EMPIRICALLY MEASURING CONCENTRATION: FUNDAMENTAL LIMITS ON INTRINSIC ROBUSTNESS

Xiao Zhang[*], Saeed Mahloujifar[*], Mohammad Mahmoody, and David Evans
University of Virginia
[saeed, xz7bc, mohammad, evans]@virginia.edu

ABSTRACT

Many recent works have shown that adversarial examples which fool classifiers can be found by minimally perturbing a normal input. Recent theoretical results, starting with Gilmer et al. (2018b), show that if the inputs are drawn from a *concentrated* metric probability space, then adversarial examples with small perturbation are inevitable. A concentrated space has the property that any subset with $\Omega(1)$ (e.g., 1/100) measure, according to the imposed distribution, has small distance to almost all (e.g., 99/100) of the points in the space. It is not clear, however, whether these theoretical results apply to actual distributions in practice such as images. This paper presents a method for empirically measuring and bounding the concentration of a concrete dataset that is proven to converge to the actual concentration. We use it to empirically estimate the intrinsic robustness to ℓ_∞ perturbations of several image classification benchmarks.

1 INTRODUCTION

Despite achieving exceptionally high accuracy on natural inputs, state-of-the-art machine learning models have been shown to be vulnerable to adversarial agents who add small perturbation to fool the classifier using so-called *adversarial examples* (Szegedy et al., 2014; Goodfellow et al., 2015). This phenomenon has motivated numerous studies (Papernot et al., 2016; Madry et al., 2017; Biggio & Roli, 2018; Gilmer et al., 2018a) to propose heuristic defenses that aim to robustify existing classifiers. However, most of defense mechanisms have been quickly broken by adaptive attacks (Carlini & Wagner, 2017; Athalye et al., 2018). To end this arms race, a recent line of research (Wong & Kolter, 2018; Raghunathan et al., 2018; Wong et al., 2018) proposes training methods that are certified to be robust for given inputs against some specific norm-bounded adversarial perturbations and empirically verified their effectiveness for toy datasets. These methods are not able to certify global robustness properties, however, but can only certify robustness for given inputs.

The above-mentioned difficulties motivate a fundamental information-theoretic question: *what are the inherent limitations of developing robust classifiers?* Recent theoretical works (Gilmer et al., 2018b; Fawzi et al., 2018; Mahloujifar et al., 2018; Shafahi et al., 2018) have shown that under certain assumptions regarding the data distribution and the perturbation metric, adversarial examples are inevitable. As a result, for a broad set of theoretically natural metric probability spaces of inputs, there is no classifier for the data distribution that achieves adversarial robustness. For example, Gilmer et al. (2018b) assumed that the input data are sampled uniformly from n-spheres and proved a model-independent theoretical bound connecting the risk to the average Euclidean distance to the "caps" (i.e., round regions on sphere). Mahloujifar et al. (2018) generalized this result to any concentrated metric probability space of inputs and showed that, e.g., if the inputs come from any Normal Lévy family (Lévy, 1951), any classifier with sub-exponentially large test error will be vulnerable to small (i.e., sublinear in the typical norm of the inputs) perturbations.

[*] Equal contribution.
The same work is also presented in the ICLR 2019 *Safe Machine Learning* workshop.

Although such theoretical findings seem discouraging to the goal of developing robust classifiers, all these impossibility results depend on assumptions about data distributions that might not hold for cases of interest. Our work aims for a general method for testing properties of concrete datasets.

Contribution. Our work shrinks the gap between the theoretical analyses of robustness of classification of theoretical data distributions and the intrinsic robustness for classification of actual datasets. Indeed, quantitative estimates of the intrinsic robustness of benchmark image datasets such as MNIST and CIFAR-10 can provide us with a better understanding of the threat of adversarial examples for natural image distributions and may suggest promising directions for practitioners to further improve classifier robustness. Our main technical contribution lies in developing a general method to evaluate the concentration of a given input distribution μ based on a set of data samples. We prove that by simultaneously increasing the sample size m and some complexity parameter T, the concentration of the empirical measure converges to the actual concentration of μ (Section 4). Using this method, we perform experiments to demonstrate the existence of robust error regions for benchmark datasets under ℓ_∞ perturbations (Section 5). For instance, we show the existence of a set/region (as a candidate set for the error region of the classifier) with risk (i.e., measure) 5.94% and adversarial risk (i.e., expanded measure) 18.13% with respect to ℓ_∞ perturbations of magnitude $\epsilon = 8/255$ on CIFAR-10. The existing robust classifier fail to achieve these rates. This suggests that the concentration of measure is *not* the *only* reason behind the vulnerability of existing classifiers to adversarial perturbations. Thus, there seems to be room for improving the robustness of image classifiers (even with non-zero classification error), at least for the datasets studied in this work.

As for related work, we are only aware of one work that tries to heuristically estimate these properties. To extend their theoretical impossibility result to the practical distributions, Gilmer et al. (2018b) studied MNIST dataset to find a region that is somewhat robust in terms of *expected* ℓ_2 distance of other images from the region. In their setting, they showed the existence of a set of measure 0.01 with average ℓ_2 distance 6.59 to all points. In comparison, our work is the first to provide a general methodology to empirically estimate the concentration of measure with provable guarantees. Moreover, we work with ℓ_∞, and *worst* case bounded perturbations for modeling adversarial risk, which is the most popular setting employed in attacks.

2 Definitions and Notations

Lower-case boldface letters such as $\boldsymbol{x}$ are used to denote vectors, and $[n]$ is used to represent $\{1, 2, \ldots, n\}$. For any set $\mathcal{A}$, let $\mathsf{Pow}(\mathcal{A})$, $|\mathcal{A}|$ and $\mathbb{1}_\mathcal{A}(\cdot)$ be the power set, cardinality and indicator function of $\mathcal{A}$, respectively. For any $\boldsymbol{x} \in \mathbb{R}^n$, the ℓ_∞-norm of $\boldsymbol{x}$ is defined as $\|\boldsymbol{x}\|_\infty = \max_{i\in[n]} |x_i|$. Let $(\mathcal{X}, \mu)$ be a probability space and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be some distance metric defined on $\mathcal{X}$. Define the empirical measure with respect to a set $\mathcal{S}$ sampled from μ as $\hat{\mu}_\mathcal{S}(\mathcal{A}) = \sum_{\boldsymbol{x}\in\mathcal{S}} \mathbb{1}_\mathcal{A}(\boldsymbol{x})/|\mathcal{S}|, \forall \mathcal{A} \subseteq \mathcal{X}$. Let $\mathrm{Ball}(\boldsymbol{x}, \epsilon) = \{\boldsymbol{x}' \in \mathcal{X} : d(\boldsymbol{x}', \boldsymbol{x}) \leq \epsilon\}$ be the ball around $\boldsymbol{x}$ with radius ϵ. For any subset $\mathcal{A} \subseteq \mathcal{X}$, define the ϵ-expansion $\mathcal{A}_\epsilon = \{\boldsymbol{x} \in \mathcal{X} \colon \exists\, \boldsymbol{x}' \in \mathrm{Ball}(\boldsymbol{x}, \epsilon) \cap \mathcal{A}\}$. The collection of the ϵ-expansions for members of any $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$ is defined and denoted as $\mathcal{G}_\epsilon = \{\mathcal{A}_\epsilon : \mathcal{A} \in \mathcal{G}\}$.

Moreover, we work with the following definitions regarding the *adversarial risk* of a classifier, the *intrinsic robustness* with respect to some family of classifiers and a specific collection of subsets characterized by *complement of union of hyperrectangles*.

Definition 2.1 (Adversarial Risk and Intrinsic Robustness)**.** *Let $(\mathcal{X}, \mu)$ be the probability space of instances and f^* be the underlying ground-truth. Given a classifier f, the* adversarial risk *of f in metric d with strength ϵ is defined as*

$$\mathit{AdvRisk}_\epsilon(f, f^*) = \Pr_{\boldsymbol{x}\sim\mu}\left[\exists\, \boldsymbol{x}' \in \mathrm{Ball}(\boldsymbol{x}, \epsilon) \text{ s.t. } f(\boldsymbol{x}') \neq f^*(\boldsymbol{x}')\right].$$

For $\epsilon = 0$, which allows no perturbation, the notion of adversarial risk coincides with traditional risk.[1] *In addition, let $\mathcal{F}$ be some family of classifiers, then the* intrinsic robustness *is defined as the*

[1] This definition is used some works including Gilmer et al. (2018b); Bubeck et al. (2018); Mahloujifar et al. (2018). Other related definitions are equivalent when we assume small perturbations preserve the ground truth. See Diochnos et al. (2018) for a taxonomy of different definitions.

maximum adversarial robustness[2] *that can be achieved within $\mathcal{F}$, namely*

$$\mathrm{Rob}_\epsilon(\mathcal{F}, f^*) = 1 - \inf_{f \in \mathcal{F}} \big\{ \mathit{AdvRisk}_\epsilon(f, f^*) \big\}.$$

In this work, we specify $\mathcal{F}$ as the family of imperfect classifiers that have risk at least $\alpha \in (0,1)$.

Definition 2.2 (Complement of union of hyperrectangles)**.** *For any positive integer T, the collection of subsets specified by* complement of union of T n-dimensional hyperrectangles *is defined as*

$$\mathcal{CR}^n_T = \Big\{ \mathbb{R}^n \setminus \cup_{t=1}^T \mathcal{R}ect(\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)}) \colon \forall t \in [T], (\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)}) \in \mathbb{R}^n \times \mathbb{R}^n_{\geq 0} \Big\},$$

where $\mathcal{R}ect(\boldsymbol{u}, \boldsymbol{r}) = \big\{ \boldsymbol{x} \in \mathcal{X} : \forall j \in [n], |\boldsymbol{x}_j - \boldsymbol{u}_j| \leq r_j/2 \big\}$ denotes the hyperrectangle centered at $\boldsymbol{u}$ with $\boldsymbol{r}$ representing the edge size vector. When n is free of context, we simply write $\mathcal{CR}_T = \mathcal{CR}^n_T$.

3 Concentration of Measure

In this paper, we focus on quantifying the effect of concentration of measure on a classification task. Previous work shows a connection between concentration of measure and maximum possible robustness of a an imperfect classifier (Gilmer et al. (2018b); Fawzi et al. (2018); Mahloujifar et al. (2018); Shafahi et al. (2018)). The concentration of measure on a metric probability space is defined by a concentration function as follows.

Definition 3.1. *(Concentration Function) Consider a metric probability space $(\mathcal{X}, \mu, d)$. Suppose $\epsilon > 0$ and $\alpha \in (0,1)$ are given parameters, then the* concentration function *of the probability measure μ with respect to ϵ, α is defined as $h(\mu, \alpha, \epsilon) = \inf_{\mathcal{E} \subseteq \mathcal{X}} \{\mu(\mathcal{E}_\epsilon) \colon \mu(\mathcal{E}) \geq \alpha\}$.*[3]

Generalizing the result of Gilmer et al. (2018b) about instances drawn from spheres, Mahloujifar et al. (2018) showed that, in general, if the metric probability space of instances are concentrated, then any classifier with 1% risk incurs large adversarial risk for small amount of perturbations.

Theorem 3.2 (Mahloujifar et al. (2018))**.** *Let $(\mathcal{X}, \mu)$ be the probability space of instances and f^* be the underlying ground-truth. For any classifier f, we have*

$$\mathit{AdvRisk}_\epsilon(f, f^*) \geq h(\mu, \mathrm{Risk}(f, f^*), \epsilon).$$

In order for this theorem to be useful, we need to know the concentration function. The behavior of this function is studied extensively for certain theoretical metric probability spaces Ledoux (2001); Milman & Schechtman (1986). However, it is not known how to measure the concentration function for arbitrary metric probability spaces. In this work, we provide a framework to (algorithmically) bound the concentration function from i.i.d. samples from a distribution. Namely, we want to solve the following optimization task using our i.i.d. samples:

$$\underset{\mathcal{E} \subseteq \mathcal{X}}{\text{minimize}} \quad \mu(\mathcal{E}_\epsilon) \quad \text{such that} \quad \mu(\mathcal{E}) \geq \alpha. \tag{1}$$

In this work, we aim to estimate the minimum possible adversarial risk, which captures the intrinsic robustness for classification in terms of the underlying distribution μ, conditioned on the fact that the original risk is at least α. Note that solving this optimization problem only shows the possibility of existence of an error region $\mathcal{E}$ with certain (small) expansion. This means that there could potentially exist a classifier with risk at least α and adversarial risk equal to the solution of the optimization problem of (1). Actually *finding* such an optimally robust classifier (with error α) using a learning algorithm might be a much more difficult task or even infeasible and is *not* the goal of our work.

4 Method for Measuring Concentration

In this section, we present a method to measure concentration of measure on a metric probability space using i.i.d. samples. To measure concentration, there are two main challenges:

[2] The term robustness is used with different meanings in previous work (e.g., in Diochnos et al. (2018), it refers to the average distances to the error region). However all such uses of the term refer to a desirable resisting property of the classifier against adversarial perturbations, which is the case here as well.

[3] Note that, the standard notion of concentration function (e.g., see Talagrand (1995)) is related to a special case of Definition 4.1 by fixing $\alpha = 1/2$.

1. Measuring concentration appears to require knowledge of the density function of the distribution, but we only have a data set sampled from the distribution.
2. Even with the density function, we have to find the best possible subset among all the subsets of the space, which seems infeasible.

We show how to overcome these challenges to find the actual concentration in the limit by first empirically simulating the distribution and then narrowing down our search space to a specific collection of subsets that are chosen in a careful way. Our results show that for such (carefully chosen) family of sets, the expansion can be bounded using polynomially many samples, while the convergence to the actual concentration (without the limits on the sets) happens in the limit $T \to \infty$ where T is the parameter related to the complexity of the collection.

Below, we introduce two useful definitions before stating our two main theorems that show how to overcome these challenges. The following definition captures the concentration function for a specific collection of subsets:

Definition 4.1. *(Concentration Function for a Collection of Subsets) Consider a metric probability space* $(\mathcal{X}, \mu, d)$*. Let* $\epsilon > 0$ *and* $\alpha \in (0, 1)$ *be given parameters, then the* concentration function *of the probability measure* μ *with respect to* ϵ*,* α *and a collection of subsets* $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$ *is defined as* $h(\mu, \alpha, \epsilon, \mathcal{G}) = \inf_{\mathcal{E}\in\mathcal{G}} \{\mu(\mathcal{E}_\epsilon)\colon \mu(\mathcal{E}) \geq \alpha\}$*. We write* $h(\mu, \alpha, \epsilon)$ *when* $\mathcal{G} = \mathsf{Pow}(\mathcal{X})$*.*

We also need to define the notion of complexity penalty for a collection of subsets. The complexity penalty for a collection of subsets capture the rate of the uniform convergence for the subsets in that collection. One can get such uniform convergence rates using VC dimension or Rademacher complexity of the collection.

Definition 4.2 (Complexity Penalty)**.** *Let* $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$ *be a collection of subsets of* $\mathcal{X}$*. A function* $\phi\colon \mathbb{N} \times \mathbb{R} \to [0, 1]$ *is a complexity penalty for* $\mathcal{G}$ *iff for any probability measure* μ *supported on* $\mathcal{X}$ *and any* $\delta \in [0, 1]$*, we have*

$$\Pr_{S\leftarrow\mu^m}[\exists\, \mathcal{E} \in \mathcal{G} \ \ s.t. \ |\mu(\mathcal{E}) - \hat{\mu}_S(\mathcal{E})| \geq \delta] \leq \phi(m, \delta).$$

The following theorem shows how to overcome the challenge of measuring concentration from finite samples when the concentration is defined with respect to specific families of subsets. Namely, it shows that the empirical concentration is close to the true concentration if the underlying collection of subsets is not too complex.

Theorem 4.3 (Generalization of Concentration)**.** *Let* $(\mathcal{X}, \mu, d)$ *be a metric probability space and* $\mathcal{G} \subseteq \mathsf{Pow}(\mathcal{X})$*. For any* $\delta, \alpha, \epsilon \in [0, 1]$*, we have*

$$\Pr_{S\leftarrow\mu^m}[h(\mu, \alpha-\delta, \epsilon, \mathcal{G}) - \delta \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) \leq h(\mu, \alpha+\delta, \epsilon, \mathcal{G}) + \delta] \geq 1 - 2\big(\phi(m, \delta) + \phi_\epsilon(m, \delta)\big)$$

where ϕ *and* ϕ_ϵ *are complexity penalties for* $\mathcal{G}$ *and* $\mathcal{G}_\epsilon$ *respectively.*

See Appendix A for the proof. The above theorem shows that if we narrow down our search to a collection of subsets $\mathcal{G}$ such that both $\mathcal{G}$ and $\mathcal{G}_\epsilon$ have small complexity penalty, then we can use the empirical distribution to measure concentration of measure for that specific collection.

Note that the generalization bound of Theorem 4.3 depends on complexity penalties for both $\mathcal{G}$ and $\mathcal{G}_\varepsilon$. Therefore, in order for this theorem to be useful, the collection $\mathcal{G}$ should be chosen in a careful way. For example, if $\mathcal{G}$ has bounded VC dimension, then $\mathcal{G}_\epsilon$ might still have very large VC dimension. Alternatively, $\mathcal{G}$ might denote the collection of subsets that are decidable by a neural network of certain size. In that case, even though there are well known complexity penalties for such collections (See Neyshabur et al. (2017)), the complexity of their *expansions* is unknown. In fact, relating the complexity penalty for expansion of a collection to that of the original collection is tightly related to generalization bounds in the adversarial settings, which has also been been the subject of recent works Attias et al. (2018); Montasser et al. (2019); Yin et al. (2018).

Theorem 4.3 showed how to estimate the concentration function with respect to a specific collection of sets. Our Theorem 4.4 below states that if we gradually increase the complexity of the collection, and the number of samples together, the empirical estimate of concentration converges to actual concentration, as long as several conditions hold. This convergence theorem and its proof techniques are inspired by the work of Scott & Nowak (2006) on learning minimum volume sets.

Theorem 4.4. *Let $\left\{\mathcal{G}^T\right\}_{T\in\mathbb{N}}$ be a family of subset collections defined over a space $\mathcal{X}$. Also let $\left\{\phi^T\right\}_{T\in\mathbb{N}}$ and $\left\{\phi^T_\epsilon\right\}_{T\in\mathbb{N}}$ be two families of complexity penalty functions such that ϕ^T and ϕ^T_ϵ are complexity penalties for $\mathcal{G}^T$ and $\mathcal{G}^T_\epsilon$ respectively, for some $\epsilon \in [0,1]$. Let $\{m(T)\}_{T\in\mathbb{N}}$ and $\{\delta(T)\}_{T\in\mathbb{N}}$ be a two series such that $m(T) \in \mathbb{N}$ and $\delta(T) \in [0,1]$. Consider a series of datasets $\{S_T\}_{T\in\mathbb{N}}$, where S_T consists of $m(T)$ i.i.d. samples from a measure μ supported on $\mathcal{X}$. Also let $\alpha \in [0,1]$ be such that h is locally continuous w.r.t the second parameter at point $(\mu, \alpha, \epsilon, \mathsf{Pow}(\mathcal{X}))$. If all the following hold,*

1. $\sum_{T=1}^{\infty} \phi^T(m(T), \delta(T)) < \infty$
2. $\sum_{T=1}^{\infty} \phi^T_\epsilon(m(T), \delta(T)) < \infty$
3. $\lim_{T\to\infty} \delta(T) = 0$
4. $\lim_{T\to\infty} h(\mu, \alpha, \epsilon, \mathcal{G}_T) = h(\mu, \alpha, \epsilon)$

then with probability 1*, we have* $\lim_{T\to\infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}_T) = h(\mu, \alpha, \epsilon)$.

See Appendix A for the proof. In the above theorem, the first two conditions are about the growth rate for the complexity of the collections. Namely, we need the complexity penalties $\phi^T(m(T), \delta(T))$ and $\phi^T_\epsilon(m(T), \delta(T))$ to rapidly approach 0 as $T \to \infty$, which means the complexity of $\mathcal{G}^T$ and $\mathcal{G}^T_\varepsilon$ should grow at a slow rate. The third condition requires that our generalization error goes to zero as we increase T. Finally, the forth condition requires that our approximation error goes to 0 as we increase T.

4.1 Special Case of ℓ_∞

In this section we show how to instantiate Theorem 4.4 for the case of ℓ_∞. Namely, we want to find a subset $\mathcal{E} \in \mathbb{R}^n$ such that $\mathcal{E}$ has measure at least α and the ϵ-expansion of $\mathcal{E}$ under ℓ_∞ has the minimum measure. To achieve this goal, we approximate the distribution μ with an empirical distribution $\hat{\mu}_S$. We also limit our search to a special collection $\mathcal{CR}_T$ (though our goal is to find the minimum concentration around arbitrary subsets). Namely, what we find is still an *upper bound* on the concentration function, and it is an upper bound that we know it converges the actual amount in the limit. Our problem thus becomes the following optimization task:

$$\underset{\mathcal{E}\in\mathcal{CR}_T}{\text{minimize}} \;\; \hat{\mu}_{\mathcal{S}}(\mathcal{E}_\epsilon) \quad \text{such that} \;\; \hat{\mu}_{\mathcal{S}}(\mathcal{E}) \geq \alpha. \tag{2}$$

The following theorem provides the key to our empirical method by providing a convergence guarantee. It states that if we increase the number of rectangles and the number of samples together in a careful way, the solution to the problem using restricted sets converges to the true concentration.

Theorem 4.5. *Consider a metric probability space $(\mathbb{R}^n, \mu, \ell_\infty)$. Let $\{S_T\}_{T\in\mathbb{N}}$ be a family of datasets such that for all $T \in \mathbb{N}$, S_T contains at least T^4 i.i.d. samples from μ. For any $\epsilon, \alpha \in [0,1]$, if h is locally continuous w.r.t the second parameter at point (μ, α, ϵ), then we probability* 1 *we get*

$$\lim_{T\to\infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{CR}_T) = h(\mu, \alpha, \epsilon).$$

See Appendix A for the proof.

5 Experiments

In this section, we provide a heuristic method to find the best possible error region that covers at least α fraction of the samples and its expansion covers the least number of points. More specifically, we first introduce our algorithm and then evaluate our approach on several benchmark image datasets: MNIST (LeCun et al., 2010), CIFAR-10 (Krizhevsky & Hinton, 2009), FASHION-MNIST (Xiao et al., 2017) and SVHN (Netzer et al., 2011).

Algorithm 1: Binary Search for Robust Error Region

```
Input  : a set of images $\mathcal{S}$; perturbation strength $\epsilon$ (in $\ell_\infty$-norm); error threshold $\alpha$; number of
         hyperrectangles $T$; number of nearest neighbours $k$; precision for binary search $\delta_{\text{bin}}$.
$r_k(\boldsymbol{x}) \leftarrow$ compute the $\ell_1$-norm distance to the $k$-th nearest neighbour for each $\boldsymbol{x} \in \mathcal{S}$;
$\mathcal{S}_{\text{sort}} \leftarrow$ sort all the images in $\mathcal{S}$ by $r_k(\boldsymbol{x})$ in an ascending order;
$q_{\text{lower}} \leftarrow 0.0, \;\; q_{\text{upper}} \leftarrow 1.0$;
while $q_{upper} - q_{lower} > \delta_{bin}$ do
    $q \leftarrow (q_{\text{lower}} + q_{\text{upper}})/2$;
    perform kmeans clustering algorithm ($T$ clusters, $\ell_1$ metric) on the top-$q$ images of $\mathcal{S}_{\text{sort}}$;
    $\{\boldsymbol{u}^{(t)}\}_{t=1}^{T} \leftarrow$ record the centroids of the resulted $T$ clusters;
    for $t = 1, 2, \ldots, T$ do
        $\mathcal{R}ect(\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)}) \leftarrow$ cover $t$-th cluster with the minimum-sized rectangle centered at $\boldsymbol{u}^{(t)}$;
    end
    $\mathcal{E}_q \leftarrow \mathcal{X} \setminus \cup_{t=1}^{T} \mathcal{R}ect_\epsilon(\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)})$ ;     // $\mathcal{R}ect_\epsilon(\boldsymbol{u}, \boldsymbol{r})$ denotes the $\epsilon$-expansion of $\mathcal{R}ect(\boldsymbol{u}, \boldsymbol{r})$
    if $|\mathcal{S} \cap \mathcal{E}_q|/|\mathcal{S}| \geq \alpha$ then
        $q_{\text{lower}} \leftarrow q, \;\; AdvRisk_q \leftarrow \big|\{\boldsymbol{x} \in \mathcal{S} : \boldsymbol{x} \notin \cup_{t=1}^{T} \mathcal{R}ect(\boldsymbol{u}^{(t)}, \boldsymbol{r}^{(t)})\}\big|/|\mathcal{S}|$;
    else
        $q_{\text{upper}} \leftarrow q$;
    end
end
$\hat{q} \leftarrow \text{argmin}_q \{AdvRisk_q\}$;
Output: ($\hat{q}$, $AdvRisk_{\hat{q}}$, $\mathcal{E}_{\hat{q}}$)
```

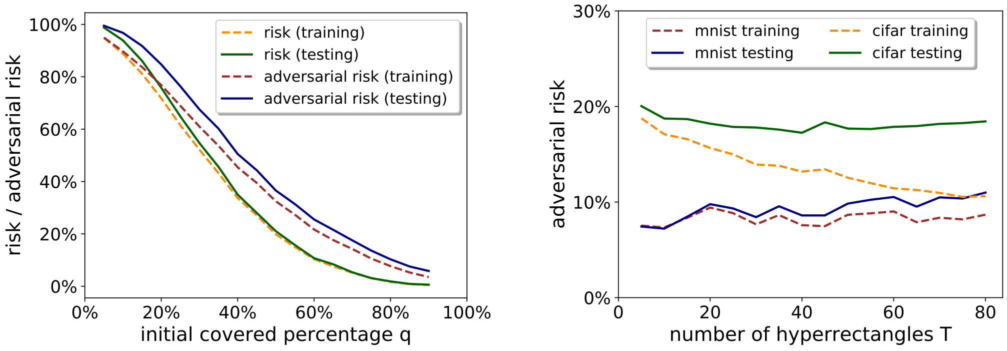
5.1 Methodology

Theorem 4.5 guarantees that the empirical concentration function $h(\hat{\mu}_{\mathcal{S}}, \alpha, \epsilon, \mathcal{CR}_T)$ converges towards the concentration of measure $h(\mu, \alpha, \epsilon)$ asymptotically. Thus, to measure the concentration of μ, it remains to solve the optimization problem (2). Although the collection of subsets is specified using simple topology, solving (2) exactly is still difficult, as the problem itself is combinatorial in nature. Borrowing techniques from clustering, we propose an empirical method, as shown in Algorithm 1, to search for desirable error region within $\mathcal{CR}_T$. We remark that any error region $\mathcal{E}$ could be used to construct a classifer $f_{\mathcal{E}}$, i.e. $f_{\mathcal{E}}(\boldsymbol{x}) = f^*(\boldsymbol{x}), \text{ if } \boldsymbol{x} \notin \mathcal{E}$; $f_{\mathcal{E}}(\boldsymbol{x}) \neq f^*(\boldsymbol{x}), \text{ if } \boldsymbol{x} \in \mathcal{E}$. However, finding such a classifier using a learning algorithm might be a very difficult task. Here we find the optimally robust error region, not the corresponding classifier. A desirable error region should have small adversarial risk[4], compared with all subsets within $\mathcal{CR}_T$ that have measure at least α.

The high-level intuition of our method is that images from different classes are likely to be concentrated in separable regions, since it is generally believed that small perturbations preserve the ground-truth class at the sampled images. Therefore, if we cluster all the images into different clusters, a desired region with low adversarial risk should exclude any image from the dense clusters, otherwise the expansion of such region will quickly cover the whole cluster. In other words, a desirable subset within $\mathcal{CR}_T$ should be ϵ-away (in ℓ_∞ norm) from all the dense image clusters, which motivates our method to cover the dense image clusters using hyperrectangles and treat the complement of them as error set.

More specifically, our algorithm starts by sorting all the training images in an ascending order based on the ℓ_1-norm distance to the k-th nearest neighbour with $k = 50$, and then obtains T hyperrectangular image clusters by performing k-means clustering (Hartigan & Wong, 1979) on the top-q densest images, where the metric is chosen as ℓ_1 and the maximum iterations is set as 30. Finally, we perform a binary search over $q \in [0, 1]$, where we set $\delta_{\text{bin}} = 0.005$ as the stopping criteria, to obtain the best robust subset (lowest adversarial risk) in $\mathcal{CR}_T$ with empirical measure at least α.

[4] The adversarial risk of an error region $\mathcal{E}$ simply refers to the adversarial risk of $f_{\mathcal{E}}$.



(a) varying q

(b) varying T

Figure 1: (a) Plots of risk and adversarial risk w.r.t. the resulted error region using our method as q varies (CIFAR-10, $\epsilon = 8/255$, $T = 30$); (b) Plots of adversarial risk w.r.t. the resulted error region using our method (best q) as T varies on MNIST ($\epsilon = 0.3$) and CIFAR-10 ($\epsilon = 8/255$).

Table 1: Summary of the main results for different experimental settings using our method

Dataset	α	ϵ	T	**Best** q	**Empirical Risk** (%)		**Empirical AdvRisk** (%)	
					training	testing	training	testing
MNIST	0.01	0.1	5	0.662	1.22 ± 0.11	1.23 ± 0.12	3.65 ± 0.29	3.64 ± 0.30
		0.2	10	0.660	1.12 ± 0.13	1.11 ± 0.10	5.76 ± 0.38	5.89 ± 0.44
		0.3	10	0.629	1.12 ± 0.12	1.15 ± 0.13	7.34 ± 0.38	7.24 ± 0.38
		0.4	10	0.598	1.15 ± 0.09	1.21 ± 0.09	9.89 ± 0.57	9.92 ± 0.60
CIFAR-10	0.05	2/255	10	0.680	5.32 ± 0.21	5.72 ± 0.25	7.29 ± 0.20	8.13 ± 0.26
		4/255	20	0.688	5.59 ± 0.25	6.05 ± 0.40	11.43 ± 0.24	13.66 ± 0.33
		8/255	40	0.734	5.55 ± 0.21	5.94 ± 0.34	13.69 ± 0.19	18.13 ± 0.30
		16/255	75	0.719	5.16 ± 0.25	5.28 ± 0.23	19.77 ± 0.22	28.83 ± 0.46
FASHION MNIST	0.05	0.1	10	0.758	5.64 ± 0.78	5.92 ± 0.85	10.30 ± 0.72	11.56 ± 0.84
		0.2	10	0.726	5.79 ± 1.00	6.00 ± 1.02	13.44 ± 0.60	14.82 ± 0.71
		0.3	10	0.668	5.90 ± 0.94	6.13 ± 0.93	17.46 ± 0.53	18.87 ± 0.66
SVHN	0.05	0.01	10	0.812	5.21 ± 0.19	8.83 ± 0.30	6.08 ± 0.20	10.17 ± 0.29
		0.02	10	0.773	5.31 ± 0.12	8.86 ± 0.20	7.76 ± 0.12	12.46 ± 0.15
		0.03	10	0.750	5.15 ± 0.13	8.55 ± 0.22	8.88 ± 0.13	13.82 ± 0.25

5.2 Results

We choose α to reflect the best accuracy achieved by state-of-the-art classifiers, using $\alpha = 0.01$ and $\epsilon \in \{0.1, 0.2, 0.3, 0.4\}$ for MNIST and selecting appropriate values to represent the best typical results on the other data sets (see Table 1). Given the number of hyperrectangles T, we obtain the resulting error region using Algorithm 1 on the training dataset, and tune T for the minimum adversarial risk on the testing dataset.

Figure 1 shows the training and testing curves regarding risk and adversarial risk for two specific experimental settings[5]. In particular, Figure 1(a) suggests that as we increase the initial covered percentage q, both risk and adversarial risk of the corresponding error region decrease. This supports our use of binary search on q in Algorithm 1. On the other hand, as can be seen from Figure 1(b), overfitting with respect to adversarial risk becomes significant as we increases the value T. According to the adversarial risk curve for testing data, the optimal value of T is selected as $T = 10$ for MNIST ($\epsilon = 0.3$) and $T = 40$ for CIFAR-10 ($\epsilon = 8/255$).

[5] Similar results are obtained under other experimental settings, which are shown in Appendix B.

Table 2: Comparisons between our method and existing robust classifier under different settings

Dataset	ϵ	Method	Empirical Risk	Empirical AdvRisk
MNIST	0.3	Madry et al. (2017) Ours ($\alpha = 0.012, T = 10$)	1.20% $1.35\% \pm 0.08\%$	10.7% $8.28\% \pm 0.22\%$
CIFAR-10	$8/255$	Madry et al. (2017) Ours ($\alpha = 0.127, T = 40$)	12.70% $14.22\% \pm 0.46\%$	52.96% $29.21\% \pm 0.35\%$

Table 1 summarizes the optimal parameters, the empirical risk and adversarial risk of the corresponding error region on both training and testing datasets for each experimental setting. Since k-means algorithm does not guarantee global optimum, we repeat our method for 10 multiple runs with random restart in terms of the best parameters, then report both the mean and the standard deviation. Our experiments provide examples of rather robust error regions for real image datsets. For instance, in Table 1 we have a case where the measure of the resulted error region increases from 5.94% to 18.13% after expansion with $\epsilon = 8/255$ in ℓ_∞ metric on CIFAR-10 dataset. This means that there could potentially exist a classifier with 5.94% risk and 18.13% adversarial risk. However, the existing adversarially robust classifiers cannot achieve these rates.

Table 2 compares the measured bounds using our method with result for an adversarially-trained classifier using PGD-attack (Madry et al., 2017) on MNIST and CIFAR-10 under ℓ_∞ perturbations. We use empirical risk and adversarial risk for Madry et al. (2017) to denote the standard test error and PGD-attack success rate of the reported robust model. Taking CIFAR-10 ($\epsilon = 8/255$) as an example, we obtain an error region with adversarial risk around 29.21% using our method, where the risk threshold is set as $\alpha = 0.127$ in order to match the test error of Madry et al. (2017). Compared with the empirical adversarial risk 52.96% attained using adversarial training, we show that only 29.21% could be explained according to the concentration of measure phenomenon, while the remaining 23.75% are due to other reasons, which requires further investigation.

All the aforementioned observations suggest that the concentration of measure phenomenon is not the sole reason behind vulnerability of the existing classifiers to adversarial examples, at least for datasets that we studied in this work. In other words, the impossibility results of (Gilmer et al., 2018b; Fawzi et al., 2018; Mahloujifar et al., 2018; Shafahi et al., 2018), should not make the community hopeless in finding more robust image classifiers.

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. *arXiv preprint arXiv:1810.02180*, 2018.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57. IEEE, 2017.

Dimitrios Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Advances in Neural Information Processing Systems*, pp. 10359–10368, 2018.

David Eisenstat and Dana Angluin. The vc dimension of k-fold union. *Information Processing Letters*, 101(5):181–184, 2007.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *arXiv preprint arXiv:1802.08686*, 2018.

Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018a.

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

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

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. `http://yann.lecun.com/exdb/mnist`, 2010.

Michel Ledoux. *The Concentration of Measure Phenomenon*. Number 89 in Mathematical Surveys and Monographs. American Mathematical Society, 2001.

Paul Lévy. *Problèmes concrets d'analyse fonctionnelle*, volume 6. Gauthier-Villars Paris, 1951.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *arXiv preprint arXiv:1809.03063*, 2018.

Vitali D Milman and Gideon Schechtman. *Asymptotic theory of finite dimensional normed spaces*, volume 1200. Springer Verlag, 1986.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597. IEEE, 2016.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Clayton D Scott and Robert D Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7(Apr):665–704, 2006.

Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

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 Etudes Scientifiques*, 81(1):73–205, 1995.

Li-Xin Wang. Universal approximation by hierarchical fuzzy systems. *Fuzzy sets and systems*, 93 (2):223–230, 1998.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5283–5292, 2018.

Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. *arXiv preprint arXiv:1805.12514*, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *arXiv preprint arXiv:1810.11914*, 2018.

A Proofs of Main Theorem

In this section, we prove Theorems 4.3, 4.4 and 4.5. We first prove Theorem 4.3.

Proof of Theorem 4.3. Define $g(\mu, \alpha, \epsilon, \mathcal{G}) = \operatorname{argmin}_{\mathcal{E} \in \mathcal{G}} \{\mu(\mathcal{E}_\epsilon) \colon \mu(\mathcal{E}) \geq \alpha\}$, and let $\mathcal{E} = g(\mu, \alpha + \delta, \epsilon, \mathcal{G})$ and $\hat{\mathcal{E}} = g(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G})$. (Note that these sets achieving the minimum might not exist, in which case we select a set for which the expansion is arbitrarily close to the infimum and every step of the proof will extend to this variant).

By the definition of the complexity penalty we have

$$\Pr_{S \leftarrow \mu^m}\left[\left|\mu(\hat{\mathcal{E}}) - \hat{\mu}_S(\hat{\mathcal{E}})\right| \geq \delta\right] \leq \phi(m, \delta)$$

which implies

$$\Pr_{S \leftarrow \mu^m}[\mu(\hat{\mathcal{E}}) \leq \alpha - \delta] \leq \phi(m, \delta).$$

Therefore, by the definition of h we have

$$\Pr_{S \leftarrow \mu^m}[\mu(\hat{\mathcal{E}}_\epsilon) \leq h(\mu, \alpha - \delta, \epsilon, \mathcal{G})] \leq \phi(m, \delta). \tag{3}$$

On the other hand, based on the definition of ϕ_ϵ we have

$$\Pr_{S \leftarrow \mu^m}\left[\left|\mu(\hat{\mathcal{E}}_\epsilon) - \hat{\mu}_S(\hat{\mathcal{E}}_\epsilon)\right| \geq \delta\right] \leq \phi_\epsilon(m, \delta). \tag{4}$$

Combining Equation 3 and Equation 4, and by a union bound we get

$$\Pr_{S \leftarrow \mu^m}[\hat{\mu}_S(\hat{\mathcal{E}}_\epsilon) \leq h(\mu, \alpha - \delta, \epsilon, \mathcal{G}) - \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta)$$

which by the definition of $\hat{\mathcal{E}}$ implies that

$$\Pr_{S \leftarrow \mu^m}[h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) \leq h(\mu, \alpha - \delta, \epsilon, \mathcal{G}) - \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta). \tag{5}$$

Now we bound the probability for the other side of our inequality. By the definition of the notion of complexity penalty we have

$$\Pr_{S \leftarrow \mu^m}[|\mu(\mathcal{E}) - \hat{\mu}_S(\mathcal{E})| \geq \delta] \leq \phi(m, \delta)$$

which implies

$$\Pr_{S \leftarrow \mu^m}[\hat{\mu}_S(\mathcal{E}) \leq \alpha] \leq \phi(m, \delta).$$

Therefore, by the definition of h we have,

$$\Pr_{S \leftarrow \mu^m}[\hat{\mu}_S(\mathcal{E}_\epsilon) \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G})] \leq \phi(m, \delta). \tag{6}$$

On the other hand, based on the definition of ϕ_ϵ we have

$$\Pr_{S \leftarrow \mu^m}[|\mu(\mathcal{E}_\epsilon) - \hat{\mu}_S(\mathcal{E}_\epsilon)| \geq \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta). \tag{7}$$

Combining Equations 6 and 7, by union bound we get

$$\Pr_{S \leftarrow \mu^m}[\mu(\mathcal{E}_\epsilon) \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) - \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta)$$

which by the definition of $\mathcal{E}$ implies

$$\Pr_{S \leftarrow \mu^m}[h(\mu, \alpha + \delta, \epsilon, \mathcal{G}) \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) - \delta] \leq \phi(m, \delta) + \phi_\epsilon(m, \delta). \tag{8}$$

Now combining Equations 5 and 8, by union bound we have

$$\Pr_{S \leftarrow \mu^m}[h(\mu, \alpha - \delta, \epsilon, \mathcal{G}) - \delta \leq h(\hat{\mu}_S, \alpha, \epsilon, \mathcal{G}) \leq h(\mu, \alpha + \delta, \epsilon, \mathcal{G}) + \delta] \geq 1 - 2\left(\phi(m, \delta) + \phi_\epsilon(m, \delta)\right).$$

□

Next, we prove Theorem 4.4 using ideas similar to ideas used in Scott & Nowak (2006).

Proof of Theorem 4.4. We use the following lemma to prove the theorem.

Lemma A.1 (Borel-Cantelli Lemma). *Let $\{E_T\}_{T\in\mathbb{N}}$ be a series of events such that*

$$\sum_{T=1}^{\infty} \Pr[E_T] < \infty$$

Then with probability 1, only finite number of events will occur.

Define E_T to be the event that

$h(\mu, \alpha - \delta(T), \epsilon, \mathcal{G}^T) - \delta(T) > h(\hat{\mu}_{S_T}, \alpha, \epsilon)$ or $h(\mu, \alpha + \delta(T), \epsilon, \mathcal{G}^T) + \delta(T) < h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G})$.

Based on Theorem 4.3 we have $\Pr[E_T] \leq 2 \cdot (\phi^T(m(T), \delta(T)) + \phi^T_\epsilon(m(T), \delta(T)))$. Therefore, by Conditions 1 and 2 we have

$$\sum_{T=1}^{\infty} \Pr[E_T] \leq 2 \left(\sum_{T=1}^{\infty} \phi^T\left(m(T), \delta(T)\right) + \phi^T_\epsilon\left(m(T), \delta(T)\right) \right) < \infty.$$

Now by Lemma A.1, we know there exist with measure 1 some $j \in \mathbb{N}$, such that for all $T \geq j$,

$$h(\mu, \alpha - \delta(T), \epsilon, \mathcal{G}^T) - \delta(T) \leq h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}^T) \leq h(\mu, \alpha + \delta(T), \epsilon, \mathcal{G}^T) + \delta(T).$$

The above implies that

$$\lim_{T\to\infty} h(\mu, \alpha - \delta(T), \epsilon, \mathcal{G}^T) - \delta(T) \leq \lim_{T\to\infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}^T) \leq \lim_{T\to\infty} h(\mu, \alpha + \delta(T), \epsilon, \mathcal{G}^T) + \delta(T).$$

Therefore, by Condition 3 and local continuity of h we have

$$\lim_{T\to\infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}^T) = \lim_{T\to\infty} h(\mu, \alpha, \epsilon, \mathcal{G}^T).$$

Now based on Condition 4 we have

$$\lim_{T\to\infty} h(\hat{\mu}_{S_T}, \alpha, \epsilon, \mathcal{G}^T) = h(\mu, \alpha, \epsilon).$$

□

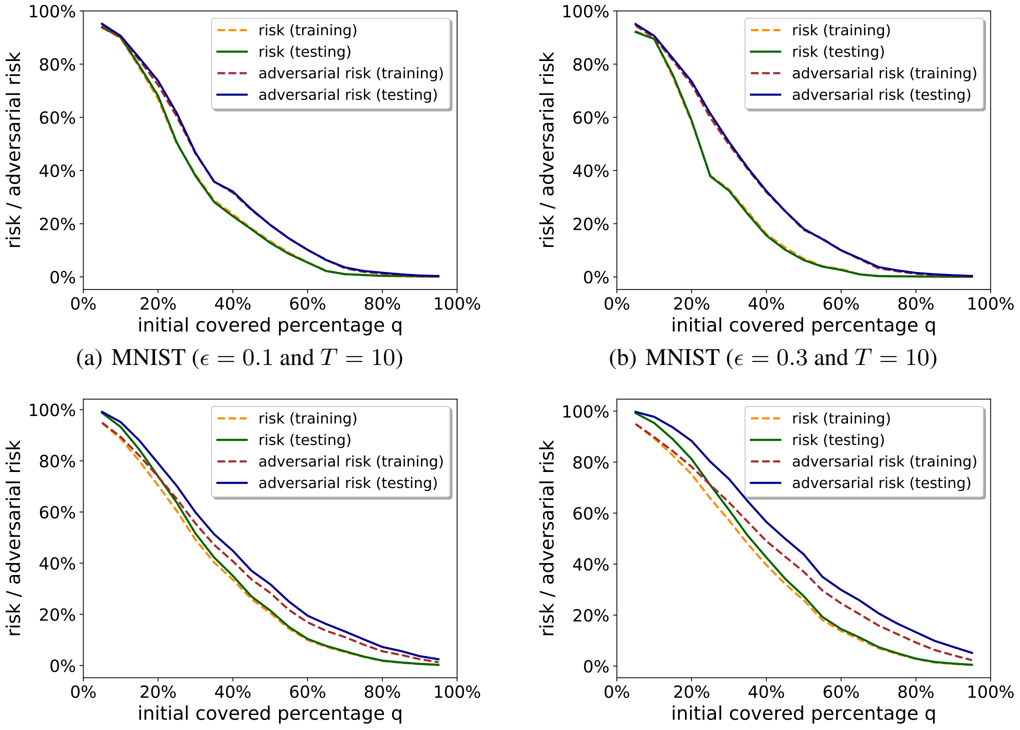
Finally, we prove Theorem 4.5 using Theorem 4.4.

Proof of Theorem 4.5. This theorem follows from our general Theorem 4.4. We show that the choice of parameters here satisfies all four conditions of Theorem 4.4. If we let $\mathcal{G}^T$ to be the collection of subsets specified by complement of union of T hyperrectangles. Then $\mathcal{G}^T_\epsilon$ will be the collection of of subsets specified by complement of union of T hyperrectangles that are bigger than ϵ in each coordinate. Therefore we have $\mathcal{G}^T_\epsilon \subset \mathcal{G}^T$. We know that the VC dimension of $\mathcal{G}^T$ is $d_T = O(nT\log(T))$ (See Eisenstat & Angluin (2007)). Therefore, by VC inequality we have

$$\Pr_{S \leftarrow \mu^m}\left[\sup_{\mathcal{E}\in\mathcal{G}^T} |\mu(\mathcal{E}) - \hat{\mu}_S(\mathcal{E})| \geq \delta\right] \leq 8e^{nT\log(T)\log(m) - m\delta^2/128}.$$

Therefore $\Phi^T(m, \delta) = 8e^{nT\log(T)\log(m) - m\delta^2/128}$ is a complexity penalty for both $\mathcal{G}^T$ and $\mathcal{G}^T_\epsilon$. Hence, if we define $\delta(T) = 1/T$ and $m(T) \geq T^4$, then the first three conditions of Theorem 4.4 are satisfied. The fourth condition is also satisfied by the universal approximation property of histograms (See Wang (1998)). □

B Other Experimental Results



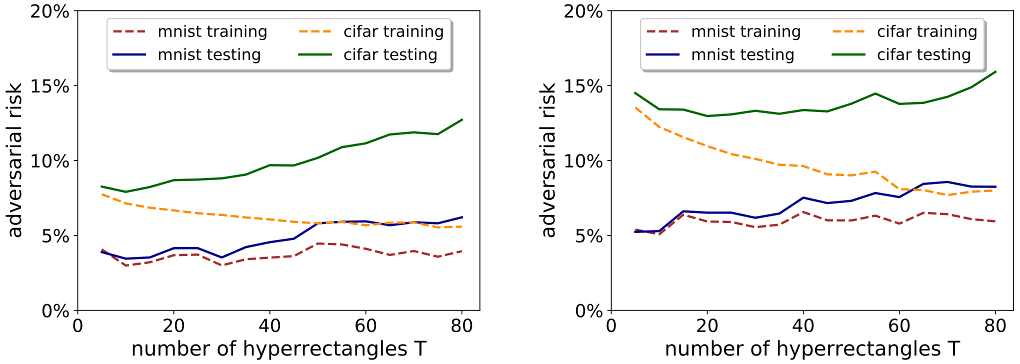
(a) MNIST ($\epsilon = 0.1$ and $T = 10$)

(b) MNIST ($\epsilon = 0.3$ and $T = 10$)

(c) CIFAR-10 ($\epsilon = 4/255$ and $T = 20$)

(d) CIFAR-10 ($\epsilon = 8/255$ and $T = 40$)

Figure 2: Risk and adversarial risk of the corresponding region as q varies under different settings.



(a)

(b)

Figure 3: Adversarial risk of the resulted error region with best q obtained using our method as T varies under different settings: (a) MNIST ($\epsilon = 0.1$, $\alpha = 0.01$) and CIFAR-10 ($\epsilon = 2/255$, $\alpha = 0.05$); (b) MNIST ($\epsilon = 0.2$, $\alpha = 0.01$) and CIFAR-10 ($\epsilon = 4/255$, $\alpha = 0.05$)