A PROOFS OF MAIN RESULTS IN SECTION 3

In this section, we provide the detailed proofs of our main theoretical results presented in Section 3.

A.1 PROOF OF THEOREM 3.3

Before proving Theorem 3.3, we first lay out the following lemma regarding the monotonicity of ℓ_p-norms. For a rigorous proof of this lemma, see Raıssouli & Jebril (2010).

Lemma A.1 (Monotonicity of ℓ_p)**.** For any vector $\boldsymbol{x} \in \mathbb{R}^n$, the mapping $p \to \|\boldsymbol{x}\|_p$ is monotonically decreasing for any $p \geq 1$ (including $p = \infty$). That said, $\|\boldsymbol{x}\|_p \leq \|\boldsymbol{x}\|_q$ holds for any $p \geq q \geq 1$.

Now, we are ready to prove Theorem 3.3. In particular, we first include a high-level proof sketch, then present the complete proof after.

Proof Sketch of Theorem 3.3. We start with the spherical Gaussian distribution where $\nu = \gamma_n$. More specifically, we are going to prove that for any $\mathcal{E} \subseteq \mathbb{R}^n$ and $\eta \geq 0$,

$$\gamma_n\big(\mathcal{E}_\eta^{(\ell_p)}\big) \geq \Phi\big(\Phi^{-1}\big(\gamma_n(\mathcal{E})\big) + \eta\big) \text{ holds for } p \geq 2. \tag{A.1}$$

Note that for any vector $\boldsymbol{x} \in \mathbb{R}^n$, the mapping $p \to \|\boldsymbol{x}\|_p$ is monotonically decreasing for any $p \geq 1$, thus we can show that $\mathcal{E}_\eta^{(\ell_q)} \subseteq \mathcal{E}_\eta^{(\ell_p)}$ holds for any $p \geq q \geq 1$. Making use of the standard Gaussian Isoperimetric Inequality (Lemma 3.2), we then immediately obtain

$$\gamma_n\big(\mathcal{E}_\eta^{(\ell_p)}\big) \geq \gamma_n\big(\mathcal{E}_\eta^{(\ell_2)}\big) \geq \Phi\big(\Phi^{-1}\big(\gamma_n(\mathcal{E})\big) + \eta\big), \text{ for any } p \geq 2.$$

Moreover, to prove the concentration bound for general case where ν is the probability measure of $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, we build connections with the spherical Gaussian case by constructing a subset $\mathcal{A} = \{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\theta}) : \boldsymbol{x} \in \mathcal{E}\}$. Based on the affine transformation of Gaussian measure, we then prove:

$$\nu(\mathcal{E}) = \gamma_n(\mathcal{A}) \quad \text{and} \quad \nu(\mathcal{E}_\epsilon^{(\ell_p)}) \geq \gamma_n(\mathcal{A}_\eta^{(\ell_p)}), \quad \text{where } \eta = \epsilon / \|\boldsymbol{\Sigma}^{1/2}\|_p. \tag{A.2}$$

Finally, combining (A.1) and (A.2) completes the proof of Theorem 3.3. □

Complete Proof of Theorem 3.3. To begin with, we consider the special case where the underlying distribution is standard Gaussian ($\nu = \gamma_n$). Specifically, we are going to prove that for any $\mathcal{E} \subseteq \mathbb{R}^n$ and $\eta \geq 0$,

$$\gamma_n\big(\mathcal{E}_\eta^{(\ell_p)}\big) \geq \Phi\big(\Phi^{-1}\big(\gamma_n(\mathcal{E})\big) + \eta\big) \text{ holds for any } p \geq 2. \tag{A.3}$$

Let $p \geq q \geq 1$. According to the definition of ϵ-expansion of a subset and Lemma A.1, we have

$$\begin{aligned} \mathcal{E}_\eta^{(\ell_q)} &= \big\{\boldsymbol{x} \in \mathbb{R}^n : \exists\, \boldsymbol{x}' \in \mathcal{E} \text{ s.t. } \|\boldsymbol{x}' - \boldsymbol{x}\|_q \leq \eta\big\} \\ &\subseteq \big\{\boldsymbol{x} \in \mathbb{R}^n : \exists\, \boldsymbol{x}' \in \mathcal{E} \text{ s.t. } \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \eta\big\} = \mathcal{E}_\eta^{(\ell_p)} \end{aligned} \tag{A.4}$$

where the inclusion is due to the fact that $\|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \|\boldsymbol{x}' - \boldsymbol{x}\|_q$ holds for any $\boldsymbol{x}'$ and $\boldsymbol{x}$. Therefore, by setting $q = 2$ in (A.4), we further obtain that for any $p \geq 2$,

$$\gamma_n\big(\mathcal{E}_\eta^{(\ell_p)}\big) \geq \gamma_n\big(\mathcal{E}_\eta^{(\ell_2)}\big) \geq \Phi\big(\Phi^{-1}\big(\gamma_n(\mathcal{E})\big) + \epsilon\big),$$

where the second inequality is due to the standard Gaussian Isoperimetric Inequality (Lemma 3.2). Thus, we have proven (A.3).

Now we turn to proving the concentration bound for the general Gaussian case. Let $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ be the eigenvalue decomposition of $\boldsymbol{\Sigma}$, where $\mathbf{U} \in \mathbb{R}^{n \times n}$ is an orthonormal matrix and $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times n}$ is a diagonal matrix consisting of all the eigenvalues. Since $\boldsymbol{\Sigma}$ is positive definite, the square root of $\boldsymbol{\Sigma}$ can be expressed as $\boldsymbol{\Sigma}^{1/2} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{U}^\top$. Let $\boldsymbol{\Sigma}^{-1/2} = \mathbf{U}\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^\top$ be the inverse matrix of $\boldsymbol{\Sigma}^{1/2}$.

Construct a subset $\mathcal{A}$ in $\mathbb{R}^n$ such that $\mathcal{A} = \{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\theta}) : \boldsymbol{x} \in \mathcal{E}\}$. Based on the construction of $\mathcal{A}$, we can then prove the following results for any $\mathcal{E} \subseteq \mathbb{R}^n$ and $\epsilon \geq 0$:

$$\nu(\mathcal{E}) = \gamma_n(\mathcal{A}) \quad \text{and} \quad \nu(\mathcal{E}_\epsilon^{(\ell_p)}) \geq \gamma_n(\mathcal{A}_\eta^{(\ell_p)}), \quad \text{where } \eta = \epsilon / \|\boldsymbol{\Sigma}^{1/2}\|_p. \tag{A.5}$$

First, we prove the equality $\nu(\mathcal{E}) = \gamma_n(\mathcal{A})$. Since ν is the probability measure of $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, we have

$$\nu(\mathcal{E}) = \Pr_{\boldsymbol{x}\sim\nu}[\boldsymbol{x} \in \mathcal{E}] = \Pr_{\boldsymbol{x}\sim\nu}[\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\theta}) \in \mathcal{A}] = \Pr_{\boldsymbol{u}\sim\gamma_n}[\boldsymbol{u} \in \mathcal{A}] = \gamma_n(\mathcal{A}), \tag{A.6}$$

where the third inequality is due to the affine transformation of Gaussian random variables.

Next, we prove the remaining inequality in (A.5). By definition, for any $\boldsymbol{u}' \in \mathcal{A}_\eta^{(\ell_p)}$, there exists $\boldsymbol{u} \in \mathcal{A}$ such that $\|\boldsymbol{u}' - \boldsymbol{u}\|_p \leq \eta$. Let $\boldsymbol{x}' = \boldsymbol{\theta} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}'$ and $\boldsymbol{x} = \boldsymbol{\theta} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}$, then we have

$$\big\|\boldsymbol{x}' - \boldsymbol{x}\big\|_p = \big\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{u}' - \boldsymbol{u})\big\|_p \leq \|\boldsymbol{\Sigma}^{1/2}\|_p \cdot \|\boldsymbol{u}' - \boldsymbol{u}\|_p \leq \eta\|\boldsymbol{\Sigma}^{1/2}\|_p \leq \epsilon, \tag{A.7}$$

where the first inequality is due to the definition of induced matrix p-norm and the last inequality holds because $\eta = \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_p$. By the construction of $\mathcal{A}$ and the fact that $\boldsymbol{u} \in \mathcal{A}$, we have $\boldsymbol{x} \in \mathcal{E}$. Combining (A.7), this further implies that for any $\boldsymbol{u}' \in \mathcal{A}_\eta^{(\ell_p)}$, $\boldsymbol{\theta} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{u}' \in \mathcal{E}_\epsilon^{(\ell_p)}$. Thus, we have

$$\nu\big(\mathcal{E}_\epsilon^{(\ell_p)}\big) \geq \nu\big(\boldsymbol{\theta} + \boldsymbol{\Sigma}^{1/2} \cdot \mathcal{A}_\eta^{(\ell_p)}\big) = \Pr_{\boldsymbol{x}\in\nu}\big[\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\theta}) \in \mathcal{A}_\eta^{(\ell_p)}\big] = \gamma_n\big(\mathcal{A}_\eta^{(\ell_p)}\big), \tag{A.8}$$

where $\boldsymbol{\theta} + \boldsymbol{\Sigma}^{1/2} \cdot \mathcal{A}_\eta^{(\ell_p)}$ denotes the transformed subset $\{\boldsymbol{\theta} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{u} : \boldsymbol{u} \in \mathcal{A}_\eta^{(\ell_p)}\}$. Therefore, based on (A.6) and (A.8), we prove the soundness of (A.5).

Finally, combining (A.3) and (A.5) completes the proof of Theorem 3.3. □

A.2 Proof of the Optimality Results in Remark 3.4

Proof. First, we prove the optimality for the spherical Gaussian case, where $\nu = \gamma_n$ and $p > 2$. Let $\mathcal{H} = \mathcal{H}_{\boldsymbol{w},b}$ be a half space with axis-aligned weight vector, that said $\boldsymbol{w} = \boldsymbol{e}_j$ for some $j \in [n]$. Intuitively speaking, the ϵ-expansion of $\mathcal{H}$ with respect to ℓ_p-norm will only happen along the j-th dimension. More rigorously, we are going to prove the following results: for any $\epsilon \geq 0$,

$$\mathcal{H}_\epsilon^{(\ell_p)} = \mathcal{H}_\epsilon^{(\ell_2)} \text{ holds for any } p \geq 1. \tag{A.9}$$

By definition, $\mathcal{H} = \{\boldsymbol{x} \in \mathbb{R}^n : x_j + b \leq 0\}$. For any $\boldsymbol{x} \notin \mathcal{H}$, let $\widehat{\boldsymbol{x}} \in \mathcal{H}$ be the closest point of $\boldsymbol{x}$ in terms of ℓ_p-norm. Since the weight vector $\mathbf{w}$ of $\mathcal{H}$ is axis-aligned, thus $\widehat{\boldsymbol{x}}$ will only differ from $\boldsymbol{x}$ by the j-th element. That said, $\widehat{x}_{j'} = x_{j'}$ for any $j' \neq j$ and $\widehat{x}_j = -b$. Thus for any $p \geq 1$, we have $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_p = \|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2 = x_j + b$. Based on this observation, we further obtain that for any $p \geq 1$,

$$\mathcal{H}_\epsilon^{(\ell_p)} = \{\boldsymbol{x} \in \mathbb{R}^n : x_j + b \leq \epsilon\} = \mathcal{H}_\epsilon^{(\ell_2)},$$

which proves (A.9). According to the Gaussian Isoperimetric Inequality (Lemma 3.2), we obtain

$$\gamma_n\big(\mathcal{H}_\epsilon^{(\ell_p)}\big) = \gamma_n\big(\mathcal{H}_\epsilon^{(\ell_2)}\big) = \Phi(\Phi^{-1}(\gamma_n(\mathcal{H})) + \epsilon).$$

Therefore, combining this with Theorem 3.3, we prove the optimality for the spherical Gaussian case.

Now we turn to prove the non-spherical Gaussian case with $p = 2$. Based on Theorem 3.3, the lower bound is $\Phi(\Phi^{-1}(\nu(\mathcal{E}) + \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_2)$ when $p = 2$. In the following, we are going to prove: if we choose $\mathcal{E} = \mathcal{H}_{\boldsymbol{v}_1,b}$, where $\boldsymbol{v}_1$ is the eigenvector with respect to the largest eigenvalue of $\boldsymbol{\Sigma}$, this lower bound is attained. Similarly to the proof of Theorem 3.3, we construct $\mathcal{A} = \{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\theta}) : \boldsymbol{x} \in \mathcal{E}\}$.

Note that when $\mathcal{E}$ is a half space, the constructed set $\mathcal{A}$ is also a half space. In particular, for the case where $\mathcal{E} = \mathcal{H}_{\boldsymbol{v}_1,b}$, for any $\boldsymbol{u} \in \mathcal{A}$, there exists an $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{u} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\theta})$ and $\boldsymbol{v}_1^\top \boldsymbol{x} + b \leq 0$. This implies that $\boldsymbol{v}_1^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{u} + \boldsymbol{v}_1^\top \boldsymbol{\theta} + b \leq 0$ for any $\boldsymbol{u} \in \mathcal{A}$. Since $\boldsymbol{v}_1$ is the eigenvector of $\boldsymbol{\Sigma}$, we further have that $\mathcal{A}$ is a half space with weight vector $\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_1 = \|\boldsymbol{\Sigma}^{1/2}\|_2 \cdot \boldsymbol{v}_1$.

Note that according to (A.2), as in the proof of Theorem 3.3, for any $\mathcal{E} \subseteq \mathbb{R}^n$, we have

$$\nu(\mathcal{E}) = \gamma_n(\mathcal{A}) \text{ and } \nu\big(\mathcal{E}_\epsilon^{(\ell_2)}\big) \geq \gamma_n\big(\mathcal{A}_\eta^{(\ell_2)}\big), \text{ where } \eta = \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_2.$$

For $\mathcal{E} = \mathcal{H}_{\boldsymbol{v}_1,b}$, based on the explicit formulation of ℓ_2-distance to a half space, we can explicitly compute the η-expansion of $\mathcal{A}$ as

$$\mathcal{A}_\eta^{(\ell_2)} = \{\boldsymbol{u} \in \mathbb{R}^n : \boldsymbol{v}_1^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{u} + \boldsymbol{v}_1^\top \boldsymbol{\theta} + b \leq \eta \cdot \|\boldsymbol{\Sigma}^{1/2}\|_2\}.$$

When we set $\eta = \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_2$, it further implies that

$$\gamma_n\big(\mathcal{A}_\eta^{(\ell_2)}\big) = \Pr_{\boldsymbol{u}\sim\gamma_n}\big[\boldsymbol{v}_1^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{u} + \boldsymbol{v}_1^\top \boldsymbol{\theta} + b \leq \epsilon\big] = \Pr_{\boldsymbol{x}\sim\nu}\big[\boldsymbol{v}_1^\top \boldsymbol{x} + b \leq \epsilon\big] = \nu\big(\mathcal{E}_\epsilon^{(\ell_2)}\big).$$

Finally, according to the optimality of the standard Gaussian Isoperimetric Inequality (Lemma 3.2), this completes the proof. □

B PROOFS OF THEORETICAL RESULTS IN SECTION 4

In this section, we present the proofs to the theoretical results presented in Section 4.

B.1 PROOF OF LEMMA 4.1

Proof of Lemma 4.1. We only consider the case when $\boldsymbol{w}^\top \boldsymbol{x} + b > 0$, because $d_p(\boldsymbol{x}, \mathcal{H}_{\boldsymbol{w},b})$ is zero trivially holds if $\boldsymbol{w}^\top \boldsymbol{x} + b \leq 0$. The problem of finding the ℓ_p-distance from a given point $\boldsymbol{x}$ to a half space $\mathcal{H}_{\boldsymbol{w},b}$ can be formulated as the following constrained optimization problem:

$$\min_{\boldsymbol{z} \in \mathbb{R}^n} \|\boldsymbol{z} - \boldsymbol{x}\|_p, \quad \text{subject to } \boldsymbol{w}^\top \boldsymbol{z} + b \leq 0. \tag{B.1}$$

Let $\widetilde{\boldsymbol{z}} = \boldsymbol{z} - \boldsymbol{x}$, then optimization problem (B.1) is equivalent to

$$\min_{\widetilde{\boldsymbol{z}} \in \mathbb{R}^n} \|\widetilde{\boldsymbol{z}}\|_p, \quad \text{subject to } \boldsymbol{w}^\top \widetilde{\boldsymbol{z}} + \boldsymbol{w}^\top \boldsymbol{x} + b \leq 0. \tag{B.2}$$

According to Hölder's Inequality, for any $\widetilde{\boldsymbol{z}} \in \mathbb{R}^n$ we have

$$-\|\boldsymbol{w}\|_q \cdot \|\widetilde{\boldsymbol{z}}\|_p \leq \boldsymbol{w}^\top \widetilde{\boldsymbol{z}} \leq \|\boldsymbol{w}\|_q \cdot \|\widetilde{\boldsymbol{z}}\|_p,$$

where $1/p + 1/q = 1$. Therefore, for any $\widetilde{\boldsymbol{z}}$ that satisfies the constraint of (B.2), we have

$$\boldsymbol{w}^\top \boldsymbol{x} + b \leq -\boldsymbol{w}^\top \widetilde{\boldsymbol{z}} \leq \|\boldsymbol{w}\|_q \cdot \|\widetilde{\boldsymbol{z}}\|_p. \tag{B.3}$$

Since $\|\boldsymbol{w}\|_2 = 1$, we have $\|\boldsymbol{w}\|_q > 0$, thus (B.3) further suggests $\|\widetilde{\boldsymbol{z}}\|_p \geq (\boldsymbol{w}^\top \boldsymbol{x} + b)/\|\boldsymbol{w}\|_q$.

Up till now, we have proven that the optimal value of (B.1) is lower bounded by $(\boldsymbol{w}^\top \boldsymbol{x} + b)/\|\boldsymbol{w}\|_q$. The remaining task is to show this lower bound can be achieved. To this end, we construct $\widehat{\boldsymbol{z}}$ as

$$\widehat{z}_j = x_j - \frac{\boldsymbol{w}^\top \boldsymbol{x} + b}{\|\boldsymbol{w}\|_q} \cdot \left(\frac{\boldsymbol{w}_j^q}{\sum_{j \in [n]} \boldsymbol{w}_j^q} \right)^{1/p}, \quad \text{for any } j \in [n],$$

where $1/p + 1/q = 1$. We remark that for the extreme case where $p = \infty$, such choice of $\widehat{\boldsymbol{z}}$ can be simplified as $\widehat{\boldsymbol{z}} = \boldsymbol{x} - (\boldsymbol{w}^\top \boldsymbol{x} + b) \cdot \mathrm{sgn}(\boldsymbol{w})/\|\boldsymbol{w}\|_q$, where $\mathrm{sgn}(\cdot)$ denotes the sign function for vectors. According to the construction, it can be verified that

$$\boldsymbol{w}^\top \widehat{\boldsymbol{z}} + b = (\boldsymbol{w}^\top \boldsymbol{x} + b) - \frac{\boldsymbol{w}^\top \boldsymbol{x} + b}{\|\boldsymbol{w}\|_q} \cdot \sum_{j \in [n]} w_j \cdot \left(\frac{\boldsymbol{w}_j^q}{\sum_{j \in [d]} \boldsymbol{w}_j^q} \right)^{1/p} = 0,$$

and $\|\widehat{\boldsymbol{z}} - \boldsymbol{x}\|_p = (\boldsymbol{w}^\top \boldsymbol{x} + b)/\|\boldsymbol{w}\|_q$. □

B.2 PROOF OF THEOREM 4.2

Proof of Theorem 4.2. We write $\mathcal{HS}$ as $\mathcal{HS}(n)$ for simplicity. Let S be a set of size m sampled from μ and $\widehat{\mu}_S$ be the corresponding empirical measure. Note that the VC-dimension of $\mathcal{HS}(n)$ is $n + 1$ (see Mohri et al. (2018)), thus according to the VC inequality, we have

$$\Pr_{S \leftarrow \mu^m} \big[\sup_{\mathcal{E} \in \mathcal{HS}(n)} \big| \widehat{\mu}_S(\mathcal{E}) - \mu(\mathcal{E}) \big| \geq \delta \big] \leq 8 e^{(n+1)\log(m+1) - m\delta^2/32}.$$

In addition, according to Lemma 4.1, the ϵ-expansion of any half space is still a half space. Therefore, we can directly apply Theorem 3.3 in Mahloujifar et al. (2019b) to bound the generalization of concentration with respect to half spaces: for any $\delta \in (0, 1)$, we have

$$\Pr_{S \leftarrow \mu^m} \Big[h_{\widehat{\mu}_S}^{(\ell_p)}(\alpha - \delta, \epsilon, \mathcal{HS}) - \delta \leq h_{\mu}^{(\ell_p)}(\alpha, \epsilon, \mathcal{HS}) \leq h_{\widehat{\mu}_S}^{(\ell_p)}(\alpha + \delta, \epsilon, \mathcal{HS}) + \delta \Big] \geq 1 - 32 e^{(n+1)\log(m+1) - m\delta^2/32}.$$

Finally, assuming the sample size $m \geq c_0 \cdot n \log n / \delta^2$ for some constant c_0 large enough, then there exists positive constant c_1 such that

$$h_{\widehat{\mu}_S}^{(\ell_p)}(\alpha - \delta, \epsilon, \mathcal{HS}) - \delta \leq h_{\mu}^{(\ell_p)}(\alpha, \epsilon, \mathcal{HS}) \leq h_{\widehat{\mu}_S}^{(\ell_p)}(\alpha + \delta, \epsilon, \mathcal{HS}) + \delta$$

holds with probability at least $1 - c_1 \cdot e^{-n \log n}$. □

Algorithm 1: Heuristic Search for Robust Half Space under ℓ_p-distance

Input : a set of samples $\{\boldsymbol{x}_i\}_{i\in[m]}$; strength ϵ (in ℓ_p-norm); risk threshold α; #iterations S.

$\mathbf{Q} \leftarrow$ compute the sample covariance matrix based on $\{\boldsymbol{x}_i\}_{i\in[m]}$;

$\mathcal{V} \leftarrow$ obtain the set of principal components by eigenvalue decomposition on $\mathbf{Q}$;

for $\boldsymbol{v} \in \mathcal{V}$ **do**

  for $s = 1, 2, \ldots, S$ **do**

    $\boldsymbol{w} \leftarrow$ select from $\{\pm\mathrm{pow}(\boldsymbol{v}, s)\}$; // pow() is defined according to (C.1)

    $b \leftarrow \alpha$-quantile of the set $\{-\boldsymbol{w}^\top \boldsymbol{x}_i : i \in [m]\}$;

    $\mathrm{AdvRisk}_\epsilon(\mathcal{H}_{\boldsymbol{w},b}) \leftarrow \sum_{i=1}^m \mathbb{1}(\boldsymbol{w}^\top \boldsymbol{x}_i + b \le \epsilon\|\boldsymbol{w}\|_q)/m$;

  end

end

$(\widehat{\boldsymbol{w}}, \widehat{b}) \leftarrow \mathrm{argmin}_{(\boldsymbol{w},b)} \mathrm{AdvRisk}_\epsilon(\mathcal{H}_{\boldsymbol{w},b})$;

Output : $\mathcal{H}_{\widehat{\boldsymbol{w}},\widehat{b}}$

C Algorithm for Estimating Concentration

To solve the empirical concentration problem (4.3), Algorithm 1 searches for a desirable half space based on the principal components of the empirical dataset and their rotations defined by a power parameter. More specifically, the function $\mathrm{pow}()$ takes a vector $\boldsymbol{v} \in \mathbb{R}^n$ and a positive integer $s \in \mathbb{Z}^+$, and returns the normalized s-th power of $\boldsymbol{v}$ (with sign preserved):

$$\mathrm{pow}(\boldsymbol{v}, s) = \mathrm{sgn}(\boldsymbol{v}) \circ [\mathrm{abs}(\boldsymbol{v})]^s / \|\boldsymbol{v}^s\|_2 = \begin{cases} \boldsymbol{v}^s/\|\boldsymbol{v}^s\|_2, & \text{if } s \text{ is odd;} \\ \mathrm{sgn}(\boldsymbol{v}) \circ \boldsymbol{v}^s/\|\boldsymbol{v}^s\|_2, & \text{otherwise.} \end{cases} \tag{C.1}$$

Note that all the functions used in (C.1) are element-wise operations for vectors, where $\mathrm{sgn}(\boldsymbol{v})$, $\mathrm{abs}(\boldsymbol{v})$, $\boldsymbol{v}^s$ represent the sign, absolute value and the s-th power of $\boldsymbol{v}$ respectively, and the operator $\circ$ denotes the Hardamard product of two vectors.

Connected with the theoretical optimum regarding Gaussian spaces in Remark 3.4, the top principal component corresponds to the optimal choice of $\boldsymbol{w}$ if the perturbation metric is ℓ_2-distance, whereas close-to-axis would be favourable for $\boldsymbol{w}$ when $p > 2$. In addition, as implied by the empirical concentration problem (4.3) and the monotonicity of ℓ_p-mapping (Lemma A.1), the value of $\|\boldsymbol{w}\|_q$ will be more influential in affecting the ϵ-expansion of half space as p grows larger. For example, the ℓ_∞-norm of $\boldsymbol{w}$ can be as large as $\sqrt{n}$ for the worst case (n denotes the input dimension), while $\|\boldsymbol{w}\|_\infty = 1$ if $\boldsymbol{w}$ aligns any axis. By searching through the region between each principal component and the closest axis, the proposed algorithm aims to find the optimal balance between $\|\boldsymbol{w}\|_q$ and the variance of the given data along $\boldsymbol{w}$ that leads to the smallest ϵ-expansion. Although there is no theoretical guarantee that our algorithm will find the optimum to (4.3) for an arbitrary dataset, we empirically show (in Section 5) its efficacy in estimating concentration across various datasets.

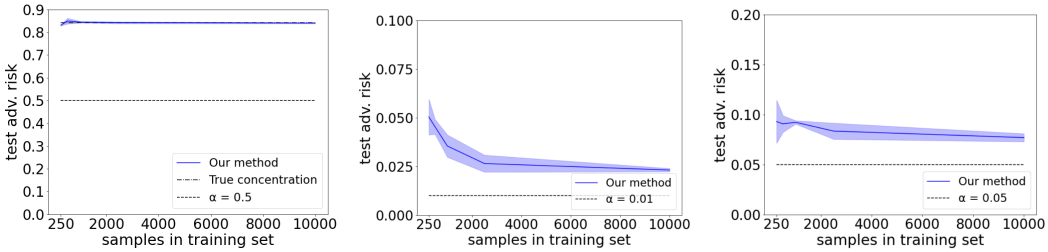
Moreover, our algorithm is efficient in terms of both time and space complexities. Precomputing the principal components requires $O(mn^2 + n^3)$ time and $O(n^2)$ space to store them, where m denotes the samples size and n is the input dimension. For each iteration step, the time complexity of computing $\boldsymbol{w}, b$ and $\mathrm{AdvRisk}_\epsilon(\mathcal{H}_{\boldsymbol{w},b})$ is $O(mn)$, while the space complexity for saving the intermediate variables and the best parameters is $O(m + n)$. With n outer iterations and S inner iterations, the total time complexity is $O(n^3 + mn^2S)$. The total space complexity is $O(n^2 + mn)$, where the extra $O(mn)$ denotes the initial space requirement for saving all the input data. For our experiments, we observe $\mathrm{AdvRisk}_\epsilon(\mathcal{H}_{\boldsymbol{w},b})$ is not sensitive to small increment of the exponent parameter s, thus we choose to increase s in a more aggressive way, which further saves computation.

D Additional Experiments

This section provides experimental results in addition to those presented in Section 5. All our experiments are conducted on a 2.4 GHz 8-Core Intel Core i9 Processor. Table 2 compares our method and the method proposed by Mahloujifar et al. (2019b) on two additional benchmark image

Table 2: Comparisons between our method of estimating concentration with ℓ_∞-norm distance and the method proposed by Mahloujifar et al. (2019b) on Fashion-MNIST and SVHN. Results for the previous method are taken directly from the original paper as a reference.

Dataset	α	ϵ	**Test Risk (%)**		**Test Adv. Risk (%)**	
			Prev. Method	Our Method	Prev. Method	Our Method
Fashion-MNIST	0.05	0.1	5.92 ± 0.85	5.33 ± 0.14	11.56 ± 0.84	6.04 ± 0.13
		0.2	6.00 ± 1.02	5.34 ± 0.14	14.82 ± 0.71	6.82 ± 0.19
		0.3	6.13 ± 0.93	5.24 ± 0.10	17.46 ± 0.53	8.01 ± 0.19
SVHN	0.05	0.01	8.83 ± 0.30	5.23 ± 0.09	10.17 ± 0.29	5.56 ± 0.08



(a) $\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$ (b) MNIST (c) CIFAR-10

Figure 2: The convergence curves of the best possible adversarial risk estimated using our method under various settings as the sample size of the training dataset increases.

datasets, Fashion-MNIST and SVHN. The results again reflect the superiority of our method in producing tighter estimates of concentration.

Figure 2 shows the convergence performance of our algorithm under different experimental settings: $\alpha = 0.5,\ \epsilon = 1$ for the simulated Gaussian dataset, $\alpha = 0.01,\ \epsilon = 0.4$ for MNIST, and $\alpha = 0.05,\ \epsilon = 16/255$ for CIFAR-10. Under these additional settings, the algorithm proposed by Mahloujifar et al. (2019b) either cannot provide meaningful estimates of concentration, or takes a substantial amount of time to run. For instance, our algorithm takes around 2 days to generate the convergence curve on CIFAR-10 ($\alpha = 0.05, \epsilon = 16/255$), whereas the previous method is at least 5 times slower, due to the large number of rectangles T needed. Thus, we only report the convergence curves of our method, where the standard deviations are calculated over 3 repeated trials.