Supplementary Material

High-dimensional analysis of the adversarial error: double-descent behavior and model-size dependency

Table of Contents

A Vector and matrix norms

A.1 Vector norms

Vector p-norms. Let $x \in \mathbb{R}^n$ and $p \in \mathbb{R}$, $p \leq 1$, the p-norm of this vector is defined as:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, \tag{24}$$

Moreover,

$$\|x\|_\infty = \max_{1\leq i \leq n} |x_i|. \tag{25}$$

The following relation between the p-norms can be established.

Lemma 6 (Relation between vector p-norms)**.** *Let p and q be norms in the range* $[1, \infty]$ *and* $x \in \mathbb{R}^m$. *And, assume that* $q > p$, *then:*

$$\|x\|_q \leq \|x\|_p \leq m^{1/p-1/q}\|x\|_q \tag{26}$$

In the case of $q = \infty$, *we have* $\|x\|_\infty \leq \|x\|_p \leq m^{1/p}\|x\|_\infty$.

The proof for the above result is provided in `https://math.stackexchange.com/q/218046` and `https://math.stackexchange.com/q/245052`. The leftmost inequality follows from an application of the Minkowski inequality and the rightmost one from an application of the Hölder inequality.

A.2 Matrix norms

Induced matrix norms. Let W be a m by d matrix.We define the (p_1, p_2) induced norm of this matrix as:

$$\|W\|_{(p_1,p_2)} = \sup\left\{\frac{\|Wx\|_{p_1}}{\|x\|_{p_2}} \middle| \text{ for } x \in \mathbb{R}^d\right\} \tag{27}$$

We use the simplified notation $\|W\|_p$ to denote $\|W\|_{(p,p)}$. The following relation between induced matrix norms follows directly from Lemma 6.

Lemma 7 (Relation between induced matrix norms)**.** *Be $W \in \mathbb{R}^{m\times d}$ a matrix, assume that and $q_1 \geq p_1$ and that $q_2 \geq p_2$, then*

$$\frac{1}{d^{\frac{1}{p_2}-\frac{1}{q_2}}}\|W\|_{(q_1,q_2)} \leq \|W\|_{(p_1,p_2)} \leq m^{\frac{1}{p_1}-\frac{1}{q_1}}\|W\|_{(q_1,q_2)} \tag{28}$$

Proof. The result follows directly from applying Lemma 6 for the pairs (p_1, q_1) and (p_2, q_2), and rearranging the inequalities, so it can be used together with the definition of the induced matrix norm. □

B Bounds on the adversarial error

B.1 Lipshitz continuous functions

In this section, we start with a general bound for arbitrary Lipshitz continuous functions and use it to derive the specific bounds studied along the main text. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function mapping inputs x to the output y. Assume this is a Lipshitz continuous function with Lipshitz constant $L_p(f)$ in the ℓ_p norm, i.e.,

$$|f(x_1) - f(x_2)| \leq L_p(f)\|x_1 - x_2\|_p. \tag{29}$$

The next theorem gives an upper bound on the adversarial risk

$$R_p^{\text{adv}}(f) = \mathbb{E}_{x_0,y_0}\left[\max_{\|\Delta x_0\|_p\leq\delta}\left(y_0 - f\left(x_0 + \Delta x_0\right)\right)^2\right]. \tag{30}$$

The bound depends on the risk $R(f) = \mathbb{E}_{x_0,y_0}\left[y_0 - f(x_0\right]$ and on the Lipshitz constant $L_p(f)$.

Theorem 8 (Adversarial error of Lipshitz continuous functions)**.** *If f is Lipshitz continuous with Lipshitz constant $L_p(f)$ the adversarial risk is upper bounded by,*

$$R_p^{adv}(f) \leq \left(\sqrt{R(f)} + L_p(f)\right)^2. \tag{31}$$

Proof. Let $\Delta\hat{y}_0 = f(x_0 + \Delta x_0) - f(x_0)$ and $e_0 = y_0 - f(x_0)$, the adversarial risk (14) can be rewritten as

$$R_p^{\text{adv}}(f) = \mathbb{E}_{x_0,y_0}\left[\max_{\|\Delta x_0\|_p\leq\delta}\left((\Delta\hat{y}_0)^2 - 2e_0\Delta\hat{y}_0\right)\right] + R(f). \tag{32}$$

Now, the Lipschtz continuity of f yields that when $\|\Delta x_0\|_p \leq \delta$ it follows that $|\Delta\hat{y}_0| \leq L_p(f)\delta$. The term inside the maximum is then upper bounded by $\delta^2 L_p(f)^2 + 2\delta|e_0|L_p(f)$. It follows that:

$$R_p^{\text{adv}} \leq \delta^2 L_p(f)^2 + 2\delta L_p(f)\mathbb{E}_{x_0,y_0}\left[|e_0|\right] + R(f). \tag{33}$$

In turn, $\mathbb{E}_{x_0,y_0}\left[|e_0|\right] \leq \sqrt{R(f)}$, by direct application of Jensen's inequality. The results follows. □

B.2 Neural network models

Neural networks are Lipshitz continuous. Let define a simple fully connected neural network with depth l. Let us denote $z_{(i)} \in \mathbb{R}^{m_i}$ the activation of the i-th layer. For a input $x \in \mathbb{R}^d$, the model can be defined recursively as

$$z_{(i)} = \begin{cases} x & i = 0; \\ \phi\left(\hat{W}_i z_{(i-1)}\right) & \text{otherwise.} \end{cases} \tag{34}$$

For which the predicted output is $\hat{y} = \hat{\beta}^{\mathsf{T}} z_{(l-1)}$. Here $\hat{W}_i \in \mathbb{R}^{m_i \times m_{i-1}}$ is the weight matrix and ϕ is an element-wise activation function which satisfies Assumption 4, i.e. it is Lipshitz continuous with constant L_ϕ.

One point that is worth clarifying is that there is no ambiguity in the definition of the Lipshitz constant of the function ϕ. That is, once Lipshitz continuity is defined for the scalar case, as in Assumption 4, it is uniquely defined for all p-norms.

Proposition 9. *If ϕ is a element-wise function. Then $x, y \in \mathbb{R}$ satisfy the inequality:*

$$|\phi(x) - \phi(y)| \leq L_\phi |x - y| \tag{35}$$

if and only if, when applied to $x, y \in \mathbb{R}^m$, it satisfy

$$\|\phi(x) - \phi(y)\|_p \leq L_\phi \|x - y\|_p \tag{36}$$

for any p norm $p \geq 1$ or $p = \infty$ for any size m.

Proof. Assume (35). Then, for $p = \infty$ we have:

$$\|\phi(x) - \phi(y)\|_\infty = \sup_{i \in 1, \cdots, m} |\phi(x^i) - \phi(y^i)| \leq \sup_{i \in 1, \cdots, m} L_\phi |x^i - y^i| = L_\phi \|x - y\|_\infty \tag{37}$$

Now, for $1 \leq p < \infty$ we obtain:

$$\|\phi(x) - \phi(y)\|_p = \left(\sum_{i=1}^{m} |\phi(x^i) - \phi(y^i)|^p \right)^{\frac{1}{p}} = \left(\sum_{i=1}^{m} L_\phi^p |x^i - y^i|^p \right)^{\frac{1}{p}} = L_\phi \|x - y\|_p \tag{38}$$

This shows necessity. The sufficiency follows directly from the definition. □

Next, we establish the Lipshitz constant of this neural network in terms of the other constants. Indeed, the proposition follows from the definitions we presented so far.

Proposition 10. *Let $f : x \mapsto \hat{y}$ be the function defined recursively by Eq.* (34) *with activation function ϕ satisfying Assumption 4. Then it satisfies the inequality* (29) *with constant:*

$$L_p(f) = (L_\phi)^{(l-1)} \|\hat{\beta}\|_q \prod_{i=1}^{l-1} \|\hat{W}_i\|_{(r_i, r_{i-1})} \tag{39}$$

Where q and all the r_i belong to the range $[1, \infty]$. *Moreover, $r_0 = p$ and q satisfies* $1/q + 1/r_{l-1} = 1$.

One note about the above proposition is that we introduce auxiliary variables r_i, $i = 1, \cdots l - 1$, and that these variables are not completely specified. Indeed, these variables can be chosen to be any value in the range $[1, \infty]$ and it is, thus, convenient to just choose it in such a way as to make the bound as tight as possible. We will go back to this idea latter, when combining the upper bounds with the matrix and vector norm inequalities. Now the proof:

Proof. Fix $r_0 = p$, and choose r_i, $i = 1, \cdots l - 1$ in the range $[1, \infty]$. Let us denote $L_{(i)}$,

$$L_{(i)} = \begin{cases} 1 & i = 0; \\ L_\phi L_{(i-1)} \|\hat{W}_i\|_{(r_i, r_{i-1})} & 0 < i < l - 1 \end{cases}$$

It follows from the Lipschtz continuity of ϕ and the definition of matrix induced norm that:

$$\|z_{(i)} - \tilde{z}_{(i)}\|_{r_i} \leq L_{(i)} \|x - \tilde{x}\|_p \tag{40}$$

where $z_{(i)}$ and $\tilde{z}_{(i)}$ are the activation functions for two different inputs x and $\tilde{x}$. Finally, Hölder inequality yields that, for q satisfying $1/q + 1/r_{l-1} = 1$,

$$|\hat{y} - \tilde{y}| \leq \|\hat{\beta}\|_q \|z_{(l-1)} - \tilde{z}_{(l-1)}\|_{r_i} \tag{41}$$

where $\hat{y}$ and $\tilde{y}$ would be the neural network outputs corresponding to x and $\tilde{x}$, respectively. The result follows. □

The next proposition follows from the previous one, but only depends on gives a constant that on the l_2 vector and induced matrix norms.

Proposition 11. *Let $f : x \mapsto \hat{y}$ be the function defined recursively by Eq.* (34) *with activation function ϕ satisfying Assumption 4. Then it satisfies the inequality* (29) *with constant:*

$$L_p(f) = J_p(b)\,(L_\phi)^{(l-1)}\,\|\hat{\beta}\|_2 \prod_{i=1}^{l-1} \|\hat{W}_i\|_2 \tag{42}$$

Where $b = \min_i(m_i)$ and J_p is defined as

$$J_p(b) = \begin{cases} 1 & p \leq 2; \\ b^{1/2-1/p} & 2 < p < \infty; \\ b^{1/2} & p = \infty. \end{cases} \tag{43}$$

As we mentioned in Section 4, b is the bottleneck, i.e., the layer with the least dimension in the neural network. The proof follows directly from from combining Proposition 10 with Lemma 7.

B.3 Linear regression

The Corollary 3 (in the main text) follows from Theorem 1 combined with Lemma 6. Moreover, we highlight that the upper bound of Theorem 1 (also introduced in the main text), can be thought as a special case of Theorem 8. The lower bound, however, is specific to the linear scenario and relies on the following proposition.

Proposition 12. *Given p in $[1, \infty]$, $\beta \in \mathbb{R}^m$, define q such that $1/p + 1/q = 1$. Let $\Delta x \in \mathbb{R}^m$ and denote its components by Δx_i. If $p = \infty$, define*

$$\Delta x_i = 1 \quad \textit{for every } i.$$

If $p = 1$, define Δx as

$$\Delta x_i = \frac{s_i}{\sum_i s_i} \quad \textit{for} \quad s_i = \begin{cases} 1 & \textit{if } \beta_i = \max_i \beta_i \\ 0 & \textit{otherwise} \end{cases}$$

Finally, if $1 < p < \infty$,

$$\Delta x_i = |\beta_i|^{q/p}$$

Then $|\beta^\mathsf{T} \Delta_x| = \|\beta\|_p \|\Delta x\|_q$

The proposition establishes that we can find a value of Δx such that equality in the Hölder inequality will hold. Moreover, it gives a constructive way to find such Δx. There is no equivalent for arbitrarily nonlinear functions or neural networks. As we add intermediary layers it becomes necessary to prove that such values are reachable (in the last layer of the neural network) to obtain the lower bound. In the case of random features, it might be possible to establish this with some given probability, but we do not do it in this work and leave the analysis for future work.

B.4 Random feature regression

Corollary 5 follows from Proposition 11 and Theorem 8 for a neural network with depth 1 and fixed weight matrix W_1.

C Asymptotic results

C.1 Linear regression with random covariates

In this section, we give the asymptotic values for $R(\hat{\beta})$ and $\|\hat{\beta}\|_2$ in linear regression. The results were originally proved by (Hastie et al., 2019), but adapted here to our notation and settings.

Assume the scenario described in Section 2.1. The training data has been generated as in (7), linearly with additive noise,

$$(x_i, \epsilon_i) \sim P_x \times P_\epsilon, \qquad y_i = x_i^\mathsf{T} \beta + \epsilon_i, \qquad i = 1, \cdots, n,$$

for which P_ϵ is a distribution in $\mathbb{R}$ such that $\mathbb{E}\left[\epsilon_i\right] = 0$ and $\mathbb{V}\left[\epsilon_i\right] = \sigma^2$ and independent from x_i. Moreover $\mathbb{E}\left[x_i\right] = 0$ and its covariance matrix will be denoted by: $\mathbb{E}\left[x_i x_i^\mathsf{T}\right] = \Sigma$. We denote by X the m by n matrix containing the training sample x_i in its i column. And, in a slight abuse of notation, we denote by $\epsilon \in \mathbb{R}^m$ the vector containing all noise terms ϵ_i that appear during training as elements.

The next lemma presents closed-form expressions for the expectation in terms of the problem matrices. Here we are taking the expectation in terms of all the noise terms ϵ, conditioned on the values of the covariates X. We denote such expectation by $\mathbb{E}_\epsilon$.

Lemma 13 (Bias-variance decomposition)**.** *We define* $\hat{\Sigma} = \frac{1}{n} X^\mathsf{T} X$, $\Phi = \hat{\Sigma}^\dagger \hat{\Sigma}$ *and* $\Pi = I - \Phi$. *Where* Φ *and* Π *are orthogonal projectors:* Π *is the projection into the null space of* X *and* Φ*, into the row space of* X*. Then:*

$$\mathbb{E}_\epsilon\left[R(\hat{\beta})\right] = \beta^\mathsf{T} \Pi \Sigma \Pi \beta + \frac{\sigma^2}{n} tr(\hat{\Sigma}^\dagger \Sigma) + \sigma^2$$

Similarly, for $q = 2$*:*

$$\mathbb{E}_\epsilon\left[\|\hat{\beta}\|_2\right] = \beta^\mathsf{T} \Phi \beta + \frac{\sigma^2}{n} tr(\hat{\Sigma}^\dagger).$$

Proof. **Proof for $\|\hat{\beta}\|_2$:** From Eq. (7) and Eq. (8) it follows that:

$$\hat{\beta} = \underbrace{(X^\mathsf{T} X)^\dagger X^\mathsf{T} X}_{\Phi} \beta + \underbrace{(X^\mathsf{T} X)^\dagger}_{\frac{1}{n}\hat{\Sigma}^\dagger} X^\mathsf{T} \epsilon. \tag{44}$$

Hence, since $\hat{\Sigma}$ is symmetric:

$$\hat{\beta}^\mathsf{T} \hat{\beta} = \beta^\mathsf{T} \Phi^\mathsf{T} \Phi \beta + \frac{1}{n^2} \epsilon^\mathsf{T} X \hat{\Sigma}^\dagger \hat{\Sigma}^\dagger X^\mathsf{T} \epsilon, \tag{45}$$

where the first term is equal to $\beta^\mathsf{T} \Phi \beta$ since, Φ is a *orthogonal projector* i.e., $\Phi^T = \Phi$ and $\Phi\Phi = \Phi$.

Now, since the second term is a scalar it is equal to its trace. Using the fact that the trace is invariant over cyclic permutations,

$$\epsilon^\mathsf{T} X \hat{\Sigma}^\dagger \hat{\Sigma}^\dagger X^\mathsf{T} \epsilon = \text{tr}\left\{\hat{\Sigma}^\dagger X^\mathsf{T} \epsilon \epsilon^\mathsf{T} X \hat{\Sigma}^\dagger\right\}. \tag{46}$$

From the assumption the noise samples are independent and have variance σ^2, we have $\mathbb{E}_\epsilon\left[\epsilon\epsilon^\mathsf{T}\right] = \sigma^2 I$, where I is the identity matrix. Since we can swap the trace and the expectation we obtain

$$\mathbb{E}_\epsilon\left[\hat{\beta}^\mathsf{T} \hat{\beta}\right] = \beta^\mathsf{T} \Phi \beta + \frac{1}{n^2} \text{tr}\left\{\hat{\Sigma}^\dagger X^\mathsf{T} \underbrace{\mathbb{E}_\epsilon\left[\epsilon\epsilon^\mathsf{T}\right]}_{\sigma^2 I} X \hat{\Sigma}^\dagger\right\}.$$

And the results follow from the definition of $\hat{\Sigma}^\dagger$ and the property of pseudoinverse: $\hat{\Sigma}^\dagger \hat{\Sigma} \hat{\Sigma}^\dagger = \hat{\Sigma}^\dagger$

Proof for R: Now,

$$R(\hat{\beta}) = \mathbb{E}_{x_0, y_0}\left[(\beta^\mathsf{T} x_0 - y_0)^2\right] = (\beta - \hat{\beta})^\mathsf{T} \Sigma (\beta - \hat{\beta}) + \sigma^2 \tag{47}$$

From (44), it follows that:

$$\beta - \hat{\beta} = \underbrace{(I - \Phi)}_{\Pi} \beta + \frac{1}{n} \hat{\Sigma}^\dagger X^\mathsf{T} \epsilon.$$

where, again Π is a orthogonal projector i.e., $\Pi^T = \Pi$ and $\Pi\,\Pi = \Pi$. We can then compute the close formula $\mathbb{E}_\epsilon\left[R(\hat{\beta})\right]$ using the same procedure.

□

Isotropic features. We start with the simplest case, the features are independent and identically distributed (i.i.d.), i.e. $\Sigma = I$. The above result is equivalent to Theorem 1 and to Corollary 1 from (Hastie et al., 2019). They do provide a prove for it in the appendix of their work.

Theorem 14. *Assume that x_i are i.i.d. and has a moment of order greater then 8 that is finite. Assume that $\|\beta\|_2^2 = r^2$. Then, as $p, n \to \infty$ $p/n \to \gamma$, it holds almost surely that:*

$$\mathbb{E}_\epsilon\left[R(\hat{\beta})\right] \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma}, \gamma < 1, \\ r^2(1 - \frac{1}{\gamma}) + \sigma^2 \frac{1}{\gamma - 1}, \gamma > 1. \end{cases} \tag{48}$$

For $q = 2$, it holds almost surely that:

$$\mathbb{E}_\epsilon\left[\|\hat{\beta}\|_2^2\right] \to \begin{cases} r^2 + \sigma^2 \frac{\gamma}{1-\gamma}, \gamma < 1, \\ r^2 \frac{1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1}, \gamma > 1. \end{cases} \tag{49}$$

Equicorrelated features. Now we consider, the case the features are ρ-equicorrelated. That is, Σ is such that its (i, j)-th entry is:

$$\Sigma_{i,j} = \begin{cases} 1 & i = j, \\ \rho & i \neq j \end{cases} \tag{50}$$

and $x = \Sigma^{1/2} z$ for z composed of i.i.d. features. The following result holds:

Theorem 15. *Assume that x_i is generated as described above for a value of z with zero mean, unitary variance and bounded moment of order greater then. Moreover, assume it has a moment of order greater then 4 that is finite. Assume that $\|\beta\|_2^2 = r^2$. Then, as $p, n \to \infty$ $p/n \to \gamma$, it holds almost surely that:*

$$\mathbb{E}_\epsilon\left[R(\hat{\beta})\right] \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma}, \gamma < 1, \\ r^2(1-\rho)(1 - \frac{1}{\gamma}) + \sigma^2 \frac{1}{\gamma - 1}, \gamma > 1. \end{cases} \tag{51}$$

For $q = 2$, it holds almost surely that:

$$\mathbb{E}_\epsilon\left[\|\hat{\beta}\|_2^2\right] \to \begin{cases} r^2 + \sigma^2 \frac{\gamma}{(1-\gamma)(1-\rho)}, \gamma < 1, \\ r^2 \frac{1}{\gamma} + \sigma^2 \frac{1}{(\gamma - 1)(1-\rho)}, \gamma > 1. \end{cases} \tag{52}$$

The above theorem can be obtained using the results from Hastie et al. (2019). It relies on the nice properties of the equicorreleted matrix. The next proposition gives the eigenvalues and eigenvectors of such a matrix. We use $\vec{\mathbf{1}}$ to denote a vector with dimension m and with all its entries equal to 1.

Proposition 16. *Let $\Sigma \in \mathbb{R}^{m \times m}$ be an equicorrelated matrix, with its entries defined as in* (50)*. Let us denote by s_i, $i = 1, \cdots, n$ its eigenvalues and by v_i the corresponding eigenvectors. Then $s_1 = 1 + (m-1)\rho$ and $s_i = (1 - \rho)$ for every $i \neq 1$. Moreover, $v_1 = \frac{1}{\sqrt{m}}\vec{\mathbf{1}}$ and v_i for $i \neq 1$ is such that the sum of its entries is equal to zero, i.e., $v_i^\mathsf{T}\vec{\mathbf{1}} = 0$.*

Proof of Theorem 15. The proof for the assymptotic of $R(\hat{\beta})$ follows from (Hastie et al., 2019) Corollary 7. The proof for the asymptotic of $\|\hat{\beta}\|_2$ follows from (Hastie et al., 2019) Corollary 2. In the case of equicorrelated features, using Hastie et al. (2019) notation $H = G$ and, moreover, $dH = \delta_{1-\rho}$. Together with the definition of c_0 yields, this yields $c_0 = \frac{1}{\gamma(\gamma-1)(1-\rho)}$ which when replaced in Corollary 2 yields the result.[1] □

C.2 Random feature regression

Here we present the result from (Mei and Montanari, 2019) adapted to our notation and setting. The result depends on additional assumptions on the activation function. The following condition is satisfied by ReLU, sigmoid functions and other commonly used activation functions.

[1] There is a small mistake in (Hastie et al., 2019) Corollary 2: The term r^2 should have appeared multiplying the integral for the overparametrized case.

Assumption 17 (Activation function)**.** *Assume that the activation function $\phi : \mathbb{R} \to \mathbb{R}$ is weakly differentiable, with weak derivative ϕ'. Moreover, assume that $|\phi(x)|$ and $|\phi'(x)|$ are upper bounded by $c_0 e^{c_1|x|}$ for some finite constants c_0 and c_1. We denote:*

$$\mu_0 \equiv \mathbb{E}_G\left[\phi(G)\right], \quad \mu_1 \equiv \mathbb{E}_G\left[G\phi(G)\right], \quad \mu_\star^2 \equiv \mathbb{E}_G\left[\phi(G)^2\right] - \mu_0^2 - \mu_1^2$$

where the expectation is with respect to $G \sim \mathrm{N}(0,1)$. And, assume that $0 < \mu_0^2, \mu_1^2, \mu_\star^2 < \infty$. And use ζ to denote $\zeta \equiv \frac{\mu_1}{\mu_\star}$.

Next, we expand on the conditions on the data generating process. An introduction to the conditions were explained in Section 3, here we present a complete list of assumptions.

Assumption 18 (Data generation process)**.** *Assume that the output is computed as*

$$y_i = y_0 + x_i^\mathsf{T}\theta + g(x_i) + \epsilon_i, \quad \textit{for } i = 1, \cdots, n,$$

where ϵ_i is sampled independently from a distribution P_ϵ in $\mathbb{R}$ such that $\mathbb{E}\left[\epsilon_i\right] = 0$ and $\mathbb{V}\left[\epsilon_i\right] = \sigma^2$ and $\mathbb{E}\left[\epsilon_i^4\right] < \infty$. The input x_i is sampled from a Gaussian distibution $\mathcal{N}(0, I)$, independently of ϵ_i. Here g is a centered Gaussian process such that

$$\mathbb{E}_g\left[g(x_1)g(x_2)\right] = \Sigma_d(x_1^\mathsf{T}x_2) \tag{53}$$

we require that this Σ_d satisfy $\mathbb{E}_x\left[\Sigma_d(x)\right] = 0$, $\mathbb{E}_x\left[x\Sigma_d(x)\right] = 0$ and that $\lim_{d\to\infty}\Sigma_d(1) = F^2$. Moreover, we denote $\tau^2 = \sigma^2 + F^2$ and $r^2 = \|\theta\|_2^2$.

Assumption 19 (Model and parameter estimation)**.** *We assume the nonlinear model $\beta^\mathsf{T}\phi(Wx_i)$. Where W is set in advanced, with each of its entries $w_{i,j}$ is sampled from a Gaussian distribution with zero mean and variance $1/\sqrt{d}$. The parameter β from* ***ridge regression*** *with regularization parameter λ, i.e., by minimizing*

$$\frac{1}{n}(\beta^\mathsf{T}\phi(Wx_i) - y)^2 + \frac{m}{d}\lambda\|\beta\|_2^2,$$

Now we are ready to enunciate the theorem:

Theorem 20 (Mei and Montanari (2019))**.** *If the above assumptions on the activation function, the data generation process, and the model and parameter estimation are satisfied, then:*

$$\mathbb{E}_{W,X,\epsilon,g}\left[\left|R(\hat\beta) - r^2\mathcal{B}(\phi,\gamma_0,\gamma_1,\lambda) + \tau^2\mathcal{V}(\phi,\gamma_0,\gamma_1,\lambda) + \tau^2\right|\right] \to 0,$$
$$\mathbb{E}_{W,X,\epsilon,g}\left[\left|\|\hat\beta\|_2^2 - r^2\mathcal{A}_1(\phi,\gamma_0,\gamma_1,\lambda) + \tau^2\mathcal{A}_2(\phi,\gamma_0,\gamma_1,\lambda)\right|\right] \to 0.$$

Where, $\mathcal{B}$, $\mathcal{V}$, $\mathcal{A}_1$ and $\mathcal{A}_2$ are defined by the fractions:

$$\mathcal{B} = \frac{B}{D}, \quad \mathcal{V} = \frac{V}{D}, \quad \mathcal{A}_1 = \frac{A_1}{D}, \quad \mathcal{A}_2 = \frac{A_2}{D}$$

for B, V, A_1 and A_2 and D polynomials,

$$\begin{aligned}
D \equiv &-\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2\\
&+ (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\\
B \equiv &\psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2\\
V \equiv &\chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2\\
A_1 \equiv &-\chi^2\left(\chi\zeta^4 - \chi\zeta^2 + \psi_2\zeta^2 + \zeta^2 - \chi\psi_2\zeta^4 + 1\right)/\mu_\star^2\\
A_2 \equiv &\chi^2\left(\chi\zeta^2 - 1\right)\left(\chi^2\zeta^4 - 2\chi\zeta^2 + \zeta^2 + 1\right)/\mu_\star^2
\end{aligned}$$

In turn, the polynomials depend on the function $\chi = \chi(\phi,\gamma_0,\gamma_1,\lambda)$. And on the parameter ζ of the activation function. For convenience, and for making our notation closer to that of Mei and Montanari (2019), we also used

$$\psi_1 = \gamma_1/\gamma_0, \quad \psi_2 = 1/\gamma_0 \tag{54}$$

Finally, χ is defined to be:

$$\chi \equiv \nu_1\left(\mathrm{i}\left(\psi_1\psi_2\lambda\right)^{1/2}/\mu_\star\right)\cdot\nu_2\left(\mathrm{i}\left(\psi_1\psi_2\lambda\right)^{1/2}/\mu_\star\right) \tag{55}$$

where $i = \sqrt{-1}$ and $\nu_1, \nu_2\ \mathbb{C}_+ \to \mathbb{C}_+$ are the uniquely defined functions by that satisfy the following conditions: $(i)\nu_1, \nu_2$ are analytic on $\mathbb{C}_+$; (ii) For $\Im(\xi) > 0$, $\nu_1(\xi), \nu_2(\xi)$ satisfy the following equations

$$\begin{aligned} \nu_1 &= \psi_1 \left(-\xi - \nu_2 - \tfrac{\zeta^2 \nu_2}{1-\zeta^2 \nu_1 \nu_2}\right)^{-1}; \\ \nu_2 &= \psi_2 \left(-\xi - \nu_1 - \tfrac{\zeta^2 \nu_1}{1-\zeta^2 \nu_1 \nu_2}\right)^{-1}, \end{aligned} \tag{56}$$

(iii) $(\nu_1(\xi), \nu_2(\xi))$ is the unique solution of these equations with $|\nu_1(\xi)| \leq \psi_1/\Im(\xi), |\nu_2(\xi)| \leq \psi_2/\Im(\xi)$ for $\Im(\xi) > C$, with C a sufficiently large constant.

A note about L_1 and in probability convergence. In the above theorem, we establish L_1 convergence, while in the main text we enunciate convergence in probability. Indeed, L_1 convergence implies convergence in probability, this follows from an application of the Markov inequality. Hence, the above theorem is indeed a stronger version of what is enunciated in the main text.

Finding the solution of Eq. (56). We choose to present the definition of ν_1 and ν_2 as in Mei and Montanari (2019) to make it easier to see the equivalence with their result. We note, however, that it is possible to rewrite the equations in a way that might be easier to solve. Some algebraic manipulation will show that $\psi_1 + \nu_1 \xi = \nu_2 \xi + \psi_2$, hence, only one equation needs to be solved. Moreover, this equation actually can be written as this 4-th order polynomial. The companion code to do this works provides an implementation of this procedure.

The relation between multivariate Gaussian and the uniform over the sphere distribution. We define x and W independent entries, each with zero mean Gaussian distribution. Mei and Montanari (2019), on the other hand, consider x and the rows of W are obtained from an uniform distribution over the sphere. Indeed, the first distribution converges weakly to to the second as $m \to \infty$. The equivalence between the two situations, then follows from an application of Portmanteau theorem.

A comparison between the way the theorem is enunciated here and in Mei and Montanari (2019) will show some small differences. Some constants do not appear in some of the equations due to different definitions of the variance. This does not alter the result and was more natural in our setting.

D Worst-case scenario

D.1 Lower bound on the l_1 parameter norm

Here, we prove Eq. (12). Again, $\beta = \left(\frac{r}{\sqrt{m}}, \cdots, \frac{r}{\sqrt{m}}\right)$ and $\hat{\beta}$ is the parameter estimated from the dataset. By the triangular inequality, $\|\beta\|_1 \leq \|\hat{\beta}\|_1 + \|\hat{\beta} - \beta\|_1$. Lemma 6 together with the definition of ℓ_1 norm yields

$$r\sqrt{m} \leq \|\hat{\beta}\|_1 + \|\hat{\beta} - \beta\|_2. \tag{57}$$

Next, we relate $\|\hat{\beta} - \beta\|_2$ with $(\beta - \hat{\beta})^\mathsf{T} \Sigma (\beta - \hat{\beta})$, for Σ representing the covariance matrix of x. The matrix Σ is symmetric, hence it allows for the spectral decomposition

$$\Sigma = \sum_{i=1}^{m} s_i v_i v_i^\mathsf{T},$$

where s_i are its eigenvalues and v_i the eigenvectors of the matrix Σ. Hence:

$$(\beta - \hat{\beta})^\mathsf{T} \Sigma (\beta - \hat{\beta}) = \sum_{i=1}^{m} s_i \left((\beta - \hat{\beta})^\mathsf{T} v_i\right)^2 \geq s_{\min} \|\beta - \hat{\beta}\|_2^2,$$

where we use $s_{\min}$ to denote the least eigenvalue and apply the property that the eigenvectors $\{v_i\}_{i=1}^m$ form an orthonormal basis, and hence preserve the ℓ_2 norm, to establish the leftmost inequality. Since, $R(\hat{\beta}) = (\beta - \hat{\beta})^\mathsf{T} \Sigma (\beta - \hat{\beta}) + \sigma^2$ (i.e., Eq. (47)) it follows that:

$$r\sqrt{m} \leq \|\hat{\beta}\|_1 + \sqrt{\frac{R(\hat{\beta})}{s_{\min}}} \tag{58}$$

For the case of an equicorrelated covariance matrix, with its entries defined as in Eq. (50) we know the close formula for the eigenvalues and eigenvectors. Using Proposition 16 it them follows that for sufficiently large m the least eigenvalue is $s_{\min} = 1 - \rho$ which could be used in (58) to obtain the bound. In this case, however, it is possible to obtain an even tighter bound, by using the exact values for the eigen-decomposition. The analysis then show that, for sufficiently large n,

$$r\sqrt{m} \leq \|\hat{\beta}\|_1 + \sqrt{R(\hat{\beta})}.$$

D.2 Lower bound on the l_q parameter norm

In the same setting, we can establish for any q norm a similar result. Indeed, the same mechanics yields

$$r(m)^{\frac{1}{q}-\frac{1}{2}} \leq \|\hat{\beta}\|_q + \sqrt{R(\hat{\beta})}.$$

Which can be then combined with Theorem 1 to analyse $R_p^{\text{adv}}(\hat{\beta})$ for arbitrary p. Hence, $R_p^{\text{adv}}(\hat{\beta}) = \Omega(m^{\frac{1}{2}-\frac{1}{p}})$ for $p > 2$ and the upper bound we establish in Corollary 3 indeed cannot be tightened in terms of its dependency with m for $p > 2$. Moreover, adversarial attacks with $p > 2$ have a behavior qualitatively similar to the case of ℓ_∞ attacks, growing polynomially with the number of features in the worst-case scenario we just presented.

D.3 Equivalence with Tsipras et al. (2019)

Here, we expand on the example presented in Section 2.3. We establishing the equivalence with the example from (Tsipras et al., 2019). The formulation is intuitive and gives a different perspective about the worst-case scenario we just presented. Moreover, it was this example that made us start looking into the relation between number of features and adversarial performance to begin with.

The formulation from Tsipras et al. (2019), considers that the covariates are conditioned on the output. The formulation presented in Section 2, on the other hand, presents the data model differently: considering the output conditioned on covariates. In the Gaussian case, it is easy to connect both, as we show next.

As in (Tsipras et al., 2019), let $y \sim \mathcal{N}(0, 1)$ and assume that the j-th feature x_j conditioned on y is also normally distributed $P(x_j \mid y) \sim \mathcal{N}(y, 1)$. Assume all features are conditionally independent and let $x = (x_1, \cdots, x_m)$ be a vector containing all of them.

Here x and y are jointly Gaussian. Hence, $P(y \mid x)$ is Gaussian variable with conditional mean $\mathbb{E}[y|\, x] = \frac{1}{\sqrt{m}}\beta^\mathsf{T} x$ for $\beta = (1/\sqrt{m}, \cdots, 1/\sqrt{m})$ and with conditional variance $\mathbb{E}\big[y^2\big|\, x\big] = 0$. Moreover, x is marginally Gaussian with $\mathbb{E}\left[x_j\right] = 0$. Hence, $x \sim \mathcal{N}(0, \Sigma)$ for a covariance matrix Σ, with entries,

$$\Sigma_{i,j} = \mathbb{E}\left[(x_j)^\mathsf{T} x_k\right] = \mathbb{E}\left[\mathbb{E}\big[(x_j)^\mathsf{T} x_k\big|\, y\big]\right] = 1. \tag{59}$$

In, the above computation, when $j = k$ the expectation inside is just the conditional variance of x^j. On the other hand, if $j \neq k$ the variables inside the inner expectation are independent and can be decomposed into $(\mathbb{E}\left[x_j|y\right]^\mathsf{T} \mathbb{E}\left[x_k|y\right]) = y^2$.

The example by Tsipras et al. (2019) is a classification example, hence direct comparison is not possible. So in a first moment, we presented an adaptation of it to the regression setting in a way we believe preserves the most interesting points in their construction. Then, in the second moment, we show the equivalence with the data generating model from Eq. 7, i.e., by showing it can be written as $y = \beta^\mathsf{T}\left(\frac{x}{\sqrt{m}}\right)$ for equicorrelated covariates.