The Curse of Concentration in Robust Learning: Evasion and Poisoning Attacks from Concentration of Measure

Saeed Mahloujifar
University of Virginia
saeed@virginia.edu

Dimitrios I. Diochnos
University of Virginia
diochnos@virginia.edu

Mohammad Mahmoody
University of Virginia
mohammad@virginia.edu

Abstract

Many modern machine learning classifiers are shown to be vulnerable to adversarial perturbations of the instances. Despite a massive amount of work focusing on making classifiers robust, the task seems quite challenging. In this work, through a theoretical study, we investigate the adversarial risk and robustness of classifiers and draw a connection to the well-known phenomenon of "concentration of measure" in metric measure spaces. We show that if the metric probability space of the test instance is concentrated, any classifier with some initial constant error is inherently vulnerable to adversarial perturbations.

One class of concentrated metric probability spaces are the so-called Lévy families that include many natural distributions. In this special case, our attacks only need to perturb the test instance by at most $O(\sqrt{n})$ to make it misclassified, where n is the data dimension. Using our general result about Lévy instance spaces, we first recover as special case some of the previously proved results about the existence of adversarial examples. However, many more Lévy families are known (e.g., product distribution under the Hamming distance) for which we immediately obtain new attacks that find adversarial examples of distance $O(\sqrt{n})$.

Finally, we show that concentration of measure for product spaces implies the existence of forms of "poisoning" attacks in which the adversary tampers with the training data with the goal of degrading the classifier. In particular, we show that for any learning algorithm that uses m training examples, there is an adversary who can increase the probability of any "bad property" (e.g., failing on a particular test instance) that initially happens with non-negligible probability to ≈ 1 by substituting only $\widetilde{O}(\sqrt{m})$ of the examples with other (still correctly labeled) examples.

1 Introduction

Learning how to classify instances based on labeled examples is a fundamental task in machine learning. The goal is to find, with high probability, the correct label $c(x)$ of a given test instance x coming from a distribution μ. Thus, we would like to find a good-on-average "hypothesis" h (also called the trained model) that minimizes the error probability $\Pr_{x\leftarrow\mu}[h(x) \neq c(x)]$, which is referred to as the risk of h with respect to the ground truth c. Due to the explosive use of learning algorithms in real-world systems (e.g., using neural networks for image classification) a more modern approach to the classification problem aims at making the learning process, from training till testing, more *robust*. Namely, even if the instance x is perturbed in a limited way into x' by an adversary A, we would like to have the hypothesis h still predict the right label for x'; hence, minimizing the "adversarial risk"

$$\Pr_{x\leftarrow\mu}[h(x') \neq c(x') \text{ for some } x' \text{ "close" to } x]$$

of the hypothesis h under such perturbations, where "close" is defined by a metric. An attack to increase the risk is called an "evasion attack" (see e.g., (Biggio, Fumera, and Roli 2014; Carlini and Wagner 2017)) due to the fact that x' "evades" the correct classification. One major motivation behind this problem comes from scenarios such as image classification, in which the adversarially perturbed instance x' would still "look similar" to the original x, at least in humans' eyes, even though the classifier h might now misclassify x' (Goodfellow, McDaniel, and Papernot 2018). In fact, starting with the work of Szegedy et al. (Szegedy et al. 2014) an active line of research investigated various attacks and possible defenses to resist such attacks. The race between attacks and defenses in this area motivates a study of whether or not such robust classifiers could ever be found.

A closely related notion of robustness for a learning algorithm deals with the *training* phase. Here, we would like to know how much the risk of the produced hypothesis h might increase, if an adversary A tampers with the training data $\mathcal{T}$ with the goal of increasing the "error" (or any "bad" event in general) during the test phase. Such attacks are referred to as *poisoning* attacks (Biggio, Nelson, and Laskov 2012), and the line of research on the power and limitations of poisoning attacks contains numerous attacks and many defenses designed against them.

The state of affairs in attacks and defenses with regard to the robustness of learning systems in both the evasion and poisoning contexts leads us to our main question:

What are the inherent limitations of defense mechanisms for evasion and poisoning attacks? Equivalently, what are the inherent power of such attacks?

Our Results

In this work, we draw a connection between the general phenomenon of "concentration of measure" in metric measured spaces and both evasion and poisoning attacks. A concentrated metric probability space $(\mathcal{X}, \mathbf{d}, \boldsymbol{\mu})$ with metric $\mathbf{d}$ and measure $\boldsymbol{\mu}$ has the property that for any set $\mathcal{S}$ of measure at least half ($\boldsymbol{\mu}(\mathcal{S}) \geq 1/2$), most of the points in $\mathcal{X}$ according to $\boldsymbol{\mu}$, are "close" to $\mathcal{S}$ according to $\mathbf{d}$ (see Definition 2.4). We prove that for any learning problem defined over such a concentrated space, no classifier with an initial constant error (e.g., $1/100$) can be robust to adversarial perturbations. Namely, we prove the following theorem.

Theorem 1.1 (Informal)**.** *Suppose $(\mathcal{X}, \mathbf{d}, \boldsymbol{\mu})$ is a concentrated metric probability space from which the test instances are drawn. Then for any classifier h with $\Omega(1)$ initial "error" probability, there is an adversary who changes the test instance x into a "close" one and increases the risk to ≈ 1.*

In Theorem 1.1, the "error" could be any undesired event over h, c, x where h is the hypothesis, c is the concept function and x is the test instance. (See Theorem 3.2.)

The intuition behind the Theorem 1.1 is as follows. Let $\mathcal{E} = \{x \in \mathcal{X} \mid h(x) \neq c(x)\}$ be the "error region" of the hypothesis h with respect to the ground truth concept $c(\cdot)$ on an input space $\mathcal{X}$. Then, by the concentration property of $\mathcal{X}$ and that $\boldsymbol{\mu}(\mathcal{E}) \geq \Omega(1)$, we can conclude that at least half of the space $\mathcal{X}$ is "close" to $\mathcal{E}$, and by one more application of the same concentration property, we can conclude that indeed most of the points in $\mathcal{X}$ are "close" to the error region $\mathcal{E}$. Thus, an adversary who launches an evasion attack, can indeed push a typical point x into the error region by little perturbations. This above argument, is indeed inspired by the intuition behind the previous results of (Gilmer et al. 2018; Fawzi, Fawzi, and Fawzi 2018), and (Diochnos, Mahloujifar, and Mahmoody 2018) all of which use isoperimetric inequalities for *specific* metric probability spaces to prove limitations of robust classification under adversarial perturbations. Indeed, one natural way of proving concentration results is to use isoperimetric inequalities that characterize the shape of sets with minimal boundaries (and thus minimal measure after expansion). However, we emphasize that bounds on concentration of measure could be proved even if no such isoperimetric inequalities are known, and e.g., *approximate* versions of such inequalities would also be sufficient. Indeed, in addition to proofs by isoperimetric inequalities, concentration of measure results are proved using tools from various fields such as differential geometry, bounds on eigenvalues of the Laplacian, martingale methods, etc, (Milman and Schechtman 1986). Thus, by proving Theorem 1.1, we pave the way for a wide range of results against robust classification for learning problems over *any* concentrated space. To compare, the results of (Gilmer et al. 2018; Fawzi, Fawzi, and Fawzi 2018; Diochnos, Mahloujifar, and Mahmoody 2018) have better *constants* due to their use of isoperimetric inequalities, while we achieve similar asymptotic bounds with worse constants but in broader contexts.

Lévy families. A well-studied class of concentrated metric probability spaces are the so-called Lévy families (see Definition 3.5) and one special case of such families are known as *normal* Lévy families. In such spaces, when the dimension (seen as the diameter of, or the typical norm of vectors in $(\mathcal{X}, \mathbf{d})$) is n, if we expand sets with measure $1/2$ by distance b, they will cover measure at least $1 - k_1 \mathrm{e}^{-k_2 b^2/n}$ for some universal constants k_1, k_2. When translated back into the context of adversarial classification using our Theorem 1.1, we conclude that any learning task defined over a normal Lévy metric space $(\mathcal{X}, \mathbf{d}, \boldsymbol{\mu})$ guarantees the existence of (misclassified) adversarial instances that are only $\widetilde{O}(\sqrt{n})$-far from the original instance x, assuming that the original error of the classifier is only polynomially large $\geq 1/\mathrm{poly}(n)$. Interestingly, all the above-mentioned classifier-independent results on the existence of adversarial instances follow as special cases by applying our Theorem 1.1 to known normal Lévy families (i.e., the n-sphere, isotropic n-Gaussian, and the Boolean hypercube under Hamming distance). However, many more examples of normal Lévy families are known in the literature (e.g., the unit cube, the unit sphere, the special orthogonal group, symmetric group under Hamming distance, etc.) for which we immediately obtain new results. In Section 3, we list some of these examples. [1]

Relation to hardness of robust image classification. Since a big motivation for studying the hardness of classifiers against adversarial perturbations comes from the challenges that have emerged in the area of image classifications, here we comment on possible ideas from our work that might be useful for such studies. Indeed, a natural possible approach is to study whether or not the metric measure space of the images is concentrated or not. We leave such studies for interesting future work. Furthermore, the work of (Fawzi, Fawzi, and Fawzi 2018) observed that vulnerability to adversarial instances over "nice" distributions (e.g., n-Gaussian in their work, and any concentrated distribution in our work) can *potentially* imply attacks on real data *assuming* that the data is generated with a smooth generative model using the mentioned nice distributions. So, as long as one such mapping could be found for a concentrated space, our impossibility results can potentially be used for deriving similar results about the generated data as well.

The special case of product distributions. One natural family of metric probability spaces for which Theorem 1.1 entails new impossibility results are *product* measure spaces under Hamming distance. Results of (Amir and Milman 1980; Milman and Schechtman 1986; Talagrand 1995) show that such metric probability spaces are indeed normal Lévy. Therefore, we immediately conclude that, in any learning task, if the instances come from any product space of dimension n, then an adversary can perturb them to be misclassi-

[1] More formally, in Definition 3.5, the concentration function is $\mathrm{e}^{-k_2 b^2 \cdot n}$, however in many natural examples that we discuss in Section 3, the original norm required to be a Lévy family is ≈ 1, while the (expected value of the) "natural" norm is $\approx n$ where n is the dimension.

fied by only changing $O(\sqrt{n})$ of the "blocks" of the input. A special case of this result covers the case of Boolean hypercube that was recently studied by (Diochnos, Mahloujifar, and Mahmoody 2018). However, here we obtain impossibilities for *any* product space. As we will see below, concentration in such spaces are useful beyond evasion attacks.

Poisoning attacks from concentration of product spaces. One intriguing application of concentration in product measure spaces is to obtain inherent *poisoning* attacks that can attack *any* deterministic learner by tampering with their *training* data and increase their error probability during the (untampered) test phase. Indeed, since the training data is always sampled as $\mathcal{T} \leftarrow (\boldsymbol{\mu}, c(\boldsymbol{\mu}))^m$ where c is the concept function and m is the sample complexity, the concentration of the space of the training data under the Hamming distance (in which the alphabet space is the full space of labeled examples) implies that an adversary can always change the training data $\mathcal{T}$ into $\mathcal{T}'$ where $\mathcal{T}'$ by changing only a "few" examples in $\mathcal{T}$ while producing a classifier h that is more vulnerable to undesired properties.

Theorem 1.2 (Informal)**.** *Let L be any* deterministic *learning algorithm for a classification task where the confidence of L in producing a "good" hypothesis h with error at most ε is $1-\delta$ for $\delta \geq 1/\mathrm{poly}(m)$. Then, there is always a poisoning attacker who substitutes only $\widetilde{O}(\sqrt{m})$ of the training data, where m is the total number of examples, with another set of* correctly labeled *training data, and yet degrades the confidence of the produced hypothesis h to almost zero. Similarly, an attack with similar parameters can increase the average error of the generated hypothesis h over any chosen test instance x from any initial probability $\frac{1}{\mathrm{poly}(m)}$ to ≈ 1.*

More generally, both attacks of 1.2 follow as special case of a more general attack in which the adversary can pick any "bad" property of the produced hypothesis h that happens with probability at least $\geq 1/\mathrm{poly}(m)$ and increase its chance to hold with probability ≈ 1 by changing only $\widetilde{O}(\sqrt{m})$ of the training examples (with other correctly labeled examples). In fact, by allowing the bad property to be defined over the *distribution* of the produced hypothesis, we will not need L to be deterministic.

Our attacks of Theorem 1.2 are *offline* in the sense that the adversary needs to know the full training set $\mathcal{T}$ before substituting some of them. We note that the so-called p-tampering attacks of (Mahloujifar, Diochnos, and Mahmoody 2018) are *online* in the sense that the adversary can decide about its choices without the knowledge of the upcoming training examples. However, in that work, they could only increase the classification error by $O(p)$ through tampering by p fraction of the training data, while here we get almost full error by only using $p \approx O(\sqrt{m})$, which is much more devastating.

Related Work

Evasion attacks. In the context of evasion attacks, the most relevant to our main question above are the recent works of Gilmer et al. (Gilmer et al. 2018), Fawzi et al. (Fawzi, Fawzi, and Fawzi 2018), and Diochnos et al. (Diochnos, Mahloujifar, and Mahmoody 2018). In all of these works, isoperimetric inequalities for specific metric probability spaces (i.e., for uniform distributions over the n-sphere by (Gilmer et al. 2018), for isotropic n-Gaussian by (Fawzi, Fawzi, and Fawzi 2018), and for uniform distribution over the Boolean hypercube by (Diochnos, Mahloujifar, and Mahmoody 2018)) were used to prove that problems on such input spaces are always vulnerable to adversarial instances.[2] Other concurrent works have also demonstrated the role of concentration of measure in guaranteeing the existence of adversarial examples in certain metric probability spaces (Dohmatob 2018; Shafahi et al. 2018). In particular, Shafahi et al. had a tighter analysis of the robustness of classification over input distributions that are uniform over the n-dimension sphere and hyper-cube, both of which are special cases of Normal Levy families studied in this work. The work of Schmidt et al. (Schmidt et al. 2018) shows that, at least in some cases, being robust to adversarial instances requires more data. However, the work of Bubeck et al. (Bubeck, Price, and Razenshteyn 2018) proved that *assuming the existence* of classifiers that are robust to evasion attacks, they *could* be found by "few" training examples in an information theoretic way. There is also a recent line of work that which provides robustness guarantees on individual instances, up to certain degree (Kolter and Wong 2017; Raghunathan, Steinhardt, and Liang 2018; Sinha, Namkoong, and Duchi 2017; Wong et al. 2018).

Poisoning attacks. In the context of poisoning attacks, some classical results about malicious noise (Valiant 1985; Kearns and Li 1993; Bshouty, Eiron, and Kushilevitz 2002) could be interpreted as limitations of learning under poisoning attacks. On the positive (algorithmic) side, the works of Diakonikolas et al. (Diakonikolas et al. 2016) and Lai et al. (Lai, Rao, and Vempala 2016) showed the surprising power of algorithmic robust inference over poisoned data with error that does not depend on the dimension of the distribution. These works led to an active line of work (e.g., see (Charikar, Steinhardt, and Valiant 2017; Diakonikolas, Kane, and Stewart 2017; 2018; Diakonikolas et al. 2018; Prasad et al. 2018; Diakonikolas, Kong, and Stewart 2018) and references therein) exploring the possibility of robust statistics over poisoned data with algorithmic guarantees. The works of (Charikar, Steinhardt, and Valiant 2017; Diakonikolas, Kane, and Stewart 2018) performed *list-docodeable* learning, and (Diakonikolas et al. 2018; Prasad et al. 2018) studied supervised learning. Demonstrating the power of poisoning *attacks*, Mahmoody and Mahloujifar (Mahloujifar and Mahmoody 2017) showed that, assuming an initial $\Omega(1)$ error, a variant of poisoning attacks

[2] More formally, Gilmer et al. (Gilmer et al. 2018) designed specific problems over (two) n-spheres, and proved them to be hard to learn robustly, but their proof extend to any problem defined over the uniform distribution over the n-sphere. Also, Fawzi et al. (Fawzi, Fawzi, and Fawzi 2018) used a different notion of adversarial risk that only considers the hypothesis h and is independent of the ground truth c, however their proofs also extend to the same setting as ours.

that tamper with $\approx p$ fraction of the training data *without* using wrong labels (called p-tampering) could always increase the error of deterministic classifiers by $\Omega(p)$ in the targeted poisoning model (Barreno et al. 2006) where the adversary knows the final test instance. Then Mahloujifar et al. (Mahloujifar, Diochnos, and Mahmoody 2018) improved the quantitative bounds of (Mahloujifar and Mahmoody 2017) and also applied those attacks to degrade the confidence parameter of any PAC learner under poisoning attacks. Both attacks of (Mahloujifar and Mahmoody 2017; Mahloujifar, Diochnos, and Mahmoody 2018) were *online*, in the sense that the adversary does not know the future examples, and as we will see their attack model is very relevant to this work. Koh and Liang (Koh and Liang 2017) studied finding training examples with most influence over the final decision over a test instance x–enabling poisoning attacks. Note that here, we *prove* the existence of $O(\sqrt{m})$ examples in the training set that can almost fully degrade the final decision on x, assuming $\Omega(1)$ initial error on x. The work of Bousquet and Elisseeff (Bousquet and Elisseeff 2002) studied how specific forms of stability of the hypothesis (which can be seen as robustness under weak forms of "attacks" that change one training example) imply *standard* generalization (under no attack). Our work, on the other hand, studies *generalization under attack* while the adversary can perturb a lot more (but still sublinear).

Other definitions of adversarial examples. The works of Madry et al. (Madry et al. 2018) and Schmidt et al. (Schmidt et al. 2018) employ an alternative definition of adversarial risk inspired by robust optimization. This definition is reminiscent of the definition of "corrupted inputs" used by Feige et al. (Feige, Mansour, and Schapire 2015) (and related works of (Mansour, Rubinstein, and Tennenholtz 2015; Feige, Mansour, and Schapire 2018; Attias, Kontorovich, and Mansour 2018)) as in all of these works, a "successful" adversarial instance x' shall have a prediction $h(x')$ that is different from the true label of the *original* (uncorrupted) instance x. However, such definitions based on corrupted instances do not always guarantee that the adversarial examples are misclassified. In fact, even going back to the original definitions of adversarial risk and robustness from (Szegedy et al. 2014), many papers (e.g., the related work of (Fawzi, Fawzi, and Fawzi 2018)) only compare the prediction of the hypothesis over the adversarial instance with its own prediction on the honest instance, and indeed ignore the ground truth defined by the concept c.) In various "natural" settings (such as image classification) the above two definition and ours coincide. We refer the reader to the work of Diochnos et al. (Diochnos, Mahloujifar, and Mahmoody 2018) where these definitions are compared and a taxonomy is given.

2 Preliminaries

Definition 2.1 (Notation for metric spaces)**.** *Let $(\mathcal{X}, \mathbf{d})$ be a metric space. We use the notation $\mathsf{Diam}^{\mathbf{d}}(\mathcal{X}) = \sup\{\mathbf{d}(x,y) \mid x, y \in \mathcal{X}_i\}$ to denote the diameter of $\mathcal{X}$ under $\mathbf{d}$, and we use $\mathcal{B}all_b^{\mathbf{d}}(x) = \{x' \mid \mathbf{d}(x,x') \leq b\}$ to denote the ball of radius b centered at x. When $\mathbf{d}$ is clear from the context, we simply write $\mathsf{Diam}(\mathcal{X})$ and $\mathcal{B}all_b(x)$. For a set $\mathcal{S} \subseteq \mathcal{X}$, by $\mathbf{d}(x, \mathcal{S}) = \inf\{\mathbf{d}(x,y) \mid y \in \mathcal{S}\}$ we denote the distance of a point x from $\mathcal{S}$.*

Unless stated, all integrals are Lebesgue integrals.

Definition 2.2 (Nice metric probability spaces)**.** *We call $(\mathcal{X}, \mathbf{d}, \boldsymbol{\mu})$ a metric probability space, if $\boldsymbol{\mu}$ is a Borel probability measure over $\mathcal{X}$ with respect to the topology defined by $\mathbf{d}$. Then, for a Borel set $\mathcal{E} \subseteq \mathcal{X}$, the* b-expansion *of $\mathcal{E}$, denoted by $\mathcal{E}_b$, is defined as*[3]

$$\mathcal{E}_b = \{x \mid \mathbf{d}(x, \mathcal{E}) \leq b\}.$$

We call $(\mathcal{X}, \mathbf{d}, \boldsymbol{\mu})$ a nice *metric probability space, if the following conditions hold.*

1. **Expansions are measurable.** *For every $\boldsymbol{\mu}$-measurable (Borel) set $\mathcal{E} \in \mathcal{X}$, and every $b \geq 0$, its b-expansion $\mathcal{E}_b$ is also $\boldsymbol{\mu}$-measurable.*
2. **Average distances exist.** *For every two Borel sets $\mathcal{E}, \mathcal{S} \in \mathcal{X}$, the average minimum distance of an element from $\mathcal{S}$ to $\mathcal{E}$ exists; namely, the integral $\int_{\mathcal{S}} \mathbf{d}(x, \mathcal{E}) \cdot d\boldsymbol{\mu}(x)$ exists.*

At a high level, and as we will see shortly, we need the first condition to define adversarial risk and need the second condition to define (a generalized notion of) robustness. Also, we remark that one can weaken the second condition above based on the first one and still have risk and robustness defined, but since our goal in this work is *not* to do a measure theoretic study, we are willing to make simplifying assumptions that hold on the actual applications, if they make the presentation simpler.

Notation on learning problems. We use calligraphic letters (e.g., $\mathcal{X}$) for sets. By $x \leftarrow \boldsymbol{\mu}$ we denote sampling x from the probability measure $\boldsymbol{\mu}$. For a randomized algorithm $R(\cdot)$, by $y \leftarrow R(x)$ we denote the randomized execution of R on input x outputting y. A classification problem $(\mathcal{X}, \mathcal{Y}, \boldsymbol{\mu}, \mathcal{C}, \mathcal{H})$ is specified by the following components. The set $\mathcal{X}$ is the set of possible *instances*, $\mathcal{Y}$ is the set of possible *labels*, $\boldsymbol{\mu}$ is a distribution over $\mathcal{X}$, $\mathcal{C}$ is a class of *concept* functions where $c \in \mathcal{C}$ is always a mapping from $\mathcal{X}$ to $\mathcal{Y}$. We did not state the loss function explicitly, as we work with classification problems. For $x \in \mathcal{X}, c \in \mathcal{C}$, the *risk* or *error* of a hypothesis $h \in \mathcal{H}$ is equal to $\mathsf{Risk}(h, c) = \Pr_{x \leftarrow \boldsymbol{\mu}}[h(x) \neq c(x)]$. We are usually interested in learning problems $(\mathcal{X}, \mathcal{Y}, \boldsymbol{\mu}, \mathcal{C}, \mathcal{H})$ with a specific metric $\mathbf{d}$ defined over $\mathcal{X}$ for the purpose of defining risk and robustness under instance perturbations controlled by metric $\mathbf{d}$. In that case, we simply write $(\mathcal{X}, \mathcal{Y}, \boldsymbol{\mu}, \mathcal{C}, \mathcal{H}, \mathbf{d})$ to include $\mathbf{d}$.

Definition 2.3 (Nice classification problems)**.** *We call $(\mathcal{X}, \mathcal{Y}, \boldsymbol{\mu}, \mathcal{C}, \mathcal{H}, \mathbf{d})$ a* nice *classification problem, if the following two conditions hold:*

1. *$(\mathcal{X}, \mathbf{d}, \boldsymbol{\mu})$ is a nice metric probability space.*
2. *For every $h \in \mathcal{H}, c \in \mathcal{C}$, their error region $\{x \in \mathcal{X} \mid h(x) \neq c(x)\}$ is $\boldsymbol{\mu}$-measurable.*

[3] The set $\mathcal{E}_b$ is also called the b-flattening or b-enlargement of $\mathcal{E}$, or simply the b-ball around A.

The second condition above is satisfied, e.g., if the set of labels $\mathcal{Y}$ (which is usually finite) is countable, and for all $y \in \mathcal{Y}$, $f \in \mathcal{H} \cup \mathcal{C}$, the set $\{x \in \mathcal{X} \mid f(x) = y\}$ is $\boldsymbol{\mu}$-measurable.

We now formally define the (standard) notion of concentration function.

Definition 2.4 (Concentration function)**.** *Let $(\mathcal{X}, \mathbf{d}, \boldsymbol{\mu})$ be a metric probability space and $\mathcal{E} \subseteq \mathcal{X}$ be a Borel set. The* concentration function *is then defined as*

$$\boldsymbol{\alpha}(b) = 1 - \inf \{\boldsymbol{\mu}(\mathcal{E}_b) \mid \boldsymbol{\mu}(\mathcal{E}) \geq 1/2\}.$$

Variations of the following Lemma 2.5 below are in (Amir and Milman 1980; Milman and Schechtman 1986), but the following version is due to Talagrand (Talagrand 1995).

Lemma 2.5 (Concentration of product spaces under Hamming distance)**.** *Let $\boldsymbol{\mu} \equiv \boldsymbol{\mu}_1 \times \cdots \times \boldsymbol{\mu}_n$ be a product probability measure of dimension n and let the metric be the Hamming distance. For any $\boldsymbol{\mu}$-measurable $\mathcal{S} \subseteq \mathcal{X}$ such that the b-expansion $\mathcal{S}_b$ of $\mathcal{S}$ under Hamming distance is measurable, it holds that*

$$\boldsymbol{\mu}(\mathcal{S}_b) \geq 1 - \frac{\mathrm{e}^{-b^2/n}}{\boldsymbol{\mu}(\mathcal{S})}.$$

Lemma 2.6 (McDiarmid inequality)**.** *Let $\boldsymbol{\mu} \equiv \boldsymbol{\mu}_1 \times \cdots \times \boldsymbol{\mu}_n$ be a product probability measure of dimension n, and let $f \colon \mathrm{Supp}(\boldsymbol{\mu}) \mapsto \mathbb{R}$ be a measurable function such that $|f(x) - f(y)| \leq 1$ whenever x and y* only *differ in one coordinate. If $a = \mathbf{E}_{x \leftarrow \boldsymbol{\mu}}[f(x)]$, then*

$$\Pr_{x \leftarrow \boldsymbol{\mu}}[f(x) \leq a - b] \leq \mathrm{e}^{-2 \cdot b^2/n}.$$

3 Evasion Attacks: Finding Adversarial Examples from Concentration

In this section, we formally prove our main results about the existence of evasion attacks for learning problems over concentrated spaces. We start by formalizing the notions of risk and robustness.

Definition 3.1 (Adversarial risk and robustness)**.** *Let $(\mathcal{X}, \mathcal{Y}, \boldsymbol{\mu}, \mathcal{C}, \mathcal{H}, \mathbf{d})$ be a nice classification problem. For $h \in \mathcal{H}$ and $c \in \mathcal{C}$, let $\mathcal{E} = \{x \in \mathcal{X} \mid h(x) \neq c(x)\}$ be the error region of h with respect to c. Then, we define:*

- **Adversarial risk.** *For $b \in \mathbb{R}_+$, the (error-region)* adversarial risk *under b-perturbation is*

  $$\mathsf{Risk}_b(h, c) = \Pr_{x \leftarrow \boldsymbol{\mu}}[\exists\, x' \in \mathcal{B}all_b(x) \cap \mathcal{E}] = \boldsymbol{\mu}(\mathcal{E}_b).$$

  We might call b the "budget" of an imaginary "adversary" who perturbs x into x'. Using $b = 0$, we recover the standard notion of risk: $\mathsf{Risk}(h, c) = \mathsf{Risk}_0(h, c) = \boldsymbol{\mu}(\mathcal{E})$.
- **Target-error robustness.** *Given a target error $\rho \in (0, 1]$, we define the ρ-error robustness as the expected perturbation needed to increase the error to ρ; namely,*

  $$\mathsf{Rob}_\rho(h, c) = \inf_{\boldsymbol{\mu}(\mathcal{S}) \geq \rho} \left\{ \mathop{\mathbf{E}}_{x \leftarrow \boldsymbol{\mu}} [\mathbf{1}_{\mathcal{S}}(x) \cdot \mathbf{d}(x, \mathcal{E})] \right\}$$

  where $\mathbf{1}_{\mathcal{S}}(x)$ is the characteristic function of membership in $\mathcal{S}$. Letting $\rho = 1$, we recover the notion of full robustness $\mathsf{Rob}(h, c) = \mathsf{Rob}_1(h, c) = \mathbf{E}_{x \leftarrow \boldsymbol{\mu}}[\mathbf{d}(x, \mathcal{E})]$ *that captures the expected amount of perturbations needed to* always *change x into a misclassified x' where $x' \in \mathcal{E}$.*

As discussed in the introduction, starting with (Szegedy et al. 2014), many papers (e.g., the related work of (Fawzi, Fawzi, and Fawzi 2018)) use a definitions of risk and robustness that *only* deal with the hypothesis/model and is independent of the concept function. In (Diochnos, Mahloujifar, and Mahmoody 2018), that definition is formalized as "prediction change" (PC) adversarial risk and robustness.

In the rest of this section, we focus on misclassification as a necessary condition for the adversarial instances. So, we use Definition 3.1 to prove our results.

Increasing Risk and Decreasing Robustness

We now formally state and prove our result that the adversarial risk can be large for any learning problem over concentrated spaces. Note that, even though the following is stated using the concentration function, having an *upper bound* on the concentration function suffices for using it. Also, we note that all the results of this section extend to settings in which the "error region" is substituted with any "bad" event modeling an undesired region of instances based on the given hypothesis h and concept function c; though the most natural bad event is that error $h(x) \neq c(x)$ occurs.

Theorem 3.2 (From concentration to large adversarial risk)**.** *Let $(\mathcal{X}, \mathcal{Y}, \boldsymbol{\mu}, \mathcal{C}, \mathcal{H}, \mathbf{d})$ be a nice classification problem. Let $h \in \mathcal{H}$ and $c \in \mathcal{C}$, and let $\varepsilon = \Pr_{x \leftarrow \boldsymbol{\mu}}[h(x) \neq c(x)]$ be the error of the hypothesis h with respect to the concept c. If $\varepsilon > \boldsymbol{\alpha}(b)$ (i.e., the original error is more than the concentration function for the budget b), then the following two hold.*

1. **Reaching adversarial risk at least half.** *Using only tampering budget b, the adversary can make the adversarial risk to be more than half; namely,* $\mathsf{Risk}_b(h, c) > 1/2$.
2. **Reaching adversarial risk close to one.** *If in addition we have $\gamma \geq \boldsymbol{\alpha}(b_2)$, then the adversarial risk for the total tampering budget $b_1 + b_2$ is* $\mathsf{Risk}_{b_1 + b_2}(h, c) \geq 1 - \gamma$.

The above theorem provides a general result that applies to *any* concentrated space. So, even though we will compute explicit bounds for spaces such as Lévy families, Theorem 3.2 could be applied to any other concentrated space as well, leading to stronger or weaker bounds than what Lévy families offer. Now, in the following, we go after finding general relations between the concentration function and the robustness of the learned models.

Simplifying notation. Suppose $(\mathcal{X}, \mathbf{d}, \boldsymbol{\mu})$ is a nice metric probability space. Since our risk and robustness definitions depend only on the error region, for any Borel set $\mathcal{E} \subseteq \mathcal{X}$ and $b \in \mathbb{R}_+$, we define its *b-tampering risk* as $\mathsf{Risk}_b(\mathcal{E}) = \boldsymbol{\mu}(\mathcal{E}_b)$, and for any such $\mathcal{E}$ and $\rho \in (0, 1]$, we define the ρ-error robustness as $\mathsf{Rob}_\rho(\mathcal{E}) = \inf_{\boldsymbol{\mu}(\mathcal{S}) \geq \rho} \left\{ \int_{\mathcal{S}} \mathbf{d}(x, \mathcal{E}) \cdot d\boldsymbol{\mu}(x) \right\}$.

The following lemma provides a very useful tool for going from adversarial risk to robustness; hence, allowing us to connect concentration of spaces to robustness. In fact, the lemma could be of independent interest, as it states a relation between *worst-case* concentration of metric probability spaces to their *average-case* concentration with a *targeted* amount of measure to cover.

Lemma 3.3 (From adversarial risk to target-error robustness)**.** *For a nice metric probability space $(\mathcal{X}, \mathbf{d}, \boldsymbol{\mu})$, let $\mathcal{E} \subseteq \mathcal{X}$ be a Borel set. If $\rho = \mathsf{Risk}_\ell(\mathcal{E})$, then we have*

$$\mathsf{Rob}_\rho(\mathcal{E}) = \rho \cdot \ell - \int_{z=0}^{\ell} \mathsf{Risk}_z(\mathcal{E}) \cdot dz.$$

Now we make a few comments on using Lemma 3.3.

Special case of full robustness. Lemma 3.3 can be used to compute the full robustness also as

$$\mathsf{Rob}(\mathcal{E}) = \mathsf{Rob}_1(\mathcal{E}) = \ell - \int_{z=0}^{\ell} \mathsf{Risk}_z(\mathcal{E}) \cdot dz, \quad (1)$$

using any $\ell \geq \mathsf{Diam}(\mathcal{X})$, because for such ℓ we will have $\mathsf{Risk}_\ell(\mathcal{E}) = 1$. In fact, even if the diameter is not finite, we can always use $\ell = \infty$ and rewrite the two terms as

$$\mathsf{Rob}(\mathcal{E}) = \int_{z=0}^{\infty} (1 - \mathsf{Risk}_z(\mathcal{E})) \cdot dz, \quad (2)$$

which might or might not converge.

When we only have *lower bounds* for adversarial risk. Lemma 3.3, as written, requires the exact amount of risk for the initial set $\mathcal{E}$. Now, suppose we only have a lower bound $L_z(\mathcal{E}) \leq \mathsf{Risk}_z(\mathcal{E})$ for the adversarial risk. In this case, we can still use Lemma 3.3, but it will only give us an *upper bound* on the ρ-error robustness using any ℓ such that $\rho \leq L_\ell(\mathcal{E})$ as follows,

$$\mathsf{Rob}_\rho(\mathcal{E}) \leq \ell - \int_{z=0}^{\ell} L_z(\mathcal{E}) \cdot dz. \quad (3)$$

Note that, even though the above bound looks similar to that of full robustness in Equation 1, in Inequality 3 we can use $\ell < \mathsf{Diam}(\mathcal{X})$, which leads to a smaller total bound on the ρ-error robustness.

We now formally state our result that concentration in the instance space leads to small robustness of classifiers. Similarly to Theorem 3.2, we note that even though the following theorem is stated using the concentration function, having an upper bound on the concentration function would suffice.

Theorem 3.4 (From concentration to small robustness)**.** *Let $(\mathcal{X}, \mathcal{Y}, \boldsymbol{\mu}, \mathcal{C}, \mathcal{H}, \mathbf{d})$ be a nice classification problem. Let $h \in \mathcal{H}$ and $c \in \mathcal{C}$, and let $\varepsilon = \Pr_{x \leftarrow \boldsymbol{\mu}}[h(x) \neq c(x)]$ be the error of the hypothesis h with respect to the concept c. Then if $\varepsilon > \boldsymbol{\alpha}(b_1)$ and $1 - \rho \geq \boldsymbol{\alpha}(b_2)$, we have*

$$\mathsf{Rob}_\rho(\mathcal{E}) \leq (1 - \varepsilon) \cdot b_1 + \int_{z=0}^{b_2} \boldsymbol{\alpha}(z) \cdot dz.$$

Normal Lévy Families as Concentrated Spaces

In this subsection, we study a well-known special case of concentrated spaces called normal Lévy families, as a rich class of concentrated spaces, leading to specific bounds on the risk and robustness of learning problems whose test instances come from any normal Lévy family. We start by formally defining normal Lévy families.

Definition 3.5 (Normal Lévy families)**.** *A family of metric probability spaces $(\mathcal{X}_n, \mathbf{d}_n, \boldsymbol{\mu}_n)_{i \in \mathbb{N}}$ with corresponding concentration functions $\boldsymbol{\alpha}_n(\cdot)$ is called a* (k_1, k_2)-normal Lévy family *if*

$$\boldsymbol{\alpha}_n(b) \leq k_1 \cdot \mathrm{e}^{-k_2 \cdot b^2 \cdot n}.$$

The following theorem shows that classifying instances that come from a normal Lévy family has the inherent vulnerability to perturbations of size $O(1/\sqrt{n})$

Theorem 3.6 (Risk and robustness in normal Lévy families)**.** *Let $(\mathcal{X}_n, \mathcal{Y}_n, \boldsymbol{\mu}_n, \mathcal{C}_n, \mathcal{H}_n, \mathbf{d}_n)_{n \in \mathbb{N}}$ be a nice classification problem with a metric probability space $(\mathcal{X}_n, \mathbf{d}_n, \boldsymbol{\mu}_n)_{n \in \mathbb{N}}$ that is a (k_1, k_2)-normal Lévy family. Let $h \in \mathcal{H}_n$ and $c \in \mathcal{C}_n$, and let $\varepsilon = \Pr_{x \leftarrow \boldsymbol{\mu}}[h(x) \neq c(x)]$ be the error of the hypothesis h with respect to the concept c.*

1. **Reaching adversarial risk at least half.** *If $b > \sqrt{\ln(k_1/\varepsilon)}/\sqrt{k_2 \cdot n}$, then $\mathsf{Risk}_b(h, c) \geq 1/2$.*
2. **Reaching Adversarial risk close to one.** *If $b > \sqrt{\ln(k_1/\varepsilon) + \ln(k_1/\gamma)}/\sqrt{k_2 \cdot n}$, then it holds that $\mathsf{Risk}_b(h, c) \geq 1 - \gamma$.*
3. **Bounding target-error robustness.** *For any $\rho \in [\frac{1}{2}, 1]$, we have*

$$\mathsf{Rob}_\rho(h, c) \leq \frac{(1-\varepsilon)\sqrt{\ln(\frac{k_1}{\varepsilon})}}{\sqrt{k_2 \cdot n}} + \frac{\mathsf{erf}\big(\sqrt{\ln(\frac{k_1}{(1-\rho)})}\big) \cdot \frac{k_1\sqrt{\pi}}{2}}{\sqrt{k_2 \cdot n}}.$$

Here we remark on its interpretation in an asymptotic sense, and discuss how much initial error is needed to achieve almost full adversarial risk.

Examples of Normal Lévy Families. Here, we list some natural metric probability spaces that are known to be normal Lévy families. For the references and more examples we refer the reader to excellent sources (Ledoux 2001; Giannopoulos and Milman 2001; Milman and Schechtman 1986). There are other variants of Lévy families, e.g., those called Lévy (without the adjective "normal") or *concentrated* Lévy families (Alon and Milman 1985) with stronger concentration, but we skip them and refer the reader to the cited sources and general tools of Theorems 3.2 and 3.4 on how to apply *any* concentration of measure results to get bounds on risk and robustness of classifiers. 1. Unit sphere with uniform distribution under Euclidean or Geodesic distance. 2. $\mathbb{R}^n$ under Gaussian distribution and Euclidean distance. 3. Unit cube and unit ball under the uniform distribution and Euclidean distance. 4. Product distributions under Hamming distance. 5. Symmetric group under with uniform distribution and under Hamming distance.

4 Poisoning Attacks from Concentration of Product Measures

In this section, we design new poisoning attacks against any deterministic learning algorithm, by using the concentration of space in the domain of training data. We start by defining the confidence and error parameters of learners.

Now, we formally define the class of poisoning attacks and their properties.

Definition 4.1 (Poisoning attacks)**.** *Let $(\mathcal{X}, \mathcal{Y}, \mathbf{\mu}, \mathcal{H}, \mathcal{C})$ be a classification with a learning algorithm L. Then, a poisoning adversary A for $(L, \mathcal{X}, \mathcal{Y}, \mathbf{\mu}, \mathcal{H}, \mathcal{C})$ is an algorithm that takes as input a training set $\mathcal{T} \leftarrow (\mathbf{\mu}, c(\mathbf{\mu}))^m$ and outputs a modified training set $\mathcal{T}' = A(\mathcal{T})$ of the same size[4]. We also interpret $\mathcal{T}$ and $\mathcal{T}$ as vectors with m coordinates with a large alphabet and let* HD *be the Hamming distance for such vectors of m coordinates. For any $c \in \mathcal{C}$, we define the following properties for A.*

- *A is called* plausible*, if $y = c(x)$ for all $(x, y) \in \mathcal{T}'$.*
- *A has* tampering budget *$b \in [m]$ if for all $\mathcal{T} \leftarrow (\mathbf{\mu}, c(\mathbf{\mu}))^m, \mathcal{T}' \leftarrow A(\mathcal{T})$, we have $\mathsf{HD}(\mathcal{T}', \mathcal{T}) \leq b$.*
- *A has* average *tampering budget b, if we have:*

$$\mathop{\mathbf{E}}_{\mathcal{T} \leftarrow (\mathbf{\mu}, c(\mathbf{\mu}))^m, \mathcal{T}' \leftarrow A(\mathcal{T})} [\mathsf{HD}(\mathcal{T}', \mathcal{T}))] \leq b.$$

Before proving our results about the power of poisoning attacks, we need to define the confidence function of a learning algorithm under such attacks.

Definition 4.2 (Confidence function and its adversarial variant)**.** *For a learning algorithm L for a classification problem $(\mathcal{X}, \mathcal{Y}, \mathbf{\mu}, \mathcal{H}, \mathcal{C})$ and for a predicate $p\colon \mathcal{H} \to \{0, 1\}$, we use Conf_A to define the* adversarial confidence *in the presence of a poisoning adversary A as follows*

$$\mathsf{Conf}_A(m, c, p) = \Pr_{\mathcal{T} \leftarrow (\mathbf{\mu}, c(\mathbf{\mu}))^m, h \leftarrow L(A(\mathcal{T}))} [p(h) = 0].$$

By $\mathsf{Conf}(\cdot)$, we denote L's confidence function *without any attack; namely, $\mathsf{Conf}(\cdot) = \mathsf{Conf}_I(\cdot)$ for the trivial (identity) attacker I that does not change the training data.*

Increasing a Bad Event's Probability by Poisoning

The following theorem formalizes (the first part of) Theorem 1.2. We emphasize that by choosing the adversary *after* the concept function is fixed, we allow the adversary to depend on the concept class. This is also the case in e.g., p-tampering poisoning attacks of (Mahloujifar, Diochnos, and Mahmoody 2018). However, there is a big distinction between our attacks here and those of (Mahloujifar, Diochnos, and Mahmoody 2018), as our attackers need to know the *entire* training sequence before tampering with them, while the attacks of Mahloujifar et al. were online.

Theorem 4.3. *Consider a classification problem $(\mathcal{X}, \mathcal{Y}, \mathbf{\mu}, \mathcal{H}, \mathcal{C})$, a deterministic learner L, a concept $c \in \mathcal{C}$ and a bad predicate $p : \mathcal{H} \to \{0, 1\}$. Also let $\mathsf{Conf}(m, c, p) = 1 - \delta$ be the original confidence of L.*

1. *For any $\gamma \in [0, 1]$, there is a plausible poisoning adversary A with tampering budget at most $\sqrt{-\ln(\delta \cdot \gamma) \cdot m}$ such that, A makes the adversarial confidence to be as small as γ. Namely, $\mathsf{Conf}_A(m, c, p) \leq \gamma$.*
2. *There is a plausible poisoning adversary A with* average *tampering budget $\sqrt{-\ln(\delta) \cdot m/2}$ eliminating all the confidence. Namely, $\mathsf{Conf}_A(m, c, p) = 0$.*

[4]Requiring the sets to be equal only makes our *negative* attacks *stronger*.

Remark 4.4 (Examples of undesired predicates)**.** *The predicate p in above theorem could be any undesirable property. For instance, it could be a Boolean function indicating weather the error of the hypothesis is greater than some threshold. Or it could be equal to one if the hypothesis outputs a wrong label on a specific target instance.*

5 Acknowledgements

The first author is supported by University of Virginia's SEAS Research Innovation Awards. The third author is supported by NSF CAREER award CCF-1350939, and two University of Virginia's SEAS Research Innovation Awards.

References

Alon, N., and Milman, V. D. 1985. $\lambda 1$, isoperimetric inequalities for graphs, and superconcentrators. *Journal of Combinatorial Theory, Series B* 38(1):73–88.

Amir, D., and Milman, V. 1980. Unconditional and symmetric sets inn-dimensional normed spaces. *Israel Journal of Mathematics* 37(1-2):3–20.

Attias, I.; Kontorovich, A.; and Mansour, Y. 2018. Improved generalization bounds for robust learning. *arXiv preprint arXiv:1810.02180*.

Barreno, M.; Nelson, B.; Sears, R.; Joseph, A. D.; and Tygar, J. D. 2006. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, 16–25. ACM.

Biggio, B.; Fumera, G.; and Roli, F. 2014. Security evaluation of pattern classifiers under attack. *IEEE transactions on knowledge and data engineering* 26(4):984–996.

Biggio, B.; Nelson, B.; and Laskov, P. 2012. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 1467–1474. Omnipress.

Bousquet, O., and Elisseeff, A. 2002. Stability and generalization. *Journal of machine learning research* 2(Mar):499–526.

Bshouty, N. H.; Eiron, N.; and Kushilevitz, E. 2002. PAC learning with nasty noise. *Theoretical Computer Science* 288(2):255–275.

Bubeck, S.; Price, E.; and Razenshteyn, I. 2018. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*.

Carlini, N., and Wagner, D. A. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, 39–57.

Charikar, M.; Steinhardt, J.; and Valiant, G. 2017. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 47–60. ACM.

Diakonikolas, I.; Kamath, G.; Kane, D. M.; Li, J.; Moitra, A.; and Stewart, A. 2016. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, 655–664. IEEE.

Diakonikolas, I.; Kamath, G.; Kane, D. M.; Li, J.; Steinhardt, J.; and Stewart, A. 2018. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*.

Diakonikolas, I.; Kane, D. M.; and Stewart, A. 2017. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, 73–84. IEEE.

Diakonikolas, I.; Kane, D. M.; and Stewart, A. 2018. List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 1047–1060. ACM.

Diakonikolas, I.; Kong, W.; and Stewart, A. 2018. Efficient algorithms and lower bounds for robust linear regression. *arXiv preprint arXiv:1806.00040*.

Diochnos, D.; Mahloujifar, S.; and Mahmoody, M. 2018. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Advances in Neural Information Processing Systems*, 10380–10389.

Dohmatob, E. 2018. Limitations of adversarial robustness: strong no free lunch theorem. *arXiv preprint arXiv:1810.04065*.

Fawzi, A.; Fawzi, H.; and Fawzi, O. 2018. Adversarial vulnerability for any classifier. *arXiv preprint arXiv:1802.08686*.

Feige, U.; Mansour, Y.; and Schapire, R. 2015. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, 637–657.

Feige, U.; Mansour, Y.; and Schapire, R. E. 2018. Robust inference for multiclass classification. In *Algorithmic Learning Theory*, 368–386.

Giannopoulos, A. A., and Milman, V. D. 2001. Euclidean structure in finite dimensional normed spaces. *Handbook of the geometry of Banach spaces* 1:707–779.

Gilmer, J.; Metz, L.; Faghri, F.; Schoenholz, S. S.; Raghu, M.; Wattenberg, M.; and Goodfellow, I. 2018. Adversarial Spheres. *arXiv preprint arXiv:1801.02774*.

Goodfellow, I. J.; McDaniel, P. D.; and Papernot, N. 2018. Making machine learning robust against adversarial inputs. *Communications of the ACM* 61(7):56–66.

Kearns, M. J., and Li, M. 1993. Learning in the Presence of Malicious Errors. *SIAM J. on Computing* 22(4):807–837.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 1885–1894.

Kolter, J. Z., and Wong, E. 2017. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851* 1(2):3.

Lai, K. A.; Rao, A. B.; and Vempala, S. 2016. Agnostic estimation of mean and covariance. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, 665–674. IEEE.

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

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.

Mahloujifar, S., and Mahmoody, M. 2017. Blockwise p-Tampering Attacks on Cryptographic Primitives, Extractors, and Learners. In *Theory of Cryptography Conference*, 245–279. Springer.

Mahloujifar, S.; Diochnos, D. I.; and Mahmoody, M. 2018. Learning under p-Tampering Attacks. In *ALT*, 572–596.

Mansour, Y.; Rubinstein, A.; and Tennenholtz, M. 2015. Robust probabilistic inference. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, 449–460. Society for Industrial and Applied Mathematics.

Milman, V. D., and Schechtman, G. 1986. *Asymptotic theory of finite dimensional normed spaces*, volume 1200. Springer.

Prasad, A.; Suggala, A. S.; Balakrishnan, S.; and Ravikumar, P. 2018. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.

Raghunathan, A.; Steinhardt, J.; and Liang, P. 2018. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*.

Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; and Madry, A. 2018. Adversarially Robust Generalization Requires More Data. *arXiv preprint arXiv:1804.11285*.

Shafahi, A.; Huang, W. R.; Studer, C.; Feizi, S.; and Goldstein, T. 2018. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*.

Sinha, A.; Namkoong, H.; and Duchi, J. 2017. Certifiable distributional robustness with principled adversarial training. *stat* 1050:29.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *ICLR*.

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

Valiant, L. G. 1985. Learning disjunctions of conjunctions. In *IJCAI*, 560–566.

Wong, E.; Schmidt, F.; Metzen, J. H.; and Kolter, J. Z. 2018. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, 8410–8419.