On the Geometric and Statistical Interpretation of Data Augmentation

by

Zhili Feng

A master's thesis submitted in partial fulfillment of
the requirements for the degree of

Master of Science

(Computer Sciences)

at the

UNIVERSITY OF WISCONSIN–MADISON

2019

© Copyright by Zhili Feng 2019
All Rights Reserved

ACKNOWLEDGMENTS

I would like to express my great gratitude to Professor Po-Ling Loh, my thesis advisor, for her kindness, patient guidance, and useful advice for my research. I want to thank Professor Dimitris Papailiopoulos for his guidance and cooperation on this data augmentation project. I would also like to thank Professor Christos Tzamos for his help in my thesis.

Figure 1.1 is taken from our ICML submission *Does data augmentation lead to positive margin?*, co-authored with Shashank Rajput, Zachary Charles, Po-Ling Loh, and Dimitris Papailiopoulos.

CONTENTS

ABSTRACT

Data augmentation (DA) is a common technique in training machine learning models. For example in image classifications, people augment image datasets by random cropping, rotating, and adding random noises. Another trending technique is the adversarial training, where the datasets are augmented by adversarial examples. Despite its empirical effectiveness, the theory behind DA is rarely known.

In this thesis, we analyze why DA generalizes and robustifies our models, from both geometric and statistical points of view. Geometrically, we provide both upper and lower bounds on the margins created by DA, via convex geometric arguments. The upper bound on the margin is distribution-independent, while the lower bound on the margin fits a wide range of probability distributions.

Statistically, we prove that DA helps generalization by controlling the stability of our learning algorithm, in a very small cost, given the training data is sufficiently large. In addition, with the same sample complexity, noise robustness is guaranteed.

1 INTRODUCTION

Machine learning models have shown their unprecedented successes in computer vision, natural language understanding, recommendation systems, and numerous other areas. With more and more tunable parameters, the traditional PAC-learning models and generalization error bounds are not particularly useful, since the enormous amount of parameters suggests that generalization requires unrealistic size of training dataset. Modern machine learning techniques like SGD and data augmentation, on the other hand, seem to perform some sort of "implicit regularization" to reduce the model complexity, thus obtain a better generalization ability with less sample complexity. 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 performance of these models.

Data augmentation (DA) creates artificial data points from the original samples. For example, in image-related tasks, translations, flips, and alternation of the RGB channels of the training data are often used, as discussed in Krizhevsky et al. (2012). Misra et al. (2015); Kuznetsova et al. (2015); Prest et al. (2012) augment sparsely annotated data to get a better supervised learning performance. Another trending technique is the adversarial training, where people create artificial samples adversarially (Szegedy et al. (2013); Bastani et al. (2016); Carlini and Wagner (2017); Szegedy et al. (2013); Goodfellow et al. (2014)). They all serve the purpose to make the model more robust: if the test data is slightly perturbed, then the prediction should not differ too much.

Although DA is quite successful in practice, the theory behind DA is not well-known. One of the first theoretical research is done by Bishop (1995), who show that in expectation, training with noise is equivalent to Tikhonov regularization. This is an asymptotic case since such regularization can only be achieved if infinite number of augmentations are added. Wager et al. (2013) prove similar regularization results on generalized linear models. Wong and Kolter (2018) provide a provable adversarial-free region from a mixed integer linear programming standpoint.

In the first part of this thesis, we analyze how DA impacts the margin of a classifier, i.e., the minimum distance from the training data to its decision boundary. We focus on the margin because it acts both as a proxy for generalization Shalev-Shwartz and Ben-David (2014) and worst-case/adversarial robustness. In the second part of the thesis, we focus on the stability and noise robustness of DA from a statistical standpoint.

Contributions

We consider the following ERM algorithm:

$$\mathcal{A}(S_{\text{train}}) = \arg\min_{f \in \mathcal{F}} \left\{ \sum_{(x,y) \in S_{\text{train}}} \ell(f(x); y) \right\} \tag{1.1}$$

where S_{train} is the input training set, $\mathcal{F}$ is the function class we have in hand, and $\ell(f(x); y) = \mathbb{1}_{\{f(x) \neq y\}}$ is the $0-1$ loss.

The first part of the thesis aims to understand how we can construct the augmented dataset S_{aug} such that when we use the vanilla ERM, our model is close to the following adversarially trained ERM:

$$\arg\min_{f \in \mathcal{F}} \left\{ \sum_{(x,y) \in S_{\text{train}}} \max_{\|r\|_2 \leqslant \epsilon} \ell(f(x+r); y) \right\}. \tag{1.2}$$

Equivalently, how can we construct an augmented training set such that our ERM model is adversarially robust?

Figure 1.1 provides a sketch of the problem for linear classification and the separable data setup.

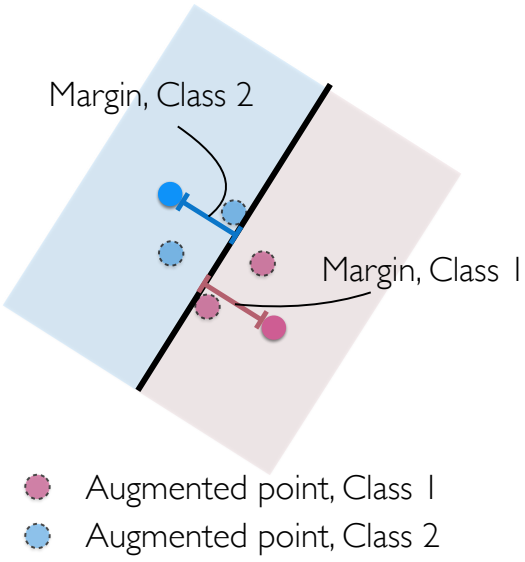

Figure 1.1: A linearly separable dataset with one data point in each class. For each point, we perturb twice. Any linear classifier that separates all 6 points has a non-zero margin. On the contrary, if no perturbation is done, then there exists a linear separator with 0 margin.

Our second goal is to understand how DA helps generalization. Mathematically, we would like to bound the following term:

$$\frac{1}{|S_{aug}|} \sum_{(x,y)\in S_{aug}} \ell(f(x), y) - \mathbb{E}_{(x,y)}[\ell(f(x), y)]$$

Last but not least, we want to analyze how DA improves noise robustness by bounding:

$$\frac{1}{|S_{aug}|} \sum_{(x,y)\in S_{aug}} \ell(f(x), y) - \mathbb{E}_{(x,y,\xi)}[\ell(f(x+\xi), y)]$$

where ξ is the test time noise. We assume the noise distribution is known during training.

Geometric lower bounds on margin. We consider linear classification for separable data. In practice, DA is often generated randomly. For image datasets, researchers often create artificial images by random crops, rotations, and flips. For generic datasets, adding random noise is also a common technique. We provide a lower bound on the linear classification margin, under mild conditions on the noise. In particular, any sub-Gaussian noise satisfies these conditions.

Distribution-independent geometric upper bounds on margin. We consider "respectful" classification model and a two-point classification scheme. We prove that as long as the perturbations are bounded in an ℓ_2 ball, with similar radius to the max margin, the worst-case margin will always be small unless we make exponentially many perturbations.

Generalization bound via stability We consider learning with convex and twice-differentiable loss functions without regularization. In this setting, it is well-known since Bishop (1995) that training with noice asymptotically (meaning adding infinite number of perturbations) is equivalent to regularization. We further show that in the non-asymptotic case, the ERM with DA is a stable algorithm, thus generalizes.

Noise robustness We consider the ERM of $0-1$ loss. Instead of comparing our empirical loss to the true risk, we consider bounding the difference between the empirical loss and the population robust risk. This is similar to learning with interdependent data. We derive a statistical upper bound on the convergence rate via the Rademacher complexity argument.

Related Work

Numerous works have tried to explain the adversarial phenomenon from the geometric perspective. Fawzi et al. (2016) study the semi-random noise robustness and quantify the difference between noise robustness and adversarial robustness. Moosavi-Dezfooli et al. (2018) study the relationship between robustness the curvature of decision boundaries. Franceschi et al. (2018) assume the decision boundaries are "locally approximately flat" and study the robustness of such classifiers. Our work, on the other hand, study the geometry of the training data, instead of the classifiers, and provide a simple augmenting algorithm.

Another line of recent works analyze the hardness of adversarial learning via the lens of statistical learning theory. Cullina et al. (2018) introduce the notion of adversarial VC-Dimension and prove that the adversarial VC-Dimension could be d times larger than the VC-Dimension in $\mathbb{R}^d$, thus getting a much slower convergence rate. Yin et al. (2018) show that the Rademacher complexity of adversarial robust generalization can be much larger than the regular Rademacher complexity, proving that adversarial learning is indeed hard.

2.1 Margin via Data Augmentation

The main setup of this thesis is the linear classifications and linearly separable datasets. Although the max-margin linear classifier, SVM, can be efficiently learned, it is not clear what margin even means in non-linear case, not to mention enforcing such constraints. However, our work is still interesting theoretically, and can be extended to the non-linear cases potentially.

Linear classification. We first recall what linear classification is. Given training set $Z \subseteq \mathbb{R}^d \times \{\pm 1\}$. For $(x, y) \in Z$, x is the feature vector, and $y \in \{\pm 1\}$ is the label.

Let $S = \{x \mid (x, 1) \in Z\}$ be the set of positive training data and $T = \{x \mid (x, -1) \in Z\}$ be the set of negative training data. A linear classifier is a function h of the following form:

$$h(x) = \operatorname{sgn}(\langle w, x \rangle + b)$$

We denote $\mathcal{H}$ as the set of all such functions.

Let $\ell(h(x), y) = \mathbb{1}_{\{h(x) \neq y\}}$ be the $0-1$ loss. If our training data $S \cup T$ is linearly separable, then $\exists h^* \in \mathcal{H}$ such that $\ell(h^*(x), y) = 0$ for all $(x, y) \in S \cup T$. Let $\mathcal{H}(S, T)$ be the set of all such ERMs.

Margin. We say that S and T are *linearly separable* if $\mathcal{H}(S, T) \neq \emptyset$. The *margin* of a linear separator $h \in \mathcal{H}(S, T)$ is defined as follows:

Definition 2.1.1. The margin of a linear classifier $h(x) = \operatorname{sgn}(\langle w, x \rangle + b)$ wrt h is

$$\gamma_h(S, T) = \inf_{x \in S \cup T} d(x, H) = \inf_{x \in S \cup T} \frac{\left|\langle w, x \rangle + b\right|}{\|w\|_2},$$

if $h \in \mathcal{H}(S, T)$, and $-\infty$ otherwise.

Note that the SVM with ℓ_2 regularization is the most robust to ℓ_2 perturbations given our definition of margin.

Margin via data augmentation. Let S' be the augmented dataset of S, T' be the augmented dataset of T. In this thesis, we analyze the margin of the worst-case ERM in our function class $\mathcal{H}(S \cup S', T \cup T')$. This idea is similar to the uniform deviation bounds in the

statistical learning theory, as we do not know which classifier will be obtained, so we have to control the convergence of all classifiers in the function class.

One can imagine that if S' and T' are carefully crafted, any $h \in \mathcal{H}(S \cup S', T \cup T')$ would have a positive margin.

Definition 2.1.2. The worst-case margin of a linear separator of $S \cup S'$ and $T \cup T'$ with respect to the original data S, T is defined as

$$\alpha(S, S', T, T') = \min_{h \in \mathcal{H}(S \cup S', T \cup T')} \gamma_h(S, T).$$

We define this to be $-\infty$ if $\mathcal{H}(S \cup S', T \cup T') = \emptyset$.

2.2 Lower Bound on Random Perturbation Margin

In this section we provide lower bounds on the random perturbation margin. Note that in the linear classification case, if $rB_2^d \subseteq S \cup S'$ and $rB_2^d \subseteq T \cup T'$, then we will have:

$$\alpha(S, S', T, T') \geqslant r$$

for any linear classifier.

Hence it is natural to derive a high probability bound on the above-mentioned event. To this end, we exploit the inclusion theorems from local theory of Banach spaces. These theorems are also relatable to geometric functional analysis and convex geometry.

Connecting Inclusion Theorems to Margin

Inclusion theorems have been important topics in local theory of Banach spaces for a long time, the main question is: given N points drawn from certain distributions in a Banach space, denote their absolute convex hull as K_N, then what is the probability that K_N contains certain convex body? In literature, mathematicians often consider the inclusion of L_q centroid body, which is beyond the scope of this thesis. For now we only consider the inclusion of a ℓ_2 ball, which coincides with our decision rules in the Euclidean space.

Let $\Gamma = [\xi_{ij}]_{1\leqslant i\leqslant N, 1\leqslant j\leqslant n}$ be a $N \times d$ random matrix. We can view d as the dimension of data point, and N as the number of perturbations. WLOG we can assume our original data point is at the origin, since every variable we will deal with is invariant to translations. Throughout the section, we consider the two-point classification scheme, that is, $S = \{x\}$ and $T = \{y\}$.

Fact 1. Let $\Gamma \in \mathbb{R}^{N\times d}$, and let K_N be the absolute convex hull of the row vectors of Γ, then $K_N = \Gamma^* B_1^N$, where B_1^N is the unit ℓ_1 ball in $\mathbb{R}^d$.

Proof. Let $\Gamma = \begin{bmatrix} \Gamma_1 \\ \vdots \\ \Gamma_N \end{bmatrix}$, then

$$\Gamma^* B_1^N = \begin{bmatrix} \Gamma_1^* & \cdots & \Gamma_N^* \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \end{bmatrix} = \sum_{i=1}^{N} \lambda_i \Gamma_i^*, \text{ where } \sum_{i=1}^{N} |\lambda_i| = 1$$

This coincides with the definition of absolute convex hull. □

Notation 1. 1. We write $\|\Gamma\|$ as the operator norm from ℓ_2^n to ℓ_2^N. Notice that the operator norm coincides with the largest singular value of Γ.

2. If X is a normed vector space, we write $X' = L(X, \mathbb{R})$ as its dual.

3. We write $|x|$ as a short term for $\|x\|_2$.

We can connect inclusion to the operator norm of Γ by the following lemma:

Lemma 1. $tB_2^n \subset K_N \iff t|x| \leqslant \|\Gamma x\|_\infty, \forall x \in \mathbb{R}^n$. *Then* $\forall x \in \mathbb{R}^n$, *we have* $t\sqrt{N}|x| \leqslant |\Gamma x| \Longrightarrow tB_2^n \subset K_N$

Proof. • $(\Longleftarrow)$.

Assume that $t\|x\|_2 \leqslant \|\Gamma x\|_\infty$ for all $x \in \mathbb{R}^n$. It suffices to show that $\mathbb{R}^n \backslash \Gamma^* B_1^N \subset \mathbb{R}^n \backslash tB_2^n$.

Pick $w \notin \Gamma^* B_1^N$, we need to show that $w \notin tB_2^n$, i.e. $\|w\|_2 > t$.

Since $\Gamma^* B_1^N$ is convex and symmetric, by the Hahn-Banach type separation theorem, there is a linear functional $\lambda : \mathbb{R}^n \to \mathbb{R}$ such that $|\lambda(x)| \leqslant 1$ for all $x \in K_N$ and $\lambda(w) > 1$, where $\lambda(x) = \langle x^*, x \rangle$ for some $x^* \in \mathbb{R}^n$.

For all $y \in B_1^N$, we have $|\lambda(\Gamma^* y)| \leqslant 1$, hence

$$|\langle y, \Gamma x^* \rangle| = |\lambda(\Gamma^* y)| \leqslant 1$$

So $\|\Gamma x^*\|_\infty \leqslant 1$, and by assumption $t\|x^*\|_2 \leqslant 1$. However, $\lambda(w) = \langle x^*, w \rangle > 1$, then

$$1 < \langle x^*, w \rangle \leqslant \|x^*\|_2 \|w\|_2 \leqslant \frac{1}{t}\|w\|_2 \Longrightarrow \|w\|_2 > t$$

($\Longrightarrow$).

$$\begin{aligned}
t|x|^2 &= \langle x, tx\rangle \\
&= \left\langle x, t\frac{x}{|x|}|x|\right\rangle \\
&\leqslant |x| \sup_{z\in tB_2^n} \langle x, z\rangle \\
&\leqslant |x| \sup_{z\in \Gamma^* B_1^N} \langle x, z\rangle \\
&= |x| \sup_{y\in B_1^N} \langle x, \Gamma^* y\rangle \\
&= |x| \sup_{y\in B_1^N} \langle \Gamma x, y\rangle \\
&= |x|\|\Gamma x\|_\infty \\
\Longrightarrow & t|x| \leqslant \|\Gamma x\|_\infty
\end{aligned}$$

Since $|x| \leqslant \sqrt{n}\|x\|_\infty$, if we have $t|x| \leqslant \frac{1}{\sqrt{N}}|\Gamma x|$, then we have $t|x| \leqslant \|\Gamma x\|_\infty$, hence $tB_2^n \subset K_N$. Notice that $t|x| \leqslant \frac{1}{\sqrt{N}}|\Gamma x|$ holds with high probability if we have a lower bound of $t\sqrt{N}$ on the smallest singular value of Γ with high probability. □

We introduce a similar lemma to show inclusion of the other direction:

Lemma 2. $K_N \subset tB_2^n \iff \|\Gamma x\|_\infty \leqslant t|x|, \forall x \in \mathbb{R}^n$.

Proof. The smallest t that satisfies the above inclusion is the circumradius of K_N, that is $\sup\{\|x\| : x \in K_N\}$. By the nature of K_N, the circumradius is the row of Γ with the largest ℓ_2 norm.

Note that $\|\Gamma\|_{\ell_2\to\ell_\infty} = \max_i \left(\sum_j |a_{ij}|^2\right)^{1/2}$. That is, the circumradius is $\max_{x\in\mathbb{R}^n} \frac{\|\Gamma x\|_\infty}{|x|}$. This finishes the proof. □

Perturbations with Independent Entries

If we assume that each entry in Γ is i.i.d, then we cannot perturb on $\mathbb{S}^{d-1}$. But sampling on the vertices of B_∞^d and i.i.d Gaussian both work.

Denote $B(\mu, \epsilon)$ the set of real random variables that satisfy

$$0 < r \leqslant \|\xi\|_{L^2}, \|\xi\|_{L^3} \leqslant \mu$$

Denote $M(N, d, \mu, \alpha_1, \alpha_2, r)$ as the family of matrices $\Gamma = [\xi_{ij}]$that satisfy:

$$\forall i, j, \xi_{ij} \in B(\mu, r), \mathbb{P}(\|\Gamma\| \geqslant \alpha_1\sqrt{N}) \leqslant e^{-\alpha_2 N}$$

Remark 1. Any sub-Gaussian random variable falls into this category. In particular, we can consider both Gaussian random matrix and ± 1 Bernoulli random matrix.

Theorem 1. *Theorem 3.1 from Litvak et al. (2005)*

Let $N = (1+\delta)n$ for some $\delta > 0$. Let Γ be an $N \times d$ random matrix from $M(N, d, \mu, \alpha_1, \alpha_2)$, for some $\mu \geqslant 1$ and $\alpha_1, \alpha_2 > 0$. There exist $\tilde{c}_1, \tilde{c}_2 > 0$ (dependent on α_1, μ) such that whenever $\delta \geqslant \tilde{c}_1/\ln(\tilde{c}_2 d)$, then

$$\mathbb{P}(s_d(\Gamma) \leqslant c_1\sqrt{N}) \leqslant \exp(-c_2 N)$$

where $c_1 > 0$ depends on δ, μ, α_1, and $c_2 > 0$ depends on μ, α_2. In particular, $s_d(\Gamma)$ is the smallest singular value of Γ.

We cite the following Corollary 4.1 from Litvak et al. (2005), which is a direct corollary of the above theorem.

Corollary 1. *Let $N = (1+\delta)d$ for some $\delta > 0$. There exist $\tilde{c}_1, \tilde{c}_2 > 0$ (dependent on α_1, μ) such that whenever $\delta \geqslant \tilde{c}_1/\ln(\tilde{c}_2 d)$, then*

$$\mathbb{P}(c_1 B_2^d \subset K_N) \geqslant 1 - \exp(-c_2 N)$$

where $c_1 > 0$ depends on δ, μ, α_1, and $c_2 > 0$ depend on μ, α_2.

The above theorem and corollary tell us that if we want a positive margin with probability at least $1-\delta$, then only $(1+\delta)d$ perturbations are needed per data. However, this is not quite satisfying, since we do not only want a positive margin, but also want a positive margin that is at least r.

To such end, we modify theorem 4.2 from Litvak et al. (2005):

Theorem 2. *Let $\Gamma \in M(N, d, \mu, \alpha_1, \alpha_2, \epsilon)$ and $K_N = \Gamma^\top B_1^N$. There exists an absolute constant $c_2 > 1$ such that $\forall \beta \in (0,1)$ and every d, N that satisfy:*

$$2^d \geqslant N \geqslant d\max\{\exp(C_\mu/\beta), (c_2\max\{\ln\alpha_1, 1/(1-\beta)^2\})^{1/(1-\beta)}\}$$

we have:

$$\mathbb{P}\left(\frac{r}{8}\left(B_\infty^d \cap \sqrt{\frac{\beta \ln N/d}{C_\mu}} B_2^d\right) \subset K_N\right) \geqslant 1 - \exp(-cd^\beta N^{1-\beta} r^3/5) - \exp(-a_2 N)$$

where $C_\mu = 12\ln(e\mu)$.

For Gaussian random matrix, we can get a slightly cleaner bound since intersection with B_∞^d is not necessary:

Remark 2. Let $C_\mu = 12\ln(e\mu), 2^n \geqslant N \geqslant n\max\{\exp(C_\mu/\beta), (c_2\max\{\ln a_1, 1/(1-\beta)^2\})^{1/(1-\beta)}\}$

$$\mathbb{P}(Cr\sqrt{\beta \ln(N/n)} B_2^n \subset K_N) \geqslant 1 - \exp(-cn^\beta N^{1-\beta} r^3)$$

For $\beta \in (0,1)$:

$$\mathbb{P}(C'r\sqrt{\beta \ln(N/n)} B_2^n \subset K_N) \leqslant 1 - \exp(-c'n^\beta N^{1-\beta} r^3)$$

where C, c, C', c' are absolute variables.

Remark 3. In particular, for ξ_{ij} that are sampled on the vertices of rB_∞^d, we have:

$$r \leqslant \|\xi\|_{L^2}, \|\xi\|_{L^3} = 0 \leqslant \mu = 1$$

Hence $C_\mu = 12$. We can find a large constant $C \geqslant \max\{\exp(12/\beta), (c_2\max\{\ln a_1, 1/(1-\beta)^2\})^{1/(1-\beta)}\}$ that only depends on β.

As a comparison, we also restate one of the theorems from our ICML submission:

Theorem 3. *Sample the augmented data on* $r\mathbb{S}^{d-1}$. *There is a universal constant* C *such that if* $|S'| = |T'| = N \geqslant Cd$ *and* $r \leqslant \beta^{-1/2}\sqrt{d/\log(N)}\gamma^*$ *for* $\beta > 1$, *with probability at least* $1 - 2e^{-d} - 2N^{1-\beta}$, *we have*

$$\alpha(S, S', T, T') \geqslant \frac{1}{8}\sqrt{\frac{\log(N/d)}{d}} r.$$

Taking $r = \beta^{-1/2}\sqrt{d/\log(N)}\gamma^*$, $\beta = 2$, and $N = \Omega(d^2)$, we can ensure that the margin achieved by DA is a constant fraction of γ^*, with high probability.

Invoking theorem 2, taking $r = \mathcal{O}(\sqrt{1/\log(N)}\gamma^*)$, $N = \Omega(d^2)$, and perturbing on rB_∞^d, we can ensure that the margin achieved by DA is a constant fraction of γ^*, with high probability.

Such perturbation does not violate linear separability because $B_\infty^d \subseteq \sqrt{d} B_2^d$, and from our ICML submission:

Theorem 4. *If* $N \geqslant 16d$ *and* $r \geqslant \frac{8e^2\sqrt{2d}}{\pi^{3/2}}\gamma^*$, *the probability that* $S \cup S'$ *and* T *are linearly inseparable is at least* $1 - 2e^{-d/6}$.

2.3 Distribution Independent Upper Bound on Margin

For the upper bound, we consider the function class of the "respectful" classifiers and the two-point classification scheme where $S = \{x\}$ and $T = \{y\}$ are singletons.

Definition 2.3.1. A function $f : \mathbb{R}^d \to \{\pm 1\}$ is respectful of S and T if $f(x) = 1$ for all $x \in \text{conv}(S)$ and $f(x) = -1$ for all $x \in \text{conv}(T)$.

For linearly separable datasets, the worst-case margin of the respectful classifiers coincides with the radius of the largest inscribed sphere of $\text{conv}(S)$:

Theorem 5. *Let* $S \cup S', T \cup T'$ *be linearly separable, and assume that* $rB_2^d \subseteq S \cup S'$ *and* $rB_2^d \subseteq T \cup T'$ *are the largest inscribed sphere of* $conv(S \cup S'), conv(T \cup T')$, *respectively. Then* $\alpha(S, S', T, T') = r$.

Proof. Consider f that classifies everything in $\text{conv}(S \cup S')$ to 1, and everything else to -1. Such f is respectful, and it has margin r. Hence $\alpha(S, S', T, T') \leqslant r$. On the other hand, since $rB_2^d \subset \text{conv}(S \cup S')$, no respectful classifier would achieve margin less than r. In conclusion, we have $\alpha(S, S', T, T') = r$. □

We cite theorem 13.2.1 from Matoušek (2002):

Theorem 6. *Let* B_2^d *denote the unit* ℓ_2 *ball in* $\mathbb{R}^d$, *and let* P *be a convex polytope contained in* B_2^d *and have at most* N *vertices. Then:*

$$\frac{vol(P)}{vol(B_2^d)} \leqslant \left(\frac{C \ln\left(\frac{N}{d} + 1\right)}{d} \right)^{d/2}$$

for some absolute constant C.

Now consider the case where $S = \{x\}$ and $T = \{y\}$ with $d_2(x, y) = 2\gamma^*$. If we perturb x and y on $\gamma^* \mathbb{S}^{d-1}$ centered on them and denote S', T' by the perturbation sets. We want to have $\frac{\gamma^*}{2} B_2^d \subseteq \text{conv}(S')$. This is the same as asking how large $|\text{conv}(S')|$ should be such that

$\frac{1}{2}B_2^d \subseteq \frac{1}{\gamma^*}\text{conv}(S') \subseteq B_2^d$. In particular we should have:

$$\frac{\text{vol}(\text{conv}(S'))}{\text{vol}(B_2^d)} \geqslant 2^{-d}$$

If $|\text{conv}(S')| < de^{\frac{d}{4C}}$, then by theorem 6, we would have $\frac{\text{vol}(\text{conv}(S'))}{\text{vol}(B_2^d)} < 2^{-d}$, which leads to a contradiction.

Hence, if we are perturbing of the order $O(\gamma^*)$, then for any algorithm, we have to augment exponential in d times to achieve a $O(\gamma^*)$ margin.

3 STATISTICAL INTERPRETATION

The geometric interpretation is useful for analyzing the adversarial robustness, by exploiting the inclusion theory from local theory of Banach spaces. We show that perturbing each datapoint by $m = O(d^2)$ times suffices to guarantee a non-zero margin. In reality, d could be huge, for example, for each picture in CIFAR-10, we need to perturb $(32 \times 32)^2 = 1048576$ times, which seems improbable. This is because the margin guarantee via DA is usually an "overkill" for adversarial robustness. Philosophically, DA should improve robustness in the average-case more efficiently, as opposed to its improvement in the worst-case. In this following chapter, we show that DA indeed helps generalization and average-case robustness, by stability and Rademacher complexity, respectively.

3.1 Data Augmentation and Stability

In this section, we discuss why learning with DA is stable. Throughout the section, we derive our bound using the squared loss for linear classification, the derivation for linear regression follows the same fashion. Similar results can be easily obtained for any convex and twice-differentiable loss function and any differentiable models. Following the regular learning conventions, we assume our original training data $S = ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{D}^n$ are generated i.i.d, where each $z_i = (x_i, y_i) \in X \times Y = Z$ is from a certain sample space. For each training data (x_i, y_i), we augment with m perturbations $((x_i + \xi_{i1}, y_i), \ldots, (x_i + \xi_{im}, y_i))$. Let each ξ_{ij} be bounded, 0 mean, and have variance σ^2. We also assume all the ξ's are drawn i.i.d. We use S to denote our training dataset, and $S^{(i)}$ to represent the dataset which we redraw the i^{th} sample from $\mathcal{D}$.

We write $f : X \to \widetilde{Y}$ as our model, where $f \in \mathcal{F}$ is from a certain pre-defined function class. Note that $\widetilde{Y}$ is not necessarily Y. We consider specifically binary classification, where $\widetilde{Y} = Y = \{\pm 1\}$. Let $\ell : \widetilde{Y} \times Y \to \mathbb{R}^+$ be our convex and twice-differentiable loss function. Denote the following symbols:

$$L_S(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

$$L_{S^{(i)}}(f) = \frac{1}{n}\left(\sum_{j\in[n], j\neq i} \ell(f(x_j), y_j) + \ell(f(x_i), y_i)\right)$$

$$\widetilde{L}_S(f) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \ell(f(x_i + \xi_{ij}), y_i)$$

$$\widetilde{L}_{S^{(i)}}(f) = \frac{1}{n}\left(\sum_{j\in[n], j\neq i} \frac{1}{m}\sum_{k=1}^{m} \ell(f(x_j + \xi_{jk}), y_j) + \frac{1}{m}\sum_{k=1}^{m} \ell(f(x_i + \xi_{ik}), y_i)\right)$$

$$R_S(f) = L_S(f) + \lambda\|f\|^2$$

$$R_{S^{(i)}}(f) = L_{S^{(i)}}(f) + \lambda\|f\|^2$$

$$A(S) = \arg\min R_S(f)$$

$$A(S^{(i)}) = \arg\min R_{S^{(i)}}(f)$$

$$\widetilde{A}(S) = \arg\min \widetilde{L}_S(f)$$

$$\widetilde{A}(S^{(i)}) = \arg\min \widetilde{L}_{S^{(i)}}(f)$$

Preliminary on Stability

The generic PAC-learning bound people try to control is the difference between the empirical loss and expected risk:

$$L_S(f) - \mathbb{E}_S[L_S(f)]$$

The most well-known bounds are derived using the Rademacher complexity and the VC-Dimension:

$$L_S(f) - \mathbb{E}_S[L_S(f)] \leqslant 2\mathcal{R}(\mathcal{F}, S) + \sqrt{\frac{\log 1/\delta}{2n}}$$

$$L_S(f) - \mathbb{E}_S[L_S(f)] \leqslant C\sqrt{\frac{VC(\mathcal{F})}{n}} + +\sqrt{\frac{\log 1/\delta}{2n}}$$

where $\mathcal{R}(\mathcal{F}, S) = \mathbb{E}_S\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i\in[n]} \sigma_i f(x, y)\right]$ is the Rademacher complexity of our function class $\mathcal{F}$, and $VC(\mathcal{F})$ is the VC-Dimension of our function class.

The drawback of these bounds is that we are essentially bound the *uniform deviation* of our function class:

$$\sup_{f\in\mathcal{F}}\left|L_S(f) - \mathbb{E}_S[L_S(f)]\right|$$

In reality, we never search throughout our entire function space $\mathcal{F}$, but rather a small portion of it. One way to get around is to study the behavior of our learning algorithm, via algorithmic stability discussed in Bousquet and Elisseeff (2002). We adopt the notions of stability in Shalev-Shwartz and Ben-David (2014):

Definition 3.1.1. Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be a monotonically decreasing function. We say that a learning algorithm A is on-average-replace-one-stable with rate $\epsilon(n)$ if for every distribution $\mathcal{D}$:

$$\mathop{\mathbb{E}}_{(S,z')\sim\mathcal{D}^{m+1},i\sim U(m)}\left[\ell\left(A\left(S^{(i)},z_i\right)\right)-\ell\left(A(S),z_i\right)\right]\leqslant\epsilon(m)$$

This notion precisely captures the generalization error:

Theorem 7. *Let $\mathcal{D}$ be a distribution. Let $S=(z_1,\ldots,z_n)$ be an i.i.d. sequence of examples and let z' be another i.i.d. example. Let $U(n)$ be the uniform distribution over $[n]$. Then, for any learning algorithm,*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S))-L_S(A(S))\right]=\mathop{\mathbb{E}}_{(S,z')\sim\mathcal{D}^{m+1},i\sim U(m)}\left[\ell\left(A\left(S^{(i)},z_i\right)\right)-\ell\left(A(S),z_i\right)\right]$$

Learning with DA is stable

Let us consider the squared loss for classification: $\ell(f(x),y)=(1-yf(x))^2$ and the linear classification rule: $f(x)=w^\top x$. Then:

$$\mathbb{E}[\ell(f(x+\xi),y)]=\mathbb{E}[(1-yw^\top(x+\xi))^2]=\ell(f(x),y)+\sigma^2\|w\|^2$$

For simplicity, we overload our symbol f to represent the induced function by the loss $\ell\circ f$, and omit the parameter y since perturbations do not alter the labels.

Fix $(x_1,\ldots,x_n)$, write $f(x_1^n,\xi_{11}^{nm})=f(x_1+\xi_{11},\ldots,x_1+\xi_{1m},\ldots,x_n+\xi_{n1},\ldots,x_n+\xi_{nm})$ as a compact notation, and write $f(x_1^n,\xi_{11}^{nm};\xi'_{lk})$ to represent we substitute ξ_{lk} with ξ'_{lk}, while keeping the other parameters unchanged. Hence:

$$\sup_{\xi_{lk},\xi'_{lk}}\left|\frac{1}{nm}\sum_{i\in[n],j\in[m]}f(x_1^n,\xi_{11}^{nm})-\frac{1}{nm}\sum_{i\in[n],j\in[m]}f(x_1^n,\xi_{11}^{nm};\xi'_{lk})\right|\leqslant\frac{C\|w\|}{nm}$$

for some constant C that depends on our perturbed dataset.

Then by McDiarmid's inequality, we have:

$$P\left(\left|\frac{1}{nm}\sum_{i\in[n]}\sum_{j\in[m]} f(x_1^n,\xi_{11}^{nm}) - \mathbb{E}\left[\frac{1}{nm}\sum_{i\in[n]}\sum_{j\in[m]} f(x_1^n,\xi_{11}^{nm})\right]\right| > \epsilon\right) \leqslant 2\exp\left(-\frac{2nm\epsilon^2}{C^2\|w\|^2}\right)$$

where the probability and the expectation are both taken with respect to ξ_{11}^{nm}.

To make the above event happen with probability less than δ, we only need $m = O(\frac{\log(1/\delta)C^2\|w\|^2}{2n\epsilon^2})$. So the more original data we have, the less perturbations we need per training data to make our model generalize.

We will need the following lemma for the formal proof:

Lemma 3. *If f is λ-strong convex, and w^* is the unique minimizer, then for any w, we have:*

$$\frac{\lambda}{2}\|w - w^*\|^2 \leqslant f(w) - f(w^*)$$

Let's also assume that $\left|\ell(f_1(x), y) - \ell(f_2(x), y)\right| \leqslant \rho\|f_1 - f_2\|$. By triangle inequality:

$$\|\widetilde{A}(S^{(i)}) - \widetilde{A}(S)\| \leqslant \underbrace{\|\widetilde{A}(S^{(i)}) - A(S^{(i)})\|}_{A} + \underbrace{\|A(S^{(i)}) - A(S)\|}_{B} + \underbrace{\|A(S) - \widetilde{A}(S)\|}_{C}$$

We will bound each term separately.

Note that $B \leqslant \frac{2\rho}{\lambda n}$ according to section 13.3.1 Shalev-Shwartz and Ben-David (2014), where ρ is the Lipschitz constant of the loss function.

For A, note that $A(S^{(i)})$ minimizes $R_{S^{(i)}}(f)$, which is 2λ-strongly convex, hence:

$$\begin{aligned}\lambda\|\widetilde{A}(S^{(i)}) - A(S^{(i)})\|^2 &\leqslant R_{S^{(i)}}\left(\widetilde{A}(S^{(i)})\right) - R_{S^{(i)}}\left(A(S^{(i)})\right)\\ &\leqslant \widetilde{L}_{S^{(i)}}(\widetilde{A}(S^{(i)})) + \epsilon - \left(\widetilde{L}_{S^{(i)}}(A(S^{(i)})) - \epsilon\right)\\ &\leqslant 2\epsilon\end{aligned}$$

where the second step happens with high probability, the last step is due to the optimality of $\widetilde{A}(S^{(i)})$. Similarly we will have:

$$\lambda\|A(S) - \widetilde{A}(S)\|^2 \leqslant 2\epsilon$$

Combine everything and we get:

$$\|\widetilde{A}(S^{(i)}) - \widetilde{A}(S)\| \leqslant \sqrt{\frac{8\epsilon}{\lambda}} + \frac{2\rho}{\lambda n}$$

By the lipschitz continuity of ℓ, and write $z = (x, y)$, we have:

$$\ell(\widetilde{A}(S^{(i)}, z) - \ell(\widetilde{A}(S), z) \leqslant \sqrt{\frac{8\epsilon\rho^2}{\lambda}} + \frac{2\rho^2}{\lambda n}$$

uniformly over the choice of z. Invoking theorem 7, we obtain:

Theorem 8. *Assume that the loss function is convex, twice-differentiable and ρ-Lipschitz, with $m = O(\frac{\log(1/\delta)C^2\|w\|^2}{2n\epsilon^2})$ perturbations per data. Then the ERM rule with DA is on-average-replace-one-stable with rate $\sqrt{\frac{8\epsilon\rho^2}{\lambda}} + \frac{2\rho^2}{\lambda n}$ with probability at least $1 - \delta$. It follows from theorem 7 that:*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) - L_S(A(S))\right] \leqslant \sqrt{\frac{8\epsilon\rho^2}{\lambda}} + \frac{2\rho^2}{\lambda n}$$

3.2 Data Augmentation and Robustness

In the previous section we show that learning with DA helps our model generalize, with only $O(\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$ perturbations in total. Another important feature that people notice in practice is that DA also helps the model to be robust, at least on average, if not adversarial.

In this chapter, we capture the difference between the training loss and the population robust risk.

Learning with Interdependent Data

In its essence we are solving a learning problem with interdependent data. A previous work of Amini and Usunier (2015) gave a Rademacher complexity upper bound on learning with ranking data. Their bound was obtained by exploiting the concentration of sum of dependent random variables, using the Janson's inequality in Janson (2004). Notice that in their setting, the interdependent datasets are generated deterministically, rather than randomly, as in our setting. We cite Janson's inequality here for individual interests:

Definition 3.2.1. Let $X = \sum_{\alpha\in\mathcal{A}} Y_\alpha$.

- A subset $\mathcal{A}'$ of $\mathcal{A}$ is independent if RV $\{Y_\alpha\}_{\alpha\in c\mathcal{A}'}$ are independent.
- A family $\{\mathcal{A}_j\}_j$ of subset of $\mathcal{A}$ is a cover of $\mathcal{A}$ if $\bigcup_j \mathcal{A}_j = \mathcal{A}$.
- A cover is proper if each $\mathcal{A}_j$ is independent.
- $\chi(A)$ is the size of the smallest proper cover of $\mathcal{A}$.

Theorem 9. *Let* $X = \sum_{\alpha\in\mathcal{A}} Y_\alpha$, *and* $Y_\alpha \in [a_\alpha, b_\alpha]$ *for every* $\alpha \in \mathcal{A}$, *then for any* $t > 0$*:*

$$\mathbb{P}(X \geqslant \mathbb{E}X + t) \leqslant \exp\left(-2\frac{t^2}{\chi(\mathcal{A})\sum_{\alpha\in\mathcal{A}}(b_\alpha - a_\alpha)^2}\right)$$

Learning with DA is Robust to Noise

Robustness to random noises means that during test time, even if we add a random noise to the test data, the prediction of our model would not differ too much. That is, we want to control the following term:

$$P_{(x,y)\sim\mathcal{D},\xi\sim\mathcal{F}}(f(x) \neq y) = \mathbb{E}_{(x,y)\sim\mathcal{D},\xi\sim\mathcal{M}}[\mathbb{1}\{f(x) \neq y\}]$$

Naturally, we consider the $0-1$ loss and bound the following term:

$$\frac{1}{nm}\sum_{i\in[n],j\in[m]} f(x_i+\xi_{ij}) - \mathbb{E}_{x\sim D,\xi\sim\mathcal{M}}[f(x+\xi)]$$

Here we assume that during the training time, we know what kind of noise will be added during testing. Note that $\forall i, j$, the training sample $x_i+\xi_{ij}$ and the test sample $x+\xi$ are from the same distribution, but our perturbed training samples are not i.i.d. So the question is: how to derive generalization error bound on non-i.i.d training samples?

To that end, we derive the following bound on the uniform deviation:

$$\begin{aligned}
&\sup_f\left(\frac{1}{nm}\sum_{i\in[n],j\in[m]} f(x_i+\xi_{ij}) - \mathbb{E}_{x,\xi}[f(x+\xi)]\right)\\
\leqslant&\sup_f\left(\frac{1}{m}\sum_{j\in[m]}\frac{1}{n}\sum_{i\in[n]} f(x_i+\xi_{ij}) - \mathbb{E}_{x,\xi}[f(x+\xi)] + \frac{1}{n}\sum_i \mathbb{E}_\xi[f(x_i+\xi)] - \frac{1}{n}\sum_i \mathbb{E}_\xi[f(x_i+\xi)]\right)\\
\leqslant&\underbrace{\sup_f\left(\frac{1}{m}\sum_{j\in[m]}\frac{1}{n}\sum_{i\in[n]} f(x_i+\xi_{ij}) - \frac{1}{n}\sum_i \mathbb{E}_\xi[f(x_i+\xi)]\right)}_{A}\\
&\qquad+\underbrace{\sup_f\left(\frac{1}{n}\sum_i \mathbb{E}_\xi[f(x_i+\xi)] - \mathbb{E}_x\mathbb{E}_\xi[[f(x+\xi)]]\right)}_{B}
\end{aligned}$$

B is an easier term to bound, note that B only depends on x_i, and they are drawn independently. We can follow the standard Rademacher complexity argument. Let $\tilde{\mathcal{F}} = \{\tilde{f} : \tilde{f}(x) = \mathbb{E}_\xi[f(x+\xi)]\}$. The caveat is that when we apply the McDiarmid's inequality to B

(we write $f(x_i + \xi; x'_k)$ to denote we substitute the k^{th} sample with x'_k):

$$\begin{aligned}
&\sup_{x_k, x'_k} \left(\sup_f \left(\frac{1}{n} \sum_i \mathbb{E}_\xi[f(x_i + \xi)] - \mathbb{E}_x \mathbb{E}_\xi[[f(x + \xi)]] \right) \right. \\
&\qquad \left. - \sup_f \left(\frac{1}{n} \sum_i \mathbb{E}_\xi[f(x_i + \xi; x'_k)] - \mathbb{E}_x \mathbb{E}_\xi[[f(x + \xi)]] \right) \right) \\
&\leqslant \sup_{x_k, x'_k} \sup_f \left(\frac{1}{n} \sum_i \mathbb{E}_\xi[f(x_i + \xi)] - \frac{1}{n} \sum_i \mathbb{E}_\xi[f(x_i + \xi; x'_k)] \right) \\
&= \sup_{x_k, x'_k} \sup_f \left(\frac{1}{n} \sum_i \left(\mathbb{E}_\xi[f(x_i + \xi)] - \mathbb{E}_\xi[f(x_i + \xi; x'_k)] \right) \right) \\
&= \sup_{x_k, x'_k, f} \left(\frac{1}{n} \mathbb{E}_\xi[f(x_k + \xi)] - \mathbb{E}_\xi[f(x'_k + \xi)] \right) \\
&\leqslant \frac{1}{n\eta(f, \xi)}
\end{aligned}$$

where $\frac{1}{\eta(f,\xi)} = \sup_{x_k, x'_k} (\tilde{f}(x_k) - \tilde{f}(x'_k))$ depends on our function class and noise. Note this number can be strictly less than 1. Now since all the x_i's are i.i.d, by the standard Rademacher complexity argument we have with probability $1 - 2\delta$:

$$B \leqslant 2\mathcal{R}(\tilde{\mathcal{F}}, x_1^n) + \sqrt{\frac{\log \frac{1}{\delta}}{2n\eta}} \leqslant 2\hat{\mathcal{R}}(\tilde{\mathcal{F}}, x_1^n) + 3\sqrt{\frac{\log \frac{1}{\delta}}{2n\eta}}$$

where $\mathcal{R}$ is the Rademacher complexity and $\hat{\mathcal{R}}$ is the empirical Rademacher complexity.

For A, fix the original training data $x_1, \ldots, x_n$, then for all i, $(x_i + \xi_{i1}, y_i), \ldots, (x_i + \xi_{im}, y_i)$ are independently distributed (although not identically). Hence we can apply the

McDiarmid's inequality wrt to $\xi_{ij}, \forall k \in [n], l \in [m]$:

$$\begin{aligned}
&\sup_{\xi_{kl},\xi'_{kl}} \left(\sup_f \left(\frac{1}{m} \sum_{j\in[m]} \frac{1}{n} \sum_{i\in[n]} f(x_i + \xi_{ij}) - \frac{1}{n} \sum_i \mathbb{E}_\xi[f(x_i + \xi)] \right) - \right. \\
&\qquad \left. \sup_f \left(\frac{1}{m} \sum_{j\in[m]} \frac{1}{n} \sum_{i\in[n]} f(x_i + \xi_{ij}; \xi'_{kl}) - \frac{1}{n} \sum_i \mathbb{E}_\xi[f(x_i + \xi)] \right) \right) \\
\leqslant &\sup_{\xi_{kl},\xi'_{kl}} \sup_f \left(\frac{1}{m} \sum_{j\in[m]} \frac{1}{n} \sum_{i\in[n]} f(x_i + \xi_{ij}) - \frac{1}{m} \sum_{j\in[m]} \frac{1}{n} \sum_{i\in[n]} f(x_i + \xi_{ij}; \xi_{kl}) \right) \\
\leqslant &\frac{1}{nm}
\end{aligned}$$

Let $\Phi(x_1^n, \xi_{11}^{nm})$ denotes the uniform deviation

$$\sup_f \left(\frac{1}{m} \sum_{j\in[m]} \frac{1}{n} \sum_{i\in[n]} f(x_i + \xi_{ij}) - \frac{1}{n} \sum_i \mathbb{E}_\xi[f(x_i + \xi)] \right)$$

we have with probability $1 - \delta$:

$$\Phi(x_1^n, \xi_{11}^{nm}) \leqslant \mathbb{E}_{\xi_{11}^{nm}}[\Phi(x_1^n, \xi_{11}^{nm}) | x_1^n] + \sqrt{\frac{\log \frac{1}{\delta}}{2nm}}$$

Due to the conditional independence of ξ_{11}^{nm} given x_1^n, we can apply the symmetrization argument:

$$
\begin{aligned}
&\mathbb{E}_{\xi_{11}^{nm}}[\Phi(x_1^n, \xi_{11}^{nm})|x_1^n] \\
=&\mathbb{E}_{\xi_{11}^{nm}}\left[\sup_f\left(\frac{1}{m}\sum_{j\in[m]}\frac{1}{n}\sum_{i\in[n]} f(x_i+\xi_{ij}) - \mathbb{E}_{\xi'_{11},\ldots,\xi'_{nm}}\frac{1}{n}\sum_i\frac{1}{m}\sum_{j=1}^m f(x_i+\xi'_{ij})\right)\right] \\
\leqslant&\mathbb{E}_{\xi_{11}^{nm},\xi'_{11},\ldots,\xi'_{nm}}\left[\sup_f\left(\frac{1}{m}\sum_{j\in[m]}\frac{1}{n}\sum_{i\in[n]} f(x_i+\xi_{ij}) - \frac{1}{n}\sum_i\frac{1}{m}\sum_{j=1}^m f(x_i+\xi'_{ij})\right)\right] \\
\leqslant&\mathbb{E}_{\xi_{11}^{nm},\xi'_{11},\ldots,\xi'_{nm},\sigma_{11}^{nm}}\left[\sup_f\left(\frac{1}{nm}\sum_{i,j}\sigma_{ij}(f(x_i+\xi_{ij}) - f(x_i+\xi'_{ij}))\right)\right] \\
\leqslant&\mathbb{E}_{\xi_{11}^{nm},\sigma_{11}^{nm}}\left[\sup_f\left(\frac{1}{nm}\sum_{i,j}\sigma_{ij}f(x_i+\xi_{ij})\right)\right] + \mathbb{E}_{\xi'_{11},\ldots,\xi'_{nm},\sigma_{11}^{nm}}\left[\sup_f\left(\frac{1}{nm}\sum_{i,j}-\sigma_{ij}f(x_i+\xi'_{ij})\right)\right] \\
\leqslant&2\mathbb{E}_{\xi_{11}^{nm}}\left[\widehat{\mathcal{R}}(\mathcal{F}, x_1^n+\xi_{11}^{nm})|x_1^n\right] \\
\leqslant&2\widehat{\mathcal{R}}(\mathcal{F}, x_1^n+\xi_{11}^{nm}) + 2\sqrt{\frac{\log\frac{1}{\delta}}{2nm}}
\end{aligned}
$$

where we write $\widehat{\mathcal{R}}$ as the empirical Rademacher complexity, and $\mathcal{F}$ as our original function class. The last step holds because of McDiarmid's inequality and the conditional independence of ξ_{11}^{nm}. Combining everything together, we get with probability $1-2\delta$:

$$
A = \Phi(x_1^n, \xi_{11}^{nm}) \leqslant 2\widehat{\mathcal{R}}(\mathcal{F}, x_1^n+\xi_{11}^{nm}) + 3\sqrt{\frac{\log\frac{1}{\delta}}{2nm}}
$$

Combining the arguments for A and B, we have with probability $1-4\delta$:

$$
\frac{1}{nm}\sum_{i\in[n],j\in[m]} f(x_i+\xi_{ij}) - \mathbb{E}_{x,\xi}[f(x+\xi)] \leqslant 2\widehat{\mathcal{R}}(\mathcal{F}, x_1^n+\xi_{11}^{nm}) + 3\sqrt{\frac{\log\frac{1}{\delta}}{2nm}} + 2\widehat{\mathcal{R}}(\tilde{\mathcal{F}}, x_1^n) + 3\sqrt{\frac{\log\frac{1}{\delta}}{2n\eta}}
$$

This gives us an upper bound on the convergence rate of learning with DA.

Estimate $\widehat{\mathcal{R}}(\mathcal{F}, x_1^n + \xi_{11}^{nm})$

We give an example to show how the second moment of the noise impacts the Rademacher complexity.

Consider linear classification rule $f = w^\top x$ where $\|w\| \leqslant W$ and $\|x\| \leqslant X$. We have:

$$\begin{aligned}
\widehat{\mathcal{R}}(\mathcal{F}, x_1^n + \xi_{11}^{nm}) &= \mathbb{E}_\sigma \sup_{w:\|w\|\leqslant W} \frac{1}{nm} \sum_{i\in[n],j\in[m]} \sigma_{ij} \left\langle w, x_i + \xi_{ij} \right\rangle \\
&= \frac{1}{nm} \mathbb{E}_\sigma \sup_{w:\|w\|\leqslant W} \left\langle w, \sum_{i\in[n],j\in[m]} \sigma_{ij}(x_i + \xi_{ij}) \right\rangle \\
&\leqslant \frac{1}{nm} \mathbb{E}_\sigma W \left\| \left\langle w, \sum_{i\in[n],j\in[m]} \sigma_{ij}(x_i + \xi_{ij}) \right\rangle \right\| \\
&\leqslant \frac{W}{nm} \sqrt{\mathbb{E}_\sigma \left[\left\langle \sum_{i,j} \sigma_{ij}(x_i + \xi_{ij}), \sum_{k,l} \sigma_{kl}(x_k + \xi_{kl}) \right\rangle \right]} \\
&= \frac{W}{nm} \sqrt{\sum_{i,j} \left\langle x_i + \xi_{ij}, x_i + \xi_{ij} \right\rangle} \\
&= \frac{W}{nm} \sqrt{\sum_{i,j} (\|x_i\|^2 + \|\xi_{ij}\|^2 + 2 \left\langle x_i, \xi_{ij} \right\rangle)} \\
&\leqslant \frac{W}{nm} \sqrt{nmX + \sum_{i,j} (\|\xi_{ij}\|^2 + 2 \left\langle x_i, \xi_{ij} \right\rangle)}
\end{aligned}$$

Now if we are dealing with the expected Rademacher complexity and assume that ξ has 0 mean and second moment σ^2:

$$\begin{aligned}
\mathbb{E}_\xi \widehat{\mathcal{R}}(\mathcal{F}, x_1^n + \xi_{11}^{nm}) &\leqslant \mathbb{E}_\xi \frac{W}{nm} \sqrt{nmX + \sum_{i,j} (\|\xi_{ij}\|^2 + 2 \left\langle x_i, \xi_{ij} \right\rangle)} \\
&\leqslant \frac{W}{nm} \sqrt{nmX + \sum_{i,j} (\mathbb{E}_\xi \|\xi_{ij}\|^2 + 2\mathbb{E}_\xi \left\langle x_i, \xi_{ij} \right\rangle)} \\
&\leqslant \frac{W}{nm} \sqrt{nm(X + \sigma^2)} \\
&\leqslant \frac{W\sqrt{X + \sigma^2}}{\sqrt{nm}}
\end{aligned}$$

4 DISCUSSION

In this thesis, we show that DA helps the adversarial robustness, but at a rather high cost. However, for generalization and noise robustness, DA helps with a small sample complexity.

A question remains: how does our Rademacher complexity upper bound indicate larger m helps more? What if we set $m = 1$ and use the i.i.d version of the Rademacher bound? By Jensen's inequality, we have:

$$\mathbb{E}_{\xi}\mathbb{E}_{\sigma}[\sup_{f}\frac{1}{n}\sum_{i}\sigma_i f(x_i+\xi_i)] \geqslant \mathbb{E}_{\sigma}\sup_{f}\frac{1}{n}\sum_{i}\sigma_i\mathbb{E}_{\xi}[f(x_i+\xi)]$$

but the magnitude of the gap remains unclear.

Another open question is the lower bound on noise robustness: can we construct one data distribution, such that our noise robust generalization lower bound decreases as m increases?

We leave these questions to future works.

REFERENCES

Amini, Massih-Reza, and Nicolas Usunier. 2015. *Learning with partially labeled and interdependent data*. Springer.

Bastani, Osbert, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya Nori, and Antonio Criminisi. 2016. Measuring neural net robustness with constraints. In *Advances in neural information processing systems*, 2613–2621.

Bishop, Chris M. 1995. Training with noise is equivalent to tikhonov regularization. *Neural computation* 7(1):108–116.

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

Carlini, Nicholas, and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Cullina, Daniel, Arjun Nitin Bhagoji, and Prateek Mittal. 2018. Pac-learning in the presence of evasion adversaries. *arXiv preprint arXiv:1806.01471*.

Fawzi, Alhussein, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2016. Robustness of classifiers: from adversarial to random noise. In *Advances in neural information processing systems*, 1632–1640.

Franceschi, Jean-Yves, Alhussein Fawzi, and Omar Fawzi. 2018. Robustness of classifiers to uniform ℓ_p and Gaussian noise. In *International conference on artificial intelligence and statistics, AISTATS 2018, 9-11 april 2018, playa blanca, lanzarote, canary islands, spain*, 1280–1288.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *CoRR* abs/1412.6572. 1412.6572.

Janson, Svante. 2004. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms* 24(3):234–248.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Kuznetsova, Alina, Sung Ju Hwang, Bodo Rosenhahn, and Leonid Sigal. 2015. 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*, 28–36.

Litvak, Alexander E, Alain Pajor, Mark Rudelson, and Nicole Tomczak-Jaegermann. 2005. Smallest singular value of random matrices and geometry of random polytopes. *Advances in Mathematics* 195(2):491–523.

Matoušek, Jiří. 2002. *Lectures on discrete geometry*, vol. 212. Springer New York.

Misra, Ishan, Abhinav Shrivastava, and Martial Hebert. 2015. Watch and learn: Semi-supervised learning for object detectors from video. In *The ieee conference on computer vision and pattern recognition (cvpr)*.

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. 2018. Robustness of classifiers to universal perturbations: A geometric perspective.

Prest, Alessandro, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. 2012. Learning object class detectors from weakly annotated video. In *Computer vision and pattern recognition (cvpr), 2012 ieee conference on*, 3282–3289. IEEE.

Shalev-Shwartz, Shai, and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Wager, Stefan, Sida Wang, and Percy S Liang. 2013. Dropout training as adaptive regularization. In *Advances in neural information processing systems*, 351–359.

Wong, Eric, and Zico Kolter. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, 5283–5292.

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

Zantedeschi, Valentina, Maria-Irina Nicolae, and Ambrish Rawat. 2017. Efficient defenses against adversarial attacks. In *Proceedings of the 10th acm workshop on artificial intelligence and security*, 39–49. ACM.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.