \mathbf{z}') \leq M + GB$.*

*Proof.* By convexity of $\ell$, we have for any $\mathbf{z}, \mathbf{z}'$

$$\ell(\mathbf{w}, \mathbf{z}, \mathbf{z}') \leq \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}') + \langle \mathbf{w}, \partial\ell(\mathbf{w}, \mathbf{z}, \mathbf{z}')\rangle \leq M + \|\mathbf{w}\|\|\partial\ell(\mathbf{w}, \mathbf{z}, \mathbf{z}')\|_2 \leq M + GB$$

where the second inequality is due to Cauchy-Schwarz inequality. The proof is complete by taking the supremum. □

The first theorem in this section is the the high probability generalization bound of UAS algorithms in pairwise learning. This theorem is an extension of Theorem 1 in Lei et al. (2020) for generalization bound of uniformly stable algorithms in pairwise learning.

**Theorem A.6.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Let $\mathcal{A}$ be a $\epsilon$-UAS randomized algorithm for pairwise learning. Suppose the output of $\mathcal{A}$ is bounded by $B$ and let $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}')$. Then we have for any $\gamma \in (0, 1)$, with probability at least $1 - \gamma$ with respect to the sample $S$ and the internal randomness of $\mathcal{A}$,*

$$R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) \leq 4\epsilon + 48\sqrt{6}eG\epsilon\lceil\ln(n)\rceil\ln(e/\gamma) + 12\sqrt{2}e(M + GB)\sqrt{\frac{\ln(e/\gamma)}{n}}.$$

*Proof.* According to Theorem 1 in Lei et al. (2020), we only need to check the expected boundedness of $\ell(\mathcal{A}(S), \cdot, \cdot)$ and the uniform stability of $\mathcal{A}$. For the boundedness part, by Lemma A.6 we know

$$\left|\mathbb{E}[\ell(\mathcal{A}(S), \mathbf{z}, \mathbf{z}')]\right| \leq M + GB$$

for any $\mathbf{z}, \mathbf{z}'$. For the uniform stability, since $\mathcal{A}$ is $\epsilon$-UAS, by the Lipschitz continuity of $\ell$ we have

$$\sup_{\mathbf{z},\mathbf{z}'} |\ell(\mathcal{A}(S), \mathbf{z}, \mathbf{z}') - \ell(\mathcal{A}(S'), \mathbf{z}, \mathbf{z}')| \leq G\|\mathcal{A}(S) - \mathcal{A}(S')\|_2 \leq G\epsilon.$$

The proof is complete. □

The next corollary is a direct application of Theorem A.6, which states if UAS holds with high probability, then so is the generalization.

**Corollary A.5.** *Let $\mathcal{A}$ be a randomized algorithm for pairwise learning. If for any $\gamma_0 \in (0, 1)$, we have, for any neighborhood datasets $S, S'$,*

$$\mathbb{P}_{\mathcal{A}}\Big[\|\mathcal{A}(S) - \mathcal{A}(S')\|_2 > \epsilon\Big] \leq \gamma_0.$$

*Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Suppose the output of $\mathcal{A}$ is bounded by $B$ and let $M = \sup_{\mathbf{z},\mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}')$. Then we have for any $\gamma \in (0, 1)$,*

$$\mathbb{P}_{S,\mathcal{A}}\Big[R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) > 4\epsilon + 48\sqrt{6}eG\epsilon\lceil\ln(n)\rceil\ln(e/\gamma) + 12\sqrt{2}e(M + GB)\sqrt{\frac{\ln(e/\gamma)}{n}}\Big] \leq \gamma + \gamma_0.$$

*Proof.* Denote $E = \{\mathcal{A} | \|\mathcal{A}(S) - \mathcal{A}(S')\|_2 > \epsilon\}$ and $F = \{S, \mathcal{A} | R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) > 4\epsilon + 48\sqrt{6}eG\epsilon\lceil\ln(n)\rceil\ln(e/\gamma) + 12\sqrt{2}e(M+GB)\sqrt{\ln(e/\gamma)/n}\}$. Then by assumption we have $\mathbb{P}_{\mathcal{A}}[\mathcal{A}\in E] \le \gamma_0$. By Theorem A.6, for any $\gamma \in (0,1)$, we have $\mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A}\in F|\mathcal{A}\notin E] \le \gamma$. Then the following identity holds

$$\begin{aligned}\mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A}\in F] =& \mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A}\in F\cap\mathcal{A}\in E] + \mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A}\in F\cap\mathcal{A}\notin E]\\ =& \mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A}\in F|\mathcal{A}\in E]\mathbb{P}[\mathcal{A}\in E] + \mathbb{P}_{S,\mathcal{A}}[S,\mathcal{A}\in F|\mathcal{A}\notin E]\mathbb{P}[\mathcal{A}\notin E]\\ \le& \gamma_0 + \gamma.\end{aligned}$$

The proof is completed. □

Combining Corollary A.5 and the stability result in Theorem 1, we arrive at the following generalization bound for Algorithm 1.

**Corollary A.6.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Let $B_T = \|\bar{\mathbf{w}}_T\|$ and $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$. If we run Algorithm 1 for $T \ge n$ iterations under random selection with replacement rule. For any $\gamma \in (0,1)$, with probability at least $1-\gamma$ with respect to the sample $S$ and the internal randomness of Algorithm 1, we have*

$$R(\bar{\mathbf{w}}_T) - R_S(\bar{\mathbf{w}}_T) \le 2\sqrt{e}\eta G(4 + 48\sqrt{6}eG\lceil\ln(n)\rceil\ln(2e/\gamma))\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(2/\gamma)}{n}\Big) + 12\sqrt{2}e(M+GB_T)\sqrt{\frac{\ln(2e/\gamma)}{n}}.$$

*Proof.* By Theorem 1, elementary inequality and the fact that stability is monotonically increasing, we have with probability at least $1-\gamma/2$,

$$\left\|\bar{\mathbf{w}}_T - \bar{\mathbf{w}}'_T\right\|_2^2 \le 4e\eta^2G^2\Big(T + \frac{3T^2\ln^2(eT)\ln^2(2/\gamma)}{n^2}\Big).$$

The proof is completed by convexity of $\|\cdot\|_2$ and applying Theorem A.6 with probability $1-\frac{\gamma}{2}$. □

The next theorem gives a bound on $R_S(\mathbf{w}_*) - R(\mathbf{w}_*)$ by Hoeffding inequality of U-statistics Hoeffding (1963).

**Theorem A.7.** *Suppose $\ell$ is convex and $G$-Lipschitz. Let $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$ and $B = \|\mathbf{w}_*\|_2$. For any $\gamma \in (0,1)$, with probability at least $1-\gamma$ with respect to the sample $S$, we have*

$$R_S(\mathbf{w}_*) - R(\mathbf{w}_*) \le (M+GB)\sqrt{\frac{\ln(1/\gamma)}{n}}.$$

*Proof.* The result is derived by applying Hoeffding inequality since $\ell(\mathbf{w}_*,\mathbf{z},\mathbf{z}') \le M + GB$ for any $\mathbf{z},\mathbf{z}'$ according to Lemma A.6. □

Next we give an upper bound on the optimization error $R_S(\bar{\mathbf{w}}_T) - R_S(\mathbf{w}_*)$. The results are inspired by Kar et al. (2013), where they consider the online-to-batch generalization bound for pairwise learning. Our bound in the next theorem is given for optimization bound on finite sample.

**Theorem A.8.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Suppose there are some non-decreasing sequence $0 \le B_t < \infty$ such that $\|\mathbf{w}_t\|_2 \le B_t$, and let $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$ and $B = \|\mathbf{w}_*\|_2$. Suppose we run Algorithm 1 for $T$ iterations, then with probability at least $1-\gamma$ with respect to the sample $S$ and the internal randomness of Algorithm 1, we have*

$$R_S(\bar{\mathbf{w}}_T) - R_S(\mathbf{w}_*) \le \frac{2}{T}\sum_{t=1}^T\mathcal{R}_t(\ell\circ\mathcal{W}_t) + \frac{2}{T}\sum_{t=1}^T\mathcal{R}_t(\ell\circ\mathcal{W}_B) + \frac{B^2}{2T\eta} + \frac{\eta G^2}{2} + (6M+3GB)\sqrt{\frac{\ln(2T/\gamma)}{T}} + 3GB_T\sqrt{\frac{\ln(2T/\gamma)}{T}},$$

*where $\mathcal{W}_t = \{\mathbf{w}\in W | \|\mathbf{w}\|_2 \le B_t\}$ and $\mathcal{W}_B = \{\mathbf{w}\in W | \|\mathbf{w}\|_2 \le B\}$ are subspaces of $\mathcal{W}$.*

In order to prove Theorem A.8, we decompose $R_S(\bar{\mathbf{w}}_T) - R_S(\mathbf{w}_*)$ as in Kar et al. (2013) and bound each part separately. In particular, recall that $\tilde{L}_{t+1}(\mathbf{w}) = \frac{1}{t}\sum_{k=1}^t\ell(\mathbf{w},\mathbf{z}_{i_{t+1}},\mathbf{z}_{i_k})$. We have the following lemmas.

**Lemma A.7.** *Assume $\ell$ is nonnegative, convex and $G$-Lipschitz. Let $\mathcal{W}_t = \big\{\mathbf{w}\in\mathcal{W} \big| \|\mathbf{w}\|_2 \le B_t\big\}$ and let $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$. With probability $1-\gamma$, we have*

$$\frac{1}{T}\sum_{t=1}^T R_S(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_t) \le \frac{2}{T}\sum_{t=1}^T\mathcal{R}_t(\ell\circ\mathcal{W}_t) + 3(M+GB_T)\sqrt{\frac{\ln(T/\gamma)}{T}}.$$

*Proof.* For any $\mathbf{w}$, denote $\tilde{L}_{t+1}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{t}\sum_{k=1}^t \ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_{i_k})$. This allows us to decompose the risk as follows

$$\frac{1}{T}\sum_{t=1}^T R_S(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_t) = \frac{1}{T}\sum_{t=1}^T \underbrace{R_S(\mathbf{w}_t) - \tilde{L}_{t+1}(\mathbf{w}_t)}_{P_{t+1}} + \underbrace{\tilde{L}_{t+1}(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_t)}_{Q_{t+1}}$$

By construction, we have $\mathbb{E}_{\mathbf{z}_{i_{t+1}}}[Q_{t+1}|\mathbf{z}_{i_1}, \cdots, \mathbf{z}_{i_t}] = 0$ and hence the sequence $Q_2, \cdots, Q_T$ forms a martingale difference sequence. By Lemma A.6 we have $Q_{t+1}$ lies in $[-M - GB_t, M + GB_t] \subseteq [-M - GB_T, M + GB_T]$ as $B_t$'s are non-decreasing. An application of the Azuma-Hoeffding inequality shows that with probability at least $1 - \gamma$,

$$\frac{1}{T}\sum_{t=1}^T Q_t \le (M + GB_T)\sqrt{\frac{2\ln(1/\gamma)}{T}}.$$

We now analyze each term $P_t$ individually. Let us start by introducing a ghost sample $\{\mathbf{z}'_{i_1}, \cdots, \mathbf{z}'_{i_t}\}$, where each $\mathbf{z}'_{i_k}$ follows the same distribution as $\mathbf{z}_{i_k}$. By linearity of expectation, we have

$$R_S(\mathbf{w}_t) = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n \frac{1}{t}\sum_{k=1}^t \ell(\mathbf{w}_t, \mathbf{z}_i, \mathbf{z}'_{i_k})\Big],$$

where the expectation is taken over $\{\mathbf{z}'_{i_k}\}_{k=1}^t$. It allows us to write $P_t$ as follow

$$\begin{aligned} P_t =& \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n \frac{1}{t}\sum_{k=1}^t \ell(\mathbf{w}_t, \mathbf{z}_i, \mathbf{z}'_{i_k})\Big] - \tilde{L}_{t+1}(\mathbf{w}_t) \le \sup_{\mathbf{w}\in\mathcal{W}_t} \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n \frac{1}{t}\sum_{k=1}^t \ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}'_{i_k})\Big] - \tilde{L}_{t+1}(\mathbf{w}) \\ \triangleq& g_{t+1}(\mathbf{z}_{i_1}, \cdots, \mathbf{z}_{i_t}). \end{aligned}$$

Since $\ell$ is bounded by $A_t$, the expression $g_{t+1}(\mathbf{z}_{i_1}, \cdots, \mathbf{z}_{i_t})$ can have a variation of at most $(M + GB_t)/t$ when changing any of its $t$ variables. Hence an application of McDiarmid's inequality gives us, with probability at least $1 - \gamma$,

$$g_{t+1}(\mathbf{z}_{i_1}, \cdots, \mathbf{z}_{i_t}) \le \mathbb{E}_{\mathbf{z}_{i_1}, \cdots, \mathbf{z}_{i_t}}[g_{t+1}(\mathbf{z}_{i_1}, \cdots, \mathbf{z}_{i_t})] + (M + GB_t)\sqrt{\frac{\ln(1/\gamma)}{2t}}.$$

For any $\mathbf{w} \in \mathcal{W}_t$, let $f(\mathbf{w}, \mathbf{z}') = \frac{1}{t}\frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}')$. Then we can write $\mathbb{E}_{\mathbf{z}_{i_1}, \cdots, \mathbf{z}_{i_t}}[g_{t+1}(\mathbf{z}_{i_1}, \cdots, \mathbf{z}_{i_t})]$ as follow

$$\begin{aligned} &\mathbb{E}_{\{\mathbf{z}_{i_k}\}}[g_{t+1}(\mathbf{z}_{i_1}, \cdots, \mathbf{z}_{i_t})] = \mathbb{E}_{\{\mathbf{z}_{i_k}\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t} \mathbb{E}_{\{\mathbf{z}'_{i_k}\}}\Big[\sum_{k=1}^t f(\mathbf{w}, \mathbf{z}'_{i_k})\Big] - \sum_{k=1}^t f(\mathbf{w}, \mathbf{z}_{i_k})\Big] \\ &\le \mathbb{E}_{\{\mathbf{z}_{i_k}, \mathbf{z}'_{i_k}\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t} \sum_{k=1}^t f(\mathbf{w}, \mathbf{z}'_{i_k}) - \sum_{k=1}^t f(\mathbf{w}, \mathbf{z}_{i_k})\Big] = \mathbb{E}_{\{\mathbf{z}_{i_k}, \mathbf{z}'_{i_k}, \sigma_k\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t} \sum_{k=1}^t \sigma_k\big(f(\mathbf{w}, \mathbf{z}'_{i_k}) - f(\mathbf{w}, \mathbf{z}_{i_k})\big)\Big] \\ &\le \frac{2}{t}\mathbb{E}_{\{\mathbf{z}_{i_k}, \sigma_k\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t} \sum_{k=1}^t \sigma_k \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_{i_k})\Big] \le \frac{2}{t}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\{\mathbf{z}_{i_k}, \sigma_k\}}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t} \sum_{k=1}^t \sigma_k \ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_{i_k})\Big] = 2\mathcal{R}_t(\ell \circ \mathcal{W}_t). \end{aligned}$$

Thus we have, with probability at least $1 - \gamma$,

$$P_t \le 2\mathcal{R}_t(\ell \circ \mathcal{W}_t) + (M + GB_t)\sqrt{\frac{\ln(1/\gamma)}{2t}}.$$

The Lemma holds by applying a union bound on $P_t$ and taking the average over $t$. □

**Lemma A.8.** *Suppose $\ell$ is nonnegative, convex and G-Lipschitz. Let $\mathcal{W}_B = \{\mathbf{w} \in \mathcal{W} \,|\, \|\mathbf{w}\|_2 \le B\}$ and let $M = \sup_{\mathbf{z}, \mathbf{z}'} \ell(0, \mathbf{z}, \mathbf{z}')$. With probability $1 - \gamma$, we have*

$$\frac{1}{T}\sum_{t=1}^T \hat{L}_{t+1}(\mathbf{w}_*) - R_S(\mathbf{w}_*) \le \frac{2}{T}\sum_{t=1}^T \mathcal{R}_t(\ell \circ \mathcal{W}_B) + 3(M + GB)\sqrt{\frac{\ln(T/\gamma)}{T}}.$$

*Proof.* Similar to the proof of Lemma A.7 by replacing $\mathbf{w}_t$ with $\mathbf{w}_*$. □

**Lemma A.9.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Suppose $\|\mathbf{w}_*\|_2 \leq B$. Suppose we run Algorithm 1 for $T$ iterations, then we have*

$$\frac{1}{T}\sum_{t=1}^{T}\hat{L}_{t+1}(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_*) \leq \frac{B^2}{2T\eta} + \frac{\eta G^2}{2}$$

*Proof.* By the update rule of Algorithm 1, we have

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 =& \big\|\mathbf{w}_t - \eta\partial\hat{L}_{t+1}(\mathbf{w}_t) - \mathbf{w}_*\big\|_2^2 = \|\mathbf{w}_t - \mathbf{w}_*\|_2^2 + \eta^2\big\|\partial\hat{L}_{t+1}(\mathbf{w}_t)\big\|_2^2 - 2\eta\big\langle\mathbf{w}_t - \mathbf{w}_*, \partial\hat{L}_{t+1}(\mathbf{w}_t)\big\rangle\\ \leq& \|\mathbf{w}_t - \mathbf{w}_*\|_2^2 + \eta^2G^2 - 2\eta\big\langle\mathbf{w}_t - \mathbf{w}_*, \partial\hat{L}_{t+1}(\mathbf{w}_t)\big\rangle.\end{aligned}$$

Therefore, by the convexity of $\hat{L}_{t+1}$, we have

$$\begin{aligned}\sum_{t=1}^{T}\hat{L}_{t+1}(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_*) \leq \sum_{t=1}^{T}\big\langle\mathbf{w}_t - \mathbf{w}_*, \partial\hat{L}_{t+1}(\mathbf{w}_t)\big\rangle \leq \sum_{t=1}^{T}\frac{\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2}{2\eta} + \frac{T\eta G^2}{2}\\ \leq \frac{\|\mathbf{w}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2},\end{aligned}$$

the Lemma holds by dividing $T$ over both sides. □

*Proof of Theorem A.8.* By the convexity of the empirical loss $R_S$, we have

$$\begin{aligned}R_S(\bar{\mathbf{w}}_T) - R_S(\mathbf{w}_*) \leq& \frac{1}{T}\sum_{t=1}^{T}R_S(\mathbf{w}_t) - R_S(\mathbf{w}_*)\\ =& \frac{1}{T}\sum_{t=1}^{T}\big(R_S(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_t) + \hat{L}_{t+1}(\mathbf{w}_*) - R_S(\mathbf{w}_*) + \hat{L}_{t+1}(\mathbf{w}_t) - \hat{L}_{t+1}(\mathbf{w}_*)\big).\end{aligned}\tag{A.15}$$

The conclusion follows from Lemma A.7, A.8 both with probability $1-\gamma/2$ and Lemma A.9. □

## A.2 Proof of Theorem 2

**Theorem A.9** (Theorem 2 restated)**.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Suppose $\mathcal{W}$ is bounded with diameter $D$ and let $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$. Assume we run Algorithm 1 for $T \geq n$ iterations under random selection with replacement rule. For any $\gamma \in (0,1)$, with probability at least $1-\gamma$, with respect to the sample $S$ and the internal randomness of Algorithm 1, we have*

$$\begin{aligned}R(\bar{\mathbf{w}}_T) - R(\mathbf{w}_*) \leq& \frac{4}{T}\sum_{t=1}^{T}\mathcal{R}_t(\ell\circ\mathcal{W}) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + 6(M+GD)\sqrt{\frac{\ln(6T/\gamma)}{n}} + 19e(M+GD)\sqrt{\frac{\ln(6e/\gamma)}{n}}\\ &+2\sqrt{e}\eta G(4+48\sqrt{6}eG\lceil\ln(n)\rceil\ln(6e/\gamma))\Big(\sqrt{T}+\frac{\sqrt{3}T\ln(eT)\ln(6/\gamma)}{n}\Big).\end{aligned}$$

*Proof of Theorem A.9.* Since $\mathcal{W}$ is bounded by $D$, we have $B = B_t = D$. Furthermore, by Lemma A.6, we have $\sup_{\mathbf{z},\mathbf{z}'}\ell(\mathbf{w}_*,\mathbf{z},\mathbf{z}') \leq M + GD$ and $\sup_{\mathbf{z},\mathbf{z}'}\ell(\mathbf{w}_t,\mathbf{z},\mathbf{z}') \leq M + GD$. The proof is completed by recalling the error decomposition (A.14), applying Corollary A.6, Theorem A.7 and A.8 each with probability $1-\gamma/3$. □

## A.3 Proof of Theorem 3

**Theorem A.10** (Theorem 3 restated)**.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Denote $M = \sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$ and $D = \|\mathbf{w}_*\|_2$. Assume we run Algorithm 1 for $T \geq n$ iterations under random selection* *with replacement rule. For any $\gamma \in (0,1)$, with probability at least $1-\gamma$ with respect to the sample $S$ and internal randomness of Algorithm 1, we have*

$$R(\bar{\mathbf{w}}_T)-R(\mathbf{w}_*)\leq\frac{2}{T}\sum_{t=1}^{T}\big(\mathcal{R}_t(\ell\circ\mathcal{W}_t)+\mathcal{R}_t(\ell\circ\mathcal{W}_D)\big)+\frac{D^2}{2T\eta}+\frac{\eta G^2}{2}+(6M+3GD)\sqrt{\frac{\ln(6T/\gamma)}{n}}+3G\sqrt{(G^2+2M)\eta\ln(6T/\gamma)}$$
$$+2\sqrt{e}\eta G(4+48\sqrt{6}eG\lceil\ln(n)\rceil\ln(6e/\gamma))\Big(\sqrt{T}+\frac{\sqrt{3}T\ln(eT)\ln(6/\gamma)}{n}\Big)$$
$$+12\sqrt{2}e(M+G\sqrt{(G^2+2M)\eta T})\sqrt{\frac{\ln(6e/\gamma)}{n}}+(M+GD)\sqrt{\frac{\ln(3/\gamma)}{n}}.$$

Although the boundedness assumption on the parameter space $\mathcal{W}$ is removed, the next lemma characterizes the bound of the iterates $\mathbf{w}_t$ by the sum of stepsizes.

**Lemma A.10.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Denote $M=\sup_{\mathbf{z},\mathbf{z}'}\ell(0,\mathbf{z},\mathbf{z}')$. Let $\{\mathbf{w}_t\}$ be the sequence of iterates by Algorithm 1 with $\eta\leq 1$. Then*

$$\|\mathbf{w}_{t+1}\|_2^2\leq(G^2+2M)\eta t.$$

*Proof.* By the update rule of Algorithm 1, we have

$$\begin{aligned}\|\mathbf{w}_{t+1}\|_2^2=&\big\|\mathbf{w}_t-\eta\partial\hat{L}_{t+1}(\mathbf{w}_t)\big\|_2^2=\|\mathbf{w}_t\|_2^2+\eta^2\big\|\partial\hat{L}_{t+1}(\mathbf{w}_t)\big\|_2^2-2\eta\big\langle\mathbf{w}_t,\partial\hat{L}_{t+1}(\mathbf{w}_t)\big\rangle\\ \leq&\|\mathbf{w}_t\|_2^2+\eta G^2-2\eta\big\langle\mathbf{w}_t,\partial\hat{L}_{t+1}(\mathbf{w}_t)\big\rangle\leq\|\mathbf{w}_t\|_2^2+\eta G^2+2\eta\big(\hat{L}_{t+1}(0)-\hat{L}_{t+1}(\mathbf{w}_t)\big)\\ \leq&\|\mathbf{w}_t\|_2^2+\eta(G^2+2M),\end{aligned}$$

where the first inequality holds since $\ell$ is $G$-Lipschitz and $\eta\leq 1$, the second inequality is due to the convexity of $\ell$ and the last inequality is due to the nonnegativity of $\ell$ and the definition of $M$. □

*Proof of Theorem A.10.* By assumption and Lemma A.10, we have $B=D$ and $B_t=\sqrt{(G^2+2M)\eta t}$. Therefore, by Lemma A.6, we also get $\sup_{\mathbf{z},\mathbf{z}'}\ell(\mathbf{w}_*,\mathbf{z},\mathbf{z}')\leq M+GD$ and $\sup_{\mathbf{z},\mathbf{z}'}\ell(\mathbf{w}_t,\mathbf{z},\mathbf{z}')\leq M+G\sqrt{(G^2+2M)\eta t}$. The proof is completed by recalling the error decomposition (A.14), applying Corollary A.6, Theorem A.7 and A.8 with probability $1-\gamma/3$ each.

□

### A.4 Proof of Theorem 5

In this section, we give utility bound of Algorithm 2. Recall the error decomposition scheme as follows

$$\begin{aligned}\epsilon_{\text{risk}}(\mathbf{w}_{\text{priv}})&=R(\mathbf{w}_{\text{priv}})-R(\mathbf{w}_*)=R(\mathbf{w}_{\text{priv}})-R(\bar{\mathbf{w}}_T)+R(\bar{\mathbf{w}}_T)-R(\mathbf{w}_*)\\&=R(\mathbf{w}_{\text{priv}})-R(\bar{\mathbf{w}}_T)+R(\bar{\mathbf{w}}_T)-R_S(\bar{\mathbf{w}}_T)+R_S(\bar{\mathbf{w}}_T)-R_S(\mathbf{w}_*)+R_S(\mathbf{w}_*)-R(\mathbf{w}_*).\end{aligned}\tag{A.16}$$

Notice that $R(\bar{\mathbf{w}}_T)-R(\mathbf{w}_*)$ yields similar excess risk as Theorem A.9. Hence the difference here is the additional term $R(\mathbf{w}_{\text{priv}})-R(\bar{\mathbf{w}}_T)$ due to the added noise $\mathbf{u}$. The next lemma is a Chernoff type bound for the $\ell_2$ norm of Gaussian vectors.

**Lemma A.11** (Chernoff bound for the $\ell_2$ norm of Gaussian vector)**.** *Let $X_1,\cdots,X_d$ be i.i.d standard Gaussian random variables and $X=[X_1,\cdots,X_d]\in\mathbb{R}^d$. Then for any $\tilde{\gamma}\in(0,1)$, with probability at least $1-\exp(-d\tilde{\gamma}^2/8)$ there holds $\|X\|_2^2\leq(1+\tilde{\gamma})d$.*

The next lemma tells us the error incurred by $R(\mathbf{w}_{\text{priv}})-R(\bar{\mathbf{w}}_T)$ is bounded by the added noise $\mathbf{u}$.

**Lemma A.12.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz. Consider $\mathbf{w}_{priv}$ and $\bar{\mathbf{w}}_T$ from Algorithm 2. For any $\gamma>0$, and for any $\gamma\in(\exp(-d/8),1)$, with probability at least $1-\gamma$, we have*

$$R(\mathbf{w}_{priv})-R(\bar{\mathbf{w}}_T)\leq 2G\sigma\sqrt{d}\ln^{1/4}(1/\gamma).$$

*Proof.* By the definition of $R$, we have

$$\begin{aligned}R(\mathbf{w}_{\text{priv}}) - R(\bar{\mathbf{w}}_T) =&\mathbb{E}_{\mathbf{z},\mathbf{z}'}[\ell(\mathbf{w}_{\text{priv}}, \mathbf{z}, \mathbf{z}') - \ell(\bar{\mathbf{w}}_T, \mathbf{z}, \mathbf{z}')]\\ \leq&\mathbb{E}_{\mathbf{z},\mathbf{z}'}[\langle \mathbf{w}_{\text{priv}} - \bar{\mathbf{w}}_T, \partial\ell(\mathbf{w}_{\text{priv}}, \mathbf{z}, \mathbf{z}')\rangle]\\ \leq&\mathbb{E}_{\mathbf{z},\mathbf{z}'}[\|\Pi_{\mathcal{W}}(\bar{\mathbf{w}}_T + \mathbf{u}) - \bar{\mathbf{w}}_T\|_2 \|\partial\ell(\mathbf{w}_{\text{priv}}, \mathbf{z}, \mathbf{z}')\|_2]\\ \leq&\mathbb{E}_{\mathbf{z},\mathbf{z}'}[\|\mathbf{u}\|_2 \|\partial\ell(\mathbf{w}_{\text{priv}}, \mathbf{z}, \mathbf{z}')\|_2]\\ \leq&G\|\mathbf{u}\|_2 \end{aligned} \tag{A.17}$$

where the first inequality is due to the convexity of $\ell$, the second inequality is by Cauchy-Schwarz inequality, the third inequality is by the non-expansiveness of projection and the last inequality is because $\ell$ is $G$-Lipschitz for any $\mathbf{w}, \mathbf{z}, \mathbf{z}'$. Now, since $\mathbf{u} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, then by Lemma A.11, for $\gamma \in (\exp(-d/8), 1)$ we have with probability $1 - \gamma$,

$$\|\mathbf{u}\|_2 \leq \sigma\sqrt{d}\Big(1 + \Big(\frac{8\ln(1/\gamma)}{d}\Big)^{1/4}\Big).$$

Plugging the above inequality back into Equation (A.17) we get the desired result. □

**Theorem A.11** (Theorem 5 restated)**.** *Suppose $\ell$ is nonnegative, convex and $G$-Lipschitz, and $\mathcal{W}$ is bounded with diameter $D$. Consider Algorithm 2 for $T$ iterations under random selection with replacement rule. For any privacy budget $\epsilon > 0$, any $\delta > 0$ and for any $\gamma \in (\max\{4\delta, \exp(-d/8)\}, 1)$, with probability at least $1 - \gamma$, we have*

$$\begin{aligned}R(\mathbf{w}_{priv}) - R(\mathbf{w}_*) \leq \frac{4}{T}\sum_{t=1}^{T}\mathcal{R}_t(\ell \circ \mathcal{W}) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + 6(M + GD)\sqrt{\frac{\ln(8T/\gamma)}{n}} + 19e(M+GD)\sqrt{\frac{\ln(8e/\gamma)}{n}}\\ + 2\sqrt{e}G\eta\big(4 + 48\sqrt{6}G\lceil\ln(n)\rceil\ln(8e/\gamma)\big)\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(2/\delta)}{n}\Big) + 2G\sigma\sqrt{d}\ln^{1/4}(4/\gamma).\end{aligned}$$

*Proof.* For any neighborhood datasets $S$ and $S'$, Theorem 1 implies with probability least $1 - \delta/2$ that

$$\|\bar{\mathbf{w}}_T - \bar{\mathbf{w}}'_T\|_2 \leq 2\sqrt{e}G\eta\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(2/\delta)}{n}\Big). \tag{A.18}$$

Since $\gamma \geq 4\delta$, we know the (A.18) holds with probability at least $1 - \gamma/8$. Applying Corollary A.6 with (A.18) we know with probability at least $1 - \gamma/4$ we have

$$\begin{aligned}R(\bar{\mathbf{w}}_T) - R_S(\bar{\mathbf{w}}_T) \leq 2\sqrt{e}G\eta\big(4 + 48\sqrt{6}G\lceil\ln(n)\rceil\ln(8e/\gamma)\big)\Big(\sqrt{T} + \frac{\sqrt{3}T\ln(eT)\ln(2/\delta)}{n}\Big)\\ + 12\sqrt{2}e(M + GD)\sqrt{\frac{\ln(8e/\gamma)}{n}}.\end{aligned} \tag{A.19}$$

Recalling the error decomposition (6) and applying Theorem A.7, Theorem A.8 and Lemma A.12 each with probability $1 - \gamma/4$ together with (A.19), we have the desired bound. □

### A.5 Rademacher Complexity for AUC Maximization and Similarity Metric Learning

Firstly we look at the Rademacher complexity for AUC maximization.

**Lemma A.13.** *Given the parameter space $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d | \|\mathbf{w}\|_2 \leq D\}$, and denote $\kappa = \sup_{\mathbf{x}}\|\mathbf{x}\|_2$. the Rademacher complexity of $\mathcal{H} = \{h_{\mathbf{w}} | \mathbf{w} \in \mathcal{W}\}$ can be upper bounded by $R_t(\mathcal{H}) \leq \frac{2D\kappa}{\sqrt{t}}$.*

*Proof.* Starting with the definition, the Rademacher complexity can be upper bounded by

$$\begin{aligned} R_t(\mathcal{H}) =& \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\sup_{h\in\mathcal{H}} \frac{1}{t}\sum_{k=1}^t \sigma_k h_{\mathbf{w}}(\mathbf{x}_i,\mathbf{x}_{i_k})\Big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t} \frac{1}{t}\sum_{k=1}^t \sigma_k \langle \mathbf{w}, \mathbf{x}_i - \mathbf{x}_{i_k}\rangle\Big] \\ \leq& \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}_t} \|\mathbf{w}\|_2 \Big\|\frac{1}{t}\sum_{k=1}^t \sigma_k(\mathbf{x}_i - \mathbf{x}_{i_k})\Big\|_2\Big] \leq \frac{D}{nt}\sum_{i=1}^n \Big(\mathbb{E}\Big[\Big\|\sum_{k=1}^t \sigma_k(\mathbf{x}_i - \mathbf{x}_{i_k})\Big\|_2^2\Big]\Big)^{\frac{1}{2}} \\ =& \frac{D}{nt}\sum_{i=1}^n \Big(\sum_{k=1}^t \mathbb{E}\Big[\|\mathbf{x}_i - \mathbf{x}_{i_k}\|_2^2\Big]\Big)^{\frac{1}{2}} \leq \frac{2D\kappa}{\sqrt{t}} \end{aligned}$$

where the first inequality is due to Cauchy-Schwarz inequality, the third identity is due to $\{\sigma_k\}_{k=1}^t$ are independent random variables with mean zero. □

Next we turn our focus to similarity metric learning.

**Lemma A.14.** *Consider the parameter space defined via the nuclear norm $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^{d\times d}, \|\mathbf{w}\|_{S_1} \leq D\}$, where $\|\mathbf{w}\|_{S_1}$ denotes the nuclear norm of a matrix $\mathbf{w}$. The complexity of $\mathcal{H} = \{h_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ is bounded by*

$$R_t(\mathcal{H}) = \mathcal{O}\Big(\frac{D\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}\sqrt{\log d}}{\sqrt{t}}\Big), \tag{A.20}$$

*where $\|\cdot\|_{S_\infty}$ denotes the largest singular value.*

*Proof.* The complexity of $\mathcal{H} = \{h_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ is bounded by

$$\begin{aligned} R_t(\mathcal{H}) =& \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \frac{1}{t}\sum_{k=1}^t \sigma_k \Big\langle \mathbf{w}, (\mathbf{x}_i - \mathbf{x}_{i_k})(\mathbf{x}_i - \mathbf{x}_{i_k})^\top\Big\rangle\Big] \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\sup_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w}\|_{S_1} \Big\|\frac{1}{t}\sum_{k=1}^t \sigma_k(\mathbf{x}_i - \mathbf{x}_{i_k})(\mathbf{x}_i - \mathbf{x}_{i_k})^\top\Big\|_{S_\infty}\Big] \\ \leq& \frac{D}{nt}\sum_{i=1}^n \mathbb{E}\Big\|\sum_{k=1}^t \sigma_k(\mathbf{x}_i - \mathbf{x}_{i_k})(\mathbf{x}_i - \mathbf{x}_{i_k})^\top\Big\|_{S_\infty} = \mathcal{O}\Big(\frac{D\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}\sqrt{\log d}}{\sqrt{t}}\Big), \end{aligned}$$

where $\|\cdot\|_{S_\infty}$ denotes the largest singular value of a matrix and we have used Lemma A.17 in the last step. □

For any $p \geq 1$, the Schatten-$p$ norm of a matrix $W \in \mathbb{R}^{d\times d}$ is defined as the $\ell_p$-norm of the vector of singular values $\sigma(W) := (\sigma_1(W), \dots, \sigma_d(W))^\top$ (the singular values are assumed to be sorted in non-increasing order), i.e., $\|W\|_{S_p} := \|\sigma(W)\|_p$. Let $\Sigma = \mathbb{E}[XX^\top]$. We assume $d \geq 3$.

The following Khintchine-Kahane inequality Lust-Piquard and Pisier (1991) provides a powerful tool to control the $q$-th norm of the summation of Rademacher series. The following form can be found in Qiu and Wicks (2014).

**Lemma A.15** (Matrix Khintchine)**.** *Let $X_1, \dots, X_n$ be a set of symmetric matrices of the same dimension and let $\sigma_1, \dots, \sigma_n$ be a sequence of independent Rademacher random variables. For all $q \geq 2$,*

$$\Big(\mathbb{E}_\sigma\Big\|\sum_{i=1}^n \sigma_i X_i\Big\|_{S_q}^q\Big)^{\frac{1}{q}} \leq 2^{-\frac{1}{4}}\sqrt{\frac{q\pi}{e}}\Big\|\Big(\sum_{i=1}^n X_i^2\Big)^{\frac{1}{2}}\Big\|_{S_q}. \tag{A.21}$$

The following inequality is the Bernstein inequality for a summation of independent matrices Tropp (2015).

**Lemma A.16** (Matrix Bernstein)**.** *Let $Z_1, \dots, Z_n$ be independent, mean-zero and symmetric random matrices in $\mathbb{R}^{d\times d}$. Assume that each one is uniformly bounded*

$$\mathbb{E}[Z_i] = 0 \quad \textit{and} \quad \|Z_i\|_{S_\infty} \leq L \quad \textit{for each } i = 1, \dots, n.$$

*Introduce the sum* $S = \sum_{i=1}^n Z_i$ *and let* $v(S)$ *denote the matrix variance statistic of the sum*

$$v(S) = \Big\| \sum_{i=1}^n \mathbb{E}[Z_i^2] \Big\|_{S_\infty}.$$

*Then*

$$\mathbb{E}\big[\|S\|_{S_\infty}\big] \le \sqrt{2v(S)\log(2d)} + \frac{L}{3}\log(2d).$$

**Lemma A.17.** *Let* $\sigma_1, \ldots, \sigma_n$ *be independent Rademacher variables. Then*

$$\frac{1}{n}\mathbb{E}\Big\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} \le 2^{\frac{1}{4}}\sqrt{\pi \log d}\Big( \frac{\sqrt{\log(2d)}\sup_{\mathbf{x}} \|\mathbf{x}\|_2^2}{n} + \frac{2\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}}{\sqrt{n}} \Big). \tag{A.22}$$

*Under the mild assumption* $\sqrt{\log(2d)}\sup_{\mathbf{x}} \|\mathbf{x}\|_2^2 \le \sqrt{n}\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}$ *we get*

$$\frac{1}{n}\mathbb{E}\Big\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} = \mathcal{O}\Big( \frac{\sqrt{\log d}\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}}{\sqrt{n}} \Big).$$

*Proof.* By the concavity of the square-root function, we know

$$\begin{aligned}
\mathbb{E}_\sigma \Big\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} &\le \Big( \mathbb{E}_\sigma \Big\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_q}^q \Big)^{\frac{1}{q}} \le 2^{-\frac{1}{4}}\sqrt{\frac{q\pi}{e}} \Big\| \Big( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \Big)^{\frac{1}{2}} \Big\|_{S_q} \\
&\le 2^{-\frac{1}{4}}\sqrt{\frac{q\pi}{e}} d^{\frac{1}{q}} \Big\| \Big( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \Big)^{\frac{1}{2}} \Big\|_{S_\infty},
\end{aligned}$$

where we have used Lemma A.15 and $\|W\|_{S_\infty} \le \|W\|_{S_q} \le d^{\frac{1}{q}}\|W\|_{S_\infty}$ for all $W \in \mathbb{R}^{d\times d}$. If we choose $q = 2\log d$ ($d \ge 3$), then

$$\sqrt{q} d^{\frac{1}{q}} = \sqrt{2\log d} d^{\frac{1}{2\log d}} = \sqrt{2e\log d}$$

and therefore

$$\mathbb{E}_\sigma \Big\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} \le 2^{\frac{1}{4}}\sqrt{\pi\log d} \Big\| \Big( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \Big)^{\frac{1}{2}} \Big\|_{S_\infty} = 2^{\frac{1}{4}}\sqrt{\pi\log d} \Big( \Big\| \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} \Big)^{\frac{1}{2}}.$$

It then follows from the concavity of the square-root function that

$$\mathbb{E}\Big\| \sum_{i=1}^n \sigma_i \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} \le 2^{\frac{1}{4}}\sqrt{\pi\log d} \Big( \mathbb{E}\Big\| \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} \Big)^{\frac{1}{2}} \tag{A.23}$$

It is clear

$$\begin{aligned}
\mathbb{E}\Big[ \Big\| \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} \Big] &\le \mathbb{E}\Big[ \Big\| \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top - \mathbb{E}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} \Big] + \Big\| \mathbb{E}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} \\
&= \mathbb{E}\Big[ \Big\| \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top - \mathbb{E}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \Big\|_{S_\infty} \Big] + n\Big\| \mathbb{E}\big[\|X\|_2^2 XX^\top\big] \Big\|_{S_\infty}.
\end{aligned} \tag{A.24}$$

For all $i \in [n]$, denote $Z_i = \|\mathbf{x}_i\|_2^2 \mathbf{x}_i \mathbf{x}_i^\top - \mathbb{E}\big[\|\mathbf{x}_i\|_2^2 \mathbf{x}_i \mathbf{x}_i^\top\big]$. It is clear that

$$\begin{aligned}
\mathbb{E}\big[ \sum_{i=1}^n Z_i^2 \big] &= \sum_{i=1}^n \mathbb{E}\Big[ \|\mathbf{x}_i\|_2^6 \mathbf{x}_i \mathbf{x}_i^\top \Big] - \sum_{i=1}^n \Big( \mathbb{E}[\|\mathbf{x}_i\|_2^2 \mathbf{x}_i \mathbf{x}_i^\top] \Big)\Big( \mathbb{E}[\|\mathbf{x}_i\|_2^2 \mathbf{x}_i \mathbf{x}_i^\top] \Big) \\
&= n\mathbb{E}\big[\|X\|_2^6 XX^\top\big] - n\mathbb{E}\big[\|X\|_2^2 XX^\top\big]\mathbb{E}\big[\|X\|_2^2 XX^\top\big] \preceq n\mathbb{E}\big[\|X\|_2^6 XX^\top\big]
\end{aligned}$$

and therefore

$$\Big\|\mathbb{E}\big[\sum_{i=1}^{n} Z_i^2\big]\Big\|_{S_\infty} \le n\Big\|\mathbb{E}\big[\|X\|_2^6 XX^\top\big]\Big\|_{S_\infty}. \tag{A.25}$$

Furthermore,

$$\|Z_i\|_{S_\infty} \le \sup_{\mathbf{x}_i} \|\mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\|_{S_\infty} \le \sup_{\mathbf{x}} \|\mathbf{x}\|_2^4. \tag{A.26}$$

We can apply Lemma A.16 with the above bound of variance (A.25) and magnitude (A.26), and derive

$$\mathbb{E}\Big[\|\sum_{i=1}^{n} Z_i\|_{S_\infty}\Big] \le \sqrt{2n\big\|\mathbb{E}\big[\|X\|_2^6 XX^\top\big]\big\|_{S_\infty}\log(2d)} + \frac{1}{3}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^4\log(2d).$$

This together with the sub-additivity of the square-root function and (A.24) implies

$$\begin{aligned}
&\Big(\mathbb{E}\Big[\|\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\|_{S_\infty}\Big]\Big)^{\frac{1}{2}} \\
&\le \bigg(\mathbb{E}\Big[\Big\|\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top - \mathbb{E}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty}\Big]\bigg)^{\frac{1}{2}} + \Big(n\Big\|\mathbb{E}[\|X\|_2^2 XX^\top]\Big\|_{S_\infty}\Big)^{\frac{1}{2}} \\
&\le \Big(2n\big\|\mathbb{E}\big[\|X\|_2^6 XX^\top\big]\big\|_{S_\infty}\log(2d)\Big)^{\frac{1}{4}} + \frac{\sqrt{\log(2d)}}{\sqrt{3}}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^2 + \sqrt{n\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}}.
\end{aligned}$$

We plug the above inequality back into (A.23), and get the inequality

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\Big\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\mathbf{x}_i^\top\Big\|_{S_\infty} \le 2^{\frac{1}{4}}\sqrt{\pi\log d}\bigg(&(2\log(2d))^{\frac{1}{4}}n^{-\frac{3}{4}}\big\|\mathbb{E}\big[\|X\|_2^6 XX^\top\big]\big\|_{S_\infty}^{\frac{1}{4}} \\
&+ \frac{\sqrt{\log(2d)}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^2}{\sqrt{3}n} + \frac{\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}}{\sqrt{n}}\bigg).
\end{aligned} \tag{A.27}$$

It is clear that

$$\Big\|\mathbb{E}\big[\|X\|_2^6 XX^\top\big]\Big\|_{S_\infty}^{\frac{1}{4}} \le \sup_{\mathbf{x}}\|\mathbf{x}\|_2\big\|\mathbb{E}[\|X\|_2^2]XX^\top\big\|_{S_\infty}^{\frac{1}{4}}.$$

This together with Cauchy-Schwartz inequality shows that

$$(2\log(2d))^{\frac{1}{4}}n^{-\frac{3}{4}}\big\|\mathbb{E}\big[\|X\|_2^6 XX^\top\big]\big\|_{S_\infty}^{\frac{1}{4}} \le \frac{\big\|\mathbb{E}[\|X\|_2^2 XX^\top]\big\|_{S_\infty}^{\frac{1}{2}}}{\sqrt{n}} + \frac{\sqrt{\log(2d)}\sup_{\mathbf{x}}\|\mathbf{x}\|_2^2}{2^{\frac{3}{2}}n}.$$

Plugging the above inequality back into (A.27) gives the stated bound (A.22) ($2^{-\frac{3}{2}} + 3^{-\frac{1}{2}} < 1$). The proof is complete. □