IMPROVED ESTIMATION OF CONCENTRATION UNDER ℓ_p-NORM DISTANCE METRICS USING HALF SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Concentration of measure has been argued to be the fundamental cause of adversarial vulnerability. Mahloujifar et al. (2019b) presented an empirical way to measure the concentration of a data distribution using samples, and employed it to find lower bounds on intrinsic robustness for several benchmark datasets. However, it remains unclear whether these lower bounds are tight enough to provide a useful approximation for the intrinsic robustness of a dataset. To gain a deeper understanding of the concentration of measure phenomenon, we first extend the Gaussian Isoperimetric Inequality to non-spherical Gaussian measures and arbitrary ℓ_p-norms ($p \geq 2$). We leverage these theoretical insights to design a method that uses half-spaces to estimate the concentration of any empirical dataset under ℓ_p-norm distance metrics. Our proposed algorithm is more efficient than Mahloujifar et al. (2019b)'s, and experiments on synthetic datasets and image benchmarks demonstrate that it is able to find much tighter intrinsic robustness bounds. These tighter estimates provide further evidence that concentration of measure is not the main reason for the adversarial vulnerability of state-of-the-art classifiers.

1 INTRODUCTION

Despite achieving exceptional performance in benign settings, modern machine learning models have been shown to be highly vulnerable to inputs, known as *adversarial examples*, crafted with targeted but imperceptible perturbations (Szegedy et al., 2014; Goodfellow et al., 2015). This discovery has prompted a wave of research studies to propose defense mechanisms, including heuristic approaches (Papernot et al., 2016; Mądry et al., 2018; Zhang et al., 2019) and certifiable methods (Wong & Kolter, 2018; Gowal et al., 2019; Cohen et al., 2019). Unfortunately, none of these methods can successfully produce adversarially-robust models, even for classification tasks on toy datasets such as CIFAR-10. To explain the prevalence of adversarial examples, a line of theoretical works (Gilmer et al., 2018; Fawzi et al., 2018; Shafahi et al., 2019; Dohmatob, 2019; Bhagoji et al., 2019) have proven upper bounds on the maximum achievable adversarial robustness by imposing different assumptions on the underlying metric probability space. In particular, Mahloujifar et al. (2019a) generalized the previous results showing that adversarial examples are inevitable as long as the input distributions are concentrated with respect to the perturbation metric. Thus, the question of whether or not natural image distributions are concentrated is highly relevant, as if they are it would rule out any possibility of there being adversarially robust image classifiers.

Recently, Mahloujifar et al. (2019b) proposed an empirical method to measure the concentration of an arbitrary distribution using data samples, then employed it to estimate a lower bound on *intrinsic robustness* (see Definition 2.2 for its formal definition) for several image benchmarks. By demonstrating the gap between the estimated bounds of intrinsic robustness and the robustness performance achieved by the best current models, they further concluded concentration of measure is not the sole reason behind the adversarial vulnerability of existing classifiers for benchmark image distributions. However, due to the heuristic nature of the proposed algorithm, it remains elusive whether the estimates it produces can serve as useful approximations of the underlying intrinsic robustness limits, thus hindering understanding of how much of the actual adversarial risk can be explained by the concentration of measure phenomenon.

In this work, we address this issue by first characterizing the optimum of the actual concentration problem for general Gaussian spaces, then using our theoretical insights to develop an alternative

algorithm for measuring concentration empirically that significantly improves both the accuracy and efficiency of estimates of intrinsic robustness. While we do not demonstrate a specific classifier which achieves this robustness upper bound, our results rule out inherent image distribution concentration as the reason for our inability to date to find adversarially robust models.

Contributions. We generalize the Gaussian Isoperimetric Inequality to non-spherical Gaussian distributions and ℓ_p-norm distance metrics with $p \geq 2$ (including ℓ_∞) (Theorem 3.3). Motivated by the optimal concentration results for special Gaussian spaces (Remark 3.4), we develop a sample-based algorithm to estimate the concentration of measure using half spaces that works for arbitrary distribution and any ℓ_p-norm distance (Section 4). Compared with prior approaches, we empirically demonstrate the significant increase in efficacy of our method under ℓ_∞-norm distance metric (Section 5). Not only does the proposed method converge to its limit with an order of magnitude fewer data (Section 5.2), it also finds a much tighter lower bound of intrinsic robustness for both simulated datasets whose underlying concentration function is analytically derivable and various benchmark image datasets (Section 5.1). In particular, we improve the best current estimated lower bound of intrinsic robustness from approximately 82% to above 93% for CIFAR-10 under ℓ_∞-norm bounded perturbations with $\epsilon = 8/255$. These tighter concentration estimates produced by our algorithm provide strong evidence that concentration of measure should not be considered as the main cause of adversarial vulnerability, at least for the image benchmarks evaluated in our experiments.

Related Work. Several prior works have sought to empirically estimate lower bounds on intrinsic robustness using data samples. The pioneering work of Gilmer et al. (2018) introduced the connection between adversarial examples and the concentration phenomenon for uniform n-spheres, then proposed a simple heuristic to find a half space that expands slowly under Euclidean distance for the MNIST dataset. Our work can be seen as a strict generalization of Gilmer et al. (2018)'s, which applies to arbitrary ℓ_p-norm distance metrics (including ℓ_∞). By characterizing the optimal transport cost between conditional distributions, Bhagoji et al. (2019) estimated a lower bound on the best possible adversarial robustness for several image datasets. However, when applied to adversaries beyond ℓ_2, such as ℓ_∞, the lower bound produced by their method is not informative (that is, it is close to zero). The most relevant previous work is Mahloujifar et al. (2019b), which proposed a general method for measuring concentration using special collections of subsets. Although the optimal value of the considered empirical concentration problem is proven to asymptotically converge to the actual concentration, there is no guarantee that the proposed searching algorithm for solving the empirical problem finds the optimum. Our approach follows the framework introduced by Mahloujifar et al. (2019b)'s, but considers a different collection of subsets for the empirical concentration problem. This not only results in optimality for theoretical Gaussian distributions, but also significantly improves the estimation performance for typical image benchmarks.

Another line of work attempts to provide estimates of intrinsic robustness upper bounds based on generative assumptions. In order to justify the theoretically-derived impossibility results, Fawzi et al. (2018) estimated the smoothness parameters of the state-of-the-art generative models on CIFAR-10 and SVHN datasets, which yield approximated upper bounds on adversarial robustness for any classifiers. Zhang et al. (2020) generalized their results to non-smoothed data manifolds, such as datasets that can be captured by a conditional generative model. However, these methods only work for simulated generative distributions, which may deviate from the actual distributions they are intended to understand.

Notation. For any $n \in \mathbb{Z}^+$, denote by $[n]$ the set $\{1, 2, \ldots, n\}$. Lowercase boldface letters denote vectors and uppercase boldface letters represent matrices. For any vector $\boldsymbol{x}$ and $p \in [1, \infty)$, let x_j, $\|\boldsymbol{x}\|_p$ and $\|\boldsymbol{x}\|_\infty$ be the j-th element, the ℓ_p-norm and the ℓ_∞-norm of $\boldsymbol{x}$. For any matrix $\mathbf{A}$, $\mathbf{B}$ is said to be a square root of $\mathbf{A}$ if $\mathbf{A} = \mathbf{B}\mathbf{B}$, and the induced matrix p-norm of $\mathbf{A}$ is defined as $\|\mathbf{A}\|_p = \sup_{\boldsymbol{x} \neq \mathbf{0}} \{\|\mathbf{A}\boldsymbol{x}\|_p / \|\boldsymbol{x}\|_p\}$. Denote by $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ the Gaussian distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$. Let γ_n be the probability measure of $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, where $\mathbf{I}_n$ denotes the identity matrix. Let $\Phi(\cdot)$ be the cumulative distribution function of $\mathcal{N}(0, 1)$ and $\Phi^{-1}(\cdot)$ be its inverse. For any set $\mathcal{A}$, let $\mathsf{pow}(\mathcal{A})$ and $\mathbb{1}_{\mathcal{A}}(\cdot)$ be all measurable subsets and the indicator function of $\mathcal{A}$. Given a metric probability space $(\mathcal{X}, \mu, \Delta)$, define the empirical measure with respect to a sample set $\{\boldsymbol{x}_i\}_{i \in [m]}$ as $\widehat{\mu}_m(\mathcal{A}) = \frac{1}{m} \sum_{i \in [m]} \mathbb{1}_{\mathcal{A}}(\boldsymbol{x}_i)$, $\forall \mathcal{A} \in \mathsf{pow}(\mathcal{X})$. Let $\mathcal{B}(\boldsymbol{x}, \epsilon, \Delta) = \{\boldsymbol{x}' \in \mathcal{X} : \Delta(\boldsymbol{x}', \boldsymbol{x}) \leq \epsilon\}$ be the ball around $\boldsymbol{x}$ with ϵ radius. Define the ϵ-expansion of $\mathcal{A}$ as $\mathcal{A}_\epsilon^{(\Delta)} = \{\boldsymbol{x} \in \mathcal{X} : \exists \boldsymbol{x}' \in \mathcal{B}(\boldsymbol{x}, \epsilon, \Delta) \cap \mathcal{A}\}$.

2 PRELIMINARIES

In this section, we introduce the problem of measuring concentration and its connection to adversarial robustness. Consider a metric probability space of instances $(\mathcal{X}, \mu, \Delta)$. Given parameters $\epsilon \geq 0$ and $\alpha > 0$, the concentration of measure problem[1] can be cast as the following optimization problem:

$$\underset{\mathcal{E} \in \mathsf{Pow}(\mathcal{X})}{\text{minimize}} \ \mu\big(\mathcal{E}_\epsilon^{(\Delta)}\big) \quad \text{subject to} \ \mu(\mathcal{E}) \geq \alpha. \tag{2.1}$$

We focus on the case where Δ is some ℓ_p-norm distance metric (including ℓ_∞) in this work.

Concentration of measure has been shown to be closely related to adversarial examples (Gilmer et al., 2018; Fawzi et al., 2018; Mahloujifar et al., 2019a). In particular, one can prove that for a given robust learning problem, if the input distribution is concentrated with respect to the perturbation metric, no adversarially robust model exists. The concentration parameter (which corresponds to the optimal value of optimization problem (2.1)) determines an inherent upper bound on the maximum adversarial robustness that *any* model can achieve for the given problem.

To explain the connection between concentration of measure and robust learning in a more formal way, we lay out the definition of *adversarial risk* that we work with. We draw this definition from several previous works, including Gilmer et al. (2018); Bubeck et al. (2019); Mahloujifar et al. (2019a;b).[2]

Definition 2.1 (Adversarial Risk)**.** Let $(\mathcal{X}, \mu, \Delta)$ be the input metric probability space. Assume f^* is the underlying ground-truth classifier that gives labels to any input. Given classifier f and $\epsilon \geq 0$, the *adversarial risk* of f with respect to ϵ-perturbations measured by Δ is defined as:

$$\mathrm{AdvRisk}_\epsilon(f) = \Pr_{\boldsymbol{x} \sim \mu}\big[\exists\, \boldsymbol{x}' \in \mathcal{B}(\boldsymbol{x}, \epsilon, \Delta) \text{ s.t. } f(\boldsymbol{x}') \neq f^*(\boldsymbol{x}')\big].$$

Correspondingly, we define the *adversarial robustness* of f as $\mathrm{AdvRob}_\epsilon(f) = 1 - \mathrm{AdvRisk}_\epsilon(f)$.

When $\epsilon = 0$, adversarial risk degenerates to standard risk. In other words, it holds for any f that $\mathrm{AdvRisk}_0(f) = \mathrm{Risk}(f) := \Pr_{\boldsymbol{x} \sim \mu}[f(\boldsymbol{x}) \neq f^*(\boldsymbol{x})]$. We remark that this definition assumes the existence of an underlying ground-truth labeling function, which does not apply to the agnostic setting where inputs can have non-deterministic labels.

Initially introduced in Mahloujifar et al. (2019b), intrinsic robustness captures the maximum adversarial robustness that can be achieved by any imperfect classifier for a robust classification problem.

Definition 2.2 (Intrinsic Robustness)**.** Consider the same setting as in Definition 2.1. For any $\alpha > 0$, let $\mathcal{F}_\alpha = \{f : \mathrm{Risk}(f) \geq \alpha\}$ be the set of imperfect classifiers whose risk is at least α. Then the *intrinsic robustness* of the given robust classification problem with respect to $\mathcal{F}_\alpha$ is defined as:

$$\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_\alpha) = 1 - \inf_{f \in \mathcal{F}_\alpha} \big\{\mathrm{AdvRisk}_\epsilon(f)\big\} = \sup_{f \in \mathcal{F}_\alpha} \{\mathrm{AdvRob}_\epsilon(f)\}.$$

It is worth noting that the value of $\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$ is only determined by the underlying input data distribution, the perturbation set and the risk threshold parameter α, which is independent of the model class one would choose for learning.

By relating the robustness of a classifier to the ϵ-expansion of its induced error region, the following lemma, proved in Mahloujifar et al. (2019a), establishes a fundamental connection between the concentration of measure and the intrinsic robustness one can hope for a robust classification problem.

Lemma 2.3. Consider the same setting as in Definition 2.2. Let $h_\mu^{(\Delta)}(\alpha, \epsilon)$ be the concentration function that captures the optimal value of the concentration of measure problem (2.1):

$$h_\mu^{(\Delta)}(\alpha, \epsilon) = \inf\big\{\mu\big(\mathcal{E}_\epsilon^{(\Delta)}\big) : \mathcal{E} \in \mathsf{pow}(\mathcal{X}) \text{ and } \mu(\mathcal{E}) \geq \alpha\big\}.$$

Then, $\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_\alpha) = 1 - h_\mu^{(\Delta)}(\alpha, \epsilon)$ holds for any $\alpha > 0$ and $\epsilon \geq 0$.

[1] The standard notion of concentration of measure (Talagrand, 1995) corresponds to the case where $\alpha = 0.5$.

[2] Other related definitions, such as the one used in most empirical works for robustness evaluation, are equivalent to this, as long as small perturbations preserve the ground truth. See Diochnos et al. (2018) for a detailed comparison of these and other definitions of adversarial robustness.

Lemma 2.3 suggests that one can characterize the intrinsic robustness limit for a robust classification problem by measuring the concentration of the input data with respect to the perturbation metric.

In this paper, we aim to understand and empirically estimate the intrinsic robustness limit for typical robust classification tasks by measuring concentration. It is worth noting that solving the concentration problem (2.1) itself only shows the existence of an error region $\mathcal{E}$ whose ϵ-expansion has certain (small) measure. This further implies the possibility of existing an optimally robust classifier (with risk at least α), whose robustness matches the intrinsic robustness limit $\overline{\mathrm{AdvRob}}_\epsilon(\mathcal{F}_\alpha)$. However, actually finding such optimal classifier using a learning algorithm might be a much more challenging task, which is beyond the scope of this work.

3 Generalizing the Gaussian Isoperimetric Inequality

Before proceeding to introduce the proposed methodology for solving the concentration of measure problem, we first present our main theoretical results of generalizing the Gaussian Isoperimetric Inequality. This theoretical result largely motivates our method.

To begin with, we introduce the Gaussian Ispoperimetric Inequality (Sudakov & Tsirelson, 1974; Borell, 1975). It characterizes the optimum of the concentration problem (2.1) with respect to standard Gaussian distribution and ℓ_2-distance, where half spaces are proven to be the optimal sets.

Definition 3.1 (Half Space)**.** Let $\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Without loss of generality, assume $\|\boldsymbol{w}\|_2 = 1$. An n-dimensional *half space* with parameters $\boldsymbol{w}$ and b is defined as:

$$\mathcal{H}_{\boldsymbol{w},b} = \{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{w}^\top \boldsymbol{z} + b \leq 0\}.$$

Lemma 3.2 (Gaussian Isoperimetric Inequality)**.** Consider the standard Gaussian space $(\mathbb{R}^n, \gamma_n)$ with ℓ_2-distance. Let $\mathcal{E} \subseteq \mathbb{R}^n$ and $\mathcal{H}$ be a half space such that $\gamma_n(\mathcal{E}) = \gamma_n(\mathcal{H})$, then for any $\epsilon \geq 0$,

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

The proof of the Gaussian Isoperimetric Inequality can be found in Ledoux (1996).

Lemma 3.2 implies the concentration function of $(\mathbb{R}^n, \gamma_n, \|\cdot\|_2)$ is $h_{\gamma_n}^{(\ell_2)}(\alpha, \epsilon) = \Phi\big(\Phi^{-1}(\alpha) + \epsilon\big)$, but only applies when the underlying distribution is a spherical Gaussian and the metric function is ℓ_2-distance. Thus, it only gives a concentration function for estimating the the intrinsic robustness limit in a very restrictive setting. To understand the concentration of measure for more general problems, we prove the following theorem that extends the standard Gaussian Isoperimetric Inequality (Lemma 3.2) to non-spherical Gaussian measure and general ℓ_p-norm distance metrics for any $p \geq 2$.

Theorem 3.3 (Generalized Gaussian Isoperimetric Inequality)**.** Let ν be the probability measure of $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}$ is a positive definite matrix in $\mathbb{R}^{n\times n}$. Consider the probability space $(\mathbb{R}^n, \nu)$ with ℓ_p-norm distance, where $p \geq 2$ (including ℓ_∞). For any subset $\mathcal{E} \subseteq \mathbb{R}^n$ and $\epsilon \geq 0$,

$$\nu\big(\mathcal{E}_\epsilon^{(\ell_p)}\big) \geq \Phi\big(\Phi^{-1}\big(\nu(\mathcal{E})\big) + \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_p\big), \tag{3.1}$$

where $\boldsymbol{\Sigma}^{1/2}$ is the square root of $\boldsymbol{\Sigma}$, and $\|\boldsymbol{\Sigma}^{1/2}\|_p$ denotes the induced matrix p-norm of $\boldsymbol{\Sigma}^{1/2}$.

Remark 3.4. Theorem 3.3 suggests that for general Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and any ℓ_p-norm distance ($p \geq 2$), the corresponding concentration function is lower bounded by $\Phi(\Phi^{-1}(\alpha) + \epsilon/\|\boldsymbol{\Sigma}^{1/2}\|_p)$. Due to the NP-hardness of approximating the matrix p-norm (Hendrickx & Olshevsky, 2010), it is generally hard to infer whether the equality of (3.1) can be attained or not. However, for specific special Gaussian spaces, we can derive optimal subsets that achieve the lower bound. In particular, for the case where $\boldsymbol{\Sigma} = \mathbf{I}_n$ and $p > 2$, the optimum is attained when $\mathcal{E}$ is a half space with axis-aligned weight vector (that is, $\boldsymbol{w} = \mathbf{e}_j$ for some $j \in [n]$). For the case where $\boldsymbol{\Sigma} \neq \mathbf{I}_n$ and $p = 2$, the optimal solution is a half space $\mathcal{H}_{\boldsymbol{v}_1,b}$, where $\boldsymbol{v}_1$ is the eigenvector with respect to the largest eigenvalue of $\boldsymbol{\Sigma}$. The proofs of these optimality results are provided in Appendix A.2.

4 Empirically Measuring Concentration using Half Spaces

Recall that the primary goal of this paper is to measure the concentration of an arbitrary distribution. However, for typical classification problems, we might not know the density function of the underlying

distribution μ, but instead we usually have access to a finite set of m instances $\{\boldsymbol{x}_i\}_{i\in[m]}$ sampled from μ. Following Mahloujifar et al. (2019b), we consider the following empirical counterpart of the actual concentration problem (2.1):

$$\underset{\mathcal{E}\in\mathcal{G}}{\text{minimize}}\ \widehat{\mu}_m\big(\mathcal{E}_\epsilon^{(\Delta)}\big) \quad \text{subject to } \widehat{\mu}_m(\mathcal{E}) \geq \alpha, \tag{4.1}$$

where $\widehat{\mu}_m$ is the empirical measure based on $\{\boldsymbol{x}_i\}_{i\in[m]}$ and $\mathcal{G} \subseteq \mathsf{pow}(\mathcal{X})$ denotes a particular collection of subsets. Mahloujifar et al. (2019b) proposed the complement of union of T hyperrectangles as $\mathcal{G}$ for ℓ_∞ and the union of T balls for ℓ_2. They proved that if one increases the complexity parameter T and the sample size m together in a careful way, the optimal value of the empirical concentration problem (4.1) converges to the actual concentration asymptotically. However, it is unclear how quickly it converges and how well the proposed heuristic algorithm in Mahloujifar et al. (2019b) finds the optimum of (4.1).

In this work, we argue that the set of half spaces is a superior choice for $\mathcal{G}$ with respect to any ℓ_p-norm distance. Apart from achieving the optimality for certain Gaussian spaces as discussed in Remark 3.4, estimating concentration using half spaces has several other advantages including the closed-form solution of ℓ_p-distance to half-space (Lemma 4.1) and its small sample complexity requirement for generalization (Theorem 4.2).

To be more specific, we focus on the following optimization problem based on the empirical measure $\widehat{\mu}_m$ and the collection of half spaces $\mathcal{HS}(n)$:

$$\underset{\mathcal{E}\in\mathcal{HS}(n)}{\text{minimize}}\ \widehat{\mu}_m\big(\mathcal{E}_\epsilon^{(\ell_p)}\big) \quad \text{subject to } \widehat{\mu}_m(\mathcal{E}) \geq \alpha, \tag{4.2}$$

where $\mathcal{HS}(n) = \{\mathcal{H}_{\boldsymbol{w},b} : \boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}, \text{ and } \|\boldsymbol{w}\|_2 = 1\}$ is the set of all half spaces in $\mathbb{R}^n$.

The following lemma, proven in Appendix B.1, characterizes the closed-form solution of the ℓ_p-norm distance between a point $\boldsymbol{x}$ and a half space. Such a formulation enables an exact computation of the empirical measure with respect to the ϵ-expansion of any half space.

Lemma 4.1 (ℓ_p-Distance to Half Space)**.** Let $\mathcal{H}_{\boldsymbol{w},b} \in \mathcal{HS}(n)$ be an n-dimensional half space. For any vector $\boldsymbol{x} \in \mathbb{R}^n$, the ℓ_p-norm distance ($p \geq 1$) from $\boldsymbol{x}$ to $\mathcal{H}_{\boldsymbol{w},b}$ is:

$$d_p(\boldsymbol{x}, \mathcal{H}_{\boldsymbol{w},b}) = \begin{cases} 0, & \boldsymbol{w}^\top \boldsymbol{x} + b \leq 0; \\ (\boldsymbol{w}^\top \boldsymbol{x} + b)/\|\boldsymbol{w}\|_q, & \text{otherwise.} \end{cases}$$

Here, q is a real number that satisfies $1/p + 1/q = 1$.

Lemma 4.1 implies that the ϵ-expansion of any half space with respect to the ℓ_p-norm is still a half space. Since the VC-dimensions of both the set of half spaces and its expansion are bounded, we can thus apply the generalization theorem in Mahloujifar et al. (2019b), which yields the following theorem, proved in Appendix B.2, that characterizes the generalization of concentration with respect to the collection of half spaces.

Theorem 4.2 (Generalization of Concentration of Half Spaces)**.** Consider the metric probability space, $(\mathcal{X}, \mu, \|\cdot\|_p)$, where $\mathcal{X} \subseteq \mathbb{R}^n$ and $p \geq 1$. Let $\{\boldsymbol{x}_i\}_{i\in[m]}$ be a set of m instances sampled from μ, and let $\widehat{\mu}_m$ be the corresponding empirical measure. Define the concentration functions regarding the collection of half spaces $\mathcal{HS}(n)$ with respect to μ as:

$$h_\mu^{(\ell_p)}\big(\alpha, \epsilon, \mathcal{HS}(n)\big) = \inf_{\mathcal{E}\in\mathcal{HS}(n)} \big\{\mu\big(\mathcal{E}_\epsilon^{(\ell_p)}\big) \colon \mu(\mathcal{E}) \geq \alpha\big\},$$

and let $h_{\widehat{\mu}_m}^{(\ell_p)}(\alpha, \epsilon, \mathcal{HS}(n))$ be its empirical counterpart with respect to $\widehat{\mu}_m$. For any $\delta \in (0, 1)$, there exists constants c_0 and c_1 such that with probability at least $1 - c_0 \cdot e^{-n \log n}$,

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

holds, provided that the sample size $m \geq c_1 \cdot n \log n/\delta^2$.

Remark 4.3. Theorem 4.2 suggests that for the concentration of measure problem with respect to half spaces, in order to achieve δ estimation error with high probability, it requires $\Omega(n \log(n)/\delta^2)$ number of samples. Compared with Mahloujifar et al. (2019b), our method using half spaces

requires fewer samples in theory to achieve the same estimation error.[3] For standard Gaussian inputs, the empirical concentration with respect to half spaces is guaranteed to converge to the actual concentration as in (2.1), i.e., $\lim_{m\to\infty} h_{\widehat{\mu}_m}^{(\ell_p)}\big(\alpha, \epsilon, \mathcal{HS}(n)\big) = h_{\mu}^{(\ell_p)}\big(\alpha, \epsilon\big)$; whereas for distributions that are not Gaussian, there might exist a gap. However, this gap of empirical and actual concentration is shown to be uniformly small across various data distributions, as will be discussed in Section 5.2.

Based on Lemma 4.1, estimating the empirical concentration using half spaces as defined in (4.2) is equivalent to solving the following constrained optimization problem:

$$\begin{aligned} \underset{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}}{\text{minimize}} \quad & \sum_{i\in[m]} \mathbb{1}\{\boldsymbol{w}^\top \boldsymbol{x}_i + b \le \epsilon \|\boldsymbol{w}\|_q\} \\ \text{subject to} \quad & \frac{1}{m}\sum_{i\in[m]} \mathbb{1}\{\boldsymbol{w}^\top \boldsymbol{x}_i + b \le 0\} \ge \alpha \ \text{ and } \ \|\boldsymbol{w}\|_2 = 1. \end{aligned} \tag{4.3}$$

The optimal solution to (4.3) would be a half space $\mathcal{H}_{\boldsymbol{w},b}$ that satisfies the following two properties: (1) approximately α-fraction of data is covered by $\mathcal{H}_{\boldsymbol{w},b}$, and (2) most of the remaining data points are at least ϵ-away from $\mathcal{H}_{\boldsymbol{w},b}$ under ℓ_p-norm distance metric.

Note that we can always set b to be the α-quantile of the projections $\{-\boldsymbol{w}^\top \boldsymbol{x}_i : i \in [m]\}$ to satisfy the first property. In addition, to satisfy the second condition, inspired by the special case optimality results in Remark 3.4, we propose to search for a weight vector $\boldsymbol{w}$ such that both the ℓ_q-norm of $\boldsymbol{w}$ is small and the variation of the given sample set along the direction of $\boldsymbol{w}$ is large. These searching criteria guarantee that the given dataset $\{\boldsymbol{x}_i\}_{i\in[m]}$, when projected onto $\boldsymbol{w}$ then normalized by $\|\boldsymbol{w}\|_q$, will have a large variance, which implies the second property.

We propose a heuristic algorithm to search for the desirable half space according to the aforementioned criteria (for the pseudocode and a detailed analysis of the algorithm, see Appendix C). It first conducts a principal component analysis with respect to the given empirical dataset, then iterates through all the principal components raised to an arbitrary power with normalization as candidate choices of the weight vector $\boldsymbol{w}$. Finally, based on the optimal choice of b, the algorithm outputs the best $\mathcal{H}_{\boldsymbol{w},b}$ that achieves the smallest empirical measure with respect to the ϵ-expansion of $\mathcal{H}_{\boldsymbol{w},b}$. As we see in the experiments in the next section, this algorithm is able to find near-optimal solutions to the concentration problem for various datasets.

5 Experiments

In this section, we evaluate our empirical method for estimating concentration under ℓ_∞-norm distance and comparing its performance to that of the method proposed by Mahloujifar et al. (2019b). We first demonstrate that the estimate produced by our algorithm is very close to the actual concentration for a spherical Gaussian distribution, and that our method is able to find much tighter bounds on the best possible adversarial risk for both MNIST and CIFAR-10 datasets. We then compare the convergence rates, and show that our method converges with substantially less data. Note that while we only provide results for the most widely-used ℓ_∞-norm perturbation metric adopted in the existing adversarial examples literature, our algorithm and experiments can be applied to any other ℓ_p-norm.

5.1 Estimation Accuracy

First, we evaluate the performance of our algorithm under ℓ_∞-norm distance metric on a generated synthetic dataset consisting of 30,000 samples from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$. Since the proposed method follows from the analytical results of concentration of multivariate Gaussian distributions, we expect results produced by our empirical method to closely approach the analytical concentration on this simulated Gaussian dataset. We initially consider the case where $\epsilon = 1.0$ and $\alpha = 0.5$ for the actual concentration problem, requiring that the feasible set contains at least half of the data samples, and the adversary can perturb each entry by precisely the standard deviation of the underlying distribution. Our algorithm is able to produce a half space whose ϵ-expansion has mean empirical measure 84.18%

[3]The proposed estimators for ℓ_∞ and ℓ_2 in Mahloujifar et al. (2019b) require $\Omega(nT\log(n)\log(T)/\delta^2)$ samples to achieve δ approximation, where T is a predefined number of hyperrectangles or balls.

Table 1: Comparisons between our method of estimating concentration with ℓ_∞-norm distance and the method proposed by Mahloujifar et al. (2019b) for different settings. For $\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$ with $\alpha = 0.5$ and $\epsilon = 1.0$, the previous method is unable to produce nontrivial estimate. Results on MNIST and CIFAR-10 for the previous method are taken directly from the original paper as a reference.

Dataset	$\boldsymbol{\alpha}$	$\boldsymbol{\epsilon}$	**Test Risk (%)**		**Test Adv. Risk (%)**	
			Prev. Method	Our Method	Prev. Method	Our Method
$\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$	0.05	1.0	6.21 ± 0.44	5.20 ± 0.16	89.75 ± 0.80	26.37 ± 0.17
	0.5	1.0	-	49.98 ± 0.25	-	84.18 ± 0.11
MNIST	0.01	0.1	1.23 ± 0.12	1.22 ± 0.05	3.64 ± 0.30	1.35 ± 0.06
		0.2	1.11 ± 0.10	1.24 ± 0.05	5.89 ± 0.44	1.52 ± 0.06
		0.3	1.15 ± 0.13	1.25 ± 0.04	7.24 ± 0.38	1.75 ± 0.05
		0.4	1.21 ± 0.09	1.27 ± 0.05	9.92 ± 0.60	1.98 ± 0.08
CIFAR-10	0.05	$2/255$	5.72 ± 0.25	5.14 ± 0.13	8.13 ± 0.26	5.28 ± 0.12
		$4/255$	6.05 ± 0.40	5.22 ± 0.20	13.66 ± 0.33	5.68 ± 0.21
		$8/255$	5.94 ± 0.34	5.22 ± 0.16	18.13 ± 0.30	6.28 ± 0.13
		$16/255$	5.28 ± 0.23	5.19 ± 0.08	28.83 ± 0.46	7.34 ± 0.15

over 5 repeated trials. According to Theorem 3.3 and Remark 3.4, the optimal value of the considered concentration problem is 84.13%. This implies that our method performs very well when the underlying distribution is Gaussian, while in stark contrast, the method by Mahloujifar et al. (2019b) is not able to find a region whose expansion has measure less than 1 on the same simulated set. In addition, we consider another setting for this dataset where $\epsilon = 1.0$ is set the same and $\alpha = 0.05$ is set to be much smaller. Similarly, we observe that our method significantly outperforms Mahloujifar et al. (2019b)'s in terms of the estimation accuracy (see Table 1 for the detailed comparison results).

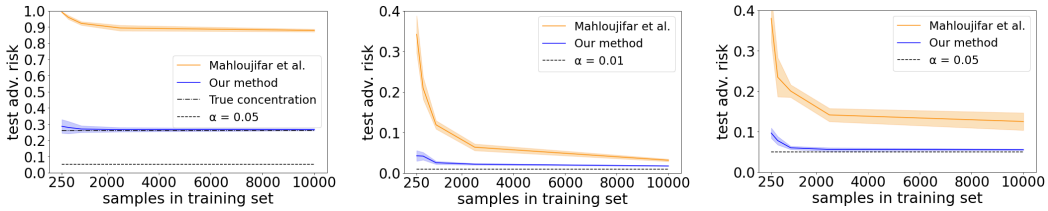
Next, we evaluate our method on typical image benchmarks. We set the values of α and ϵ to be the same as in Mahloujifar et al. (2019b) for the ℓ_∞ case, specifically $\alpha = 0.01,\ \epsilon \in \{0.1, 0.2, 0.3, 0.4\}$ for MNIST, and $\alpha = 0.05,\ \epsilon \in \{2/255, 4/255, 8/255, 16/255\}$ for CIFAR-10. These α values were selected because they roughly represent the standard error of the state-of-the-art classifiers.

Table 1 demonstrates the risk and adversarial risk with respect to the best produced subsets using both methods, computed on a separate test dataset. In our context of measuring concentration, risk refers to the empirical measure of the produced subset, while adversarial risk corresponds to the empirical measure of its ϵ-expansion. We use a $50/50$ train-test split over the whole dataset to perform our evaluation, and determine the best exponent of each principal component based on a brute-force search. Though our method is deterministic for a given pair of training and testing sets, we account for the variance of our method over different train-test splits by repeating our experiments 5 times and reporting the mean and standard deviation of the results for each (α, ϵ). It is worth noting that the randomness of the previous method is derived not only from the selection of the training and test sets, but also from the inherent randomness of the employed k-means algorithm.

We observe from Table 1 that in every case, the estimated adversarial risk is significantly lower for our method than for the one found by Mahloujifar et al. (2019b)'s. Since both methods restrict the search space to some special collection of subsets, these estimates can be viewed as valid empirical upper bounds of the actual concentration as defined in (2.1). Therefore, the fact that our results are significantly lower indicates that our algorithm is able to produce estimates that are much closer to the optimum of the targeted problem. In addition, when translated to adversarial robustness, these tighter estimates prove the existence of a rather robust classifier[4] that has risk at least α, which further suggests that the underlying intrinsic robustness limit of each of these image benchmarks is actually much higher than previously thought.

For example, the best classifier produced by Mahloujifar et al. (2019b) has 18.1% adversarial risk under ℓ_∞-perturbations bounded by $\epsilon = 8/255$ on CIFAR-10. However, our results demonstrate

[4]Based on the ground-truth f^* and the returned set $\mathcal{E}$ of our algorithm, this classifier can be simply constructed by setting $f(\boldsymbol{x}) = f^*(\boldsymbol{x})$ for $\boldsymbol{x} \notin \mathcal{E}$ and $f(\boldsymbol{x}) \neq f^*(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{E}$. Without knowing the ground-truth, we note that such classifier may or may not be learnable. The learnability of such f is beyond the scope of this paper.



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

Figure 1: The convergence curves of the best possible adversarial risk estimated using our method and the method proposed by Mahloujifar et al. (2019b) as the number of training samples grows.

that the adversarial risk of the best possible robust classifier can be as low as 6.3% given the same risk constraint, indicating the underlying intrinsic robustness to be above 93.7%. As the intrinsic robustness limits are shown to be very close to the trivial upper bound $1 - \alpha$ across all the settings, our results reveal that the concentration of measure phenomenon is not an important factor that causes the adversarial vulnerability of existing classifiers on these image benchmarks.

5.2 Convergence

Figure 5.2 shows the convergence rate of our method compared with that of the method proposed by Mahloujifar et al. (2019b) under ℓ_∞-distance for Gaussian data from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{784})$ $(\alpha = 0.05,\ \epsilon = 1)$, as well as for MNIST $(\alpha = 0.01,\ \epsilon = 0.1)$ and CIFAR-10 $(\alpha = 0.05,\ \epsilon = 2/255)$. We observe similar trends for other settings, thus we defer these results to Appendix D. For each graph, the horizontal x-axis represents the size of the dataset used to train the estimator, and the vertical y-axis shows the concentration bounds estimated for a separate test set, which is of size $30,000$ for each case. We generate the means and standard deviations for these convergence curves by repeating both methods 5 times for different randomly-selected training and test tests. For the proposed algorithm in Mahloujifar et al. (2019b), we tune the number of the hyperrectangles T for the optimal performance based on the empirically-observed adversarial risk.

For the simulated Gaussian datasets, we include a horizontal line at $y = 0.2595$ to represent the true concentration of the underlying distribution, derived from Theorem 3.3 and Remark 3.4. This allows us to more accurately assess the convergence of our method, as it is the only case for which we know the optimal value that our empirical estimates should be converging to. We see that our estimates approach this line very quickly, coming within 0.01 of the true value given only about 1,000 samples.

While we do not have such a theoretical limit for other datasets, the risk threshold α can be viewed as a lower bound of the actual concentration, which is useful in visually assessing the convergence performance of our method. We can see that for both MNIST and CIFAR-10, our estimates get very close to the horizontal line at $y = \alpha$ given several thousand training samples. Since the actual concentration must be no less than α and our estimated upper bound is approaching α from the above, we immediate infer that the actual concentration of these data distributions with ℓ_∞-norm distance should be a value sightly greater than α. These results not only demonstrate the superiority of our method over the method of Mahloujifar et al. (2019b) in estimating concentration, but also show that concentration of measure is not the reason for our inability to find adversarially robust models for these image benchmarks.

6 Conclusion

Our results advance understanding of the intrinsic limits of adversarial robustness, strengthening the conclusion from Mahloujifar et al. (2019b) which asserted that concentration of measure is not the sole cause of the vulnerability of existing classifiers to adversarial attacks. We generalized the standard Gaussian Isoperimetric Inequality, and then leveraged theoretical insights from that to construct an efficient method for empirically estimating the concentration of arbitrary data distribution using samples. Our method is able to generalize to any ℓ_p-norm distance metric, and surpasses previous approaches in both estimation accuracy and data-efficiency.

REFERENCES

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *NeurIPS*, 2019.

Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones mathematicae*, 30(2): 207–216, 1975.

Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, 2019.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320, 2019.

Dimitrios Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *NeurIPS*, 2018.

Elvis Dohmatob. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning*, 2019.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *NeurIPS*, 2018.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv:1801.02774*, 2018.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *IEEE International Conference on Computer Vision*, 2019.

Julien M Hendrickx and Alex Olshevsky. Matrix p-norms are NP-hard to approximate if $p \neq 1, 2, \infty$. *SIAM Journal on Matrix Analysis and Applications*, 31(5), 2010.

Michel Ledoux. Isoperimetry and Gaussian analysis. In *Lectures on Probability Theory and Statistics*. Springer, 1996.

Saeed Mahloujifar, Dimitrios Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *AAAI Conference on Artificial Intelligence*, 2019a.

Saeed Mahloujifar, Xiao Zhang, Mohammad Mahmoody, and David Evans. Empirically measuring concentration: Fundamental limits on intrinsic robustness. In *NeurIPS*, 2019b.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, 2016.

Mustapha Raıssouli and Iqbal H Jebril. Various proofs for the decrease monotonicity of the Schatten's power norm, various families of $\mathbb{R}^n$-norms and some open problems. *International Journal of Open Problems in Computer Science and Mathematics*, 3(2):164–174, 2010.

Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019.

Vladimir N Sudakov and Boris S Tsirelson. Extremal properties of half-spaces for spherically invariant measures. *Zapiski Nauchnykh Seminarov Leningrad Otdel Mathematical Institute Steklov* (LOMI), 41:14–24, 1974.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 81(1):73–205, 1995.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Xiao Zhang, Jinghui Chen, Quanquan Gu, and David Evans. Understanding the intrinsic robustness of image distributions using conditional generative models. In *International Conference on Artificial Intelligence and Statistics* (AISTATS), 2020.