Consistent Adversarially Robust Linear Classification: Non-Parametric Setting

Elvis Dohmatob [1]

Abstract

For binary classification in d dimensions, it is known that with a sample size of n, an excess adversarial risk of $O(d/n)$ is achievable under strong parametric assumptions about the underlying data distribution (e.g., assuming a Gaussian mixture model). In the case of well-separated distributions, this rate can be further refined to $O(1/n)$. Our work studies the non-parametric setting, where very little is known. With only mild regularity conditions on the conditional distribution of the features, we examine adversarial attacks with respect to arbitrary norms and introduce a straightforward yet effective estimator with provable consistency w.r.t adversarial risk. Our estimator is given by minimizing a series of smoothed versions of the robust 0/1 loss, with a smoothing bandwidth that adapts to both n and d. Furthermore, we demonstrate that our estimator can achieve the minimax excess adversarial risk of $\widetilde{O}(\sqrt{d/n})$ for linear classifiers, at the cost of solving possibly rougher optimization problems.

1. Introduction

Supervised machine-learning models like (generalized) linear models, kernel methods, and neural networks have enjoyed tremendous practical success in many applications involving high-dimensional data, such as images. Yet, these models are highly sensitive to small perturbations known as adversarial examples (Szegedy et al., 2013; Tsipras et al., 2019; Su et al., 2017; Schmidt et al., 2018), which are often imperceptible by humans. While various strategies such as adversarial training (Madry et al., 2018) can mitigate this empirically, lack of robustness remains highly problematic for many safety-critical applications (e.g, autonomous vehicles), and motivates a better understanding of the phenomena at play. In the era of so-called large-language models (LLMs) like ChatGPT and their rapid adaptation by the general public for all kinds of duties like image generation, text generation, chatbots, encyclopedic knowledge assistants, etc., the impact of adversarial examples (in the form of adversarial *prompts*) will be even more devastating. For example see (Carlini et al., 2023) and the references therein.

In this work, we consider the statistical problem of computing a Bayes-optimal robust linear classifier for a binary classification problem on $\mathbb{R}^d$ without any parametric assumptions on the underlying distribution of the data, beyond the requirement of some regularity of the conditional density of the distribution of the features. Efficiently computable consistent estimators in the adversarial setting are a big deal because it is known that unlike the case of ordinary classification, convex surrogates don't exist in the adversarial case (Bao et al., 2020). To summarize,

– *Risk-Consistency.* We construct a sequence of linear classifiers (indexed by the sample size n and input dimension d), whose adversarial risk / classification error approaches the optimal value over all linear models, in the limit $n \to \infty$ such that $d \log(n)/n \to 0$. See Theorem 4.1. Our estimator relies on Gaussian smoothing of the data, wherein the smoothing bandwidth h_n gracefully adapts to the sample size n and the input dimension d. In the special case of Euclidean-norm attacks, our proposed estimator can be efficiently computed via smooth optimization methods on the Euclidean unit-sphere (a smooth manifold) (Boumal et al., 2018). The bandwidth parameter h_n trades between ease of optimization and rate of statistical convergence of the adversarial risk estimator to the Bayes-optimal value.

– *Minimax Optimality.* Under a stronger smoothness condition on the conditional distribution of the features, we recover in Theorem 4.2 rates which are uniformly optimal over all attack strengths, and match the minimax optimal rate of classification in the absence of adversaries.

2. Preliminaries and Problem Setting

2.1. Linear Classification and Adversarial Attacks

Consider a binary classification problem in $\mathbb{R}^d$, consisting of learning the label $y \in \{\pm 1\}$ of a random data point with feature vector $x \in \mathbb{R}^d$. Let P be the unknown joint distribution of the pair (x, y). We are given access to an iid

[1] Meta FAIR. Correspondence to: Elvis Dohmatob <dohmatob@meta.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

sample $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subseteq \mathbb{R}^d \times \{\pm 1\}$ from P. For any $w \in \mathbb{R}^d$, consider the linear (in fact, affine) function $f_w : \mathbb{R}^d \to \mathbb{R}$ defined by $f_w(x) := x^\top w$. This induces a linear classifier (aka half-space)

$$C_f(x) := \operatorname{sign}(f(x)) = \begin{cases} 1, & \text{if } f(x) \geq 0, \\ -1, & \text{else.} \end{cases} \tag{1}$$

Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ used to measure the strength of the attacker, and let $\|\cdot\|_\star$ be the dual norm, defined by

$$\|w\|_\star := \sup_{z \in \mathbb{R}^d,\, \|z\| \leq 1} z^\top w. \tag{2}$$

Popular examples in the literature include ℓ_p norm $\|\cdot\|_p$, with $p \in [1, \infty]$, especially for $p = 1$ (sparse attacks), $p = \infty$ (entry-wise bounded attacks), and $p = 2$. Note that if $p^\star \in [1, \infty]$ is the *harmonic conjugate* of p, i.e., if $1/p + 1/p^\star = 1$, then the dual norm for $\|\cdot\|_p$ is $\|\cdot\|_{p^\star}$. For example, the ℓ_2-norm is auto-dual, whilst the ℓ_1-norm and the ℓ_∞-norm are duals of one another.

If $w \neq 0$, we can always normalize it so that $\|w\|_\star = 1$, without modifying the induced classifier C_{f_w}. We therefore consider the following restricted subset of linear models

$$\mathcal{F}_{\text{lin}} := \{f_w \mid w \in \mathbb{R}^d,\, \|w\|_\star = 1\}. \tag{3}$$

Note that the same function space was considered in (Bao et al., 2020) in the specific case of Euclidean-norm attacks. Though the scope of our work is linear models, our results are applicable to the case of models with frozen representations (i.e., pre-training), since we do not make any structural / parametric assumptions on the distribution of the features. Equivalently, this case corresponds to scenarios where the attacker can inject perturbations into the feature map $\phi(x)$ of an input x at test time (Bietti & Mairal, 2019).

2.2. Adversarial Risk and Adversarial Regret

Definition 2.1 (Adversarial Risk). *Given an attack budget $\epsilon \geq 0$, the adversarial risk (aka robust classification error) $E_\epsilon(f) = E_\epsilon(f; P)$ of a model $f \in \mathcal{F}_{\text{lin}}$ is defined by*

$$E_\epsilon(f; P) := \mathbb{P}(\exists \delta \in \mathbb{R}^d,\, \|\delta\| \leq \epsilon \mid C_f(x + \delta) \neq y), \tag{4}$$

where (x, y) is a labelled random test example from P.

Note that $E_0(f)$ is the non-adversarial risk (aka ordinary classification error) of f. It is clear that $E_\epsilon(f) \geq E_0(f)$ for any $\epsilon \geq 0$. More generally, the function $\epsilon \mapsto E_\epsilon(f)$ is nondecreasing.

Definition 2.2 (Adversarial Regret). *Given a linear model $f \in \mathcal{F}_{\text{lin}}$, its adversarial regret is the quantity*

$$R_\epsilon(f) := E_\epsilon(f) - \inf_{g \in \mathcal{F}_{\text{lin}}} E_\epsilon(g), \tag{5}$$

i.e the excess adversarial risk of f relative to the optimum value over the function class $\mathcal{F}_{\text{lin}}$ of linear classifiers.

Note that, by design $\inf_{f \in \mathcal{F}_{\text{lin}}} R_\epsilon(f) = 0$ for all $\epsilon \geq 0$. Observe that, for any $f \in \mathcal{F}_{\text{lin}}$,

$$E_\epsilon(f) = \mathbb{P}(yf(x) \leq \epsilon) = \mathbb{E}_{x,y}\left[\theta_\epsilon(yf(x))\right], \tag{6}$$

where $\theta_\epsilon : \mathbb{R} \to \{0, 1\}$ is the unit-step function shifted to the right by an amount ϵ, i.e

$$\theta_\epsilon(t) := \theta_0(t - \epsilon) = \begin{cases} 0, & \text{if } t \geq \epsilon, \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

Indeed, for any $w \in \mathbb{R}^d$ with $\|w\|_\star = 1$, one computes

$$\begin{aligned} \inf_{\|\delta\| \leq \epsilon} y((x + \delta)^\top w) &= yx^\top w + \inf_{\|\delta\| \leq \epsilon} y\delta^\top w \\ &= yx^\top w - \epsilon\|w\|_\star = yx^\top w - \epsilon. \end{aligned}$$

Also see Proposition 3 of (Bao et al., 2020). The quantity $m_w(x, y) := yf_w(x) = yx^\top w$ has a geometric meaning as the *margin* of the data point (x, y) w.r.t to the decision-boundary (a hyperplane) induced by f_w. Indeed, in the euclidean setting, $|m_w(x, y)|/\|w\|_2$ is exactly equal to the distance of x to the decision boundary of the classifier f_w. Moreover, $m_w(x, y) > 0$ for points which are correctly classified by f_w, whilst $m_w(x, y) \leq 0$ for wrong classified points. Thus, (6) demands that even correctly classified points whose margin is not greater ϵ are counted as incorrectly classified, in the adversarial setting.

Given the training dataset $\mathcal{D}_n$ and the attack budget ϵ, the task of the machine-learner is to output a binary linear classifier $\widehat{f}_n \in \mathcal{F}_{\text{lin}}$.

2.3. The Adversarial 0/1 Loss

Our goal is to find the linear model $f \in \mathcal{F}_{\text{lin}}$ with the least possible adversarial risk $E_\epsilon(f)$. Since $E_\epsilon(f)$ depends on the unknown distribution P of (x, y), this cannot be done directly. Vanilla empirical risk minimization (ERM) corresponds to computing $\widehat{f}_n \in \mathcal{F}_{\text{lin}}$ which minimizes the empirical version of $E_\epsilon(f)$, namely

$$\begin{aligned} \widehat{E}_{n,\epsilon}(f) = E_\epsilon(f; \widehat{P}_n) &= \frac{1}{n} \sum_{i=1}^n \theta_\epsilon(y_i f(x_i)) \\ &= \frac{1}{n} \#\{i \in [n] \mid y_i f(x_i) \leq \epsilon\} \in [0, 1], \end{aligned} \tag{8}$$

where $\widehat{P}_n := (1/n) \sum_{i=1}^n \delta_{(x_i, y_i)}$ is the empirical distribution corresponding to the random dataset $\mathcal{D}_n$. Because of the discontinuity of the margin loss $t \mapsto \theta_\epsilon(t)$ at $t = \epsilon$, the empirical risk $\widehat{E}_{n,\epsilon}(f)$ cannot be optimized directly. Moreover, as established in (Bao et al., 2020; Dan et al., 2020), there is no convex surrogate for this loss which is Bayes-consistent. Our idea is to replace θ_ϵ by a smoothed surrogate, which is computationally tractable.

Definition 2.3. *An estimator $\widehat{f}_{n,\epsilon} \in \mathcal{F}_{\rm lin}$ computed on the training dataset $\mathcal{D}_n$ is said to be consistent if it has vanishing adversarial regret, i.e $R_\epsilon(\widehat{f}_{n,\epsilon}) \overset{n\to\infty}{\longrightarrow} 0$ for all $\epsilon \geq 0$.*

We stress that our definition of consistency is relative to linear models, which are the main focus of this paper.

2.4. Notations

The maximum of two real numbers a and b will be denoted $a \vee b$. The maximum of a and 0 is denoted a_+. As usual, Φ (resp. φ) denotes the Gaussian cumulative distribution (resp. probability density) function. Given a real matrix $A \in \mathbb{R}^{d\times m}$, its singular-values $\sigma_1(A) \geq \sigma_d(A) \geq \ldots \geq \sigma_d(A)$ correspond to the sorted list of positive square-roots eigenvalues of the positive-semidefinite matrix $A^\top A$. The operator norm of A, denoted $\|A\|_{op}$, corresponds to $\sigma_1(A)$, while the the Frobenius (aka Hilbert-Schmidt) norm of A, denoted $\|A\|_F$, corresponds to $\sqrt{\sum_{j=1}^d \sigma_j(A)^2}$. The number of nonzero singular-values of A, correspond to its rank, denoted $\mathrm{rank}(A)$. The effective rank of A, denote $r(A)$, is defined by $r(A) := \|A\|_F^2/\|A\|_{op}^2$. Note that if A is a nonzero matrix, then $1 \leq r(A) \leq \mathrm{rank}(A) \leq \min(m,d)$.

$O(n)$, $\Omega(n)$, $o(n)$, etc. are standard standard notations, while the avatars $\widetilde{O}(n)$, $\widetilde{\Omega}(n)$, etc., mean that there are hidden factors which are at most poly-logarithmic in n. All asymptotics will be w.r.t the limit $n \to \infty$. An event occurs with high probability (w.h.p) if its occurrence probability is at least $1 - o(1)$ in the limit $n \to \infty$.

3. The Proposed Estimator

3.1. Smooth Surrogate for Adversarial Loss

Let $Q : \mathbb{R} \to [0,1]$ be a function verifying the following conditions: (1) Q is strictly increasing. (2) The tail of Q behaves like $\lim_{t\to\infty} Q(t) = 0$. For example, the survival function of any random variable is such a function. Given a bandwidth parameter $h > 0$, define a function $\theta_{\varepsilon,h}$ by

$$\theta_{\varepsilon,h}(u) := Q_h(u - \epsilon), \text{ where } Q_h(u) := Q(u/h). \tag{9}$$

Note that if in addition Q is continuous, then $\theta_{\varepsilon,h}$ converges pointwise to θ_ε in the limit $h \to 0^+$. Let $\ell_{\varepsilon,h}$ be the loss function induced by $\theta_{\varepsilon,h}$, i.e., $\varepsilon \geq 0$ and $h > 0$,

$$\ell_{\varepsilon,h}(y,y') := \theta_{\varepsilon,h}(yy'), \tag{10}$$

This leads to a smoothed version of the empirical adversarial 0/1 risk defined in (11), namely

$$\begin{aligned}\widehat{E}_{n,\varepsilon,h}(f) &:= \frac{1}{n}\sum_{i=1}^n \ell_{\varepsilon,h}(y_i, f(x_i)) \\ &= \frac{1}{n}\sum_{i=1}^n \theta_{\varepsilon,h}(y_i f(x_i)) \in [0,1].\end{aligned} \tag{11}$$

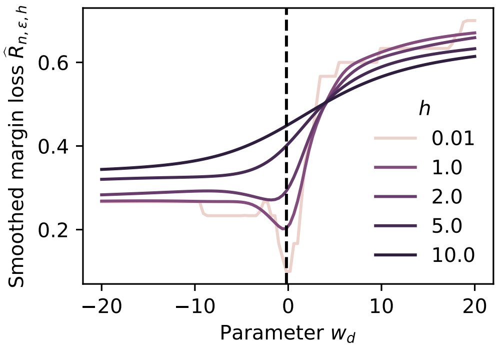

Figure 1. **The Effect of Smoothing.** Mild smoothing (small h) makes the underlying optimization problem computationally tractable without changing the value of the optimal empirical adversarial risk. Refer to Lemma 3.1. Extreme smoothing (large h) makes the optimization problem very easy but essentially destroys the information contained in the training dataset $\mathcal{D}_n$.

Note that in the nonrobust case $\epsilon = 0$, choosing $Q = Q_G =$ Gaussian survival function corresponds to the *probit loss* considered in (Keshet et al., 2011) in another context.

3.2. Link with Kernel Density Estimation

An important example is when Q is the *survival function* of a random variable with density. In this case, one can alternatively write the smoothed margin function (9) as a convolution product $\theta_{\varepsilon,h} = \theta_\varepsilon \star Q(\cdot/h)$, the smoothed loss then corresponds replacing the empirical marginal distribution of the data points $x_1,\ldots,x_n \in \mathbb{R}^d$ by its kernel density estimate (KDE), constructed via the smoothing function $s_Q := Q'$, i.e

$$\widehat{E}_{n,\varepsilon,h}(f) = E_{n,\varepsilon,h}(f;\widehat{P}_n) = E_\varepsilon(f; \underbrace{\widehat{P}_n \star s_Q(\cdot/h)}_{\text{KDE}}), \tag{12}$$

where $\star$ denotes convolution. Because of the way the adversarial risk functional E_ε is defined, the RHS in the above can be seen to correspond to a 1-dimensional kernel density estimation (KDE) for the distribution of margins $m_f(x,y) := yf(x)$, with $(x,y) \sim P$.

3.3. The Estimator

For an appropriate choice of smoothing bandwidth $h = h_n$, our proposed estimator $\widehat{f}_{n,\varepsilon,h}$ is defined as any minimizer of the smoothed empirical adversarial risk functional $\widehat{E}_{n,\varepsilon,h}$ over the linear function class $\mathcal{F}_{\rm lin}$, that is

$$\widehat{f}_{n,\varepsilon,h} \in \arg\min_{f\in\mathcal{F}_{\rm lin}} \widehat{E}_{n,\varepsilon,h}(f). \tag{13}$$

The rest of this section will be spent discussing algorithmic aspects of the above estimator. We begin with the following lemma establishes an explicit formula for a sub-gradient of the empirical smoothed risk $\widehat{E}_{n,\varepsilon,h}(f_w)$.

Lemma 3.1. *For any $w \in \mathbb{R}^d$, let $m_i(w) := y_i x_i^\top w$ be the margin of the data point (x_i, y_i) w.r.t the linear model f_w. Then, the smoothed empirical adversarial risk functional $w \mapsto \widehat{E}_{n,\varepsilon,h}(f_w)$ is differentiable on $\mathbb{R}^d$ with gradient*

$$\nabla_w \widehat{E}_{n,\varepsilon,h}(f_w) = \frac{1}{nh} \sum_{i=1}^{n} Q'\left((m_i - \varepsilon)/h\right) y_i x_i. \tag{14}$$

Moreover, if Q is twice differentiable with bounded second derivative, i.e $\|Q''\|_\infty \leq \alpha$ for some $\alpha \geq 0$, then for any $\varepsilon \geq 0$, the smoothed empirical adversarial risk functional $w \mapsto E_{n,\varepsilon,h}(f_w)$ is L-smooth on $\mathbb{R}^d$, with $L = \alpha\|\Sigma_n\|_{op}/h^2$ and Hessian given by

$$\nabla_w^2 \widehat{E}_{n,\epsilon,h}(f_w) = \frac{1}{nh^2} X^\top D_\epsilon(w) X, \tag{15}$$

where X is the $n \times d$ matrix with rows $x_1, \ldots, x_n$ (i.e the design matrix), $\Sigma_n := X^\top X/n$ is the empirical covariance matrix (aka the normalized Gram matrix), and $D_\epsilon(w)$ is the $n \times n$ diagonal matrix given by $D_\epsilon(w)_{ii} := Q''((m_i(w) - \varepsilon)/h)$ for all $i \in [n]$.

For example, in the cause of the Gaussian survival function $Q = Q_G$, it is easy to see that the hypothesis of proposition are verified with $\alpha = \|Q_G''\|_\infty = 1/\sqrt{2\pi e}$. Likewise, in the case where $Q = Q_S := 1 - \sigma$, with $\sigma(t) := 1/(1 + e^{-t})$ being the sigmoid, the hypothesis holds with $\alpha = \|\sigma''\|_\infty \leq 1$.

Proof of Lemma 3.1. Indeed, one can write

$$\widehat{E}_{n,\varepsilon,h}(f_w) = \frac{1}{n} \sum_{i=1}^{n} \theta_{\varepsilon,h}(m_i) = \frac{1}{n} \sum_{i=1}^{n} Q\left((m_i - \varepsilon)/h\right),$$

and so differentiating w.r.t w and noting that $\nabla m_i(w) = y_i x_i$ gives the first part of the result. Now, differentiating (14) to w gives

$$nh^2 \nabla_w^2 \widehat{E}_{n,\varepsilon,h}(f_w) = \sum_{i=1}^{n} Q''(\frac{m_i - \varepsilon}{h}) x_i x_i^\top = X^\top D_\epsilon(w) X.$$

It follows that uniformly on $w \in \mathbb{R}^d$, the operator norm of the Hessian $\nabla_w \widehat{E}_{n,\varepsilon,h}(f_w)$ is upper-bounded by

$$\begin{aligned}\|\nabla_w^2 \widehat{E}_{n,\varepsilon,h}(f_w)\|_{op} &\leq \|D_\epsilon(w)\|_{op} \frac{\|X\|_{op}^2}{nh^2} \\ &= \frac{\|D_\epsilon(w)\|_{op}\|\Sigma_n\|_{op}}{h^2} \leq \frac{\alpha\|\Sigma_n\|_{op}}{h^2},\end{aligned}$$

which completes the proof. □

3.4. Algorithm for Computing the Estimator

In the case where the attack is measured w.r.t Euclidean norm, or more generally, Mahalanobis norms, (13) thus corresponds to a smooth optimization optimization on a smooth sphere-like Riemannian manifold. This is a standard problem, and there are lots of efficient algorithms (Boumal et al., 2018). We use trust-region-based methods (Absil et al., 2007) implemented in the Manopt library (Boumal et al., 2014).

In the case of general ℓ_p-norms with $p \notin \{2, \infty\}$, the problem (13) can be tackled using the methods developed in (Sato, 2023).

Remark 3.1. Note that we are only able to guarantee convergence to a stationary point of $\widehat{E}_{n,\epsilon,h}$. However, in all our experiments, we observe that the numerically obtained stationary point is also the global optimum. A rigorous study of this aspect will done in a future work, possibly leveraging ideas from (Diakonikolas et al., 2019).

4. Statistical Analysis

We now establish a deviation bound which proves that, for an appropriate sample-size dependent sequence of bandwidths $h_n \to 0$, our proposed estimator $\widehat{f}_{n,\varepsilon,h_n} \in \mathcal{F}_{\text{lin}}$ defined in (13) is Bayes-consistent over the linear function class $\mathcal{F}_{\text{lin}}$ in the sense that its adversarial risk converges with large n to that of the optimal robust linear classifier.

4.1. Assumptions

We will consider the regime where the sample size n and the input-dimension d scale like so

$$n \to \infty \text{ such that } (d/n)\log n \to 0. \tag{16}$$

This includes the fixed-dimensional regime where d is bounded. The above condition is of course more general as d is allowed to grow with n, as long as the condition $(d/n)\log n \to 0$ is respected.

Definition 4.1. Given a random variable z, its *Lévy concentration function (LCF)* is defined for any $\delta \geq 0$ by

$$\operatorname{LCF}(z, \delta) := \sup_{u \in \mathbb{R}} \mathbb{P}(|z - u| \leq \delta). \tag{17}$$

LCFs control how much mass a random variable can concentrate around any given point. They are encountered in convex geometry, problems of random matrix theory, and concentration of random polynomials, just to name a few. We refer the reader to (Rudelson & Vershynin, 2014) and the numerous references therein.

For the statistical analysis of the adversarial regret of our proposed estimator (13), we will make the following mild regularity assumption on the distribution P of the data

Assumption 4.1 ("Small-Ball" Assumption)**.** *For a random labelled test data point $(x, y) \sim P$, assume that*

$$\sup_{\|w\|_\star = 1} \operatorname{LCF}(yx^\top w, \delta) \leq \eta(\delta), \tag{18}$$

for some increasing function η verifying $\lim_{\delta\to 0}\eta(\delta) = 0$.

Assumption 4.1 is very mild, and holds under rather very general conditions. We will see that the assumption holds if the distribution of the feature vector x conditioned $y = l$ for any fixed value $l \in \{\pm 1\}$ of the label y has density (w.r.t to Lebesgue measure on $\mathbb{R}^d$). The assumption will allow us control the variations of the margin loss $\theta_\epsilon : t \mapsto 1_{t\le\epsilon}$ in the vicinity of the discontinuity at $t = \epsilon$, uniformly over the linear function class $\mathcal{F}_{\text{lin}}$ and the attack strength $\varepsilon > 0$.

Remark 4.1. *In a completely different context, a condition similar to Assumption 4.1 has been used in the analysis of benign overfitting in linear neural networks regression (Chatterji et al., 2022) (see Assumption A.4 therein).*

It turns our that for Assumption 4.1 to hold, the following condition is sufficient.

Condition 4.1. *For any label $l \in \{\pm 1\}$, the distribution of the feature vector x conditioned the event $y = l$, admits a density (w.r.t Lebesgue measure on $\mathbb{R}^d$).*

Proposition 4.1. *If Condition 4.1 holds, then Assumption 4.1 prevails.*

4.2. Risk-Consistency of Our Proposed Estimator

The following is one of our main theoretical results. It establishes that the adversarial risk of the linear classifier $\widehat{f}_{n,\varepsilon,h_n}$ (13) converges to the optimal linear classifier. Note that $\widehat{f}_{n,\varepsilon,h_n}$ is a random function, as it depends on the random training data $\mathcal{D}_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$.

Theorem 4.1. *Let Assumption 4.1 be in order. In the limit (16), if $(h_n)_n$ is a sequence of bandwidths tending to 0, then w.h.p over training data $\mathcal{D}_n$, it holds that*

$$\sup_{\varepsilon\ge 0} R_\varepsilon(\widehat{f}_{n,\varepsilon,h_n}) = O\left(\lambda(h_n) + \sqrt{\frac{d\log n}{n}}\right) \overset{n\to\infty}{\longrightarrow} 0, \tag{19}$$

where $\lambda(h_n) := \eta(h_n\sqrt{\log 1/h_n}) + Q(\sqrt{\log 1/h_n})$ and the function η is as in Assumption 4.1.

4.3. Minimax Optimality

We now obtain quantitative rates for the excess adversarial risk under Assumption 4.1. We do this via an appropriate choice of bandwidths h_n. The following important consequence of Theorem 4.1.

Theorem 4.2. *Suppose Assumption 4.1 is in order. Let the smoothing function Q have exponential-tail (e.g Gaussian survival function Q_G). Consider the choice of bandwidth*

$$h_n \asymp \eta^{-1}\left(\sqrt{\frac{d\log n}{n}}\right)^{1+\Omega(1)}. \tag{20}$$

Then, in the limit (16), it holds w.h.p over $\mathcal{D}_n$ that

$$\sup_{\varepsilon\ge 0} R_\varepsilon(\widehat{f}_{n,\varepsilon,h_n}) = O\left(\sqrt{\frac{d\log n}{n}}\right) \overset{n\to\infty}{\longrightarrow} 0. \tag{21}$$

Note that in Theorem 4.2, the rate $\sqrt{d/n}$ (ignoring log-factors) is matches the minimax rate for non-parameteric classification. Let us examplify Theorem 4.2 in the case where the class-conditional distributions of the the features are multivariate Gaussians.

Corollary 4.1. *Suppose that for $(x, y) \sim P$, the distribution of yx is a mixture of (multivariate) Gaussians, and let the smoothing function Q_G and bandwidth $h_n \asymp (\sqrt{(d/n)\log n})^{1+\Omega(1)}$. Then, w.h.p over $\mathcal{D}_n$,*

$$\sup_{\varepsilon\ge 0} R_\varepsilon(\widehat{f}_{n,\varepsilon,h_n}) = O\left(\sqrt{\frac{d\log n}{n}}\right). \tag{22}$$

The result follows from Theorem 4.2 above and the following lemma by which $\eta(\delta) = O(\delta)$ for Gaussian mixtures.

Lemma 4.2. *Suppose there exists an absolute constants $b \ge 0$ such that for all $w \in \mathbb{R}^d$, the random variable $yx^\top w$ has density bounded by b. Then, Assumption 4.1 holds with $\eta(\delta) = 2b\delta$.*

In particular, this is the case when the distribution of $z := yx$ is a mixture of (multivariate) Gaussians.

Proof. Indeed, if $f_w \le b$ is the density of $yx^\top w$, then

$$\begin{aligned}\mathrm{LCF}(yx^\top w, \delta) &:= \sup_{u\in\mathbb{R}} \mathbb{P}(|yx^\top w - u| \le \delta)\\ &= \sup_{u\in\mathbb{R}} \int_{u-\delta}^{u+\delta} f_w(z)\mathrm{d}z \le b\cdot 2\delta = 2b\delta,\end{aligned}$$

which proves the first part of the claim.

In particular, if the distribution of $z \sim \sum_k \pi_k N(\mu_k, \Sigma)$ (a setting considered in (Dan et al., 2020)), then for every $w \in \mathbb{R}^d$, the random variable $Z^\top w$ has distribution $\sum_k \pi_k N(\mu_k^\top w, \|w\|_\Sigma)$, which has density bounded by $O(1/\|w\|_\Sigma) = O(1)$, over the sphere $\|w\|_\star = 1$. □

For example, in the case of euclidean-norm attacks, this bound reduces to the spectral norm $O(\|\Sigma^{-1}\|_{op})$.

5. Related Work

There has been a sustained interest in characterizing the statistical hardness of adversarial learning. In this section, we review the works which are most relevant to ours.

5.1. Non-Existence of Consistent Convex Surrogates

Consistency and calibration results (Bartlett et al., 2005) provide reassurance that optimizing a surrogate does not ultimately hinder the search for a function that achieves the risk, and thus allow such a search to proceed within the scope of computationally efficient algorithms. In the case of ordinary learning corresponding to the 0/1 loss $\theta_0 : t \mapsto 1_{t \le 0}$, it has been shown in (Bartlett et al., 2005) that consistent convex loss functions exist. In fact, all the classically used convex loss functions (logistic loss, etc.) are consistent for classification.

However, Bao et al. (2020) recently established negative results which expose a stark separation between tractability of learning half spaces in the aforementioned ordinary regime, and the case of robust / adversarial classification. More precisely, they show that in contrast to the case of ordinary learning where calibrated and consistent convex surrogates to the 0/1 loss $\theta_0 : t \mapsto 1_{t \le 0}$ exist Bartlett et al. (2006), no consistent loss functions exist for the adversarial 0/1 $\theta_\varepsilon : t \mapsto 1_{t \le \varepsilon}$ (with $\varepsilon \neq 0$), over linear models $f \in \mathcal{F}_{\text{lin}}$. The issue is the discontinuity at the point $t = \varepsilon$. In the case of ordinary / non-adversarial classification where $\varepsilon = 0$, this discontinuity can be handled by using a convex loss function. In the case of adversarial classification where $\varepsilon > 0$, such a convexification of θ_ε is not possible (Bao et al., 2020). In particular, this implies that regularization based heuristics for training would-be robust models is neither calibrated nor consistent for robustness, in general! An exception studied in (Awasthi et al., 2021) is the realizable case, where the data distribution is separable in the general sense (23). Our results don't contradict the observations of (Bao et al., 2020; Awasthi et al., 2021) , since actually, our surrogate losses will be non-convex.

5.2. Known Positive Results

Well-Separated and Parametric Settings. A number of works have proposed linear classifiers which provably achieve consistency with quantitative rates in favorable settings. Notably, in the case of Gaussian and Bernoulli mixture models which are reminiscent of a parametric assumption, Schmidt et al. (2018) established a d/n minimax lower-bound on the adversarial regret (5). Corresponding upper-bounds of $\sqrt{d}/n$ and d/n were established for these data distributions. Moreover, the bound d/n has been shown to be tight in general (Schmidt et al., 2018; Dan et al., 2020; Bhattacharjee et al., 2021). Similar results were obtained in (Dan et al., 2020) in the case of Gaussian mixtures. In the case of well-separated data, (Bhattacharjee et al., 2021) proposed a modified max-margin algorithm which achieves a $1/n$ bound on the adversarial risk. Such so-called fast rates are reminiscent large-margin assumptions, which can be fully captured by the so-called *Massart noise condition* (Massart & Nédélec, 2006)

$$|\mathbb{P}(y = 1 \mid x) - 1/2| \ge \gamma, \tag{23}$$

for some $\gamma \in (0, 1/2]$ and almost all $x \in \mathbb{R}^d$. In contrast to the aforementioned works, we consider the case where no such parametric or large-margin assumption on the underlying distribution is made, and establish a non-parametric error rate of $\sqrt{d/n}$ (ignoring log-factors) (Vapnik & Chervonenkis, 1971; Massart & Nédélec, 2006). Moreover, this rate is optimal since it is well-known that $\sqrt{d/n}$ is optimal in classical / non-robust classification without further distributional assumptions like (23).

The General Setting. Departing from the aforementioned parametric of well-separatedness assumptions, (Bhattacharjee & Chaudhuri, 2020) proposed a two-stage estimator which is provably consistent. The first stage consists in pruning the dataset $\mathcal{D}_n$ so as to only keep a maximal ϵ'-separated subset, for some $\epsilon' > \epsilon$. The second stage amounts to running a weighted nearest neighbors classifier on the pruned dataset. Let us mention some weaknesses of the above method. First, though quite general, the resulting estimator suffers a notable irreducible computational drawback due to the first stage. Indeed, computing a maximal ϵ'-separated subset is reminiscent of finding a maximal independent set on a bipartite graph with n vertices. Due to NP-hardness (Feige, 2002), the complexity of this problem is impractical in the large-n limit. Furthermore, the consistency results in (Bhattacharjee & Chaudhuri, 2020) for their proposed estimator are non-quantitative (i.e no rates).

Since the setting of the method proposed in ((Bhattacharjee & Chaudhuri, 2020) described above is most similar to the setting considered in our work, we now provide a detailed comparison with our proposed estimator (13). First, from a computational standpoint, observe from Lemma 3.1 that our proposed estimator corresponds to a non-convex but L-smooth optimization problem with $L = O(\|\Sigma_n\|_{op}/h_n^2)$, where we have the choice of smoothing bandwidth h_n such that $h_n \to 0$. As in Lemma 3.1, $\Sigma_n := X^\top X/n$ is the empirical covariance matrix of the features. For concreteness if we take $\|\Sigma_n\|_{op} = O(1)$, then $L = O(1/h_n^2)$ which controls the computational complexity associated with computing our estimator. On the other hand, the adversarial regret of our estimator is of order $\lambda(h_n) + \sqrt{d\log(n)/n}$ (where λ is a decreasing function depending on η in Assumption 4.1) thanks to Theorem 4.1, allowing us to trade-off between computational complexity / optimization difficulty (small large L) and rate of statistical convergence by appropriately tuning h_n. In particular, if we are only interested in consistency (and no quantitative rates), then our proposed estimator can be made to run in linear time (i.e fast optimization) by tuning the smoothing bandwidth h_n such that $h_n \to 0$ only very slowly.

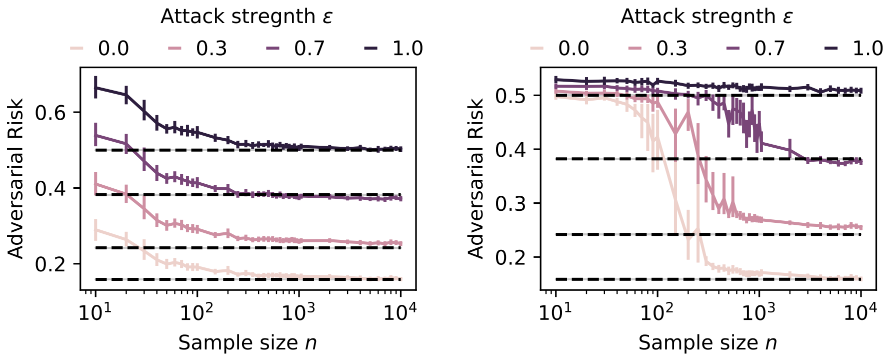

Figure 2. **Gaussian Experiment.** Showing the adversarial risk (6) of our proposed estimator $\widehat{f}_{n,\epsilon,h_n}$ defined in (13) as a function of sample size n and Euclidean-norm attack strength ϵ. **Left:** isotropic features (i.e $\Sigma = I_d$). **Right:** Non-isotropic features. The horizontal broken lines correspond to the Bayes-optimal adversarial risk (25) the given value of ϵ. Error-bars correspond to 10 runs of computing our estimator, each one corresponding to a different draw of the training dataset $\mathcal{D}_n$. As the sample size n is increased, notice how the adversarial risk of our estimator approaches the Bayes-optimal adversarial risk, for each level of the attack strength ϵ, in accordance to Theorem 4.2 and Corollary 4.1.

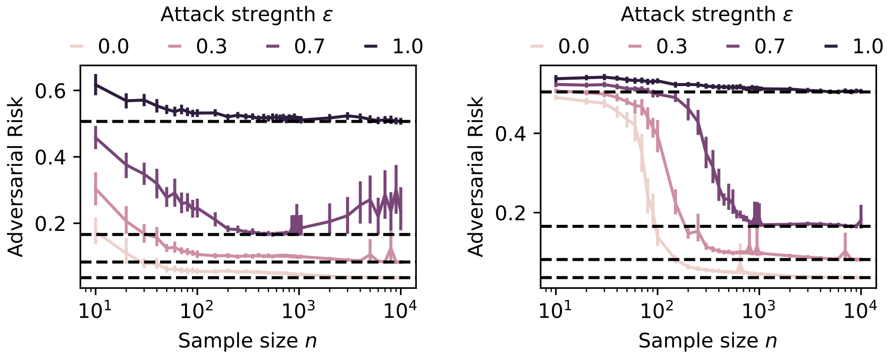

Figure 3. **Non-Gaussian Experiment.** As in Figure 2, error-bars correspond to 10 runs of computing our estimator $\widehat{f}_{n,\epsilon,h_n}$ (13), each one corresponding to a different draw of training dataset $\mathcal{D}_n$. **Left:** isotropic features (i.e $\Sigma = I_d$). **Right:** Non-isotropic features. As the sample size n is increased, notice how the adversarial risk of our estimator approaches the Bayes-optimal adversarial risk, for each level of the attack strength ϵ, in accordance to Theorem 4.2.

5.3. The Case of Regression

In the setting of linear regression, (Xing et al., 2021) studied Euclidean-norm attacks with general covariance matrices. They showed that the optimal robust model is a ridge regression whose ridge parameter depends implicitly on the strength of the attacks. (Javanmard et al., 2020) studied tradeoffs between ordinary and adversarial risk in linear regression, and computed exact Pareto optimal curves in the case of Euclidean-norm attacks on isotropic features. Their results show a tradeoff between ordinary and adversarial risk for adversarial training. (Javanmard & Mehrabi, 2021) also revisited this tradeoff for latent models and show that this tradeoff is mitigated when the data enjoys a low-dimensional structure.

Recently, the study of robustness in linear regression for general norms and feature covariance matrices has been initiated in (Scetbon & Dohmatob, 2023; Dohmatob & Scetbon, 2023) in a student-teach setup.

Finally, let us mention (Hassani & Javanmard, 2022) which established tradeoffs between accuracy and robustness to Euclidean-norm attacks on two-layer neural networks in random features regime. (Dohmatob & Bietti, 2022) extended this analysis to other learning regimes including stochastic gradient descent (SGD) and the so-called ”lazy training” regime (Chizat et al., 2019).

6. Empirical Validation

In this section, we present some empirical verification of our theoretical results. All experiments were run on a single modern CPU laptop.

Gaussian Experiment. Consider the distribution

$$\mathbb{P}(y = \pm 1) = 1/2,\; x \mid y \sim N(y\mu, \Sigma), \tag{24}$$

where $\mu \in \mathbb{R}^d$ with $\|\mu\|_\Sigma = 1$ is a fixed vector and the $d \times d$ positive-definite matrix Σ is the feature covariance matrix. We try both $\Sigma = I_d$ and $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, with $\sigma_k^2 = k^{-2}$ for all $k \in [d]$. This data model corresponds to the setting of Schmidt et al. (2018); Dobriban & Wager (2018); Dan et al. (2020), and the Bayes-optimal robust linear classifier is known analytically. Note that for this problem, it is well-known (Bhagoji et al., 2019; Dobriban et al., 2020) that the optimal value of the adversarial risk is attained over linear models and is explicitly given by

$$E_\epsilon^{opt} = \Phi(-(1 - \epsilon/\|\mu\|_2)_+ \cdot \mathrm{SNR}), \tag{25}$$

where $\mathrm{SNR} := \|\mu\|_2/\sigma_d$ is a measure of signal-to-noise ratio. We set the input-dimension to $d = 20$ for this experiment. For each value of sample size $n \in \{100, 200, \ldots, 1000, 2000, 3000, \ldots, 10000\}$, we generate a dataset $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of n iid samples, and then compute the estimator $\widehat{f}_{n,\epsilon,h_n}$ described in Section 3.3, where h_n is the bandwidth parameter, taken as $h_n = \sqrt{(d/n)\log n}$, in accordance with the choice in Corollary 4.1. We consider Euclidean-norm attacks of strength ϵ ranging in $\{0.1, 0.2, \ldots, 0.9, 1\}$.

The results of this experiment are shown in Figure 2.

Non-Gaussian Experiment. Next we turn from the Gaussian mixture setting and consider a setup where

$$z \mid y \sim N(y\mu, I_d),\ x = \max(z, 0) \text{ (entry-wise)}. \tag{26}$$

This setup captures the structure of a two-layer neural network with ReLU activation in the so-called random features regime where only the parameters of the output layer are learned. The rest of the experimental setup is as in the previous experiment (the ”Gaussian Experiment”).

The results of this experiment are shown in Figure 3.

Comparison with (Bao et al., 2020). It should be noted that (Bao et al., 2020) doesn't attempt any analysis of consistency, and therefore contains no regret analysis (rates). In particular, minimizing their proposed surrogate losses (sigmoid, etc.) don't provably converge to Bayes-optimum adversarial risk. In contrast, our work principally explores the problem of consistency and we show that for a class of non-convex smooth loss function (e.g sigmoid, Gaussian, etc.), the resulting estimator is consistent w.r.t adversarial 0/1 risk; we also establish an explicit regret of order $\widetilde{O}(1/\sqrt{n})$ rate of convergence. A key idea in our work is to not consider static surrogate losses as in (Bao et al., 2020), but to adapt a smoothing bandwidth $h = h_n$ so that it vanishes with sample size n at a precise rate (20).

Nothwithstanding, Figure 4 (**Right**) shows the result of a small experiment comparing our proposed estimator with that proposed in (Bao et al., 2020). The dataset, is a 2D Gaussian dataset given by $x \mid y \sim N(y\mu, \Sigma)$ and $\mathbb{P}(y = \pm 1) = 1/2$, where $\Sigma = \operatorname{diag}(\sigma_1^2, \sigma_2^2)$, with $\sigma_1 = 1$, $\sigma_2 = 0.1$. This is a modification of the dataset ”twonorm” described in Section 10 of (Bao et al., 2020) where they had $\sigma_1 = \sigma_2 = 0.1$ (left plot). For a smoothing bandwidth parameter, we use the surrogate loss function $\theta_{\epsilon,h}(u) = Q((u - \epsilon)/h)$, with $Q = Q_S$. The case $h = 1$ corresponds to the the estimator proposed in (Bao et al., 2020), while $h = h_n$ as in (20) corresponds to the estimator proposed in our work. As we can see (right plot), the estimator from (Bao et al., 2020) fails to be consistent here, while our adaptive estimator (left plot) is consistent, as predicted by Theorem 4.2. Similar results are observed with the choice $Q = Q_G$.

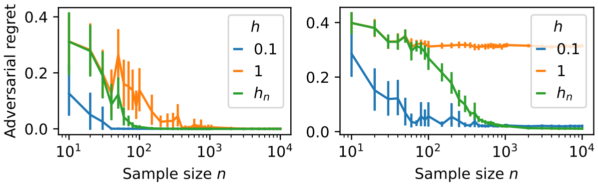

Figure 4. Inconsistency of non-adaptive smoothing bandwidth. **Left.** Gaussian mixture problem with $\sigma_1 = \sigma_2 = 0.1$. **Right.** $\sigma_1 = 1$, $\sigma_2 = 0.1$. Non-adaptive bandwidth $h = 1$ corresponds to (Bao et al., 2020); unlike our proposed estimator (adaptive bandwidth h_n), it fails to achieve the optimal adversarial risk.

7. Conclusion

In this work, we have considered the problem of consistent adversarially robust classification in a setting where no parametric or well-separatedness assumptions are made, in contrast with previous works We have proposed a estimator which simultaneously enjoys low-computational complexity and statistical guarantees in the form of quantitative rates of convergence to the Bayes-optimal adversarial risk. Our estimator is based on smoothing the adversarial 0/1 loss, with a smoothing bandwidth which effectively adapts to the sample complexity and input-dimension.

Non-linear Models. Our work can be extended in a number of directions. Notably, our work has focused on linear models. Even though our theoretical results developed here are already very interesting in this regime, analyzing neural networks (at least its so-called linearized and kernel regimes) would be an interesting direction to pursue.

Statistics vs Optimization. Finally, the theoretical analysis in this work has focused on the statistical properties of the proposed estimator. The optimization related aspects of computing our proposed estimator (provable convergence to global optimum instead of just stationary points) will be pursued in a future extension of our work.

Impact Statement

As we usher in the era of large language models (LLMs) and ChatGPT, the challenge of adversarial examples gains unprecedented importance, magnified by the widespread distribution of unsanitized and untraceable content and models across the internet. Our research delves into theoretical and algorithmic analyses of robust regression within a non-parametric framework, introducing the first estimator proven to be optimally robust in such a complex environment.

This paper aims to contribute significantly to the advancement of Machine Learning, meticulously examining the societal impacts of our findings. We are confident that our work, which carefully navigates the intricacies of adversarial robustness, offers exclusively positive implications for the field and society at large.

References

Absil, P.-A., Baker, C. G., and Gallivan, K. A. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7(3), Jul 2007.

Awasthi, P., Frank, N., Mao, A., Mohri, M., and Zhong, Y. Calibration and consistency of adversarial surrogate losses. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9804–9815. Curran Associates, Inc., 2021.

Bao, H., Scott, C., and Sugiyama, M. Calibrated surrogate losses for adversarially robust classification. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 408–451. PMLR, 09–12 Jul 2020.

Bartlett, P., Jordan, M., and McAuliffe, J. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 02 2006. doi: 10.1198/016214505000000907.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local Rademacher complexities. *The Annals of Statistics*, 33 (4):1497 – 1537, 2005.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport, 2019.

Bhattacharjee, R. and Chaudhuri, K. When are nonparametric methods robust? In *ICML*, pp. 832–841. PMLR, 2020.

Bhattacharjee, R., Jha, S., and Chaudhuri, K. Sample complexity of robust linear classification on separated data. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR, 2021.

Bietti, A. and Mairal, J. Group invariance, stability to deformations, and complexity of deep convolutional representations. *Journal of Machine Learning Research*, 20 (25):1–49, 2019.

Boumal, N., Mishra, B., Absil, P.-A., and Sepulchre, R. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42): 1455–1459, 2014.

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 2018.

Carlini, N., Nasr, M., Choquette-Choo, C. A., Jagielski, M., Gao, I., Awadalla, A., Koh, P. W., Ippolito, D., Lee, K., Tramèr, F., and Schmidt, L. Are aligned neural networks adversarially aligned? *ArXiv*, abs/2306.15447, 2023.

Chatterji, N. S., Long, P. M., and Bartlett, P. L. The interplay between implicit bias and benign overfitting in two-layer linear networks. *Journal of Machine Learning Research*, 23(263):1–48, 2022.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. 2019.

Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guaratees for adversarially robust Gaussian classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2345–2355. PMLR, 13–18 Jul 2020.

Diakonikolas, I., Kane, D., and Manurangsi, P. Nearly tight bounds for robust proper learning of halfspaces with a margin. In *Advances in Neural Information Processing Systems*, 2019.

Dobriban, E. and Wager, S. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247 – 279, 2018.

Dobriban, E., Hassani, H., Hong, D., and Robey, A. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

Dohmatob, E. and Bietti, A. On the (non-)robustness of two-layer neural networks in different learning regimes, 2022.

Dohmatob, E. and Scetbon, M. Robust linear regression: Phase-transitions and precise tradeoffs for general norms. *ArXiv*, abs/2308.00556, 2023.

Feige, U. Relations between average case complexity and approximation complexity. STOC. Association for Computing Machinery, 2002.

Hassani, H. and Javanmard, A. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *arXiv preprint arXiv:2201.05149*, 2022.

Javanmard, A. and Mehrabi, M. Adversarial robustness for latent models: Revisiting the robust-standard accuracies tradeoff. *arXiv preprint arXiv:2110.11950*, 2021.

Javanmard, A., Soltanolkotabi, M., and Hassani, H. Precise tradeoffs in adversarial training for linear regression. In *COLT*, 2020.

Keshet, J., McAllester, D. A., and Hazan, T. Pac-bayesian approach for minimization of phoneme error rate. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2224–2227, 2011.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. 2018.

Massart, P. and Nédélec, É. Risk bounds for statistical learning. *The Annals of Statistics*, 34, 2006.

Rudelson, M. and Vershynin, R. Small Ball Probabilities for Linear Images of High-Dimensional Distributions. *International Mathematics Research Notices*, 2015(19): 9594–9617, 12 2014.

Sato, H. Riemannian optimization on unit sphere with p-norm and its applications. *Computational Optimization and Applications*, 2023.

Scetbon, M. and Dohmatob, E. Robust linear regression: Gradient-descent, early-stopping, and beyond. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *CoRR*, abs/1804.11285, 2018.

Su, J., Vargas, D. V., and Sakurai, K. One pixel attack for fooling deep neural networks. *CoRR*, abs/1710.08864, 2017.

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.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, volume abs/1805.12152, 2019.

van der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Process: With Applications to Statistics*. 1996.

Vapnik, V. N. *The nature of statistical learning theory.* New York, NY: Springer, 2nd ed. edition, 2000. ISBN 0-387-98780-0.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2), 1971.

Xing, Y., Zhang, R., and Cheng, G. Adversarially robust estimate and risk analysis in linear regression. In *AISTATS*, volume 130, 2021.

Supplementary Material

Consistent Adversarially Robust Linear Classification: Non-Parametric Setting

Contents

A An Oracle Bound: Proof of Theorem 4.1 11

B Controlling the Bias in Adversarial Risk Due to Smoothing 12

B.1 Proof of Theorem 4.2 14

C VC Dimension Computations 14

D Proof of Proposition 4.1 16

A. An Oracle Bound: Proof of Theorem 4.1

Recall the definition of adversarial risk $E_\epsilon(f)$ and excess adversarial risk $R_\epsilon(f)$ from Section 2.1

Theorem 4.1. *Let Assumption 4.1 be in order. In the limit (16), if $(h_n)_n$ is a sequence of bandwidths tending to 0, then w.h.p over training data $\mathcal{D}_n$, it holds that*

$$\sup_{\varepsilon \geq 0} R_\varepsilon(\widehat{f}_{n,\varepsilon,h_n}) = O\left(\lambda(h_n) + \sqrt{\frac{d \log n}{n}}\right) \overset{n\to\infty}{\longrightarrow} 0, \tag{19}$$

where $\lambda(h_n) := \eta(h_n\sqrt{\log 1/h_n}) + Q(\sqrt{\log 1/h_n})$ and the function η is as in Assumption 4.1.

Proof. Let $\widehat{f} = \widehat{f}_{n,\epsilon,h}$ be a minimizer of $\widehat{E}_{n,\epsilon,h}$ over the linear function class $\mathcal{F}_{\text{lin}}$ and let $f_\star$ be a minimizer of E_ϵ over $\mathcal{F}_{\text{lin}}$. Then, for any $f \in \mathcal{F}_{\text{lin}}$, one has

$$\begin{aligned} 0 \leq E_\epsilon(\widehat{f}) - E_\epsilon(f_\star) &= E_\epsilon(\widehat{f}) - \widehat{E}_{n,\epsilon,h}(\widehat{f}) + \underbrace{\widehat{E}_{n,\epsilon,h}(\widehat{f}) - \widehat{E}_{n,\epsilon,h}(f_\star)}_{\leq 0} + \widehat{E}_{n,\epsilon,h}(f_\star) - E_\epsilon(f_\star) \\ &\leq E_\epsilon(\widehat{f}) - \widehat{E}_{n,\epsilon,h}(\widehat{f}) + \widehat{E}_{n,\epsilon,h}(f_\star) - E_\epsilon(f_\star) \leq 2\|E_\epsilon - \widehat{E}_{n,\epsilon,h}\|_{\mathcal{F}_{\text{lin}}}. \end{aligned} \tag{27}$$

We deduce that

$$R_\epsilon(\widehat{f}) := E_\epsilon(\widehat{f}) - \inf_{f \in \mathcal{F}_{\text{lin}}} E_\epsilon(f) \leq 2\|E_\epsilon - \widehat{E}_{n,\epsilon,h}\|_{\mathcal{F}_{\text{lin}}}. \tag{28}$$

On the other hand, we can further decompose

$$\begin{aligned} \|E_\epsilon - \widehat{E}_{n,\epsilon,h}\|_{\mathcal{F}_{\text{lin}}} &= \|E_\epsilon - \widehat{E}_{n,\epsilon} + \widehat{E}_{n,\epsilon} - \widehat{E}_{n,\epsilon,h}\|_{\mathcal{F}_{\text{lin}}} \\ &\leq \underbrace{\|E_\epsilon - \widehat{E}_{n,\epsilon}\|_{\mathcal{F}_{\text{lin}}}}_{\text{gap due to ERM on margin loss}} + \underbrace{\|\widehat{E}_{n,\epsilon} - \widehat{E}_{n,\epsilon,h}\|_{\mathcal{F}_{\text{lin}}}}_{\text{gap due to smoothing of margin loss}}, \end{aligned} \tag{29}$$

where $\|U - V\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |U(f) - V(f)|$.

Combining (28) and (29) from the sketch of the proof in the main text, we get

$$E_\epsilon(\widehat{f}_{n,\epsilon,h}) - \inf_{f \in \mathcal{F}_{\text{lin}}} R_\epsilon(f) \le \underbrace{\|E_\epsilon - \widehat{E}_{n,\epsilon}\|_{\mathcal{F}_{\text{lin}}}}_{\text{usual ERM error}} + \underbrace{\|\widehat{E}_{n,\epsilon} - \widehat{E}_{n,\epsilon,h}\|_{\mathcal{F}_{\text{lin}}}}_{\text{empirical error due to smoothing}} \tag{30}$$

We now bound the different terms in the RHS of (30) separately.

Bounding the first term in (30). Consider the binary function class $\mathcal{G}_\epsilon$ on $\mathbb{R}^d \times \{\pm 1\}$ given by

$$\mathcal{G}_\epsilon := \{(x, y) \mapsto \theta_\epsilon(-yf(x)) \mid f \in \mathcal{F}_{\text{lin}}\}. \tag{31}$$

Thanks to Lemma C.2, we know that the VC dimension of $\mathcal{G}_\epsilon$ is at most $4d$. Then, applying the Vapnik-Chervonenkis (VC) theory uniform convergence (Vapnik & Chervonenkis, 1971) gives

$$\|E_\epsilon - \widehat{E}_{n,\epsilon}\|_{\mathcal{F}_{\text{lin}}} \lesssim \sqrt{\frac{d \log n}{n}} \text{ w.h.p over training data } \mathcal{D}_n. \tag{32}$$

Bounding the second term in (30). In Proposition B.1, it is established that for small h and large n, the following orcale bound holds w.h.p,

$$\|\widehat{E}_{n,\epsilon,h} - \widehat{E}_{n,\epsilon}\|_{\mathcal{F}_{\text{lin}}} \lesssim \overline{\lambda}(h) + \sqrt{\frac{d \log n}{n}}. \tag{33}$$

Combining with (30) and (32), we obtain that w.h.p,

$$E_\epsilon(\widehat{f}) - \inf_{f \in \mathcal{F}_{\text{lin}}} E_\epsilon(f) \lesssim \lambda(h) + \sqrt{\frac{d \log n}{n}}, \tag{34}$$

which completes the proof. □

B. Controlling the Bias in Adversarial Risk Due to Smoothing

We now prove inequality (33), which was instrumental in the proof of Theorem 4.1.

Proposition B.1. *With the same conditions and notations as in Theorem 4.1, it holds w.h.p over the training data $\mathcal{D}_n$ that* $\|\widehat{E}_{n,\epsilon,h} - \widehat{E}_{n,\epsilon}\|_{\mathcal{F}_{\text{lin}}} = O\left(\lambda(h_n) + \sqrt{\frac{d \log n}{n}}\right)$.

Proof. Vary $h = h_n$ and $t = t_n > 0$ such that

$$h \to 0^+,\ t \to \infty,\ ht \to 0^+. \tag{35}$$

For each $f \in \mathcal{F}_{\text{lin}}$ and $i \in [n]$, consider the binary random variable

$$z_{t,i}(f) := 1_{|y_i f(x_i) - \epsilon| \ge ht} = \begin{cases} 1, & \text{if } |y_i f(x_i) - \epsilon| \ge ht, \\ 0, & \text{else.} \end{cases} \tag{36}$$

Note that the sequence $z_{t,1}(f), \ldots, z_{t,n}(f)$ is an iid sequence of Bernoulli random variables with mean $\alpha_t(f) \in [0, 1]$ given by

$$\alpha_t(f) := \mathbb{E}\,[z_{t,i}(f)] = \mathbb{P}(|y_1 f(x_1) - \epsilon| \ge ht). \tag{37}$$

Recall the function $\theta_\epsilon(u) := 1_{u \le \epsilon}$ and $\theta_{\epsilon,h}(u) := Q((u - \epsilon)/h)$ for any $u \in \mathbb{R}$, introduced in (7) and (9) respectively, and observe that if $|u - \epsilon| \ge ht$, then

$$\begin{aligned} |\theta_{\epsilon,h}(u) - \theta_\epsilon(u)| &\le \begin{cases} |1 - Q((u - \epsilon)/h)| = Q(-(u - \epsilon)/h) = Q(|u - \epsilon|/h), & \text{if } u - \epsilon \le -ht, \\ |Q((u - \epsilon)/h) - 0| = Q((u - \epsilon)/h) = Q(|u - \epsilon|/h), & \text{if } u - \epsilon \ge ht \end{cases} \\ &\le Q(t), \end{aligned} \tag{38}$$

Where have used the property that Q is non-increasing and $Q(-a) \ge 1/2$ for all $a \ge 0$. We deduce that: every f and i,

$$|\theta_{\epsilon,h}(y_i f(x_i)) - \theta_\epsilon(y_i f(x_i))| \le 1 - (1 - Q(t)) z_{t,i}(f). \tag{39}$$

Now, consider the $\{0,1\}$-valued function class on $\mathbb{R}^d \times \{\pm 1\}$ given by

$$\mathcal{H}_t := \{(x,y) \mapsto 1_{|yf(x)-\epsilon| \ge ht} \mid f \in \mathcal{F}_{\text{lin}}\}. \tag{40}$$

Thanks to Lemma C.1, we know that the VC dimension of $\mathcal{H}_t$ is at most $4d$, and so by (Vapnik & Chervonenkis, 1971), it holds w.h.p over training data $\mathcal{D}_n$ that[1]

$$\sup_{f \in \mathcal{F}_{\text{lin}}} \left| \frac{1}{n} \sum_{i=1}^n z_{t,i}(f) - \alpha_t(f) \right| \lesssim \sqrt{\frac{d \log n}{n}}, \tag{41}$$

where all the hidden constants are independent of d, n, h, and t. Combining with (39), we deduce that, for large n and small h, the following chain of inequalities holds w.h.p over training data $\mathcal{D}_n$

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n |\theta_{\epsilon,h}(y_i f(x_i)) - \theta_\epsilon(y_i f(x_i))| &\le 1 - (1 - Q(t)) \cdot \frac{1}{n}\sum_{i=1}^n z_{t,i}(f) \\ &= \frac{1}{n}\sum_{i=1}^n (1 - z_{t,i}(f)) + \frac{1}{n}\sum_{i=1}^n Q(t) z_{t,i}(f) \\ &\lesssim 1 - \alpha_t(f) + \sqrt{\frac{d\log n}{n}} + Q(t)\alpha_t(f) + \sqrt{\frac{d \log n}{n}} \\ &\lesssim \beta_t(f) + \sqrt{\frac{d\log n}{n}}, \end{aligned} \tag{42}$$

where $\beta_t(f)$ is defined by

$$\beta_t(f) := 1 - \alpha_t(f) + Q(t)\alpha_t(f). \tag{43}$$

Taking $t = \sqrt{k \log 1/h}$ and applying Lemma B.1 ensures that, for any fixed positive k and for $h \to 0^+$,

$$\beta_t(f) \lesssim \eta(\sqrt{k}h\sqrt{\log 1/h}) + Q(\sqrt{k \log 1/h}),\ \text{for all } f \in \mathcal{F}_{\text{lin}}. \tag{44}$$

We deduce that, for large n and small h, it holds w.h.p over the training data $\mathcal{D}_n$, that

$$\begin{aligned} \|\widehat{E}_{n,\epsilon,h} - \widehat{E}_{n,\epsilon}\|_{\mathcal{F}_{\text{lin}}} &= \sup_{f \in \mathcal{F}_{\text{lin}}} \frac{1}{n}\sum_{i=1}^n |\theta_{\epsilon,h}(y_i f(x_i)) - \theta_\epsilon(y_i f(x_i)))| \\ &\lesssim \eta(h\sqrt{\log 1/h}) + Q(\sqrt{k\log 1/h}) + \sqrt{\frac{d \log n}{n}}, \end{aligned} \tag{45}$$

which completes the proof. □

Lemma B.1. *Suppose the function class $\mathcal{F}_{\text{lin}}$ is nice. Then, for any fixed positive k and in the limit $h \to 0^+$ with $t = \sqrt{3 \log 1/h}$, it holds that $\|\beta_t\|_{\mathcal{F}_{\text{lin}}} \lesssim \eta(\sqrt{k}h\sqrt{\log 1/h}) + Q(\sqrt{k\log 1/h})$.*

Proof. For small $ht > 0$ with $ht \to 0^+$, one computes

$$\begin{aligned} \beta_t(f) &= 1 - \alpha_t(f) + Q(t)\alpha_t(f) \\ &\le 1 - \alpha_t(f) + Q(t) \\ &\le \eta(ht) + Q(t), \end{aligned} \tag{46}$$

where,

[1] Note that the $\widetilde{O}(1/\sqrt{n})$ rate cannot be improved to $o(1/\sqrt{n})$ even for a single $f \in \mathcal{F}_{\text{lin}}$ since this would contradict the CLT.

- The first inequality is a basic Gaussian tail-bound, and the fact that $\alpha_t(f) \le 1$.
- The second inequality invokes Assumption 4.1 to get $1 - \alpha_t(f) = \mathbb{P}(|yf(x) - \epsilon| \le ht) \le \eta(ht)$ for $(x, y) \sim P$.

Finally, choosing $t = \sqrt{k \log 1/h}$ with $h \to 0^+$ ensures that $ht = \sqrt{k}\lambda(h) \asymp \lambda(h)$, and so $\beta_t(f) \le \eta(\sqrt{k}h\sqrt{\log 1/h}) + Q(\sqrt{k \log 1/h})$ as claimed. □

B.1. Proof of Theorem 4.2

Let $\widehat{f} = \widehat{f}_{n,\varepsilon,h}$ be a minimizer of $\widehat{E}_{n,\varepsilon,h}$ over the function class $\mathcal{F}_{\text{lin}}$. For any $f \in \mathcal{F}_{\text{lin}}$, one computes

$$\begin{aligned}R_\varepsilon(\widehat{f}) \le E_\varepsilon(\widehat{f}) - E_\varepsilon(f) &= E_\varepsilon(\widehat{f}) - \widehat{E}_{n,\varepsilon,h}(\widehat{f}) + \underbrace{\widehat{E}_{n,\varepsilon,h}(\widehat{f}) - \widehat{E}_{n,\varepsilon,h}(f)}_{\le 0} + \widehat{E}_{n,\varepsilon,h}(f) - E_\varepsilon(f)\\
&\le E_\varepsilon(\widehat{f}) - \widehat{E}_{n,\varepsilon,h}(\widehat{f}) + \widehat{E}_{n,\varepsilon,h}(f) - E_\varepsilon(f)\\
&\le 2 \sup_{f \in \mathcal{F}_{\text{lin}}} |E_\varepsilon(f) - \widehat{E}_{n,\varepsilon,h}(f)|.\end{aligned}$$

Thus, defining $\|E_\epsilon - \widehat{E}_{n,\epsilon,h}\|_{\mathcal{F}_{\text{lin}}} := \sup_{f \in \mathcal{F}_{\text{lin}}} |\widehat{E}_\epsilon(f) - \widehat{E}_{n,\epsilon,h}(f)|$, we deduce that

$$R_\varepsilon(\widehat{f}) \le 2\|E_\varepsilon - \widehat{E}_{n,\varepsilon,h}\|_{\mathcal{F}_{\text{lin}}}. \tag{47}$$

On the other hand, we can further decompose

$$\|E_\varepsilon - \widehat{E}_{n,\varepsilon,h}\|_{\mathcal{F}_{\text{lin}}} = \|E_\varepsilon - \widehat{E}_{n,\varepsilon} + \widehat{E}_{n,\varepsilon} - \widehat{E}_{n,\varepsilon,h}\|_{\mathcal{F}_{\text{lin}}} \le \underbrace{\|E_\varepsilon - \widehat{E}_{n,\varepsilon}\|_{\mathcal{F}_{\text{lin}}}}_{\text{regret due to ERM}} + \underbrace{\|\widehat{E}_{n,\varepsilon} - \widehat{E}_{n,\varepsilon,h}\|_{\mathcal{F}_{\text{lin}}}}_{\text{regret due to smoothing}}. \tag{48}$$

The first term in the above decomposition is controlled using classical *uniform convergence* arguments (Vapnik & Chervonenkis, 1971; Vapnik, 2000); it is of order $O(\sqrt{(d/n)\log n})$ since the VC pseudo-dimension of $\mathcal{F}_{\text{lin}}$ is at most d. Thanks to Proposition B.1, we know that the second term is of order $\lambda(h) + \sqrt{(d/n)\log n}$ w.h.p. Combining with (47) then gives (4.1).

Finally, taking the bandwidth h_n as in (20) balances both terms in (4.1) and we obtain (21). □

C. VC Dimension Computations

Let $\mathcal{F}$ be a real-valued function class on an abstract set $\mathcal{X}$.

Lemma C.1. *Let $\alpha, \beta \in \mathbb{R}$ with $\beta \ge 0$, and define a binary-valued function class $\mathcal{H} := \{(x, y) \mapsto 1_{|yf(x)-\alpha|\ge\beta} \mid f \in \mathcal{F}\}$. Then, we have the upper bound*

$$\mathrm{VCdim}(\mathcal{H}) \le 4 \cdot \mathrm{VCpdim}(\mathcal{F}). \tag{49}$$

In particular, if $\mathcal{F} = \mathcal{F}$ is the linear function class on $\mathbb{R}^d$, then $\mathrm{VCdim}(\mathcal{H}) \le 4d$.

Proof. First observe that, in terms of VC dimension, the function class $\mathcal{H}$ can be equivalently defined as a collection of sets like so

$$\mathcal{H} = \{T_{\alpha,\beta}(f) \mid f \in \mathcal{F}\}, \text{ where } T_{\alpha,\beta}(f) := \{(x, y) \in \mathcal{X} \times \{\pm 1\} \mid |yf(x) - \alpha| \ge \beta\}. \tag{50}$$

Now, for any $t \in \mathbb{R}$ and $f \in \mathcal{F}$, define $S_t(f) := \{(x, y) \in \mathcal{X} \times \{\pm 1\} \mid yf(x) \le t\}$, and consider the function class $S_t(\mathcal{F})$ on $\mathcal{X} \times \{\pm 1\}$ given by

$$S_t(\mathcal{F}) := \{S_t(f) \mid f \in \mathcal{F}\}. \tag{51}$$

Since $|yf(x) - \alpha| \ge \beta$ iff $yf(x) \le \alpha - \beta$ or $-yf(x) \le -\alpha - \beta$, it is clear that $(x, y) \in T_{\alpha,\beta}(f)$ iff $(x, y) \in S_{\alpha-\beta}(f) \cup S_{-\alpha-\beta}(f)$, and so $T_{\alpha,\beta}(f) = S_{\alpha-\beta}(f) \cup S_{-\alpha-\beta}(f)$. It follows that,

$$\mathcal{H} \subseteq \{A \cup B \mid A \in S_{\alpha-\beta}(\mathcal{F}),\ B \in S_{-\alpha-\beta}(-\mathcal{F})\}, \tag{52}$$

where $-\mathcal{F} := \{-f \mid f \in \mathcal{F}\}$.

Thanks to van der Vaart & Wellner (1996, Lemma 2.6.17, part (iii)), we deduce that

$$\begin{aligned}\mathrm{VCdim}(\mathcal{H}) &\le \mathrm{VCdim}(S_{\alpha-\beta}(\mathcal{F})) + \mathrm{VCdim}(S_{-\alpha-\beta}(-\mathcal{F}))\\ &= \mathrm{VCdim}(S_{\alpha-\beta}(\mathcal{F})) + \mathrm{VCdim}(S_{-\alpha-\beta}(-\mathcal{F}))\\ &= \mathrm{VCdim}(S_{\alpha-\beta}(\mathcal{F})) + \mathrm{VCdim}(\neg S_{\alpha+\beta}(\mathcal{F}))\\ &= 2\cdot \mathrm{VCdim}(S_{\alpha-\beta}(\mathcal{F})).\end{aligned} \tag{53}$$

where we have used the fact that $\mathrm{VCdim}(\{S' \mid S \in \mathcal{S}\}) = \mathrm{VCdim}(\mathcal{S})$ for any collection of sets $\mathcal{S}$.

Finally, If thanks to Lemma C.2, we have

$$\mathrm{VCdim}(S_t(\mathcal{F})) \le 2\cdot \mathrm{VCpdim}(\mathcal{F}). \tag{54}$$

Combining with (53) then gives $\mathrm{VCdim}(\mathcal{H}) \le 4\mathrm{VCpdim}(\mathcal{F})$.

In particular, for the half-space function class $\mathcal{F}$ on $\mathbb{R}^d$, we have $\mathrm{VCpdim}(\mathcal{F}) \le d$, which then gives $\mathrm{VCdim}(\mathcal{H}(\mathcal{F})) \le 4d$, and the proof is complete. □

Lemma C.2. *Let $\epsilon \in \mathbb{R}$ be fixed and define a collection of subsets of $\mathcal{X} \times \{\pm 1\}$ by*

$$\mathcal{H} := \{\Lambda(f) \mid f \in \mathcal{F}\},\ \textit{where}\ \Lambda(f) := \{(x,y) \in \mathcal{X} \times \{\pm 1\} \mid yf(x) \le \epsilon\}. \tag{55}$$

Then, $\mathrm{VCdim}(\mathcal{H}) \le 2\cdot \mathrm{VCpdim}(\mathcal{F})$.

Proof. For any $f \in \mathcal{F}$ and $y \in \mathcal{Y}$, define $f_y : \mathcal{X} \to \mathbb{R}$ by $f_y(x) := y(x) - \epsilon + y$, and let $\mathcal{F}_y := \{f_y \mid f \in \mathcal{F}\}$. Thus, $\mathcal{F}_y$ is an affine translation of $\mathcal{F}$. Then, one computes

$$\begin{aligned}\Lambda(f) &= \cup_{y\in\{\pm1\}}\{(x,y) \mid x \in \mathcal{X},\ yf(x) \le \epsilon\}\\ &= \cup_{y\in\{\pm1\}}\{(x,y) \mid x \in \mathcal{X},\ f_y(x) \le y\}\\ &= \cup_{y\in\{\pm1\}}\{(x,t) \in \mathcal{X}\times\mathbb{R} \mid f_y(x) \le t\} \cap (\mathcal{X} \times \{y\})\\ &= \cup_{y\in\{\pm1\}}\mathrm{SG}(\mathcal{F}_y) \cap (\mathcal{X}\times\{y\})\\ &= \cup_{y\in\{\pm1\}}\Lambda_y(f),\end{aligned} \tag{56}$$

where $\Lambda_y(f) := \mathrm{SG}(\mathcal{F}_y) \cap (\mathcal{X}\times\{y\})$. For every y, let $\Lambda_y(\mathcal{F}) := \{\Lambda_y(f) \mid f \in \mathcal{F}\}$. We deduce from the previous display that,

$$\mathcal{H} = \{\cup_{y\in\{\pm1\}}\Lambda_y(f) \mid f \in \mathcal{F}\} \subseteq \{A \cup B \mid A \in \Lambda_+(\mathcal{F}),\ B \in \Lambda_-(\mathcal{F})\}, \tag{57}$$

and so, thanks to (**?**)Lemma 2.6.17, part (iii))]vaartwellner96book, we obtain

$$\mathrm{VCdim}(\mathcal{H}) \le \sum_{y\in\{\pm1\}} \mathrm{VCdim}(\Lambda_y(\mathcal{F})). \tag{58}$$

Now, observe that $\Lambda_y(\mathcal{F}) = \{A \cap (\mathcal{X}\times\{y\}) \mid A \in \mathrm{SG}(\mathcal{F}_y)\}$, and so by (**?**)Lemma 2.6.17, part (ii))]vaartwellner96book, we get

$$\mathrm{VCdim}(\Lambda_y(\mathcal{F})) \le \mathrm{VCdim}(\mathrm{SG}(\mathcal{F}_y)). \tag{59}$$

Now, since the transformation $f \mapsto f_y$ is obviously a bijection between $\mathcal{F}$ and $\mathcal{F}_y$, the VC pseudo-dimensions of $\mathcal{F}$ and $\mathcal{F}_y$ are equal. We deduce from (59) that

$$\mathrm{VCdim}(\Lambda_y(\mathcal{F})) \le \mathrm{VCdim}(\mathrm{SG}(\mathcal{F}_y)) = \mathrm{VCdim}(\mathrm{SG}(\mathcal{F})) =: \mathrm{VCpdim}(\mathcal{F}). \tag{60}$$

Combining with (58) gives the claimed result. □

D. Proof of Proposition 4.1

Proof. Under Condition 4.1, it is clear that for $(x, y) \sim P$, the random vector $z := yx$ has density. Indeed, if $N \subseteq \mathbb{R}^d$ is a null-set w.r.t the Lebesgue measure on $\mathbb{R}^d$, then one computes

$$\mathbb{P}(z \in N) = \mathbb{P}(y = 1)\mathbb{P}(x \in N \mid y = 1) + \mathbb{P}(y = -1)\mathbb{P}(-x \in N \mid y = -1) = 0 + 0 = 0, \tag{61}$$

since $-N := \{-v \mid v \in N\}$ is also a null-set. It follows that every continuous transformation of z also has density, and thus has a continuous CDF. In particular, for every $w \in \mathbb{R}^d$, the ”margin” random variable $m(x, y; w) := yx^\top w = z^\top w$ has a continuous CDF.

Now, let R be a large positive number. By the preceding argument and the compactness of the set

$$S_R := \{(w, u) \in \mathbb{R}^{d+1} \mid \|w\|_\star = 1,\ |u| \le R\},$$

the function $\delta \mapsto \sup_{(w,u)\in S_R} \mathbb{P}(|z^\top w - u| \le \delta)$ is continuous on $(0, \infty)$, and so

$$\begin{aligned}\lim_{\delta\to 0^+} \sup_{(w,u)\in S_R} \mathbb{P}(|z^\top w - u| \le \delta) &= \sup_{(w,u)\in S_R} \lim_{\delta\to 0^+} \mathbb{P}(|z^\top w - u| \le \delta) \\ &= \sup_{(w,u)\in S_R} \mathbb{P}(z^\top w = u) = 0.\end{aligned} \tag{62}$$

On the other hand, if $|u| > R$, then $|z^\top w - u| \ge |u| - |z^\top w| > R - |z^\top w|$, and so

$$\begin{aligned}\sup_{\|w\|_\star=1,\ |u|>R} \mathbb{P}(|z^\top w - u| \le \delta) &\le \sup_{\|w\|_\star=1} \mathbb{P}(R - |z^\top w| \le \delta) \le \sup_{\|w\|_\star=1} \mathbb{P}(|z^\top w| > R - \delta) \\ &\le \mathbb{P}(\sup_{\|w\|_\star=1} |z^\top w| > R - \delta) = \mathbb{P}(\|z\| > R - \delta) \\ &= \mathbb{P}(\|z\| > R - \delta) \overset{\delta\to 0^+}{\longrightarrow} \mathbb{P}(\|z\| > R) \overset{R\to\infty}{\longrightarrow} 0.\end{aligned} \tag{63}$$

In the last two steps, we have used the fact that the random variable $\|z\|$ has density since every norm on $\mathbb{R}^d$ is continuous. Combining (62) and (63) completes the proof of the claim. □