A Fundamental Accuracy–Robustness Trade-off in Regression and Classification

Anonymous authors
Paper under double-blind review

Abstract

We derive a fundamental trade-off between standard and adversarial risk in a rather general situation that formalizes the following simple intuition:

> *If no (nearly) optimal predictor is smooth, adversarial robustness comes at the cost of accuracy.*

As a concrete example, we evaluate the derived trade-off in regression with polynomial ridge functions under mild regularity conditions.

1 Introduction

The study of *adversarial robustness* of machine learning models is concerned with finding prediction schemes whose accuracy gracefully degrades when minute but adversarial perturbations are applied to the data. We focus on a simple prevalent mathematical abstraction of the problem that can be described as follows. Given a pair of random variables $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^k$ that represent a data sample and its corresponding label, as well as $\mathcal{F}$, a class of functions from $\mathbb{R}^d$ to $\mathbb{R}^k$, the goal is to find an $f \in \mathcal{F}$ that achieves a low adversarial risk $\mathbb{E}\left(\sup_{\Delta\colon \|\Delta\| \le \epsilon} \ell\left(f\left(X + \Delta\right), Y\right)\right)$ where $\ell\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_{\ge 0}$ is the loss function and $\|\cdot\|$ is a certain norm defined over $\mathbb{R}^d$.

Intuitively, if a prediction function $f \in \mathcal{F}$ is very non-smooth (e.g., it has a large *Lipschitz seminorm*), then we can expect its adversarial risk

$$R_\epsilon(f) = \mathbb{E}\left(\sup_{\Delta\colon \|\Delta\| \le \epsilon} \ell\left(f\left(X + \Delta\right), Y\right)\right),$$

to be much larger than its standard risk

$$R(f) = \mathbb{E}\left(\ell\left(f\left(X\right), Y\right)\right).$$

We formalize this intuition and show a simple fundamental trade-off between the standard and adversarial risks of any candidate function $f \in \mathcal{F}$ as quantified by certain notion of local smoothness of f. It is worth mentioning that the constraints on the intensity of the data perturbation are formulated using a norm merely for the sake of a simpler exposition; many other types of perturbation constraints can be addressed by straightforward adaptation of the presented arguments.

In the finite sample setting we only access n independent draws of (X, Y) which we denote by $(X_1, Y_1), \dots, (X_n, Y_n)$. Then, the empirical risk and its adversarial version are defined respectively by

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i \in [n]} \ell(f(X_i), Y_i),$$

and

$$\widehat{R}_{n,\epsilon}(f) = \frac{1}{n} \sum_{i \in [n]} \sup_{\Delta_i\colon \|\Delta_i\| \le \epsilon} \ell(f(X_i + \Delta_i), Y_i).$$

Empirical risk minimization (ERM) is the most common mechanism to construct a predictor f from n data samples that "generalizes" well in the sense that $R(f)$ is close to $\min_{f\in\mathcal{F}} R(f)$. Specifically, based on the premise that $R(\cdot)$ is nowhere much larger than $\widehat{R}_n(\cdot)$, ERM provides the predictor

$$\widehat{f}_n = \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}_n(f)\,.$$

Analogously, empirical adversarial risk minimization, also referred to as *adversarial training*, provides the predictor

$$\widehat{f}_{n,\epsilon} = \operatorname*{argmin}_{f\in\mathcal{F}} \widehat{R}_{n,\epsilon}(f)\,. \tag{1}$$

If $\widehat{R}_{n,\epsilon}(f)$ concentrates around $R_\epsilon(f)$ uniformly for all $f\in\mathcal{F}$, which can be shown, e.g., using a variety of tools from the theory of empirical processes (van der Vaart & Wellner, 2012) or PAC-Bayesian arguments (Catoni, 2007), under reasonable regularity conditions, then we can guarantee with high probability that $R_\epsilon(\widehat{f}_{n,\epsilon}) \le \min_{f\in\mathcal{F}} R_\epsilon(f) + o_n(1)$ where $o_n(1)$ is a term that vanishes to zero typically at rate $n^{-1/2}$. Therefore, the adversarial risk of $\widehat{f}_{n,\epsilon}$ is nearly-optimal and we only need to examine the gap between the standard risk of $\widehat{f}_{n,\epsilon}$ and the optimal standard risk $\min_{f\in\mathcal{F}} R(f)$. We do not pursue the finite sample scenario any further in this paper. Instead, we exclusively focus on a fundamental trade-off between the standard population risk $R(f)$ and its adversarial analog $R_\epsilon(f)$ that exist for any arbitrary predictor $f\in\mathcal{F}$ even if the access to the data distribution is not restricted by finite samples.

Related Work

A comprehensive review of the literature on adversarial robustness is beyond the scope of this work, but we summarize some of the results in this area that are most relevant for us. For a broader view of the literature interested readers are referred to (Carlini et al., 2019; Bai et al., 2021) and references therein.

A common theme in the literature is to analyze adversarial robustness in classification or regression problems assuming a curated data distribution (Fawzi et al., 2017; Schmidt et al., 2018; Tsipras et al., 2019; Dan et al., 2020; Javanmard et al., 2020; Dobriban et al., 2023; Javanmard & Mehrabi, 2024). Among the results that study the trade-offs between standard and adversarial risk, (Javanmard et al., 2020) considers the least squares linear regression with standard Gaussian covariates under an ℓ_2 adversarial perturbation. Leveraging the convex Gaussian min-max theorem (Thrampoulidis et al., 2015) they provide a precise (asymptotic) trade-off formula between $R_\epsilon(\widehat{f}_{n,\epsilon})$ and $R(\widehat{f}_{n,\epsilon})$ where $\widehat{f}_{n,\epsilon}$ is the linear function obtained by adversarial training on n samples as in (1). This trade-off formula approaches a fundamental limit for any estimator by increasing the considered "sampling ratio" n/d.

Adversarial ℓ_2 and ℓ_∞ robustness for classification of a mixture of two or three Gaussian distributions with colinear means and identical isotropic covariance matrices is analyzed in (Dobriban et al., 2023). In the mentioned setting, optimal and approximately optimal robust classifiers are derived, and it is shown that the trade-off between accuracy and robustness for any classifier deteriorates as the imbalance of the classes increase.

Binary classification with ℓ_p-adversaries for two particular low-dimensional manifolds is analyzed in (Javanmard & Mehrabi, 2024) where the covariates are modeled as $X = \varphi(WZ)$ with φ denoting a monotonic coordinatewise nonlinearity, $W\in\mathbb{R}^{d\times k}$ being a tall matrix, and $Z\in\mathbb{R}^k$ being the low-dimensional latent variable. The first model is a Gaussian mixture model where Y is a biased ±1-valued random variable, and conditioned on Y we have $Z\sim \mathrm{Normal}(\mathrm{Y}\mu,\mathrm{I})$ for a fixed $\mu\in\mathbb{R}^k$. The second model is a *generalized linear model* where $Z\sim\mathrm{Normal}(0,I)$ and $\mathbb{P}(Y=1\mid Z) = 1-\mathbb{P}(Y=-1\mid Z) = g(\beta^\intercal Z)$ for a monotonic "link function" $g\colon\mathbb{R}\to[0,1]$ and a fixed $\beta\in\mathbb{R}^k$. With $\sigma_{\min}(W)$ denoting the smallest singular value of W, it is shown in (Javanmard & Mehrabi, 2024) that if $\sigma_{\min}(W)$ dominates $\epsilon d^{1/2-1/p}$ as $d\to\infty$, then for both of the considered models the gap between the adversarial risk and the standard risk (i.e., the "boundary risk") of the optimal standard classifier vanishes asymptotically.

In contrast, we derive a simple yet fundamental trade-off between the robustness and accuracy, or more generally the adversarial and standard risks, that holds in rather general scenarios under minimal assumptions.

As an example, we quantify this trade-off for the problem of regression over polynomial ridge functions, which includes linear regression as a special case.

2 Problem Setup and the Main Result

The basic property that we will exploit is that there often exist functions $A : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ and $B : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ such that $A(u,u) = B(u,u) = 0$ for all $u \in \mathbb{R}^k$, and for every two pairs (u,v) and (u',v') in $\mathbb{R}^k \times \mathbb{R}^k$ we have

$$\ell(u,v) + \ell(u',v') + A(u,u') \geq B(v,v')\,, \tag{2}$$

and

$$\ell(u,v) + \ell(u',v') + A(v,v') \geq B(u,u')\,. \tag{3}$$

For the sake of concreteness, we distinguish three special scenarios in our notation. The first scenario is the *least-squares regression* where $k = 1$ and

$$\ell(u,v) = \ell_{\mathsf{LS}}(u,v) \stackrel{\text{def}}{=} (u-v)^2/2$$

for which, using the Cauchy-Schwarz inequality,

$$A(u,v) = A_{\mathsf{LS}}(u,v) \stackrel{\text{def}}{=} (u-v)^2/2$$

and

$$B(u,v) = B_{\mathsf{LS}}(u,v) \stackrel{\text{def}}{=} (u-v)^2/6$$

fulfill the conditions (2) and (3). The second scenario is the *multiclass classification* where the functions $f \in \mathcal{F}$ map their input to the unit simplex $\triangle^{k-1} \subset \mathbb{R}^k$ for some $k > 1$ and the response variable Y is some extreme point of $\triangle^{k-1}$, i.e., a canonical basis vector in $\mathbb{R}^k$. In this setting, the Kullback–Leibler (KL) divergence can be used as the loss function, i.e.,

$$\ell(u,v) = \ell_{\mathsf{KL}}(u,v) = \sum_{j=1}^{k} v_j \log \frac{v_j}{u_j}\,,$$

for which, by Pinsker's inequality (see, e.g., (Boucheron et al., 2013, Theorem 4.19)), the functions

$$A(u,v) = A_{\mathsf{KL}}(u,v) \stackrel{\text{def}}{=} \frac{1}{2}\|u-v\|_1^2\,,$$

and

$$B(u,v) = B_{\mathsf{KL}}(u,v) \stackrel{\text{def}}{=} \frac{1}{6}\|u-v\|_1^2\,,$$

satisfy the conditions (2) and (3).

Another suitable loss function for multiclass classification whose corresponding risk is the actual misclassification probability, rather than a surrogate of it, is

$$\ell(u,v) = \ell_{0/1}(u,v) = \begin{cases} 0\,, & \exists i \in [k] \text{ such that } u_i > u_j \text{ and } v_i > v_j \text{ for all } j \in [k]\backslash\{i\} \\ 1\,, & \text{otherwise}\,. \end{cases}$$

Clearly, $\ell(u,v) = \ell(v,u)$, and we have

$$\ell_{0/1}(u,v) + \ell_{0/1}(u',v') + \ell_{0/1}(u,u') \geq \ell_{0/1}(v,v')\,,$$

since the left-hand side is either greater than 1 or the vectors u, u', v, and v' all have their unique maximum at the same coordinate. Therefore, in this case we can choose

$$A(u,v) = A_{0/1}(u,v) = \ell_{0/1}(u,v)\,,$$

and

$$B(u,v) = B_{0/1}(u,v) = \ell_{0/1}(u,v)\,.$$

Theorem 1. *Assuming that the loss function $\ell\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_{\geq 0}$, and the functions $A\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ and $B\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ meet the conditions (2) and (3), then for every $f \in \mathcal{F}$ we have*

$$R(f) + R_\epsilon(f) \geq \max\left\{\mathbb{E}\left(\sup_{\Delta\colon \|\Delta\|\leq\epsilon} B(f(X), f(X+\Delta))\right), \mathbb{E}\, B(Y, Y')\right\}.$$

Proof. Under (3), we can write

$$\begin{aligned}\ell(f(X), Y) + \ell(f(X+\Delta), Y) &= \ell(f(X), Y) + \ell(f(X+\Delta), Y) + A(Y, Y)\\ &\geq B(f(X), f(X+\Delta)).\end{aligned}$$

Taking the supremum with respect to Δ subject to $\|\Delta\| \leq \epsilon$, and then taking the expectation on both sides of the inequality yields

$$\begin{aligned}R(f) + R_\epsilon(f) &= \mathbb{E}\,\ell(f(X), Y) + \mathbb{E}\left(\sup_{\Delta\colon \|\Delta\|\leq\epsilon} \ell(f(X+\Delta), Y)\right)\\ &\geq \mathbb{E}\left(\sup_{\Delta\colon \|\Delta\|\leq\epsilon} B(f(X), f(X+\Delta))\right). \qquad (4)\end{aligned}$$

Similarly, with Y' being independent and identically distributed as Y conditioned on X, i.e., $Y \perp\!\!\!\perp Y' \mid X$ and $\mathbb{P}_{Y'|X} = \mathbb{P}_{Y|X}$, it follows from (2) that

$$\ell(f(X), Y) + \ell(f(X+\Delta), Y') \geq B(Y, Y') - A(f(X), f(X+\Delta)).$$

Again, taking the supremum with respect to Δ subject to $\|\Delta\| \leq \epsilon$, and then taking the expectation on both sides of the inequality yields

$$\begin{aligned}R(f) + R_\epsilon(f) &= \mathbb{E}\,\ell(f(X), Y) + \mathbb{E}\left(\sup_{\Delta\colon \|\Delta\|\leq\epsilon} \ell(f(X+\Delta), Y')\right)\\ &\geq \mathbb{E}\, B(Y, Y') - \mathbb{E}\left(\inf_{\Delta\colon \|\Delta\|\leq\epsilon} A(f(X), f(X+\Delta))\right)\\ &\geq \mathbb{E}\, B(Y, Y'), \qquad (5)\end{aligned}$$

where the second inequality follows from the fact that $A(\cdot)$ is nonnegative. The claimed lower bound on $R(f) + R_\epsilon(f)$ follows by choosing the better lower bound between (4) and (5).

□

The lower bound (4) becomes important if $f(\cdot)$ is not (locally) smooth, whereas the lower bound (5) becomes important if $f(\cdot)$ is (locally) smooth, but the measurement noise is significant.

The following immediate corollary addresses the cases of multiclass classification and least-squares regression mentioned above where the risk of a function $f \in \mathcal{F}$ is $R(f) = \mathbb{E}\,\ell_{\mathsf{KL}}(f(X), Y)$ and $R(f) = \mathbb{E}\,\ell_{\mathsf{LS}}(f(X), Y)$, respectively.

Corollary 1. *Define the* "mean local smoothness factor" *of $f \in \mathcal{F}$ with respect to X as*

$$L_\epsilon(f) = \mathbb{E} \sup_{\Delta\colon \|\Delta\|\leq\epsilon} \|f(X+\Delta) - f(X)\|_1^2,$$

both in the case of least-squares regression and multiclass classification. Then, we have

$$R(f) + R_\epsilon(f) \geq \frac{1}{6}\max\left\{L_\epsilon(f),\, \mathbb{E}\,\|Y - Y'\|_1^2\right\}, \qquad (6)$$

with Y and Y' being i.i.d. conditioned on X.

Proof. The claim follows immediately from Theorem 1 by specializing $B(\cdot)$ to $B_{\mathsf{LS}}(\cdot)$ in the case of least squares regression, and to $B_{\mathsf{KL}}(\cdot)$ in the case of multiclass classification. □

Corollary 1 shows that for every $f \in \mathcal{F}$ the standard risk $R(f)$ and the adversarial risk $R_\epsilon(f)$ cannot be both less than $\frac{1}{12} \max\left\{L_\epsilon(f),\, \mathbb{E}\,\|Y - Y'\|_1^2\right\}$. In particular, because the adversarial risk of a function is always greater than its standard risk, any function $f \in \mathcal{F}$ whose standard risk $R(f)$ is nearly optimal in the sense that[1] $R(f) \lesssim R_\star = \inf_{\widetilde{f} \in \mathcal{F}} R(\widetilde{f})$, cannot achieve an adversarial risk $R_\epsilon(f)$ comparable to $R_\star$ if it is not sufficiently smooth as quantified by $L_\epsilon(f) \gg R_\star$. From a different perspective, the derived trade-off can be interpreted as the necessity of ϵ to be sufficiently small such that $L_\epsilon(f) \lesssim R_\star$, to make $R(f) + R_\epsilon(f) \lesssim R_\star$ possible.

The following is a similar corollary in the case of multiclass classification using $\ell_{0/1}(\cdot)$ as the loss function.

Corollary 2. *For any $\mathcal{S} \subset \mathbb{R}^d$ define the ϵ-core of $\mathcal{S}$ with respect to the norm $\|\cdot\|$ as*

$$\mathrm{core}_\epsilon(\mathcal{S}) \overset{\mathrm{def}}{=} \left\{x \in \mathbb{R}^d \colon x + \epsilon\mathcal{B} \subseteq \mathcal{S}\right\} ,$$

where $\mathcal{B}$ denotes the unit ball of $\|\cdot\|$. For any $f \colon \mathbb{R}^d \to \mathbb{R}^k$ let

$$\mathcal{S}_i(f) = \left\{x \in \mathbb{R}^d \colon f_i(x) > \max_{j \in [k]\setminus\{i\}} f_j(x)\right\} ,$$

denote the set of points $x \in \mathbb{R}^d$ at which the ith coordinate of $f(x)$ is the unique largest entry of $f(x)$. Then, for multiclass classification using the $\ell_{0/1}(\cdot)$ loss, we have

$$R(f) + R_\epsilon(f) \geq \max\left\{\mathbb{P}\left(X \notin \cup_{i\in[k]}\mathrm{core}_\epsilon\left(\mathcal{S}_i(f)\right)\right),\, \frac{1}{2}\,\mathbb{E}\,\|Y - Y'\|_1\right\} ,$$

where, conditioned on X, Y' is an independent copy of Y (whose domain is the canonical basis vectors in $\mathbb{R}^k$).

Proof. In view of Theorem 1, it suffices to show that

$$\mathbb{E} \sup_{\Delta\colon \|\Delta\|\leq\epsilon} \ell_{0/1}(f(X), f(X+\Delta)) = \mathbb{P}\left(X \notin \cup_{i\in[k]}\mathrm{core}_\epsilon\left(\mathcal{S}_i(f)\right)\right) , \tag{7}$$

and

$$\mathbb{E}\,\ell_{0/1}(Y, Y') = \frac{1}{2}\|Y - Y'\|_1 . \tag{8}$$

Recalling the definition of $\ell_{0/1}(\cdot)$, we have $\sup_{\Delta\colon \|\Delta\|\leq\epsilon} \ell_{0/1}(f(x), f(x+\Delta)) = 0$ if for all $\Delta \in \epsilon\mathcal{B}$, the maximum entry of $f(x+\Delta)$ occurs at the same unique coordinate. This can be equivalently expressed as $x \in \cup_{i\in[k]}\mathrm{core}_\epsilon\left(\mathcal{S}_i(f)\right)$, from which (7) follows. Furthermore, (8) holds because the fact that Y and Y' canonical basis vectors in $\mathbb{R}^k$ guarantees that

$$\ell_{0/1}(Y, Y') = \frac{1}{2}\|Y - Y'\|_1 .$$

□

An intuitive interpretation of Corollary 2 is that if the functions $f \in \mathcal{F}$ that provide near-optimal standard accuracy, in the sense that $R(f) \lesssim R_\star = \inf_{\widetilde{f}\in\mathcal{F}} R(\widetilde{f})$, have small ϵ-cores in regions where each coordinate of f is dominant, then adversarial robustness comes at the cost of losing the accuracy.

Of course, the bounds established in Theorem 1, Corollary 1, and Corollary 2 inevitably contain abstract terms due to the generality of these bounds. However, these abstract terms can be approximated appropriately using the structure of the special prediction problems of interest. In the next section, we consider a special regression problem and express the derived bounds in terms of more explicit quantities.

[1] We write $a \lesssim b$ if $a \leq Cb$ for some absolute constant $C > 0$.

3 Least Squares Regression over Polynomial Ridge Functions

Consider the case of least squares regression over the set of polynomial ridge functions

$$\mathcal{F}_p = \left\{ x \mapsto f_\theta(x) \stackrel{\text{def}}{=} \langle \theta, x\rangle^p \colon \theta \in \mathbb{R}^d \right\} ,$$

for some integer $p \geq 1$. In particular, for a parameter $\theta_\star \in \mathbb{R}^d$ the observations have the form

$$Y = f_{\theta_\star}(X) + Z ,$$

where X is a zero-mean random variable in $\mathbb{R}^d$ whose marginals have finite moments of order at least $2p$, and the noise term Z is a random scalar independent of X such that $\mathbb{E}\, Z = 0$ and $\mathbb{E}\, Z^2 = \sigma^2$. This model reduces to the standard linear regression model for $p = 1$. With Σ denoting the covariance matrix of X and $\|\theta\|_\Sigma \stackrel{\text{def}}{=} (\theta^\intercal \Sigma \theta)^{1/2}$, we further assume that for some constant $C_p > 0$, for every $\theta \in \mathbb{R}^d$ we have

$$\left(\mathbb{E}\, |\langle \theta, X\rangle|^{2p}\right)^{1/(2p)} \leq C_p \|\theta\|_\Sigma \,. \tag{9}$$

For $f_\theta(x) = \langle \theta, x\rangle^p \in \mathcal{F}_p$ the quantity $L_\epsilon(f_\theta)$ can be expressed as

$$\begin{aligned} L_\epsilon(f_\theta) &= \mathbb{E} \sup_{\Delta \colon \|\Delta\| \leq \epsilon} |\langle \theta, X + \Delta\rangle^p - \langle \theta, X\rangle^p|^2 \\ &= \mathbb{E}\left((|\langle \theta, X\rangle| + \|\theta\|_* \epsilon)^p - |\langle \theta, X\rangle|^p \right)^2 \end{aligned}$$

where the second line follows from the facts that

$$(|z_0| + \delta)^p - |z_0|^p \leq \sup_{z \colon |z - z_0| \leq \delta} |z^p - z_0^p| ,$$

as one can specialize the right-hand side to the case $z = z_0 + \operatorname{sgn}(z_0)\delta$, and

$$\begin{aligned} \sup_{z \colon |z - z_0| \leq \delta} |z^p - z_0^p| &= \sup_{z \colon |z - z_0| \leq \delta} \left| \sum_{k=1}^p \binom{p}{k} (z - z_0)^k z_0^{p-k} \right| \\ &\leq \sum_{k=1}^p \binom{p}{k} \delta^k |z_0|^{p-k} \\ &= (|z_0| + \delta)^p - |z_0|^p \,. \end{aligned}$$

To obtain upper and lower bounds for $L_\epsilon(f_\theta)$ we will use the following identities

$$\mathbb{E}\left((|\langle \theta, X\rangle| + \|\theta\|_* \epsilon)^p - |\langle \theta, X\rangle|^p \right)^2 = \mathbb{E}\left(\sum_{k=1}^p \binom{p}{k} |\langle \theta, X\rangle|^{p-k} \|\theta\|_*^k \epsilon^k \right)^2 \tag{10}$$

$$\mathbb{E}\left(\sum_{k=1}^K \binom{p}{k} |\langle \theta, X\rangle|^{p-k} \|\theta\|_*^k \epsilon^k \right)^2 = \sum_{j=1}^K \sum_{k=1}^K \binom{p}{j} \binom{p}{k} \mathbb{E}\, |\langle \theta, X\rangle|^{2p-j-k} \|\theta\|_*^{j+k} \epsilon^{j+k} , \tag{11}$$

for $K \in [p]$. Using the fact that for any pair of nonnegative numbers a and b we have $(a+b)^2 \geq a^2 + b^2$, it follows from (10) and (11) that

$$\begin{aligned} \mathbb{E}\left((|\langle \theta, X\rangle| + \|\theta\|_* \epsilon)^p - |\langle \theta, X\rangle|^p \right)^2 &\geq \|\theta\|_*^{2p} \epsilon^{2p} + \mathbb{E}\left(\sum_{k=1}^{p-1} \binom{p}{k} |\langle \theta, X\rangle|^{p-k} \|\theta\|_*^k \epsilon^k \right)^2 \\ &= \|\theta\|_*^{2p} \epsilon^{2p} + \sum_{j=1}^{p-1} \sum_{k=1}^{p-1} \binom{p}{j} \binom{p}{k} \mathbb{E}\, |\langle \theta, X\rangle|^{2p-j-k} \|\theta\|_*^{j+k} \epsilon^{j+k} \end{aligned}$$

$$
\begin{aligned}
&\geq \|\theta\|_*^{2p}\epsilon^{2p} + \sum_{j=1}^{p-1}\sum_{k=1}^{p-1}\binom{p}{j}\binom{p}{k}\|\theta\|_\Sigma^{2p-j-k}\|\theta\|_*^{j+k}\epsilon^{j+k} \\
&= \|\theta\|_*^{2p}\epsilon^{2p} + \left(\sum_{k=1}^{p-1}\binom{p}{k}\|\theta\|_\Sigma^{p-k}\|\theta\|_*^{k}\epsilon^{k}\right)^2 \\
&\geq \frac{1}{2}\left(\sum_{k=1}^{p}\binom{p}{k}\|\theta\|_\Sigma^{p-k}\|\theta\|_*^{k}\epsilon^{k}\right)^2 ,
\end{aligned}
$$

where we used the power mean inequality on the third line and the Cauchy–Schwarz inequality on the fifth line. Therefore, we have shown that

$$
L_\epsilon(f_\theta) \geq \frac{1}{2}\left((\|\theta\|_\Sigma + \|\theta\|_*\epsilon)^p - \|\theta\|_\Sigma^p\right)^2 .
$$

Furthermore, we have

$$
\mathbb{E}(Y - Y')^2 = 2\sigma^2 .
$$

Applying the inequalities above in (6), we obtain

$$
R(f_\theta) + R_\epsilon(f_\theta) \geq \max\left\{\frac{1}{12}\left((\|\theta\|_\Sigma + \|\theta\|_*\epsilon)^p - \|\theta\|_\Sigma^p\right)^2 , \frac{\sigma^2}{3}\right\} .
$$

Because the standard risk is of the form

$$
R(f_\theta) = \mathbb{E}\left(\langle X, \theta\rangle^p - \langle X, \theta_\star\rangle^p\right)^2 + \sigma^2 ,
$$

if $R(f_\theta) + R_\epsilon(f_\theta) \lesssim R(f_{\theta_\star}) = \sigma^2$, meaning that $f_\theta(X)$ is an accurate and robust predictor for Y, then we must have

$$
\mathbb{E}\left(\langle X, \theta\rangle^p - \langle X, \theta_\star\rangle^p\right)^2 \lesssim \sigma^2 ,
$$

and

$$
\left((\|\theta\|_\Sigma + \|\theta\|_*\epsilon)^p - \|\theta\|_\Sigma^p\right)^2 \lesssim \sigma^2 .
$$

With

$$
\lambda_* = \sup_{\vartheta \neq 0} \frac{\|\vartheta\|_\Sigma^2}{\|\vartheta\|_*^2} ,
$$

we can write

$$
\begin{aligned}
\left((\|\theta\|_\Sigma + \|\theta\|_*\epsilon)^p - \|\theta\|_\Sigma^p\right)^2 &= \|\theta\|_\Sigma^{2p}\left(\left(1 + \frac{\|\theta\|_*}{\|\theta\|_\Sigma}\epsilon\right)^p - 1\right)^2 \\
&\geq \|\theta\|_\Sigma^{2p}\left(\left(1 + \frac{\epsilon}{\sqrt{\lambda_*}}\right)^p - 1\right)^2 .
\end{aligned}
$$

Furthermore, using the triangle inequality and the moments equivalence assumption (9), we also have

$$
\begin{aligned}
\sqrt{\mathbb{E}\langle X, \theta_\star\rangle^{2p}} &\leq \sqrt{\mathbb{E}\left(\langle X, \theta\rangle^p - \langle X, \theta_\star\rangle^p\right)^2} + \sqrt{\mathbb{E}\langle X, \theta\rangle^{2p}} \\
&\leq \sqrt{\mathbb{E}\left(\langle X, \theta\rangle^p - \langle X, \theta_\star\rangle^p\right)^2} + C_p^p\|\theta\|_\Sigma^p .
\end{aligned}
$$

Combining the derived bounds we obtain

$$
\mathbb{E}\langle X, \theta_\star\rangle^{2p} \lesssim \left(1 + \frac{C_p^p}{(1 + \epsilon/\sqrt{\lambda_*})^p - 1}\right)^2 \sigma^2 .
$$

Interpreting $\mathsf{SNR}_p = \mathbb{E}\left\langle X, \theta_\star \right\rangle^{2p} / \sigma^2$ as the Signal-to-Noise-Ratio and using the inequality $(1+\epsilon/\sqrt{\lambda_*})^p - 1 \geq \max\{p\epsilon/\sqrt{\lambda_*}, (\epsilon/\sqrt{\lambda_*})^p\}$, the bound above shows that if

$$\epsilon \gg \min\left\{ \frac{C_p^p}{p}\sqrt{\frac{\lambda_*}{\mathsf{SNR}_p}}, C_p\sqrt{\frac{\lambda_*}{\mathsf{SNR}_p^{1/p}}} \right\}, \tag{12}$$

adversarial robustness is impossible unless we are operating at low SNR_p meaning that the optimal standard risk is also relatively large.

In particular, for linear regression which corresponds to the case of $p = 1$, where we have $\mathsf{SNR}_1 = \|\theta_\star\|_\Sigma^2/\sigma^2$ and $C_1 = 1$, adversarial robustness is impossible if

$$\epsilon \gg \sqrt{\frac{\lambda_*}{\|\theta_\star\|_\Sigma^2}}\,\sigma\,,$$

unless $\|\theta_\star\|_\Sigma^2/\sigma^2$ is low.

It is worth mentioning that the threshold for ϵ specified by (12) is in general dependent on the dimension d through the parameter λ_*. For example, if X has the identity matrix as its covariance (i.e., $\Sigma = I$), and the perturbations are bounded in ℓ_∞ norm (i.e., $\|\cdot\| = \|\cdot\|_\infty$), then we have $\lambda_* = \sup_{\theta\neq 0}\|\theta\|_2^2/\|\theta\|_1^2 = 1/d$. Therefore, if p, C_p, and SNR_p are constants independent of the dimension d, robustness against adversarial ℓ_∞ perturbations of size greater than $O(d^{-1/2})$ cannot be guaranteed.

References

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4312–4321. International Joint Conferences on Artificial Intelligence Organization, August 2021. doi: 10.24963/ijcai.2021/591. URL https://doi.org/10.24963/ijcai.2021/591. Survey Track.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford, February 2013. ISBN 978-0199535255.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019. URL https://arxiv.org/abs/1902.06705.

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. In *IMS Lecture Notes Monograph Series*, volume 56. Institute of Mathematical Statistics, 2007. doi: 10.1214/074921707000000391.

Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guaratees for adversarially robust Gaussian classification. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2345–2355. PMLR, July 2020. URL https://proceedings.mlr.press/v119/dan20b.html.

Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *IEEE Transactions on Information Theory*, 69(12):7793–7822, 2023. doi: 10.1109/TIT.2022.3205449.

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, August 2017. ISSN 1573-0565. doi: 10.1007/s10994-017-5663-3.

Adel Javanmard and Mohammad Mehrabi. Adversarial robustness for latent models: Revisiting the robust-standard accuracies tradeoff. *Operations Research*, 72(3):1016–1030, 2024. doi: 10.1287/opre.2022.0162. URL https://doi.org/10.1287/opre.2022.0162.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2034–2078. PMLR, July 2020. URL https://proceedings.mlr.press/v125/javanmard20a.html.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f708f064faaf32a43e4d3c784e6af9ea-Paper.pdf.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1683–1709, Paris, France, July 2015. PMLR. URL https://proceedings.mlr.press/v40/Thrampoulidis15.html.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Mądry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes*. Springer New York, 2012.