Appendix

A Omitted Proofs

A.1 Notation

We begin with notations. For a random variable X, its sub-gaussian norm/Orlicz norm is defined as $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[e^{X^2/t^2}] \leq 1\}$. For a d-dimensional random vector Y, the sub-gaussian norm of Y is defined as $\|Y\|_{\psi_2} = \sup_{v\in S^{d-1}} \|\langle Y, v\rangle\|$, where S^{d-1} denotes the sphere of a unit ball in $\mathbb{R}^d$. For two sequences of positive numbers a_n and b_n, $a_n \lesssim b_n$ means that for some constant $c > 0$, $a_n \leqslant cb_n$ for all n, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Further, we use the notion o_p and O_p, where for a sequence of random variables X_n, $X_n = o_p(a_n)$ means $X_n/a_n \to 0$ in probability, $X_n = O_p(b_n)$ means that for any $\varepsilon > 0$, there is a constant K, such that $\mathbb{P}(|X_n| \leq K \cdot b_n) \geq 1 - \varepsilon$, and $X_n = \Omega_p(b_n)$ means that for any $\varepsilon > 0$, there is a constant K, such that $\mathbb{P}(|X_n| \geq K \cdot b_n) \geq 1 - \varepsilon$. Finally, we use $c_0, c_1, c_2, C_1, C_2, \ldots$ to denote generic positive constants that may vary from place to place.

Besides, let $L = L_1L_2$, then $\phi(\cdot)$ is L_1L_2-Lipchitz in ℓ_2-norm.

A.2 Proof of Theorem 1

We firstly consider to prove a bound for

$$\mathbb{P}\Big(\big\|\frac{1}{n}\sum_{i=1}^n \phi(z_i) - \mathbb{E}\phi(z)\big\| \geqslant t\Big),$$

where $z_i \sim \mathcal{N}(\mu, \sigma^2 I)$.

Lemma 1. *There exists a universal constant c, such that*

$$\mathbb{P}\Big(\big\|\frac{1}{n}\sum_{i=1}^n \phi(z_i) - \mathbb{E}\phi(z)\big\| \geqslant c\sigma\big(\frac{\sqrt{d}\tilde{L}}{\sqrt{n/2}} + L\sqrt{\frac{2\log(2/\delta)}{n}}\big)\Big) \leqslant \delta.$$

From the above inequality, we can immediately obtain

$$\mathbb{P}\Big(\big\|\frac{1}{n}\sum_{i=1}^n \phi(z_i)\big\| \geqslant \big\|\mathbb{E}\phi(z)\big\| + c\sigma\big(\frac{\sqrt{d}\tilde{L}}{\sqrt{n/2}} + L\sqrt{\frac{2\log(2/\delta)}{n}}\big)\Big) \leqslant \delta.$$

Remark 5. *Note that the concentration bound still holds for $y\phi(yz)$ and $\phi(yz)$ by simply applying conditional probability.*

Proof. Let $\vartheta_u(z) = \big\langle \phi(z) - \mathbb{E}\phi(z), u\big\rangle$

$$\begin{aligned}
\mathbb{P}\Big(\big\|\frac{1}{n}\sum_{i=1}^n \phi(z_i) - \mathbb{E}\phi(z)\big\| \geqslant t\Big) &= \mathbb{P}\Big(\sup_{\|u\|=1} \big\langle \frac{1}{n}\sum_{i=1}^n \phi(z_i) - \mathbb{E}\phi(z), u\big\rangle \geqslant t\Big) \\
&= \mathbb{P}\Big(\sup_{\|u\|=1} \frac{1}{n}\sum_{i=1}^n \vartheta_u(z_i) \geqslant t\Big) \\
&\leqslant \mathbb{P}\Big(\sup_{u,u'\in\mathbb{B}(0,1)} \big|\frac{1}{n}\sum_{i=1}^n \vartheta_u(z_i) - \sum_{i=1}^n \vartheta_{u'}(z_i)\big| \geqslant t\Big)
\end{aligned}$$

Let us use chaining and Orlicz-processes to obtain a bound. We prove $\{\frac{1}{n}\sum_{i=1}^n \vartheta_u(z_i), u \in \mathbb{B}(0,1)\}$ is a ϑ_2-process with respect to a rescaled distance $\|\cdot\|/\lambda$ for some $\lambda > 0$. If so, we will have

$$\mathbb{E}\exp\Big(\frac{\lambda^2|\frac{1}{n}\sum_{i=1}^n \vartheta_u(z_i) - \frac{1}{n}\sum_{i=1}^n \vartheta_{u'}(z_i)|^2}{\|u - u'\|^2}\Big) \leqslant 2.$$

The LHS

$$\mathbb{E}\exp\Big(\frac{\lambda^2|\frac{1}{n}\sum_{i=1}^n\vartheta_u(z_i)-\frac{1}{n}\sum_{i=1}^n\vartheta_{u'}(z_i)|^2}{\|u-u'\|^2}\Big)\leqslant\int_0^\infty e^t\mathbb{P}\Big(\frac{|\frac{1}{n}\sum_{i=1}^n\vartheta_u(z_i)-\frac{1}{n}\sum_{i=1}^n\vartheta_{u'}(z_i)|}{\|u-u'\|}\geqslant\frac{\sqrt{t}}{\lambda}\Big)dt$$
$$\leqslant\int_0^\infty e^t 2\exp\big(-\frac{tn}{\lambda^2\sigma^2L^2}\big)dt.$$

As long as

$$\frac{tn}{\lambda^2\sigma^2L^2}\geqslant 2,\quad\text{i.e.}\quad\lambda\leqslant\frac{\sqrt{n/2}}{\sigma L},$$

we would obtain the Dudley entropy integral as

$$J(D)=\frac{1}{\lambda}\int_0^1\sqrt{\log(1+\exp(d\log\frac{1}{\delta}))}d\delta\approx const\cdot\frac{\sqrt{d}}{\lambda}.$$

We let $\lambda=\sqrt{n/2}/(\sigma L)$, it gives us $J(D)=(\sigma\tilde{L})/\sqrt{n/2}$, where $\tilde{L}=const\cdot L$. □

Next, let us consider bounding

$$\mathbb{P}\big(\hat{w}^\top(\phi(z)-\hat{b})\leqslant 0\big)=\mathbb{P}\Big(\frac{\hat{w}^\top}{\|\hat{w}\|}(\phi(z)-\hat{b})\leqslant 0\Big).\quad\text{(notice }\hat{w}=0\text{ is of zero probability)}$$

We further denote $\nu_w=\big[\mathbb{E}\phi(z_i^+)-\mathbb{E}\phi(z_i^-)\big]/2$, $\nu_b=\big[\mathbb{E}\phi(z_i^+)+\mathbb{E}\phi(z_i^-)\big]/2$ and

$$\gamma=c\sigma\big(\frac{\sqrt{d}\tilde{L}}{\sqrt{n/2}}+L\sqrt{\frac{2\log(2/\delta)}{n}}\big).$$

From Lemma 1, we can obtain

$$\mathbb{P}\left(\big|\|\hat{w}-\nu_w\|\big|\geqslant\gamma\right)\leqslant\delta,$$
$$\mathbb{P}\left(\|\hat{b}-\nu_b\|\geqslant\gamma\right)\leqslant\delta.$$

Notice that for any unit vector v, $v^\top(\phi(z)-\hat{b})$ is a σL-Lipschitz function of $(z^\top,z_1^\top,\cdots,z_n^\top)^\top\sim\mathcal{N}(0,I_{(n+1)m})$, by standard concentration, we have the following lemma.

Lemma 2. *For any $t>0$ and unit vector v*

$$\mathbb{P}\Big(|v^\top(\phi(z)-\hat{b})-v^\top\nu_w|\geqslant t\Big)\leqslant 2\exp(-\frac{t^2}{2\sigma^2L^2}).$$

Next, we provide a bound for $\langle\hat{w},\nu_w\rangle$.

Lemma 3. *For any $t>0$*

$$\mathbb{P}\big(|\langle\hat{w},\nu_w\rangle-\|\nu_w\|^2|\geqslant t\big)\leqslant 2\exp(\frac{-nt^2}{2\sigma^2L^2\|\nu_w\|^2})$$

Taking $\delta=2\exp(\frac{-nt^2}{2\sigma^2L^2\|\nu_w\|^2})$, we have

$$\mathbb{P}\big(|\langle\hat{w},\nu_w\rangle-\|\nu_w\|^2|\geqslant\sigma L\|\nu_w\|\sqrt{\frac{2\log(2/\delta)}{n}}\big)\leqslant\delta$$

Proof. LHS is equivalent to

$$\mathbb{P}\big(|\langle\hat{w}-\nu_w,\nu_w\rangle|\geqslant t\big).$$

Besides, we have $\langle\hat{w}-\nu_w,\nu_w/\|\nu_w\|\rangle=\frac{1}{n}\sum_{i=1}^n\langle y_i\phi(y_iz_i)-\nu_w,\nu_w/\|\nu_w\|\rangle$ is a sum of sub-gaussian variables with constant σL, then by sub-gaussian tail bound we have

$$\mathbb{P}\big(|\langle\hat{w}-\nu_w,\nu_w\rangle|\geqslant t\big)\leqslant 2\exp(\frac{-nt^2}{2\sigma^2L^2\|\nu_w\|^2}).$$

□

[Proof of Theorem 1] If we denote $E = A \cup B$, where $A = \left\{|\langle \hat{w}, \nu_w \rangle - \|\nu_w\|^2| \leqslant \sigma L \|\nu_w\| \sqrt{\frac{2\log(2/\delta_1)}{n}}\right\}$, $B = \{|\|\hat{w} - \nu_w\|\big| \leqslant \gamma\}$, then with probability $\mathbb{P}(E)$, we have for $t > 0$

$$\mathbb{P}\Big(\frac{\hat{w}^\top}{\|\hat{w}\|}(\phi(z) - \hat{b}) \leqslant 0\Big|E\Big) \leqslant \mathbb{P}\Big(\frac{\hat{w}^\top \nu_w}{\|\nu_w\| + \gamma} \leqslant t\Big|E\Big) + 2\exp(-\frac{t^2}{2\sigma^2 L^2}).$$

As long as we choose γ and t such that

$$\frac{\hat{w}^\top \nu_w}{\|\nu_w\| + \gamma} > t$$

we have

$$\mathbb{P}\Big(\frac{\hat{w}^\top}{\|\hat{w}\|}(\phi(z) - \hat{b}) \leqslant 0\Big|E\Big) \leqslant 2\exp(-\frac{t^2}{2\sigma^2 L^2}).$$

We take

$$\delta = 2\exp(-(\tilde{L}/L)^2 d), \quad \delta_1 = 2\exp(-\frac{d}{8\sigma^2 L^2})$$

then

$$\gamma = 2\sqrt{2}c\sigma\tilde{L}\sqrt{\frac{d}{n}}.$$

As a result, we obtain

$$|\langle \hat{w}, \nu_w \rangle - \|\nu_w\|^2| \leqslant \frac{d}{2\sqrt{n}}, \quad |\|\hat{w} - \nu_w\|\big| \leqslant \gamma$$

with probability at least $1 - \delta_1 - \delta$.

We can choose

$$t = \frac{\sqrt{d}(\sqrt{n} - 1/2)}{\sqrt{n} + 2\sqrt{2}c\sigma\tilde{L}},$$

so that

$$\mathbb{P}\Big(\frac{\hat{w}^\top}{\|\hat{w}\|}(\phi(z) - \hat{b}) \leqslant 0\Big|E\Big) \leqslant 2\exp\left(\frac{-d(\sqrt{n} - 1/2)^2}{2\sigma^2 L^2(\sqrt{n} + 2\sqrt{2}c\sigma\tilde{L})^2}\right).$$

Thusly,

$$\mathbb{P}\left(\beta(\hat{w}, \hat{b}) \geqslant 2\exp\left(\frac{-d(\sqrt{n} - 1/2)^2}{2\sigma^2 L^2(\sqrt{n} + 2\sqrt{2}c\sigma\tilde{L})^2}\right)\right) \leqslant \mathbb{P}(E^c),$$

which gives us the final result stated in the theorem.

A.3 Proof of Theorem 2

Since we know ϑ is L_2'-Lipchitz continuous in ℓ_∞-norm. Then, we know

$$\mathbb{P}_{(x,y)\sim\mathcal{P}}[\exists u \in \mathbb{B}_p(x, \varepsilon) : \langle \hat{w}, y \cdot (\vartheta(u) - \hat{b})\rangle \leqslant 0] \geqslant \mathbb{P}_{(zy,y)\sim\mathcal{P}'}[\exists u \in \mathbb{B}_p(yz, \varepsilon/L_2') : \langle \hat{w}, y \cdot (\phi(u) - \hat{b})\rangle \leqslant 0]$$

since the pre-image of $bB_p(x, \varepsilon)$ via ϑ includes the set $\mathbb{B}_p(yz, \varepsilon/L_2')$. Then following the argument in Schmidt et al. (2018), the result follows.

Remark 6. *As a side interest, we also provide an analysis to show the lower bound result in Theorem 3.2 is achievable up to a logarithm factor, by purely using labeled data. This scale matches the result in Schmidt et al. (2018), but under a more general model considered in our paper.*

$$\begin{aligned}
\mathbb{P}_{(x,y)\sim\mathcal{P}}[\exists u \in \mathbb{B}_p(x, \varepsilon) : f_{\hat{w},\hat{b}}(u) \neq y] &= \mathbb{P}_{(x,y)\sim\mathcal{P}}[\exists u \in \mathbb{B}_p(x, \varepsilon) : \langle \hat{w}, y \cdot (\vartheta(u) - \hat{b})\rangle \leqslant 0] \\
&= \mathbb{P}_{(x,y)\sim\mathcal{P}}[\exists \eta \in \mathbb{B}_p(0, \varepsilon) : \langle \hat{w}, y \cdot (\vartheta(x + \eta) - \hat{b})\rangle \leqslant 0] \\
&\leqslant \mathbb{P}_{(x,y)\sim\mathcal{P}}[\langle \hat{w}, (\phi(z) - \hat{b})\rangle + \min_{\eta\in\mathbb{B}_p(0,\varepsilon)} \langle \eta L_1, \hat{w}\rangle \leqslant 0] \\
&= \mathbb{P}_{(x,y)\sim\mathcal{P}}[\langle \hat{w}, (\phi(z) - \hat{b})\rangle \leqslant \varepsilon L_1 \|\hat{w}\|_q]
\end{aligned}$$

where $1/p + 1/q = 1$.

When $p=\infty$, $q=1$, it leads to $\|\hat{w}\|_1/\|\hat{w}\| \leqslant \sqrt{d}$. Recall in Theorem 1, $E = A \cup B$, where $A = \Big\{|\langle \hat{w}, \nu_w\rangle - \|\nu_w\|^2| \leqslant \sigma L \|\nu_w\| \sqrt{\frac{2\log(2/\delta_1)}{n}}\Big\}$, $B = \{|\|\hat{w}-\nu_w\|\big| \leqslant \gamma\}$, then with probability $\mathbb{P}(E)$, we have

$$\begin{aligned}\mathbb{P}\Big(\frac{\hat{w}^\top}{\|\hat{w}\|}(\phi(z)-\hat{b}) \leqslant \varepsilon L_1 \frac{\|\hat{w}\|_q}{\|\hat{w}\|}\Big|E\Big) &\leqslant \mathbb{P}\Big(\frac{\hat{w}^\top \nu_w}{\|\nu_w\|+\gamma} \leqslant t + \varepsilon L_1 \frac{\|\hat{w}\|_q}{\|\hat{w}\|}\Big|E\Big) + 2\exp(-\frac{t^2}{2\sigma^2 L^2}) \\ &\leqslant \mathbb{P}\Big(\frac{\hat{w}^\top \nu_w}{\|\nu_w\|+\gamma} \leqslant t + \varepsilon L_1 \sqrt{d}\Big|E\Big) + 2\exp(-\frac{t^2}{2\sigma^2 L^2}).\end{aligned}$$

We still choose

$$\delta = 2\exp(-(\tilde{L}/L)^2 d), \quad \delta_1 = 2\exp(-\frac{d}{8\sigma^2 L^2})$$

such that

$$\gamma = 2\sqrt{2}c\sigma\tilde{L}\sqrt{\frac{d}{n}}.$$

As a result, we obtain

$$|\langle \hat{w}, \nu_w\rangle - \|\nu_w\|^2| \leqslant \frac{d}{2\sqrt{n}}, \quad |\|\hat{w}-\nu_w\|\big| \leqslant 2\sqrt{2}c\sigma\tilde{L}\sqrt{\frac{d}{n}}$$

with probability at least $1-\delta_1-\delta$. We choose t such that

$$\frac{\hat{w}^\top \nu_w}{\|\nu_w\|+\gamma} > t + \varepsilon L_1\sqrt{d}.$$

We let

$$t = \frac{(\sqrt{d})^2 - d/(2\sqrt{n})}{\sqrt{d}/2 + 2\sqrt{2}c\sigma\tilde{L}\sqrt{\frac{d}{n}}} - \varepsilon L_1\sqrt{d} = \frac{\sqrt{d}(\sqrt{n}-1/2)}{\sqrt{n}/2 + 2\sqrt{2}c\sigma\tilde{L}} - \varepsilon L_1\sqrt{d}.$$

As long as

$$\varepsilon \leqslant (\frac{\sqrt{n}-1/2}{\sqrt{n}/2 + 2\sqrt{2}c\sigma\tilde{L}} - \frac{\sigma L\sqrt{2\log(1/\beta)}}{\sqrt{d}})/L_1,$$

we have

$$\beta^{\mathcal{R}}(\hat{w},\hat{b}) \leqslant 2\exp(-\frac{t^2}{2\sigma^2 L^2}) \leqslant \beta$$

A.4 Statistical Measures

Recall the definition of

$$d_\nu = \max\Big\{\frac{\|\tilde{\mu}_1-\mu_1\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}, \frac{\|\tilde{\mu}_2-\mu_2\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}\Big\}.$$

We now make connections to commonly used statistical measures and provide a sketch of proof.

(a). *Wasserstein Distance*: the Wasserstein Distance induced by metric ρ between distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ over $\mathbb{R}^d$ is defined as

$$W_\rho(\mathcal{P}_1,\mathcal{P}_2) = \sup_{\|f\|_{\text{Lip}}\leqslant 1}[\int f d\mathcal{P}_1 - f d\mathcal{P}_2],$$

where $\|f\|_{\text{Lip}} \leqslant 1$ indicates the class of $f: \mathbb{R}^d \mapsto \mathbb{R}$ such that for any $x, x' \in \mathbb{R}^d$, $|f(x)-f(x)| \leqslant \rho(x,x')$. Let us consider $\rho(x,x') = \|x-x'\|$.

Proposition 3. *Suppose* $\max\{W_\rho(\mathcal{P}_1,\tilde{\mathcal{P}}_1), W_\rho(\mathcal{P}_2,\tilde{\mathcal{P}}_2)\} \leqslant \tau$, *for* $\tau \geqslant 0$, *then we have* $\|\mu_i - \tilde{\mu_i}\| \leqslant \tau, \quad i=1,2$. *As a result,*

$$d_\nu \leqslant \frac{\tau}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}.$$

If we further have $\tau \leqslant \|\mu_1-\mu_2\|/2$, *we have* $d_\nu \leqslant \tau/(\|\mu_1-\mu_2\|-2\tau)$.

Proof. Notice $f(x)=x$ also satisfies $\|f\|_{Lip} \leqslant 1$, then we know $\|\mu_i-\tilde{\mu_i}\| \leqslant \tau, \quad i=1,2$. If we further have $\tau \leqslant \|\mu_1-\mu_2\|/2$, plugging into the denominator, the result follows. □

(b). *Maximal Information*: Maximal Information between distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ over $\mathbb{R}^d$ is defined as

$$MI(\mathcal{P}_1, \mathcal{P}_2) = \sup_{O \subseteq \mathbb{R}^d} \max \left\{ \frac{\mathbb{P}_{x\sim\mathcal{P}_1}(x \in O)}{\mathbb{P}_{x\sim\mathcal{P}_2}(x \in O)}, \frac{\mathbb{P}_{x\sim\mathcal{P}_2}(x \in O)}{\mathbb{P}_{x\sim\mathcal{P}_1}(x \in O)} \right\}.$$

Proposition 4. *Suppose* $\max\{MI(\mathcal{P}_1, \tilde{\mathcal{P}}_1), MI(\mathcal{P}_2, \tilde{\mathcal{P}}_2)\} \leqslant \tau$ *for* $1 \leqslant \tau \leqslant 1 + \|\mu_1 - \mu_2\|/(2\|\mu_1\| + 2\|\mu_2\|)$*, then we have* $\|\mu_i - \tilde{\mu}_i\| \leqslant (\tau - 1)\|\mu_i\|,\ i = 1, 2$*. As a result, we have*

$$d_\nu \leqslant \frac{(\tau - 1)\max\{\|\mu_1\|, \|\mu_2\|\}}{\|\mu_1 - \mu_2\| - 2(\tau - 1)(\|\mu_1\| + \|\mu_2\|)}.$$

As we can see, as $\tau \to 1$, $d_\nu \to 0$.

Proof. Let $X_1 \sim \mathcal{P}_1$, $X_2 \sim \mathcal{P}_2$. By the definition of Maximal Information,

$$\sup_{x \in \mathbb{R}^d} \max \left\{ \frac{\mathbb{P}(X_1 = x)}{\mathbb{P}(X_2 = x)}, \frac{\mathbb{P}(X_2 = x)}{\mathbb{P}(X_1 = x)} \right\} \leqslant \tau.$$

Then, we know $\|\mu_i - \tilde{\mu}_i\| \leqslant (\tau - 1)\|\mu_i\|,\ i = 1, 2$ once we notice for all corresponding entries of the vector of X_1 and X_2, their maximal information is bounded by τ. So,

$$d_\nu \leqslant \frac{(\tau - 1)\max\{\|\mu_1\|, \|\mu_2\|\}}{\|\tilde{\mu}_1 - \tilde{\mu}_2\|}.$$

If we further have $\tau \leqslant 1 + \|\mu_1 - \mu_2\|/(2\|\mu_1\| + 2\|\mu_2\|)$, plugging into the denominator, the result follows. □

(c). $\mathcal{H}$-Divergence: let $\mathcal{H}$ be a class of binary classifiers, then $\mathcal{H}$-divergence between distributions $\mathcal{P}$ and $\mathcal{P}'$ over $\mathbb{R}^d$ is defined as

$$D_{\mathcal{H}}(\mathcal{P}, \mathcal{P}') = \sup_{h \in \mathcal{H}} |\mathbb{P}_{x\sim\mathcal{P}}(h(x) = 1) - \mathbb{P}_{x\sim\mathcal{P}'}(h(x) = 1)|.$$

To illustrate the connection between Theorem 3 and $\mathcal{H}$-divergence, we consider a specific hypothesis class

$$\mathcal{H} = \left\{ h | h(t) = sgn(w^\top (t - b)), (w, b) \in \mathbb{R}^d \times \mathbb{R}^d \right\}. \tag{3}$$

Proposition 5. *Suppose for* $X_i \sim \mathcal{P}_i$ *and* $\tilde{X}_i \sim \tilde{\mathcal{P}}_i$ $i = 1, 2$*, the sub-gaussian norm of* $\|X_i - \mu_i\|_{\psi_2}$ *and* $\|\tilde{X}_i - \tilde{\mu}_i\|_{\psi_2}$ *are bounded by* σ *and* $\tilde{\sigma}$*, where* $X_i \sim \mathcal{P}_i$*,* $\tilde{X}_i \sim \tilde{\mathcal{P}}_i$ *and* μ_i*,* $\tilde{\mu}_i$ *are the corresponding means. Let* $\alpha = \zeta\sqrt{\log(4/(1 - \tau))}$*, where* $\zeta = \max\{\sigma, \tilde{\sigma}\}$*, if* $\max\{D_{\mathcal{H}}(\mathcal{P}_1, \tilde{\mathcal{P}}_1), D_{\mathcal{H}}(\mathcal{P}_2, \tilde{\mathcal{P}}_2)\} \leqslant \tau$*, for* $\tau \leqslant 1$*, we have* $\|\mu_i - \tilde{\mu}_i\| \leqslant \alpha$*,* $i = 1, 2$*. As a result,*

$$d_\nu \leqslant \frac{\alpha}{\|\tilde{\mu}_1 - \tilde{\mu}_2\|}.$$

If we further have $\tau \leqslant 1 - 4\exp(-\|\mu_1 - \mu_2\|^2/4\zeta^2)$*, then* $d_\nu \leqslant \alpha/(\|\mu_1 - \mu_2\| - 2\alpha)$*.*

Proof. It follows a simple geometric argument – a hyperplane cannot distinguish the two distributions too well. Recall if $\|X_i - \mu_i\|_{\psi_2} \leqslant \sigma_i$ and $\|\tilde{X}_i - \tilde{\mu}_i\|_{\psi_2} \leqslant \tilde{\sigma}_i$, then for $i = 1, 2$

$$\mathbb{P}(\|X_i - \mu_i\| \geqslant t) \leqslant 2\exp(-\frac{t^2}{\sigma^2}), \quad \mathbb{P}(\|\tilde{X}_i - \tilde{\mu}_i\| \geqslant t) \leqslant 2\exp(-\frac{t^2}{\tilde{\sigma}^2})$$

Consider t^* such that

$$2\exp(-\frac{t^{*2}}{\zeta^2}) = (1 - \tau)/2, \quad i.e.\ t^* = \alpha/2.$$

It is easy to see the distance $\|(\mu_1 - \mu_2)/2 - (\tilde{\mu}_1 - \tilde{\mu}_2)/2\|$ should be upper bounded by $2t^*$, otherwise, there exists a hyperplane such that the probability mass of $X_1 \sim \mathcal{P}_1$ and $\tilde{X}_1 \sim \tilde{\mathcal{P}}_1$ has high probability mass on difference side of the hyperplane.

□

As we can see in the case for Wasserstein Distance, as $\tau \to 0$, $d_\nu \to 0$. However, for $\mathcal{H}$-Divergence when $\tau \to 0$, d_ν will not go to 0. That is due to the constraint of capacity of $\mathcal{H}$. Even if $\tau = 0$, $\mathcal{P}_i$ and $\tilde{\mathcal{P}}_i$ can still be quite different.

A.5 Proof of Theorem 3 and Proposition 1

Let us recall the statement of Theorem 3 with some specified constants.

Theorem 7 (Robust accuracy)**.** *Consider the Gaussian generative model, where the marginal distribution of the input x of labeled domain is a uniform mixture of two distributions with mean $\mu_1 = \mathbb{E}[\phi(z)]$ and $\mu_2 = \mathbb{E}[\phi(-z)]$ respectively, where $z \sim \mathcal{N}(0, \sigma^2 I_{s_1})$. Suppose the marginal distribution of the input of unlabeled domain is a mixture of two sub-gaussian distributions with mean $\tilde{\mu}_1$ and $\tilde{\mu}_2$ with mixing probabilities q and $1-q$ and $\left\|\mathbb{E}\left[\vartheta(\tilde{x}_i) - \mathbb{E}[\vartheta(\tilde{x}_i)] \mid a^T\vartheta(\tilde{x}_i) = b\right]\right\| \le c_\varepsilon \cdot (\sqrt{d} + |b|)$ for fixed unit vector a. Assuming the sub-gaussian norm for both labeled and unlabeled data are upper bounded by a universal quantity $\sigma_{\max} := C_\sigma d^{1/4}$, $c_q < q < 1 - c_q$, $\|\tilde{\mu}_1 - \tilde{\mu}_2\|_2 = C_\mu\sqrt{d}$, for some constants $C_\sigma > 0, 0 < c_q < 1/2$, $C_\mu > 0$ sufficiently large, and*

$$d_\nu = \max\left\{\frac{\|\tilde{\mu}_1 - \mu_1\|}{\|\tilde{\mu}_1 - \tilde{\mu}_2\|}, \frac{\|\tilde{\mu}_2 - \mu_2\|}{\|\tilde{\mu}_1 - \tilde{\mu}_2\|}\right\} < c_0,$$

for some constant $c_0 \le 1/4$, then the robust classification error is at most 1% when d is sufficiently large, $n \ge C$ for some constant C (not depending on d and ε) and

$$\tilde{n} \gtrsim \varepsilon^2 \log d\sqrt{d}.$$

Now let us proceed to the proof.

For simplicity of presentation. We first denote the distributions for the two classes of labeled data as $subGaussian(\mu_1, \sigma^2_{\max})$, and $subGaussian(\mu_2, \sigma^2_{\max})$ respectively. Similarly, we also denote the distributions for the two classes of unlabeled data as $subGaussian(\tilde{\mu}_1, \sigma^2_{\max})$, and $subGaussian(\tilde{\mu}_2, \sigma^2_{\max})$ respectively. Also, to avoid the visual similarity and emphasize the estimates constructed by the labeled and unlabeled data respectively, we write $\hat{w}$ as $\hat{w}_{\text{intermediate}}$, $\hat{b}$ as $\hat{b}_{\text{intermediate}}$, $\tilde{w}$ as $\tilde{w}_{\text{final}}$ and $\tilde{b}$ as $\tilde{b}_{\text{final}}$.

Then, let us write out the robust error of misclassifying class 1 against the ℓ_∞ attack (the robust error of misclassifying class 2 can be bounded similarly) as

$$\begin{aligned}
&\max_{\|\delta\|_\infty \le L_1'\varepsilon} \mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\tilde{x} + \delta - \tilde{b}_{\text{final}}) \leqslant 0 \mid \tilde{x} \sim subGaussian(\tilde{\mu}_1, \sigma^2_{\max})\Big)\\
=&\mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\mu_1 - \tilde{b}_{\text{final}}) + \varepsilon\frac{\|\hat{\theta}_{\text{final}}\|_1}{\|\hat{\theta}_{\text{final}}\|}\Big)\\
=&\mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\mu_1 - \tilde{b}_{\text{final}}) + L_1' \cdot \varepsilon\sqrt{d}\Big)
\end{aligned}$$

Denote $\tilde{b}_{\text{final}} := \frac{\tilde{\mu}_1 + \tilde{\mu}_2}{2} + d_b$, we then have

$$\begin{aligned}
&\mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\mu_1 - \tilde{b}_{\text{final}}) + L_1' \cdot \varepsilon\sqrt{d}\Big)\\
=&\mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}((\mu_1 - \tilde{\mu}_1) + \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{2} - d_b) + L_1' \cdot \varepsilon\sqrt{d}\Big)
\end{aligned}$$

We are going to bound $|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\mu_1 - \tilde{\mu}_1)|$, $|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\tilde{\mu}_1 - \tilde{\mu}_2)/2|$, and $|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}d_b|$ respectively.

Let $\tilde{\mu} = (\tilde{\mu}_1 - \tilde{\mu}_2)/2$, $\tilde{\nu} = (\tilde{\mu}_1 + \tilde{\mu}_2)/2$, $\mu = (\mu_1 - \mu_2)/2$, $\nu = (\mu_1 + \mu_2)/2$, and b_i be the indicator that the ith pseudo-label $\tilde{y}_i$ is incorrect, so that $\tilde{x}_i \sim \tilde{\nu} + subGaussian\left((1 - 2b_i)\,\tilde{y}_i\tilde{\mu}, \sigma^2\right)$

Let $\tilde{n}_1 = \sum_{i=1}^{\tilde{n}} 1\{\tilde{y}_i = 1\}$, $\tilde{n}_2 = \sum_{i=1}^{\tilde{n}} 1\{\tilde{y}_i = -1\}$. We recall the final direction estimator as

$$\begin{aligned}
\hat{\theta}_{\text{final}} =&\frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\tilde{x}_i - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\tilde{x}_i\\
=&\frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1 - 2b_i)\,\tilde{\mu} + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1 - 2b_i)\,\tilde{\mu} - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i,
\end{aligned}$$

where $\epsilon_i \sim subGaussian\left(0, \sigma^2\right)$ independent of each other.

Now let

$$\gamma \coloneqq \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i) + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i),$$

and define

$$\tilde{\delta} \coloneqq \hat{\theta}_{\text{final}} - \gamma\tilde{\mu} = \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i.$$

We then have the decomposition and bound

$$\begin{aligned}\frac{\left\|\hat{\theta}_{\text{final}}\right\|^2}{\left(\tilde{\mu}^\top\hat{\theta}_{\text{final}}\right)^2} &= \frac{\left\|\tilde{\delta}+\gamma\tilde{\mu}\right\|^2}{\left(\gamma\left\|\tilde{\mu}\right\|^2+\tilde{\mu}^\top\tilde{\delta}\right)^2} = \frac{1}{\left\|\tilde{\mu}\right\|^2} + \frac{\left\|\tilde{\delta}+\gamma\tilde{\mu}\right\|^2 - \frac{1}{\|\tilde{\mu}\|^2}\left(\gamma\left\|\tilde{\mu}\right\|^2+\tilde{\mu}^\top\tilde{\delta}\right)^2}{\left(\gamma\left\|\tilde{\mu}\right\|^2+\tilde{\mu}^\top\tilde{\delta}\right)^2} \\ &= \frac{1}{\left\|\tilde{\mu}\right\|^2} + \frac{\|\tilde{\delta}\|^2 - \frac{1}{\|\tilde{\mu}\|^2}\left(\tilde{\mu}^\top\tilde{\delta}\right)^2}{\left(\gamma\left\|\tilde{\mu}\right\|^2+\tilde{\mu}^\top\tilde{\delta}\right)^2} \le \frac{1}{\left\|\tilde{\mu}\right\|^2} + \frac{\|\tilde{\delta}\|^2}{\|\tilde{\mu}\|^4\left(\gamma+\frac{1}{\|\tilde{\mu}\|^2}\tilde{\mu}^\top\tilde{\delta}\right)^2}. \end{aligned} \tag{4}$$

To write down concentration bounds for $\|\tilde{\delta}\|^2$ and $\tilde{\mu}^\top\tilde{\delta}$ we must address their sub-Gaussianity. To do so, write

$$\tilde{\delta} = \frac{1}{2\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)\varepsilon_i - \frac{1}{2\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=-1)}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=-1)\varepsilon_i,$$

and

$$\begin{aligned}\tilde{y}_i &\overset{i.i.d.}{\sim} \text{sign}\left((z_i\tilde{\mu}+\tilde{\nu}-\hat{b}_{intermediate}+\varepsilon_i)^\top\hat{\theta}_{\text{intermediate}}\right), \\ \tilde{y}_i\epsilon_i &\overset{i.i.d.}{\sim} \text{sign}\left((z_i\tilde{\mu}+\tilde{\nu}-\tilde{b}_{intermediate}+\varepsilon_i)^\top\hat{\theta}_{\text{intermediate}}\right)\cdot\epsilon_i,\end{aligned}$$

where z_i is the true label of $\tilde{x}_i$ (taken value from ±1).

We then have

$$\begin{aligned}\mathbb{E}[1(\tilde{y}_i=1)] =& \mathbb{P}((z_i\tilde{\mu}+\tilde{\nu}-\tilde{b}_{intermediate}+\varepsilon_i)^\top\hat{\theta}_{\text{intermediate}}>0) \\ =& \frac{1}{2}\mathbb{P}((\tilde{\mu}_1-\tilde{b}_{intermediate}+\varepsilon_i)^\top\hat{\theta}_{\text{intermediate}}>0) + \frac{1}{2}\mathbb{P}((\tilde{\mu}_2-\tilde{b}_{intermediate}+\varepsilon_i)^\top\hat{\theta}_{\text{intermediate}}>0) \\ =& \frac{1}{2}\mathbb{P}((\tilde{\mu}_1-\frac{\mu_1+\mu_2}{2}+e_b+\varepsilon_i)^\top\hat{\theta}_{\text{intermediate}}>0) + \frac{1}{2}\mathbb{P}((\tilde{\mu}_2-\frac{\mu_1+\mu_2}{2}+e_b+\varepsilon_i)^\top\hat{\theta}_{\text{intermediate}}>0) \\ \ge& \frac{1}{2}\mathbb{P}((\tilde{\mu}_1-\frac{\mu_1+\mu_2}{2}+e_b)^\top\hat{\theta}_{\text{intermediate}}+\varepsilon_i^\top\hat{\theta}_{\text{intermediate}}>0) \\ =& \frac{1}{2}\mathbb{P}(\varepsilon_i^\top\hat{\theta}_{\text{intermediate}} > -(\tilde{\mu}_1-\mu_1+\frac{\mu_1-\mu_2}{2}+e_b)^\top\hat{\theta}_{\text{intermediate}})\end{aligned}$$

The term in the last line can be bounded as follows. Let us recall $d_\nu = \max\left\{\frac{\|\tilde{\mu}_1-\mu_1\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}, \frac{\|\tilde{\mu}_2-\mu_2\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}\right\} < c_0$ implies that $\|\tilde{\mu}-\mu\| < c_0\|\tilde{\mu}\|$, and therefore $\|\mu\| \ge \|\tilde{\mu}\| - c_0\|\tilde{\mu}\| = (1-c_0)\|\tilde{\mu}\|$. We then obtain

$$|(\tilde{\mu}_1-\mu_1)^\top\mu| \le \|\tilde{\mu}_1-\mu_1\|\cdot\|\mu\| \le c_0\|\tilde{\mu}\|\cdot\|\mu\| \le \frac{c_0}{1-c_0}\|\mu\|^2.$$

As a result, we have

$$\begin{aligned}|(\tilde{\mu}_1-\mu_1+\frac{\mu_1-\mu_2}{2}+e_b)^\top\hat{\theta}_{\text{intermediate}}| =& |(\tilde{\mu}_1-\mu_1+\frac{\mu_1-\mu_2}{2}+e_b)^\top(\mu+e_w)| \\ \ge& \|\mu\|^2 - |(\tilde{\mu}_1-\mu_1)^\top\mu| - |e_b^\top\mu| - |(\tilde{\mu}_1-\mu_1+\frac{\mu_1-\mu_2}{2}+e_b)^\top e_w| \\ \gtrsim& \Omega_p(\sqrt{d}).\end{aligned}$$

We then have

$$\begin{aligned}\mathbb{E}[1(\tilde{y}_i=1)] \geq & \frac{1}{2}\mathbb{P}(\varepsilon_i^\top \hat{\theta}_{\text{intermediate}} > -(\tilde{\mu}_1 - \mu_1 + \frac{\mu_1-\mu_2}{2} + e_b)^\top \hat{\theta}_{\text{intermediate}}) \\ \geq & \frac{1}{2}\mathbb{P}(\varepsilon_i^\top \frac{\mu_1-\mu_2}{2} > 0) \geq \frac{c}{2}, \end{aligned} \tag{5}$$

for some constant c close to 1 when d is sufficiently large.

Therefore, we have

$$\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1) \geq c + o_p(1),$$

and

$$\|\frac{1}{\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\varepsilon_i\| \lesssim \|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\varepsilon_i\|$$

In addition, we have

$$\|\mathbb{E}[1(\tilde{y}_i=1)\epsilon_i]\| = \|\mathbb{E}[\mathbb{E}[1(\tilde{y}_i=1)\epsilon_i \mid \hat{\theta}_{\text{intermediate}}^\top \varepsilon_i]]\| \leq \mathbb{E}[\sqrt{d} + |\hat{\theta}_{\text{intermediate}}^\top \varepsilon_i|] \lesssim \sqrt{d}.$$

Since $\|1(\tilde{y}_i=1)\epsilon_i^{(j)} - \mathbb{E}[1(\tilde{y}_i=1)\epsilon_i^{(j)}]\|_{\psi_2} \leq 2\|1(\tilde{y}_i=1)\epsilon_i^{(j)}\|_{\psi_2} \leq C\|\epsilon_i^{(j)}\|_{\psi_2} \leq C\sigma_{\max}$, we have

$$\mathbb{P}\left(\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\epsilon_i^{(j)}\right)^2 \geq (\mathbb{E}[1(\tilde{y}_i=1)\epsilon_i^{(j)}])^2 + t^2\cdot\sigma_{\max}^2\right) \leq e^{-C\tilde{n}t^2}.$$

Therefore, by union bound, with probability at least $1-d^{-1}$,

$$\|\frac{1}{\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\varepsilon_i\|^2 \lesssim \|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\epsilon_i\|^2 = \sum_{j=1}^{d}\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\epsilon_i^{(j)}\right)^2 \lesssim d + d\cdot\frac{\log d}{\tilde{n}}\sigma_{\max}^2.$$

Similarly, we have

$$\|\frac{1}{\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=-1)}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=-1)\varepsilon_i\|^2 \lesssim \|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=-1)\epsilon_i\|^2 = \sum_{j=1}^{d}\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=-1)\epsilon_i^{(j)}\right)^2 \lesssim d+d\cdot\frac{\log d}{\tilde{n}}\sigma_{\max}^2.$$

Then, since $\|\tilde{\delta}\|^2 \leq \|\frac{1}{2\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\varepsilon_i\|^2 + \|\frac{1}{2\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=-1)}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=-1)\varepsilon_i\|^2$, we have

$$\|\tilde{\delta}\|^2 = O_p(d\cdot(1+\frac{\log d}{\tilde{n}}\sigma^2)).$$

The same technique also yields a crude bound on $\tilde{\mu}^\top\tilde{\delta} = \frac{1}{2\tilde{n}_1}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\tilde{\mu}^\top\varepsilon_i - \frac{1}{2\tilde{n}_2}\sum_{i=1} 1(\tilde{y}_i=-1)\tilde{\mu}^\top\varepsilon_i$. We can write

$$1(\tilde{y}_i=1)\tilde{\mu}^\top\epsilon_i \overset{i.i.d.}{\sim} 1\left((z_i\tilde{\mu}+\tilde{\nu}-\hat{b}_{intermediate}+\varepsilon_i)^\top\hat{\theta}_{\text{intermediate}} > 0\right)\cdot\tilde{\mu}^\top\epsilon_i.$$

Since $\|1(\tilde{y}_i=1)\tilde{\mu}^\top\epsilon_i\|_{\vartheta_2} \leq C\|\tilde{\mu}^\top\epsilon_i^{(j)}\|_{\vartheta_2} \leq C\|\tilde{\mu}\|_2\sigma$, we have

$$\mathbb{P}\left(\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\tilde{\mu}^\top\epsilon_i\right)^2 \geq t^2\cdot\|\tilde{\mu}\|^2\sigma^2 + \|\tilde{\mu}\|^2\sigma^2\right) \leq e^{-C\tilde{n}t^2}.$$

and by the fact that $\left(\frac{1}{\tilde{n}_1}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\tilde{\mu}^\top\epsilon_i\right)^2 \lesssim \left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i=1)\tilde{\mu}^\top\epsilon_i\right)^2$, we have

$$\mathbb{P}\left(\left|\tilde{\mu}^\top\tilde{\delta}\right| \geq \sqrt{2}\sigma\|\tilde{\mu}\| + \|\tilde{\mu}\|\sigma\right) = \mathbb{P}\left(\left|\tilde{\mu}^\top\tilde{\delta}\right|^2 \geq C\sigma^2\|\tilde{\mu}\|^2\right) \leq e^{-\tilde{n}/8}.$$

Finally, we need to argue that γ is not too small. Recall that $\gamma = \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i) + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)$ where b_i is the indicator that $\tilde{y}_i$ is incorrect and therefore

$$\begin{aligned}\mathbb{E}\left[1-2b_i \mid \hat{\theta}_{\text{intermediate}}, \tilde{y}_i = 1\right] &= 1 - 2\mathbb{P}(f_{\hat{\theta}_{\text{intermediate}}} \mid \tilde{x} \sim subGaussian(\tilde{\mu}_1, \sigma^2)) \\ &= 2\mathbb{P}(\varepsilon_i^\top \hat{\theta}_{\text{intermediate}} > -(\tilde{\mu}_1 - \mu_1 + \frac{\mu_1-\mu_2}{2} + e_b)^\top \hat{\theta}_{\text{intermediate}}) - 1.\end{aligned}$$

This term can be lower bounded similarly as equation 7, which satisfies

$$\begin{aligned}&\mathbb{E}\left[1-2b_i \mid \hat{\theta}_{\text{intermediate}}, \tilde{y}_i = 1\right] \\ =&2\mathbb{P}(\varepsilon_i^\top \hat{\theta}_{\text{intermediate}} > -(\tilde{\mu}_1 - \mu_1 + \frac{\mu_1-\mu_2}{2} + e_b)^\top \hat{\theta}_{\text{intermediate}}) - 1 \geq \frac{4}{5},\end{aligned}$$

with high probability when d is sufficiently large.

Similarly, we have

$$\mathbb{E}\left[1-2b_i \mid \hat{\theta}_{\text{intermediate}}, \tilde{y}_i = -1\right] \geq \frac{4}{5},$$

with high probability when d is sufficiently large.

Therefore we expect γ to be reasonably large as long as $\mathbb{E}[\gamma] \geq \frac{4}{5}$. Indeed, define

$$\tilde{\gamma} = \frac{1}{\tilde{n}}\sum_{i=1}^{n}(1-2b_i).$$

We then have

$$\begin{aligned}\mathbb{E}[\tilde{\gamma}] \geq&\mathbb{E}[\frac{1}{\tilde{n}}\sum_{y_i=1}(1-2b_i) + \frac{1}{\tilde{n}}\sum_{y_i=-1}(1-2b_i)] \\ \geq&\mathbb{E}[\frac{1}{\tilde{n}}\cdot\frac{4}{5}\tilde{n}_1 + \frac{1}{\tilde{n}}\cdot\frac{4}{5}\tilde{n}_2] \geq \frac{4}{5}.\end{aligned}$$

By using $\gamma \geq \frac{1}{2}\tilde{\gamma}$, we have

$$\begin{aligned}\mathbb{P}(\gamma \geq \frac{1}{5}) \geq \mathbb{P}(\tilde{\gamma} \geq \frac{2}{5}) = 1 - \mathbb{P}(\tilde{\gamma} < \frac{2}{5}) \\ \geq 1 - \mathbb{P}(|\tilde{\gamma} - \mathbb{E}[\tilde{\gamma}]| > \frac{2}{5}) \geq 1 - e^{-c\tilde{n}},\end{aligned}$$

where the last inequality is due to Hoeffding's inequality.

As a result, we have $\gamma \geq \frac{2}{5}$ with high probability.

Define the event,

$$\mathcal{E} = \left\{\|\tilde{\delta}\|^2 \leq \|\gamma(\frac{\|\mu\|}{\|\tilde{\mu}\|}\tilde{\mu} - \mu)\|^2 + \frac{d\cdot\sigma_{\max}^2}{\tilde{n}}\log d + d\xi_n^2,\ \left|\mu^\top\tilde{\delta}\right| \leq \sqrt{2}\sigma_{\max}\|\mu\| + \gamma\mu^\top(\frac{\|\mu\|}{\|\tilde{\mu}\|}\tilde{\mu} - \mu) + \xi_n\|\mu\| \ \text{ and } \gamma \geq \frac{2}{5}\right\};$$

by the preceding discussion,

$$\mathbb{P}\left(\mathcal{E}^C\right) \leq \frac{1}{d} + e^{-\tilde{n}/8} + e^{-c\|\mu\|^2/8\sigma_{\max}^2} + 2e^{-cn\|\mu\|/2\sigma_{\max}} + e^{-c\tilde{n}}$$

Moreover, by the bound (6), $\mathcal{E}$ implies

$$\frac{\left\|\hat{\theta}_{\text{final}}\right\|^2}{\left(\tilde{\mu}^\top\hat{\theta}_{\text{final}}\right)^2} \leq \frac{1}{\|\tilde{\mu}\|^2} + \frac{\|\tilde{\delta}\|^2}{\|\tilde{\mu}\|^4\left(\gamma + \frac{1}{\|\tilde{\mu}\|^2}\tilde{\mu}^\top\tilde{\delta}\right)^2} \leq \frac{1}{\|\tilde{\mu}\|^2} + \frac{d\cdot(1+\frac{\log d}{\tilde{n}}\sigma_{\max}^2)}{\|\tilde{\mu}\|^4\left(\frac{2}{5} + \frac{1}{\|\tilde{\mu}\|^2}\cdot\|\tilde{\mu}\|\sigma_{\max}\right)^2}.$$

Therefore,

$$\frac{\tilde{\mu}^\top \hat{\theta}_{\text{final}}}{\sigma_{\max}\left\|\hat{\theta}_{\text{final}}\right\|} \geq \left(\frac{\sigma_{\max}^2}{\|\tilde{\mu}\|^2} + \frac{d\cdot(1+\frac{\log d}{\tilde{n}}\sigma_{\max}^2)}{\|\tilde{\mu}\|^4\left(\frac{2}{5}-\frac{\sigma_{\max}}{\|\tilde{\mu}\|_2}\right)^2}\right)^{-1/2}$$

with probability $\geq 1-(\frac{1}{d}+e^{-\tilde{n}/8}+e^{-c\|\mu\|^2/8\sigma_{\max}^2}+2e^{-cn\|\mu\|/2\sigma_{\max}})$.

Recall that we take $\sigma_{\max} := C_\sigma d^{1/4}$ and $\|\tilde{\mu}_1-\tilde{\mu}_2\|_2 = C_\mu\sqrt{d}$ for sufficiently large C_μ, we than have when $\tilde{n}\gtrsim \varepsilon^2 d\log d$,

$$\frac{\tilde{\mu}^\top \hat{\theta}_{\text{final}}}{\left\|\hat{\theta}_{\text{final}}\right\|} = \Omega_P(\sqrt{d}).$$

Then let us consider

$$\begin{aligned}
\hat{b}_{\text{final}} =& \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\tilde{x}_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\tilde{x}_i \\
=& \tilde{\nu} + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i)\,\tilde{\mu} - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)\,\tilde{\mu} + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i \\
=& \tilde{\nu} + \left[\frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i) - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)\right]\tilde{\mu} + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i,
\end{aligned}$$

Let

$$\lambda := \frac{1}{\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i) - \frac{1}{\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)$$

When $n > C$ for sufficiently large C, we have $\lambda \leq 0.01$.

Also, let us denote $\tilde{\delta}_2 = \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i$, we then have

$$|\tilde{\delta}^\top\tilde{\delta}_2| = \|\frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i\|^2 - \|\frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i\|^2 \leq \|\frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i\|^2 \leq c_\varepsilon(d + d\cdot\frac{\log d}{\tilde{n}}\sigma_{\max}^2)$$

We also have

$$\begin{aligned}
|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}d_b| \leq& \lambda|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\tilde{\mu}| + |\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\tilde{\delta}_2| \\
\leq& \lambda\|\tilde{\mu}\| + |\frac{(\tilde{\delta}+\gamma\tilde{\mu})^\top}{\|\tilde{\delta}+\gamma\tilde{\mu}\|}\tilde{\delta}_2| \\
\leq& \lambda\|\tilde{\mu}\| + \frac{|\tilde{\delta}^\top\tilde{\delta}_2 + \gamma\tilde{\mu}^\top\tilde{\delta}_2|}{\|\gamma\tilde{\mu}\| - \|\tilde{\delta}\|} \\
\leq& \lambda\|\tilde{\mu}\| + \frac{c_\varepsilon(d + d\cdot\frac{\log d}{\tilde{n}}\sigma^2) + O_P(\|\tilde{\mu}\|\sigma)}{\gamma\|\tilde{\mu}\| - c_\varepsilon(d + d\cdot\frac{\log d}{\tilde{n}}\sigma^2)} \\
\leq& (\lambda C_\mu + \frac{c_\varepsilon}{\gamma C_\mu - c_\varepsilon})\cdot\sqrt{d}
\end{aligned}$$

Therefore, when the constant C_μ is sufficiently large,

$$\begin{aligned}
&\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}((\mu_1-\tilde{\mu}_1)+\frac{\tilde{\mu}_1-\tilde{\mu}_2}{2}-d_b)\\
\geq&|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\tilde{\mu}|-|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\mu_1-\tilde{\mu}_1)|-|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}d_b|\\
\geq&\left(\frac{\sigma^2}{\|\tilde{\mu}\|^2}+\frac{d\cdot(1+\frac{\log d}{\tilde{n}}\sigma_{\max}^2)}{\|\tilde{\mu}\|^4\left(\frac{1}{6}-\frac{\sigma_{\max}}{\|\tilde{\mu}\|_2}\right)^2}\right)^{-1/2}-2c_0\|\tilde{\mu}\|-(\lambda C_\mu+\frac{c_\varepsilon}{\gamma C_\mu-c_\varepsilon})\cdot\sqrt{d}\\
=&\Omega_P(\sqrt{d}).
\end{aligned}$$

The robust error is then

$$\begin{aligned}
&\mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\varepsilon\leqslant-\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\mu_1-\hat{b})+L_1'\cdot\varepsilon\sqrt{d}\Big)\\
=&\mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\varepsilon\leqslant-\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}((\mu_1-\tilde{\mu}_1)+\frac{\tilde{\mu}_1-\tilde{\mu}_2}{2}-d_b)+L_1'\cdot\varepsilon\sqrt{d}\Big)\\
\leq&\exp(-C\sqrt{d})\leq 0.01,
\end{aligned}$$

when d is sufficiently large.

A.6 Proof of Proposition 1

The proof of Proposition 1 is very similar to those of Theorem 3 except for the tail probabilities changed from subgaussian to $g(\cdot)$. For completeness, we present the proof below.

We first recall the definition of D_g:

$$\begin{aligned}
D_g(\mu,\sigma^2)=\{X\in\mathbb{R}^d:\forall v\in\mathbb{R}^d,\|v\|_2=1,\text{Var}(X_j)\leq\sigma^2\\
\mathbb{P}(|v^T(X-\mu)|>\sigma\cdot t)\leq g(t)\},
\end{aligned}$$

and restate Proposition 1.

Proposition 1 Suppose D_g is closed under independent summation, and assume $\left\|\mathbb{E}\left[\tilde{x}_i-\mathbb{E}[\tilde{x}_i]\mid a^T\tilde{x}_i=b\right]\right\|\lesssim\sqrt{d}+|b|$ for fixed unit vector a, $\tilde{\sigma}\leq\sigma_{\max}\asymp d^{1/4}$, $\|\tilde{\mu}_1-\tilde{\mu}_2\|_2\asymp\sqrt{d}$, $c<q<1-c$ for some constant $0<c<1/2$, and

$$d_\nu=\max\left\{\frac{\|\tilde{\mu}_1-\mu_1\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|},\frac{\|\tilde{\mu}_2-\mu_2\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}\right\}<c_0,$$

for some constant $c_0\leq 1/4$, then the robust classification error is at most 1% when d is sufficiently large, $n\geq C$ for some constant C (not depending on d and ε) and

$$\tilde{n}\gtrsim\varepsilon^2\cdot(g^{-1}(1/d\log d))^2\cdot\sqrt{d}.$$

Now let us proceed to the proof.

We first recall the distributions for the two classes of labeled data as $D_g(\mu_1,\sigma_{\max}^2)$, and $D_g(\mu_2,\sigma_{\max}^2)$ respectively. Similarly, we also denote the distributions for the two classes of unlabeled data as $D_g(\tilde{\mu}_1,\sigma_{\max}^2)$, and $D_g(\tilde{\mu}_2,\sigma_{\max}^2)$ respectively. Also, to avoid the visual similarity and emphasize the estimates constructed by the labeled and unlabeled data respectively, we write $\hat{w}$ as $\hat{w}_{\text{intermediate}}$, $\hat{b}$ as $\hat{b}_{\text{intermediate}}$, $\tilde{w}$ as $\tilde{w}_{\text{final}}$ and $\tilde{b}$ as $\tilde{b}_{\text{final}}$.

Then, let us write out the robust error of misclassifying class 1 against the ℓ_∞ attack (the robust error of

misclassifying class 2 can be bounded similarly) as

$$\begin{aligned}
&\max_{\|\delta\|_\infty \le L_1'\varepsilon} \mathbb{P}\Big(\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}(\tilde x + \delta - \tilde b_{\text{final}}) \leqslant 0 \mid \tilde x \sim D_g(\tilde\mu_1, \sigma_{\max}^2)\Big)\\
=&\mathbb{P}\Big(\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}(\mu_1 - \tilde b_{\text{final}}) + \varepsilon\frac{\|\hat\theta_{\text{final}}\|_1}{\|\hat\theta_{\text{final}}\|}\Big)\\
=&\mathbb{P}\Big(\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}(\mu_1 - \tilde b_{\text{final}}) + L_1'\cdot\varepsilon\sqrt{d}\Big)
\end{aligned}$$

Denote $\tilde b_{\text{final}} := \frac{\tilde\mu_1+\tilde\mu_2}{2} + d_b$, we then have

$$\begin{aligned}
&\mathbb{P}\Big(\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}(\mu_1 - \tilde b_{\text{final}}) + L_1'\cdot\varepsilon\sqrt{d}\Big)\\
=&\mathbb{P}\Big(\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}((\mu_1 - \tilde\mu_1) + \frac{\tilde\mu_1 - \tilde\mu_2}{2} - d_b) + L_1'\cdot\varepsilon\sqrt{d}\Big)
\end{aligned}$$

We are going to bound $|\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}(\mu_1 - \tilde\mu_1)|$, $|\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}(\tilde\mu_1 - \tilde\mu_2)/2|$, and $|\frac{\hat\theta_{\text{final}}^\top}{\|\hat\theta_{\text{final}}\|}d_b|$ respectively.

Let $\tilde\mu = (\tilde\mu_1 - \tilde\mu_2)/2$, $\tilde\nu = (\tilde\mu_1 + \tilde\mu_2)/2$, $\mu = (\mu_1 - \mu_2)/2$, $\nu = (\mu_1 + \mu_2)/2$, and b_i be the indicator that the ith pseudo-label $\tilde y_i$ is incorrect, so that $\tilde x_i \sim \tilde\nu + D_g\left((1-2b_i)\,\tilde y_i\tilde\mu, \sigma^2\right)$

Let $\tilde n_1 = \sum_{i=1}^{\tilde n} 1\{\tilde y_i = 1\}$, $\tilde n_2 = \sum_{i=1}^{\tilde n} 1\{\tilde y_i = -1\}$. We recall the final direction estimator as

$$\begin{aligned}
\hat\theta_{\text{final}} =&\frac{1}{2\tilde n_1}\sum_{\tilde y_i=1}\tilde x_i - \frac{1}{2\tilde n_2}\sum_{\tilde y_i=-1}\tilde x_i\\
=&\frac{1}{2\tilde n_1}\sum_{\tilde y_i=1}(1-2b_i)\,\tilde\mu + \frac{1}{2\tilde n_1}\sum_{\tilde y_i=1}\varepsilon_i + \frac{1}{2\tilde n_2}\sum_{\tilde y_i=-1}(1-2b_i)\,\tilde\mu - \frac{1}{2\tilde n_2}\sum_{\tilde y_i=-1}\varepsilon_i,
\end{aligned}$$

where $\epsilon_i \sim D_g\left(0, \sigma^2\right)$ independent of each other.

Now let

$$\gamma := \frac{1}{2\tilde n_1}\sum_{\tilde y_i=1}(1-2b_i) + \frac{1}{2\tilde n_2}\sum_{\tilde y_i=-1}(1-2b_i),$$

and define

$$\tilde\delta := \hat\theta_{\text{final}} - \gamma\tilde\mu = \frac{1}{2\tilde n_1}\sum_{\tilde y_i=1}\varepsilon_i - \frac{1}{2\tilde n_2}\sum_{\tilde y_i=-1}\varepsilon_i.$$

We then have the decomposition and bound

$$\begin{aligned}
\frac{\left\|\hat\theta_{\text{final}}\right\|^2}{\left(\tilde\mu^\top\hat\theta_{\text{final}}\right)^2} &= \frac{\left\|\tilde\delta + \gamma\tilde\mu\right\|^2}{\left(\gamma\left\|\tilde\mu\right\|^2 + \tilde\mu^\top\tilde\delta\right)^2} = \frac{1}{\|\tilde\mu\|^2} + \frac{\left\|\tilde\delta + \gamma\tilde\mu\right\|^2 - \frac{1}{\|\tilde\mu\|^2}\left(\gamma\left\|\tilde\mu\right\|^2 + \tilde\mu^\top\tilde\delta\right)^2}{\left(\gamma\left\|\tilde\mu\right\|^2 + \tilde\mu^\top\tilde\delta\right)^2}\\
&= \frac{1}{\|\tilde\mu\|^2} + \frac{\|\tilde\delta\|^2 - \frac{1}{\|\tilde\mu\|^2}\left(\tilde\mu^\top\tilde\delta\right)^2}{\left(\gamma\left\|\tilde\mu\right\|^2 + \tilde\mu^\top\tilde\delta\right)^2} \le \frac{1}{\|\tilde\mu\|^2} + \frac{\|\tilde\delta\|^2}{\|\tilde\mu\|^4\left(\gamma + \frac{1}{\|\tilde\mu\|^2}\tilde\mu^\top\tilde\delta\right)^2}.
\end{aligned} \tag{6}$$

To write down concentration bounds for $\|\tilde\delta\|^2$ and $\tilde\mu^\top\tilde\delta$ we must control their tail bound. To do so, write

$$\tilde\delta = \frac{1}{2\sum_{i=1}^{\tilde n}1(\tilde y_i = 1)}\sum_{i=1}^{\tilde n}1(\tilde y_i = 1)\varepsilon_i - \frac{1}{2\sum_{i=1}^{\tilde n}1(\tilde y_i = -1)}\sum_{i=1}^{\tilde n}1(\tilde y_i = -1)\varepsilon_i,$$

and

$$\tilde{y}_i \overset{i.i.d.}{\sim} \text{sign}\left((z_i\tilde{\mu} + \tilde{\nu} - \hat{b}_{intermediate} + \varepsilon_i)^\top \hat{\theta}_{\text{intermediate}}\right),$$
$$\tilde{y}_i\epsilon_i \overset{i.i.d.}{\sim} \text{sign}\left((z_i\tilde{\mu} + \tilde{\nu} - \tilde{b}_{intermediate} + \varepsilon_i)^\top \hat{\theta}_{\text{intermediate}}\right) \cdot \epsilon_i,$$

where z_i is the true label of $\tilde{x}_i$ (taken value from ± 1).

We then have

$$\begin{aligned}
\mathbb{E}[1(\tilde{y}_i = 1)] =& \mathbb{P}((z_i\tilde{\mu} + \tilde{\nu} - \tilde{b}_{intermediate} + \varepsilon_i)^\top \hat{\theta}_{\text{intermediate}} > 0) \\
=& \frac{1}{2}\mathbb{P}((\tilde{\mu}_1 - \tilde{b}_{intermediate} + \varepsilon_i)^\top \hat{\theta}_{\text{intermediate}} > 0) + \frac{1}{2}\mathbb{P}((\tilde{\mu}_2 - \tilde{b}_{intermediate} + \varepsilon_i)^\top \hat{\theta}_{\text{intermediate}} > 0) \\
=& \frac{1}{2}\mathbb{P}((\tilde{\mu}_1 - \frac{\mu_1+\mu_2}{2} + e_b + \varepsilon_i)^\top \hat{\theta}_{\text{intermediate}} > 0) + \frac{1}{2}\mathbb{P}((\tilde{\mu}_2 - \frac{\mu_1+\mu_2}{2} + e_b + \varepsilon_i)^\top \hat{\theta}_{\text{intermediate}} > 0) \\
\geq& \frac{1}{2}\mathbb{P}((\tilde{\mu}_1 - \frac{\mu_1+\mu_2}{2} + e_b)^\top \hat{\theta}_{\text{intermediate}} + \varepsilon_i^\top \hat{\theta}_{\text{intermediate}} > 0) \\
=& \frac{1}{2}\mathbb{P}(\varepsilon_i^\top \hat{\theta}_{\text{intermediate}} > -(\tilde{\mu}_1 - \mu_1 + \frac{\mu_1-\mu_2}{2} + e_b)^\top \hat{\theta}_{\text{intermediate}})
\end{aligned}$$

The term in the last line can be bounded as follows. Let us recall $d_\nu = \max\left\{\frac{\|\tilde{\mu}_1-\mu_1\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}, \frac{\|\tilde{\mu}_2-\mu_2\|}{\|\tilde{\mu}_1-\tilde{\mu}_2\|}\right\} < c_0$ implies that $\|\tilde{\mu} - \mu\| < c_0\|\tilde{\mu}\|$, and therefore $\|\mu\| \geq \|\tilde{\mu}\| - c_0\|\tilde{\mu}\| = (1-c_0)\|\tilde{\mu}\|$. We then obtain

$$|(\tilde{\mu}_1 - \mu_1)^\top \mu| \leq \|\tilde{\mu}_1 - \mu_1\| \cdot \|\mu\| \leq c_0\|\tilde{\mu}\| \cdot \|\mu\| \leq \frac{c_0}{1-c_0}\|\mu\|^2.$$

As a result, we have

$$\begin{aligned}
|(\tilde{\mu}_1 - \mu_1 + \frac{\mu_1-\mu_2}{2} + e_b)^\top \hat{\theta}_{\text{intermediate}}| =& |(\tilde{\mu}_1 - \mu_1 + \frac{\mu_1-\mu_2}{2} + e_b)^\top (\mu + e_w)| \\
\geq& \|\mu\|^2 - |(\tilde{\mu}_1 - \mu_1)^\top \mu| - |e_b^\top \mu| - |(\tilde{\mu}_1 - \mu_1 + \frac{\mu_1-\mu_2}{2} + e_b)^\top e_w| \\
\gtrsim& \Omega_p(\sqrt{d}).
\end{aligned}$$

We then have

$$\begin{aligned}
\mathbb{E}[1(\tilde{y}_i = 1)] \geq& \frac{1}{2}\mathbb{P}(\varepsilon_i^\top \hat{\theta}_{\text{intermediate}} > -(\tilde{\mu}_1 - \mu_1 + \frac{\mu_1-\mu_2}{2} + e_b)^\top \hat{\theta}_{\text{intermediate}}) \\
\geq& \frac{1}{2}\mathbb{P}(\varepsilon_i^\top \frac{\mu_1-\mu_2}{2} > 0) \geq \frac{c}{2},
\end{aligned} \tag{7}$$

for some constant c close to 1 when d is sufficiently large.

Therefore, we have

$$\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i = 1) \geq c + o_p(1),$$

and

$$\|\frac{1}{\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i = 1)} \sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i = 1)\varepsilon_i\| \lesssim \|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i = 1)\varepsilon_i\|$$

In addition, we have

$$\|\mathbb{E}[1(\tilde{y}_i = 1)\epsilon_i]\| = \|\mathbb{E}[\mathbb{E}[1(\tilde{y}_i = 1)\epsilon_i \mid \hat{\theta}_{\text{intermediate}}^\top \varepsilon_i]]\| \leq \mathbb{E}[\sqrt{d} + |\hat{\theta}_{\text{intermediate}}^\top \varepsilon_i|] \lesssim \sqrt{d}.$$

By the definition of D_g, we have

$$\mathbb{P}\left(\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} 1(\tilde{y}_i = 1)\epsilon_i^{(j)}\right)^2 \geq (\mathbb{E}[1(\tilde{y}_i = 1)\epsilon_i^{(j)}])^2 + t^2 \cdot \sigma_{\max}^2\right) \leq g(C\sqrt{\tilde{n}}t).$$

Therefore, by union bound, with probability at least $1-(\log d)^{-1}$,

$$\|\frac{1}{\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)\varepsilon_i\|^2 \lesssim \|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)\epsilon_i\|^2 = \sum_{j=1}^{d}\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)\epsilon_i^{(j)}\right)^2 \lesssim d+d\cdot\frac{(g^{-1}(1/d\log d))^2}{\tilde{n}}\sigma_{\max}^2.$$

Similarly, we have

$$\|\frac{1}{\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=-1)}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=-1)\varepsilon_i\|^2 \lesssim \|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=-1)\epsilon_i\|^2 = \sum_{j=1}^{d}\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=-1)\epsilon_i^{(j)}\right)^2 \lesssim d+d\cdot\frac{(g^{-1}(1/d\log d))^2}{\tilde{n}}\sigma_{\max}^2.$$

Then, since $\|\tilde{\delta}\|^2 \le \|\frac{1}{2\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)\varepsilon_i\|^2 + \|\frac{1}{2\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=-1)}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=-1)\varepsilon_i\|^2$, we have

$$\|\tilde{\delta}\|^2 = O_p(d\cdot(1+\frac{(g^{-1}(1/d\log d))^2}{\tilde{n}}\sigma^2)).$$

The same technique also yields a crude bound on $\tilde{\mu}^\top\tilde{\delta} = \frac{1}{2\tilde{n}_1}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)\tilde{\mu}^\top\varepsilon_i - \frac{1}{2\tilde{n}_2}\sum_{i=1}1(\tilde{y}_i=-1)\tilde{\mu}^\top\varepsilon_i$. We can write

$$1(\tilde{y}_i=1)\tilde{\mu}^\top\epsilon_i \overset{i.i.d.}{\sim} 1\left((z_i\tilde{\mu}+\tilde{\nu}-\hat{b}_{intermediate}+\varepsilon_i)^\top\hat{\theta}_{\text{intermediate}}>0\right)\cdot\tilde{\mu}^\top\epsilon_i.$$

By definition of D_g, we have

$$\mathbb{P}\left(\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)\tilde{\mu}^\top\epsilon_i\right)^2 \ge t^2\cdot\|\tilde{\mu}\|^2\sigma^2+\|\tilde{\mu}\|^2\sigma^2\right) \le g(C\sqrt{\tilde{n}}t).$$

and by the fact that $\left(\frac{1}{\tilde{n}_1}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)\tilde{\mu}^\top\epsilon_i\right)^2 \lesssim \left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}1(\tilde{y}_i=1)\tilde{\mu}^\top\epsilon_i\right)^2$, we have

$$\mathbb{P}\left(\left|\tilde{\mu}^\top\tilde{\delta}\right| \ge \sqrt{2}\sigma\|\tilde{\mu}\|+\|\tilde{\mu}\|\sigma\right) = \mathbb{P}\left(\left|\tilde{\mu}^\top\tilde{\delta}\right|^2 \ge C\sigma^2\|\tilde{\mu}\|^2\right) \le g(C\sqrt{\tilde{n}}).$$

Finally, we need to argue that γ is not too small. Recall that $\gamma = \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i)+\frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)$ where b_i is the indicator that $\tilde{y}_i$ is incorrect and therefore

$$\begin{aligned}\mathbb{E}\left[1-2b_i \mid \hat{\theta}_{\text{intermediate}}, \tilde{y}_i=1\right] &= 1-2\mathbb{P}(f_{\hat{\theta}_{\text{intermediate}}} \mid \tilde{x}\sim subGaussian(\tilde{\mu}_1,\sigma^2))\\ &= 2\mathbb{P}(\varepsilon_i^\top\hat{\theta}_{\text{intermediate}} > -(\tilde{\mu}_1-\mu_1+\frac{\mu_1-\mu_2}{2}+e_b)^\top\hat{\theta}_{\text{intermediate}})-1.\end{aligned}$$

This term can be lower bounded similarly as equation 7, which satisfies

$$\begin{aligned}&\mathbb{E}\left[1-2b_i \mid \hat{\theta}_{\text{intermediate}}, \tilde{y}_i=1\right]\\ =&2\mathbb{P}(\varepsilon_i^\top\hat{\theta}_{\text{intermediate}} > -(\tilde{\mu}_1-\mu_1+\frac{\mu_1-\mu_2}{2}+e_b)^\top\hat{\theta}_{\text{intermediate}})-1 \ge \frac{4}{5},\end{aligned}$$

with high probability when d is sufficiently large.

Similarly, we have

$$\mathbb{E}\left[1-2b_i \mid \hat{\theta}_{\text{intermediate}}, \tilde{y}_i=-1\right] \ge \frac{4}{5},$$

with high probability when d is sufficiently large.

Therefore we expect γ to be reasonably large as long as $\mathbb{E}[\gamma]\ge\frac{4}{5}$. Indeed, define

$$\tilde{\gamma} = \frac{1}{\tilde{n}}\sum_{i=1}^{n}(1-2b_i).$$

We then have

$$\begin{aligned}\mathbb{E}[\tilde{\gamma}] \geq & \mathbb{E}[\frac{1}{\tilde{n}}\sum_{y_i=1}(1-2b_i)+\frac{1}{\tilde{n}}\sum_{y_i=-1}(1-2b_i)] \\ \geq & \mathbb{E}[\frac{1}{\tilde{n}}\cdot\frac{4}{5}\tilde{n}_1+\frac{1}{\tilde{n}}\cdot\frac{4}{5}\tilde{n}_2] \geq \frac{4}{5}.\end{aligned}$$

By using $\gamma \geq \frac{1}{2}\tilde{\gamma}$, we have

$$\begin{aligned}\mathbb{P}(\gamma \geq \frac{1}{5}) \geq & \mathbb{P}(\tilde{\gamma} \geq \frac{2}{5}) = 1-\mathbb{P}(\tilde{\gamma} < \frac{2}{5}) \\ \geq & 1-\mathbb{P}(|\tilde{\gamma}-\mathbb{E}[\tilde{\gamma}]| > \frac{2}{5}) \geq 1-e^{-c\tilde{n}},\end{aligned}$$

where the last inequality is due to Hoeffding's inequality.

As a result, we have $\gamma \geq \frac{2}{5}$ with high probability.

Define the event,

$$\mathcal{E} = \left\{\|\tilde{\delta}\|^2 \leq \|\gamma(\frac{\|\mu\|}{\|\tilde{\mu}\|}\tilde{\mu}-\mu)\|^2 + \frac{d\cdot\sigma_{\max}^2}{\tilde{n}}\log d + d\xi_n^2,\ \left|\mu^\top\tilde{\delta}\right| \leq \sqrt{2}\sigma_{\max}\|\mu\| + \gamma\mu^\top(\frac{\|\mu\|}{\|\tilde{\mu}\|}\tilde{\mu}-\mu) + \xi_n\|\mu\| \ \text{ and } \gamma \geq \frac{2}{5}\right\};$$

by the preceding discussion,

$$\mathbb{P}\left(\mathcal{E}^C\right) \leq \frac{1}{\log d} + g(C\sqrt{\tilde{n}}) + g(C\|\mu\|/\sigma_{\max}) + 2g(C\sqrt{\tilde{n}\|\mu\|/\sigma_{\max}}) + e^{-c\tilde{n}}$$

Moreover, by the bound (6), $\mathcal{E}$ implies

$$\frac{\left\|\hat{\theta}_{\text{final}}\right\|^2}{\left(\tilde{\mu}^\top\hat{\theta}_{\text{final}}\right)^2} \leq \frac{1}{\|\tilde{\mu}\|^2} + \frac{\|\tilde{\delta}\|^2}{\|\tilde{\mu}\|^4\left(\gamma+\frac{1}{\|\tilde{\mu}\|^2}\tilde{\mu}^\top\tilde{\delta}\right)^2} \leq \frac{1}{\|\tilde{\mu}\|^2} + \frac{d\cdot(1+\frac{(g^{-1}(1/d\log d))^2}{\tilde{n}}\sigma_{\max}^2)}{\|\tilde{\mu}\|^4\left(\frac{2}{5}+\frac{1}{\|\tilde{\mu}\|^2}\cdot\|\tilde{\mu}\|\sigma_{\max}\right)^2}.$$

Therefore,

$$\frac{\tilde{\mu}^\top\hat{\theta}_{\text{final}}}{\sigma_{\max}\left\|\hat{\theta}_{\text{final}}\right\|} \geq \left(\frac{\sigma_{\max}^2}{\|\tilde{\mu}\|^2} + \frac{d\cdot(1+\frac{(g^{-1}(1/d\log d))^2}{\tilde{n}}\sigma_{\max}^2)}{\|\tilde{\mu}\|^4\left(\frac{2}{5}-\frac{\sigma_{\max}}{\|\tilde{\mu}\|_2}\right)^2}\right)^{-1/2}$$

with probability $\geq 1-(\frac{1}{\log d}+g(C\sqrt{\tilde{n}})+g(C\|\mu\|/\sigma_{\max})+2g(C\sqrt{\tilde{n}\|\mu\|/\sigma_{\max}})+e^{-c\tilde{n}})$.

Recall that we take $\sigma_{\max} := C_\sigma d^{1/4}$ and $\|\tilde{\mu}_1-\tilde{\mu}_2\|_2 = C_\mu\sqrt{d}$ for sufficiently large C_μ, we than have when $\tilde{n} \gtrsim \varepsilon^2 d(g^{-1}(1/d\log d))^2$,

$$\frac{\tilde{\mu}^\top\hat{\theta}_{\text{final}}}{\left\|\hat{\theta}_{\text{final}}\right\|} = \Omega_P(\sqrt{d}).$$

Then let us consider

$$\begin{aligned}\hat{b}_{\text{final}} = & \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\tilde{x}_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\tilde{x}_i \\ = & \tilde{\nu} + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i)\,\tilde{\mu} - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)\,\tilde{\mu} + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i \\ = & \tilde{\nu} + \left[\frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i) - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)\right]\tilde{\mu} + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i,\end{aligned}$$

Let

$$\lambda \coloneqq \frac{1}{\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i) - \frac{1}{\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)$$

When $n > C$ for sufficiently large C, we have $\lambda \le 0.01$.

Also, let us denote $\tilde{\delta}_2 = \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i$, we then have

$$|\tilde{\delta}^\top\tilde{\delta}_2| = \|\frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i\|^2 - \|\frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i\|^2 \le \|\frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i\|^2 \le c_\varepsilon(d + d\cdot\frac{(g^{-1}(1/d\log d))^2}{\tilde{n}}\sigma_{\max}^2)$$

We also have

$$\begin{aligned}
|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}d_b| \le& \lambda|\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\tilde{\mu}| + |\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\tilde{\delta}_2| \\
\le& \lambda\|\tilde{\mu}\| + |\frac{(\tilde{\delta}+\gamma\tilde{\mu})^\top}{\|\tilde{\delta}+\gamma\tilde{\mu}\|}\tilde{\delta}_2| \\
\le& \lambda\|\tilde{\mu}\| + \frac{|\tilde{\delta}^\top\tilde{\delta}_2 + \gamma\tilde{\mu}^\top\tilde{\delta}_2|}{\|\gamma\tilde{\mu}\| - \|\tilde{\delta}\|} \\
\le& \lambda\|\tilde{\mu}\| + \frac{c_\varepsilon(d + d\cdot\frac{(g^{-1}(1/d\log d))^2}{\tilde{n}}\sigma^2) + O_P(\|\tilde{\mu}\|\sigma)}{\gamma\|\tilde{\mu}\| - c_\varepsilon(d + d\cdot\frac{(g^{-1}(1/d\log d))^2}{\tilde{n}}\sigma^2)} \\
\le& (\lambda C_\mu + \frac{c_\varepsilon}{\gamma C_\mu - c_\varepsilon})\cdot\sqrt{d}
\end{aligned}$$

Therefore, when the constant C_μ is sufficiently large,

$$\begin{aligned}
&\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}((\mu_1 - \tilde{\mu}_1) + \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{2} - d_b) \\
\ge& |\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\tilde{\mu}| - |\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\mu_1 - \tilde{\mu}_1)| - |\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}d_b| \\
\ge& \left(\frac{\sigma^2}{\|\tilde{\mu}\|^2} + \frac{d\cdot(1+\frac{(g^{-1}(1/d\log d))^2}{\tilde{n}}\sigma_{\max}^2)}{\|\tilde{\mu}\|^4\left(\frac{1}{6} - \frac{\sigma_{\max}}{\|\tilde{\mu}\|_2}\right)^2}\right)^{-1/2} - 2c_0\|\tilde{\mu}\| - (\lambda C_\mu + \frac{c_\varepsilon}{\gamma C_\mu - c_\varepsilon})\cdot\sqrt{d} \\
=& \Omega_P(\sqrt{d}).
\end{aligned}$$

The robust error is then

$$\begin{aligned}
&\mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\mu_1 - \hat{b}) + L_1'\cdot\varepsilon\sqrt{d}\Big) \\
=& \mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\varepsilon \leqslant -\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}((\mu_1 - \tilde{\mu}_1) + \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{2} - d_b) + L_1'\cdot\varepsilon\sqrt{d}\Big) \\
\le& \exp(-C\sqrt{d}) \le 0.01,
\end{aligned}$$

when d is sufficiently large.

A.7 Proof of Theorem 4

Let us consider the following modelr: $x \sim N(y\mu, \sigma^2 I)$ with y uniform on $\{-1, 1\}$ and $\mu \in \mathbb{R}^d$. Consider a linear classifier $f_w(x) = sgn(x^\top w)$.

It's easy to see that the robust error probability is

$$err_{robust}^{\infty}(f) = Q(\frac{\mu^\top w}{\sigma\|w\|} - \frac{\varepsilon\|w\|_1}{\sigma\|w\|}),$$

where $Q = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2}\, dt$.

Therefore

$$\begin{aligned}\arg\min_{\|w\|=1} err^{\infty}_{robust}(f_w) =& \arg\max_{\|w\|=1} \frac{\mu^\top w}{\sigma\|w\|} - \frac{\varepsilon\|w\|_1}{\sigma\|w\|} \\ =& \arg\max_{\|w\|=1} \mu^\top w - \varepsilon\|w\|_1 \\ =& \arg\max_{\|w\|=1} \sum_{j=1}^{d} \mu_j w_j - \varepsilon|w_j|\end{aligned}$$

By observation, when reaching maximum, we have to have $sgn(w_j) = sgn(\mu_j)$, therefore

$$\begin{aligned}&\arg\max_{\|w\|=1} \sum_{j=1}^{d} \mu_j w_j - \varepsilon|w_j| \\ =& \arg\max_{\|w\|=1} \sum_{j=1}^{d} (\mu_j - \varepsilon \cdot sgn(\mu_j)) w_j \\ =& \frac{T_\varepsilon(\mu)}{\|T_\varepsilon(\mu)\|},\end{aligned}$$

where $T_\varepsilon(\mu)$ is the hard-thresholding operator with $(T_\varepsilon(\mu))_j = sgn(\mu_j) \cdot \max\{|\mu_j| - \varepsilon, 0\}$.

Now let us consider the example: μ with $\mu_j > \varepsilon$ for all $j = 1, 2, ...d$. For the shifted domain, we let $\tilde{\mu}_1 = -\tilde{\mu}_2 = \tilde{\mu} = \mu - \varepsilon \cdot 1_p$, and the mixing proportion is half-half.

Let b_i be the indicator that the ith pseudo-label $\tilde{y}_i = sgn(\tilde{x}_i^\top \hat{w}_{intermediate})$ is incorrect, so that $\tilde{x}_i \sim N\left((1-2b_i)\,\tilde{y}_i\tilde{\mu}, \sigma^2 I\right)$, and let

$$\gamma \coloneqq \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (1 - 2b_i) \in [-1, 1].$$

We may write the final direction estimator as

$$\hat{\theta}_{\text{final}} = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \tilde{y}_i \tilde{x}_i = \gamma\tilde{\mu} + \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \tilde{y}_i \epsilon_i$$

where $\epsilon_i \sim N\left(0, \sigma^2 I\right)$ independent of each other.

By orthogonal invariance of Gaussianality, we choose a coordinate system such that the first coordinate is in the direction of $\hat{w}_{intermediate}$, we then have

$$\begin{aligned}\|\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \tilde{y}_i \epsilon_i\|_2^2 =& |\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \tilde{y}_i \epsilon_{i1}|_2^2 + \sum_{j=2}^{d} |\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \tilde{y}_i \epsilon_{ij}|_2^2 \\ \leq& \frac{\sigma^2}{\tilde{n}} \chi^2_{\tilde{n}} + \frac{\sigma^2}{\tilde{n}} \chi^2_{d-1}.\end{aligned}$$

In addition, we have $\|\tilde{\mu}\|_2^2 = d\varepsilon^2$. Therefore, if $(1/d + 1/\tilde{n}) \cdot \frac{\sigma^2}{\varepsilon^2} \to 0$, we will then have

$$\frac{\hat{\theta}_{\text{final}}}{\|\hat{\theta}_{\text{final}}\|} \to \tilde{\mu},$$

and therefore

$$err^{\infty}_{robust}(f_{\hat{w}_{final}}) \leq err^{\infty}_{robust}(f_\mu).$$

A.8 Proof of Theorem 5

Suppose the labeled domain distribution is $\frac{1}{2}N(\mu, \sigma^2 I_p) + \frac{1}{2}N(-\mu, \sigma^2 I_p)$. Let $v \in \mathbb{R}^p$ be a vector such that $v^\top \mu = 0$, $\|\mu\| = a\|v\|$ for some $a > 0$, and let the unlabeled domain distribution be $\frac{1}{2}N(v, \sigma^2 I_p) + \frac{1}{2}N(-v, \sigma^2 I_p)$. That is, $\mu_1 = -\mu_2 = \mu, \tilde{\mu}_1 = -\tilde{\mu}_2 = v$.

We then have

$$d_\nu = \max\{|\frac{\|\tilde{\mu}_1 - \mu_1\|}{\|\tilde{\mu}_1 - \tilde{\mu}_2\|}|, \frac{\|\tilde{\mu}_2 - \mu_2\|}{\|\tilde{\mu}_1 - \tilde{\mu}_2\|}|\} = \frac{\sqrt{1+a^2} \cdot \|v\|}{2\|v\|} = \frac{\sqrt{1+a^2} \cdot \|\tilde{\mu}_1 - \tilde{\mu}_2\|}{2},$$

which falls into the specified class.

Now let us consider the case where $\mu = e_1, v = a^{-1}e_2$, where e_1, e_2 are the canonical basis, and study the performance of the classifier $\text{sgn}(\hat{\theta}_{\text{final}}^\top (z - \hat{b}))$, where

$$\hat{b}_{\text{final}} = \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\tilde{x}_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\tilde{x}_i; \quad \hat{\theta}_{\text{final}} = \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\tilde{x}_i - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\tilde{x}_i.$$

Similar to the proof in the last section, let b_i be the indicator that $\tilde{y}_i$ is incorrect and we decompose $\hat{\theta}_{\text{final}}$ and $\hat{b}$ into

$$\begin{aligned}
\hat{\theta}_{\text{final}} =& \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\tilde{x}_i - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\tilde{x}_i \\
=& \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i)\,v + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)\,v - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i,
\end{aligned}$$

$$\begin{aligned}
\hat{b}_{\text{final}} =& \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\tilde{x}_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\tilde{x}_i \\
=& \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}(1-2b_i)\,\tilde{\mu} - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}(1-2b_i)\,\tilde{\mu} + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i.
\end{aligned}$$

Now let us investigate b_i carefully. When $\tilde{y}_i = 1$, we have

$$\begin{aligned}
&\mathbb{E}[b_i \mid \tilde{y}_i = 1] = \mathbb{P}(\tilde{x} \sim subGaussian(\tilde{\mu}_1, \sigma^2) \mid (\tilde{x} - \hat{b}_{\text{intermediate}})^\top \hat{\theta}_{\text{intermediate}} > 0) \\
=& \frac{\mathbb{P}((\tilde{x} - \hat{b}_{\text{intermediate}})^\top \hat{\theta}_{\text{intermediate}} > 0) \mid \tilde{x} \sim subGaussian(\tilde{\mu}_1, \sigma^2)}{\mathbb{P}((\tilde{x} - \hat{b}_{\text{intermediate}})^\top \hat{\theta}_{\text{intermediate}} > 0) \mid \tilde{x} \sim (\tilde{\mu}_1, \sigma^2)) + \mathbb{P}((\tilde{x} - \hat{b}_{\text{intermediate}})^\top \hat{\theta}_{\text{intermediate}} > 0) \mid \tilde{x} \sim (\tilde{\mu}_2, \sigma^2))} \\
=& \frac{1}{2} + O_p(\sqrt{\frac{d}{n}}).
\end{aligned}$$

As a result, we have

$$\tilde{w}_{\text{final}} = (O_p(\sqrt{\frac{d}{n}}) + O_p(\sqrt{\frac{1}{\tilde{n}}}))v + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i - \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i;$$

$$\hat{b}_{\text{final}} = (O_p(\sqrt{\frac{d}{n}}) + O_p(\sqrt{\frac{1}{\tilde{n}}}))v + \frac{1}{2\tilde{n}_1}\sum_{\tilde{y}_i=1}\varepsilon_i + \frac{1}{2\tilde{n}_2}\sum_{\tilde{y}_i=-1}\varepsilon_i.$$

Then let us study $\varepsilon_i \mid \tilde{y}_i = 1$. When $\tilde{y}_i = 1$, we have

$$(\tilde{x} - \hat{b}_{\text{intermediate}})^\top \hat{\theta}_{\text{intermediate}} > 0.$$

Recall that $\mu = e_1, v = a^{-1}e_2$ the inequality is equivalent to

$$\langle e_1, \varepsilon_i \rangle + O_P(\sqrt{\frac{d}{n}}) > 0.$$

and put no constraint on other coordinates. Similarly, when $\tilde{y}_i = -1$, we have

$$\langle e_1, \varepsilon_i \rangle + O_P(\sqrt{\frac{d}{n}}) < 0,$$

and put no constraint on other coordinates.

As a result, we have

$$\frac{\tilde{w}_{\text{final}}}{\|\tilde{w}_{\text{final}}\|_2} = (O_p(\sqrt{\frac{d}{n}}) + O_p(\sqrt{\frac{1}{\tilde{n}}}))v + e_1;$$
$$\hat{b}_{\text{final}} = (O_p(\sqrt{\frac{d}{n}}) + O_p(\sqrt{\frac{1}{\tilde{n}}}))v + O_p(\sqrt{\frac{d}{\tilde{n}}}).$$

Then we write out the misclassification error

$$\mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}(\tilde{x} - \hat{b}) \leqslant 0 \mid \tilde{x} \sim subGaussian(v, \sigma^2)\Big) = \mathbb{P}\Big(\frac{\hat{\theta}_{\text{final}}^\top}{\|\hat{\theta}_{\text{final}}\|}\varepsilon \leqslant O_p(\sqrt{\frac{d}{\tilde{n}}}) + O_p(\sqrt{\frac{d}{n}})\Big)$$
$$= 1/2 + O_p(\sqrt{\frac{d}{\tilde{n}}}) + O_p(\sqrt{\frac{d}{n}}).$$

Therefore, when $\frac{d}{n}$ and $\frac{d}{\tilde{n}}$ sufficiently small, we then have

$$\beta^{\varepsilon,\infty}(\tilde{w}, \tilde{b}) \geq \beta^{\infty}(\tilde{w}, \tilde{b}) \geq 49\%.$$

A.9 The high-dimensional EM algorithm mentioned in the main paper

The algorithm used in the main paper to extract the support information from the unlabeled domain is presented in the following in Algorithm 1, which is adapted from Cai et al. (2019).

A.10 Proof of Theorem 6

Let us first adapt the Theorem 3.1 in Cai et al. (2019), which states the convergence rate of Algorithm 1

Lemma 4 (adapted from Theorem 3.1 in Cai et al. (2019))**.** *Under the same conditions of Theorem 3.6, if we choose the initializations of Algorithm 1 according to Hardt & Price (2015). Then there is a constant $\kappa \in (0, 1)$, such that the estimator $\hat{\boldsymbol{\beta}}^{(T_0)}$ satisfies*

$$\|\hat{\boldsymbol{\beta}}^{(T_0)} - (\tilde{\mu}_1 - \tilde{\mu}_2)\|_2 \lesssim \kappa^{T_0} \cdot (\|\hat{\boldsymbol{\beta}}^{(0)} - (\tilde{\mu}_1 - \tilde{\mu}_2)\|_2 + |\hat{\omega}^{(0)} - q|) + \sigma\sqrt{\frac{m \log d}{\tilde{n}}}.$$

In particular, if we let $T_0 \gtrsim (-\log(\kappa))^{-1} \log(n \cdot (\|\hat{\boldsymbol{\beta}}^{(0)} - (\tilde{\mu}_1 - \tilde{\mu}_2)\|_2 + |\hat{\omega}^{(0)} - q|))$, we have

$$\|\hat{\boldsymbol{\beta}}^{(T_0)} - (\tilde{\mu}_1 - \tilde{\mu}_2)\|_2 \lesssim \sigma\sqrt{\frac{m \log d}{\tilde{n}}}.$$

As a direct consequence of Lemma 4, we have

$$\|\hat{\boldsymbol{\beta}}^{(T_0)} - (\tilde{\mu}_1 - \tilde{\mu}_2)\|_\infty \lesssim \sigma\sqrt{\frac{m \log d}{\tilde{n}}}.$$

Using the condition that $\min_{\tilde{\mu}_{1,j} - \tilde{\mu}_{2j} \neq 0} |\tilde{\mu}_{1j} - \tilde{\mu}_{2,j}| \geq C\sigma\sqrt{2m \log d/\tilde{n}}$ for sufficiently large C, we then have, with high probability,

$$\hat{S} = Supp(\hat{\boldsymbol{\beta}}^{(T_0)}) = Supp(\tilde{\mu}_1 - \tilde{\mu}_2).$$

Therefore, when we project the labeled data to this support $\hat{S}$, it reduce the model to the previous setting considered in Theorem 3.1 and 3.2 with the dimension of $\nu(\tilde{x})$ reduced to m. Combing the proofs of Theorem 3.1, 3.2, and 3,3m we then have the desired result that if $n \gtrsim \varepsilon^2 \log d\sqrt{m}$, we have

$$\beta^{\varepsilon,\infty}(\hat{w}_{sparse}, \hat{b}_{sparse}) \leq 10^{-3} + O_P(\frac{1}{n} + \frac{1}{d}).$$

Algorithm 1 **C**lustering of **HI**gh-dimensional Gaussian **M**ixtures with the **E**M (CHIME)

Inputs: Initializations $\hat{\omega}^{(0)}, \hat{\mu}_1^{(0)}$, and $\hat{\mu}_2^{(0)}$, maximum number of iterations T_0, and a constant $\kappa \in (0,1)$. Set

$$\hat{\boldsymbol{\beta}}^{(0)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \boldsymbol{\beta}^\top(\hat{\mu}_1^{(0)} - \hat{\mu}_2^{(0)}) + \lambda_n^{(0)}\|\boldsymbol{\beta}\|_1 \right\},$$

where the tuning parameter $\lambda_n^{(0)} = C_1 \cdot (|\hat{\omega}| \vee \|\hat{\mu}_1^{(0)} - \hat{\mu}_2^{(0)}\|_{2,s})/\sqrt{s} + C_\lambda\sqrt{\log p/n}$.

for $t = 0, 1, \ldots, T_0 - 1$ **do**

Let

$$\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\tilde{x}_i) = \frac{\hat{\omega}^{(t)}}{\hat{\omega}^{(t)} + (1-\hat{\omega}^{(t)})\exp\left\{((\hat{\mu}_2^{(t)} - \hat{\mu}_1^{(t)}))^\top\big(\tilde{x}_i - \frac{\hat{\mu}_1^{(t)}+\hat{\mu}_2^{(t)}}{2}\big)\right\}}.$$

Update $\hat{\omega}^{(t+1)}, \hat{\mu}_1^{(t+1)}$, and $\hat{\mu}_2^{(t+1)}$, by

$$\begin{aligned}
\hat{\omega}^{(t+1)} &= \hat{\omega}(\hat{\boldsymbol{\theta}}^{(t)}) = \frac{1}{n}\sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\tilde{x}_i), \\
\hat{\mu}_1^{(t+1)} &= \hat{\mu}_1(\hat{\boldsymbol{\theta}}^{(t)}) = \Big\{n - \sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\tilde{x}_i)\Big\}^{-1}\Big\{\sum_{i=1}^n (1-\gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\tilde{x}_i))\tilde{x}_i\Big\}, \\
\hat{\mu}_2^{(t+1)} &= \hat{\mu}_2(\hat{\boldsymbol{\theta}}^{(t)}) = \Big\{\sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\tilde{x}_i)\Big\}^{-1}\Big\{\sum_{i=1}^n \gamma_{\hat{\boldsymbol{\theta}}^{(t)}}(\tilde{x}_i)\tilde{x}_i\Big\},
\end{aligned}$$

and update $\hat{\boldsymbol{\beta}}^{(t+1)}$ via

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \boldsymbol{\beta}^\top(\hat{\mu}_1^{(t+1)} - \hat{\mu}_2^{(t+1)}) + \lambda_n^{(t+1)}\|\boldsymbol{\beta}\|_1 \right\},$$

with

$$\lambda_n^{(t+1)} = \kappa\lambda_n^{(t)} + C_\lambda\sqrt{\frac{\log p}{n}}.$$

end for

Output the support of $\hat{\boldsymbol{\beta}}^{(T_0)}$, $\hat{S} = Supp(\hat{\boldsymbol{\beta}}^{(T_0)})$.

Project x_i's to $\hat{S} = Supp(\hat{\boldsymbol{\beta}}^{(T_0)})$, estimate $\hat{w}$ and $\hat{b}$ on these projected samples

Construct the classifier $sgn(\hat{w}^\top(z_{\hat{S}}) - \hat{b})$

B Experimental Implementation on Synthetic Data

We here complement our theory result in Theorem 4 and Theorem 6 by experiments with synthetic data.

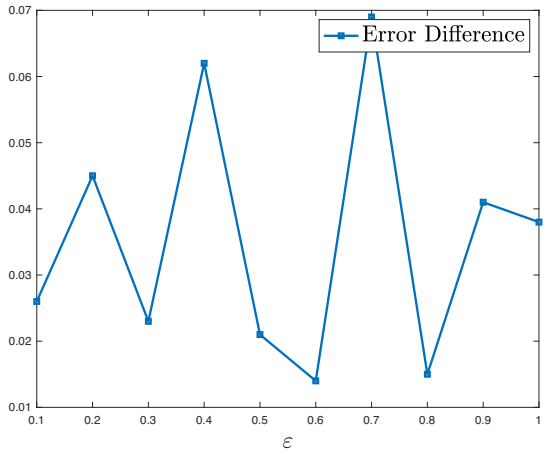

Figure 2: Error Difference vs ε: we use synthetic data described in Theorem 4, where the Error Difference = Adversarial Error when Unlabeled Data from the Same Domain - Adversarial Error when Unlabeled Data from the Shifted Domain. We can see that error difference is positive for all the ε we take, which implies unlabeled data from a shifted domain works even better.

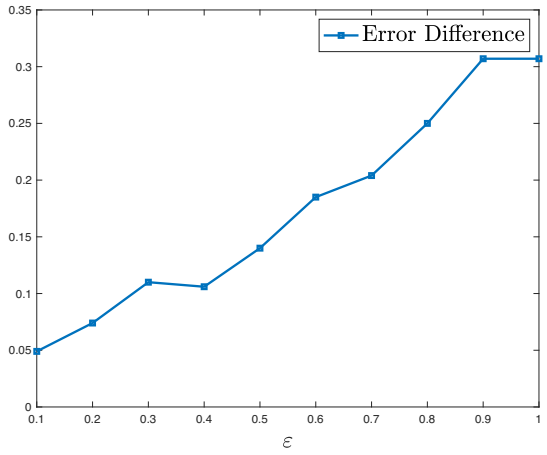

Figure 3: Error Difference vs ε: we use gaussian synthetic data with sparsity structure, where we create 100 dimensional Gaussian data for both original domain and shifted domain, and only the first 10 coordinates matters (mean difference of positive and negative distribution of data from each domain is non-zero for only the first 10 coordinates). The Error Difference = Adversarial Error with Semi-supervised Learning Algorithm - Adversarial Error with Algorithm with Unknown Sparsity. We can see that error difference is positive for all the ε we take, which implies our algorithm with unknown sparsity works better when there is some common sparsity structure.

C Experimental Implementation on Real Data

We use the implementation from Carmon et al. (2019) that can be accessed from `https://github.com/yaircarmon/semisup-adv`

C.1 Experimental setup

We follow the implementation in Carmon et al. (2019) for our experiments:

C.1.1 CIFAR10/CINIC-10

Architecture We use a Wide ResNet 28-10 Zagoruyko & Komodakis (2016) architecture.

Training hyperparemeters We use a batch size of 256 with SGD optimizer (along with Nesterov momentum of 0.1). We use cosine learning rate annealing Loshchilov & Hutter (2016) with initial rate of 0.01 and no restarts. The weight decay parameter is set to 0.0005. We define an epoch to be a pass over 50000 training points. For normal training we run 200 epochs. For adversarial training and stability training we run 100 and 400 epochs respectively.

Data Augmentation We do a 4-pixel random cropping and a random horizontal flip.

Adversarial attacks We use the recommended parameters in Carmon et al. (2019). In test time, we run 40 iterations of projected gradient descent with step-size of 0.01 and do 5 restarts.

Stability training We et noise variance to $\sigma = 0.25$. In test time, we set $N_0 = 100$ and $N = 10000$ with $\alpha = 0.001$.

C.1.2 SVHN

We use similar parameters to CIFAR-10/CINIC-10 except the following.

architecture We use a Wide ResNet 16-8.

Training hyper-parameters The same as CIFAR-10/CINIC-10 except we use batch size of 128 and run $98k$ gradient steps for all models.

Data Augmentation We do not perfom any augmentation.

C.2 Cheap-10 dataset creation pipeline

To create the Cheap-10 dataset, for each CIFAR-10 class, we create 50 related keywords to search for on Bing image search engine. Using an existing image downloding API implementation [4], we were able to download ~ 1000 images for each key-word search. CIFAR-10 dataset is made of 10 classes. For animal classes (bird, cat, deer, dog, frog, horse), our keyboards were made of names of different breeds and different colors or adjectives known to accompany the specific animal. For instance, Parasitic Jaeger, Scottish Fold cat, Pygmy Brocket Deer, Spinone Italiano Dog, Northern Leopard Frog, and Belgian Horse. For other classes (airplane, automobile, ship, truck), we search for different brands or classes. For example, Lockheed Martin F-22 Raptor, Renault automobile, Tanker ship, and Citroën truck. We then downsize images to the original CIFAR-10 size of 32x32. We show example images of the dataset compared to CIFAR-10 images in Fig. 4.

[4] `https://github.com/hardikvasa/google-images-download`

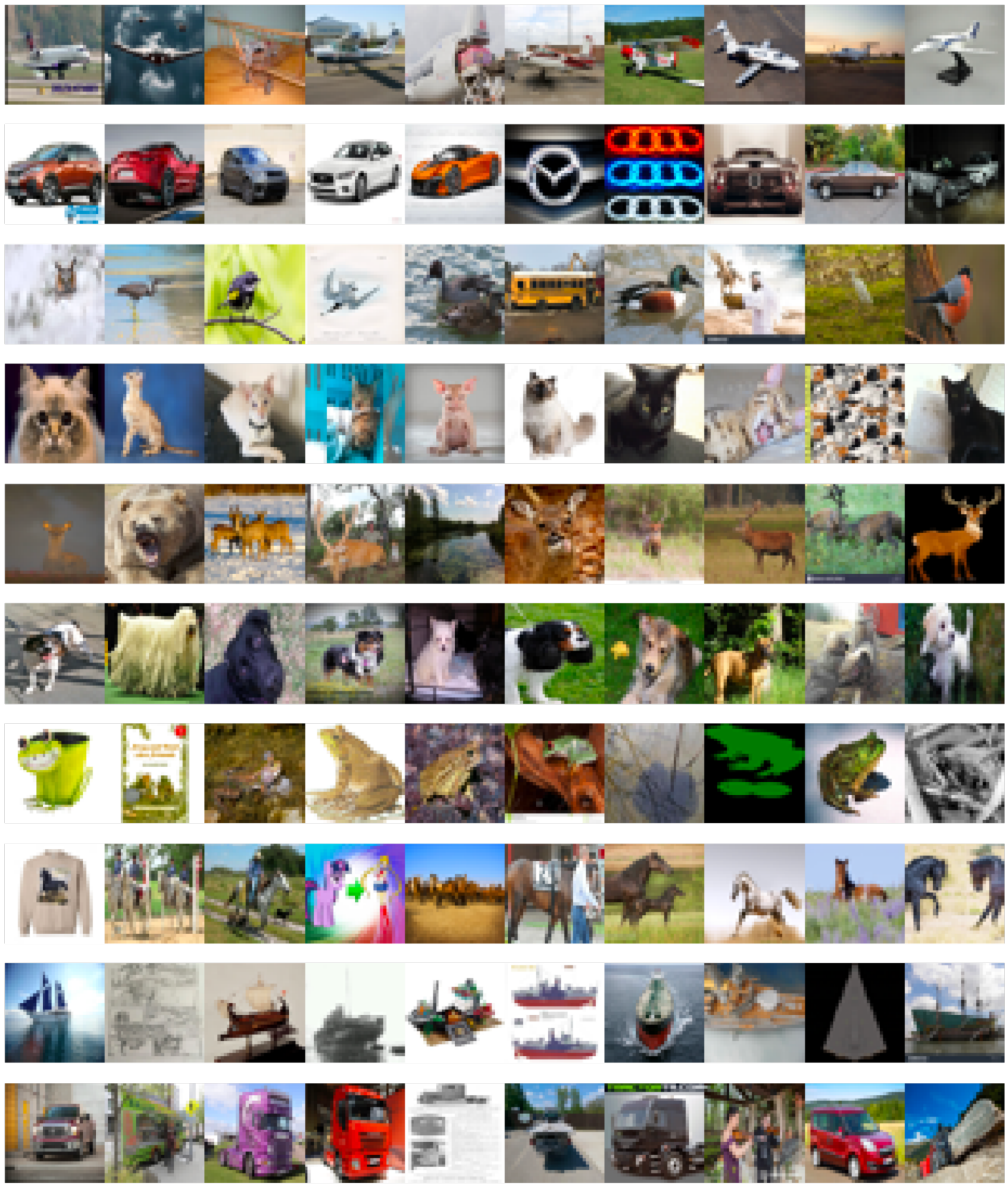

Figure 4: **Cheap-10 examples** Each row shows 10 examples of Cheap-10 dataset.