Abstract—We study the error of linear regression in the face of adversarial attacks. In this framework, an adversary changes the input to the regression model in order to maximize the prediction error. We provide bounds on the prediction error in the presence of an adversary as a function of the parameter norm and the error in the absence of such an adversary. We show how these bounds make it possible to study the adversarial error using analysis from non-adversarial setups. The obtained results shed light on the robustness of overparameterized linear models to adversarial attacks. Adding features might be either a source of additional robustness or brittleness. On the one hand, we use asymptotic results to illustrate how double-descent curves can be obtained for the adversarial error. On the other hand, we derive conditions under which the adversarial error can grow to infinity as more features are added, while at the same time, the test error goes to zero. We show this behavior is caused by the fact that the norm of the parameter vector grows with the number of features. It is also established that \( \ell_\infty \) and \( \ell_2 \)-adversarial attacks might behave fundamentally different due to how the \( \ell_1 \) and \( \ell_2 \)-norms of random projections concentrate. We also show how our reformulation allows for solving adversarial training as a convex optimization problem. This fact is then exploited to establish similarities between adversarial training and parameter-shrinking methods and to study how the training might affect the robustness of the estimated models.

Index Terms—Adversarial machine learning, parameter estimation, regression analysis.

I. INTRODUCTION

A S MACHINE learning models start to be considered for critical applications such as medical settings [1] or autonomous driving [2], their vulnerabilities and brittleness become a pressing concern [3]. The adversarial attack framework is popular for studying these issues. It considers inputs contaminated with small disturbances deliberately chosen to maximize the model error. The susceptibility of state-of-the-art neural network models to very small input modifications [4] gave the framework a lot of attention from the research community.

There is a conflicting view on the relationship between high-dimensionality and model robustness to adversarial attacks that served as the driving force for this work. On the one hand, high-dimensionality is pointed out as a source of vulnerability to adversarial attacks [5], [6], [7]. On another hand, a new line of work has established the advantages of high-dimensionality: the study of double-descent curves show that it is sometimes possible to obtain improvements in performance if we continue to increase the model size beyond the point of interpolation. The phenomenon is counterintuitive because it shows that under certain conditions overfitting the dataset can also be benign to model generalization and performance; precise conditions for benign overfitting are provided by Bartlett et al. [8]. The idea also applies to robustness analysis and increasing the model size can be a recipe to obtain more robust models—as shown by Bubeck et al. [9] using isoperimetric inequalities.

Linear models are a natural setting to study the role of high-dimensions in the robustness to adversarial attacks. Not only can linear models be made vulnerable to adversarial attack [10], but the double-descent and the benign overfitting phenomenon can also be observed in a purely linear setting [8], [11]. Indeed, there is a growing body of work that study fundamental properties of adversarial attacks in linear models [12], [13], [14], [15], [16].

In this paper, we consider adversarial attacks in the context of linear regression. Given an input \( x_0 \in \mathbb{R}^m \) and an output \( y_0 \in \mathbb{R} \), a linear model \( \hat{\beta} \) makes a prediction \( \hat{\beta}^T x \). The adversary modifies the input with a disturbance \( \Delta x \) such that, for a magnitude of at most \( \delta \), it maximizes the squared adversarial error,

\[
\max_{\|\Delta x\|_2 \leq \delta} (y_0 - (x_0 + \Delta x)^T \hat{\beta})^2. \tag{1}
\]

We refer to the above attack as \( \ell_2 \)-adversarial attack since it constrains the attack to a ball in the \( \ell_2 \)-norm. More generally, we consider the framework of \( \ell_p \)-adversarial attacks, which contains the commonly used \( \ell_2 \) and \( \ell_\infty \) adversaries as special cases [4], [5], [6], [7], [10], [17], [18], [19].

Adversarial examples in linear regression have been studied before, e.g. Javanmard et al., [13] presents an exact asymptotic analysis. Here we use an approximate analysis instead. The quantity of interest is the adversarial risk, i.e., the expected value of the squared adversarial error displayed in (1). We approximate the adversarial risk by the sum of risk and the parameter norm \( \|\hat{\beta}\| \) and show that the analysis of these two components is enough to explain high-dimensionality both as a source of brittleness and as a potential recipe for producing robust models. This setting can be used to gain insight into...
we show how adversarial training affects the conclusion obtained for minimum-norm solutions.

a) We show how our formulation allows for solving adversarial training in linear regression as a convex optimization problem.

b) We compare and establish similarities with ridge regression and lasso [22].

c) We study how adversarial training and parameter shrinking methods affect the parameter norm grow with the number of features. We use our observations to explain when it effectively prevents vulnerability to adversarial attacks.

II. PROBLEM FORMULATION

Consider a training dataset $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ consisting of $n$ i.i.d. datapoints of dimension $\mathbb{R}^m \times \mathbb{R}$, sampled from the distribution $(x_i, y_i) \sim P_{x,y}$. To this data, we fit a linear model from the function class $\{\beta^T x \mid \beta \in \mathbb{R}^m\}$. We use $\hat{\beta}$ to denote the parameter estimated from the training data. The estimation method is detailed in what follows.

We will use subscripts to denote the source of randomness considered in the conditional expectation. For instance, let $x, y, z, w$ be random variables, we use $\mathbb{E}_{x,y}[f(x, y, z, w)]$ to denote the expectation with respect to $x, y$ and conditioned on the variables that are not explicitly mentioned in the subscript, here $z$ and $w$.

Let $(x_0, y_0) \sim P_{x,y}$ be a point not seen during training and independently sampled from the same distribution as the rest of the data. We denote the out-of-sample prediction risk by

$$R(\hat{\beta}) = \mathbb{E}_{x_0,y_0} \left[ (y_0 - x_0^T \hat{\beta})^2 \right].$$

(2)

The $\ell_p$-adversarial risk is defined as

$$P_p^{\text{adv}}(\hat{\beta}) = \mathbb{E}_{x_0,y_0} \left[ \max_{\|\Delta x_0\|_p \leq \delta} (y_0 - (x_0 + \Delta x_0)^T \hat{\beta})^2 \right],$$

(3)

which is the risk when the model is subject to a disturbance $\Delta x_0$ that results in the worst possible performance inside the region $\|\Delta x_0\|_p \leq \delta$. Here, we use $\|\cdot\|_p$ to denote the $\ell_p$-norm of a vector. That is, given a vector $a \in \mathbb{R}^n$, $\|a\|_p = (\sum_{i=1}^n |a_i|^p)^{\frac{1}{p}}$ and $\|a\|_{\infty} = \max_{1 \leq i \leq n} |a_i|$. The design parameter $\delta$ is the maximum size for the adversarial perturbation.

Moreover, the empirical risk is denoted by

$$\hat{R}(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \hat{\beta})^2,$$

(4)

where the expectation w.r.t. the true distribution is replaced by the average over the observed training samples. We use a similar notation for the empirical adversarial risk, $\hat{P}_p^{\text{adv}}(\hat{\beta})$. 

Fig. 1. Double-descent vs robustness-accuracy trade-off. On the $x$-axis we have the ratio between the number of features $m$ and the number of training datapoints $n$. The model out-of-sample prediction risk (when there is no adversarial disturbance) continually decreases in the overparameterized region, achieving significantly better results than in the underparameterized region. On the other hand, increasing the number of features continuously produce worse adversarial robustness for the risk of $\ell_\infty$-adversarial attacks. For an $\ell_2$-adversary, however, the model actually benefits from operating in the overparameterized region and larger models yield better adversarial robustness. The precise setup for this example is provided in Section IV-G.

Contributions

This paper makes the following contributions:

1) We show that the analysis of adversarial attacks can be simplified in linear regression problems. The adversarial error in (1) can be rewritten as:

$$\left( y_0 - x_0^T \hat{\beta} + \delta \|\hat{\beta}\|_2 \right)^2.$$

This is proved in Section III, where we also discuss the implications of the reformulation.

2) In Section IV, we show that the ratio between $\ell_2$-adversarial risk and $(\text{risk} + \delta^2 \|\hat{\beta}\|_2^2)$ is always a factor between 1 and 2. We use this approximation to analyse the adversarial risk:

a) The minimum-norm solution is commonly used to select overparameterized models in connection to the study of the double-descent phenomenon [8], [20]. We use asymptotic and non-asymptotic analysis to show for this estimate that $\|\hat{\beta}\|_2$ decreases as we add more features. This fact is used to obtain cases where we can observe double-descent in adversarial scenarios.

b) We generalize the analysis to other $\ell_p$-norms. While $\|\hat{\beta}\|_2$ decreases as we add more features, the $\ell_1$-norm, $\|\hat{\beta}\|_1$, does not. This is used to explain why more features increase the models vulnerability to some types of adversarial attacks; i.e., in Fig. 1 the model becomes more vulnerable to $\ell_\infty$-adversarial attacks as more features are added, but not to $\ell_2$-adversarial attacks.

3) Adversarial training is a standard method to produce models that are robust to adversarial attacks [21]. In Section V, we show how adversarial training affects the conclusion obtained for minimum-norm solutions.

a) We show how our formulation allows for solving adversarial training in linear regression as a convex optimization problem.

b) We compare and establish similarities with ridge regression and lasso [22].

c) We study how adversarial training and parameter shrinking methods affect the parameter norm grow with the number of features. We use our observations to explain when it effectively prevents vulnerability to adversarial attacks.
III. ADVERSARIAL RISK IN LINEAR REGRESSION

In this paper, we show that the adversarial risk for linear regression can be simplified. The following lemma gives a quadratic form for the adversarial risk defined in (3).

**Lemma 1:** Let \( q \) be a positive real number for which \( \frac{1}{p} + \frac{1}{q} = 1 \), then

\[
R_p^{\text{adv}}(\beta) = \mathbb{E}_{x_0, y_0} \left[ \left( |y_0 - x_0^T \hat{\beta} + \delta \| \hat{\beta} \|_q \right)^2 \right].
\]

The proof of the lemma above turns out to be a useful tool when analysing robustness to adversarial attacks and adversarial training. A contribution of this paper is to show how it can be used in various situations. In Section IV, Lemma 1 is used to analyse the adversarial robustness of linear models and the interplay between overparameterization and robustness. In Section V, we show how the formula allows for an efficient method for adversarial training using convex programming.

The proof of the lemma is based on Hölder’s inequality \( |\beta^T \Delta x| \leq \|\beta\|_p \|\Delta x\|_q \) and on the fact that, given \( \beta \), we have \( \Delta x \) such that the equality holds. The next proposition gives the precise construction that we will make use of.

**Proposition 2:** Given \( p, q \in (1, \infty) \) such that \( 1/p + 1/q = 1 \) and \( \beta \in \mathbb{R}^m \), \( \Delta x \in \mathbb{R}^m \). We have \( |\beta^T \Delta x| = \|\beta\|_q \|\Delta x\|_p \) when the \( i \)-th component is \( \Delta x_i = \text{sign}(\beta_i) \beta_i^{1/p} \). Moreover, if \( \Delta x_i = \text{sign}(\beta_i) \delta \) for every \( i \), then \( |\beta^T \Delta x| = \|\beta\|_1 \|\Delta x\|_\infty \).

\[
\Delta x_i = \frac{s_i}{\sum s_i} \text{ for } s_i = \begin{cases} 1 & \text{if } \beta_i = \max_i \beta_i \\ 0 & \text{otherwise} \end{cases}
\]

then \( |\beta^T \Delta x| = \|\beta\|_\infty \|\Delta x\|_1 \).

The above proposition is well-known. See, for instance, Exercise 4, Section 2.4 of Ash et al. [23]. Indeed, most adversarial attacks are constructed based on it. For instance, the Fast Gradient Sign Method (FGSM) [10] linearizes the neural network and applies the above construction for \( \rho = \infty \) to obtain the adversarial perturbation. For completeness, we also provide a proof of the proposition in the appendix. We are now ready to prove Lemma 1.

**Proof of Lemma 1:** For \( 1 \leq p \leq \infty \), let \( q \) be a positive real number such that \( \frac{1}{p} + \frac{1}{q} = 1 \). Let \( e_0 = y_0 - x_0^T \hat{\beta} \). After some algebraic manipulation, (3) can be rewritten as

\[
R_p^{\text{adv}}(\beta) = \mathbb{E}_{x_0, y_0} \left[ e_0^2 + \max_{\|\Delta x_0\|_p \leq \delta} \left( (\Delta x_0^T \hat{\beta})^2 - 2e_0 \Delta x_0^T \hat{\beta} \right) \right].
\]

In turn, if we define \( r = \Delta x_0^T \hat{\beta} \) and use the fact that \( \|\Delta x_0\|_p \leq \delta \), Hölder’s inequality yields

\[
|r| \leq \delta \|\hat{\beta}\|_q,
\]

for \( q \) satisfying \( 1/p + 1/q = 1 \). Since Proposition 2 guarantees that we can always choose vectors such that the equality holds, the term inside the expectation is equal to \( M + e_0^2 \), where

\[
M = \max_{|r| \leq \delta \|\beta\|_q} (r^2 - 2e_0 r).
\]

\(1\) The result still holds for the pair of values \((p = 1, q = \infty)\) and \((p = \infty, q = 1)\).

Now, the maximum is attained at \( r = -\delta \|\hat{\beta}\|_q \) if \( e_0 \geq 0 \) and at \( r = \delta \|\hat{\beta}\|_q \) if \( e_0 < 0 \). Hence, \( M = \delta^2 \|\hat{\beta}\|_q^2 + 2\delta \|\hat{\beta}\|_q |e_0| \) and

\[
R_p^{\text{adv}}(\beta) = \mathbb{E}_{x_0, y_0} \left[ \delta^2 \|\hat{\beta}\|_q^2 + 2\delta \|\hat{\beta}\|_q |e_0| + |e_0|^2 \right].
\]

\(\blacksquare\)

IV. ADVERSARIAL ROBUSTNESS OF LINEAR MODELS

In this section, we analyse the adversarial risk, by exploiting Lemma 1. Expanding (5) and using the linearity of the expectation operator we obtain

\[
R_p^{\text{adv}}(\beta) = \left( R(\beta) + \delta^2 \|\hat{\beta}\|_q^2 \right) + 2\delta \|\hat{\beta}\|_q \mathbb{E}_{x_0, y_0} \left[ |y_0 - x_0^T \hat{\beta}| \right].
\]

The term inside the parenthesis is present in most regularized settings and we can naturally see similarities with the cost function of lasso and ridge regression for \( q = 1 \) and 2, respectively. The last term is always positive and can be upper bounded using Jensen’s inequality. Hence, it follows that

\[
R(\beta) + \delta^2 \|\hat{\beta}\|_q^2 \leq R_p^{\text{adv}}(\beta) \leq \left( \sqrt{R(\beta) + \delta \|\hat{\beta}\|_q^2} \right)^2.
\]

The inequality can be further simplified using the fact that for all \( a, b \geq 0 \) the inequality \( (a + b)^2 \leq 2(a^2 + b^2) \) holds. Hence,

\[
1 \leq \frac{R_p^{\text{adv}}(\beta)}{R(\beta) + \delta^2 \|\hat{\beta}\|_q^2} \leq 2.
\]

That is, the adversarial risk is between 1 and 2 times the quantity \( R(\beta) + \delta^2 \|\hat{\beta}\|_q^2 \). Such bounds allow for the analysis of adversarial robustness from values that are often obtained from other analyses (the risk and the parameter norm). For instance, from the double-descent literature, it is well-known that the \( \ell_2 \)-norm of the estimated parameter often also exhibits a double-descent behavior as a function of the number of features. This is observed experimentally, for instance, by Belkin et al. [20]. Hence, the above approximation offers an easy way to understand why we can expect double-descent behavior for the \( \ell_2 \)-adversarial risk.

The approximation is also enough to justify the potential brittleness of high-dimensional models. It gives sufficient and necessary conditions for a model with good test performance to be made vulnerable to adversarial attacks as more features are added. For instance, (8) implies that: \( \text{For a sufficiently small risk } R(\beta) < \epsilon, \text{ the adversarial risk } R_p^{\text{adv}}(\beta) \to \infty \text{ as } m \to \infty \text{ if and only if } \delta \|\hat{\beta}\|_q \to \infty \). The next example demonstrates a fairly simple case where such behavior can be observed.

A. Motivating Example: Weak Features

In this section, we show brittleness and vulnerability to \( \ell_\infty \)-adversarial attacks arising from high-dimensionality. Tsipras et al. [6] makes use of a linear example to motivate their argument that robustness and accuracy might be at odds. We show a modified construction\(^2\) to motivate how brittleness might...
risk is minimized (i.e., as in Fig. 2(b)). The second aspect is the scaling $\eta$, we will motivate different choices and show how they can yield quite different results.

**B. Preliminaries**

Here we focus on the minimum norm solution, which is often used when studying the behavior of overparameterized models in connection with the double-descent phenomenon [20]. We will assume that the training and test data have been generated linearly with additive noise:

$$y_i = x_i^\top \beta + \epsilon_i,$$

(10)

where $P_x$ is a distribution in $\mathbb{R}$ such that $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{V}[\epsilon_i] = \sigma^2$ and $\epsilon_i$ is assumed to be independent of $x_i$. Moreover, $\mathbb{E}[x_i] = 0$ and $\text{Cov}[x_i] = \Sigma$. The $\ell_2$-norm of the data generation parameter is denoted by $\|\beta\|_2 = r^2$.

Let $X \in \mathbb{R}^{n \times m}$ denote the matrix consisting of stacked training inputs $x_i^\top$ and similarly let $y \in \mathbb{R}^n$ denote the output vector. The parameters are estimated as

$$\tilde{\beta} = (X^\top X)^\dagger X^\top y,$$

(11)

where $(X^\top X)^\dagger$ represents the pseudo-inverse of $X^\top X$. In the underparameterized ($m < n$) region, it corresponds to the least-square solution. In the overparameterized ($m > n$) case, where more than one solution is possible, this corresponds to the solution for which the parameter norm $\|\beta\|_2$ is minimum, the **minimum-norm solution**.

From (10) and (11) it follows that:

$$\tilde{\beta} = (X^\top X)^\dagger X^\top \beta + (X^\top X)^\dagger X^\top \epsilon,$$

(12)

where we have introduced the following notation $\Sigma = {\frac{1}{m}}X^\top X$, $\Phi = \Sigma^2 \Sigma$ and $\Pi = I - \Phi$. Here, $\Phi$ and $\Pi$ are orthogonal projectors: $\Pi$ is the projection into the null space of $X$ and $\Phi$ into the row space of $X$. The first term in (12) can be understood as a projection of the original parameter into the row space of the regressors and it is the parameters estimated in a noiseless scenario. The second term is the consequence of the noise. It follows that the risk and the expected parameter norm can be decomposed as in the subsequent Lemma. The proof is provided in the Supplementary Material.

**Lemma 3 (Bias-variance decomposition):** Denote $\|z\|_2^2 = z^\top \Sigma z$. The expected risk and $\ell_2$ parameter norm are

$$\mathbb{E}_x[R(\beta)] = \|\Pi \beta\|_2^2 + \frac{\sigma^2}{n} \text{tr}(\Sigma) + \sigma^2,$$

(13)

$$\mathbb{E}_x[\|\hat{\beta}\|_2^2] = \|\Phi \beta\|_2^2 + \frac{\sigma^2}{n} \text{tr}(\Sigma^2).$$

(14)

The following bounds can be used in conjunction with Lemma 3 to analyze the adversarial risk

$$R + \delta^2 L_q \leq L_p^\text{adv} \leq (\sqrt{R} + \delta \sqrt{L_q})^2,$$

(15)

where $L_q = \mathbb{E}_x[\|\hat{\beta}\|_2^4]$, $R = \mathbb{E}_x[R(\beta)]$ and $R_p^\text{adv} = \mathbb{E}_x[R_p^\text{adv}(\tilde{\beta})]$. 

![Image](Fig. 2) **Motivating example: weak features.** For the data generated as in (9), we show the risk and the $\ell_\infty$-adversarial risk on the test dataset for: (a) the optimal predictor $\hat{\beta} = [\frac{1}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}}]$; and, (b) the predictor obtained from a training dataset using the minimum-norm solution, i.e. (11).
parameters are added to the model. This naturally yields models more robust to $\ell_2$ perturbations. This can be observed in Fig. 3: after the local minima $\gamma = \sqrt{n}$ in the standard risk, the risk is increasing with $\gamma$. However, the adversarial risk for, say, $\delta = 2$ is decreasing due to the tendency of the minimum-norm solution to select smoother solutions. Moreover, while the standard risk does not have better results in the overparameterized region than in the underparameterized region, the adversarial risk in the overparameterized region can actually be better than the adversarial risk in the underparameterized region.

D. Non-Asymptotic Results for the Parameter $\ell_2$-norm

Central to our analysis of $\ell_2$-adversarial attacks is the idea that the parameter norm decays with the rate $O((\frac{m}{n})^{-\frac{1}{2}})$ even when the data generator parameter remains constant, i.e. $\|\beta\|_2 = r$. This intuition is formalized in the theorem below. We do not attempt to provide the most general result, instead, our choice is motivated by the fact that many of the steps in proving this theorem can be carried on to the $\ell_1$-norm.

Theorem 5: Let the data be generated according to (10). Assume additionally that $m > n$ and:

1) The noise $\epsilon$ and the regressor $x$ are sub-Gaussian.
2) the regressor $x$ is sampled from a rotationally invariant distribution. Then there exists constants $c, C > 0$ such that for all $t < \sqrt{\frac{m}{n}}$ with probability $1 - 2\exp(-ct^2n)$ we have:

$$\|\hat{\beta}\|_2 \leq r \frac{1 + t}{\sqrt{m/n}} + \sigma \frac{1 + t}{C\sqrt{m/n} - 1}.$$  

This theorem provides a non-asymptotic result and an exponential rate of convergence. It strengthens the assumptions from the previous section in two ways. First, it assumes the variables are sub-Gaussian, which is used to obtain an exponential rate of convergence. If this assumption is relaxed lower rates of concentration are obtained. For instance, Hastie et al. [11] does not assume this, which results in a convergence rate of $n^{-1/7}$. Secondly, it assumes the regressor to be rotationally invariant. Hence, given an orthogonal matrix $Q$, multiplication by this matrix does not change the distribution, i.e. $x \sim Qx$. Standard examples where $x$ is rotationally invariant are values sampled from standard Gaussian or from the uniform distribution over the sphere. Rotational invariance implies isotropy, but not all isotropic distributions are rotationally invariant. Hence, this is again a stronger assumption.

From Lemma 3 we have:

$$E_x \left[ \|\hat{\beta}\|_2^2 \right] = \|\Phi \beta\|_2^2 + \sigma^2 \frac{\text{tr}(\Sigma)}{n}.$$  

The next Lemma gives concentration inequalities for the eigenvalues of $\hat{\Sigma}$. If we use $\lambda_i(\hat{\Sigma})$ to denote the $i$-th eigenvalue of $\hat{\Sigma}$, we have that $\text{tr}(\Sigma) = \frac{2}{n} \sum_i \lambda_i(\Sigma)$. Hence, the following result immediately implies that the second term of the above expression concentrates around $\Theta((\frac{m}{n})^{-\frac{1}{2}})$. The proof is provided in the Supplementary Material.

Lemma 6: Let $x_i \in \mathbb{R}^m$ be independently sampled sub-Gaussian random vectors, $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$, and let $m > n$. Then there exists a constant $C > 0$ such that, with probability
greater than \(1 - 2 \exp(-m)\),
\[
\frac{1}{(C \sqrt{\frac{m}{n}} + 1)^2} \leq \lambda_i(\hat{\Sigma}) \leq \frac{1}{(C \sqrt{\frac{m}{n}} - 1)^2}.
\]

**Proof:** From the lemma statement: \(x_i\) are independently sampled sub-Gaussian vectors. Let \(X\) be a matrix containing the vectors \(x_i\) as its rows. Let \(s_i(X)\) denote the \(i\)-th singular value of \(X\). From Vershynin [24, Theorem 4.6.1] we have that there exist a constant \(C\) such that with probability larger then \(1 - 2 \exp(-t^2)\),
\[
\sqrt{n} - C(\sqrt{m} + t) \leq s_i(X) \leq \sqrt{n} + C(\sqrt{m} + t), \forall i = 1, \ldots, n.
\]

Set \(t = \sqrt{m}\) then, since \(\lambda_i(\Sigma) = \left(\frac{1}{\sqrt{n}} s(n-i)(X)\right)^{-2}\) we obtain with probability greater than \(1 - 2 \exp(-m)\) that
\[
\frac{1}{(1 + C \sqrt{\frac{m}{n}})^2} \leq \lambda_i(\hat{\Sigma}) \leq \frac{1}{(1 - C \sqrt{\frac{m}{n}})^2}, \forall i = 1, \ldots, n.
\]

Finally, since \(m > n\), we have that \(1 - 2 \exp(-m) > 1 - 2 \exp(-n)\) and the result follows.

Next, we turn to the analysis of the first term in (14). In the case \(m > n\), \(\Phi\) is a projection matrix that projects a vector from \(\mathbb{R}^m\) into a subspace of dimension \(n\). The set of all possible subspaces of dimension \(n\) in \(\mathbb{R}^m\) is well studied and known as the Grassmannian manifold \(G(m, n)\). There is a one-to-one relationship between the projection matrices \(\Phi\) and the points in this manifold. For the case when \(X\) is rotationally invariant, we have, given any orthogonal matrix \(Q\) that
\[
\Phi = (X^T X) \dagger (X^T X) \sim Q(X^T X) \dagger (X^T X)Q^T = Q\Phi Q^T.
\]

Hence, the subspace is invariant under rotation and it is possible to establish that the matrix \(\Phi\) is a random projection that projects into a subspace sampled uniformly (i.e., Haar measure) from the Grassmannian \(G(m, n)\). The next result is from [24, Lemma 5.3.2] and it states that the norm \(\|\Phi \beta\|_2\) of the projection of \(\beta\) into this \(n\)-dimensional subspace concentrates around \(\sqrt{m}\|\beta\|_2\).

**Lemma 7 (Vershynin [24, Lemma 5.3.2]):** Let \(\beta \in \mathbb{R}^m\) be a vector and \(\Phi \in \mathbb{R}^{m \times m}\) be a projection from \(\mathbb{R}^m\) onto a random \(m\)-dimensional subspace uniformly sampled from \(G(m, n)\). Then,

1. \(\mathbb{E}[\|\Phi \beta\|_2^2] = \frac{n}{m} \|\beta\|_2^2\);
2. There exist a constant \(c\), such that with probability greater than \(1 - 2 \exp(-ct^2n)\), we have:

\[
(1 - t) \sqrt{\frac{n}{m}} \|\beta\|_2 \leq \|\Phi \beta\|_2 \leq (1 + t) \sqrt{\frac{n}{m}} \|\beta\|_2.
\]  

(19)

The following proposition will also be needed.

**Proposition 8:** We have that:
\[
\left\|\hat{\beta} - \Phi \beta\right\|_2 \leq \frac{1}{\sqrt{n}} \sqrt{\|\Sigma\|_2 \|\epsilon\|_2}.
\]

**Proof:** From (12) and the triangular inequality:
\[
\left\|\hat{\beta} - \Phi \beta\right\|_2 \leq \frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}} \left\|\Sigma^\dagger X^T \epsilon\right\|_2.
\]

(21)

In turn, we have that
\[
\left\|\frac{1}{\sqrt{n}} \Sigma^\dagger X^T \epsilon\right\|_2 \leq \left\|\frac{1}{\sqrt{n}} \Sigma^\dagger X^T \right\|_2 \|\epsilon\|_2.
\]

Here, \(\|\cdot\|_2\) is used to denote the operator norm, such that for a matrix \(A \in \mathbb{R}^{m \times n}, \|A\|_2 = \max_{i \leq n} \sqrt{\lambda_i(A^T A)}\). Hence:
\[
\left\|\frac{1}{\sqrt{n}} \Sigma^\dagger X^T \right\|_2 = \max_i \sqrt{\lambda_i(\Sigma \Sigma^\dagger \Sigma)} = \|\Sigma^\dagger\|_2,
\]

where the second equality follows from direct use of the property \(\Sigma \Sigma^\dagger \Sigma = \Sigma^\dagger\) and the fact that \(\Sigma^\dagger\) is positive semidefinite.

Equipped with the proposition and the lemmas we are now ready to prove the theorem. Note that Lemma 3 could provide another possible route to prove similar results and (maybe) tighter bounds. Nonetheless, here, we choose to use Proposition 8 for two reasons: 1) it can easily be combined with non-asymptotic results; 2) the argument applied above can be extended for any other \(p\)-norms. We also point out that an analogous procedure could be used to provide a lower bound on \(\|\hat{\beta}\|_2\).

**Proof of Theorem 5:** From Proposition 8 we have \(\|\hat{\beta}\|_2 \leq \|\Phi \beta\|_2 + \sqrt{\frac{m}{n}} \sqrt{\lambda_{\max}(\Sigma^\dagger)}\). Due to the fact that the noise is sub-Gaussian, a straightforward application of Theorem 3.1.1 from [24] implies that \(\frac{|c|}{\sqrt{n}} - a^2 \leq t\) with probability greater than \(1 - 2 \exp(ct^2n)\). This together with Lemma 6 yields the desired upper bound for the second term. The first term can be bounded using Lemma 7. The result follows.

**E. \(\ell_p\) Adversaries**

We now turn to the study of \(\ell_p\)-adversarial attacks when \(p \neq 2\). As in (15), let \(q\) be the complement of \(p\). The following well-known relationship between vector norms will be useful in our development.

**Lemma 9 (Relationship between vector \(p\)-norms):** Let \(p\) and \(q\) be values in the range \([1, \infty]\) and \(\hat{\beta} \in \mathbb{R}^m\). Assume that \(q > p\), then:
\[
\|\hat{\beta}\|_q \leq \|\hat{\beta}\|_p \leq m^{1/p - 1/q} \|\hat{\beta}\|_q.
\]

(22)

The leftmost inequality follows from an application of Minkowski’s inequality and the rightmost from an application of Hölder’s inequality.

The asymptotic results from Lemma 4 to compute \(R\) and \(L_2\) can now be used in conjunction with the above inequalities to find the upper and lower bounds on the adversarial risk. Hence, for any \(1 \leq p < 2\), the upper bound is the same as the upper bound obtained for \(\ell_2\) attacks. However, there is a new multiplicative term in the lower bound. For instance, the \(\ell_1\) adversarial risk is bounded by
\[
R(\hat{\beta}) + \frac{\delta^2}{m} \|\hat{\beta}\|_2^2 \leq R_{\text{adv}}(\hat{\beta}) \leq \left(\sqrt{R(\hat{\beta})} + \delta \|\hat{\beta}\|_2\right)^2.
\]

(23)

On the other hand, for \(\ell_p\)-adversarial attacks with \(p > 2\), we obtain an asymptotic upper bound that grows with \(m^{1 - \frac{1}{p}}\). As an example, for \(\ell_\infty\) attacks,
\[
R(\hat{\beta}) + \delta^2 \|\hat{\beta}\|_2^2 \leq R_{\text{adv}}(\hat{\beta}) \leq \left(\sqrt{R} + \delta \sqrt{m} \|\hat{\beta}\|_2\right)^2.
\]

(24)
In Fig. 4(a) we illustrate the bounds obtained in this way. We note that the $\ell_\infty$-adversarial risk follows the upper bound closely. Moreover, Lemma 9 implies that
\[ \|\hat{\beta}\|_2 \leq \|\hat{\beta}\|_1 \leq \sqrt{m}\|\hat{\beta}\|_2. \] (25)

From Fig. 4(b) we see that the $\ell_1$-norm of the estimated parameter seems to follow the upper bound closely. At the same time, the adversarial risk is also close to the upper bound closely. Next, we provide some insight into this observation, by following the same steps used in the non-asymptotic analysis of the $\ell_2$ parameter norm.

Lemma 7 show how $\|\Phi\beta\|_2$ concentrate around $\sqrt{\frac{m}{n}}\|\beta\|_2$. One might wonder whether similar concentration inequalities can be obtained also for the $\ell_1$-norm. In Fig. 5 we illustrate the experiments for both $\|\Phi\beta\|_1$ and $\|\Phi\beta\|_2$. The first plot just illustrates the results known for the $\ell_2$-norm from Lemma 7; the second plot suggests that the $\ell_1$-norm of the projection has mean $c\sqrt{n}\|\beta\|_2$. From the experiments, we also estimate that $c \approx 0.8$. We state this result as a conjecture.

**Conjecture 10:** Let $\beta \in \mathbb{R}$ and $\Phi$ be a projection from $\mathbb{R}^m$ onto a random $n$-dimensional subspace uniformly sampled from the Grassmannian manifold $G(m, n)$. Then, $\mathbb{E}_\Phi[\|\Phi\beta\|_1] = c\sqrt{n}\|\beta\|_2$.

Since $\|\Phi\beta\|_1$ concentrates around its mean, the conjecture also implies a high probability statement. Indeed, in Appendix A it is proved that with probability greater than $1 - \exp(-t^2n)$
\[ \|\Phi\beta\|_1 - \mathbb{E}_\Phi[\|\Phi\beta\|_1] < t\sqrt{n}\|\beta\|_2. \] (26)

Combined with this result, the conjecture implies that with probability greater than $1 - \exp(-t^2n)$,
\[ \|\Phi\beta\|_1 - c\sqrt{n}\|\beta\|_2 < t\sqrt{n}\|\beta\|_2. \] (27)

We point out that this result does have important consequences for the study of overparameterized models. We obtain smoother models by increasing the number of parameters for the $\ell_2$-norm, but the conjecture implies that this does not happen for the $\ell_1$-norm. Indeed, it implies that with high probability
\[ (c - t)\sqrt{n}\|\beta\|_2 < \|\Phi\beta\|_1 < (c + t)\sqrt{n}\|\beta\|_2. \] (28)

Now, using exactly the same argument as in Lemma 8, we obtain
\[ \|\Phi\beta\|_1 - \|\epsilon\|_2 \sqrt{\frac{m}{n}}\|\Sigma\|_2 \leq \|\hat{\beta}\|_1 \leq \|\Phi\beta\|_1 + \|\epsilon\|_2 \sqrt{\frac{m}{n}}\|\Sigma\|_2 \]

The conjecture implies that $\|\Phi\beta\|_1 = \Theta(\sqrt{n})\|\beta\|_2$, and Lemma 6 implies that the second term is $\Theta(\sqrt{n})\|\beta\|_2$. Hence, for a sufficiently large signal-to-noise ratio ($r > \sigma$) it follows from the conjecture that $\|\hat{\beta}\|_1 = \Theta(\sqrt{n})$. Since we obtained in the Theorem 5 that $\|\hat{\beta}\|_2 = \Theta(\sqrt{n})$, it follows that $\|\hat{\beta}\|_1 = \Theta(\sqrt{n})\|\hat{\beta}\|_2$, which is consistent with the results we are experimentally observing.

**F. Scaling**

The scaling of variables plays an important role in the analysis. Assume that a given $\hat{\beta}$ was estimated and that the corresponding model prediction is $\hat{\beta}^T\tilde{x}$. By simply redefining the input variable as $\tilde{x} = \frac{1}{\eta}x$ we could obtain an equivalent model $\hat{\beta}^T\tilde{x}$ that, for $\hat{\beta} = \eta\hat{\beta}$, would yield exactly the same predictions.

Notice that while the standard risk $R$ for this new, rescaled, model is exactly the same as the first, the norm of the estimated
parameter $\| \tilde{\beta} \|_q$ is $\eta$ times larger. The adversarial risk is not the same for the two models, as an inspection of (15) reveals. The difference is because the relative magnitude of the adversarial disturbance is larger in the second model (even though it is the same in absolute value).

Since we are interested in the impact that the number of parameters $m$ has on adversarial robustness, we will let the scaling factor depend on this parameter, i.e. $\eta = \eta(m)$. The next proposition motivates two choices of scaling. The proof is provided in the Supplementary Material.

**Proposition 11:** Let $x$ be an isotropic random vector, $\mathbb{E}[\| x \|_2^2] = m$. Additionally, if $x$ is a sub-Gaussian random vector, then $\mathbb{E}[\| x \|_\infty] = \Theta(\sqrt{\log(m)})$.

Hence, $\eta(m) = \sqrt{m}$ or $\eta(m) = \sqrt{\log m}$ are both quite natural choices of the scaling factor. They render, respectively, the expected $\ell_2$- and $\ell_\infty$-norms of the input vector constant as the number of features $m$ varies.

Assume that the inputs are redefined as $\tilde{x}_i = \frac{1}{\eta(m)} x_i$. A quick inspection of (11) reveals that the estimated parameter is $\tilde{\beta} = \eta(m) \hat{\beta}$. The risk $R(\tilde{\beta})$ does not change by the transformation, but the expected squared norm of the parameter does, $\| \tilde{\beta} \|_2^2 = (\eta(m))^2 \| \hat{\beta} \|_2^2$. Hence, when $\eta(m) = \sqrt{\log m}$, it follows from (18) that

$$\| \tilde{\beta} \|_2^2 \approx \log m \left( r^2 \frac{1}{m/n} + \sigma^2 \frac{1}{m/n - 1} \right).$$

Here, the logarithmic term changes slowly compared to the linear term in the denominator. Hence, the result is similar to what was obtained without any scaling. On the other hand, the square root scaling $\eta(m) = \sqrt{m}$ yields:

$$\| \tilde{\beta} \|_2^2 \approx r^2 n + \sigma^2 \frac{1}{n - 1/m}.$$  \hfill (30)

Here, the parameter norm does not go to zero. Instead, it approaches a constant as $m \to \infty$. The behavior is illustrated in Fig. S.2 in the Supplementary material. One interesting consequence of (30) is that $\| \tilde{\beta} \|_2$ grows with the number of training datapoints. Hence, the $\ell_2$-adversarial performance degrades as we add more training data points.

The situation is even more pathological in the case of $\ell_\infty$-adversarial attacks. In Fig. 6, we show the $\ell_\infty$-adversarial risk as a function of $m$ when the input is scaled by $\eta(m) = \sqrt{m}$ and $\eta(m) = \sqrt{\log m}$. We also provide the upper bound obtained from Lemma 4 and the inequality in (24). The behavior of $\ell_\infty$-adversarial attacks is governed by $\| \tilde{\beta} \|_1$ (see (8)) and we have $\| \tilde{\beta} \|_1 = \Theta(\sqrt{m}) \| \hat{\beta} \|_2$ (recall Section IV-E). Hence, (30) and (29) yield, respectively, $\| \tilde{\beta} \|_1 = \Theta(\sqrt{m})$ and $\| \tilde{\beta} \|_1 = \Theta(\log m)$, which explain the behavior observed in the figure.

We end this section with another interpretation of rescaling. Let us consider the following change of variables $\tilde{x} = \frac{1}{\eta(m)} x$ and $\tilde{\beta} = \eta(m) \hat{\beta}$. The next proposition states that this can also be interpreted as keeping the input and parameter constant while re-scaling the adversarial disturbance region $\delta$ by a factor $\eta(m)$.

**Proposition 12:** Let

$$\text{adv-error}(x, \beta, \delta) = \max_{\| \Delta x \|_2 \leq \delta} \left\{ y - (x + \Delta x)^T \beta \right\}^2,$$

then we have that:

$$\text{adv-error} \left( \frac{x}{\eta(m)}, \eta(m) \beta, \delta \right) = \text{adv-error} (x, \beta, \eta(m) \delta).$$

**G. Latent Space Model**

For the model studied in the previous section, it is in general possible to achieve where the test error is smaller in the underparameterized region than in the overparameterized region. Thus, it could be argued that the lack of $\ell_\infty$-adversarial robustness in the overparameterized region should not be a problem in practice. Let us now illustrate a different data generation procedure for which we have better performance in the overparameterized regime and where the performance is continuously improved as more features are added. However, it is still possible to observe that the $\ell_\infty$-adversarial robustness degrades indefinitely with the number of features (recall Fig. 1).

We consider a data model where the features $x$ are noisy observations of a lower-dimensional subspace of dimension $d$. A vector in this latent space is represented by $z \in \mathbb{R}^m$. This vector is indirectly observed via the features $x \in \mathbb{R}^m$ according to

$$x = Wz + u,$$

where $W$ is an $m \times d$ matrix, for $m \geq d$. We assume that the responses are described by a linear model in this latent space

$$y = \theta^T z + \xi,$$

where $\xi \in \mathbb{R}$ and $u \in \mathbb{R}^m$ are mutually independent noise variables. Moreover, $\xi \sim \mathcal{N}(0, \sigma_x^2)$ and $u \sim \mathcal{N}(0, I_m)$. We consider the features in the latent space to be isotropic and normal

$$z_i = \mathcal{N}(0, I_d).$$

To facilitate the analysis, we choose $W$ such that its columns are orthogonal, $W^TW = \frac{m}{d} I_d$, where the factor $\frac{m}{d}$
setting where we also include the standard risk in the same plot (scaling $\eta(m) = \sqrt{m}$). Additional results are presented in Supplementary Material, Section E.

V. ADVERSARIAL TRAINING AND REGULARIZATION

Empirical risk minimization (ERM) is a popular paradigm for estimating predictive models [25]. In the last section, the model was trained to minimize the empirical risk $\hat{R}(\beta)$ but evaluated according to an adversarial criteria. One natural idea to obtain models that are more robust to adversarial attacks is to instead minimize the empirical adversarial risk,

$$\hat{R}_p^{\text{adv}}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \max_{|\Delta x_i|_p \leq \delta} (y_i - (\hat{x}_i + \Delta x_i)^T \hat{\beta})^2. \quad (33)$$

This method is commonly called adversarial training [21].

In this section, we use Lemma 1 to develop a convex formulation of adversarial training for linear regression problems. With this tool in hand, we explore the effect of adversarial training on how the model robustness changes the number of features. We also compare it to ridge regression,

$$\hat{R}_\text{ridge}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \delta \| \beta \|_2^2, \quad (34)$$

and lasso,

$$\hat{R}_\text{lasso}(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \delta \| \beta \|_1. \quad (35)$$

A. Adversarial Training Using Convex Programming

Using Lemma 1 we can show the convexity of the adversarial risk (defined in (3)).

**Proposition 13:** For $p \in [1, \infty]$, $R_p^{\text{adv}}(\beta)$ is convex in $\beta$.

**Proof:** Let $\gamma \in [0, 1]$, for $q \in [1, \infty]$, $\| \beta \|_q$ is a norm and from the triangular inequality, we have that:

$$\| y \| \gamma \hat{\beta}_1 + (1 - \gamma) \hat{\beta}_1 \|_q \leq \gamma \| \hat{\beta}_1 \|_q + (1 - \gamma) \| \hat{\beta}_2 \|_q. \quad (36)$$

Moreover,$$
|y_0 - x_0^T (\gamma \hat{\beta}_1 + (1 - \gamma) \hat{\beta}_1)| \leq \gamma |y_0 - x_0^T \hat{\beta}_1| + (1 - \gamma) |y_0 - x_0^T \hat{\beta}_2|.
$$

Hence, $h(\beta) = |y_0 - x_0^T \beta| + \| \beta \|_q$ is convex and, also, $h(\beta) \geq 0$ for all $\beta$. Now, since $g(x) = x^2$ is convex and non-decreasing for $x \geq 0$, the composition $g \circ h$ is convex – See Boyd et al. [26, Section 3.2.4]; moreover, the expected value of a convex function is also convex [26, Section 3.2.1] and it follows that the right-hand side of (5) is convex.

The results obtained for the adversarial risk are also valid for the empirical adversarial risk. Hence, it follows from Lemma 1 that:

$$\hat{R}_p^{\text{adv}}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left( |y_i - x_i^T \beta| + \delta \| \beta \|_q \right)^2, \quad (37)$$

and that it is convex. The above expression can be entered into a standard convex modeling language to obtain the adversarial training solution. In the numerical examples that follows we use CVXPY [27] to train the model.
B. Overparameterized Models: Latent Space Feature Model

In this example, we consider artificially generated data from the latent space feature model described in Section IV. The same experiment for the isotropic feature model is provided in the Supplementary Material, Section F. In Section IV, we saw the unfortunate effect that if the input variables scale with \( \eta(m) = \sqrt{m} \) (which corresponds to keeping \( \mathbb{E}[\|x\|_2^2] \) constant as we vary the number of features \( m \)) we observe that \( \|\hat{\beta}\|_1 \) grows indefinitely with \( m \) when \( \hat{\beta} \) was estimated using the minimum-norm solution. We also showed how this makes the \( \ell_\infty \)-adversarial risk grow indefinitely with the number of features (i.e., Fig. 1).

Let us now investigate if the same effect can be observed for models trained with ridge regression, lasso and adversarial training. In Fig. 8 we show the norm \( \|\hat{\beta}\|_1 \) in these cases. For ridge regression the parameter norm grows with \( O(m) \) regardless of how large the regularization parameter \( \delta \) is. We notice that \( \ell_2 \)-adversarial training has a similar behavior for \( \delta \) smaller then a certain threshold, in these cases it displays curves similar to ridge regression that grow with \( O(m) \). However, for sufficiently large values of the regularization parameter, the parameter norm of the solution is zero for all values of \( m \).

For lasso, we see that the parameter norm goes to zero for overparameterized models with sufficiently large \( m \). Looking at lasso as a bi-objective optimization problem helps interpret this behavior: as the number of features \( m \) increases, the scaling affects the two objectives differently and the objective of keeping \( \|\hat{\beta}\|_1 \) starts to be prioritized over the objective of keeping the square training error low, the more \( m \) is increased. Interestingly, the \( \ell_\infty \)-adversarial training seems to behave in a very similar way.

In Fig. 9 we provide the adversarial test error for models trained with ridge regression, lasso and adversarial training, respectively. As expected by our analysis of \( \|\hat{\beta}\|_1 \), lasso and \( \ell_\infty \)-adversarial training yield solutions that do not deteriorate indefinitely. We believe this observation adds to our discussion about the role of scaling. It highlights the fact that, even in the case of a mismatch between disturbance and how the input scales with the number of variables (i.e., \( \mathbb{E}[\|x\|_2^2] \) constant while we evaluate it under an \( \ell_\infty \)-adversary) it is still possible to avoid brittleness by considering a type of regularization that acts under the right norm.

VI. DISCUSSION

A. Related Work

a) Adversarial attacks: The study of adversarial attacks predates the widespread use of deep neural networks [28], [29]. An overview of earlier work is provided by Biggio et al. [30]. Nonetheless, the susceptibility of high-performance neural networks to adversarial attacks gave this framework higher visibility [4]. The framework of adversarial attacks has generated striking examples of the vulnerability of such models to very small input perturbations. Small changes in the input can cause a substantial drop in performance in otherwise state-of-the-art models, see for instance [4], [7], [10], [17], [18], [19].

b) Robustness and the role of high-dimensionality: The conflict between robustness and high-performance models is explored by Tsipras et al. [6] and Ilyas et al. [7]. Indeed, one of the examples we give for the worst-case scenario of the \( \ell_\infty \)-adversarial error is motivated by an idea presented in [6]. Moreover, simple examples where high-dimensional inputs yield easy-to-construct adversarial examples are abundant in the literature [5]. The analysis of the robustness of more general nonlinear models, such as neural networks is provided in [31] and extended in Bubeck et al. [9]. They show how over-parametrization can be a recipe for robustness. An alternative view is provided in [32], [33], where it is shown that ReLU
neural networks can be made vulnerable since what these models learn is locally very similar to random linear functions. As we mentioned in the introduction, our work tries to reconcile these somewhat conflicting views in the context of linear models.

c) Double-descent: The double-descent performance curve has been experimentally observed for a variety of machine learning models, such as random Fourier features, random forests, shallow networks, transformers, convolutional networks and nonlinear ARX models; and for datasets obtained in diverse contexts, including image classification, natural language processing datasets and the identification of nonlinear dynamical systems [20], [34], [35], [36], [37]. For instance, we illustrate the double-descent phenomena using random Fourier features in the Supplementary Material Fig. S.1. Theoretical models for such phenomena are also often pursued: Bartlett et al. [8] derive non-asymptotic bounds for linear regression models using concentration inequalities. In [35] the authors draw connections with the physical phenomena of “jamming” in a class of glassy systems. Deng et al. [38] characterize logistic regression test error using the Gaussian min-max theorem. Muthukumar et al. [39] provides bounds on the risk.

Random matrix theory has been a useful tool for studying statistical phenomena. The framework and its potential for explaining and studying neural networks have been the focus of recent work [40], [41], [42]. It has also been a powerful tool in producing theoretical models for the double-descent phenomenon [11], [43], [44], [45], [46]. In our study of linear regression with random covariates, we make direct use of the asymptotic results obtained by Hastie et al. [11].

d) Analysis of adversarial attacks: Theoretical analysis of models under adversarial attacks is currently a rather popular topic. Hassani et al. [14] obtain exact asymptotics for random feature regressions. Also in the context of random feature regression and D’Amour et al. [47] provide asymptotics based on Mei et al. [44]. They study a scenario where the adversarial attack is constrained to not change the risk. The theoretical model is used to explain how underspecification might present a challenge in deployment and is backed by experiments. Diochnos et al. [48] consider the adversarial risk when the instances are uniformly distributed over \{0, 1\}^n. Dohmatob [49] provides a no-free-lunch theorem where it is shown that any classifier can be adversarially fooled with high probability when the perturbations are slightly greater than the natural noise level in the problem.

e) Adversarial attacks in linear models: While a lot of current research focuses on adversarial examples for deep learning models, there is a growing body of work that study the fundamental properties of adversarial attacks in linear models. There is a sound reason for this focus: linear models allow for analytical analysis while still reproducing phenomena of interest.

Bhagoji et al. [50] obtain optimal transport-based lower bounds for adversarial examples in classification problems. They consider Gaussian data and norm-bounded adversaries. Taheri et al. [12] derived asymptotics for adversarial training in binary classification. Moreover, Javanmard et al. [13] provide asymptotics for adversarial attacks in linear regression. Javanmard et al. [51] study classification settings.

These asymptotics are often used to gain insight into the effect of adversarial training and adversarial robustness. Javanmard et al. [13] studies the trade-off between adversarial risk and standard risk. Note that Javanmard et al. [51] studies how overparametrization affects robustness to perturbations in the input and Min et al. [15] studies how the size of the dataset affects adversarial performance. We corroborate their observation that the adversarial performance might degrade as the size of the dataset increases.

The derivation of exact asymptotics is an impressive technical development, but we point out that it is not always trivial to gain insight from these results. The asymptotics obtained often do not have closed-form expressions and require the solution of either polynomials or integral equations. Here, we advocate a simpler approach: approximating the adversarial risk using terms that often appear in other contexts. We believe that this is a powerful tool to gain insight into the problem, providing extra flexibility for quickly navigating between different setups. We use (8) and reduce the analysis to the risk and the parameter norm. Lemma 1 is an important tool for this analysis. We point out that Xing et al. [52] proved a version of Lemma 1 specialized to the Gaussian case and \ell_2-norm and that Javanmard et al. [13] state a version of the same lemma for the \ell_2-norm.

Fig. 9. Adversarial \(\ell_\infty\)-risk. Top: ridge and lasso. Bottom: \(\ell_2\) and \(\ell_\infty\) adversarial training. On the y-axis we show the \(\ell_\infty\)-adversarial risk for models obtained by different training methods. On the x-axis we have the ratio between the number of features \(m\) and the number of training datapoints \(n\). The error bars give the median and the 0.25 and 0.75 quantiles obtained from numerical experiments (6 realizations). We use \(\delta = 0.01\) both during inference (to compute the adversarial risk) and during the adversarial training, as in (33). We also use \(\delta = 0.03\) for lasso and ridge regression, see (34) and (35). Here the shaded region indicates the upper and lower bounds obtained empirically from (8). The empirical risk in the test and the parameter norm are obtained from the experiments.
Rademacher complexity analysis: Close to our work is that of Yin et al. [16], which provided an analysis of $\ell_\infty$-adversarial attack on linear classifiers based on the Rademacher complexity. Their Theorem 1 resembles (8): we show that the adversarial risk and $(\text{risk} + \delta^2 \|\widehat{\beta}\|_q^2)$ are upper and lower bounded by constant factors. Yin et al. [16] prove a similar relation for the adversarial Rademacher complexity of a linear classifier. In their proof, they use a reformulation of the adversarial loss similar to that of Lemma 1 but for classifiers. Similar to our results, they showed an unavoidable dimension dependence unless the weight vector has a bounded norm. On the one hand, our work extends their results for regression. On the other hand, by analysing general $\ell_p$-adversarial attacks and different covariate scalings, we studied a wide variety of possible behaviors that they do not observe by focusing only on $\ell_\infty$-adversarial attacks and training.

B. Connections to Neural Networks

The success of deep neural networks is an important reason for digging deeper into the properties of overparameterized models. Here, however, we study the phenomenon in linear models. To motivate the relevance of our study also for neural networks, we appeal to a recent line of work that has pointed out a direct connection between linear models and more complex models such as neural networks [53], [54], [55], [56]. The idea can be understood in simple terms. Let the parameterized function $f(\cdot; \theta)$ denote the neural network, where $\theta \in \mathbb{R}^m$ denote the vector of parameters. Assume that the number of parameters is very large and that training the neural network moves each of them just by a small amount w.r.t. its initialization $\theta_0$. A linearization of the model around $\theta_0$, yields

$$f(x; \theta) \approx f(x; \theta_0) + \nabla_{\theta} f(x; \theta_0)^T \delta,$$

(38)

where $\delta = \theta - \theta_0$. Hence, the problem can be approximated by an affine problem that could be solved using linear regression. Indeed, it can be established that as the neural network becomes infinitely wide the training of the neural network actually becomes solving a problem similar to that in (38).

Linear models are also a natural setup to study adversarial attacks. Indeed, while there was initial speculation that the highly nonlinear nature of deep neural networks was the cause of its vulnerabilities to adversarial attacks [4], the idea was later dismissed and the vulnerabilities can be observed already in purely linear settings [10].

C. Extension to Nonlinear Models

A question that naturally comes to mind is if parts of this analysis can be generalized to nonlinear settings. For that, define the adversarial risk associated with a given function $f$ by:

$$R_{p}\text{adv}(f) = \mathbb{E}_{x_0, y_0} \left[ \max_{\|\Delta x_0\|_p \leq \delta} (y_0 - f(x_0 + \Delta x_0))^2 \right].$$

(39)

Let $L(f)$ be the Lipschitz constant of $f$, i.e., $|f(x_1) - f(x_2)| \leq L(f) \|x_1 - x_2\|$ for all $x_1, x_2$. The idea of using the Lipschitz constant as a proxy for robustness is quite standard and a common procedure to obtain robust models is to jointly optimize the risk $R(f)$ and the Lipschitz constant $L(f)$, see e.g. [57]. Indeed, an analysis equivalent to the one used in the proof of Lemma 1 yields

$$R_{p}\text{adv}(f) \leq \mathbb{E}_{x_0, y_0} \left[ (y_0 - f(x_0)) + \delta L(f) \right]^2.$$  

(40)

In this case, equality does not necessarily hold. Proposition 2 was used in the proof for the linear case, but there is not an obvious equivalent in the nonlinear case. Hence, instead of the approximation (8) we would only have an upper bound (and no lower bound).

VII. Conclusion

In this paper, we focus on the behavior of the adversarial risk as we change the number of features. Our analysis is based on the fact that $\ell_p$-adversarial risk is between 1 and 2 times $(\text{risk} + \delta^2 \|\widehat{\beta}\|_q^2)$, where $q$ is the complementary norm to $p$. Hence, the behavior of the adversarial risk can be studied by analysing these two components. We use such results to analyse the role of high-dimensionality in the performance of linear models under adversarial attacks.

On the other hand, the result implies that $\ell_2$-adversarial risk presents a double-descent curve when both the risk and $\|\widehat{\beta}\|_2$ present such behavior. We use asymptotic results from Hastie et al. [11] to illustrate a double-descent curve in the adversarial risk for models with features randomly generated with isotropic, equicorrelated and spiked (i.e., latent space model) covariance matrices.

On the other hand, we focus on the analyse of cases where the risk is small but the $\ell_p$-adversarial risk grows with the number of features. In our setup, as a direct consequence of the aforementioned approximation of the adversarial risk, this happens if and only if $\delta \|\widehat{\beta}\|_q \to \infty$ as the number of features $m \to \infty$. In order to analyse the term $\delta \|\widehat{\beta}\|_q$,

- we use non-asymptotic analysis for the norm of the estimated parameter obtained by the minimum-norm solution. For Gaussian covariates, we show that: $\|\widehat{\beta}\|_2 = O(1/m)$, while $\|\widehat{\beta}\|_1 = O(1)$.

- we show that for isotropic covariates, if $\delta \propto E[\|x\|_2]$ then $\delta = O(m)$. Furthermore, for sub-Gaussian covariates we have that if $\delta \propto E[\|x\|_\infty]$, then $\delta = O(\sqrt{\log(m)})$.

We combine the two results to show examples that are robust to $\ell_2$-adversarial attacks but can be made vulnerable to $\ell_\infty$-adversarial attacks as we increase the number of features. The most pathological results are usually obtained in a mismatched situation, where we apply an $\ell_\infty$-adversarial attack with magnitude $\delta \propto E[\|x\|_2]$. In this case, we have shown that the adversarial risk can be made arbitrarily large (i.e., the model is arbitrarily vulnerable to an adversary) as the number of features grows. Such a mismatched setup (with a $\ell_\infty$-adversarial attack with scaling proportional to the $\ell_2$-norm) is present in influential examples such as those in [6], [10], and the mismatch often appears hidden in the argument. Finally, we also provided a convex optimization formulation of adversarial training and studied similarities between adversarial training and parameter-shrinking methods.
ACKNOWLEDGMENT
The authors would like to thank Dave Zachariah for very fruitful discussions.

REFERENCES
[1] P. Rajpurkar, E. Chen, O. Barjere, and E. J. Topol, “AI in health and medicine,” Nature Med., vol. 28, pp. 31–38, Jan. 2022.
[2] R. Hassain and S. Zeadally, “Autonomous cars: Research results, issues, and future challenges,” IEEE Commun. Surv. Tut., vol. 21, no. 2, pp. 1275–1313, Apr.–Jun. 2019.
[3] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, “Unsolved problems in ML safety,” Jun. 2022. [Online]. Available: http://arxiv.org/abs/2201.01196.
[4] J. Bruna et al., “Intriguing properties of neural networks,” in Proc. 2nd Int. Conf. Learn. Representations, 2014.
[5] J. Gilmer et al., “Adversarial spheres,” Sep. 2018. [Online]. Available: http://arxiv.org/abs/1801.02774
[6] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Ma, “Robustness may be at odds with accuracy,” in Proc. Int. Conf. Learn. Representations, 2019, Art. no. 23.
[7] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” in Proc. 33rd Int. Conf. Neural Inf. Process. Syst., 2019, pp. 125–136. [Online]. Available: http://arxiv.org/abs/1905.02175
[8] Y. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, “Benign overfitting in linear regression,” Proc. Nat. Acad. Sci., vol. 117, no. 48, pp. 30063–30070, Apr. 2020.
[9] S. Bubeck and M. Sellke, “A universal law of robustness via isoperimetry,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2021. [Online]. Available: http://arxiv.org/abs/2105.12806
[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proc. 3rd Int. Conf. Learn. Representations, 2015.
[11] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, “Surprises in high-dimensional ridgeless least squares interpolation,” Ann. Statist., vol. 50, no. 2, pp. 949–986, 2022. doi: 10.1214/21-AOS2133.
[12] H. Taheri, R. Pedarsani, and C. Thrampoulidis, “Asymptotic behavior of adversarial training in binary classification,” IEEE Int. Symp. Inf. Theory (ISIT), vol. 127-132, 2022. doi: 10.1109/ISIT50566.2022.9834717.
[13] A. Javanmard, M. Soltanolkotabi, and H. Hassani, “Precise tradeoffs in adversarial training for linear regression,” in Proc. 33rd Conf. Learn. Theory, 2020, pp. 2034–2078. [Online]. Available: http://proceedings.mlr.press/v125/javanmard20a.html
[14] H. Hassani and A. Javanmard, “The curse of overparameterization in adversarial training: Precise analysis of robust generalization for random features regression,” Jan. 2022. [Online]. Available: http://arxiv.org/abs/2201.05149
[15] Y. Min, L. Chen, and A. Karbasi, “The curious case of adversarially robust models: More data can help, double descend, or hurt generalization,” in Proc. 37th Conf. Uncertainty Artif. Intell., 2021, pp. 129–139. [Online]. Available: http://arxiv.org/abs/2002.11080
[16] D. Yin, R. Kannan, and P. Bartlett, “Rademacher complexity for adversarially robust generalization,” in Proc. 36th Int. Conf. Mach. Learn. Res., 2019, pp. 7085–7094. [Online]. Available: https://proceedings.mlr.press/v97/yin19b.html
[17] A. B. Ash and C. Doles-Dale, “Probability and Measure Theory,” 2nd ed.
[18] R. B. Ash and C. Doles-Dale, Probability and Measure Theory, 2nd ed.
[19] M. Geiger et al., “Scaling description of generalization with number of parameters in deep learning,” J. Stat. Mechanics: Theory Exp., vol. 2020, no. 2, pp. 949–986, 2020. doi: 10.1088/1742-5468/ab77d2.
[20] A. Daniely and H. Shacham, “Most ReLU networks suffer from \ell^2 adversarial perturbations,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2020, pp. 6629–6636. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/497f766a1816251905e88a0df5a25-Abstract.html
[21] S. Bubeck, Y. Li, and D. Nagaraj, “A law of robustness for two-layers neural networks,” Proc. Mach. Learn. Res., Conf. Learn. Theory, 2021, pp. 804–820.
[22] H. Taheri, R. Pedarsani, and C. Thrampoulidis, “Asymptotic behavior of adversarial training in binary classification,” IEEE Int. Symp. Inf. Theory (ISIT), vol. 127-132, 2022. doi: 10.1109/ISIT50566.2022.9834717.
[23] A. Javanmard, M. Soltanolkotabi, and H. Hassani, “Precise tradeoffs in adversarial training for linear regression,” in Proc. 33rd Conf. Learn. Theory, 2020, pp. 2034–2078. [Online]. Available: http://proceedings.mlr.press/v125/javanmard20a.html
[24] H. Hassani and A. Javanmard, “The curse of overparameterization in adversarial training: Precise analysis of robust generalization for random features regression,” Jan. 2022. [Online]. Available: http://arxiv.org/abs/2201.05149
[25] Y. Min, L. Chen, and A. Karbasi, “The curious case of adversarially robust models: More data can help, double descend, or hurt generalization,” in Proc. 37th Conf. Uncertainty Artif. Intell., 2021, pp. 129–139. [Online]. Available: http://arxiv.org/abs/2002.11080
[48] D. Diochnos, S. Mahloujifar, and M. Mahmoody, “Adversarial risk and robustness: General definitions and implications for the uniform distribution,” in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 10380–10389. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/34835e5ec0489e5c394b028ec4e153e1-Paper.pdf

[49] E. Dohmatob, “Generalized no free lunch theorem for adversarial robustness,” in *Proc. 36th Int. Conf. Mach. Learn. Res.*, 2019, pp. 1646–1654.

[50] A. N. Bhagoji, D. Cullina, and P. Mittal, “Lower bounds on adversarial robustness from optimal transport,” in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 7498–7510. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/02bfb6214e264555e3412283e817deca-Article.html

[51] A. Javanmard and M. Soltanolkotabi, “Precise statistical analysis of classification accuracies for adversarial training,” *Ann. Statist.*, vol. 50, no. 4, pp. 2127–2156, 2022, doi: 10.1214/22-AOS2180.

[52] Y. Xing, Q. Song, and G. Cheng, “On the generalization properties of adversarial training,” in *Proc. 24th Int. Conf. Artif. Intell. Statist.*, 2021, pp. 505–513. [Online]. Available: https://proceedings.mlr.press/v130/xing21b.html

[53] Z. Allen-Zhu, Y. Li, and Z. Song, “A convergence theory for deep learning via over-parameterization,” in *Proc. 36th Int. Conf. Mach. Learn.*, 2019. [Online]. Available: http://arxiv.org/abs/1811.03962

[54] L. Chizat and F. Bach, “On the global convergence of gradient descent for over-parameterized models using optimal transport,” in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 3040–3050. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/a1af586ca9540d0579299ec361d726-Paper.pdf

[55] S. Du, J. Lee, H. Li, L. Wang, and X. Zhou, “Gradient descent finds global minima of deep neural networks,” in *Proc. 36th Int. Conf. Mach. Learn. Res.*, 2019, pp. 1675–1685. [Online]. Available: http://proceedings.mlr.press/v97/du19c.html

[56] A. Jacot, F. Gabriel, and C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018. [Online]. Available: http://arxiv.org/abs/1806.07572

[57] M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. J. Pappas, “Efficient and accurate estimation of Lipschitz constants for deep neural networks,” in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 11427–11438. [Online]. Available: http://arxiv.org/abs/1906.04893

Antônio H. Ribeiro received the B.Sc, M.Sc., and Ph.D. degrees in electrical engineering from the Federal University of Minas Gerais, Belo Horizonte, Brazil, in 2016, 2017 and 2020, respectively. He is currently a Postdoctoral Associate with Uppsala University, Uppsala, Sweden. His research interests include system theory, signal processing, and machine learning and its applications to health care and medicine. He was the recipient of the best Ph.D. thesis in engineering and physical sciences by the Federal University of Minas Gerais, and Benzelius Award from the Royal Society of Sciences in Uppsala, Sweden, in 2022.

Thomas B. Schön (Senior Member, IEEE) received the B.Sc. degree in business administration and economics, the M.Sc. degree in applied physics and electrical engineering, and the Ph.D. degree in automatic control from Linköping University, Linköping, Sweden, in January 2001, in September 2001, and in February 2006, respectively. He is currently a Beijer Professor of artificial intelligence with the Department of Information Technology, Uppsala University, Uppsala, Sweden. He has held visiting positions with the University of Cambridge, Cambridge, U.K., University of Newcastle, Callaghan NSW, Australia, and Universidad Técnica Federico Santa María, Valparaíso, Chile. In 2018, he was elected to The Royal Swedish Academy of Engineering Sciences and The Royal Society of Sciences at Uppsala. His research interests include developing new algorithms and mathematical models capable of learning and acting based on data. He was the recipient of the Tage Erlander prize for natural sciences and technology in 2017 and Armborg prize in 2016, both awarded by the Royal Swedish Academy of Sciences, Automatica Best Paper Prize in 2014, best Ph.D. thesis award by The European Association for Signal Processing (EURASIP) in 2013, and Best Teacher Award at the Institute of Technology, Linköping University in 2009. He is a Fellow of the ELLIS Society.