A Supplementary

A.1 Supporting Theorems and Lemmas

Let us recall the excess risk of a randomized algorithm \( \mathcal{A} \) defined as 
\[
\epsilon_{\text{risk}}(\mathcal{A}(S)) = R(\mathcal{A}(S)) - R(w_0),
\]
which can be decomposed by
\[
\epsilon_{\text{risk}}(\mathcal{A}(S)) = R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) + R_S(w_0) - R_S(\mathcal{A}(S)) - R_S(w_0).
\]
(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 \( \| w \| \). A simple lemma indicates that if \( w \) is bounded, then \( \ell(w, \cdot, \cdot) \) is also bounded. In the subsequent sections, we will characterize the bounds for the iterates \( \{ w_t \} \) whenever the parameter space \( \mathcal{W} \) is bounded or unbounded.

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

**Proof.** By convexity of \( \ell \), we have for any \( z, z' \)
\[
\ell(w, z, z') \leq \sup_{z,z'} \ell(0, z, z') + (w, \partial \ell(w, z, z')) \leq M + \| w \| \| \ell(w, z, z') \|_2 \leq M + GB
\]
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 \( M = \sup_{z,z'} \ell(0, z, z') \). 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 \( \mathcal{A} \),
\[
R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) \leq 4\epsilon + 48\sqrt{6}eGe[\ln(n)] \ln(\epsilon/\gamma) + 12\sqrt{2}e(M + GB) \sqrt{\frac{\ln(\epsilon/\gamma)}{n}}.
\]

**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
\[
|E[\ell(\mathcal{A}(S), z, z')]| \leq M + GB
\]
for any \( z, z' \). For the uniform stability, since \( \mathcal{A} \) is \( \epsilon \)-UAS, by the Lipschitz continuity of \( \ell \) we have
\[
\sup_{z,z'} |\ell(\mathcal{A}(S), z, z') - \ell(\mathcal{A}(S'), z, z')| \leq G\| \mathcal{A}(S) - \mathcal{A}(S') \|_2 \leq G\epsilon.
\]
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 \( S, S' \),
\[
\Pr_{\mathcal{A}}[\| \mathcal{A}(S) - \mathcal{A}(S') \|_2 > \epsilon] \leq \gamma_0.
\]

Suppose \( \ell \) is nonnegative, convex and \( G \)-Lipschitz. Suppose the output of \( \mathcal{A} \) is bounded by \( B \) and let \( M = \sup_{z,z'} \ell(0, z, z') \). Then we have for any \( \gamma \in (0, 1) \),
\[
\Pr_{S,\mathcal{A}}[R(\mathcal{A}(S)) - R_S(\mathcal{A}(S)) > 4\epsilon + 48\sqrt{6}eGe[\ln(n)] \ln(\epsilon/\gamma) + 12\sqrt{2}e(M + GB) \sqrt{\frac{\ln(\epsilon/\gamma)}{n}}] \leq \gamma + \gamma_0.
\]
In order to prove Theorem A.8, we decompose Algorithm 1 for sequence Theorem A.8. et al. (2013), where they consider the online-to-batch generalization bound for pairwise learning. Our bound in the next theorem gives a bound on probability at least 1 − γ with probability at least 1 − γ. If we run Algorithm 1 for Corollary A.6. for Algorithm 1. Combining Corollary A.5 and the stability result in Theorem 1, we arrive at the following generalization bound Theorem A.6, for any Proof. Denote $E = \{A\|A(S) - A(S')\| > \epsilon\}$ and $F = \{S, A| R(A(S)) - R_S(A(S)) > 4\epsilon + 48\sqrt{6G}[\ln(n)]\ln(\epsilon/\gamma) + 12\sqrt{2}\epsilon(M + GB)\sqrt{\ln(\epsilon/\gamma)/n}\}$. Then by assumption we have $P_{A}[A \in E] \leq \gamma_0$. By Theorem A.6. for any $\gamma \geq 0$, we have $P_{A}[A \in E|A \notin F] \leq \gamma$. Then the following identity holds

$$
P_{S,A}[S, A \in F] = P_{S,A}[S, A \in F \cap A \in E] + P_{S,A}[S, A \in F \cap A \notin E]$$

$$= P_{S,A}[S, A \in F|A \in E]P[A \in E] + P_{S,A}[S, A \in F|A \notin E]P[A \notin E] \leq \gamma_0 + \gamma.$$

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 $B_T = \|\bar{\nu}_T\|$ and $M = \sup_{z, z'} \ell(0, z, z')$. If 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

$$R(\bar{w}_T) - R_S(\bar{w}_T) \leq 2\sqrt{\epsilon n}G(4 + 48\sqrt{6G}[\ln(n)]\ln(2\epsilon/\gamma))\left(\sqrt{T} + \frac{\sqrt{3T}\ln(\epsilon T)\ln(2\epsilon/\gamma)}{n}\right) + 12\sqrt{2}\epsilon(M + GB_T)\sqrt{\ln(2\epsilon/\gamma)/n}.$$

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

$$\|\bar{w}_T - \bar{w}_T\|_2^2 \leq 4\epsilon n^2G^2\left(T + \frac{3T^2\ln^2(\epsilon T)\ln^2(2\epsilon/\gamma)}{n^2}\right).$$

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

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

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

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

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

Next we give an upper bound on the optimization error $R_S(\bar{w}_T) - R_S(\bar{w}_*)$. 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 $0 \leq B_t < \infty$ such that $\|\bar{w}_t\|_2 \leq B_t$, and let $M = \sup_{z, z'} \ell(0, z, z')$ and $B = \|w_*\|_2$. 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

$$R_S(\bar{w}_T) - R_S(\bar{w}_*) \leq \frac{2}{T} \sum_{t=1}^{T} R_t(\ell \circ \bar{w}) + \frac{2}{T} \sum_{t=1}^{T} R_t(\ell \circ \bar{w}_B) + \frac{B^2}{2T\gamma} + \frac{\eta^2}{2} + (6M + 3GB)\sqrt{\frac{\ln(2T/\gamma)}{T}} + 3GBT\sqrt{\frac{\ln(2T/\gamma)}{T}},$$

where $\bar{w}_t = \{w \in W||w||_2 \leq B_t\}$ and $\bar{w}_B = \{w \in W||w||_2 \leq B\}$ are subspaces of $W$.

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

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

$$\frac{1}{T} \sum_{t=1}^{T} R_S(\bar{w}_t) - \hat{L}_{i+1}(w_t) \leq \frac{2}{T} \sum_{t=1}^{T} R_t(\ell \circ \bar{w}_t) + (M + GB_T)\sqrt{\frac{\ln(T/\gamma)}{T}}.$$
Proof. For any \( w \), denote \( \hat{L}_{t+1}(w) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{t} \ell(w, z_i, z_i^k) \). This allows us to decompose the risk as follows

\[
\frac{1}{T} \sum_{t=1}^{T} R_S(w_t) - \hat{L}_{t+1}(w_t) = \frac{1}{T} \sum_{t=1}^{T} R_S(w_t) - \hat{L}_{t+1}(w_t) + \hat{L}_{t+1}(w_t) - \hat{L}_{t+1}(w_t)
\]

By construction, we have \( \mathbb{E}_{z_i^1, \ldots, z_i^n} [Q_{t+1} | z_i, \ldots, z_i] = 0 \) and hence the sequence \( Q_2, \ldots, Q_T \) 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 \( 1 - \gamma \),

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

We now analyze each term \( P_t \) individually. Let us start by introducing a ghost sample \( \{z_i', \ldots, z_i'\} \), where each \( z_i' \) follows the same distribution as \( z_i \). By linearity of expectation, we have

\[
R_S(w_t) = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{t} \ell(w_t, z_i, z_i^k) \right],
\]

where the expectation is taken over \( \{z_i'\}_{k=1}^{t} \). It allows us to write \( P_t \) as follow

\[
P_t = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{t} \ell(w_t, z_i, z_i^k) - \hat{L}_{t+1}(w_t) \right] \leq \sup_{w \in W_t} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{t} \ell(w, z_i, z_i^k) - \hat{L}_{t+1}(w) \right]
\]

\[
\triangleq g_{t+1}(z_i, \ldots, z_i).
\]

Since \( \ell \) is bounded by \( A_t \), the expression \( g_{t+1}(z_i, \ldots, z_i) \) 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 \( 1 - \gamma \),

\[
g_{t+1}(z_i, \ldots, z_i) \leq \mathbb{E}_{z_i, \ldots, z_i} [g_{t+1}(z_i, \ldots, z_i)] + (M + GB_t) \sqrt{\frac{\ln(1/\gamma)}{2t}}.
\]

For any \( w \in W_t \), let \( f(w, z') = \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i, z_i') \). Then we can write \( \mathbb{E}_{z_i, \ldots, z_i} [g_{t+1}(z_i, \ldots, z_i)] \) as follow

\[
\mathbb{E}_{\{z_i\}} [g_{t+1}(z_i, \ldots, z_i)] = \mathbb{E}_{\{z_i\}} \left[ \sup_{w \in W_t} \mathbb{E}_{\{z_i'\}} \left[ \sum_{k=1}^{t} f(w, z_i^k) \right] - \sum_{k=1}^{t} f(w, z_i) \right]
\]

\[
\leq \mathbb{E}_{\{z_i, z_i'\}} \left[ \sup_{w \in W_t} \sum_{k=1}^{t} f(w, z_i^k) - \sum_{k=1}^{t} f(w, z_i) \right] = \mathbb{E}_{\{z_i, z_i'\}} \left[ \sup_{w \in W_t} \sum_{k=1}^{t} \sigma_k (f(w, z_i^k) - f(w, z_i)) \right]
\]

\[
\leq \frac{2}{t} \mathbb{E}_{\{z_i, z_i'\}} \left[ \sup_{w \in W_t} \sum_{k=1}^{t} \sigma_k \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i, z_i) \right] \leq \frac{2}{t} \mathbb{E}_{\{z_i, z_i'\}} \left[ \sup_{w \in W_t} \sum_{k=1}^{t} \sigma_k \ell(w, z_i, z_i) \right] = 2R_t(\ell \circ W_t).
\]

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

\[
P_t \leq 2R_t(\ell \circ W_t) + (M + GB_t) \sqrt{\frac{\ln(1/\gamma)}{2t}}.
\]

The Lemma holds by applying a union bound on \( P_t \) and taking the average over \( t \).

\[\square\]

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

\[
\frac{1}{T} \sum_{t=1}^{T} \hat{L}_{t+1}(w_t) - R_S(w_t) \leq \frac{2}{T} \sum_{t=1}^{T} R_t(\ell \circ W_B) + 3(M + GB) \sqrt{\frac{\ln(T/\gamma)}{T}}.
\]
Proof. Similar to the proof of Lemma A.7 by replacing \( w_t \) with \( w_* \).

**Lemma A.9.** Suppose \( \ell \) is nonnegative, convex and G-Lipschitz. Suppose \( \|w_*\|_2 \leq B \). Suppose we run Algorithm 1 for \( T \) iterations, then we have

\[
\frac{1}{T} \sum_{t=1}^{T} \hat{L}_{t+1}(w_t) - \hat{L}_{t+1}(w_*) \leq \frac{B^2}{2T\eta} + \frac{\eta G^2}{2}
\]

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

\[
\|w_{t+1} - w_*\|_2 = \|w_t - \eta \partial \hat{L}_{t+1}(w_t) - w_*\|_2 = \|w_t - w_*\|_2 + \eta^2 \|\partial \hat{L}_{t+1}(w_t)\|_2^2 - 2\eta \langle w_t - w_*, \partial \hat{L}_{t+1}(w_t) \rangle
\]

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

\[
\sum_{t=1}^{T} \hat{L}_{t+1}(w_t) - \hat{L}_{t+1}(w_*) \leq \sum_{t=1}^{T} \langle w_t - w_*, \partial \hat{L}_{t+1}(w_t) \rangle \leq \sum_{t=1}^{T} \frac{\|w_t - w_*\|^2_2 - \|w_{t+1} - w_*\|^2_2}{2\eta} + \frac{T\eta G^2}{2}
\]

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

\[
R_S(w_T) - R_S(w_*) \leq \frac{1}{T} \sum_{t=1}^{T} R_S(w_t) - R_S(w_*)
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} (R_S(w_t) - \hat{L}_{t+1}(w_t) + \hat{L}_{t+1}(w_*) - R_S(w_*) + \hat{L}_{t+1}(w_t) - \hat{L}_{t+1}(w_*)). \tag{A.15}
\]

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 \( W \) is bounded with diameter \( D \) and let \( M = \sup_{z,z'} \ell(0, z, z') \). 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

\[
R(w_T) - R(w_*) \leq \frac{4}{T} \sum_{t=1}^{T} R_t(\ell \circ W) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + 6(M + GD)\sqrt{\frac{\ln(6T/\gamma)}{n}} + 19e(M + GD)\sqrt{\frac{\ln(6e/\gamma)}{n}} + 2\sqrt{T\eta G(4 + 48\sqrt{6eG\ln(n)}) \ln(6e/\gamma)} \left( \sqrt{T} + \frac{\sqrt{3T\ln(eT)\ln(6/\gamma)}}{n} \right).
\]

**Proof of Theorem A.8** Since \( W \) is bounded by \( D \), we have \( B = B_t = D \). Furthermore, by Lemma A.6, we have \( \sup_{z,z'} \ell(w_*, z, z') \leq M + GD \) and \( \sup_{z,z'} \ell(w, z, z') \leq M + GD \). 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 \( M = \sup_{z,z'} \ell(0, z, z') \) and \( D = \|w_*\|_2 \). Assume we run Algorithm 1 for \( T \geq n \) iterations under random selection
Stability and Differential Privacy of Non-smooth Pairwise Learning

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

$$R(\bar{w}_T) - R(w^*_T) \leq \frac{2}{T} \sum_{t=1}^{T} (R_t(\ell \circ W_t) + R_t(\ell \circ W_D)) + \frac{D^2}{2T\eta} + \frac{G^2}{2} + (6M + 3GD)\sqrt{\frac{\ln(6T/\gamma)}{n}} + 3G\sqrt{(G^2 + 2M)\eta\ln(6T/\gamma)} + 2\sqrt{\epsilon}G(4 + 48\sqrt{6}\epsilon G[\ln(n)] \ln(6\epsilon/\gamma)) \left( \sqrt{T} + \frac{\sqrt{3T\ln(G^2/\gamma)}}{n} \right) + 12\sqrt{2}\epsilon(M + G\sqrt{(G^2 + 2M)\eta T})\sqrt{\frac{\ln(6\epsilon/\gamma)}{n}} + (M + GD)\sqrt{\frac{\ln(3\epsilon/\gamma)}{n}}.$$  

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

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

$$\|w_{t+1}\|^2 \leq (G^2 + 2M)\eta t.$$  

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

$$\|w_{t+1}\|^2 = \|w_t - \eta \partial \hat{L}_{t+1}(w_t)\|^2 = \|w_t\|^2 + \eta^2 \|\partial \hat{L}_{t+1}(w_t)\|^2 - 2\eta\langle w_t, \partial \hat{L}_{t+1}(w_t) \rangle \leq \|w_t\|^2 + \eta G^2 - 2\eta\langle w_t, \partial \hat{L}_{t+1}(w_t) \rangle \leq \|w_t\|^2 + \eta G^2 + 2\eta (\hat{L}_{t+1}(0) - \hat{L}_{t+1}(w_t)) \leq \|w_t\|^2 + \eta(G^2 + 2M),$$

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 $\sup_{z, z'} \ell(w_t, z, z') \leq M + GD$ and $\sup_{z, z'} \ell(w_t, z, z') \leq M + G\sqrt{(G^2 + 2M)\eta T}$. 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

$$\epsilon_{risk}(w_{priv}) = R(w_{priv}) - R(w^*_T) = R(w_{priv}) - R(\bar{w}_T) + R(\bar{w}_T) - R(w^*_T).$$

Notice that $R(\bar{w}_T) - R(w^*_T)$ yields similar excess risk as Theorem A.9. Hence the difference here is the additional term $R(w_{priv}) - R(\bar{w}_T)$ due to the added noise $u$. 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 $X_1, \ldots, X_d$ be i.i.d standard Gaussian random variables and $X = [X_1, \ldots, X_d] \in \mathbb{R}^d$. Then for any $\tilde{\gamma} \in (0, 1)$, with probability at least $1 - \exp(-d\tilde{\gamma}^2/8)$ there holds $\|X\|^2 \leq (1 + \tilde{\gamma})d$.

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

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

$$R(w_{priv}) - R(\bar{w}_T) \leq 2G\sigma\sqrt{d}ln^{1/4}(1/\gamma).$$
Proof. By the definition of $R$, we have

$$R(w_{\text{priv}}) - R(w_T) = \mathbb{E}_{z,z'}[\ell(w_{\text{priv}}, z, z') - \ell(w_T, z, z')]$$

$$\leq \mathbb{E}_{z,z'}[(w_{\text{priv}} - w_T, \partial \ell(w_{\text{priv}}, z, z'))]$$

$$\leq \mathbb{E}_{z,z'}[\|\Pi_W(w_T + u) - w_T\|_2 \|\partial \ell(w_{\text{priv}}, z, z')\|_2]$$

$$\leq \mathbb{E}_{z,z'}[\|u\|_2 \|\partial \ell(w_{\text{priv}}, z, z')\|_2]$$

$$\leq G\|u\|_2$$ \hspace{1cm} (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 $w, z, z'$. Now, since $u \sim N(0, \sigma^2 I_d)$, then by Lemma A.11 for $\gamma \in (\exp(-d/8), 1)$ we have with probability $1 - \gamma$,

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

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 $W$ 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

$$R(w_{\text{priv}}) - R(w_*) \leq \frac{4}{T} \sum_{t=1}^{T} R_t(\ell \circ W) + \frac{D^2}{2T\eta} + \frac{\eta G^2}{2} + 6(M + GD)\sqrt{\frac{\ln(8T/\gamma)}{n}} + 19e(M + GD)\sqrt{\frac{\ln(8e/\gamma)}{n}}$$

$$+ 2\sqrt{\epsilon G\eta(4 + 48\sqrt{6G[\ln(n)]} \ln(8e/\gamma))} \left(\sqrt{T} + \frac{\sqrt{3T \ln(\epsilon T) \ln(2/\delta)}}{n}\right) + 2G\sigma \sqrt{d \ln^{1/4}(4/\gamma)}.$$

Proof. For any neighborhood datasets $S$ and $S'$, Theorem 5 implies with probability $1 - \delta/2$ that

$$\|w_T - w'_T\|_2 \leq 2\sqrt{\epsilon G\eta\left(\sqrt{T} + \frac{\sqrt{3T \ln(\epsilon T) \ln(2/\delta)}}{n}\right)}.$$ \hspace{1cm} (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

$$R(w_T) - R_S(w_T) \leq 2\sqrt{\epsilon G\eta(4 + 48\sqrt{6G[\ln(n)]} \ln(8e/\gamma))} \left(\sqrt{T} + \frac{\sqrt{3T \ln(\epsilon T) \ln(2/\delta)}}{n}\right)$$

$$+ 12\sqrt{2e(M + GD)}\sqrt{\frac{\ln(8e/\gamma)}{n}}.$$ \hspace{1cm} (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 $W = \{w \in \mathbb{R}^d \|w\|_2 \leq D\}$, and denote $\kappa = \sup_x \|x\|_2$. the Rademacher complexity of $\mathcal{H} = \{h_w | w \in W\}$ can be upper bounded by $R_\epsilon(\mathcal{H}) \leq \frac{2D\kappa}{\sqrt{T}}$. 
The following inequality is the Bernstein inequality for a summation of independent matrices Tropp (2015).

\[ R_t(\mathcal{H}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{h \in \mathcal{H}} \frac{1}{t} \sum_{k=1}^{t} \sigma_k h w(x_i, x_{i_k}) \right] \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{w \in \mathcal{W}_i} \frac{1}{t} \sum_{k=1}^{t} \sigma_k \langle w, x_i - x_{i_k} \rangle \right] \]

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 \( \mathcal{W} = \{ w \in \mathbb{R}^{d \times d}, ||w||_{S_1} \leq D \} \), where \( ||w||_{S_1} \) denotes the nuclear norm of a matrix \( w \). The complexity of \( \mathcal{H} = \{ h_w : w \in \mathcal{W} \} \) is bounded by

\[ R_t(\mathcal{H}) = \mathcal{O} \left( \frac{D \| \mathbb{E}[||X||_2^2 XX^\top] \|_{S_{\infty}}^{\frac{1}{2}}}{\sqrt{t}} \sqrt{\log d} \right), \]  

(A.20)

where \( || \cdot ||_{S_{\infty}} \) denotes the largest singular value.

**Proof.** The complexity of \( \mathcal{H} = \{ h_w : w \in \mathcal{W} \} \) is bounded by

\[ R_t(\mathcal{H}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{w \in \mathcal{W}_i} \frac{1}{t} \sum_{k=1}^{t} \sigma_k \langle w, (x_i - x_{i_k})(x_i - x_{i_k})^\top \rangle \right] \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{w \in \mathcal{W}_i} \frac{1}{t} \sum_{k=1}^{t} \sigma_k \langle w, (x_i - x_{i_k})(x_i - x_{i_k})^\top \rangle \right] \]

\[ \leq \frac{D}{nt} \sum_{i=1}^{n} \mathbb{E} \left[ \sup_{w \in \mathcal{W}_i} \frac{1}{t} \sum_{k=1}^{t} \sigma_k \langle w, (x_i - x_{i_k})(x_i - x_{i_k})^\top \rangle \right] \leq \mathcal{O} \left( \frac{D \| \mathbb{E}[||X||_2^2 XX^\top] \|_{S_{\infty}}^{\frac{1}{2}}}{\sqrt{t}} \sqrt{\log d} \right), \]

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 \( \sigma(W) := (\sigma_1(W), \ldots, \sigma_d(W))^\top \) (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 \( X_1, \ldots, X_n \) be a set of symmetric matrices of the same dimension and let \( \sigma_1, \ldots, \sigma_n \) be a sequence of independent Rademacher random variables. For all \( q \geq 2 \),

\[ \left( \mathbb{E}[\| \sum_{i=1}^{n} \sigma_i X_i \|_{S_q}^q] \right)^{\frac{1}{q}} \leq 2^{-\frac{q}{2}} \sqrt{\frac{q}{e}} \left( \sum_{i=1}^{n} X_i^2 \right)^{\frac{1}{2}} \| S_q \]. \]  

(A.21)

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

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

\[ \mathbb{E}[Z_i] = 0 \quad \text{and} \quad \| Z_i \|_{S_{\infty}} \leq L \quad \text{for each} \quad i = 1, \ldots, n. \]
Introduce the sum $S = \sum_{i=1}^{n} Z_i$ and let $v(S)$ denote the matrix variance statistic of the sum

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

Then

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

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

$$
\frac{1}{n} \mathbb{E} \left\| \sum_{i=1}^{n} \sigma_i x_i x_i^T \right\|_{S_{\infty}} \leq \frac{2^{\frac{1}{4}} \sqrt{\pi} \log d}{n} \left( \frac{\log(2d) \sup_x \|x\|_2}{n} + \frac{2 \left\|\mathbb{E}[\|X\|_2 X X^T]\right\|_{S_{\infty}}}{\sqrt{n}} \right).
$$

Under the mild assumption $\sqrt{\log(2d) \sup_x \|x\|_2} \leq \sqrt{n} \left\|\mathbb{E}[\|X\|_2 X X^T]\right\|_{S_{\infty}}\frac{3}{4}$ we get

$$
\frac{1}{n} \mathbb{E} \left\| \sum_{i=1}^{n} \sigma_i x_i x_i^T \right\|_{S_{\infty}} = O \left( \frac{\sqrt{\log d} \left\|\mathbb{E}[\|X\|_2 X X^T]\right\|_{S_{\infty}}}{\sqrt{n}} \right).
$$

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

$$
\mathbb{E}_{\sigma} \left\| \sum_{i=1}^{n} \sigma_i x_i x_i^T \right\|_{S_{\infty}} \leq \left( \mathbb{E}_{\sigma} \left\| \sum_{i=1}^{n} \sigma_i x_i x_i^T \|_{S_{\infty}}^{\eta} \right\|_{S_{\infty}} \right)^{\frac{1}{\eta}} \leq 2^{-\frac{1}{4}} \sqrt{\frac{n}{e}} \left\| \left( \sum_{i=1}^{n} x_i x_i^T \right)^{\frac{1}{2}} \right\|_{S_{\infty}},
$$

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

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

and therefore

$$
\mathbb{E}_{\sigma} \left\| \sum_{i=1}^{n} \sigma_i x_i x_i^T \right\|_{S_{\infty}} \leq 2^{\frac{1}{2}} \sqrt{\pi} \log d \left\| \left( \sum_{i=1}^{n} x_i x_i^T \right)^{\frac{1}{2}} \right\|_{S_{\infty}} = 2^{\frac{1}{2}} \sqrt{\pi} \log d \left( \left\| \sum_{i=1}^{n} x_i x_i^T \right\|_{S_{\infty}} \right)^{\frac{1}{2}}.
$$

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

$$
\mathbb{E} \left\| \sum_{i=1}^{n} \sigma_i x_i x_i^T \right\|_{S_{\infty}} \leq 2^{\frac{1}{2}} \sqrt{\pi} \log d \left( \mathbb{E} \left\| \sum_{i=1}^{n} x_i x_i^T \right\|_{S_{\infty}} \right)^{\frac{1}{2}}
$$

**A.23**

It is clear

$$
\mathbb{E} \left[ \left\| \sum_{i=1}^{n} x_i x_i^T \right\|_{S_{\infty}} \right] \leq \mathbb{E} \left[ \left\| \sum_{i=1}^{n} x_i x_i^T - \mathbb{E} \sum_{i=1}^{n} x_i x_i^T \right\|_{S_{\infty}} \right] + \mathbb{E} \left\| \sum_{i=1}^{n} x_i x_i^T \right\|_{S_{\infty}}
$$

**A.24**

Then

$$
\mathbb{E} \left[ \left\| \sum_{i=1}^{n} Z_i^2 \right\|_{S_{\infty}} \right] = \sum_{i=1}^{n} \mathbb{E} \left( \left\| X_i \right\|_2^2 \right) - \sum_{i=1}^{n} \mathbb{E} \left[ \left\| X_i \right\|_2^2 \right] \left( \mathbb{E} \left[ \left\| X \right\|_2^2 X X^T \right] \right) \leq n \mathbb{E} \left[ \left\| X \right\|_2^2 \right] \mathbb{E} \left[ \left\| X \right\|_2^2 X X^T \right] \mathbb{E} \left[ \left\| X \right\|_2^2 X X^T \right] \leq n \mathbb{E} \left( \left\| X \right\|_2^2 \right).
and therefore
\[ \left\| \mathbb{E}\left[ \sum_{i=1}^{n} Z_i^2 \right] \right\|_{S_\infty} \leq n \left\| \mathbb{E}\left[ \|X\|_2^6 X X^\top \right] \right\|_{S_\infty}. \tag{A.25} \]

Furthermore,
\[ \|Z_i\|_{S_\infty} \leq \sup_{x_i} \|x_i x_i^\top x_i x_i^\top\|_{S_\infty} \leq \sup_{x_i} \|x_i\|_2^4. \tag{A.26} \]

We can apply Lemma A.16 with the above bound of variance (A.25) and magnitude (A.26), and derive
\[
\mathbb{E}\left[ \left\| \sum_{i=1}^{n} x_i x_i^\top x_i x_i^\top \right\|_{S_\infty} \right] \leq \sqrt{2n \mathbb{E}\left[ \|X\|_2^6 X X^\top \right]_{S_\infty} \log(2d)} + \frac{1}{3} \sup_{x_i} \|x_i\|_2^4 \log(2d).
\]

This together with the sub-additivity of the square-root function and (A.24) implies
\[
\left( \mathbb{E}\left[ \left\| \sum_{i=1}^{n} x_i x_i^\top x_i x_i^\top \right\|_{S_\infty} \right] \right)^{\frac{1}{2}} \leq \left( \mathbb{E}\left[ \left\| \sum_{i=1}^{n} x_i x_i^\top x_i x_i^\top - \mathbb{E}\sum_{i=1}^{n} x_i x_i^\top x_i x_i^\top \right\|_{S_\infty} \right] \right)^{\frac{1}{2}} + \left( n \mathbb{E}\left[ \|X\|_2^6 X X^\top \right]_{S_\infty} \right)^{\frac{1}{2}} \leq (2n \mathbb{E}\left[ \|X\|_2^6 X X^\top \right]_{S_\infty} \log(2d))^{\frac{1}{2}} + \frac{\sqrt{\log(2d)}}{\sqrt{3}} \sup_{x_i} \|x_i\|_2^4 \leq \sqrt{n \mathbb{E}\left[ \|X\|_2^6 X X^\top \right]_{S_\infty}}.
\]

We plug the above inequality back into (A.23), and get the inequality
\[
\frac{1}{n} \mathbb{E}\left[ \sum_{i=1}^{n} \sigma_i x_i x_i^\top \right]_{S_\infty} \leq 2^{\frac{1}{4}} \sqrt{n} \log d \left( (2 \log(2d))^{\frac{1}{2}} n^{-\frac{1}{4}} \|\mathbb{E}\left[ \|X\|_2^6 X X^\top \right]_{S_\infty} \|_{S_\infty} \right) + \frac{\sqrt{\log(2d)}}{\sqrt{3}} \sup_{x_i} \|x_i\|_2^4 \leq \frac{\mathbb{E}\left[ \|X\|_2^6 X X^\top \right]_{S_\infty}}{\sqrt{n}}. \tag{A.27}
\]

It is clear that
\[ \left\| \mathbb{E}\left[ \|X\|_2^6 X X^\top \right] \right\|_{S_\infty} \leq \sup_{x_i} \|x_i\|_2 \|\mathbb{E}\left[ \|X\|_2^6 X X^\top \right] \|_{S_\infty}^{\frac{1}{4}}. \]

This together with Cauchy-Schwartz inequality shows that
\[ (2 \log(2d))^{\frac{1}{2}} n^{-\frac{1}{4}} \|\mathbb{E}\left[ \|X\|_2^6 X X^\top \right]_{S_\infty} \|_{S_\infty} \leq \frac{\mathbb{E}\left[ \|X\|_2^6 X X^\top \right]_{S_\infty}}{\sqrt{n}} + \frac{\sqrt{\log(2d)}}{\sqrt{3}} \sup_{x_i} \|x_i\|_2^4. \]

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