Supplementary material: Margin based PU Learning

We give the complete proofs of Theorem 1 and 2 in Section 4. We first introduce the well-known concentration inequality, so the covariance estimator can be bounded. Then we analyze the convergence of PMPU.

Matrix Concentration Inequalities

Lemma 1. *(Matrix Bernstein's inequality) Consider a finite sequence $\{S_i\}$ of independent random matrices of dimension $d_1 \times d_2$. Assume that each matrix has uniformly bounded deviation from its mean:*

$$\|S_i - \mathbb{E}S_i\| \leq L \quad \text{for each index } i.$$

Introduce the random matrix $Z = \sum_i S_i$, and let $\nu(Z)$ be the matrix variance of Z where

$$\begin{aligned}\nu(Z) &= \max\{\|\mathbb{E}(Z-\mathbb{E}Z)(Z-\mathbb{E}Z)^\top)\|, \|\mathbb{E}(Z-\mathbb{E}Z)^\top(Z-\mathbb{E}Z)\|\} \\ &= \max\{\|\sum_i \mathbb{E}(S_i-\mathbb{E}S_i)(S_i-\mathbb{E}S_i)^\top)\|, \|\sum_i \mathbb{E}(S_i-\mathbb{E}S_i)^\top(S_i-\mathbb{E}S_i)\|\}.\end{aligned}$$

Then

$$\mathbb{E}\|Z-\mathbb{E}Z\| \leq \sqrt{2\nu(Z)\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2).$$

Furthermore, for all $t > 0$,

$$\mathbb{P}\{\|Z-\mathbb{E}Z\| \geq t\} \leq (d_1+d_2)\exp\left\{-\frac{t^2/2}{\nu(Z)+Lt/3}\right\}.$$

With matrix Bernstein's inequality, it is standard to get the concentration of covariance estimation:

Proposition 1. *Suppose $\{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^d$ are independent and identical distributed (i.i.d.) sub-gaussian random vectors and $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$, then with probability at least $1-\delta$,*

$$\|\frac{1}{N}XX^\top - I\|_2 \leq \epsilon$$

provided $N \geq C_\delta d\log(2d)/\epsilon^2$.

Lemma 2. *Let $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N] \in \mathbb{R}^{d\times N}$. Suppose each $\mathbf{x}_i$'s are independently sampled from the truncated Gaussian distribution with positive margin τ, then for $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\|_2 = 1$, we have*

$$\mathbb{E}\mathrm{sign}(\langle \mathbf{x}, \mathbf{w}\rangle)\mathbf{x} = \lambda_\tau \mathbf{w},$$

where $\lambda_\tau = \sqrt{\frac{2}{\pi}} + \frac{\exp(-\frac{\tau^2}{2})-1}{2}$.

Proof. It is well known that when $\mathbf{x}_i$ is the standard Gaussian random variable, $\lambda = \sqrt{\frac{2}{\pi}}$. In our setting, the 1st dimension of $\mathbf{x}$ is a truncated Gaussian, hence

$$\mathbb{E}\mathrm{sign}(\langle \mathbf{x}, \mathbf{w}\rangle)\mathbf{x} = \mathbb{E}|x_1| \cdot \mathbf{w} = \left(\sqrt{\frac{2}{\pi}} + \frac{\exp(-\frac{\tau^2}{2})-1}{2}\right)\mathbf{w}.$$

$\square$

Lemma 3. *Let $\mathbf{g} = [g_1, g_2, \cdots, g_d]^\top$, g_1 be a truncated Gaussian random variable, and the remaining $d-1$ dimensions are i.i.d. from standard Gaussian distribution. For two different vectors $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^d$, if $\arccos(\langle \mathbf{w}, \mathbf{w}'\rangle) \leq \frac{\pi}{2}$, we have*

$$\begin{aligned}\|\mathbb{E}\mathbf{g}\mathbf{g}^\top|\mathrm{sign}(\langle \mathbf{g}, \mathbf{w}\rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}'\rangle)|^2\|_2 &\leq C_1 d\left[\frac{1}{2} + e^{-\frac{\tau^2}{2}}\left(\frac{\tau}{\sqrt{2\pi}} + \frac{1}{2}\right)\right]\|\mathbf{w}-\mathbf{w}'\|_2 \\ \|\mathbb{E}\|\mathbf{g}\|_2^2|\mathrm{sign}(\langle \mathbf{g}, \mathbf{w}\rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}'\rangle)|^2\|_2 &\leq C_2 d\left[\frac{1}{2} + e^{-\frac{\tau^2}{2}}\left(\frac{\tau}{\sqrt{2\pi}} + \frac{1}{2}\right)\right]\|\mathbf{w}-\mathbf{w}'\|_2.\end{aligned}$$

Proof. Define $\alpha = \arccos(\langle \mathbf{w}, \mathbf{w}' \rangle)$ and $\mathbf{w}\|_2 = 1$, $\|\mathbf{w}'\|_2 = 1$. We will prove the two inequalities under the condition $\alpha \leq \frac{\pi}{2}$.
(a) Since

$$\begin{aligned}
&\|\mathbb{E}\mathbf{g}\mathbf{g}^\top |\mathrm{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle)|^2\|_2 \\
=&\left\| \mathbb{E} \begin{pmatrix} g_1^2 & g_1 g_2 & \cdots & g_1 g_d \\ g_2 g_1 & g_2^2 & \cdots & g_2 g_d \\ \vdots & \vdots & & \vdots \\ g_d g_1 & g_d g_2 & \cdots & g_d g_d \end{pmatrix} |\mathrm{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle)|^2 \right\|_2,
\end{aligned}$$

we need to estimate each $\mathbb{E}(g_i g_j |\mathrm{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle)|^2)$. Observe that only when $g_1 > 0 \wedge g_1 \cos\alpha + g_2 \sin\alpha < 0$ or $g_1 < 0 \wedge g_1 \cos\alpha + g_2 \sin\alpha > 0$, $|\mathrm{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle)|^2 = 4$. Otherwise it is 0. Hence, the domain of the expectation is

$$\Omega = \{(g_1, g_2) : g_1 > 0 \wedge g_1 \cos\alpha + g_2 \sin\alpha < 0\} \cup \{g_1 < 0 \wedge g_1 \cos\alpha + g_2 \sin\alpha > 0\}$$

with all other Gaussian variables $g_3, \cdots, g_d \in (-\infty, \infty)$.

For $i = j = 1$ ($i, j \in [d]$),

$$\mathbb{E}(g_1^2 |\mathrm{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle)|^2) = 1 + \frac{1}{\sqrt{2\pi}} \tau \exp\Big(-\frac{\tau^2}{2}\Big) - \frac{\mathrm{erf}(\frac{\tau}{\sqrt{2}})}{2}.$$

For $i = j = 2$,

$$\begin{aligned}
&\mathbb{E}(g_2^2 |\mathrm{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle)|^2) \\
=&4 \int_{(g_1, g_2) \in \Omega} g_2^2 \phi(g_1) \phi(g_2) dg_1 dg_2 \int_{g_3, \cdots, g_d} \phi(g_3) \cdots \phi(g_d) dg_3 \cdots dg_d \\
=&4 \int_{(g_1, g_2) \in \Omega} g_2^2 \phi(g_1) \phi(g_2) dg_1 dg_2 \\
=&8 \int_{\pi/2}^{\pi/2+\alpha} \int_0^\infty \sin^2(\theta) e^{-r^2} r^3 dr d\theta \quad \text{(by polar transformation)} \\
=&c_1(2\alpha + \sin\alpha),
\end{aligned}$$

since $\alpha < \frac{\pi}{2}$, we have

$$c_1(2\alpha + \sin\alpha) \leq 3c_1\alpha.$$

For $i = j \geq 3$, we have

$$\mathbb{E}(g_i^2 |\mathrm{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle)|^2) = \frac{4\alpha}{\pi}$$

For $i = 1, j = 3, \cdots, d$ or $j = 1, i = 3, \cdots, d$, we get

$$\begin{aligned}
&\mathbb{E}(g_i g_j |\mathrm{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle)|^2) \\
=&\quad 4\mathbb{E}g_1 \mathbb{E}g_2 \int_{g_3, \cdots, g_d} \phi(g_3) \cdots \phi(g_d) dg_3 \cdots dg_d \\
=&\quad 4 * \sqrt{\frac{2}{\pi}} \Big(\sqrt{\frac{2}{\pi}} + \frac{\exp(-\tau^2/2 - 1)}{2}\Big) \\
=&\quad \frac{8}{\pi} + \frac{4\exp(-\tau^2/2) - 4}{\sqrt{2\pi}}.
\end{aligned}$$

For all the other cases that $i \neq j$, we can see that

$$\mathbb{E}(g_i g_j |\mathrm{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \mathrm{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle)|^2) = 0.$$

Therefore,

$$
\begin{aligned}
&\|\mathbb{E}\mathbf{g}\mathbf{g}^\top|\mathrm{sign}(\langle\mathbf{g},\mathbf{w}\rangle)-\mathrm{sign}(\langle\mathbf{g},\mathbf{w}'\rangle)|^2\|_2\\
=&\left\|\begin{pmatrix}
1+\frac{\tau\exp\left(-\frac{\tau^2}{2}\right)}{\sqrt{2\pi}}-\frac{\mathrm{erf}(\frac{\tau}{\sqrt{2}})}{2} & \frac{8}{\pi}+\frac{4\exp(-\tau^2/2)-4}{\sqrt{2\pi}} & \frac{8}{\pi}+\frac{4\exp(-\tau^2/2)-4}{\sqrt{2\pi}} & \cdots & \frac{8}{\pi}+\frac{4\exp(-\tau^2/2)-4}{\sqrt{2\pi}}\\
\frac{8}{\pi}+\frac{4\exp(-\tau^2/2)-4}{\sqrt{2\pi}} & c_1(2\alpha+\sin\alpha) & 0 & \cdots & 0\\
\frac{8}{\pi}+\frac{4\exp(-\tau^2/2)-4}{\sqrt{2\pi}} & 0 & \frac{4\alpha}{\pi} & \cdots & 0\\
\vdots & \vdots & \vdots & & \vdots\\
\frac{8}{\pi}+\frac{4\exp(-\tau^2/2)-4}{\sqrt{2\pi}} & 0 & 0 & \cdots & \frac{4\alpha}{\pi}
\end{pmatrix}\right\|_2\\
\le&\ \max\left\{\frac{1}{2}+\exp(-\frac{\tau^2}{2})(\frac{\tau}{\sqrt{2\pi}}+\frac{1}{2})+(d-1)\Big(\frac{8}{\pi}+\frac{4\exp(-\tau^2/2)-4}{\sqrt{2\pi}}\Big),\right.\\
&\left.3c_1\alpha+\frac{8}{\pi}+\frac{4\exp(-\tau^2/2)-4}{\sqrt{2\pi}},\frac{4\alpha}{\pi}+\frac{8}{\pi}+\frac{4\exp(-\tau^2/2)-4}{\sqrt{2\pi}}\right\}\\
\le&\ C_1 d\exp\big(\frac{-\tau^2}{2}\big)\Big(\frac{4+\tau}{\sqrt{2\pi}}+\frac{1}{2}\Big)\|\mathbf{w}-\mathbf{w}'\|_2,
\end{aligned}
$$

in which the first inequality holds because $\|A\|_2\le\sqrt{\|A\|_1\cdot\|A\|_\infty}$.

(b): The proof is similar to that of **(a)**. We have

$$
\begin{aligned}
&\|\mathbb{E}\|\mathbf{g}\|_2^2|\mathrm{sign}(\langle\mathbf{g},\mathbf{w}\rangle)-\mathrm{sign}(\langle\mathbf{g},\mathbf{w}'\rangle)|^2\|_2\\
=&\sum_i g_i^2|\mathrm{sign}(\langle\mathbf{g},\mathbf{w}\rangle)-\mathrm{sign}(\langle\mathbf{g},\mathbf{w}'\rangle)|^2\\
\le&\frac{1}{2}+\exp\big(-\frac{\tau^2}{2}\big)\Big(\frac{\tau}{\sqrt{2\pi}}+\frac{1}{2}\Big)+3c_1\alpha+(d-2)\frac{4\alpha}{\pi}\\
\le&\ C_2 d\exp\big(\frac{-\tau^2}{2}\big)\Big(\frac{4+\tau}{\sqrt{2\pi}}+\frac{1}{2}\Big)\|\mathbf{w}-\mathbf{w}'\|_2
\end{aligned}
$$

□

Proof of Theorem 1

Proof. According to the rotation invariance of the Euclidean space, there exists a rotation matrix Q^* such that $Q^*\mathbf{w}^*=[1,0,\cdots,0]$. Without loss of generality, we assume that $\mathbf{w}^*=[1,0,\cdots,0]\in\mathbb{R}^d$. For simplicity, we will discard the superscript t in $X^{(t)}$ but the reader should aware that the feature matrix X is always re-sampled in each iteration. Let $\mathbf{x}=[x_1,\bar{\mathbf{x}}_2]$ where x_1 denotes the 1st dimension of $\mathbf{x}$ and $\bar{\mathbf{x}}_2$ denotes the remaining $d-1$ dimension. Similarly, we denote $\mathbf{w}^{(0)}=[w_1^{(0)},\mathbf{w}_2^{(0)}]$. Denote by $\Delta\mathbf{y}^{(1)}=\mathbf{y}^{(1)}-\hat{\mathbf{y}}^{(1)}$ the initial error. Since at the t-th iteration,

$$
\begin{aligned}
\mathbf{w}^{(t)}&=\mathbf{w}^{(t-1)}-\frac{1}{\lambda_\tau m_t}X\Delta\mathbf{y}^{(t-1)}\\
&=\mathbf{w}^{(t-1)}-\frac{1}{\lambda_\tau m_t}X(\mathbf{y}^{(t-1)}-\hat{\mathbf{y}}^{(t-1)}),
\end{aligned}
$$

we have

$$
\begin{aligned}
\mathbf{w}^{(t)}-\mathbf{w}^*&=(\mathbf{w}^{(t-1)}-\mathbf{w}^*)-\frac{1}{\lambda_\tau m_t}X(\mathrm{sign}(X^\top\mathbf{w}^{(t-1)})-\mathrm{sign}(X^\top\mathbf{w}^*))+\mathrm{sign}(X^\top\mathbf{w}^*)-\mathcal{S}_\tau(X^\top\mathbf{w}^{(t-1)}))\\
&=(\mathbf{w}^{(t-1)}-\mathbf{w}^*)-\frac{1}{\lambda_\tau m_t}X(\mathrm{sign}(X^\top\mathbf{w}^{(t-1)})-\mathrm{sign}(X^\top\mathbf{w}^*))+\frac{1}{\lambda_\tau m_t}X\Delta_t
\end{aligned}
$$

where $\Delta_t=\mathrm{sign}(X^\top\mathbf{w}^*)-\mathcal{S}_\tau(X^\top\mathbf{w}^{(t-1)})$.

To bound the first two terms, using Lemma 2 and Lemma 3, we have with probability at least $1-\delta$,

$$
\begin{aligned}
&\|(\mathbf{w}^{(t-1)}-\mathbf{w}^*)-\frac{1}{\lambda_\tau m_t}X(\mathrm{sign}(X^\top\mathbf{w}^{(t-1)})-\mathrm{sign}(X^\top\mathbf{w}^*))\|_2\\
\le&\epsilon\max(\|\mathbf{w}^{(t-1)}-\mathbf{w}^*\|_2,\|\mathbf{w}^{(t-1)}-\mathbf{w}^*\|_2^{1/2})\ .
\end{aligned}
$$

provided $m_t \geq O(d \log d \exp(-\tau^2/2)/\epsilon^2)$.

As we assume m_0 is sufficiently large, it is easy to satisfy that $\|\mathbf{w}^{(t-1)} - \mathbf{w}^*\|_2 \leq 1$. Then

$$\begin{aligned}&\|(\mathbf{w}^{(t-1)} - \mathbf{w}^*) - \frac{1}{\lambda_\tau m_t} X(\text{sign}(X^\top \mathbf{w}^{(t-1)}) - \text{sign}(X^\top \mathbf{w}^*))\|_2 \\ \leq& \epsilon \ .\end{aligned}$$

Next let us first consider Δ_1. It's clear that with probability at least $1-\delta$,

$$\|\frac{1}{m_t} X \Delta_1\|_2 \leq C_\delta \sqrt{\frac{d \log(d)}{m_t}} + \|\mathbb{E}[\text{sign}(\mathbf{x}^\top \mathbf{w}^*) - \mathcal{S}_\tau(x^\top \mathbf{w}^{(0)})]\|_2 \ .$$

The estimation of Δ_1 includes two cases, i.e. E_+ and E_- where E_+ is the error on $\{\mathbf{x}^\top \mathbf{w} \leq \eta_0 \wedge z > \tau\}$ and E_- is the error on $\{\mathbf{x}^\top \mathbf{w} > \eta_0 \wedge z < 0\}$, where $z = \mathbf{x}^\top \mathbf{w}^*$. Denote the cumulative distribution function of standard Gaussian distribution by

$$\Phi(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt \ .$$

We obtain

$$\begin{aligned}
E_+ &= \int_\tau^\infty \mathbb{P}\Big(x_1 w_1 + \bar{\mathbf{x}}_2^\top \mathbf{w}_2 < \eta_0, x_1 = \alpha\Big) d\alpha \\
&= \int_\tau^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{\alpha^2}{2}} \Phi\Big(\frac{\eta_0 - \alpha w_1}{\|\mathbf{w}_2\|_2}\Big) d\alpha \\
&= \int_\tau^\infty \frac{1}{2\sqrt{2\pi}} e^{-\frac{\alpha^2}{2}} \Big[1 + \text{erf}\Big(\frac{\eta_0 - \alpha w_1}{\|\mathbf{w}_2\|}\Big)\Big] d\alpha \\
&= \int_\tau^\infty \frac{1}{2\sqrt{2\pi}} e^{-\frac{\alpha^2}{2}} \text{erfc}\Big(\frac{\alpha w_1 - \eta_0}{\sqrt{2}\|\mathbf{w}_2\|}\Big) d\alpha \\
&\leq \int_\tau^\infty \frac{1}{2\sqrt{2\pi}} e^{-\frac{\alpha^2}{2}} e^{-\Big(\frac{\alpha w_1 - \eta_0}{\sqrt{2}\|\mathbf{w}_2\|}\Big)^2} d\alpha \\
&= \frac{\|\mathbf{w}_2\| e^{-\frac{\eta_0^2}{2(w_1^2 + \|\mathbf{w}_2\|^2)}}}{4\sqrt{w_1^2 + \|\mathbf{w}_2\|^2}} \text{erfc}\left(\frac{(w_1^2 + \|\mathbf{w}_2\|^2)\tau - w_1 \eta_0}{\|\mathbf{w}_2\| \sqrt{2(w_1^2 + \|\mathbf{w}_2\|^2)}}\right) \\
&= \frac{\|\mathbf{w}_2\| e^{-\frac{\eta_0^2}{2}}}{4} \text{erfc}\Big(\frac{\tau - w_1 \eta_0}{\sqrt{2}\|\mathbf{w}_2\|}\Big)
\end{aligned}$$

where $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2} dx$ denotes the error function and $\text{erfc}(z) = 1 - \text{erf}(z)$ is the complementary error function. The 4-th equality holds because cumulative function

$$\Phi(z) = \frac{1}{2}\Big(1 + \text{erf}\Big(\frac{z}{\sqrt{2}}\Big)\Big).$$

Similarly, we have

$$\begin{aligned}
E_- &= \int_{-\infty}^0 \mathbb{P}\Big(x_1 w_1 + \bar{\mathbf{x}}_2^\top \mathbf{w}_2 \geq \eta_0, x_1 = \beta\Big) d\beta \\
&= \int_{-\infty}^0 \frac{1}{2\sqrt{2\pi}} e^{-\frac{\beta^2}{2}} \text{erfc}\Big(\frac{\eta_0 - \beta w_1}{\sqrt{2}\|\mathbf{w}_2\|}\Big) d\beta \\
&\leq \int_\tau^\infty \frac{1}{2\sqrt{2\pi}} e^{-\frac{\beta^2}{2}} e^{-\Big(\frac{\eta_0 - \beta w_1}{\sqrt{2}\|\mathbf{w}_2\|}\Big)^2} d\beta \\
&= \frac{\|\mathbf{w}_2\| e^{-\frac{\eta_0^2}{2(w_1^2 + \|\mathbf{w}_2\|^2)}}}{4\sqrt{w_1^2 + \|\mathbf{w}_2\|^2}} \text{erfc}\left(\frac{w_1 \eta_0}{\|\mathbf{w}_2\| \sqrt{2(w_1^2 + \|\mathbf{w}_2\|^2)}}\right) \\
&= \frac{\|\mathbf{w}_2\| e^{-\frac{\eta_0^2}{2}}}{4} \text{erfc}\left(\frac{w_1 \eta_0}{\sqrt{2}\|\mathbf{w}_2\|}\right),
\end{aligned}$$

Then,

$$
\begin{aligned}
\|\mathbb{E}[\text{sign}(\mathbf{x}^\top \mathbf{w}^*) - \mathcal{S}_\tau(x^\top \mathbf{w}^{(0)})]\|_2 &= E_+ + E_- \\
&= \frac{\|\mathbf{w}_2\| e^{-\frac{\eta_0^2}{2}}}{4}\Big[\text{erfc}\Big(\frac{\tau - w_1\eta_0}{\sqrt{2}\|\mathbf{w}_2\|}\Big) + \text{erfc}\Big(\frac{w_1\eta_0}{\sqrt{2}\|\mathbf{w}_2\|}\Big)\Big]. \\
&\overset{(a)}{\leq} \frac{\|\mathbf{w}_2\|}{4}\Big[\exp\Big(\frac{-\tau^2 + 2\tau w_1\eta_0 - w_1^2\eta_0^2}{2\|\mathbf{w}_2\|^2}\Big) + \exp\Big(\frac{-w_1^2\eta_0^2}{2\|\mathbf{w}_2\|^2}\Big)\Big] \qquad (1) \\
&\overset{(b)}{\leq} \hat{c}_1[\exp(-c_2\tau^2) + \bar{\delta}_{m_0}]\|\mathbf{w}^{(0)} - \mathbf{w}^*\|_2 \\
&\leq c_1 \exp(-c_2\tau^2)\|\mathbf{w}^{(0)} - \mathbf{w}^*\|_2 \; .
\end{aligned}
$$

(b) is a simplification of (a). The constant $\hat{c}$ and c_2 actually depend on τ and many other factors. However once we fixed the parameters, they will be constants and do not control the order of our bound. $\bar{\delta}_{m_0}$ is a small number if m_0 is large because when m_0 is large $w_1 \approx 1$ and $\mathbf{w}_2 \approx 0$. As we always assume m_0 is sufficiently large, $\bar{\delta}_{m_0} < 0.1$ due to the exponential decaying.

Similarly the upper bound of the error at the t-th step is

$$\|\mathbb{E}[\text{sign}(\mathbf{x}^\top \mathbf{w}^*) - \mathcal{S}_\tau(x^\top \mathbf{w}^{(t)})]\|_2 \leq c_1 \exp(-c_2\tau^2)\|\mathbf{w}^{(t-1)} - \mathbf{w}^*\|_2$$

Combine everything above, we have with probability at least $1-\delta$,

$$\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \leq \epsilon + C_\delta\sqrt{\frac{d\log(d)}{m_t}} + c_1\exp(-c_2\tau^2)\|\mathbf{w}^{(t-1)} - \mathbf{w}^*\|_2 \; . \qquad (2)$$

As m_t is sampled on unlabeled dataset, it can be as large as we want. Therefore the above inequality can be simplified when m_t is sufficiently large, that is, $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \leq c_1\exp(-c_2\tau^2)\|\mathbf{w}^{(t-1)} - \mathbf{w}^*\|_2$

□

Proof of Theorem 2

Proof. Let

$$B_i = \frac{1}{\lambda_\tau}[\mathbf{x}_i\text{sign}(\langle \mathbf{x}_i, \mathbf{w}\rangle) - \mathbf{x}_i\text{sign}(\langle \mathbf{x}_i, \mathbf{w}'\rangle)],$$

then by Lemma 2, we have

$$\mathbb{E}B_i = \mathbf{w} - \mathbf{w}'.$$

Further, we set

$$Z_i = B_i - \mathbb{E}B_i,$$

where

$$\sum_{i=1}^m B_i = \frac{1}{\lambda_\tau}[X\text{sign}(\langle X, \mathbf{w}\rangle) - X\text{sign}(\langle X, \mathbf{w}'\rangle)]$$

In order to utilize matrix Bernstein inequality, we need to bound the terms $\max_i \|Z_i\|_2$, $\|\mathbb{E}Z_i^\top Z_i\|$ and $\|\mathbb{E}Z_iZ_i^\top\|_2$ respectively.

For the first term, we have

$$
\begin{aligned}
&\max_i \|Z_i\|_2 \\
=&\max_i \|B_i - \mathbb{E}B_i\|_2 \\
\leq&\max_i(\|B_i\|_2 + \|\mathbb{E}B_i\|_2) \\
\leq&\max_i \frac{1}{\lambda_\tau}\|\mathbf{x}_i\text{sign}(\langle \mathbf{x}_i, \mathbf{w}\rangle) - \mathbf{x}_i\text{sign}(\langle \mathbf{x}_i, \mathbf{w}'\rangle)\|_2 + \|\mathbf{w} - \mathbf{w}'\|_2 \\
\leq&\frac{2\sqrt{d}}{\lambda_\tau} + \|\mathbf{w} - \mathbf{w}'\|_2.
\end{aligned}
$$

When $\mathbf{w} - \mathbf{w}'$ is sufficient small, then

$$\frac{2\sqrt{d}}{\lambda_\tau} + \|\mathbf{w} - \mathbf{w}'\|_2 \leq c\frac{2\sqrt{d}}{\lambda_\tau}.$$

For the second term, we get

$$\begin{aligned}
&\|\mathbb{E}Z_i^\top Z_i\|_2\\
=&\|\mathbb{E}(B_i-\mathbb{E}B_i)^\top(B_i-\mathbb{E}B_i)\|_2\\
=&\|\mathbb{E}B_i^\top B_i - B_i^\top\cdot\mathbb{E}B_i-\mathbb{E}B_i^\top\cdot B_i+\mathbb{E}B_i^\top\mathbb{E}B_i\|_2\\
\leq&\|\mathbb{E}B_i^\top B_i\|_2+\|\mathbb{E}B_i^\top\mathbb{E}B_i\|_2.
\end{aligned}$$

Since

$$\|\mathbb{E}B_i^\top\mathbb{E}B_i\|_2=\|\mathbf{w}-\mathbf{w}'\|_2^2,$$

and

$$\begin{aligned}
&\|\mathbb{E}B_i^\top B_i\|_2\\
=&\frac{1}{\lambda_\tau^2}\|\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\top|\text{sign}(\langle\mathbf{x}_i,\mathbf{w}\rangle)-\text{sign}(\langle\mathbf{x}_i,\mathbf{w}'\rangle)|^2\|_2.\\
\leq&\frac{C_2 d}{\lambda_\tau^2}\|\mathbf{w}-\mathbf{w}'\|_2
\end{aligned}$$

Thus, we have

$$\|\mathbb{E}Z_i^\top Z_i\|_2\leq\frac{C_2 d}{\lambda_\tau^2}\|\mathbf{w}-\mathbf{w}'\|_2+\|\mathbf{w}-\mathbf{w}'\|_2^2.$$

Note that if $\|\mathbf{w}-\mathbf{w}'\|_2<1$, then $\|\mathbf{w}-\mathbf{w}'\|_2>\|\mathbf{w}-\mathbf{w}'\|_2^2$, and $\|\mathbf{w}-\mathbf{w}'\|_2\geq 1$, then $\|\mathbf{w}-\mathbf{w}'\|_2\leq\|\mathbf{w}-\mathbf{w}'\|_2^2$. Hence, the above inequality can be rewritten as

$$\|\mathbb{E}Z_i^\top Z_i\|_2\leq\frac{C_2 d}{\lambda_\tau^2}\max\{\|\mathbf{w}-\mathbf{w}'\|_2,\|\mathbf{w}-\mathbf{w}'\|_2^2\}$$

For the third term, we have

$$\begin{aligned}
&\|\mathbb{E}Z_iZ_i^\top\|_2\\
=&\|\mathbb{E}(B_i-\mathbb{E}B_i)(B_i-\mathbb{E}B_i)^\top\|_2\\
\leq&\|\mathbb{E}B_iB_i^\top\|_2+\|\mathbb{E}B_i\mathbb{E}B_i^\top\|_2.
\end{aligned}$$

Since

$$\|\mathbb{E}B_i\mathbb{E}B_i^\top\|_2=\|\mathbf{w}-\mathbf{w}'\|_2^2,$$

and

$$\begin{aligned}
&\|\mathbb{E}B_iB_i^\top\|_2\\
=&\frac{1}{\lambda_\tau^2}\|\mathbb{E}\|\mathbf{x}_i\|_2^2|\text{sign}(\langle\mathbf{x}_i,w\rangle)-\text{sign}(\langle\mathbf{x}_i,w'\rangle)|^2\|_2\\
\leq&\frac{C_1}{\lambda_\tau^2}\|\mathbf{w}-\mathbf{w}'\|_2,\quad(\text{by Lemma 3})
\end{aligned}$$

Then, we derive

$$\|\mathbb{E}Z_iZ_i^\top\|_2\leq\frac{C_1}{\lambda_\tau^2}\|\mathbf{w}-\mathbf{w}'\|_2+\|\mathbf{w}-\mathbf{w}'\|_2^2,$$

which can be rewritten as

$$\|\mathbb{E}Z_i^\top Z_i\|_2\leq\frac{C_1}{\lambda_\tau^2}\max\{\|\mathbf{w}-\mathbf{w}'\|_2,\|\mathbf{w}-\mathbf{w}'\|_2^2\}.$$

Now we can apply matrix Bernstein inequality to obtain the final result. □