Precise Accuracy / Robustness Tradeoffs in Regression: Case of General Norms

Elvis Dohmatob [1] **Meyer Scetbon** [2]

Abstract

In this paper, we investigate the impact of test-time adversarial attacks on linear regression models and determine the optimal level of robustness that any model can reach while maintaining a given level of standard predictive performance (accuracy). Through quantitative estimates, we uncover fundamental tradeoffs between adversarial robustness and accuracy in different regimes. We obtain a precise characterization which distinguishes between regimes where robustness is achievable without hurting standard accuracy and regimes where a tradeoff might be unavoidable. Our findings are empirically confirmed with simple experiments that represent a variety of settings. This work covers feature covariance matrices and attack norms of any nature, extending previous works in this area.

1. Introduction

Machine learning models are known to be highly sensitive to small perturbations known as *adversarial examples* (Szegedy et al., 2013), which are often imperceptible to humans. While various strategies such as adversarial training (Madry et al., 2018) can mitigate this vulnerability empirically, the situation remains highly problematic for many safety-critical applications like autonomous vehicles or health, and motivates a better theoretical understanding of what mechanisms may be causing this.

From a theoretical perspective, the case of classification is rather well-understood. (Tsipras et al., 2019) showed that adversarial robustness could be at odds with accuracy. The hardness of classification under adversarial attacks has been crisply characterized (Bhagoji et al., 2019; Bubeck et al., 2018). In the special case of linear classification, explicit lower-bounds on sample complexity have been obtained (Schmidt et al., 2018; Bhattacharjee et al., 2021). However, the case of linear regression is relatively understudied. Recently, in the setting of Euclidean-norm attacks on isotropic features, (Javanmard et al., 2020) has initiated a theoretical study of a possible tradeoff between standard risk (a.k.a. generalization error) and adversarial risk (a.k.a. robust generalization error) for linear regression with isotropic features, where an adversary is allowed to attack the input data point at test-time. The authors computed exact Pareto optimal curves that reveal a tradeoff between standard and adversarial risk.

Our work mostly considers the following question:

Question 1. *In the context of linear regression, if a model has "small" standard risk, how "small" can its adversarial risk be? Is it possible to be robust while being accurate?*

In the context of classification, an analogous question was considered in (Tsipras et al., 2019), where the authors constructed a high-dimensional problem for which every model with standard accuracy $1-\epsilon$ has adversarial accuracy at most $c\epsilon$, where c is an absolute constant. Such a result is reminiscent of a tradeoff between accuracy and robustness, and our results have the same flavor.

Summary of Our Contributions. The main contributions of this work precise quantitative estimates which distinguishes between regimes where robustness is achievable without hurting standard accuracy and regimes where a tradeoff might be unavoidable. Importantly, unlike previous works like (Javanmard et al., 2020; Xing et al., 2021), our analysis applies to general attack norms (not just Euclidean) and covariance matrices (not just isotropic). Our main findings can be broken down as follows.

- *Analytic Formula for Optimal Robustness.* As a function of the attack strength, we obtain analytic estimates of the optimal adversarial risk. Importantly, the model which attains optimal robustness is a regularized version of the generative model (the labelling function) with explicit regularization parameter. In the special case of Euclidean-norm attacks, it is a ridge estimator, and we recover a simplified formulation of the result obtained in (Xing et al., 2021).
- *Free Lunch and Tradeoffs for Robustness.* At any given attack strength, we establish a threshold on the

[1] Meta FAIR [2] Microsoft Research. Correspondence to: Elvis Dohmatob <dohmatob@meta.com>.

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

standard risk above which no tradeoff between standard accuracy and robustness is needed. These results answer Question 1 quantitatively. Importantly, we show that the model achieving the above accuracy / robustness tradeoff is a regularized estimate of the ground-truth / generative model, with regularization parameter given explicitly in terms of the attack strength and the accuracy tolerance.

- *Phase-Transition Diagrams.* Our analytic results identify phase-transitions for robustness in different regimes. As concrete examples, we focus on two settings: (i) the case of Euclidean-norm attacks on linear regression under polynomially-decaying spectral and source conditions, and (ii) the setting of ℓ_p-norm attacks on distributions where the covariance matrix is isotropic, with various structural assumptions on the generative model. For both settings, we compute the complete phase-transition diagram illustrating the tradeoffs between standard accuracy and adversarial robustness.

Implications of Our Results. Trade-offs between standard and adversarial risk provide valuable insights on the problem of adv. robustness and how it interferes with the standard objective of training models which perform well w.r.t to standard risk (non-robust test error). Indeed, from a conceptual viewpoint, such trade-offs mean that practitioners should treat the problem of robustness as seriously as the problem of good performance (in the classical sense of achieving good test error), and not as some minor quirk / side effect. From the practical standpoint, such tradeoffs provide guidance for deciding which loss function or algorithm (e.g regularization / no regularization; adv. training or normal training; etc.) to use. For example, our work shows that it is always optimal to consider a regularized model (see definition of (9) and its implication in our main results. In the case of Euclidean-norm attacks, this translates to early-stopping gradient-descent at an intermediate-time.

Related Work. The theoretical understanding of adversarial examples is now an active area of research. Below we discuss the works which are most relevant to our current paper. A more detailed overview of the literature can be found in Appendix A.

In the setting of classification, (Tsipras et al., 2019) considers a specific data distribution where good accuracy implies poor robustness. (Shafahi et al., 2018; Mahloujifar et al., 2018; Gilmer et al., 2018; Dohmatob, 2019) show that for high-dimensional data distributions which have concentration property (e.g., multivariate Gaussians, distributions satisfying log-Sobolev inequalities), an imperfect classifier will admit adversarial examples. (Dobriban et al., 2020) studies tradeoffs in Gaussian mixture classification problems, highlighting the impact of class imbalance. Additionally, (Yang et al., 2020) observed empirically that natural images are well-separated, and so locally-lipschitz classifiers should not suffer from potential attacks.

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

The study of robustness in linear regression for general norms and feature covariance matrices has been initiated in (Scetbon & Dohmatob, 2023) which gave sufficient conditions for the generative model w_0 and gradient-descent based estimators to be robust. However, the the question of tradeoffs was not considered.

Finally, (Dohmatob & Bietti, 2022) established tradeoffs between accuracy and robustness to Euclidean-norm attacks on two-layer neural networks in different regimes.

2. Problem Formulation

Notations. Given a positive-definite matrix M, the induced Mahahanobis norm $\|\cdot\|_M$ is define by $\|z\|_M := \|M^{1/2}z\|_2$. The notation $f(d) = O(g(d))$, also written $f(d) \lesssim g(d)$, means there exists an absolute constant K such that $f(d) \le K \cdot g(d)$. Likewise, $f(d) = \Omega(g(d))$ (or $f(d) \gtrsim g(d)$) means $g(d) = O(f(d))$. We write $f(d) \asymp g(d)$ to mean $f(d) \lesssim g(d) \lesssim f(d)$. Finally, $f(d) = o_d(g(d))$ (also written $f(d) \ll g(d)$) means $f(d)/g(d) \to 0$ when $d \to +\infty$. In particular, $f(d) = o_d(1)$ means that $f(d) \to 0$.

2.1. Data Distribution

In this work, we consider linear regression problem given by the following distribution P over a d-dimensional feature vector $x \in \mathbb{R}^d$ and labels $y \in \mathbb{R}$

$$\begin{aligned}&\textbf{(Features) } x \sim P_x := N(0, \Sigma),\\ &\textbf{(Label) } y = x^\top w_0 + z, \text{ with } z \sim N(0, \sigma^2) \text{ indep. of } x.\end{aligned} \tag{1}$$

Thus, the marginal distribution P_x of the features is a multivariate Gaussian with $d \times d$ positive-definite covariance matrix Σ. The generative model for the labels is a linear model defined by $x \mapsto f_{w_0}(x) := x^\top w_0$, for some fixed vector of parameters $w_0 \in \mathbb{R}^d$. To avoid trivialities, we will assume WLOG that $w_0 \neq 0$. The scalar $\sigma \geq 0$ is the strength of the label-noise $z \sim N(0, \sigma^2)$. The input-dimension d is not assumed fixed, and in fact, for the large part of this paper, we shall consider phenomena happening in the limit $d \to \infty$. One should keep in mind that in such a case, we are actually considering a sequence of problems, i.e. distributions $P(d)$ indexed by d.

Such a data distribution is also the setup of previous works like (Javanmard et al., 2020; Xing et al., 2021; Scetbon & Dohmatob, 2023). Note however that in (Javanmard et al., 2020), the covariance matrix is identity, i.e. $\Sigma = I_d$. In contrast, as in (Scetbon & Dohmatob, 2023), our work considers general covariance matrices Σ.

2.2. Linear Models, Standard and Adversarial Risks

This work considers regression over linear models $f_w(x) := x^\top w$, parametrized by a weights vector $w \in \mathbb{R}^d$. An adversarial attack replaces a clean data point $(x, y) \sim P$ by a perturbed version $(x + \delta, y)$. The size of the perturbation $\delta = \delta(x, y)$ is measured w.r.t a pre-specified norm $\|\cdot\|$ on the feature space $\mathbb{R}^d$. Note that the attacker is only allowed to change the feature vector x, and not the label y. By an attack of strength $r \geq 0$, we mean that the constraint $\|\delta\| \leq r$ is enforced. The goal of the attacker is to make the prediction $f_w(x + \delta)$ on the corrupted feature vector $x + \delta$ deviates from the ground-truth label y of clean feature vector x, as much as possible.

Definition 2.1 (Standard and Adversarial Risks)**.** *Given $w \in \mathbb{R}^d$, attack budget $r \geq 0$ w.r.t a arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$, the adversarial risk of the linear model f_w is*

$$E(w, r) := \mathbb{E}\left[\sup_{\|\delta\| \leq r} (f_w(x + \delta) - y)^2\right], \qquad (2)$$

for $(x, y) \sim P$. Also, recall the definition of the standard risk of f_w, namely

$$E(w) := \mathbb{E}[(f_w(x) - y)^2] = \|w - w_0\|_\Sigma^2 + \sigma^2. \qquad (3)$$

Of course, $E(w, r) \geq E(w, 0) = E(w)$ for any $w \in \mathbb{R}^d$ and $r \geq 0$, with equality if $r = 0$.

Remark 2.1. *Note that by definition, $E(w, r)$ depends on the attacker's norm $\|\cdot\|$. To simplify notations, we omit its dependency and precise the norm when needed.*

Let $\|\cdot\|_\star$ be the dual of $\|\cdot\|$, defined by $\|w\|_\star := \sup_{\|\delta\| \leq 1} \delta^\top w$. The following lemma established in (Scetbon & Dohmatob, 2023) gives an analytic formula for the adversarial risk which will be useful in the sequel.

Lemma 2.1. *For any $w \in \mathbb{R}^d$ and $r \geq 0$, it holds that $E(w, r) = E(w) + r^2\|w\|_\star^2 + 2\sqrt{2/\pi} r \|w\|_\star \sqrt{E(w)}$.*

Imperceptible Adversarial Attacks. In practice, one is usually only concerned with adversarial attacks which are *small* to the eye. Formally, this means that the attack strength r is restricted to be much smaller than the average norm of a random data point, i.e.

$$r = o_d(R(\Sigma)), \text{ with } R(\Sigma) := \mathbb{E}\,\|x\| \text{ for } x \sim P_x. \qquad (4)$$

For example, in the case of Euclidean-norm attack on isotropic features with covariance matrix $\Sigma = I_d$, we have $R(\Sigma) = R(I_d) \asymp \sqrt{\operatorname{tr}\Sigma} \asymp \sqrt{d}$ in the limit $d \to \infty$. Thus, in this case, an attack is only small in the sense of the above definition i.f.f $r/\sqrt{d} = o_d(1)$. For example, $r = \sqrt{d/\log d}$ would be small. On the other hand, if $\Sigma = (1/d)I_d$, then $R(\Sigma) \asymp \sqrt{\operatorname{tr} I_d/d} = 1$, thus, a Euclidean-norm attack would be small i.f.f $r = o_d(1)$.

3. Analysis of the Optimal Robustness

Definition 3.1. *Given an attack strength $r \geq 0$, let $E_{opt}(r)$ be the optimal adversarial risk,*

$$E_{opt}(r) := \min_{w \in \mathbb{R}^d} E(w, r). \qquad (5)$$

Furthermore, let $w_{opt}(r)$ denotes any $w \in \mathbb{R}^d$ achieving this optimum.

Observe that the expression for adversarial risk $E(w, r)$ given in Lemma 2.1 exhibits a tension between the standard risk $E(w)$, which is minimized by the generative model w_0, and the dual norm $\|w\|_\star$, which is minimized by the null model $w = 0$. Thus, for a given attack strength r, one would expect that the optimal robust model $w_{opt}(r)$ would have to somehow interpolate between w_0 and 0. In this section, we show that this is indeed the case (Theorem 3.1).

3.1. Adversarial Risk Proxies

Even though the adversarial risk functional E admits an analytic formula thanks to Lemma 2.1, that expression does not lend itself well to analysis. Following (Scetbon & Dohmatob, 2023), we shall resort to multiplicative approximations defined as follows. For any linear model $w \in \mathbb{R}^d$ and attack strength $r \geq 0$, set

$$\overline{E}(w, r) := \sigma^2 + \|w - w_0\|_\Sigma^2 + r^2\|w\|_\star^2, \qquad (6)$$

$$\widetilde{E}(w, r) := \sigma^2 + K(w, r)^2, \qquad (7)$$

with $K(w, r) := \|w - w_0\|_\Sigma + r\|w\|_\star$. Note that just like E, both $\overline{E}$ and $\widetilde{E}$ implicitly depend on the underlying norm $\|\cdot\|$. The following lemma shows that $\overline{E}(w, r)$ and $\widetilde{E}(w, r)$ are indeed proxies (i.e. multiplicative approximations) of the adversarial risk $E(w, r)$.

Lemma 3.1. *There exists absolute constants c_1 and c_2 such* *that for a general attacker norm $\|\cdot\|$, and $w \in \mathbb{R}^d$, $r \geq 0$,*

$$\begin{aligned}\widetilde{E}(w,r) \leq E(w,r) \leq c_1\widetilde{E}(w,r),\\ \overline{E}(w,r) \leq E(w,r) \leq c_2\overline{E}(w,r).\end{aligned} \tag{8}$$

The first part was established in (Scetbon & Dohmatob, 2023). The second part follows similar arguments as the first one. See Appendix 3.1 for the proof, including explicit values c_1 and c_2.

Remark 3.1. *The multiplicative approximations given in Lemma 3.1 suffice for our purposes whereby we only are interested in the orders of magnitude of the adversarial risk of models relative to the optimum value $E_{opt}(r)$, as a function of the attack strength r.*

3.2. Robustness via Regularization

For any $\lambda \geq 0$, let $w^{prox}(\lambda)$ be the unique minimizer of $\overline{E}(w, \sqrt{\lambda})$ over $w \in \mathbb{R}^d$, that is

$$w^{prox}(\lambda) := \arg\min_{w\in\mathbb{R}^d} \|w - w_0\|_\Sigma^2 + \lambda\|w\|_\star^2. \tag{9}$$

Thus, $w^{prox}(\lambda)$ is the *proximal operator* w.r.t the squared-Mahalanobis norm $\|\cdot\|_\Sigma^2$, of the square of rescaled squared dual norm $\lambda\|\cdot\|_\star^2$, evaluated at the point w_0. Of course, it implicitly depends on the choice of the norm $\|\cdot\|$ of the attacker. For example, in the special case of Mahalanobis-norm attacks w.r.t. any positive definite matrix B, that is when $\|\cdot\| = \|\cdot\|_B$, we have the closed-form expression $w^{prox}(\lambda) = (\Sigma + \lambda B^{-1})^{-1}\Sigma w_0$.

Define auxiliary functions

$$\begin{aligned}G(\lambda) = G^{\|\cdot\|}(\lambda) := \|w^{prox}(\lambda) - w_0\|_\Sigma^2,\\ F(r,\lambda) = F^{\|\cdot\|}(r,\lambda) := G(\lambda) + r^2\|w^{prox}(\lambda)\|_\star^2,\end{aligned} \tag{10}$$

The following result which holds for any choice of the attacker's norm $\|\cdot\|$ is one of our main results.

Theorem 3.1. *With $\lambda = r^2$, it holds that $E_{opt}(r) \asymp E(w^{prox}(\lambda), r) \asymp \sigma^2 + F(r, r^2)$. That is, up to within multiplicative absolute constants, $w^{prox}(\lambda = r^2)$ attains the optimal adversarial risk $E_{opt}(r)$.*

Note that the above result is valid for any attacker norm. The special case of Euclidean-norm attacks was handled in (Xing et al., 2021) where it was shown that $w_{opt}(r) = (\Sigma + \lambda I_d)^{-1}\Sigma w_0$, for some $\lambda \in [0, \infty]$ which depends on r, w_0, and Σ, via a fixed-point equation that must be solved numerically. Even, in this scenario, our result above gives a much clearer understanding, since it proposes to use the explicit ridge parameter $\lambda = r^2$, which clearly highlights the dependence on the attack strength r.

4. Phase-Transitions and Accuracy vs Robustness Tradeoffs

4.1. Preliminaries

In view of addressing Question 1, our objective is to compute bounds on the optimal adversarial risk with respect to any given norm over all linear models which attain a certain level of standard risk. For any linear model $w \in \mathbb{R}^d$, let $\Delta(w) := (E(w) - \sigma^2)/\|w_0\|_\Sigma^2$ be the normalized excess standard risk of w. The division by $\|w_0\|_\Sigma^2$ ensures that $\Delta(w_0) = 0$ while $\Delta(0) = 1$.

Optimal Robustness of Accurate Models. For any $r, \epsilon \geq 0$, let $\mathcal{W}_\epsilon$ be the set of all ϵ-accurate models, i.e.

$$\begin{aligned}\mathcal{W}_\epsilon &:= \{w \in \mathbb{R}^d \mid \Delta(w) \leq \epsilon^2\}\\ &= \{w \in \mathbb{R}^d \mid \|w - w_0\|_\Sigma \leq \epsilon\|w_0\|_\Sigma\},\end{aligned} \tag{11}$$

and let $E_{opt}(r, \epsilon)$ be the optimal adversarial risk of such models against attacks of strength r, i.e.

$$E_{opt}(r, \epsilon) := \min_{w\in\mathcal{W}_\epsilon} E(w, r). \tag{12}$$

Finally, let $w_{opt}(r, \epsilon)$ denotes any $w \in \mathbb{R}^d$ achieving the optimum in (12). $E_{opt}(r, \epsilon)$ will be the main object of study of our paper as it captures the sacrifice in robustness that must be made by accurate models.

The following lemma shows that this constrained formulation of the adversarial risk minimization problem can only differ from the unconstrained one when $\epsilon \in [0, 1)$.

Lemma 4.1. *For any $r \geq 0$ and $\epsilon \geq 1$, it holds that $E_{opt}(r, \epsilon) = E_{opt}(r)$.*

Thus, when $\epsilon \geq 1$, $\mathcal{W}_\epsilon$ contains an optimally robust linear model achieving the optimal adversarial risk over all the linear models. Let us now introduce a critical value for ε.

Free Lunch Threshold. Given any $r \geq 0$, the quantity $\epsilon_{FL}(r) \in [0, 1]$ defined by

$$\epsilon_{FL}(r) := \sqrt{\Delta(w^{prox}(r^2))} = \sqrt{G(r^2)}/\|w_0\|_\Sigma, \tag{13}$$

will be called the "free lunch" threshold for attacks of strength r, a terminology that will become clear shortly in Theorem 4.1. In addition we also need to introduce the following lemma.

Lemma 4.2. *For any $r \geq 0$ and $0 \leq \epsilon \leq \epsilon_{FL}(r)$, the scalar equation*

$$G(\lambda) = \epsilon^2\|w_0\|_\Sigma^2 \tag{14}$$

has a unique solution $\lambda_{opt}(r, \epsilon)$ in $[0, r^2]$.

We extend the definition of $\lambda_{opt}(r, \epsilon)$ to all $\epsilon \in [0, 1]$ by setting $\lambda_{opt}(r, \epsilon) = r^2$ whenever $\epsilon \geq \epsilon_{FL}(r)$.

4.2. Main Result

We are now ready to present the main result of this work and show in the following theorem tradeoffs between standard accuracy and adversarial robustness in the setting of general feature covariance matrix Σ and attacker norm $\|\cdot\|$.

Theorem 4.1. *For any attack strength $r \geq 0$ and tolerance $\epsilon \in [0,1]$, the following hold.*

*(A) (**Accuracy vs Robustness Tradeoff**) It holds that*

$$E_{opt}(r,\epsilon) \asymp E(w^{prox}(\lambda_{opt}(r,\epsilon)),r),$$

where $\lambda_{opt}(r,\epsilon) \in [0,r^2]$ is as in Lemma 4.2 and $w^{prox}(\lambda)$ is as defined in (9). That is, up to within multiplicative absolute constants, with the choice $\lambda = \lambda_{opt}(r,\epsilon)$ the vector $w^{prox}(\lambda)$ attains the optimal adversarial risk $E_{opt}(r,\epsilon)$ over all ϵ-accurate models.

*(B) (**Free Lunch**) If $\epsilon \geq \epsilon_{FL}(r)$, then it holds that $E_{opt}(r,\epsilon) \asymp E_{opt}(r)$. That is, no accuracy / robustness tradeoff is needed when the excess risk level ϵ is greater than the threshold $\epsilon_{FL}(r)$: there is always an ϵ-accurate model which achieves the absolute optimal (up to within multiplicative absolute constants) adversarial risk $E_{opt}(r)$.*

Comparison to (Javanmard & Soltanolkotabi, 2022). Note that the above theorem is more general that the results of (Javanmard & Soltanolkotabi, 2022). Indeed, the latter only focuses on Euclidean-norm attacks on isotropic features where $\Sigma = I_d$. In contrast, Theorem 4.1 (and corollaries) covers the case of general covariance matrices Σ and general attack norms (not just Euclidean). Also, the techniques used in our work are very different from those in (Javanmard & Soltanolkotabi, 2022). Indeed, the analysis in the latter is based on *Gordon's Comparison Inequality* (Gordon & Milman, 1988; Thrampoulidis et al., 2015; 2018), which is a very versatile tool in the analysis of regularized estimators but fails to produce analytic results when one deviates from the setting of Euclidean-norm attacks on isotropic features. In contrast, our analysis is based on basic Langrangian duality. It relies on some approximations which turn out to only introduce multiplicative absolute constants in the final result, but are completely harmless for the final analysis and interpretation.

Remark 4.1. Note that our Theorem 4.1 can be put in the language of Pareto fronts as discussed in the Appendix H.

5. Some Direct Consequences of Our Results

We now apply our Theorems 3.1 and 4.1, to obtain some concrete consequences in a variety of settings. Section 6 will provide some empirical confirmation of these predicted consequences.

5.1. When the Ground-Truth is Sparse

Consider the case where the feature covariance matrix Σ and the generative model w_0 are given by

$$\Sigma = (1/d)I_d,\ w_j = 1\,\forall j \leq s,\ w_j = 0\,\forall j \geq s+1, \tag{15}$$

for some sparsity parameter $s \in [d]$. We consider ℓ_p-norm attacks, for some fixed $p \in [1,\infty]$. First observe, that for ℓ_∞-norm attacks of strength r, the adversarial risk of the generative mode w_0 is given by $E(w_0,r) = \sigma^2 + r^2\|w_0\|_1^2 = \sigma^2 + s^2r^2$, while for ℓ_p-norm attacks with $p \in [1,\infty)$, we have

$$E(w_0,r) = \sigma^2 + r^2\|w_0\|_q^2 = \sigma^2 + r^2 s^{2/q},$$

where $q :=\in [1,\infty]$ is the harmonic conjugate of p.

Theorem 5.1. *Recall the notations of Theorem 4.1. Let the attack norm be an ℓ_p with $p \in [1,\infty]$. For any $r \geq 0$ and $\epsilon \in [0,1]$, the robustness profile is given as in Table 1.*

In particular, in the limit $d \to \infty$, we have that, if

- *$p \in [1,\infty)$, $1 \ll s \leq d$, and we take $r \asymp 1/s^{1/q}$, OR*
- *$p = \infty$, $\sqrt{d/\log d} \ll s \leq d$, and we take $r \asymp 1/s$,*

then for $\epsilon \in [0,1)$, it holds that

$$E_{opt}(r) = o_d(1),\ \text{and}\ E_{opt}(r,\epsilon) = \Theta((1-\epsilon)^2). \tag{16}$$

5.2. Polynomial Spectral Decay

Let $\Sigma = \sum_{k\geq 1}\lambda_k\phi_k\phi_k^\top$ be the spectral decomposition of the feature covariance matrix Σ and $c_k = \phi_k^\top w_0$ be the k-th alignment coefficient of the generative model w_0, so that $w_0 = \sum_{k\geq 1} c_k\phi_k$. We place ourselves in the high-dimensional setting ($d \to \infty$), and assume spectral information $(\lambda_k, c_k)_{k\geq 1}$ is given by the following polynomial (aka power-law) scalings

$$\lambda_k \asymp k^{-\beta},\ \text{ and } c_k^2 \asymp k^{-\delta} \text{ for all } k, \tag{17}$$

where $\beta > 1$ and $\delta \geq 0$ are constants. This model is well-studied in the literature (Caponnetto & Vito, 2007; Liang & Rakhlin, 2020) because (1) it usually leads to tractable analysis, and (2) it can be used to approximate the the macroscopic structure of certain neural networks in the kernel regime (Bahri et al., 2021; Cui et al., 2022; Wei et al., 2022). In this setting, observe that tr$(\Sigma) \asymp \sum_k k^{-\beta} = \Theta(1)$, $\|w_0\|_\Sigma^2 \asymp \sum_k k^{-\beta-\delta} = \Theta(1)$, while

$$\|w_0\|_2^2 \asymp \sum_k k^{-\delta} \asymp \begin{cases} d^{1-\delta}, & \text{if } 0 \leq \delta < 1, \\ \log d, & \text{if } \delta = 1, \\ 1, & \text{if } \delta > 1. \end{cases} \tag{18}$$

	$\epsilon_{FL}(r)$	$\lambda_{opt}(r,\epsilon)$	$E_{opt}(r)$	$E_{opt}(r,\epsilon)$
$p=2$	$r^2/(1+r^2)$	$\epsilon/(1-\epsilon)$	$\sigma^2+(s/d)\min(r\sqrt{d},1)^2$	$\sigma^2+(s/d)H(r\sqrt{d},\epsilon)^2$
$p\neq 2$	–	–	$\sigma^2+(s/d)\min(r/r_0(p),1)^2$	$\sigma^2+(s/d)H(r/r_0(p),\epsilon)^2$

Table 1. Details of Theorem 5.1. Here, $r_0(p) = s^{1/p-1/2}/\sqrt{d}$. In particular, $r_0(2) = 1/\sqrt{d}$, $r_0(\infty) = 1/\sqrt{sd}$. The function H is defined by $H(r,\epsilon) = r$ if $r \le 1$; else $H(r,\epsilon) = \epsilon + (1-\epsilon)r$.

Also note that for Euclidean-norm attacks, we have $R(\Sigma) \asymp \sqrt{\operatorname{tr}(\Sigma)} = \Theta(1)$.

The following result is one of our main contributions.

Theorem 5.2. *For Euclidean-norm attacks of small strength $r > 0$, the conclusions of Theorem 4.1 prevail, and the quantities $\epsilon_{FL}(r)$, $\lambda_{opt}(r,\epsilon)$, $E_{opt}(r)$, and $E_{opt}(r,\epsilon)$ are as given in Table 2.*

Consider the particular regime where $0 \le \delta \le 1$. For small $\sigma^2 \ge 0$, $\epsilon > 0$, and $r = r(\epsilon)$ given by

$$r = \begin{cases} \epsilon^{\phi}, & \text{if } 0 \le \delta < 1, \\ \sqrt{1/\log(1/\epsilon)}, & \text{if } \delta = 1 \end{cases} \tag{19}$$

with $\theta := (1-\delta)/\beta \ge 0$, $\phi := \theta/(1-\theta) \ge 0$, it holds that

$$E_{opt}(r) = o(1),\ E_{opt}(r,\epsilon) = \Theta(1). \tag{20}$$

Thus, as regards robustness, $\delta = 1$ is a critical value for the source exponent in (17): For $\delta \in [0,1]$, accuracy (controlled by the excess risk tolerance ϵ) has to be traded for robustness, while for $\delta \in (1,\infty)$, the generative model w_0 is so smooth that robustness and accuracy are aligned. In the regime $0 \le \delta < 1$, the theorem predicts that even though robustness to imperceptible attacks is achievable in this setting, accurate models (especially the generative model w_0 itself) are non-robust.

Therefore, there is a phase-transition at $\delta = 1$ whereby accurate models switch from non-robust to robust. This is also empirically confirmed in Section 6. Notice the power-law behavior dependence on $E_{opt}(r,\epsilon)$ on $1/\epsilon$ in the case of nonsmooth ground-truth models w_0 where $\delta \in [0,1)$.

5.3. A Non-Euclidean Setting

Let us now present an example of non-Euclidean setting where generative model w_0 fails to be robust to genuinely small adversarial perturbations. Still in high dimensions ($d \to \infty$), consider the setting where the feature covariance matrix is $\Sigma = I_d$ while the coefficients of the generative model have the "harmonic" distribution w_0, i.e

$$d \to \infty,\ \Sigma = I_d,\ (w_0)_k = 1/k,\ \text{for all } k \in [d]. \tag{21}$$

For sup-norm (i.e ℓ_∞-norm) attacks, observe that $R(\Sigma) \asymp \sqrt{\log \operatorname{tr}\Sigma} \asymp \sqrt{\log d}$. We have the following.

Theorem 5.3. *Consider the setting (21). For ℓ_∞-norm attacks of strength r with $1/\sqrt{d} \le r = o(1)$, it holds that*

$$E_{opt}(r) \asymp \sigma^2 + r^2\log(1/r)^2. \tag{22}$$

In particular, for $r \asymp 1/\log d$ and $\sigma^2 = o(1)$, it holds that

$$E(w_0,r) = \Theta(1),\ E_{opt}(r) = o(1). \tag{23}$$

That is, even though robustness is achievable, the generative model w_0 is itself non-robust.

5.4. What about Neural Networks ?

Notwithstanding, our theoretical insights carry over to wide neural networks in the so-called kernel regime (e.g infinite-width NTK). In this limit, the model is $y = \langle\phi(x), w_0\rangle_H + N(0,\sigma^2)$, where $\phi(x)$ is a feature corresponding to the RKHS H induced by the limiting kernel K. Such a regime has been considered in [25] in the study of scaling laws for test error in neural networks. If we assume $\phi(x) \sim N(0,\Sigma)$ as in (Cui et al., 2022), then we can use section 3 of our paper to obtain insights on the robustness-accuracy tradeoffs for this regime. Applying this to the settings considered in Section 5 of (Cui et al., 2022), namely MNIST and FashionMNIST with kernel $K(x,z) = (1 + 10^{-3}x^\top z)^5$ (the task is to predict the class label via kernel regression), we see from Table 1 of (Cui et al., 2022) that $\delta = 1+\alpha(2r-1) = 1+1.3(2(0.13)-1) = \mathbf{0.038} \in (0,1)$ for MNIST and $\delta = 1+\alpha(2r-1) = 1+1.2(2(0.15)-1) = \mathbf{0.16} \in (0,1)$ for FashionMNIST. Equipped with this back-of-envelop calculation, Theorem 5.2 of our paper (refer to the first row of Table 2 therein) then predicts that for both of these kernelized NN settings, there is an unavoidable tradeoff between accuracy and robustness.

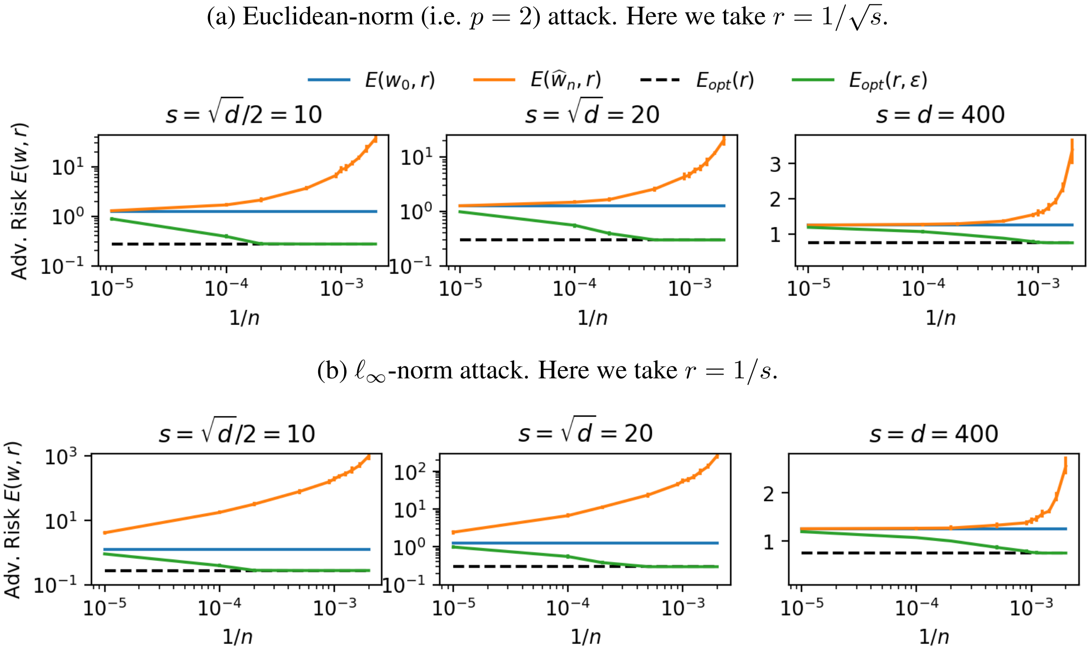
6. Empirical Verification

We provide a series of simple experiments on simulated data to empirically verify our theoretical results. Additional experiments are provided in the appendix.

Experiment 1 (Verification of Theorem 5.1). For this experiment, we fix the input dimension $d = 400$, while the covariance matrix Σ and generative model w_0 are as in (15) for different values of the sparsity parameter $s \in \{10, 20, d = 400\}$. For different values of sample size n from d to 10^4, we generate (5 runs) an iid dataset

Regime	$\epsilon_{FL}(r)$	$\lambda_{opt}(r,\epsilon)$	$E_{opt}(r)$	$E_{opt}(r,\epsilon)$	Free lunch ?
$0 \le \delta < 1$	$r^{2(1-\theta)}$	$\epsilon^{2/(1-\theta)}$	$\sigma^2 + r^{2(1-\theta)}$	$\sigma^2 + \epsilon^2 + r^2\epsilon^{-2\phi}$	No
$\delta = 1$	$r^4 \log(1/r)$	$e^{W(-\Theta(\epsilon^2))/2}$	$\sigma^2 + r^2 \log(1/r)$	$\sigma^2 + \epsilon^2 + r^2 \log(1/\epsilon)$	No
$\delta > 1$	r^4	ϵ	$\sigma^2 + r^2$	$\sigma^2 + \epsilon^2 + r^2$	Yes!

Table 2. Details for Theorem 5.2. Here, W is an appropriate branch of the Lambert function. Note that except for the first column, all the entries in the table are given only within multiplicative absolute constants. The last column records whether there is free lunch (FL), wherein robustness is achievable without sacrificing accuracy.


(a) Euclidean-norm (i.e. $p = 2$) attack. Here we take $r = 1/\sqrt{s}$.

(b) ℓ_∞-norm attack. Here we take $r = 1/s$.

Figure 1. (**Experiment 1**) Empirical verification of Theorem 5.1, for different levels of sparsity s of the generative model w_0. Here the input dimension is set to $d = 400$ and n is the sample size. The theoretical curves $E_{opt}(r)$ and $E_{opt}(r,\epsilon)$ are as given by the theorem. Error bars correspond to 5 different runs of computing $\widehat{w}_n$ (OLS). Notice the conformity with the theorem's predictions.

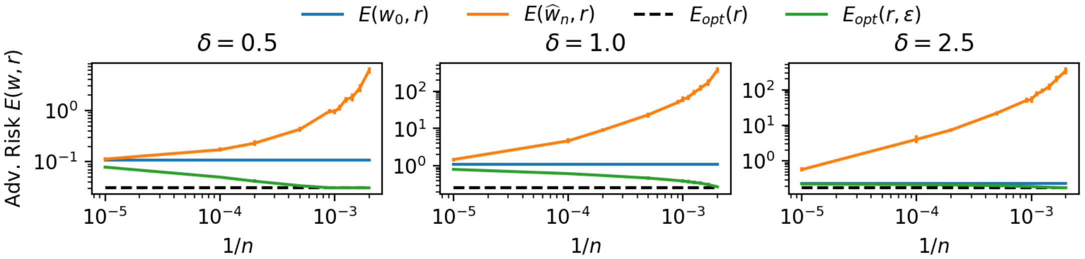


Figure 2. (**Experiment 2**) Empirical verification of Theorem 5.2. Here $\beta = 2$ and $d = 10^4$, while n is the sample size. Error bars correspond to 5 different runs of computing the OLS estimator $\widehat{w}_n$. The curves $E_{opt}(r)$ and $E_{opt}(r,\epsilon)$ are as given by the theorem. Notice the conformity of these empirical results with the theorem.

$\mathcal{D}_n = \{(x_1,y_1),\ldots,(x_n,y_n)\}$ and construct a ordinary least-squares (OLS) estimate $\widehat{w}_n$ for w_0. Note that this estimator is known to be consistent in the regime considered. Next, we compute the excess standard risk of $\widehat{w}_n$, namely $\epsilon = \epsilon_n := (\|\widehat{w}_n - w_0\|_\Sigma^2 - \sigma^2)/\|w_0\|_\Sigma^2$. We consider ℓ_p-norm attacks with $p \in \{2,\infty\}$. The attack strength r is set as in the second part of Theorem 5.1. We compute the adversarial risk of $\widehat{w}_n$, alongside the adversarial risk of the generative model w_0, via the formula given in Lemma 2.1. We approximate the optimally robust ϵ-accurate model via the regularization scheme (9) with $\lambda = \lambda_{opt}(r,\epsilon)$ given as in Theorem 5.1. The results for this experiment are shown in Figure 1. From the figure, we clearly see that the predictions of Theorem 5.1 are confirmed.

Experiment 2 (Verification of Theorem 5.2). The setup

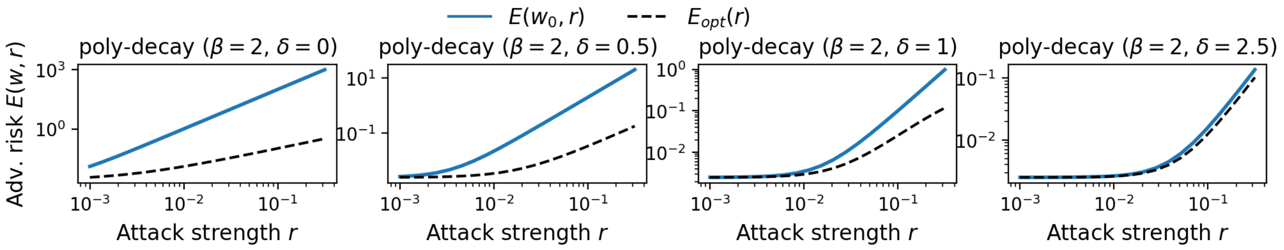


Figure 3. **(Experiment 2)** Empirical validation of Theorem 5.2 for the case $\epsilon = 0$. Notice the conformity of the results with the theorem.

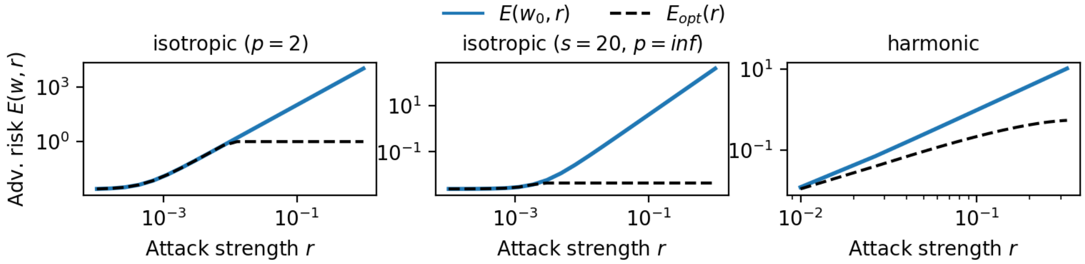


Figure 4. Failure of ground-truth model w_0 to be robust to small adversarial perturbations (indicated by the separation between the solid and the broken curves). From left to right, the plots show results for **Experiment 3(a–c)**. Notice perfect confirmation of Theorem 5.1 for $\epsilon = 0$ and $p \in \{2, \infty\}$ (**Left** and **Middle** plots), and Theorem 5.3 (**Right** plot).

for this experiment is as in **Experiment 1**, but with input dimension $d = 10^4$, feature covariance matrix Σ and generative model w_0 given as in (17). We consider Euclidean-norm attacks with strength r as given in Theorem 5.2. The results for the experiment are shown in Figure 2. As predicted by the theorem, we see that for $\delta \in [0, 1]$, the ground-truth model w_0 is non-robust. Furthermore, for $\delta > 1$, w_0 becomes robust to small adversarial perturbations (blue curve and broken black line coincide) as predicted. As $n \to \infty$, the adversarial risk $E(\widehat{w}_n, r)$ of the estimator $\widehat{w}_n$ approaches that of the ground-truth model, namely $E(w_0, r)$; we see from the figure that is optimal in the smooth regime where $\delta > 1$, but catastrophic in the non-smooth regime ($\delta \in [0, 1]$), in conformity the theorem.

We also consider the case $\epsilon = 0$, corresponding the ground-truth model, and vary the attack strength r. The results are shown in Figure 3. Here again, the predictions of Theorem 5.2 are confirmed.

Experiment 3: Non-Robustness of Generative Model w_0. This experiment is meant to verify Theorem 5.3 and Theorem 5.1 for the case $\epsilon = 0$ (corresponding to the generative model w_0). We fix the input dimension to $d = 400$ as **Experiment 1**, and consider 3 different scenarios for the attacker's norm $\|\cdot\|$, the generative model w_0, and the feature covariance matrix Σ.

- **Experiment 3(a)**: Here, the feature covariance matrix is $\Sigma = (1/d)I_d$ and the generative model is $w_0 = 1_d = (1, \ldots, 1)$. The attacker's norm is Euclidean, i.e. $p = 2$ in Theorem 5.1.
- **Experiment 3(b)**: Here, $\Sigma = (1/d)I_d$ and w_0 is as in (15) with $s = 20$. The attacker's norm is ℓ_∞, i.e. $p = \infty$ in Theorem 5.1.
- **Experiment 3(c)**: Here, Σ and w_0 are as in Theorem 5.3 and the attacker's norm is ℓ_∞.

For different values of attack strength r, we compute the adversarial risk $E(w_0, r)$ of the generative model w_0 and compare it to the optimum adversarial risk $E_{opt}(r)$. The results of the experiment are shown in Figure 4. We see that the predictions of Theorem 5.1 (**Left** and **Middle** plots) and Theorem 5.3 (**Right** plot) are perfectly confirmed.

7. Concluding Remarks

Our work has considered the problem of adversarial robustness in linear regression and have obtained precise quantitative estimates that allow us to uncover fundamental tradeoffs between adversarial robustness and standard accuracy in different regimes. Unlike previous works, our results apply to arbitrary covariance structures and attack norms.

Future Directions. An interesting future direction of our work will be to extend the scope to neural networks in linearized regimes like random features. As in (Hassani & Javanmard, 2022), such an analysis would rely on a careful application of random matrix theory, to reduce things to the linear case.

Impact Statement

This paper is dedicated to advancing the field of Machine Learning, specifically by enhancing the theoretical understanding of adversarial robustness—a critical, yet unresolved issue poised to become even more prominent in the era of large language models (LLMs) and ChatGPT. We carefully examine the societal implications of our research, confidently asserting that our contributions are entirely positive. Through a detailed exploration of adversarial robustness, our work not only addresses a fundamental challenge but also sets the stage for more secure and reliable AI systems in the face of evolving adversarial threats. This research is particularly relevant as we enter a period marked by the widespread adoption of LLMs, underscoring the urgent need for advancements in our theoretical frameworks to safeguard the integrity and utility of these powerful tools.

References

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining neural scaling laws. *ArXiv*, abs/2102.06701, 2021.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 1st edition, 2011.

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

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

Blais, E., Canonne, C. L., and Gur, T. Distribution testing lower bounds via reductions from communication complexity. *ACM Trans. Comput. Theory*, 11(2), feb 2019.

Bubeck, S. and Sellke, M. A universal law of robustness via isoperimetry. In *Advances in Neural Information Processing Systems*, 2021.

Bubeck, S., Price, E., and Razenshteyn, I. P. Adversarial examples from computational constraints. *CoRR*, abs/1805.10204, 2018.

Bubeck, S., Li, Y., and Nagaraj, D. A law of robustness for two-layers neural networks. *arXiv e-prints*, art. arXiv:2009.14444, September 2020b.

Caponnetto, A. and Vito, E. D. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 2007.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Generalization error rates in kernel regression: the crossover from the noiseless to noisy regime*. *Journal of Statistical Mechanics: Theory and Experiment*, pp. 114004, 2022.

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

Dohmatob, E. Generalized no free lunch theorem for adversarial robustness. In *International Conference on Machine Learning (ICML)*, 2019.

Dohmatob, E. Fundamental tradeoffs between memorization and robustness in random features and neural tangent regimes. *arXiv preprint arXiv:2106.02630*, 2021.

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

Gao, R., Cai, T., Li, H., Hsieh, C.-J., Wang, L., and Lee, J. D. Convergence of adversarial training in overparametrized neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. J. Adversarial spheres. *CoRR*, abs/1801.02774, 2018.

Gordon, Y. and Milman, V. D. On milman's inequality and random subspaces which escape through a mesh in R^n. In *Geometric Aspects of Functional Analysis*, pp. 84–106. Springer Berlin Heidelberg, 1988.

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

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

Javanmard, A. and Soltanolkotabi, M. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, 2022.

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

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. *ANNALS OF STATISTICS*, pp. 1329–1347, 2020.

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

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *CoRR*, abs/1809.03063, 2018.

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

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

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. Are adversarial examples inevitable? *CoRR*, abs/1809.02104, 2018.

Sra, S. Nonconvex proximal splitting: batch and incremental algorithms, 2011.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, 2015.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 2018.

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

Wei, A., Hu, W., and Steinhardt, J. More than a toy: Random matrix models predict how real-world neural representations generalize. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23549–23588. PMLR, 2022.

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

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. In *Advances in Neural Information Processing Systems*, 2020.

Appendix / Supplementary Material

Precise Accuracy / Robustness Tradeoffs in Regression: Case of General Norms

A. Detailed Overview of Related Works

Adversarial Examples From High-Dimensional Geometry. In the setting of classification, (Tsipras et al., 2019) considers a specific data distribution where good accuracy implies poor robustness. (Shafahi et al., 2018; Mahloujifar et al., 2018; Gilmer et al., 2018; Dohmatob, 2019) show that for high-dimensional data distributions which have concentration property (e.g., multivariate Gaussians, distributions satisfying log-Sobolev inequalities, etc.), an imperfect classifier will admit adversarial examples. (Dobriban et al., 2020) studies tradeoffs in Gaussian mixture classification problems, highlighting the impact of class imbalance. On the other hand, (Yang et al., 2020) observed empirically that natural images are well-separated, and so locally-lipschitz classifiers should not suffer any kind of test error vs robustness tradeoff.

The Impact of Over-Parametrization. (Gao et al., 2019; Bubeck et al., 2020b; Bubeck & Sellke, 2021) show that over-parameterization may be necessary for robust interpolation in the presence of noise. In contrast, our paper considers a structured problem with noiseless signal and infinite training data, where the network width m and the input dimension d tend to infinity proportionately. In this under-complete asymptotic setting, our results show a systematic and precise tradeoff between approximation (test error) and robustness in different learning regimes. Thus, our work nuances the picture presented by previous works by exhibiting a nontrivial interplay between robustness and test error, which persists even in the case of infinite training data where the resulting model isn't affected by label noise. (Dohmatob, 2021; Hassani & Javanmard, 2022) study the tradeoffs between interpolation, predictive performance (test error), and robustness for finite-width over-parameterized networks in kernel regimes with noisy linear target functions. In contrast, we consider structured quadratic target functions and compare different learning settings, including SGD optimization in a non-kernel regime, as well as lazy/linearized models.

Precise Analysis of Robustness in Linear Regression. (Xing et al., 2021) studied Euclidean-norm attacks with general covariance matrices. They showed that the optimal robust model is a ridge regression whose ridge parameter depends implicitly on the strength of the attacks. (Javanmard et al., 2020) studied tradeoffs between ordinary and adversarial risk in linear regression, and computed exact Pareto optimal curves in the case of Euclidean-norm attacks on isotropic features. Their results show a tradeoff between ordinary and adversarial risk for adversarial training. (Javanmard & Mehrabi, 2021) also revisited this tradeoff for latent models and show that this tradeoff is mitigated when the data enjoys a low-dimensional structure. The analysis in (Javanmard et al., 2020) is based on *Gordon's Comparison Inequality* (Gordon & Milman, 1988; Thrampoulidis et al., 2015; 2018), which is a very versatile tool in the analysis of regularized estimators but fails to produce analytic results when one deviates from the setting of Euclidean-norm attacks on isotropic features. In contrast, our analysis is based on basic Langrangian duality. It relies on some approximations which turn out to only introduce multiplicative absolute constants in the final result, but are completely harmless for the final analysis and interpretation.

Finally, the study of robustness of gradient-descent in the context of linear regression under general-norms attacks and feature covariance matrices has been initiated in (Scetbon & Dohmatob, 2023) which gave sufficient conditions for the generative model w_0 (and its estimators like gradient descent, ridge regression, etc.) to be robust. However, the the question of tradeoffs was not considered.

B. Additional Experimental Results

We provide further empirical confirmation for Theorem 5.2. Figures 5 and 6 are complementary to Figures 2 and 3 respectively in the main text. They show results for **Experiment 2** (refer to Section 6) for other values of the exponents β and δ. We stress that all experiments in our paper were run on a single modern CPU laptop. See attached Jupyter (Python) notebook.

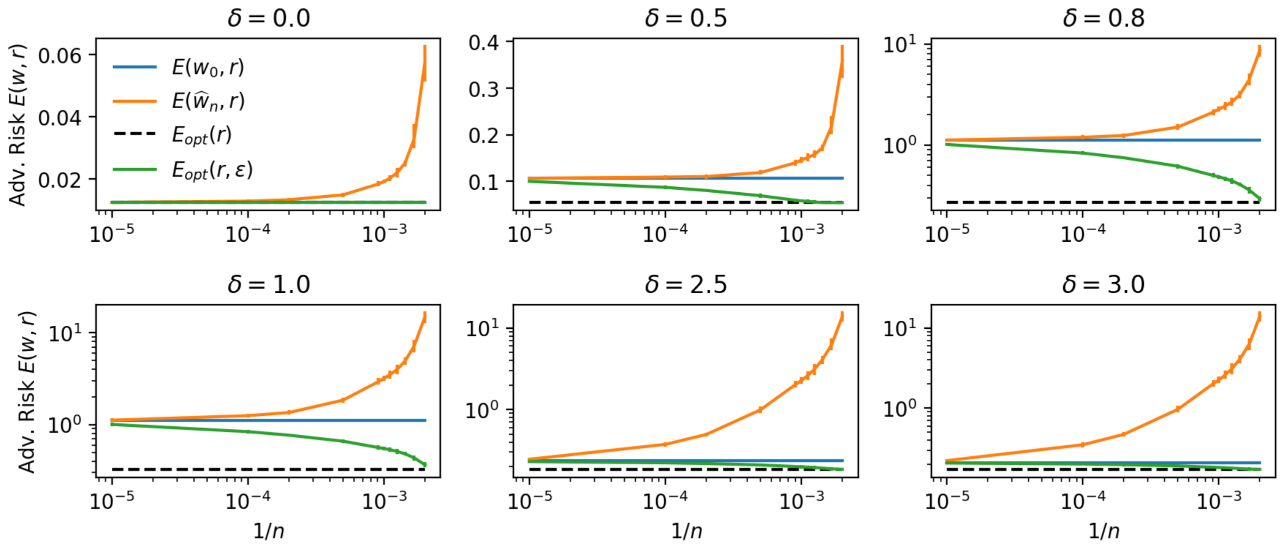


Figure 5. (**Experiment 2, extended**) Empirical verification of Theorem 5.2. Here $\beta = 1.4$ and $d = 10^4$. As in Figure 2, notice the conformity of the results with the theorem, namely: if the the model $\widehat{w}_n$ is accurate (small ϵ), then it is robust (compared to the optimal achievable adversarial risk $E_{opt}(r)$) for $\delta \in (1, \infty)$, but non-robust for $\delta \in [0, 1)$.

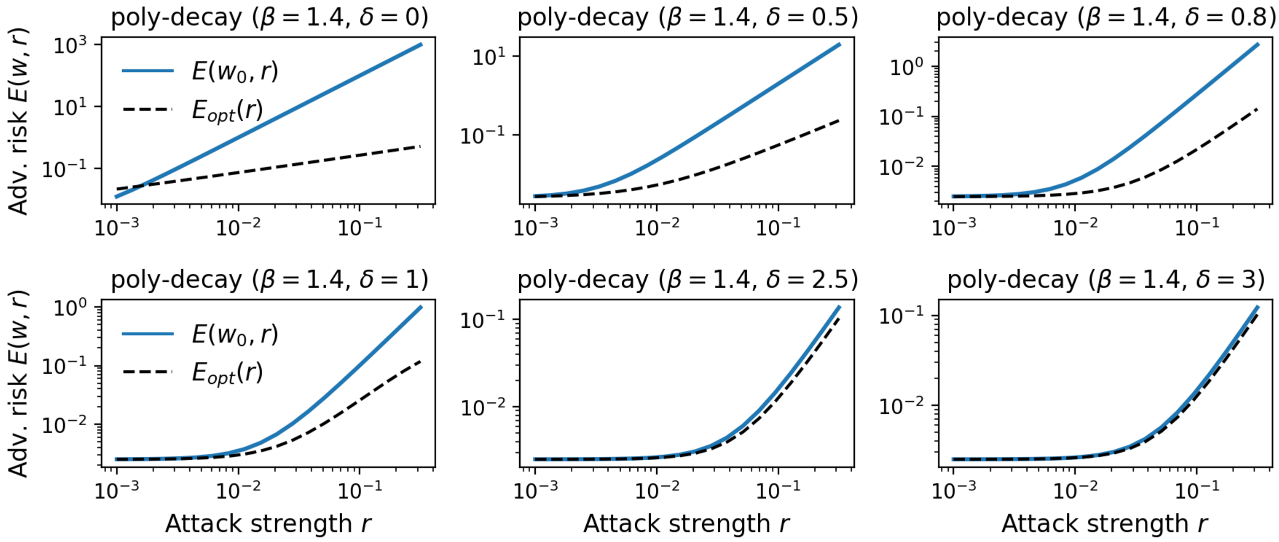


Figure 6. (**Experiment 2, extended**) Empirical validation of Theorem 5.2 for the case when $\epsilon = 0$. Here $\beta = 1.4$. As in Figure 3, notice the conformity of the results with the theorem, namely: the generative model w_0 is robust (compared to the optimal achievable adversarial risk $E_{opt}(r)$) for $\delta \in (1, \infty)$ but non-robust for $\delta \in [0, 1)$.

C. Proof of Theorem 4.1 (Main Result) and Theorem 3.1

Theorem 4.1. *For any attack strength $r \geq 0$ and tolerance $\epsilon \in [0, 1]$, the following hold.*

*(A) (**Accuracy vs Robustness Tradeoff**) It holds that*

$$E_{opt}(r, \epsilon) \asymp E(w^{prox}(\lambda_{opt}(r, \epsilon)), r),$$

where $\lambda_{opt}(r, \epsilon) \in [0, r^2]$ is as in Lemma 4.2 and $w^{prox}(\lambda)$ is as defined in (9). That is, up to within multiplicative absolute constants, with the choice $\lambda = \lambda_{opt}(r, \epsilon)$ the vector $w^{prox}(\lambda)$ attains the optimal adversarial risk $E_{opt}(r, \epsilon)$ over *all ϵ-accurate models.*

*(B) (**Free Lunch**) If $\epsilon \geq \epsilon_{FL}(r)$, then it holds that $E_{opt}(r,\epsilon) \asymp E_{opt}(r)$. That is, no accuracy / robustness tradeoff is needed when the excess risk level ϵ is greater than the threshold $\epsilon_{FL}(r)$: there is always an ϵ-accurate model which achieves the absolute optimal (up to within multiplicative absolute constants) adversarial risk $E_{opt}(r)$.*

Proof. Write $\overline{E}_{opt}(r,\epsilon) := \inf_{w \in \mathcal{W}_\epsilon} \overline{E}(w,r)$, where we recall that

$$\mathcal{W}_\epsilon := \{w \in \mathbb{R}^d \mid \Delta(w) \leq \epsilon^2\} = \{w \in \mathbb{R}^d \mid \|w - w_0\|_\Sigma^2 \leq \epsilon^2 \|w_0\|_\Sigma^2\}.$$

Thus, if $\epsilon \geq \epsilon_{FL}(r) := \sqrt{\Delta(w^{prox}(r^2))} = \sqrt{G(r^2)}/\|w_0\|_\Sigma$, then $w^{prox}(r^2) \in \mathcal{W}_\epsilon$, and so we deduce from Lemma 3.1 that

$$E_{opt}(r,\epsilon) \asymp \overline{E}_{opt}(r,\epsilon) = \overline{E}(w^{prox}(r^2), r) \asymp E(w^{prox}(r^2), r). \tag{24}$$

Henceforth, suppose $0 \leq \epsilon \leq \epsilon_{FL}(r)$. First observe that, for every $\lambda \in [0, r^2]$,

$$F(r,\lambda) = \inf_{\|w - w_0\|_\Sigma^2 \leq G(\lambda)} \overline{E}(w,r). \tag{25}$$

Indeed, let $w \in \mathbb{R}^d$ such that $\|w - w_0\|_\Sigma^2 \leq G(\lambda) := \|w^{prox}(\lambda) - w_0\|_\Sigma^2$. Let $t \geq 0$ such that $\lambda = r^2/(1+t)$. By definition of $w^{prox}(\lambda)$ in (9), one has

$$\begin{aligned}
F(r,\lambda) + tG(\lambda) &= \|w^{prox}(\lambda) - w_0\|_\Sigma^2 + r^2\|w^{prox}(\lambda)\|_\star^2 + t\|w^{prox}(\lambda) - w_0\|_\Sigma^2 \\
&= (1+t)(\|w^{prox}(\lambda) - w_0\|_\Sigma^2 + \lambda\|w^{prox}(\lambda)\|_\star^2) \\
&\leq (1+t)(\|w - w_0\|_\Sigma^2 + \lambda\|w\|_\star^2) \\
&\leq (1+t)(\|w - w_0\|_\Sigma^2 + \lambda\|w\|_\star^2) \text{ by definition of } w^{prox}(\lambda) \\
&= \overline{E}(w,r) + t\|w - w_0\|_\Sigma^2,
\end{aligned}$$

and it follows that $\overline{E}(w^{prox}(\lambda), r) = F(r,\lambda) \leq \overline{E}(w,r) + t(\|w - w_0\|_\Sigma^2 - G(\lambda)) \leq \overline{E}(w,r)$.

Now, equipped with (25) and the definition of $\overline{E}_{opt}(r,\epsilon)$, observe that

- If $G(\lambda) \leq \epsilon^2\|w_0\|_\Sigma^2$, then $\overline{E}_{opt}(r,\epsilon) \geq F(r,\lambda)$.
- Analogously, if $G(\lambda) \geq \epsilon^2\|w_0\|_\Sigma^2$, then $\overline{E}_{opt}(r,\epsilon) \leq F(r,\lambda)$.

On the other hand, Lemma 4.2 tells us that the equation $G(\lambda) = \epsilon^2\|w_0\|_\Sigma^2$ has a unique solution $\lambda_{opt}(r,\epsilon)$ in $[0, r^2]$. Thus, $\overline{E}_{opt}(r,\epsilon) = \overline{E}(w^{prox}(\lambda_{opt}(r,\epsilon)), r)$ and we deduce from Lemma 3.1 that

$$E_{opt}(r,\epsilon) \asymp \overline{E}_{opt}(r,\epsilon) = \overline{E}(w^{prox}(\lambda_{opt}(r,\epsilon)), r) \asymp E(w^{prox}(\lambda_{opt}(r,\epsilon)), r),$$

which completes the proof. □

C.1. Proof of Theorem 3.1

Theorem 3.1. *With $\lambda = r^2$, it holds that $E_{opt}(r) \asymp E(w^{prox}(\lambda), r) \asymp \sigma^2 + F(r, r^2)$. That is, up to within multiplicative absolute constants, $w^{prox}(\lambda = r^2)$ attains the optimal adversarial risk $E_{opt}(r)$.*

Proof. Indeed, by Lemma 4.1 we know that $E_{opt}(r) = E_{opt}(r,1)$. Also, by Lemma 4.2, we know that $\lambda_{opt}(r,1) = r^2$. Combining these with part (A) of Theorem 4.1 then gives the result. □

C.2. A Corollary: Isotropic Features

As an important corollary to Theorem 4.1 (not stated in the main manuscript), consider the case of isotropic features considered in (Javanmard et al., 2020), where $\Sigma = I_d$.

Theorem C.1. *Consider the isotropic setting where $\Sigma = I_d$. For Euclidean-norm attack of strength $r \ge 0$, it holds for any tolerance level $\epsilon \in [0,1]$ that*

(i) ***(Free Lunch Threshold)*** $\epsilon_{FL}(r) = r^2/(1+r^2) \in [0,1)$.

(ii) ***(Free Lunch)*** *If $\epsilon \ge \epsilon_{FL}(r)$, then the optimal regularization is $\lambda_{opt}(r,\epsilon) = r^2$, and we have*

$$E_{opt}(r,\epsilon) \asymp E_{opt}(r) \asymp \sigma^2 + \|w_0\|_2^2 \min(r^2,1). \tag{26}$$

(ii) ***(Accuracy / Robustness Tradeoff)*** *If $\epsilon < \epsilon_{FL}(r)$, then the optimal regularization parameter is given by $\lambda_{opt}(r,\epsilon) = \epsilon/(1-\epsilon)$, and we have*

$$E_{opt}(r,\epsilon) \asymp \sigma^2 + \|w_0\|_2^2(\epsilon^2 + (1-\epsilon)^2 \min(r^2,1)). \tag{27}$$

D. Structure of Optima

In this section, we explore the structure of the curve $\lambda \mapsto w^{prox}(\lambda)$ given in (9) for some notable choices of the attacker's norm $\|\cdot\|$.

D.1. Mahalanobis-Norm Attacks

Suppose the attacker's norm $\|\cdot\|$ is the Mahalanobis norm $\|\cdot\|_B$ induced by a positive-definite $d \times d$ matrix B. Then, for any $\lambda \ge 0$, $w^{prox}(\lambda)$ minimizes $\|w - w_0\|_\Sigma^2 + \lambda\|w\|_\star^2 = \|w-w_0\|_\Sigma^2 + \lambda\|w\|_{B^{-1}}^2$, which gives the closed-form solution

$$w^{prox}(\lambda) = (\Sigma + \lambda B^{-1})^{-1}\Sigma w_0 = (B\Sigma + \lambda I_d)^{-1}B\Sigma w_0. \tag{28}$$

Also, note that the functions F and G defined in (10) are now reduced to

$$\begin{aligned} G(\lambda) &:= \|w^{prox}(\lambda) - w_0\|_\Sigma^2 = \|((\Sigma + \lambda B^{-1})^{-1}\Sigma - I_d)w_0\|_\Sigma^2 \\ &= \lambda^2\|(B\Sigma + \lambda I_d)^{-1}w_0\|_\Sigma^2 \\ F(r,\lambda) - G(\lambda) &= r^2\|w^{prox}(\lambda)\|_2^2 = r^2\|(\Sigma + \lambda B^{-1})^{-1}\Sigma w_0\|_2^2 \\ &= r^2\|(B\Sigma + \lambda I_d)^{-1}B\Sigma w_0\|_2^2. \end{aligned} \tag{29}$$

Link with (Scetbon & Dohmatob, 2023). Note that (28) recovers the structure established in (Scetbon & Dohmatob, 2023), where $1/\lambda$ should be thought of as the time parameter in the population-wise *adapted* (i.e. pre-conditioned) gradient-flow (GD+) proposed in that work, with the choise $M = B^{1/2}$. We deduce the following:

- GD+ started from zero and run for time $O(1/r^2)$ achieves the optimal adversarial risk $E_{opt}(r)$ (up to within multiplicative absolute constants). This follows from Theorem 3.1 and the preceding argument.
- More generally, for a tolerance parameter $\epsilon \in [0,1]$, GD+ started from zero and run for time $O(1/\lambda_{opt}(r,\epsilon))$ achieves the optimal adversarial risk, where $\lambda_{opt}(r,\epsilon) \in [0, r^2]$ is as given in Lemma 4.2.

Link with (Xing et al., 2021). In particular, for Euclidean-norm attacks corresponding to $B = I_d$, (28) reduces to

$$w^{prox}(\lambda) = (\Sigma + \lambda I_d)^{-1}\Sigma w_0, \tag{30}$$

which recovers the structure established in (Xing et al., 2021).

D.2. ℓ_p-norm Attacks on Diagonal Feature Covariance Matrix

Suppose the feature covariance matrix is $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ and the attacker's norm is $\|\cdot\|_p$, for some $p \in [1, \infty]$. Let $q \in [1, \infty]$ be the harmonic conjugate of p. By (9), $w^{prox}(\lambda)$ is the minimizer of $\|w - w_0\|_\Sigma^2 + \lambda\|w\|_q^2$ over $w \in \mathbb{R}^d$. If $R_q(\lambda) := \|w^{prox}(\lambda)\|_q$, then by first order optimality conditions, we have $\Sigma w_0 \in \Sigma w + \lambda R(\lambda)\partial\|\cdot\|_q(w)$, i.e

$$w^{prox}(\lambda) = (I + R_q(\lambda)\lambda\partial\|\cdot\|_q)^{-1}(\Sigma w_0) = \mathrm{prox}_{t_q(\lambda)\|\cdot\|_q}(\Sigma w_0), \tag{31}$$

where $t_q(\lambda) := R_q(\lambda)\lambda$.

The Case of ℓ_∞-Norm Attacks. In the special case where $p = \infty$, we have $q = 1$, giving

$$w^{prox}(\lambda)_k = \mathrm{ST}(\mu_k; t_1(\lambda)) = \begin{cases} \mu_k + t_1(\lambda), & \text{if } \mu_k < -t_1(\lambda), \\ 0, & \text{if } |\mu_k| \le t_1(\lambda), \\ \mu_k - t_1(\lambda), & \text{if } \mu_k > t_1(\lambda), \end{cases} \tag{32}$$

where $\mu_k := \lambda_k \cdot (w_0)_k$ for all $k \in [d]$ (i.e $\mu = \Sigma w_0$) and ST is the well-known *soft-thresholding (ST)* operator. Thus, if $t_1(\lambda) \ge \|\mu\|_\infty$, then $w^{prox}(\lambda) = 0$. This means that we can always restrict our search of the optimal threshold $t_1(\lambda)$ to a compact interval,

$$t_1(\lambda) \in [0, \|\mu\|_\infty]. \tag{33}$$

The structure of the optimal (32) is instructive: components $(w_0)_k$ of w_0 corresponding to to features with small values of $|\mu_k|$ are zeroed-out.

E. Well-Conditioned Problems

E.1. Estimating $E_{opt}(r)$

We now give a complete analysis of $E_{opt}(r)$ for the case of so-called "well-conditioned" problems (formally defined later). To develop an intuition, first consider the simple case of Euclidean-norm attacks on isotropic features (i.e. $\Sigma = I_d$). In this case, Lemma I.1 tells us that

$$G(\lambda) = \lambda^2\|w_0\|_2^2/(1+\lambda)^2,\ F(r, \lambda) = G(\lambda) + r^2\|w_0\|_2^2/(1+\lambda)^2. \tag{34}$$

Theorem 3.1 then predicts that

$$\begin{aligned} E_{opt}(r) &\asymp \sigma^2 + F(r, r^2) = \sigma^2 + \frac{r^4}{(1+r^2)^2}\|w_0\|_2^2 + \frac{r^2}{(1+r^2)^2}\|w_0\|_2^2 \\ &= \sigma^2 + \frac{r^2}{1+r^2}\|w_0\|_2^2 \asymp \sigma^2 + \|w_0\|_2^2\min(r^2, 1) \asymp \min(E(w_0, r), E(0, r)). \end{aligned} \tag{35}$$

Thus, for $r \le 1$, the generative model w_0 attains the optimal adversarial risk $E_{opt}(r)$ (upto within multiplicative constant); for $r \ge 1$, the optimal adversarial risk is attained by the null model $w = 0$. This recovers a result of (Scetbon & Dohmatob, 2023).

We now consider the situation of general norms and covariance matrices. Define $r_0, r_1, \eta > 0$ by

$$r_0 := \|w_0\|_\Sigma/\|w_0\|_\star,\ r_1 := \|\Sigma w_0\|/\|w_0\|_\Sigma,\ \eta_0 := r_1/r_0. \tag{36}$$

Note that $\eta_0 \ge 1$ by Cauchy-Schwarz inequality. This scalar should be thought of as a kind of *condition number* for Σ w.r.t the attacker's norm $\|\cdot\|$. In particular, when this norm is Euclidean, then η_0 is upper-bounded by the usual linear-algebraic condition number of Σ. In general, $r_1 = r_0\eta_0 \ge r_0$, with equality when $\eta_0 = 1$, which is the case when for example one considers Euclidean-norm attacks on isotropic features, i.e. $\Sigma = I_d$.

Definition E.1. *By* well-conditioned *problems, we mean scenarios where $\eta_0 = O(1)$.*

Theorem E.1. *For an attack strength $r \geq 0$ w.r.t a general norm $\|\cdot\|$, it holds that*

$$\sigma^2 + \|w_0\|_\Sigma^2 \min(r/r_1, 1)^2 \lesssim E_{opt}(r) \lesssim \sigma^2 + \|w_0\|_\Sigma^2 \min(r/r_0, 1)^2. \tag{37}$$

In particular, for well-conditioned problems (i.e. $\eta_0 = O(1)$), it holds that

$$E_{opt}(r) \asymp \sigma^2 + \|w_0\|_\Sigma^2 \min(r/r_0, 1)^2 \asymp \min(E(w_0, r), E(0, r)). \tag{38}$$

Thus, for well-conditioned problems, the generative model w_0 is optimally robust (up to an absolute multiplicative constant) for small values of r (i.e. $r \leq r_0$), while for large values of r ($r \geq r_0$), the null model $w = 0$ is optimally robust (up to an absolute multiplicative constant).

E.2. Estimating $E_{opt}(r, \epsilon)$

We now generalize the results of the previous subsection and establish some results which are complementary to the results in Sections 4. These use different techniques but arrive at qualitatively and quantitatively similar results in certain settings.

Define an auxiliary function $H : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ by

$$H(r, \epsilon) := \begin{cases} r, & \text{if } 0 \leq r < 1, \\ \delta + (1-\delta) r, & \text{else,} \end{cases}, \tag{39}$$

where $\delta = \delta(\epsilon) := \min(1, \epsilon)$. This is the same function which appears in Theorem 5.1 (see Table 1). Note that $r = H(r, 0) = H(r, \epsilon) \geq H(r, 1) = \min(r, 1)$, for all $r \geq 0$ and $\epsilon \in [0, 1]$. The following which holds for any attacker norm, is one of our main contributions.

Theorem E.2. *For any $r \geq 0$ and $\epsilon \in [0, 1]$, the following bounds hold*

$$\sigma^2 + \|w_0\|_\Sigma^2 H(r/r_1, \epsilon)^2 \lesssim E_{opt}(r, \epsilon) \lesssim \sigma^2 + \|w_0\|_\Sigma^2 H(r/r_0, \epsilon)^2, \tag{40}$$

where r_0 and r_1 are as defined in (36). In particular, if $r_0 \asymp r_1$, then $E_{opt}(r, \epsilon) \asymp \sigma^2 + \|w_0\|_\Sigma^2 H(r/r_0, \epsilon)^2$.

Note that when $\epsilon = 1$, we have $H(r, \epsilon) = H(r, 1) = \min(r, 1)$ for any $r \geq 0$, and the above result recovers Theorem E.1 as a special case.

E.3. The Case of Euclidean-Norm Attacks

In the special case of Euclidean-norm attacks, one computes

$$1 \leq \eta_0 = \frac{r_0}{r_1} = \frac{\|\Sigma w_0\|_2 \|w_0\|_2}{\|w_0\|_\Sigma^2} = \frac{\|\Sigma^{1/2}\Sigma^{1/2} w_0\|_2}{\|\Sigma^{1/2} w_0\|_2} \frac{\|w_0\|_2}{\|\Sigma^{1/2} w_0\|_2} \leq \sqrt{\kappa(\Sigma)},$$

where $\kappa(\Sigma)$ is the ordinary condition number of the covariance matrix Σ.

A Well-Conditioned Example. In particular, in the case of isotropic features where $\Sigma = I_d$, we have $r_0 = r_1 = \|\Sigma^{1/2}\|_{op} = 1$ and $\eta_0 = 1$, and Theorem E.2 then gives

$$E_{opt}(r, \epsilon) \asymp \sigma^2 + \|w_0\|_\Sigma^2 \cdot H(r/r_0, \epsilon)^2 = \sigma^2 + (\epsilon^2 + (1-\epsilon)^2 r^2) \|w_0\|_\Sigma^2, \tag{41}$$

for all $r \geq r_0$. This is exactly the result obtained in Theorem C.1.

A Case of Failure for Theorem E.2. Note that even in the Euclidean case, Theorem E.2 might become vacuous when the "condition number" η_0 is too large. This is for example, the case of polynomial decaying eigenvalues of the covariance matrix Σ, considered in Section 5.2. Indeed, that example with $\delta \in (1, \infty)$, one easily computes $r_0 = \|w_0\|_\Sigma / \|w_0\|_2 \to 0$ in the limit $\delta \to 1^+$, and $r_1 = \|\Sigma w_0\|_2 / \|w_0\|_\Sigma = \Theta(1)$, and the lower-bound in (40) becomes vacuous. However, the results of Section 5.2 (Theorem 5.2) remain valid even in this ill-conditioned limit.

E.4. Sketch of Proof of Theorem E.2

We now outline the key ideas underlying the proof of Theorem E.2, split into various steps. The details are provided in Section F.

Step 1: Proxy for Adversarial Risk. From Lemma 3.1, we know that $E(w,r) \asymp \widetilde{E}(w,r)$, and so

$$E_{opt}(r,\epsilon) \asymp \sigma^2 + K_{opt}(r,\epsilon)^2, \tag{42}$$

where $\widetilde{E}(w,r) := \|w-w_0\|_\Sigma^2 + r^2\|w\|_\star^2$ and $K_{opt}(r,\epsilon) := \inf_{w\in\mathcal{W}_\epsilon} K(w,r)$.

Step 2: Restricting the Search to a Chord. Computing $K_{opt}(r,\epsilon)$, even though conceivably easier than $E_{opt}(r,\epsilon)$, is still difficult. Instead, we will restrict the optimization to a line / chord in $\mathcal{W}_\epsilon$, parallel to the generative model w_0. It will turn out that up to within multiplicative constants, this strategy gives the correct value of $K_{opt}(r,\epsilon)$ as a function of all relevant problem parameters. To this end, let $K_{shrink}(r,\epsilon)$ be the optimal adversarial risk achieved by a linear model which is co-linear with the generative model w_0, i.e

$$K_{shrink}(r,\epsilon) := \inf_{w\in\mathcal{W}_\epsilon\cap\langle w_0\rangle} K(w,r), \tag{43}$$

where $\langle w_0\rangle := \{tw_0 \mid t\in\mathbb{R}\}$ is the one-dimensional subspace of $\mathbb{R}^d$ spanned by the generative model w_0.

Proposition E.1. *For any $r,\epsilon \geq 0$, the following bounds hold*

$$K_{shrink}(r\eta_0,\epsilon) \leq K_{opt}(r,\epsilon) \leq K_{shrink}(r,\epsilon). \tag{44}$$

This auxiliary result, which is the main component of the proof of Theorem E.2, is proved in Section F.

Step 3: Computing the Value of $K_{shrink}(r,\epsilon)$. To complete the proof of Theorem E.2, it remains to show that the proxy $K_{shrink}(r,\epsilon)$ equals $H(r/r_0,\epsilon)^2$ up to within multiplicative absolute constants. The proof of Theorem E.2 would then follow upon plugging such estimates into the bounds given in Proposition E.1.

Proposition E.2. *For any $r,\epsilon \geq 0$, it holds that*

$$\frac{K_{shrink}(r,\epsilon)}{\|w_0\|_\Sigma^2} \asymp \frac{K(t_{opt}w_0,r)}{\|w_0\|_\Sigma^2} \asymp H(\frac{r}{r_0},\epsilon)^2, \tag{45}$$

where where the function H is as defined in (39), and r_0 is the scalar defined in (36), and $t_{opt} = t_{opt}(r/r_0,\epsilon) \in [0,1]$ is the optimaland where the function T is as defined in (46).

Comparing Propositions E.1 and E.2, it becomes clear how the function H and T enter the bounds in Theorem E.2.

F. Details of the Proof of Theorem E.2

Theorem E.2. *For any $r \geq 0$ and $\epsilon \in [0,1]$, the following bounds hold*

$$\sigma^2 + \|w_0\|_\Sigma^2 H(r/r_1,\epsilon)^2 \lesssim E_{opt}(r,\epsilon) \lesssim \sigma^2 + \|w_0\|_\Sigma^2 H(r/r_0,\epsilon)^2, \tag{40}$$

where r_0 and r_1 are as defined in (36). In particular, if $r_0 \asymp r_1$, then $E_{opt}(r,\epsilon) \asymp \sigma^2 + \|w_0\|_\Sigma^2 H(r/r_0,\epsilon)^2$.

The proof was sketched in Section E.4 with the help of auxiliary propositions, namely Proposition E.1 and E.2. Here we just need to provide the proofs for these propositions.

F.1. Proof of Proposition E.2

Proposition E.2. *For any $r,\epsilon \geq 0$, it holds that*

$$\frac{K_{shrink}(r,\epsilon)}{\|w_0\|_\Sigma^2} \asymp \frac{K(t_{opt}w_0,r)}{\|w_0\|_\Sigma^2} \asymp H(\frac{r}{r_0},\epsilon)^2, \tag{45}$$

where where the function H is as defined in (39), and r_0 is the scalar defined in (36), and $t_{opt} = t_{opt}(r/r_0,\epsilon) \in [0,1]$ is the optimaland where the function T is as defined in (46).

Proof. Define an auxiliary function $t_{opt} : \mathbb{R}_+^2 \to \mathbb{R}_+$ by

$$t_{opt}(r,\epsilon) := \begin{cases} 1, & \text{if } 0 \le r < 1, \\ 1-\delta, & \text{if } r \ge 1, \end{cases} \tag{46}$$

with $\delta = \delta(\epsilon) := \min(1,\epsilon)$ as before. Also define $K_{shrink}(r,\epsilon)$ by

$$K_{shrink}(r,\epsilon) := \inf_{w \in \mathcal{W}_\epsilon \cap \langle w_0 \rangle} K(w,r). \tag{47}$$

By definition of the set $\mathcal{W}_\epsilon$, note that $w \in \mathcal{W}_\epsilon \cap \langle w_0 \rangle$ iff $w = tw_0$ for some $t \in \mathbb{R}$ such that $|t-1| \le \epsilon$. Thus, noting that $K(w,r) := \|w-w_0\|_\Sigma^2 + r^2\|w\|_\star^2 \asymp (\|w-w_0\|_\Sigma + r\|w\|_\star)^2$, one computes

$$\begin{aligned}\sqrt{K_{shrink}(r,\epsilon)} &:= \inf_{w \in \mathcal{W}_\epsilon \cap \langle w_0 \rangle} \sqrt{K(w,r)} \\ &\asymp \inf_{|t-1|\le\epsilon} \|w_0\|_\Sigma |t-1| + r\|w_0\|_\star |t| \\ &= \|w_0\|_\Sigma \cdot \inf_{|t-1|\le\epsilon} |t-1| + \frac{r\|w_0\|_\star}{\|w_0\|_\Sigma}|t| \\ &= \|w_0\|_\Sigma \cdot \inf_{|t-1|\le\epsilon} |t-1| + \frac{r}{r_0}|t| \\ &= \|w_0\|_\Sigma \cdot H(r/r_0,\epsilon),\end{aligned}$$

where the last line is thanks to Lemma G.3. □

F.2. Proof of Proposition E.1

Proposition E.1. *For any $r,\epsilon \ge 0$, the following bounds hold*

$$K_{shrink}(r\eta_0,\epsilon) \le K_{opt}(r,\epsilon) \le K_{shrink}(r,\epsilon). \tag{44}$$

Proof. Let $C = \langle w_0 \rangle := \{tw_0 \mid t \in \mathbb{R}\} \subseteq \mathbb{R}^d$ be the one-dimensional subspace spanned by the generative model w_0, and let $P_{C,\Sigma} : \mathbb{R}^d \to C$ be the projection operator onto C, w.r.t the the Mahalanobis norm $\|\cdot\|_\Sigma$. Then, by non-expansiveness of $P_{C,\Sigma}$ (see (Bauschke & Combettes, 2011), for example), one has for any $w \in \mathbb{R}^d$,

$$\|P_{C,\Sigma}(w) - w_0\|_\Sigma = \|P_{C,\Sigma}(w) - P_{C,\Sigma}(w_0)\|_\Sigma \le \|w-w_0\|_\Sigma.$$

Now, for any $w \in \mathbb{R}^d$, we have $P_{C,\Sigma}(w) = tw_0$, where $t \in \mathbb{R}$ minimizes $f(t) := \|w-tw_0\|_\Sigma^2$. Now, $f'(t) = 2w_0^\top\Sigma(tw_0 - w) = 2(\|w_0\|_\Sigma^2 t - w^\top\Sigma w_0)$. Thus, the optimal t is $w^\top\Sigma w_0/\|w_0\|_\Sigma^2$, and so

$$P_{C,\Sigma}(w) = \frac{w^\top\Sigma w_0}{\|w_0\|_\Sigma^2} w_0. \tag{48}$$

Let us now bound the operator norm of $P_{C,\Sigma}$ w.r.t the dual norm $\|\cdot\|_\star$. For any $w \in \mathbb{R}^d$, one has

$$\frac{\|P_{C,\Sigma}(w)\|_\star}{\|w\|_\star} = \frac{\|(w^\top\Sigma w_0)w_0\|_\star}{\|w\|_\star\|w_0\|_\Sigma^2} = \frac{|w^\top\Sigma w_0|}{\|w\|_\star\|w_0\|_\Sigma}\frac{\|w_0\|_\star}{\|w_0\|_\Sigma} \le \frac{\|\Sigma w_0\|}{\|w_0\|_\Sigma}\frac{\|w_0\|_\star}{\|w_0\|_\Sigma} =: \eta_0,$$

where the second line is an application of the Cauchy-Schwarz inequality. We deduce that

$$\begin{aligned}\sqrt{K(P_{C,\Sigma}(w),r)} &\asymp \|P_{C,\Sigma}(w) - w_0\|_\Sigma + r\|P_{C,\Sigma}(w)\|_\star \\ &\le \|w-w_0\|_\Sigma + r\eta_0\|w\|_\star \asymp \sqrt{K(w,r\eta_0)}.\end{aligned}$$

Thus, $K(w,r) \gtrsim K(P_{C,\Sigma}(w), r/\eta_0)$. On the other hand, if $w \in \mathcal{W}_\epsilon$, then the non-expansiveness of $P_{C,\Sigma}$ (again!) gives

$$\|P_{C,\Sigma}(w) - w_0\|_\Sigma = \|P_{C,\Sigma}(w) - P_{C,\Sigma}(w_0)\|_\Sigma \le \|w-w_0\|_\Sigma \le \epsilon\|w_0\|_\Sigma,$$

that is, $P_{C,\Sigma}(w) \in \mathcal{W}_\epsilon$. Putting things together yields: for any $w \in \mathcal{W}_\epsilon$, there exists $z \in \mathcal{W}_\epsilon \cap C$ such that $K(z, r/\eta_0) \leq K(w, r)$. Therefore,

$$K_{opt}(r, \epsilon) := \inf_{w \in \mathcal{W}_\epsilon} K(w, r) \gtrsim \inf_{z \in \mathcal{W}_\epsilon \cap C} K(z, r/\eta_0) =: K_{shrink}(r/\eta_0, \epsilon).$$

This establishes the lower-bound Proposition E.1.

As for the upper-bound, one computes

$$K_{opt}(r, \epsilon) := \inf_{w \in \mathcal{W}_\epsilon} K(w, r) \leq \inf_{w \in \mathcal{W}_\epsilon \cap C} K(w, r) =: K_{shrink}(r, \epsilon),$$

as claimed. □

G. Technical Proofs

G.1. Proof of Lemma 2.1: Analytic Formula for Adversarial Risk

Lemma 2.1. *For any $w \in \mathbb{R}^d$ and $r \geq 0$, it holds that $E(w, r) = E(w) + r^2\|w\|_\star^2 + 2\sqrt{2/\pi} r \|w\|_\star \sqrt{E(w)}$.*

For the proof, we will need the following auxiliary lemma.

Lemma G.1. *For any $x, w \in \mathbb{R}^d$, $r \geq 0$, and $y \in \mathbb{R}$, the following identity holds*

$$\sup_{\|\delta\| \leq r} |(x+\delta)^\top w - y| = |x^\top w - y| + r\|w\|_\star. \tag{49}$$

Proof. Note that $h(x, y, \delta)/2 = \eta(x, y)^2/2 + g(x, y, \delta)/2$, where $g(x, y, \delta) := w(\delta)^2 - 2\eta(x, y)w(\delta)$, and $\eta(x, y) := w(x) - y$, and $w(x) := x^\top w$. Now, because the real function $z \mapsto z^2/2$ is its own Fenchel-Legendre conjugate, we can ”dualize” our problem as follows

$$\begin{aligned}
\sup_{\|\delta\| \leq r} g(x, y, \delta)/2 &= \sup_{\|\delta\|_\star \leq r} -\eta(x, y)w(\delta) + \sup_{z \in \mathbb{R}} zw(\delta) - z^2/2 \\
&= \sup_{z \in \mathbb{R}} -z^2/2 + \sup_{\|\delta\| \leq r} (z - \eta(x, y))w(\delta) \\
&= \sup_{z \in \mathbb{R}} r\|w\|_\star |z - \eta(x, y)| - z^2/2 \\
&= \sup_{s \in \{\pm 1\}} \sup_{z \in \mathbb{R}} rs(z - \eta(x, y)) - z^2/2 \\
&= \sup_{s \in \{\pm 1\}} -r\|w\|_\star s\eta(x, y) + \sup_{z \in \mathbb{R}} r\|w\|_\star sz - z^2/2 \\
&= \sup_{s \in \{\pm 1\}} -r\|w\|_\star s\eta(x, y) + r^2\|w\|_\star^2/2 \\
&= r\|w\|_\star |\eta(x, y)| + r^2\|w\|_\star^2/2.
\end{aligned}$$

We deduce that

$$\sup_{\|\delta\| \leq r} h(x, y, \delta)/2 = \eta(x, y)^2/2 + r\|w\|_\star |\eta(x, y)| + r^2\|w\|_\star^2/2 = (|\eta(x, y)| + r\|w\|_\star)^2/2,$$

from which the result follows. □

Proof of Lemma 2.1. Indeed, thanks to Lemma G.1, one has

$$E(w, r) := \mathbb{E} \sup_{\|\delta\| \leq r} h(x, y, \delta) = \mathbb{E}[(\eta(x, y) + r\|w\|_\star)^2], \tag{50}$$

where the functions h and η are as in the proof of Lemma G.1. The result then follows upon noting that, for $x \sim N(0, \Sigma)$ and $y|x \sim N(x^\top w_0, \sigma^2)$,

$$\begin{aligned}\mathbb{E}[\eta(x,y)^2] &= \mathbb{E}[(x^\top w - y)^2] = E(w) = \|w - w_0\|_\Sigma^2 + \sigma^2,\\ \mathbb{E}|\eta(x,y)| &= \mathbb{E}_x|x^\top w - y| = c_0\sqrt{\|w - w_0\|_\Sigma^2 + \sigma^2} = c_0\sqrt{E(w)},\end{aligned}$$

where $c_0 := \sqrt{2/\pi}$ as in the lemma. □

G.2. Proof of Lemma 3.1

Lemma 3.1. *There exists absolute constants c_1 and c_2 such that for a general attacker norm $\|\cdot\|$, and $w \in \mathbb{R}^d$, $r \geq 0$,*

$$\begin{aligned}\widetilde{E}(w,r) \leq E(w,r) \leq c_1\widetilde{E}(w,r),\\ \overline{E}(w,r) \leq E(w,r) \leq c_2\overline{E}(w,r).\end{aligned} \tag{8}$$

We will need the following elementary lemma.

Lemma G.2. *For any $a, b, c \geq 0$ with $c \leq 1$, it holds that*

$$\begin{aligned}(a+b)^2 \geq a^2 + b^2 + 2abc \geq \frac{1+c}{2}(a+b)^2,\\ a^2 + b^2 \leq a^2 + b^2 + 2abc \leq (1+c)(a^2+b^2).\end{aligned} \tag{51}$$

Proof. Let $h(a,b,c) := a^2 + b^2 + 2abc$. For the LHS, it suffices to observe that $h(a,b,c) \leq h(a,b,1) = (a+b)^2$. For the RHS, WLOG assume that $a \neq 0$, and set $t := b/a \geq 0$. Observe

$$1 \geq \frac{h(a,b,c)}{(a+b)^2} = \frac{1+t^2+2ct}{(1+t)^2},$$

and the RHS is minimized when $t = 1$, because $0 \leq c \leq 1$ by assumption. We deduce that $h(a,b,c)/(a+b)^2 \geq (1+1+2c)/(1+1)^2 = (1+c)/2$. This proves the first line of inequalities in the lemma.

On the other hand, observe that

$$1 \leq \frac{h(a,b,c)}{a^2+b^2} = \frac{1+t^2+2ct}{1+t^2} = 1 + \frac{2ct}{1+t^2}. \tag{52}$$

Clearly, the RHS attains a maximum value of $1 + c$ at $t = 1$. This proves the second line of inequalities in the lemma. □

We are now ready to proof Lemma 3.1.

Proof of Lemma 3.1. From Lemma G.2 above applied with $c = c_0 = \sqrt{2/\pi}$, we deduce that Lemma 3.1 holds with $c_1 = 2/(1+c_0) \approx 1.11$ and $c_2 = 1 + c_0 \approx 1.8$. □

G.3. Proof of Lemma 4.1

Lemma 4.1. *For any $r \geq 0$ and $\epsilon \geq 1$, it holds that $E_{opt}(r,\epsilon) = E_{opt}(r)$.*

Proof. Recall that $E_{opt}(r) := \inf_{w \in \mathbb{R}^d} E(w,r)$ and $E_{opt}(r,\epsilon) := \inf_{w \in \mathcal{W}_\epsilon} E(w,r)$, where

$$\mathcal{W}_\epsilon := \{w \in \mathbb{R}^d \mid \|w - w_0\|_\Sigma \leq \epsilon\|w_0\|_\Sigma\}.$$

Observe that, if $w \in \mathbb{R}^d \setminus \mathcal{W}_1$, then $\|w - w_0\|_\Sigma > \|w_0\|_\Sigma$. We deduce that

$$E(w,r) \geq E(w) = \|w - w_0\|_\Sigma^2 + \sigma^2 > \|w_0\|_\Sigma^2 + \sigma^2 = E(0) = E(0,r).$$

On the other hand, if $\epsilon \geq 1$, then $0 \in \mathcal{W}_1 \subseteq \mathcal{W}_\epsilon$. Combining with the above inequality gives

$$\begin{aligned} E_{opt}(r,\epsilon) &= \inf_{w \in \mathcal{W}_\epsilon} E(w,r) = \min\left(\inf_{w \in \mathcal{W}_1} E(w,r), \inf_{w \in \mathcal{W}_\epsilon \setminus \mathcal{W}_1} E(w,r)\right) \\ &= \inf_{w \in \mathcal{W}_1} E(w,r) =: E_{opt}(r,1). \end{aligned}$$

and the proof is complete. □

G.4. Proof of Lemma 4.2

Lemma 4.2. *For any $r \geq 0$ and $0 \leq \epsilon \leq \epsilon_{FL}(r)$, the scalar equation*

$$G(\lambda) = \epsilon^2 \|w_0\|_\Sigma^2 \tag{14}$$

has a unique solution $\lambda_{opt}(r,\epsilon)$ in $[0,r^2]$.

We extend the definition of $\lambda_{opt}(r,\epsilon)$ to all $\epsilon \in [0,1]$ by setting $\lambda_{opt}(r,\epsilon) = r^2$ whenever $\epsilon \geq \epsilon_{FL}(r)$.

Proof. Indeed, thanks to (Sra, 2011, Lemma 4) and the definition of $w^{prox}(\lambda)$ in (9) and $G(\lambda)$ in (10), the function G is increasing on $[0,r^2]$ with minimal value $G(0) =$ and maximal value $G(r^2) = \epsilon_{FL}(r)\|w_0\|_\Sigma^2$. Thus, if $0 \leq \epsilon \leq \epsilon_{FL}(r)$, then $0 \leq \epsilon^2\|w_0\|_\Sigma^2 \leq G(r^2)$, and so $\epsilon^2\|w_0\|_\Sigma^2$ is in the range of G over $\lambda \in [0,r^2]$. □

G.5. On the Auxiliary Function H

Remark G.1. *The following properties of the function H are easily verified*

(i) $H(r,\epsilon) \geq \delta + (1-\delta)r$ *for all* $r \geq 1$ *and* $\epsilon \geq 0$.

(ii) $H(r,\epsilon) = H(r,1) = \min(r,1)$ *for all* $r \geq 0$ *and* $\epsilon \geq 1$.

(iii) $H(\eta r,\epsilon) \geq \eta H(r,\epsilon)$ *for all* $r,\epsilon,\eta \geq 0$.

The functions H and T are linked by the following lemma.

Lemma G.3. *For any $r,\epsilon \geq 0$, we have*

$$\inf_{|t-1|\leq\epsilon} k(t) = H(r,\epsilon) = k(t_{opt}(r,\epsilon)), \tag{53}$$

where $k : \mathbb{R} \to \mathbb{R}$ is the function defined by $k(t) := |t-1| + r|t|$.

Proof. First notice that $k(-t) \geq k(t)$ and $|-t-1| = t+1 \geq |t-1|$ if $t \geq 0$. Thus, WLOG we may assume $t \geq 0$ in the optimization problem. Now, consider the change of variable $t = t(u) := 1 - \sqrt{u}$, for $u \in [0,1]$, so that $k(t) = h(u) := \sqrt{u} + r(1-\sqrt{u})$. Note that $|t-1| \leq \epsilon$ iff $0 \leq u \leq \delta^2$, where $\delta = \delta(\epsilon) := \min(1,\epsilon)$.

Now, $2h'(u) = (1-r)/\sqrt{u}$. We deduce that h is non-decreasing if $r \in [0,1)$ and non-increasing if $r \geq 1$. Thus,

$$\begin{aligned} \inf_{|t-1|\leq\epsilon} k(t) = \inf_{0\leq u\leq\delta^2} h(u) &= \begin{cases} h(0) = r, & \text{if } 0 \leq r < 1, \\ h(\delta^2) = \delta + r(1-\delta), & \text{if } r \geq 1. \end{cases} \\ &=: H(r,\epsilon), \end{aligned}$$

as claimed.

To conclude the proof, one manually checks that $k(T(r,\epsilon)) = H(r,\delta) = H(r,\epsilon)$. □

H. An Alternate View of Theorem of Results in terms of Pareto Fronts

Let us now link results to the idea of Pareto Fronts employed in (Javanmard et al., 2020). For any $r \geq 0$, consider the *Pareto Front* $\overline{\mathcal{C}}_r \subseteq \mathbb{R}_+^2$ of the standard risk $E(w)$ and the adversarial risk proxy $\overline{E}(w,r)$, i.e.

$$\overline{\mathcal{C}}_r = \Big\{\Big(E(w(r,t)), \overline{E}(w(r,t),r)\Big) \mid t \geq 0\Big\}, \tag{54}$$

where $w(r,t)$ is the unique minimizer of $L_t(w,r) := tE(w) + \overline{E}(w,r)$ over $w \in \mathbb{R}^d$, and $\overline{E}(w,r) := \|w - w_0\|_\Sigma^2 + r^2\|w\|_\star^2$ as defined in (6).

Proposition H.1. *For an attack strength* $r \geq 0$ *w.r.t to general norm* $\|\cdot\|$, *it holds that*

$$\overline{\mathcal{C}}_r = \Big\{\Big(\sigma^2 + G(\lambda), \sigma^2 + F(r,\lambda)\Big) \mid 0 \leq \lambda \leq r^2\Big\}, \tag{55}$$

$$\overline{\mathcal{C}}_r = \Big\{\Big(\sigma^2 + \epsilon^2\|w_0\|_\Sigma^2, F(r,\lambda_{opt}(r,\epsilon))\Big) \mid 0 \leq \epsilon \leq 1\Big\} \tag{56}$$

$$= \Big\{\Big(\sigma^2 + \epsilon^2\|w_0\|_\Sigma^2, F(r,\lambda_{opt}(r,\epsilon))\Big) \mid 0 \leq \epsilon \leq \epsilon_{FL}(r)\Big\} \cup \overline{L}_r, \tag{57}$$

where $\lambda_{opt}(r,\epsilon) \in [0, r^2]$ *is as defined in* (14) *and* $\overline{L}_r \subseteq \mathbb{R}_+^2$ *the horizontal line segment defined by*

$$\overline{L}_r := \Big\{\Big(\sigma^2 + \epsilon^2\|w_0\|_\Sigma^2, F(r,r^2)\Big) \mid \epsilon_{FL}(r) \leq \epsilon \leq r\Big\}.$$

Proof. Consider the bijective map $t \mapsto \lambda(r,t) := r^2/(1+t)$ from $[0,\infty]$ to $[0,r^2]$ and observe that $w(r,t) = w^{prox}(\lambda(r,t))$. The first part of the result then follows from the definition of $\overline{\mathcal{C}}_r$.

For the second part, observe from the definition of $\lambda_{opt}(r,\epsilon)$ in Lemma 4.2 that $G(\lambda_{opt}(r,\epsilon)) = \epsilon^2\|w_0\|_\Sigma^2$. The result then follows from the first part. □

For example, consider the setting of Euclidean-norm attacks on isotropic features. We know that $w^{prox}(\lambda) = (\Sigma + \lambda I_d)^{-1}\Sigma w_0 = w_0/(1+\lambda)$.

$$\begin{aligned} G(\lambda) &:= \|w^{prox}(\lambda) - w_0\|_\Sigma^2 = \left(\frac{\lambda}{1+\lambda}\right)^2, \\ F(r,\lambda) &= G(\lambda) + r^2\|w^{prox}(\lambda)\|_2^2 = \frac{\lambda^2 + r^2}{(1+\lambda)^2}. \end{aligned} \tag{58}$$

The Free Lunch threshold is then $\epsilon_{FL}(r) := \sqrt{(G(r^2)/\|w_0\|_\Sigma^2} = r^2/(1+r^2)$. Now, given $\epsilon \in [0,1]$, if $\epsilon \leq \epsilon_{FL}(r)$, then solving $G(\lambda) = \epsilon^2\|w_0\|_\Sigma^2 = \epsilon^2$ for $\lambda \in [0,r^2]$ gives $\lambda = \epsilon/(1-\epsilon)$. We deduce that

$$\lambda_{opt}(r,\epsilon) = \begin{cases} \epsilon/(1-\epsilon), & \text{if } 0 \leq \epsilon \leq \epsilon_{FL}(r), \\ r^2, & \text{if } \epsilon_{FL}(r) \leq \epsilon \leq 1. \end{cases} \tag{59}$$

We deduce that

$$F(r,\lambda_{opt}(r,\epsilon)) = \begin{cases} F(r,\epsilon/(1-\epsilon)) = \frac{\epsilon^2/(1-\epsilon)^2 + r^2}{(1+\epsilon/(1-\epsilon))^2} = \epsilon^2 + r^2(1-\epsilon)^2, & \text{if } \epsilon \leq \epsilon_{FL}(r), \\ F(r,r^2) = \frac{r^4+r^2}{(1+r^2)^2} = \frac{r^2}{1+r^2} \asymp \min(r^2,1), & \text{else} \end{cases}. \tag{60}$$

Thus, the Pareto front is

$$\overline{\mathcal{C}}_r = \Big\{(\sigma^2 + \epsilon^2, \epsilon^2 + r^2(1-\epsilon)^2) \mid 0 \leq \epsilon \leq \epsilon_{FL}(r)\Big\} \cup \overline{L}_r, \tag{61}$$

with $\overline{L}_r = \Big\{\Big(\sigma^2 + \epsilon^2, \min(r^2,1)\Big) \mid \epsilon_{FL}(r) \leq \epsilon \leq 1\Big\}$.

I. Spectral Analysis of Euclidean-Norm Attacks

In the case of Euclidean-norm attacks, it turns out that the functions F and G are completely given in terms of spectral information as we now show. Let $\Sigma = \sum_{k\geq 1} \lambda_k \phi_k \phi_k^\top$ be the eigenvalue-decomposition of the feature covariance matrix Σ and $c_k = \phi_k^\top w_0$ be the kth alignment coefficient of the generative model w_0, so that $w_0 = \sum_{k\geq 1} c_k \phi_k$. We shall occasionally consider the infinite-dimensional case where $d = \infty$, and we will require

$$\operatorname{tr}(\Sigma) = \sum_k \lambda_k < \infty, \tag{62}$$

which ensures that the covariance operator Σ is of *trace class*. Furthermore, it is easy to see that

$$\|w_0\|_2^2 = \sum_k c_k^2,\ \|w_0\|_\Sigma^2 = \sum_k \lambda_k c_k^2. \tag{63}$$

In this setting, the following result shows that the functions F and G (10) are given explicitly in terms of the spectral information $(\lambda_k, c_k^2)_{k\geq 1}$.

Lemma I.1. *For Euclidean-norm attacks, it holds for any $r, \lambda \geq 0$ that*

$$G(\lambda) = \lambda^2 \sum_k \frac{\lambda_k c_k^2}{(\lambda_k+\lambda)^2},\ F(r,\lambda) = G(\lambda) + r^2 \sum_k \frac{\lambda_k^2 c_k^2}{(\lambda_k+\lambda)^2}. \tag{64}$$

In particular, for $\lambda = r^2$, it holds that

$$F(r, r^2) = r^2 \sum_k \frac{\lambda_k c_k^2}{\lambda_k + r^2}. \tag{65}$$

The proof (which will be provided shortly) relies on observing that $w^{prox}(\lambda) = (\Sigma + \lambda I)^{-1}\Sigma w_0$.

I.1. Robustness and Statistical Dimension

Recall that, for any $\lambda \geq 0$, the statistical dimension of Σ is defined by $d_\Sigma(\lambda) := \operatorname{tr}(\Sigma(\Sigma+\lambda I)^{-1}) = \sum_k \lambda_k/(\lambda_k+\lambda)$. The following result highlights the role of the statistical dimension on adversarial robustness in the case of uniform source condition.

Corollary I.1. *If $c_k^2 \asymp c^2$ for some constant $c > 0$, then $E_{opt}(r) \asymp \sigma^2 + c^2 r^2 d_\Sigma(r^2)$ for all $r \geq 0$.*

Proof. Indeed, applying the second part of Lemma I.1 gives

$$F(r, r^2) = r^2 \sum_k \lambda_k c_k^2/(\lambda_k + r^2) = c^2 r^2 d_\Sigma(r^2).$$

The result then follows directly from Theorem 3.1. □

I.2. Proof of Lemma I.1

Indeed, in this setting, for any $\lambda \geq 0$, the solution of (9) is explicitly given by $w^{prox}(\lambda) = (\Sigma + \lambda I)^{-1}\Sigma w_0$. Plugging this into the definition of F and G given in (10) respectively then gives

$$\begin{aligned} G(\lambda) &= \|w^{prox}(\lambda) - w_0\|_\Sigma^2 = \|(\Sigma + \lambda I_d)\Sigma w_0 - w_0\|_\Sigma^2 \\ &= \lambda^2 \|(\Sigma+\lambda I)^{-1} w_0\|_\Sigma^2 = \lambda^2 \sum_k \frac{\lambda_k c_k^2}{(\lambda_k+\lambda)^2} \end{aligned}$$

and

$$\begin{aligned} F(r,\lambda) &= G(\lambda) + r^2\|w^{prox}(\lambda)\|_2^2 = G(\lambda) + r^2 \|(\Sigma+\lambda I)^{-1}\Sigma w_0\|_2^2 \\ &= G(\lambda) + r^2 \sum_k \frac{\lambda_k^2 c_k^2}{(\lambda_k+\lambda)^2}, \end{aligned}$$

which proves the first part of the claim.

For the second part, applying the first part with $\lambda = r^2$ gives

$$F(r,r^2) = \sum_k \frac{(r^4\lambda_k + r^2\lambda_k^2)c_k^2}{(\lambda_k + r^2)^2} = r^2 \sum_k \frac{\lambda_k c_k^2}{\lambda_k + r^2}, \tag{66}$$

where the second step is a basic algebraic manipulation.

I.3. Proof of Theorem C.1

Theorem C.1. *Consider the isotropic setting where $\Sigma = I_d$. For Euclidean-norm attack of strength $r \geq 0$, it holds for any tolerance level $\epsilon \in [0,1]$ that*

(i) ***(Free Lunch Threshold)*** $\epsilon_{FL}(r) = r^2/(1+r^2) \in [0,1)$.

(ii) ***(Free Lunch)*** *If $\epsilon \geq \epsilon_{FL}(r)$, then the optimal regularization is $\lambda_{opt}(r,\epsilon) = r^2$, and we have*

$$E_{opt}(r,\epsilon) \asymp E_{opt}(r) \asymp \sigma^2 + \|w_0\|_2^2 \min(r^2, 1). \tag{26}$$

(ii) ***(Accuracy / Robustness Tradeoff)*** *If $\epsilon < \epsilon_{FL}(r)$, then the optimal regularization parameter is given by $\lambda_{opt}(r,\epsilon) = \epsilon/(1-\epsilon)$, and we have*

$$E_{opt}(r,\epsilon) \asymp \sigma^2 + \|w_0\|_2^2(\epsilon^2 + (1-\epsilon)^2 \min(r^2,1)). \tag{27}$$

Proof. Indeed, for any $\lambda \geq 0$, one easily computes

$$G(\lambda) = \lambda^2 \sum_{k=1}^d \frac{c_k^2}{(1+\lambda)^2} = \frac{\lambda^2 \|w_0\|_\Sigma^2}{(1+\lambda)^2}, \tag{67}$$

$$F(r,\lambda) = G(\lambda) + r^2 \sum_{k=1}^d \frac{c_k^2}{(1+\lambda)^2} = \frac{(r^2+\lambda^2)\|w_0\|_\Sigma^2}{(1+\lambda)^2}. \tag{68}$$

Now, one easily computes the free lunch threshold as $\epsilon_{FL}(r) = \sqrt{G(r^2)}/\|w_0\|_\Sigma = r^2/(1+r^2) \in [0,1]$. Observe that $\epsilon_{FL}(r) \asymp \min(r^2,1) \in [0,1]$. We then deduce from Theorem 4.1 that the absolute optimal adversarial risk is given by

$$\begin{aligned} E_{opt}(r) &\asymp \sigma^2 + F(r,r^2) = \sigma^2 + \frac{r^2+r^4}{(1+r^2)^2}\|w_0\|_2^2 = \sigma^2 + \frac{r^2}{1+r^2}\|w_0\|_2^2 \\ &\asymp \sigma^2 + \min(r^2,1)\|w_0\|_\Sigma^2. \end{aligned}$$

Moreover, if $\epsilon \geq \epsilon_{FL}(r)$, then $E_{opt}(r,\epsilon) \asymp E_{opt}(r)$ and there is free lunch: no tradeoff is required between accuracy and robustness. On the other hand, if $\epsilon \in [0, \epsilon_{FL}(r))$, then solving the equation $G(\lambda) = \epsilon^2\|w_0\|_\Sigma^2$, we deduce that the optimal regularization parameter is given by

$$\lambda_{opt}(\epsilon) = \frac{\epsilon}{1-\epsilon}. \tag{69}$$

Consequently, Theorem 4.1 tells us that that

$$\begin{aligned} E_{opt}(r,\epsilon) \asymp \sigma^2 + F(r, \lambda_{opt}(\epsilon)) &= \sigma^2 + \frac{(r^2 + \epsilon^2/(1-\epsilon)^2))\|w_0\|_\Sigma^2}{(1+\epsilon/(1-\epsilon))^2} \\ &= \sigma^2 + \|w_0\|_\Sigma^2(\epsilon^2 + (1-\epsilon)^2 r^2) \\ &= \sigma^2 + \|w_0\|_\Sigma^2(\epsilon^2 + (1-\epsilon)^2 r^2), \end{aligned}$$

from which the result follows. □

Regime	$\epsilon_{FL}(r)$	$\lambda_{opt}(r,\epsilon)$	$E_{opt}(r)$	$E_{opt}(r,\epsilon)$	Free lunch ?
$0 \le \delta < 1$	$r^{2(1-\theta)}$	$\epsilon^{2/(1-\theta)}$	$\sigma^2 + r^{2(1-\theta)}$	$\sigma^2 + \epsilon^2 + r^2\epsilon^{-2\phi}$	No
$\delta = 1$	$r^4 \log(1/r)$	$e^{W(-\Theta(\epsilon^2))/2}$	$\sigma^2 + r^2 \log(1/r)$	$\sigma^2 + \epsilon^2 + r^2 \log(1/\epsilon)$	No
$\delta > 1$	r^4	ϵ	$\sigma^2 + r^2$	$\sigma^2 + \epsilon^2 + r^2$	Yes!

Table 3. Details for Theorem 5.2. Here, W is an appropriate branch of the Lambert function. Note that except for the first column, all the entries in the table are given only within multiplicative absolute constants. The last column records whether there is free lunch (FL), wherein robustness is achievable without sacrificing accuracy.

I.4. Proof of Theorem 5.2

Theorem 5.2. *For Euclidean-norm attacks of small strength $r > 0$, the conclusions of Theorem 4.1 prevail, and the quantities $\epsilon_{FL}(r)$, $\lambda_{opt}(r,\epsilon)$, $E_{opt}(r)$, and $E_{opt}(r,\epsilon)$ are as given in Table 2.*

Consider the particular regime where $0 \le \delta \le 1$. For small $\sigma^2 \ge 0$, $\epsilon > 0$, and $r = r(\epsilon)$ given by

$$r = \begin{cases} \epsilon^{\phi}, & \text{if } 0 \le \delta < 1, \\ \sqrt{1/\log(1/\epsilon)}, & \text{if } \delta = 1 \end{cases} \tag{19}$$

with $\theta := (1-\delta)/\beta \ge 0$, $\phi := \theta/(1-\theta) \ge 0$, it holds that

$$E_{opt}(r) = o(1),\, E_{opt}(r,\epsilon) = \Theta(1). \tag{20}$$

We will need the following crucial lemma.

Lemma I.2. *Let the sequence $(\lambda_k)_{k\ge1}$ of positive numbers be such that $\lambda_k \asymp k^{-\beta}$ for some constant $\beta > 0$, and let $m, n \ge 0$ with $n\beta > 1$. Then, for $D \gg 1$, it holds that*

$$\sum_{k=1}^{\infty} \frac{\lambda_k^n}{(1+D\lambda_k)^m} \asymp D^{-c} \begin{cases} \log D, & \text{if } m = n - 1/\beta, \\ 1, & \text{else,} \end{cases} \tag{70}$$

where $c := \min(m, n - 1/\beta) \ge 0$.

Proof. First observe that

$$\begin{aligned} \lambda_k^n/(1+D\lambda_k)^m &\asymp \lambda_k^n \min(1, (D\lambda_k)^{-m}) \\ &= \begin{cases} \lambda_k^n = k^{-n\beta}, & \text{if } D\lambda_k < 1, \text{ i.e if } k > D^{1/\beta}, \\ D^{-m}\lambda_k^{-(m-n)} = D^{-m}k^{(m-n)\beta}, & \text{else.} \end{cases} \end{aligned}$$

We deduce that

$$\sum_{k=1}^{\infty} \frac{\lambda_k^n}{(1+D\lambda_k)^m} \asymp D^{-m} \sum_{1 \le k \le D^{1/\beta}} k^{(m-n)\beta} + \sum_{k > D^{1/\beta}} k^{-n\beta}. \tag{71}$$

By comparing with the corresponding integral, one can write the first sum in (71) as

$$\begin{aligned}
D^{-m}\sum_{1\le k\le D^{1/\beta}} k^{(m-n)\beta} &\asymp D^{-m}\int_1^{D^{1/\beta}} u^{(m-n)\beta}\mathrm{d}u\\
&\asymp D^{-m}\begin{cases}(D^{1/\beta})^{1+(m-n)\beta}=D^{-(n-1/\beta)}, & \text{if } n-1/\beta<m,\\ \log D, & \text{if } m=n-1/\beta,\\ 1, & \text{else.}\end{cases}\\
&=\begin{cases}D^{-(n-1/\beta)}, & \text{if } n-1/\beta<m,\\ D^{-m}\log D, & \text{if } m=n-1/\beta,\\ D^{-m}, & \text{else.}\end{cases}\\
&=D^{-c}\begin{cases}\log D, & \text{if } m=n-1/\beta,\\ 1, & \text{else,}\end{cases}
\end{aligned}$$

where $c\ge 0$ is as given in the lemma.

Analogously, one can write the second sum in (71) as

$$\sum_{k>D^{1/\beta}} k^{-n\beta}\asymp\int_{D^{1/\beta}}^{\infty} u^{-n\beta}\mathrm{d}u\asymp (D^{1/\beta})^{1-n\beta}=D^{-(n-1/\beta)},$$

and the result follows upon putting things together. □

Corollary I.2. *Let β and δ be as Theorem 5.2. For any $r\ge 0$ and small $\lambda>0$, the functions G and F defined in (10) satisfy*

$$G(\lambda)\asymp\begin{cases}\lambda^{1-\theta}, & \textit{if } 0\le\delta<\beta+1,\\ \lambda^2\log(1/\lambda), & \textit{if } \delta=\beta+1,\\ \lambda^2, & \textit{if } \delta>\beta+1.\end{cases}\tag{72}$$

$$F(r,\lambda)\asymp\begin{cases}\lambda^{1-\theta}+r^2\lambda^{-\theta}, & \textit{if } 0\le\delta<1,\\ \lambda^{1-\theta}+r^2\log(1/\lambda), & \textit{if } \delta=1,\\ \lambda^{1-\theta}+r^2, & \textit{if } 1<\delta<\beta+1,\,.\\ \lambda^2\log(1/\lambda)+r^2, & \textit{if } \delta=\beta+1,\\ \lambda^2+r^2, & \textit{if } \delta>\beta+1.\end{cases}\tag{73}$$

Moreover, for small $r>0$, it holds that

$$F(r,r^2)\asymp\begin{cases}r^{2(1-\theta)}, & \textit{if } 0\le\delta<1,\\ r^2\log(1/r), & \textit{if } \delta=1,\\ r^2, & \textit{if } \delta>1.\end{cases}\tag{74}$$

Proof. Set $D:=1/\lambda$. One can write

$$G(\lambda)=\lambda^2\sum_{k\ge1}\frac{\lambda_k c_k^2}{(\lambda+\lambda_k)^2}=\sum_{k\ge1}\frac{\lambda_k c_k^2}{(1+D\lambda_k)^2}\asymp\sum_{k\ge1}\frac{\lambda_k^{1+\delta/\beta}}{(1+D\lambda_k)^2},\tag{75}$$

where we have used the fact that $\lambda_k c_k^2\asymp k^{-\beta-\delta}=k^{-(1+\delta/\beta)\beta}\asymp\lambda_k^{1+\delta/\beta}$. Applying Lemma I.2 with $n=1+\delta/\beta$ and $m=2$, we deduce that

$$G(\lambda)\asymp D^{-c}\begin{cases}\log D, & \text{if } m=n-1/\beta,\text{ i.e if } \delta=\beta+1,\\ 1, & \text{else,}\end{cases}\tag{76}$$

where $c = \min(m, n - 1/\beta) = \min(2, 1 + \delta/\beta - 1/\beta) = \min(2, 1 - \theta)$. This proves (72).

Analogously, one can rewrite

$$(F(r,\lambda) - G(\lambda))/r^2 = \sum_{k \ge 1} \frac{\lambda_k^2 c_k^2}{(\lambda + \lambda_k)^2} = D^2 \sum_{k \ge 1} \frac{\lambda_k^2 c_k^2}{(1 + D\lambda_k)^2} \asymp D^2 \sum_{k \ge 1} \frac{\lambda_k^{2+\delta/\beta}}{(1 + D\lambda_k)^2}.$$

Applying Lemma I.2 to the RHS with $n = 2 + \delta/\beta$ and $m = 2$ then gives

$$(F(r,\lambda) - G(\lambda))/r^2 \asymp D^2 \sum_{k \ge 1} \frac{\lambda_k^{2+\delta/\beta}}{(1 + D\lambda_k)^2} \asymp D^C \begin{cases} \log D, & \text{if } m = n - 1/\beta, \text{ i.e if } \delta = 1, \\ 1, & \text{else,} \end{cases}$$

where $C = 2 - \min(m, n - 1/\beta) = 2 - \min(2, 2 - \theta) = -\min(0, -\theta) = \max(\theta, 0)$. Combining with (72) proves (73).

Finally, (74) follows by pluggin $\lambda = r^2$ in (73) and simplifying. □

We are now ready to prove Theorem 5.2

Proof of Theorem 5.2. Equipped with Corollary I.2, first observe that if $\delta > 1$, then $E_{opt}(r) \asymp \sigma^2 + F(r, r^2) \asymp \sigma^2 + r^2$ which matches $E(w_0, r) \asymp \sigma^2 + r^2 \|w_0\|_2^2 \asymp \sigma^2 + r^2$, and so no tradeoff is needed: the ground-truth model w_0 achieves the optimal level of robustness $E_{opt}(r)$.

Now, if $\delta \in [0, \beta + 1)$, we deduce that for $r = o(1)$, the free lunch threshold is given by

$$\epsilon_{FL}(r) := \frac{\sqrt{G(r^2)}}{\|w_0\|_\Sigma} \asymp r^{(1-\theta)} = o(1). \tag{77}$$

For $\epsilon \in [0, \epsilon_{FL}(r))$, solving the equation $G(\lambda) = \epsilon^2 \|w_0\|_\Sigma^2$ for $\lambda \in [0, r^2]$ gives

$$\lambda_{opt}(r, \epsilon) \asymp \epsilon^{2/(1-\theta)}. \tag{78}$$

On the other hand, if $\delta = \beta + 1$, then $G(\lambda) \asymp \lambda^2 \log(1/\lambda)$ for $\lambda = o(1)$, and so

$$\epsilon_{FL}(r) := \frac{\sqrt{G(r^2)}}{\|w_0\|_\Sigma} \asymp r^2 \log(1/r). \tag{79}$$

For $\epsilon \in [0, \epsilon_{FL}(r))$, solving (14) for $\lambda \in [0, r^2]$ then gives

$$\lambda_{opt}(r, \epsilon) \asymp e^{W(-\Theta(\epsilon^2))/2}, \tag{80}$$

where W is an appropriate branch of the Lambert function.

Finally, if $\delta > \beta + 1$, then $G(\lambda) \asymp \lambda^2$ for $\lambda = o(1)$, and so

$$\epsilon_{FL}(r) := \frac{\sqrt{G(r^2)}}{\|w_0\|_\Sigma} \asymp r^2. \tag{81}$$

For $\epsilon \in [0, \epsilon_{FL}(r))$, solving (14) for $\lambda \in [0, r^2]$ then gives

$$\lambda_{opt}(r, \epsilon) \asymp r. \tag{82}$$

Combining with Theorem 4.1 and putting things together gives the estimates stated in Table 2. □

J. Other Proofs

J.1. Proof of Theorem 5.1

Theorem 5.1. *Recall the notations of Theorem 4.1. Let the attack norm be an ℓ_p with $p \in [1, \infty]$. For any $r \geq 0$ and $\epsilon \in [0, 1]$, the robustness profile is given as in Table 1.*

In particular, in the limit $d \to \infty$, we have that, if

- *$p \in [1, \infty)$, $1 \ll s \leq d$, and we take $r \asymp 1/s^{1/q}$, OR*
- *$p = \infty$, $\sqrt{d/\log d} \ll s \leq d$, and we take $r \asymp 1/s$,*

then for $\epsilon \in [0, 1)$, it holds that

$$E_{opt}(r) = o_d(1), \text{ and } E_{opt}(r, \epsilon) = \Theta((1 - \epsilon)^2). \tag{16}$$

Proof. First observe that $\|w_0\|_\Sigma^2 = \|w_0\|_2^2 = s$ and $\|\Sigma w_0\|_p = \|w_0\|_p = s^{1/p}$. We deduce that $r_0 := \|w_0\|_\Sigma / \|w_0\|_\star = s^{1/2}/\|w_0\|_q = s^{1/2-1/q}$. Likewise, $r_1 := \|\Sigma w_0\|/\|w_0\|_\Sigma = s^{1/p-1/2} = r_0$, since $1/p + 1/q = 1$ by definition. We conclude that the problem is well-conditioned in the sense of Section xyz, and the claimed formulae for $E_{opt}(r)$ and $E_{opt}(r, \epsilon)$ follow from Theorem E.1.

Furthermore, in the special case of Euclidean-norm attacks (i.e $p = 2$), the claimed formula for the free lunch threshold $\epsilon_{FL}(r)$ and the optimal regularization parameter $\lambda_{opt}(r, \epsilon)$ follow from Theorem C.1.

To conclude the proof, we know establish (16). Indeed, if $\sigma^2 = o(1)$, $p \in [1, \infty)$, and $1 \ll s \leq d$ and we take $r \asymp 1/s^{1/q}$ in the limit $d \to \infty$, then we see from Table 1 that $E_{opt}(r) \asymp (s/d) \min(r\sqrt{d}, 1)^2 \asymp \sigma^2 + (s/d) \min(s^{1/q}\sqrt{d}, 1)^2 = \sigma^2 + (s/d) = o(1)$ since $\sigma^2 = o(1)$ and $s/d = o(1)$.

Also from the same table, one reads $E_{opt}(r, \varepsilon) \asymp (s/d)H(r/r_0(p), \epsilon)^2$ with $r_0(p) = s^{1/p-1/2}/\sqrt{d}$. Now, for any fixed $\epsilon \in [0, 1)$, one computes

$$H(r/r_0(p), \epsilon) \asymp H(s^{1/2-1/p-1/q}\sqrt{d}, \epsilon) = H(\sqrt{d/s}, \epsilon) = (\epsilon + (1 - \epsilon)\sqrt{d/s}) \asymp (1 - \epsilon)\sqrt{d/s},$$

and so $E_{opt}(r, \epsilon) \asymp (s/d)H(r/r_0(p), \epsilon)^2 \asymp (1 - \epsilon)^2$ as claimed. □

J.2. Proof of Theorem 5.3

Theorem 5.3. *Consider the setting (21). For ℓ_∞-norm attacks of strength r with $1/\sqrt{d} \leq r = o(1)$, it holds that*

$$E_{opt}(r) \asymp \sigma^2 + r^2 \log(1/r)^2. \tag{22}$$

In particular, for $r \asymp 1/\log d$ and $\sigma^2 = o(1)$, it holds that

$$E(w_0, r) = \Theta(1),\ E_{opt}(r) = o(1). \tag{23}$$

That is, even though robustness is achievable, the generative model w_0 is itself non-robust.

Proof. For any $a \in \mathbb{R}^d$ and $t \geq 0$, define $\kappa_a(t) := \inf_{u \in \mathbb{R}^d} t\|u - a\|_2 + \|u\|_1$. Let $H_d := \sum_{k=1}^d 1/k \asymp \log d$ be dth harmonic number. In the particular case where $a = (1/(kH_d))_{k \in [d]} \in \mathbb{R}^d$, it was shown in (Blais et al., 2019) that: $\kappa_a(t) = \dfrac{2 \log t + O(1)}{\log d}$ for $1 \ll t \leq \sqrt{d}$. Noting that $w_0 = H_d a$ and taking $t = 1/r$, we deduce that

$$\begin{aligned} K(w, r) &:= \inf_{w \in \mathbb{R}^d} \|w - w_0\|_2 + r\|w\|_1 = \inf_{w \in \mathbb{R}^d} H_d \|w/H_d - a\|_2 + rH_d\|w/H_d\|_1 \\ &= rH_d \cdot \inf_{u \in \mathbb{R}^d} r^{-1}\|u - a\|_2 + \|u\|_1 \text{ with change of variable } u = w/H_d \\ &= rH_d \cdot \kappa_a(1/r). \end{aligned} \tag{83}$$

Thus, for $1/\sqrt{d} \le r = o(1)$, taking $t = 1/r$ gives

$$K(w, r) = \frac{r(\log(1/r) + O(1))H_d}{\log d} \asymp r(\log(1/r) + o(1)),$$

from which the first part of the claim follows.

For the second part, taking $r = 1/\log d$ gives

$$E_{opt}(r) \asymp \sigma^2 + r^2 \log(1/r)^2 = \sigma^2 + (\log^2 d/\log d)^2 = \sigma^2 + o(1) = o(1).$$

On the other hand, it is clear that for any $r \ge 0$,

$$E(w_0, r) = \sigma^2 + r^2\|w_0\|_1^2 = \sigma^2 + r^2(\sum_{k=1}^{d} 1/k)^2 \asymp \sigma^2 + (r \log d)^2,$$

Thus, for $r = 1/\log d$ and $\sigma^2 = o(1)$, then $E(w_0, r) \asymp \sigma^2 + 1 = \Theta(1)$. □