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

\[
\nu(Z) = \max\{\|\mathbb{E}(Z - \mathbb{E}Z)(Z - \mathbb{E}Z)^\top\|, \|\mathbb{E}(Z - \mathbb{E}Z)^\top (Z - \mathbb{E}Z)\|\}
\]

\[
= \max\{\| \sum_i \mathbb{E}(S_i - \mathbb{E}S_i)(S_i - \mathbb{E}S_i)^\top\|, \| \sum_i \mathbb{E}(S_i - \mathbb{E}S_i)^\top (S_i - \mathbb{E}S_i)\|\}.
\]

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 \( t > 0 \),

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

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

Proposition 1. Suppose \( \{x_i\}_{i=1}^N \subseteq \mathbb{R}^d \) are independent and identical distributed (i.i.d.) sub-gaussian random vectors and \( X = [x_1, x_2, \ldots, x_N] \), then with probability at least \( 1 - \delta \),

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

provided \( N \geq C_d d \log(2d)/\epsilon^2 \).

Lemma 2. Let \( X = [x_1, x_2, \ldots, x_N] \subseteq \mathbb{R}^{d \times N} \). Suppose each \( x_i \)'s are independently sampled from the truncated Gaussian distribution with positive margin \( \tau \), then for \( w \in \mathbb{R}^d \) with \( \|w\|_2 = 1 \), we have

\[
\mathbb{E}\text{sign}(\langle x, w \rangle)x = \lambda_\tau w,
\]

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

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

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

\[\square\]

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

\[
\|\mathbb{E}gg^\top \text{sign}(\langle g, w \rangle) - \text{sign}(\langle g, w' \rangle)\|_2 \leq C_1d\left[\frac{1}{2} + e^{-\frac{\tau^2}{2}}\left(\frac{\tau}{\sqrt{2\pi}} + \frac{1}{2}\right)\right]\|w - w'\|_2
\]

\[
\|\mathbb{E}\|g\|_2^2 \text{sign}(\langle g, w \rangle) - \text{sign}(\langle g, w' \rangle)\|^2\|_2 \leq C_2d\left[\frac{1}{2} + e^{-\frac{\tau^2}{2}}\left(\frac{\tau}{\sqrt{2\pi}} + \frac{1}{2}\right)\right]\|w - w'\|_2.
\]

\[\square\]
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

\[
\|E_{\mathbf{g}} [\mathbf{g}]^\top | \text{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \text{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle) |^2 \|_2
\]

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

\[
\Omega = \{ (g_1, g_2) : g_1 > 0 \land g_1 \cos \alpha + g_2 \sin \alpha < 0 \} \cup \{ g_1 < 0 \land g_1 \cos \alpha + g_2 \sin \alpha > 0 \}
\]

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

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

\[
E(g_1^2 | \text{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \text{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle) |^2) = 1 + \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{\tau^2}{2} \right) - \frac{\text{erf}(\frac{\tau}{\sqrt{2}})}{2}.
\]

For \( i = j = 2 \),

\[
E(g_1^2 g_2 | \text{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \text{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle) |^2)
\]

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

\[
c_1 (2\alpha + \sin \alpha) \leq 3c_1 \alpha.
\]

For \( i = j \geq 3 \), we have

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

For \( i = 1, j = 3, \ldots, d \) or \( j = 1, i = 3, \ldots, d \), we get

\[
E(g_1 g_j | \text{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \text{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle) |^2)
\]

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

\[
E(g_i g_j | \text{sign}(\langle \mathbf{g}, \mathbf{w} \rangle) - \text{sign}(\langle \mathbf{g}, \mathbf{w}' \rangle) |^2) = 0.
\]
Therefore,
\[
\|Eg^T \text{sign}((g, w)) - \text{sign}((g, w'))\|^2_2
\]
\[
= \left( \begin{array}{cccc}
1 + \frac{\tau \exp\left(-\frac{\tau^2}{2}\right)}{\sqrt{2\pi}} - \frac{\text{erf}(\frac{\tau}{\sqrt{2}})}{2} & \frac{8}{\pi} + \frac{4 \exp(-\tau^2/2) - 4}{\sqrt{2\pi}} & \frac{8}{\pi} + \frac{4 \exp(-\tau^2/2) - 4}{\sqrt{2\pi}} & \cdots & \frac{8}{\pi} + \frac{4 \exp(-\tau^2/2) - 4}{\sqrt{2\pi}} \\
\frac{8}{\pi} + \frac{4 \exp(-\tau^2/2) - 4}{\sqrt{2\pi}} & c_1(2\alpha + \sin \alpha) & 0 & \cdots & 0 \\
\frac{8}{\pi} + \frac{4 \exp(-\tau^2/2) - 4}{\sqrt{2\pi}} & 0 & \frac{4\alpha}{\pi} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{8}{\pi} + \frac{4 \exp(-\tau^2/2) - 4}{\sqrt{2\pi}} & 0 & 0 & \cdots & \frac{4\alpha}{\pi}
\end{array} \right) \leq \max \left\{ \frac{1}{2} + \exp\left(-\frac{\tau^2}{2}\right) \left( \frac{\tau}{\sqrt{2\pi}} + \frac{1}{2} \right) + (d - 1) \left( \frac{8}{\pi} + \frac{4 \exp(-\tau^2/2) - 4}{\sqrt{2\pi}} \right) \right\}
\]
\[\leq C_1 d \exp\left(\frac{-\tau^2}{2}\right) \left( \frac{4 + \tau}{\sqrt{2\pi}} + \frac{1}{2} \right) \|w - w'\|_2,
\]
in which the first inequality holds because \(\|A\|_2 \leq \sqrt{\|A\|_1 \cdot \|A\|_\infty}\).

(b): The proof is similar to that of (a). We have
\[
\|E\|_2 \text{sign}((g, w)) - \text{sign}((g, w'))\|^2_2
\]
\[
= \sum_i g_i^2 \text{sign}((g, w)) - \text{sign}((g, w'))^2
\]
\[\leq \frac{1}{2} + \exp\left(-\frac{\tau^2}{2}\right) \left( \frac{\tau}{\sqrt{2\pi}} + \frac{1}{2} \right) + 3c_1\alpha + (d - 2) \frac{4\alpha}{\pi}
\]
\[\leq C_2 d \exp\left(\frac{-\tau^2}{2}\right) \left( \frac{4 + \tau}{\sqrt{2\pi}} + \frac{1}{2} \right) \|w - w'\|_2.
\]

\(\square\)

**Proof of Theorem 1**

*Proof.\* According to the rotation invariance of the Euclidean space, there exists a rotation matrix \(Q^*\) such that \(Q^* w^* = [1, 0, \ldots, 0]\). Without loss of generality, we assume that \(w^* = [1, 0, \ldots, 0] \in \mathbb{R}^d\). 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 \(x = [x_1, \bar{x}_d]\) where \(x_1\) denotes the 1st dimension of \(x\) and \(\bar{x}_d\) denotes the remaining \(d - 1\) dimension. Similarly, we denote \(w^{(t)} = [w_1^{(t)}, w_2^{(t)}]\). Denote by \(\Delta y^{(1)} = y^{(1)} - \hat{y}^{(1)}\) the initial error. Since at the \(t\)-th iteration,
\[
w^{(t)} = w^{(t-1)} - \frac{1}{\lambda_r m_t} X \Delta y^{(t-1)}
\]
\[
= w^{(t-1)} - \frac{1}{\lambda_r m_t} X (y^{(t-1)} - \hat{y}^{(t-1)}),
\]
we have
\[
w^{(t)} - w^* = (w^{(t-1)} - w^*) - \frac{1}{\lambda_r m_t} X (\text{sign}(X^T w^{(t-1)}) - \text{sign}(X^T w^*) + \text{sign}(X^T w^*) - S_r(X^T w^{(t-1)}))
\]
\[
= (w^{(t-1)} - w^*) - \frac{1}{\lambda_r m_t} X (\text{sign}(X^T w^{(t-1)}) - \text{sign}(X^T w^*) + \frac{1}{\lambda_r m_t} X \Delta_t)
\]
where \(\Delta_t = \text{sign}(X^T w^*) - S_r(X^T w^{(t-1)})\).

To bound the first two terms, using Lemma 2 and Lemma 3, we have with probability at least \(1 - \delta\),
\[
\|(w^{(t-1)} - w^*) - \frac{1}{\lambda_r m_t} X (\text{sign}(X^T w^{(t-1)}) - \text{sign}(X^T w^*))\|_2
\]
\[\leq \epsilon \max(\|w^{(t-1)} - w^*\|_2, \|w^{(t-1)} - w^*\|_2^{1/2})
\].
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 \( \| w^{(t-1)} - w^* \|_2 \leq 1 \). Then

\[
\| (w^{(t-1)} - w^*) - \frac{1}{\lambda_t m_t} X (\text{sign}(X^T w^{(t-1)}) - \text{sign}(X^T w^*)) \|_2 
\leq \epsilon .
\]

Next let us first consider \( \Delta_1 \). It’s clear that with probability at least \( 1 - \delta \),

\[
\| \frac{1}{m_t} X \Delta \|_2 \leq C \sqrt{\frac{d \log(d)}{m_t}} + \| \mathbb{E}[\text{sign}(x^T w^*) - S_t(x^T w^{(0)})] \|_2 .
\]

The estimation of \( \Delta_1 \) includes two cases, i.e. \( E_+ \) and \( E_- \) where \( E_+ \) is the error on \( \{ x^T w \leq \eta_0 \wedge z > \tau \} \) and \( E_- \) is the error on \( \{ x^T w > \eta_0 \wedge z < 0 \} \), where \( z = x^T w^* \). Denote the cumulative distribution function of standard Gaussian distribution by

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

We obtain

\[
E_+ = \int_{\tau}^{\infty} \mathbb{P}(x_1 w_1 + x_2^T w_2 < \eta_0, x_1 = \alpha) d\alpha
\]

\[
= \int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \Phi\left( \frac{\eta_0 - \alpha w_1}{\|w_2\|} \right) d\alpha
\]

\[
= \int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \left[ 1 + \text{erf}\left( \frac{\eta_0 - \alpha w_1}{\|w_2\|} \right) \right] d\alpha
\]

\[
= \int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \text{erfc}\left( \frac{\eta_0 - \alpha w_1}{\sqrt{2}\|w_2\|} \right) d\alpha
\]

\[
\leq \int_{\tau}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} e^{-\left( \frac{\eta_0 - \alpha w_1}{\|w_2\|} \right)^2} d\alpha
\]

\[
= \|w_2\| e^{-\frac{\tau^2}{4\|w_2\|^2}} \text{erfc}\left( \frac{\|w_2\|^2}{\|w_2\|^2} \right)
\]

\[
= \|w_2\| e^{-\frac{\tau^2}{4\|w_2\|^2}} \text{erfc}\left( \frac{\|w_2\|^2}{\|w_2\|^2} \right)
\]

where \( \text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-x^2} dx \) 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

\[
\Phi(z) = \frac{1}{2} \left( 1 + \text{erf}\left( \frac{z}{\sqrt{2}} \right) \right).
\]

Similarly, we have

\[
E_- = \int_{-\infty}^{0} \mathbb{P}(x_1 w_1 + x_2^T w_2 \geq \eta_0, x_1 = \beta) d\beta
\]

\[
= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \text{erfc}\left( \frac{\eta_0 - \beta w_1}{\sqrt{2}\|w_2\|} \right) d\beta
\]

\[
\leq \int_{0}^{\tau} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} e^{-\left( \frac{\eta_0 - \beta w_1}{\|w_2\|} \right)^2} d\beta
\]

\[
= \|w_2\| e^{-\frac{\tau^2}{4\|w_2\|^2}} \text{erfc}\left( \frac{\|w_2\|^2}{\|w_2\|^2} \right)
\]

\[
= \|w_2\| e^{-\frac{\tau^2}{4\|w_2\|^2}} \text{erfc}\left( \frac{\|w_2\|^2}{\|w_2\|^2} \right)
\]
Then,
\[ \| \mathbb{E}[\text{sign}(x^Tw^*) - S_r(x^T w^{(0)})]\|_2 = E_+ + E_- \]
\[ = \frac{\|w_2\| e^{-\frac{\beta^2}{2}}}{4} \left[ \text{erfc}(\frac{\tau w_1 \eta_0}{\sqrt{2\|w_2\|}}) + \text{erfc}(\frac{w_1 \eta_0}{\sqrt{2\|w_2\|}}) \right]. \]
\[ \leq \frac{\|w_2\|}{4} \left[ \exp \left( -\frac{\tau^2 + 2\tau w_1 \eta_0 - w_2^2 \eta_0^2}{2\|w_2\|^2} \right) + \exp \left( \frac{-w_1^2 \eta_0^2}{2\|w_2\|^2} \right) \right] \tag{1} \]
\[ \leq \tilde{c}_1 \exp(-c_2 \tau^2) \|w^{(0)} - w^*\|_2 \]
\[ \leq c_1 \exp(-c_2 \tau^2) \|w^{(0)} - w^*\|_2 . \]

(b) is a simplification of (a). The constant \( \tilde{c}_1 \) and \( c_2 \) actually depend on \( \tau \) and many other factors. However once we fixed the parameters, they will be constants and do not control the order of our bound. \( \delta_{m_0} \) is a small number if \( m_0 \) is large because when \( m_0 \) is large \( w_1 \approx 1 \) and \( w_2 \approx 0 \). As we always assume \( m_0 \) is sufficiently large, \( \delta_{m_0} < 0.1 \) due to the exponential decaying.

Similarly the upper bound of the error at the \( t \)-th step is
\[ \| \mathbb{E}[\text{sign}(x^Tw^*) - S_r(x^T w^{(t)})]\|_2 \leq c_1 \exp(-c_2 \tau^2) \|w^{(t-1)} - w^*\|_2 \]
Combine everything above, we have with probability at least \( 1 - \delta \),
\[ \|w^{(t)} - w^*\|_2 \leq \epsilon + C_\delta \sqrt{\frac{\log(d)}{m_t}} + c_1 \exp(-c_2 \tau^2) \|w^{(t-1)} - w^*\|_2 . \tag{2} \]
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, \( \|w^{(t)} - w^*\|_2 \leq c_1 \exp(-c_2 \tau^2) \|w^{(t-1)} - w^*\|_2 \)

\[ \square \]

**Proof of Theorem 2**

**Proof.** Let
\[ B_i = \frac{1}{\lambda_r} [x_i \text{sign}(\langle x_i, w \rangle) - x_i \text{sign}(\langle x_i, w' \rangle)], \]
then by Lemma 2, we have
\[ \mathbb{E}B_i = w - w'. \]
Further, we set
\[ Z_i = B_i - \mathbb{E}B_i, \]
where
\[ \sum_{i=1}^m B_i = \frac{1}{\lambda_r} [X \text{sign}(\langle X, w \rangle) - X \text{sign}(\langle X, w' \rangle)] \]
In order to utilize matrix Bernstein inequality, we need to bound the terms \( \max_i \|Z_i\|_2, \|\mathbb{E}Z_i^T Z_i\|_2 \) and \( \|\mathbb{E}Z_i^T \|_2 \) respectively. For the first term, we have
\[ \max_i \|Z_i\|_2 \]
\[ = \max_i \|B_i - \mathbb{E}B_i\|_2 \]
\[ \leq \max_i (\|B_i\|_2 + \|\mathbb{E}B_i\|_2) \]
\[ \leq \max_i \frac{1}{\lambda_r} \|x_i \text{sign}(\langle x_i, w \rangle) - x_i \text{sign}(\langle x_i, w' \rangle)\|_2 + \|w - w'\|_2 \]
\[ \leq \frac{2\sqrt{d}}{\lambda_r} + \|w - w'\|_2 . \]
When \( w - w' \) is sufficient small, then
\[ \frac{2\sqrt{d}}{\lambda_r} + \|w - w'\|_2 < \epsilon \frac{2\sqrt{d}}{\lambda_r} . \]
For the second term, we get
\[
\|E Z_i^\top Z_i\|_2 \\
= \|E(B_i - EB_i)(B_i - EB_i)\|_2 \\
\leq \|EB_i^\top B_i\|_2 + \|EB_i^\top EB_i\|_2.
\]
Since
\[
\|EB_i^\top EB_i\|_2 = \|w - w'\|_2^2,
\]
and
\[
\|EB_i^\top B_i\|_2 \\
= \frac{1}{\lambda_f^2} \|E[x_i x_i^\top | \text{sign}(\langle x_i, w \rangle) - \text{sign}(\langle x_i, w' \rangle)]^2\|_2 \\
\leq C_2d \lambda_f^2 \|w - w'\|_2.
\]
Thus, we have
\[
\|E Z_i^\top Z_i\|_2 \leq C_2d \lambda_f^2 \|w - w'\|_2 + \|w - w'\|_2^2.
\]
Note that if \( \|w - w'\|_2 < 1 \), then \( \|w - w'\|_2 > \|w - w'\|_2^2 \), and \( \|w - w'\|_2 \geq 1 \), then \( \|w - w'\|_2 \leq \|w - w'\|_2^2 \). Hence, the above inequality can be rewritten as
\[
\|E Z_i^\top Z_i\|_2 \leq \frac{C_2d}{\lambda_f^2} \max\{\|w - w'\|_2, \|w - w'\|_2^2\}
\]
For the third term, we have
\[
\|EZ_i Z_i^\top\|_2 \\
\leq \|E(B_i - EB_i)(B_i - EB_i)^\top\|_2 \\
\leq \|EB_i^\top B_i\|_2 + \|EB_i^\top EB_i\|_2.
\]
Since
\[
\|EB_i^\top EB_i\|_2 = \|w - w'\|_2^2,
\]
and
\[
\|EB_i^\top B_i\|_2 \\
= \frac{1}{\lambda_f^2} \|E|x_i|_2^2 | \text{sign}(\langle x_i, w \rangle) - \text{sign}(\langle x_i, w' \rangle)]^2\|_2 \\
\leq \frac{C_1}{\lambda_f^2} \|w - w'\|_2, \quad \text{(by Lemma 3)}
\]
Then, we derive
\[
\|EZ_i Z_i^\top\|_2 \leq \frac{C_1}{\lambda_f^2} \|w - w'\|_2 + \|w - w'\|_2^2,
\]
which can be rewritten as
\[
\|EZ_i Z_i^\top\|_2 \leq \frac{C_1}{\lambda_f^2} \max\{\|w - w'\|_2, \|w - w'\|_2^2\}.
\]
Now we can apply matrix Bernstein inequality to obtain the final result.