Does Data Augmentation Lead to Positive Margin?

Shashank Rajput [*1] **Zhili Feng** [*1] **Zachary Charles** [2] **Po-Ling Loh** [3] **Dimitris Papailiopoulos** [2]

Abstract

Data augmentation (DA) is commonly used during model training, as it significantly improves test error and model robustness. DA artificially expands the training set by applying random noise, rotations, crops, or even adversarial perturbations to the input data. Although DA is widely used, its capacity to provably improve robustness is not fully understood. In this work, we analyze the robustness that DA begets by quantifying the *margin* that DA enforces on empirical risk minimizers. We first focus on linear separators, and then a class of nonlinear models whose labeling is constant within small convex hulls of data points. We present lower bounds on the number of augmented data points required for non-zero margin, and show that commonly used DA techniques may only introduce significant margin after adding exponentially many points to the data set.

1. Introduction

Modern machine learning has ushered in a plethora of advances in data science and engineering, which leverage models with millions of tunable parameters and achieve unprecedented accuracy on many vision, speech, and text prediction tasks. For state-of-the-art performance, model training involves stochastic gradient descent (SGD), combined with regularization, momentum, data augmentation, and other heuristics. Several empirical studies (Zhang et al., 2016; Zantedeschi et al., 2017) observe that among these methods, data augmentation plays a central role in improving the test error performance and robustness of these models.

Data augmentation (DA) expands the training set with artificial data points. For example, Krizhevsky et al. (2012) augmented ImageNet using translations, horizontal reflections, and altered intensities of the RGB channels of images in the training set. Others have augmented datasets by adding labels to sparsely annotated videos (Misra et al., 2015; Kuznetsova et al., 2015; Prest et al., 2012). Another important class of data augmentation methods are referred to broadly as *adversarial training*. Such methods use adversarial examples (Szegedy et al., 2013; Madry et al., 2017) to enlarge the training set. Many works have since shown that by training models on these adversarial examples, we can increase the robustness of learned models (Bastani et al., 2016; Carlini & Wagner, 2017; Szegedy et al., 2013; Goodfellow et al., 2014). Recently, (Ford et al., 2019) studied the use of additive Gaussian DA in ensuring robustness of learned classifiers. While they showed the approach can have some limited success, ensuring robustness to adversarial attacks requires augmenting the data set with Gaussian noise of particularly high variance.

The high-level motivation of DA is clear: a reliable model should be trained to predict the same class even if an image is slightly perturbed. Despite its empirical effectiveness, relatively few works theoretically analyze the performance and limitations of DA. Bishop (1995) shows that training with noise is equivalent to Tikhonov regularization in expectation. Wager et al. (2013) show that training generalized linear models while randomly dropping features is approximately equivalent to ℓ_2-regularization normalized by the inverse diagonal Fisher information matrix. Dao et al. (2018) study data augmentation as feature-averaging and variance regularization, using a Markov process to augment the dataset. Wong & Kolter (2018) provide a provable defense against bounded ℓ_∞-attacks by training on a convex relaxation of the "adversarial polytope," which is also a form of DA.

We take a different path by analyzing how DA impacts the margin of a classifier, *i.e.*, the minimum distance from the training data to its decision boundary. We focus on margin since it acts as a proxy for both generalization (Shalev-Shwartz & Ben-David, 2014) and worst-case robustness. In particular, we analyze how much data augmentation is necessary in order to ensure that any empirical risk minimization algorithm achieves positive, or even large, margin. To the best of our knowledge, no existing work has explicitly analyzed data augmentation through the lens of margin.

[*]Equal contribution [1]Department of Computer Science, University of Wisconsin-Madison [2]Department of Electrical and Computer Engineering, University of Wisconsin-Madison [3]Department of Statistics, University of Wisconsin-Madison. Correspondence to: Shashank Rajput <rajput3@wisc.edu>, Zhili Feng <zfeng49@cs.wisc.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

1.1. Contributions

We consider the following empirical risk minimization (ERM) problem:

$$\mathcal{A}(S) = \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \sum_{i=1}^{n} \ell(f(x_i), y_i) \right\}$$

where $S = \{(x_i, y_i)\}_{i=1}^n$ is the training set, $x_i \in \mathbb{R}^d$ are the feature vectors, and $y_i \in \{-1, +1\}$ their labels. $\mathcal{F}$ is the set of classifiers we are optimizing over, and $\ell(f(x), y) = \mathbf{1}_{\{f(x) \neq y\}}$ is the $0/1$ loss quantifying the discrepancy between the predicted label $f(x)$ and the truth.

For the purpose of better generalization and robustness, we often desire an ERM solution with large margin. A classifier f has margin ϵ with respect to some p-norm, if $(x, y) \in S$ then for any $\delta \in \mathbb{R}^d$ with $\|\delta\|_p \leq \epsilon$, $f(x) = f(x + \delta) = y$. While margin can be explicitly enforced through constraints or regularization for linear classifiers, doing so efficiently and provably for general classifiers remains a challenging open problem. Since data augmentation has had success in offering better robustness in practice, we ask the following question:

Can data augmentation guarantee non-zero margin?

That is, can we use an augmented data set S^{aug}, such that by applying any ERM to it, the output classifier $\mathcal{A}(S^{\text{aug}})$ has some margin? Figure 1 provides a sketch of this problem for linear classification.

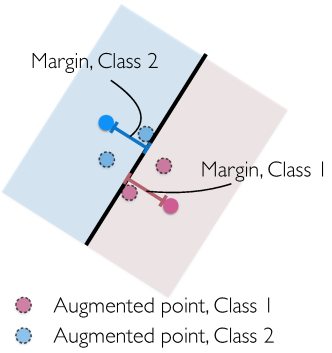

Figure 1. A linearly separable data set with two data points, each in its own class, and two input dimensions. If we wish to guarantee a positive margin for all feasible linear separators, i.e., all linear ERMs, we need to augment the training set with additional data points. Otherwise, a linear separator exists with zero margin.

Lower bounds on the number of augmentations. We first consider linear classification of linearly separable data. We develop lower bounds on the number of augmented data points needed to guarantee that any linear separator of the augmented data has positive margin with respect to the original data set. We show that in d dimensions, $d+1$ augmented data points are necessary for any data augmentation strategy to achieve positive margin. Moreover, there is some strategy that achieves the best possible margin with only $d + 1$ augmented points. However, if the augmented points are formed by bounded perturbations of the training set, we need at least as many augmented data points as true training points to ensure positive margin.

Upper bounds for additive random perturbations. In practice, many data augmentation methods employ random perturbations, including random crops, rotations, and additive noise. As a first step towards analyzing these methods, we focus on the setting that the augmented data set is formed by adding spherical random noise to the original training data. We specifically quantify how the dimension of the data, the number of augmentations per data point, and their norm can impact the worst-case margin. Our results show that if the norm of the additive noise is proportional to the margin, then the number of augmented data points must be exponential to ensure a constant factor approximation of the best possible margin. However, if the norm of the additive noise is carefully chosen, then polynomially many augmentations are sufficient to guarantee that any sperateor of the augmented data set has margin that is a constant fraction of the max margin of the original data set.

Nonlinear classification and margin. Finally, we extend our results to nonlinear classifiers that assign the same label within small convex hulls of the training data. We provide lower bounds on the number of augmentations needed for such "respectful" classifiers to achieve positive margin, and also analyze their margin under random DA methods. Despite respectful classifiers being significantly more general than linear ones, we show that their worst-case margin after augmentation can be comparable to that of linear classifiers.

1.2. Related Work

DA is closely related to robust optimization methods (Xu et al., 2009; Caramanis et al., 2012; Sinha et al., 2018; Wong & Kolter, 2018). While DA aims at improving model robustness via finitely many perturbations of the input data, robust optimization methods solve robust versions of ERM, which typically involve worst-case perturbations over infinite sets. Our work has particularly strong connections to Xu et al. (2009), which shows that regularized SVMs are equivalent to robust versions of linear classification. Our results can be viewed as attempting to train robust models without the need to perform robust optimization.

Our work may also be viewed as quantifying the robustness of classifiers trained with DA against adversarial (i.e., worst-case) perturbations. Many recent works have analyzed the robustness of various classifiers to adversarial perturbations from a geometric perspective. Fawzi et al. (2016) introduce

a notion of semi-random noise and study the robustness of classifiers to this noise in terms of the curvature of the decision boundary. Moosavi-Dezfooli et al. (2018) also relate the robustness of a classifier to the local curvature of its decision boundary, and provide an empirical analysis of the curvature of decision boundaries of neural networks. Fawzi et al. (2018a) relate the robustness of a classifier to its empirical risk and show that guaranteeing worst-case robustness is much more difficult than robustness to random noise. Franceschi et al. (2018) provide a geometric characterization of the robustness of linear and "locally approximately flat" classifiers. Their results analyze the relation between the robustness of a classifier to noise and its robustness to adversarial perturbations.

2. Margin via Data Augmentation

Our work aims to quantify the potential of DA to guarantee margin for generic ERMs. We first examine linear classification on linearly separable data, and then extend our results to nonlinear classification. Although we can find max-margin linear classifiers efficiently through quadratic programming (Shalev-Shwartz & Ben-David, 2014), generalizing this to nonlinear classifiers has proved difficult; if this was a simple task for neural networks, the problem of adversarial examples would be non-existent. Hence linear classification serves as a valuable entry point for our study of data agumentation.

We first introduce some notation. Let $A, B \subseteq \mathbb{R}^d$, $x, y \in \mathbb{R}^d$, and $r \geq 0$. Let $d(x, y)$ denote the ℓ_2 distance between x, y, and let $d(A, B) = \inf_{x \in A, y \in B} d(x, y)$. Define $A_r := \{z \in \mathbb{R}^d \mid d(z, A) \leq r\}$. Let $|A|$, $\int(A)$, and $\text{conv}(A)$ denote the cardinality, interior, and convex hull of A. Let $\mathcal{S}^{d-1}$ denote the unit sphere in $\mathbb{R}^d$, and for $r > 0$ let $r\mathcal{S}^{d-1}$ denote the sphere of r.

Let $S \subseteq \mathbb{R}^d \times \{\pm 1\}$ be our training set. For $(x, y) \in S$, x is the feature vector, and $y \in \{\pm 1\}$ is the label. For any such S, we define

$$X_+ = \{x \mid (x, 1) \in S\}, X_- = \{x \mid (x, -1) \in S\}. \quad (2.1)$$

Linear classification. We next recall some background on linear classification. As in Section 1.1, we assume we have access to an algorithm $\mathcal{A}$ that solves the ERM problem over the set of linear classifiers.

A linear classifier is a function of the form $h(x) = \text{sign}(\langle w, x \rangle + b)$, for $w \in \mathbb{R}^d, b \in \mathbb{R}$. We often identify h with the hyperplane $H = \{x \mid \langle w, x \rangle + b = 0\}$. We say that h linearly separates S if $\forall x \in X_+, h(x) \geq 0$ and $\forall x \in X_-, h(x) \leq 0$. If such h exists, S is linearly separable. Let $\mathcal{H}(S)$ denote the set of linear separators of S.

Margin. Suppose S is linearly separable. The *margin* of a linear separator $h \in \mathcal{H}(S)$ is defined as follows:

Definition 1. *The margin of a linear separator $h(x) = \text{sign}(\langle w, x \rangle + b)$ with associated hyperplane H is*

$$\gamma_h(S) = \inf_{(x,y) \in S} d(x, H) = \inf_{(x,y) \in S} \frac{|\langle w, x \rangle + b|}{\|w\|_2}.$$

We define $\gamma_h(S) = -\infty$ if h does not linearly separate S.

If S is linearly separable, there is a linear classifier h^* corresponding to (w^*, b^*) with maximal margin γ^*. This classifier is the most robust linear classifier with respect to bounded ℓ_2 perturbations of samples in S.

In this work, we analyze the margin of ERMs that are trained without any explicit margin constraints or regularization. Let S denote the *true dataset*. To achieve margin, we create an *artificial dataset* S'. We then assume we have access to an algorithm that outputs (if possible) a linear separator h of the *augmented dataset* $S^{\text{aug}} := S \cup S'$. We define $X'_\pm, X^{\text{aug}}_\pm$ analogously to $X_\pm$ in (2.1).

We will analyze the margin of h with respect to the true training data S. If S is linearly separable and we add no artificial points, then some $h \in \mathcal{H}(S)$ must have 0 margin. If S' is designed properly, one might hope that S^{aug} is still linearly separable and that any $h \in \mathcal{H}(S^{\text{aug}})$ has positive margin with respect to S. The following notion formalizes this idea, illustrated in Figure 2.

Definition 2. *The worst-case margin of a linear separator of S^{aug} with respect to the original data S is defined as*

$$\alpha(S, S') = \min_{h \in \mathcal{H}(S^{\text{aug}})} \gamma_h(S).$$

We define this to be $-\infty$ if $\mathcal{H}(S^{\text{aug}}) = \emptyset$.

We are generally interested in the following question:

Question. *How do we design S' so that $\alpha(S, S')$ is as large as possible?*

In Section 3.1, we analyze how large S' must be to ensure that $\alpha(S, S')$ is positive. We show that $|S'| > d$ is necessary to ensure positive worst-case margin. Moreover, if S' is formed via bounded perturbations of S, we need $|S'| \geq |S|$ to guarantee positive margin. In Section 3.2, we analyze the setting where S' is formed by spherical random perturbations of S of radius r, a technique that mirrors random noise perturbations used in practice. We show that if r is not well-calibrated, exponentially many perturbations are required to achieve a margin close to γ^*. However, if r is set correctly, then it suffices to have $|S'|$ polynomial in n and d to ensure that any linear separator of S^{aug} will achieve margin close to γ^* on S. In Section 4, we generalize this notion to a class of nonlinear classifiers, which we refer to as

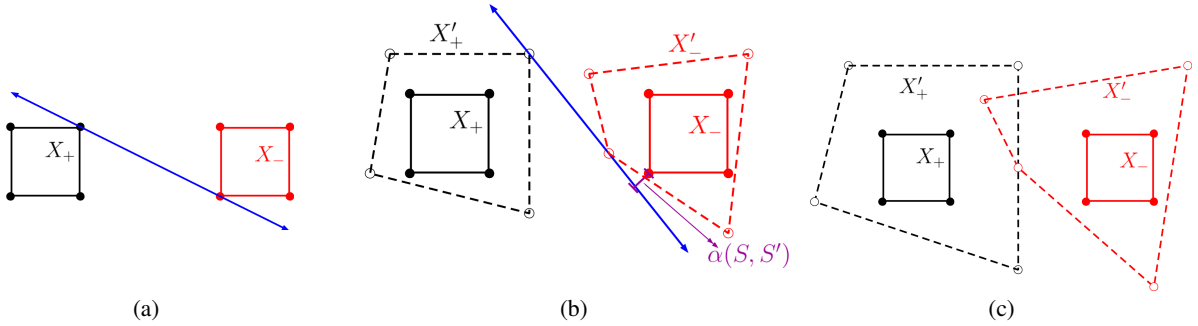

Figure 2. Solid dots represent the true data points and hollow dots represent artificial data points. Convex hulls of the true and augmented data are represented by solid and dashed lines, respectively. Classifiers are shown in blue. (a) Without DA, we may obtain a zero margin classifier. (b) Carefully chosen augmentations can guarantee positive margin. (c) Large augmentations may violate linear separability.

"respectful" classifiers, and derive analogous results to those described above. We show that this class includes classifiers of general interest, such as nearest neighbors classifiers.

3. Linear Classifiers

3.1. How Much Augmentation Is Necessary?

Suppose S is linearly separable with max-margin γ^*. We wish to determine the required size of S' to ensure that $\alpha(S, S') > 0$. We first show that to achieve a positive worst-case margin, the total number of perturbations must exceed the ambient dimension.

Theorem 1. *If* $|S'| < d + 1$, *then* $\alpha(S, S') \leq 0$.

Therefore, we need to augment by at least $d + 1$ points to ensure positive margin. We now wish to understand what margin is possible using data augmentation. We have the following lemma.

Lemma 1. *Let* γ^* *be the maximum margin on* S. *For all* $S' \subseteq \mathbb{R}^d$, $\alpha(S, S') \leq \gamma^*$.

In fact, if we know the max-margin hyperplane, then $d + 1$ points are sufficient to achieve $\alpha(S, S') = \gamma^*$.

Theorem 2. *Let* S *be linearly separable with max-margin* γ^*. *Then* $\exists S'$ *such that* $|S'| = d + 1$ *and* $\alpha(S, S') = \gamma^*$.

The augmentation method in the proof (see Section **??**) requires explicit knowledge of the maximum-margin hyperplane. In practice, most augmentation methods avoid such global computations, and instead apply bounded perturbations to the true data. Recall that for $A \subseteq \mathbb{R}^d$, $A_r = \{x | d(x, A) \leq r\}$. For $S \subseteq \mathbb{R}^d \times \{\pm 1\}$, we define

$$S_r = \Big((X_+)_r \times \{1\}\Big) \bigcup \Big((X_-)_r \times \{-1\}\Big). \quad (3.1)$$

If S' is formed from S by perturbations of size at most r, then $S' \subseteq S_r$. The following result shows that if $S' \subseteq S_r$, then $|S'| \geq |S|$ is necessary to guarantee that $\alpha(S, S') > 0$.

Theorem 3. *Fix* $(n, m) \in \mathbb{N}^2$ *and* $r > 0$. *Then* $\exists S \subseteq \mathbb{R}^d$ *with* $|X_+| = n$ *and* $|X_-| = m$, *such that if* $S' \subseteq S_r$, *and* $|X'_+| < n$, *then* $\alpha(S, S') = 0$.

Figure 3 provides an illustration. Given r, we choose X_+ to lie on a parabola P such that the tangent lines at these points are at distance at least r from other points. We choose X_- to be far enough below the x-axis so that these tangent lines linearly separate X_+ from X_-^{aug}. Suppose we do not augment some point $s \in X_+$. Then the tangent at that point linearly separates X_+ from X_-^{aug}, while being at distance 0 away from s. Thus, we need augmentations at every point in X_+ to guarantee positive margin.

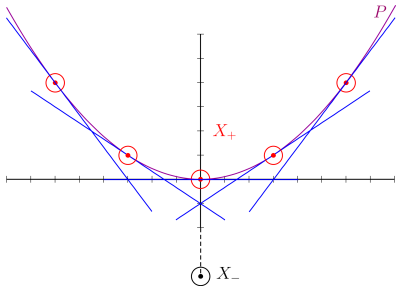

Figure 3. Points in X_+ lie on the the parabola P defined by $y = 9x^2$. The tangent at each point $s \in X_+$ does not intersect the ball of radius r around any other point in X_+. We choose X_- to have points far enough below the x-axis so that the tangents at X_+ separate X'_+ from any $X'_- \subseteq (X_-)_r$. Points in X_+ and their r-balls are in red, their tangents are in blue, and X_- is in black.

3.2. Random Perturbations

We now analyze the setting where S' is formed by random perturbations of S. Our results reveal a fundamental trade-off between the size of perturbations, number of perturbations, margin achieved, and whether or not linear separability is maintained. If we construct many large perturbations, we may violate linear separability, but if we use too few perturbations that are too small in size, we may only

achieve small margin guarantees.

In the rest of this section, we assume that each point in S' is of the form $(x+z, y)$ where $(x, y) \in S$ and z is drawn uniformly at random from $r\mathcal{S}^{d-1}$, the sphere of radius r. Due to the construction of S', the following lemma about the inner products of random points on the sphere $\mathcal{S}^{d-1}$ will be useful throughout.

Lemma 2. *Let a be a unit vector and z be generated uniformly at random from the sphere of radius γ. Then with probability at least $1 - e^{-d\epsilon^2/2\gamma^2}$, $\langle a, z\rangle \leq \epsilon$.*

For further reference, see Chapter 3 of (Vershynin, 2011).

Upper bounds on margin. By Theorem 1, we know that $|S'| \geq d+1$ is necessary to achieve positive margin on S. Since $S' \subseteq S_r$, we must have $\alpha(S, S') \leq r$. In general, we hope that high probability, $\alpha(S, S') \approx r$. We show below that the margin and perturbation size can be close only if $|S'|$ is exponential in d. The result follows using results on the measure of spherical cap densities to bound the distance between S and the max-margin hyperplane.

Theorem 4. *For all $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\alpha(S, S') \leq \left(\sqrt{\frac{2\ln(|S'|) + 2\ln(1/\delta)}{d}}\right) r.$$

This result shows that to achieve minimum-margin close to r, we need the number of perturbations to be exponential in d. Thus, if $r \approx \gamma^*$, we require exponentially many augmentations. However, by making r much larger than γ^*, we may be able to achieve a large margin, provided linear separability is maintained.

Maintaining linear separability. We now show that if r is too large, the augmented sets will often not be linearly separable. Specifically, we show that when S just has two points, if $r = \Omega(\sqrt{d}\gamma^*)$ and $|X'_+| = \Omega(d)$, then linear separability is violated with high probability. For Theorem 5, suppose $S = \{(x_1, 1), (x_2, -1)\}$ where $d(x_1, x_2) = 2\gamma^*$ (i.e., the max-margin is γ^*).

Theorem 5. *If $|X'_+| \geq 16d$ and $r \geq \frac{8e^2\sqrt{2d}}{\pi^{3/2}}\gamma^*$, with probability at least $1 - 2e^{-d/6}$, S^{aug} is not linearly separable.*

To prove this, we first show that with high probability, there are $\Omega(d)$ points in X'_+ labeled -1 by the max-margin classifier. We then use estimates of when random points on the sphere are contained in a hemisphere to show that with high probability, the convex hull of the these points contains x_2. This analysis can be extended directly to the setting where X_+ and X_- are contained in balls of sufficiently small radius compared to $\sqrt{d}\gamma^*$.

On the other hand, we show that if r is slightly smaller than $\sqrt{d}\gamma^*$, linear separability holds with high probability.

Theorem 6. *Suppose S is linearly separable and $|S'| \leq N$. If $r \leq \beta^{-1/2}\sqrt{d/\log(N)}\gamma^*$ for $\beta > 1$, then with probability at least $1 - N^{1-\beta}$, S^{aug} is linearly separable.*

A short proof sketch is as follows: Let w^* be a unit vector orthogonal to the max-margin hyperplane H^*. Suppose $(x+z, y) \in S'$ where $(x, y) \in S$ and z is sampled uniformly on the sphere of radius r. By Lemma 2, with high probability $\langle w^*, x+z\rangle$ will be close to $\langle w^*, x\rangle$, and so $x, x+z$ will fall on the same side of H^*. The result then follows by a union bound.

Theorems 5 and 6 together imply that if $r = \Omega(\sqrt{d}\gamma^*)$, we cannot hope to maintain linear separability. Instead, setting $r = O(\sqrt{d/\log N}\gamma^*)$, we will maintain linear separability with high probability. We will use the latter result in the next section to show that for such r, we can actually provide lower bounds on the adversarial margin $\alpha(S, S')$ achieved.

Lower bounds on margin. By Theorem 4, we know that if $r \approx \gamma^*$, we need N to be exponential in d to achieve a margin close to γ^*. By Theorem 6, we can set r to be as large as $O(\sqrt{d/\log N}\gamma^*)$ and maintain linear separability. We might hope that in this latter setting, we can achieve a margin close to γ^* with substantially fewer points than when $r \approx \gamma^*$.

Suppose S' is formed by taking N perturbations of each point in $S = \{(x_i, y_i)\}_{i\in[n]}$. Formally, for $i \in [n], j \in [N]$ let $z_i^{(j)}$ be drawn uniformly at random from $r\mathcal{S}^{d-1}$. Then,

$$S' = \{(x_i + z_i^{(j)}, y_i)\}_{i\in[n], j\in[N]}. \tag{3.2}$$

We show following theorem:

Theorem 7. *Suppose S is linearly separable with max-margin γ^*. Let S' be as in (3.2). There is a universal constant C such that if $N \geq Cd$ and $r \leq \beta^{-1/2}\sqrt{d/\log N}\gamma^*$ for $\beta > 1$, then with probability at least $1 - ne^{-d} - nN^{1-\beta}$, we have*

$$\alpha(S, S') \geq \frac{1}{2\sqrt{2}}\sqrt{\frac{\log(N/d)}{d}}r.$$

Taking $r = \beta^{-1/2}\sqrt{d/\log N}\gamma^*$ and β sufficiently large, we can ensure that the worst-case margin among linear separators is a constant fraction of the max-margin. Thus, with high probability, we can achieve a constant approximation of the best possible margin with $|S'| = O(nd^2)$. While Theorems 1 and 3 indicate that $|S'|$ should grow linearly in n and d, determining whether $O(nd^2)$ is tight for some S is an open problem.

Remark 1. *Theorem 7 can be extended to the setting where we only take perturbations of each point in a τ-cover of X_+*

and X_-. Recall that A is a τ-cover of B if $\forall x \in B, \exists x' \in A$ where $d(x, x') \leq \epsilon$. The same result (with the constant $2\sqrt{2}$ replaced by $4\sqrt{2}$) holds when S' is formed according to (3.2), but with S replaced by $A_+ \times \{1\} \cup A_- \times \{-1\}$ where A_+, A_- are τ-covers of X_+, X_- for

$$\tau = \frac{1}{4\sqrt{2}} \sqrt{\frac{\log(N/d)}{d}} r. \tag{3.3}$$

Thus, we only need $|S'| = O(md^2)$ perturbations, where $m = \max\{|X_+|, |X_-|\}$. When S is highly clustered, this could result in a much smaller sample complexity, as m may be much smaller than n.

To give a sketch of the proof, suppose $(0, 1) \in S$. Thus, S' contains N points of the form $(z_i, 1)$ where $z_i \sim r\mathcal{S}^{d-1}$. We wish to guarantee that any linear separator, with associated hyperplane H, has some margin at 0. Consider $K = \text{conv}(\{z_i\}_{i \in [N]})$. Since each z_i has label 1, we know that H cannot intersect the interior of K. Then, if 0 is in the interior of K, then H has positive margin at 0. In fact, we extract a strengthening of this from the proof of Lemma 3.1 of (Alonso-Gutierrez, 2008):

Lemma 3. *Let $z_1, \ldots, z_N$ be drawn uniformly at random on $r\mathcal{S}^{d-1}$. Let $K = \text{conv}(z_1, \ldots, z_N)$. Then there exists a constant $C > 0$ such that if $N \geq Cd$, then*

$$\mathbb{P}\left(\frac{1}{2\sqrt{2}} \sqrt{\frac{\log(N/d)}{d}} \mathcal{B}_r(0) \not\subseteq K \right) \leq e^{-d}.$$

Thus, with high probability $\mathcal{B}_\rho(0) \subseteq K$ where $\rho = \Omega(\sqrt{\log(N/d)/d} r)$. The margin of H at 0 is therefore at least ρ. Applying Theorem 6, we derive Theorem 7. A pictorial explanation of the proof is given in Figure 4.

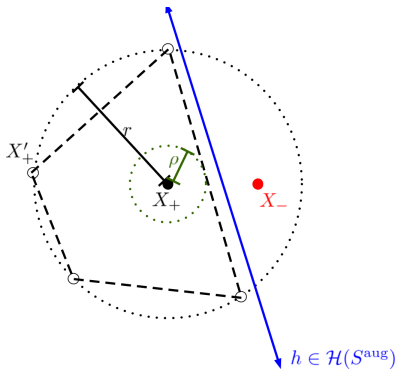

Figure 4. A pictorial explanation of the proof of Theorem 7. Suppose X'_+ is drawn uniformly at radius r from X_+. With r as in the theorem statement, with high probability X'_+ will not prevent linear separability of S^{aug}. Moreover, with high probability $\text{conv}(X'_+)$ will contain a ball of radius ρ around each point in X_+. This then implies that any $h \in \mathcal{H}(S^{\text{aug}})$ has margin at least ρ.

4. Nonlinear Classifiers

We now consider more general binary-valued classifiers. Given $S \subseteq \mathbb{R}^d \times \{\pm 1\}$, a classifier $f : \mathbb{R}^d \to \{\pm 1\}$ separates S if $f(x) = y$ for all $(x, y) \in S$. Let $\mathcal{R}(S)$ denote the collection of separators of S. If $\mathcal{R}(S)$ is non-empty, we say that S is separable. Given $f : \mathbb{R}^d \to \{\pm 1\}$, we define a generalization of the notion of margin in 1.

Definition 3. *If $f \in \mathcal{R}(S)$, its margin on S is given by*

$$\gamma_f(S) := \min_{(x,y) \in S} d(x, f^{-1}(-y)).$$

We define $\gamma_f(S) = -\infty$ if $f \notin \mathcal{R}(S)$.

Suppose we have a function class $\mathcal{F}$ and we wish to find an ERM of the $0-1$ loss on S (more generally, any nonnegative loss function where $\ell(f(x), y) = 0$ iff $f(x) = y$). The set of ERMs is simply $\mathcal{R}(S) \cap \mathcal{F}$.

To find ERMs with positive margin, we will again form a perturbed dataset S', and then find some ERM of $S^{\text{aug}} = S \cup S'$. We define the margin of f with respect to S and S' as follows.

Definition 4. *The margin $\gamma_f(S, S')$ of f with respect to S, S' is defined by $\gamma_f(S, S') = \gamma_f(S)$ if $f \in \mathcal{R}(S^{\text{aug}})$ and $-\infty$ otherwise.*

If S^{aug} is separable and $\mathcal{F}$ is sufficiently expressive, one can always find an ERM with zero margin. Instead, we will restrict to a collection of functions that is expressive, but still have meaningful margin guarantees. We refer to these as *respectful* functions.

Respectful classifiers. If $x_1, x_2 \in \mathbb{R}^d$ are sufficiently close and have the same label, it is reasonable to expect a well-behaved classifier to assign the same label to every point between x_1 and x_2. In fact, (Fawzi et al., 2018b) shows that empirically, state-of-the-art deep nets often remain constant on straight lines connecting different points of the same class. For a linear classifier f labels all points in A as 1, we know that f assigns 1 to the entire set $\text{conv}(A)$. With this in mind, we give the following definition:

Definition 5. *A function $f : \mathbb{R}^d \to \{\pm 1\}$ is respectful of S if $\forall x \in \text{conv}(X_+), f(x) = 1$ and $\forall x \in \text{conv}(X_-), f(x) = -1$.*

Intuitively, f must respect the operation of taking convex hulls of points with the same label. However, assigning all of $\text{conv}(X_+)$ and $\text{conv}(X_-)$ the same label is a relatively strict condition. To relax this condition, we define a class of functions that are respectful only on small clusters of points. Recall the notion of a circumradius:

Definition 6. *The circumradius $R(A)$ of a set $A \subseteq \mathbb{R}^d$ is the radius of the smallest ball containing A.*

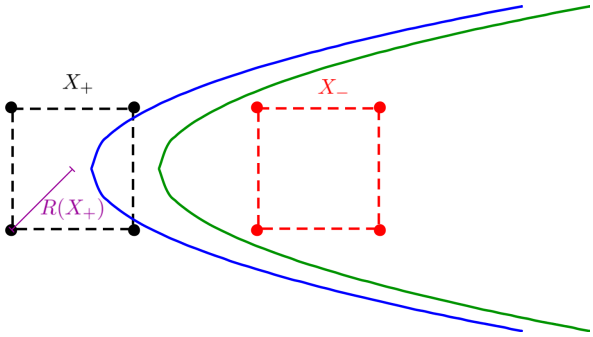

Figure 5. Suppose that $R(X_+) \leq \epsilon$. The classifier with a blue decision boundary is not ϵ-respectful of S, but the classifier with a green decision boundary is ϵ-respectful of S.

We now define ϵ-respectful classifiers:

Definition 7. *For $\epsilon \in [0, \infty]$, we say that a classifier $f : \mathbb{R}^d \to \{\pm 1\}$ is ϵ-respectful of S if $\forall A \subseteq X_+$ such that $R(A) \leq \epsilon$, and $\forall x \in \text{conv}(A)$, $f(x) = 1$; and $\forall B \subseteq X_-$ such that $R(B) \leq \epsilon$, and $\forall x \in \text{conv}(B)$, $f(x) = -1$. Let $\mathcal{R}_\epsilon(S)$ denote the set of ϵ-respectful classifiers.*

An illustration is provided in Figure 5. Note that the set of separators of S is simply $\mathcal{R}_0(S)$, and the set of respectful classifiers is $\mathcal{R}_\infty(S)$. Smaller values of ϵ lead to more expressive function classes $\mathcal{R}_\epsilon(S)$. We now show that this definition includes some function classes of interest:

Example 1 (Linear Classifiers)**.** *Recall that $\mathcal{H}(S)$ is the set of linear separators of S. It is straightforward to see that such functions are respectful of S, so $\mathcal{H}(S) \subseteq \mathcal{R}_\infty(S)$. By the hyperplane separation theorem (see Lemma **??**), we have $\mathcal{H}(S) \neq \emptyset$ if and only if $\mathcal{R}_\infty(S) \neq \emptyset$. In general, $\mathcal{H}(S)$ is a proper subset of $\mathcal{R}_\infty(S)$.*

Example 2 (Nearest Neighbor)**.** *Let f_{NN} denote the 1-nearest neighbor classifier on S: For $x \in \mathbb{R}^d$, we have $f_{NN}(x) = 1$ if $d(x, X_+) \leq d(x, X_-)$, and $f_{NN}(x) = -1$ otherwise. For $\epsilon \in [0, \frac{d(X_+, X_-)}{2})$, we can argue that $f_{NN} \in \mathcal{R}_\epsilon(S)$, as follows: Suppose $x \in \text{conv}(A)$ where $A \subseteq X_+$ and $R(A) \leq \epsilon$. Then $d(x, X_+) \leq \epsilon$. For all $u \in X_-$, we have $d(u, X_+) \geq d(X_+, X_-)$, so $d(x, u) > \frac{d(X_+, X_-)}{2}$. Hence, $f_{NN}(x) = 1$.*

We now consider the following adversarial problem. Given S, we form a perturbed version S'. An adversary can pick an ϵ-respectful classifier $f \in \mathcal{R}(S^{\text{aug}})$. The smaller the value of ϵ, the more powerful the adversary. We hope that no matter which f the adversary chooses, the value of $\gamma_f(S, S')$ is not too small.

We first provide bounds on how large S' must be to ensure a positive margin, and then derive results for random perturbations when S is (non)-linearly separable. Our results are versions of Theorem 7 for respectful classifiers. Finally, we will show that for respectful classifiers, our bounds for random perturbations are tight up to constants for some S.

4.1. How Much Augmentation Is Necessary?

We first show that for any $\epsilon \in [0, \infty]$, we must have $|S'| > 2d$ in order to achieve a positive margin.

Theorem 8. *Suppose S is separable. If $|X'_+| \leq d$ or $|X'_-| \leq d$, then for any $\epsilon \in [0, \infty]$, either $\mathcal{R}_\epsilon(S^{\text{aug}}) = \emptyset$, or $\exists f \in \mathcal{R}_\epsilon(S^{\text{aug}})$ such that $\gamma_f(S, S') = 0$.*

Suppose we limit ourselves to bounded perturbations of S, so that $S' \subseteq S_r$ for some $r > 0$. We will show that in this setting, we may need as many as $|S|(d+1)$ perturbations to guarantee a positive margin.

Theorem 9. *For all $n \geq 1$ and $\epsilon, r \in (0, \infty)$, there is some S of size n such that if $|S'| \leq |S|(d+1)$, then $\exists f \in \mathcal{R}_\epsilon(S^{\text{aug}})$ such that $\gamma_f(S, S') = 0$.*

Next, we consider the problem of ensuring a positive margin with bounded perturbations. The following lemma shows that if $\epsilon < r$, there is some S such that the adversary can find a zero margin classifier for any $S' \subseteq S_r$.

Lemma 4. *For any $\epsilon \in (0, \infty)$ and $r > \epsilon$, there is S such that for any $S' \subseteq S_r$, $\exists f \in \mathcal{R}_\epsilon(S^{\text{aug}})$ such that $\gamma_f(S, S') = 0$.*

Therefore, for $S' \subseteq S_r$, to ensure that any $f \in \mathcal{R}_\epsilon(S^{\text{aug}})$ has positive margin, we need $r \leq \epsilon$, $|S'| \geq 2d + 2$, and $|S'| \geq |S|(d+1)$. In fact, these three conditions are sufficient to ensure positive margin.

Theorem 10. *For any S, if $\epsilon \in (0, \infty]$ and $r \leq \epsilon$, then $\exists S' \subseteq S_r$ with $|S'| = |S|(d+1)$, such that $\forall f \in \mathcal{R}_\epsilon(S^{\text{aug}})$, $\gamma_f(S, S') > 0$.*

While this theorem does not guarantee that $\mathcal{R}_\epsilon(S^{\text{aug}}) \neq \emptyset$, we will show in Lemma **??** that if $S' \subseteq S_r$ for $r \leq \epsilon < \frac{d(X_+, X_-)}{4}$, then $\mathcal{R}_\epsilon(S^{\text{aug}})$ is guaranteed to be nonempty.

4.2. Random Perturbations

We now analyze how random perturbations affect the margin of ϵ-respectful classifiers. Just as in the linear setting, we focus on the case where the points in S' are of the form $(x + z, y)$ where z is drawn uniformly at random from the sphere of radius r. We provide lower bounds on the margin that are analogous to the linear setting, and show that our margin bounds are tight up to constants in some settings.

Linearly separable data. We first show that when S is linearly separable and we perform random augmentations, the results in Section 3.2 still hold, even though the adversary is allowed to select classifiers in the larger set $\mathcal{R}_\infty$.

Theorem 11. *Let S' be generated as in (3.2). There is a universal constant C such that if $N \geq Cd$ and $r \leq \beta^{-1/2}\sqrt{d/\log N}\gamma^*$ for $\beta > 1$, then with probability at least $1 - ne^{-d} - nN^{1-\beta}$, we have $\mathcal{R}_\infty(S) \neq \emptyset$.*

Furthermore, $\forall f \in \mathcal{R}_\infty(S^{\text{aug}})$*, we have*

$$\gamma_f(S, S') \geq \frac{1}{2\sqrt{2}}\sqrt{\frac{\log(N/d)}{d}}r.$$

The proof uses a generalization of Theorem 7 to respectful functions. We show in Theorem 13 that this bound is tight up to constants under certain assumptions on S.

As in the linear case, a perturbation radius of $r = O(\sqrt{d}\gamma^*)$ is necessary to maintain separability. Suppose $S = \{(x_1, 1), (x_2, -1)\}$ with $d(x_1, x_2) = 2\gamma^*$ and S' is as in (3.2). Applying the hyperplane separation theorem and Theorem 5, we have the following result:

Theorem 12. *If* $N \geq 16d$ *and* $r \geq \frac{8e^2\sqrt{2d}}{\pi^{3/2}}\gamma^*$*, then*

$$\mathbb{P}(\mathcal{R}_\infty(S^{\text{aug}}) = \emptyset) \geq 1 - 2e^{-d/6}.$$

In short, spherical random data augmentation behaves similarly when the adversary selects linear classifiers or classifiers in $\mathcal{R}_\infty(S^{\text{aug}})$, both in terms of margin achieved and upper bounds on perturbation size to maintain separability.

Nonlinearly separable data. When S consists of more than two points, the margin obtained by some $f \in \mathcal{R}_\epsilon(S)$ may be much larger than the max-margin linear classifier. Moreover, $\mathcal{R}_\epsilon(S)$ may be non-empty even though S is not linearly separable. Thus, we would like to derive versions of the results in Section 3.2 for settings where S may not be linearly separable, but $\mathcal{R}_\epsilon(S) \neq \emptyset$. In fact, if $\mathcal{R}_\epsilon(S) \neq \emptyset$ and we generate S' as in (3.2), we can derive the following theorem, comparable to Theorem 7 above:

Theorem 13. *If* $r \leq \epsilon$*, then there is a universal constant* C *such that if* $N \geq Cd$*, then with probability at least* $1 - ne^{-d}$*,* $\forall f \in \mathcal{R}_\epsilon(S^{\text{aug}})$*,*

$$\gamma_f(S, S') \geq \frac{1}{2\sqrt{2}}\sqrt{\frac{\log(N/d)}{d}}r.$$

Furthermore, if $\epsilon < \frac{d(X_+, X_-)}{4}$ *then* $\mathcal{R}_\epsilon(S^{\text{aug}}) \neq \emptyset$*.*

The first part of the proof proceeds similarly to that of Theorem 7, using the definition of ϵ-respectful classifiers. For the second, we use nearest neighbor classifiers (as in Example 2) to construct ϵ-respectful classifiers of S^{aug}.

Although $r \leq \epsilon < \frac{d(X_+, X_-)}{4}$ is sufficient to guarantee that $\mathcal{R}_\epsilon(S^{\text{aug}}) \neq \emptyset$, this may be overly conservative. Whereas Theorems 11 and 12 provide a characterization of the range on r for which $\mathcal{R}_\infty(S^{\text{aug}})$ is non-empty with high probability, a tighter characterization for $\epsilon < \infty$ remains open.

Upper bounds on margin. Finally, we show that for certain S, the margin bounds in Theorems 11 and 13 are tight up to constants. While it is as yet unknown whether Theorem 7 is asymptotically tight, the increased expressive capability of respectful classifiers allows us to exhibit upper bounds on the worst-case margin matching the lower bounds above. Suppose $S = \{(x_1, 1), (x_2, -1)\}$, and S' is generated as in (3.2). We have the following result:

Theorem 14. *Fix* $\epsilon \in [0, \infty]$ *and* $r > 0$*. There are absolute constants* C_1, C_2 *such that if* $N > d$ *and* $\mathcal{R}_\epsilon(S^{\text{aug}}) \neq \emptyset$*, then with probability at least* $1 - 2e^{-C_2 d \log(N/d)}$*,* $\exists f \in \mathcal{R}_\epsilon(S^{\text{aug}})$ *such that*

$$\gamma_f(S, S') \leq \sqrt{C_1\frac{\log(2N/d)}{d}}r. \tag{4.1}$$

The proof relies on estimates of the inradius of random convex polytopes from (Alonso-Gutierrez, 2008). The theorem can also be extended to settings where X_+ and X_- are not singletons. Suppose we can decompose X_+ and X_- into clusters $\{A_i\}_{i=1}^k$ and $\{B_j\}_{j=1}^l$ such that each cluster has size at most m, circumradius at most $O(\sqrt{\log(N/d)/d}r)$, and the distance between any two clusters is $\Omega(\epsilon)$. If S' is generated as in (3.2), then with high probability there is some $f \in \mathcal{R}_\epsilon(S^{\text{aug}})$ satisfying (4.1) where N is replaced by mN.

5. Conclusion and Open Problems

Data augmentation is commonly used in practice, since it significantly improves test error and model robustness. In this work, we have analyzed the performance of data augmentation through the lens of margin. We have demonstrated how data augmentation can guarantee positive margin for unconstrained empirical risk minimizers. For both linear and nonlinear "respectful" classifiers, we provided lower bounds on the number of points needed to ensure positive margin, and analyzed the margin attained by additive spherical data augmentation.

There are several interesting open problems that we plan to tackle in the future. First, it would be interesting to theoretically analyze practical state-of-the-art augmentation methods, such as random crops, flips, and rotations. Such perturbations often fall outside our framework, as they are not bounded in the ℓ_2 norm. Another fruitful direction would be to examine the performance of adaptive data augmentation techniques. For example, robust adversarial training, (such as in (Madry et al., 2017)), can be viewed as a form of adaptive data augmentation. By taking a data augmentation viewpoint, we hope to derive theoretical benefits of using adversarial training methods. One final direction would be to develop improved augmentation methods. In particular, we would like methods that can exploit domain knowledge and the geometry of the underlying problem in order to find models with better robustness and generalization properties.

References

Alonso-Gutierrez, D. On the isotropy constant of random convex sets. *Proceedings of the American Mathematical Society*, 136(9):3293–3300, 2008.

Andrews, G. E., Askey, R., and Roy, R. *Special functions*, volume 71. Cambridge university press, 2000.

Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A., and Criminisi, A. Measuring neural net robustness with constraints. In *Advances in Neural Information Processing Systems*, pp. 2613–2621, 2016.

Bishop, C. M. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge university press, 2004.

Caramanis, C., Mannor, S., and Xu, H. Robust optimization in machine learning. *Optimization for Machine Learning*, pp. 369, 2012.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Dao, T., Gu, A., Ratner, A. J., Smith, V., De Sa, C., and Ré, C. A kernel theory of modern data augmentation. *arXiv preprint arXiv:1803.06084*, 2018.

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. Robustness of classifiers: From adversarial to random noise. In *Advances in Neural Information Processing Systems*, pp. 1632–1640, 2016.

Fawzi, A., Fawzi, O., and Frossard, P. Analysis of classifiers: Robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018a.

Fawzi, A., Moosavi-Dezfooli, S.-M., Frossard, P., and Soatto, S. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3762–3770, 2018b.

Ford, N., Gilmer, J., Carlini, N., and Cubuk, D. Adversarial examples are a natural consequence of test error in noise. *arXiv preprint arXiv:1901.10513*, 2019.

Franceschi, J., Fawzi, A., and Fawzi, O. Robustness of classifiers to uniform ℓ_p and Gaussian noise. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018*, pp. 1280–1288, 2018.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. URL http://arxiv.org/abs/1412.6572.

Huber, G. Gamma function derivation of n-sphere volumes. *The American Mathematical Monthly*, 89(5):301–302, 1982.

Klartag, B. and Kozma, G. On the hyperplane conjecture for random convex sets. *Israel Journal of Mathematics*, 170(1):253–268, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Kuznetsova, A., Ju Hwang, S., Rosenhahn, B., and Sigal, L. Expanding object detector's horizon: Incremental learning framework for object detection in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 28–36, 2015.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Misra, I., Shrivastava, A., and Hebert, M. Watch and learn: Semi-supervised learning for object detectors from video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P., and Soatto, S. Robustness of classifiers to universal perturbations: A geometric perspective. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByrZyglCb.

Prest, A., Leistner, C., Civera, J., Schmid, C., and Ferrari, V. Learning object class detectors from weakly annotated video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3282–3289. IEEE, 2012.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Vershynin, R. Lectures in geometric functional analysis. *Preprint, University of Michigan*, 2011.

Wager, S., Wang, S., and Liang, P. S. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, pp. 351–359, 2013.

Wendel, J. G. A problem in geometric probability. *Mathematica Scandinavica*, 11(1):109–111, 1963. ISSN 00255521, 19031807. URL http://www.jstor.org/stable/24490189.

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

Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10(Jul):1485–1510, 2009.

Zantedeschi, V., Nicolae, M.-I., and Rawat, A. Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 39–49. ACM, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.