On the existence of consistent adversarial attacks in high-dimensional linear classification

Matteo Vilucchio[1], Lenka Zdeborová[2], and Bruno Loureiro[3]

[1]Information Learning and Physics Laboratory, École Polytechnique Fédérale de Lausanne (EPFL)
[2]Statistical Physics of Computation Laboratory, École Polytechnique Fédérale de Lausanne (EPFL)
[3]Département d'Informatique, École Normale Supérieure - PSL & CNRS, France

June 17, 2025

Abstract

What fundamentally distinguishes an adversarial attack from a misclassification due to limited model expressivity or finite data? In this work, we investigate this question in the setting of high-dimensional binary classification, where statistical effects due to limited data availability play a central role. We introduce a new error metric that precisely capture this distinction, quantifying model vulnerability to consistent adversarial attacks — perturbations that preserve the ground-truth labels. Our main technical contribution is an exact and rigorous asymptotic characterization of these metrics in both well-specified models and latent space models, revealing different vulnerability patterns compared to standard robust error measures. The theoretical results demonstrate that as models become more overparameterized, their vulnerability to label-preserving perturbations grows, offering theoretical insight into the mechanisms underlying model sensitivity to adversarial attacks.

1 Introduction

Machine learning models, despite their remarkable performance across various domains, remain vulnerable to adversarial examples — inputs specifically crafted to mislead models while appearing innocuous to humans. While adversarial robustness has attracted significant research attention, a critical distinction often overlooked is between *consistent* (or *proper*) and *inconsistent* (or *improper*) adversarial examples. Consistent adversarial examples maintain the ground-truth label despite perturbations, whereas inconsistent ones change the true classification.

To illustrate this distinction, consider the classic example from [1]: an image of a panda that, after subtle perturbations, is misclassified by a neural network. This represents a consistent adversarial example because the image still depicts a panda to human observers. In contrast, if the perturbation were to transform the image to genuinely resemble a different animal, it would be an inconsistent adversarial example. This distinction is crucial: vulnerability to consistent attacks represents a genuine failure of the model to capture invariant features that humans naturally perceive.

The assumption that adversarial perturbations do not alter the true class (i.e., remain consistent) underlies most practical approaches to adversarial robustness in computer vision [1, 2]. While this assumption has been explored in theoretical works on robust generalization [3, 4], a mathematical understanding of their properties, such as existence and effectiveness in tricking even simple linear classifiers remains elusive.

Following a large body of work in high-dimensional statistics [5–10], we analyze this problem through the lens of exact asymptotics of linear classifiers. We develop novel metrics that precisely quantify vulnerability to both consistent and inconsistent adversarial attacks. We

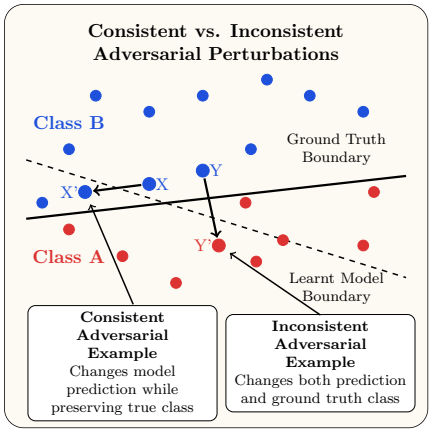

Figure 1: Illustration of the difference between consistent Adversarial Perturbation and inconsistent Adversarial Perturbation.

define and analyze these metrics in two complementary settings: first, a *well-specified* model where all input covariates are directly available; second, a *latent space* model where the available covariates are feature transformations of underlying latent variables. For both settings, we derive closed-form expressions for the consistent robustness metrics in the high-dimensional limit—where the latent space dimension d, the number of features p, and the sample size n all scale to infinity at fixed ratios. In the latent space model, we further derive exact asymptotic descriptions for the performance of robust empirical risk minimization [1, 2], the mostly adopted way of finding a robust model.

Furthermore, the effect of over or under-parameterization — using more or less parameters than strictly necessary to encode the data features — is still unclear in the adversarial settings. Some recent works [11] consider the regression case for squared loss but in the context of classification the question is still open. While overparameterization generally improves standard generalization [12, 13], its effects on adversarial robustness are less understood, particularly when considering consistent versus inconsistent attacks. Conventional wisdom suggests that overparameterized models might be more vulnerable to adversarial examples due to their flexibility in fitting noise and the more parameters that can be changed to flip the model prediction. However, our analysis reveals a more nuanced picture: more overparameterization can improve an attacker's ability to craft effective adversarial examples, but this relationship depends critically on considering attacks on already correctly classified data points. If instead we consider consistent attacks on all possible inputs (also the misclassified by the model) we notice that increasing overparameterization leads to an improvement because of the beneficial role that overparameterization has on the clean generalization.

Our **main contributions** can be summarized as

1. We establish necessary and sufficient conditions for the existence of consistent perturbations in two classes of binary classifiers: well-specified linear classifiers, and a latent variable model that accounts for misspecification and overparametrization in linear estimation, independently of the data distribution. Under a Gaussian design, this leads to an exact formula for the probability that consistent attacks exist in these models.

2. We introduce novel consistent robust error metrics quantifying the impact of consistent attacks. For the classifiers of interest, we derive an asymptotic formula that exactly characterize their high-dimensional limits under a Gaussian design assumption.

3. We study how robust empirical risk minimization can mitigate the impact of consistent attacks in this high-dimensional limit, for both the well-specified and latent variable model. For the latter, this requires an exact asymptotic characterization of the robust ERM estimator under misspecification which is novel and of independent interest.

Our work reveals that overparameterization plays a nuanced but crucial role in building resistance against consistent adversarial attacks. Contrary to conventional wisdom, our theoretical analysis demonstrates that higher degrees of parameterization can be beneficial for overall robustness, though this benefit must be balanced against increased vulnerability on specific subsets of inputs. These insights can provide guidance for system design, highlighting the importance of considering the consistent/inconsistent attack distinction when evaluating and optimizing model robustness.

1.1 Further Related Works

Exact Asymptotics: Our analysis builds upon the previous literature characterizing the properties of predictors in the high-dimensional proportional regime. This approach spans multiple theoretical frameworks: high-dimensional probability theory [14–16], statistical physics approaches [17–23], and random matrix theory [24–31]. Our work is particularly motivated by recent advances in Gaussian universality [32–34], which demonstrate that simple Gaussian models often provide surprisingly accurate predictions for more complex data distributions in high dimensions. This phenomenon emerges from concentration properties in high-dimensional spaces, leading to universality in generalization behavior across different covariate distributions [35–38].

Adversarial Robustness: Robust empirical risk minimization, commonly known as adversarial training, was pioneered in computer vision [1, 2] and has since evolved into a primary defense against adversarial attacks. Researchers have developed numerous approaches to improve its computational efficiency [39, 40] and statistical properties [41–43]. On the theoretical front, several works have investigated the properties of robust empirical risk minimization for linear models [4, 44–48], including sharp proportional asymptotics under different data designs [3, 11, 47, 49–54].

Of particular relevance to our work is [47], which derives high-dimensional asymptotics for binary classification in the well-specified model — a result which we build upon in our analysis in Section 3.

Consistent attack: The idea that adversarial attacks should be imperceptible to some metric of interest (e.g. the human eyes in vision) underlies most of the empirical literature [1, 55, 56]. The notion of a consistent attack in the theoretical literature was formalized in [3, 4].

Notation

We denote vectors by bold letters $\boldsymbol{x} \in \mathbb{R}^d$. $\mathbb{S}^{d-1}(r) = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}||_2 = r\}$ denote the Euclidean sphere of radius r, and $\text{span}(\boldsymbol{x}) = \boldsymbol{x}\mathbb{R} = \{\mu\boldsymbol{x}, \mu \in \mathbb{R}\}$. For $q \geq 1$, $||\boldsymbol{x}||_q = \left(\sum_{j=1}^d x_j^q\right)^{1/q}$ denote the ℓ^q-norm, and $B_q(r) = \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}||_p \leq r\}$ the ℓ^q-ball of radius $r > 0$. We denote by $q^\star$ the dual of q in the ℓ^q sense: the number $q^\star$ such that $1/q + 1/q^\star = 1$. We denote by $\mathcal{N}(0,1)$ the standard normal distribution, and $\mathbb{P}[Z \leq t] = \Phi(t)$ its c.d.f.

2 Consistent adversarial perturbations and how to quantify them

As motivated in Section 1, the key distinction between an adversarial attack and a random perturbation of the data is the underlying assumption that adversarial attacks leave the ground truth data distribution unchanged. Our starting point is to formalize this notion in the context of binary classifiers.

Consider a binary classification task $(\mu, f_\star)$ defined by a covariate distribution μ over $\mathbb{R}^d$ and a ground-truth classifier $f_\star : \mathbb{R}^d \to [0,1]$, such that for a given $\boldsymbol{x} \sim \mu$, we can assign a binary label $y \in \{-1, +1\}$ with probability given by $f_\star(\boldsymbol{x}) = \mathbb{P}(y = +1|\boldsymbol{x})$.

Definition 1 (Consistent attack)**.** Let $f_\star, \hat{f} : \mathbb{R}^d \to [0,1]$ denote two binary classifiers, refereed to as the *target* and the *model*, $\boldsymbol{x} \in \mathbb{R}^d$ a covariate and $\hat{y} : [0,1] \to \{\pm 1\}$ a decision rule associated to $\hat{f}$. We say a perturbation $\boldsymbol{\delta} \in B_q(\varepsilon)$ of the model $\hat{f}$ is a *consistent* adversarial attack with respect to the target $f_\star$, the covariate $\boldsymbol{x} \in \mathbb{R}^d$ and the decision rule $\hat{y}$ if the following two conditions hold:

- **Model deception:** $\hat{y}(\hat{f}(\boldsymbol{x})) \neq \hat{y}(\hat{f}(\boldsymbol{x} + \boldsymbol{\delta}))$.
- **Target invariance:** $f_\star(\boldsymbol{x}) = f_\star(\boldsymbol{x} + \boldsymbol{\delta})$.

Otherwise, we say that the attack is *inconsistent*.

See Figure 1 for an illustration of a consistent vs. inconsistent attack in the case of linear classifiers.

Remark 1. Note that the second condition (target invariance) is equivalent to label invariance $y(f_\star(\boldsymbol{x}+\boldsymbol{\delta})) = y(f_\star(\boldsymbol{x}))$ in the case of a deterministic ground-truth rule. In the presence of label noise, this condition rules out label swapping due to noise.

One of our central goals in this work is to investigate the properties of consistent adversarial attacks for particular classes of problems. Before moving to specific tasks, we introduce the central metrics allowing us to quantify these properties.

Definition 2 (Adversarial errors)**.** Let $(\mu, f_\star)$ denote a binary classification task. Given a classifier $\hat{f} : \mathbb{R}^d \to [0,1]$ and its associated predictor $\hat{y}(\boldsymbol{x})$, we define the following three metrics

- **Robust error**: This is the standard notion of robust generalization error in the adversarial literature [45, 57, 58], and simply quantifies how vulnerable $\hat{f}$ is to arbitrary perturbations in a ℓ^q-ball:

$$E_{\text{rob}}(\hat{f}) = \mathbb{E}\left[\max_{\boldsymbol{\delta} \in B_q(\varepsilon)} \mathbb{1}\{y \neq \hat{y}(\boldsymbol{x} + \boldsymbol{\delta})\}\right], \tag{1}$$

- **Consistent robust error:** The standard robust error considers both consistent and inconsistent perturbations. In order to quantify the role of consistent attacks, we define the *consistent robust error* by excluding inconsistent perturbations:

$$E_{\text{rob}}^{\text{cns}}(\hat{f}) = \mathbb{E}\left[\max_{\boldsymbol{\delta} \in B_q(\varepsilon) : f_\star(\boldsymbol{x}) = f_\star(\boldsymbol{x}+\boldsymbol{\delta})} \mathbb{1}\{y \neq \hat{y}(\boldsymbol{x} + \boldsymbol{\delta})\}\right], \tag{2}$$

Note that the critical difference between E_{rob} and $E^{\text{cns}}_{\text{rob}}$ lies in the constraint in the inner maximization that satisfies the target invariance from Definition 1 (c.f. Remark 1).

- **Consistent boundary error:** Finally, note that the consistent robust error does not distinguish between labels that are originally misclassified by the model and labels that become misclassified under the attack perturbation. This motivates the introduction of a more nuanced metric, accounting only for labels that are misclassified due to the attack:

$$E^{\text{cns}}_{\text{bnd}}(\hat{f}) = \mathbb{E}\left[\max_{\boldsymbol{\delta}\in B_q(\varepsilon): f_\star(\boldsymbol{x})=f_\star(\boldsymbol{x}+\boldsymbol{\delta})} \mathbb{1}\{y \neq \hat{y}(\boldsymbol{x}+\boldsymbol{\delta})\}\mathbb{1}\{y = \hat{y}(\boldsymbol{x})\}\right]. \tag{3}$$

Remark 2. Note that for any $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \{-1, +1\}$, the constraint sets:

$$\begin{aligned} C_{\text{rob}} &= \{\boldsymbol{\delta} \in B_q(\varepsilon) : \hat{y}(\boldsymbol{x}) \neq \hat{y}(\boldsymbol{x}+\boldsymbol{\delta})\}, \\ C^{\text{cns}}_{\text{rob}} &= \{\boldsymbol{\delta} \in B_q(\varepsilon) : \hat{y}(\boldsymbol{x}) \neq \hat{y}(\boldsymbol{x}+\boldsymbol{\delta}) \text{ and } f_\star(\boldsymbol{x}) = f_\star(\boldsymbol{x}+\boldsymbol{\delta})\}, \\ C^{\text{cns}}_{\text{bnd}} &= C^{\text{cns}}_{\text{rob}} \cap \{y = \hat{y}(\boldsymbol{x})\}, \end{aligned} \tag{4}$$

are nested $C^{\text{cns}}_{\text{bnd}} \subset C^{\text{cns}}_{\text{rob}} \subset C_{\text{rob}}$. Therefore, we generally have:

$$0 \leq E^{\text{cns}}_{\text{bnd}} \leq E^{\text{cns}}_{\text{rob}} \leq E_{\text{rob}}. \tag{5}$$

3 Consistent attacks in well-specified linear classification

Despite a established literature studying robust training schemes, the fundamental properties of consistent attacks remain poorly understood. Our goal in the following is to fill this gap by studying their behavior in the context of high-dimensional binary linear classifiers.

Definition 3 (Linear classifiers)**.** A linear binary classifier in $\mathbb{R}^d$ is a function

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = \mathbb{P}_{\boldsymbol{w}}(y = +1|\boldsymbol{x}) = \varphi(\langle \boldsymbol{w}, \boldsymbol{x}\rangle), \tag{6}$$

defined by the *weight vector* $\boldsymbol{w} \in \mathbb{R}^d$ and a monotonic *link function* $\varphi : \mathbb{R} \to [0, 1]$.

The class of linear binary classifiers encompass several models of interest in statistics, such as the logit $\varphi(t) = (1 + e^{-t})^{-1}$, the probit $\varphi(t) = 1/2(\text{erf}(t) + 1)$ and the noiseless $\varphi(t) = 1_{t\geq 0}$ model.

3.1 Geometry of consistent attacks

As a first step, we consider the geometry of consistent attacks in the class of linear classifiers. Let $f_{\boldsymbol{w}_\star}$ denote a reference linear classifier with weights $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$ and link function $\varphi_\star$, which we will refer to as the *ground-truth*. Since $\varphi_\star : \mathbb{R} \to [0, 1]$ is monotonic, the target invariance condition $f_{\boldsymbol{w}_\star}(\boldsymbol{x}) = f_{\boldsymbol{w}_\star}(\boldsymbol{w} + \boldsymbol{\delta})$ is equivalent to $\langle \boldsymbol{\delta}, \boldsymbol{x}\rangle = 0$, i.e. the attack must be orthogonal to the covariate. Therefore, the set of admissible consistent adversarial attacks with respect to the target classifier defines a hyperplane:

$$H_q(\varepsilon) \coloneqq \{\boldsymbol{\delta} \in B_q(\varepsilon) : \langle \boldsymbol{w}_\star, \boldsymbol{\delta}\rangle = 0\}. \tag{7}$$

Consider a second linear classifier $f_{\hat{\boldsymbol{w}}}$ with weights $\hat{\boldsymbol{w}} \in \mathbb{R}^d$ and link function φ, which we will refer to as the *model*. A successful attack should flip the predictor labels $\hat{y}(\boldsymbol{x}) \neq \hat{y}(\boldsymbol{x} + \boldsymbol{\delta})$. For the standard decision function $\hat{y}(\boldsymbol{x}) = \text{sign}(2f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}) - 1)$ this condition is equivalent to having $\langle \hat{\boldsymbol{w}}, \boldsymbol{x}\rangle(\langle \hat{\boldsymbol{w}}, \boldsymbol{x}\rangle + \langle \hat{\boldsymbol{w}}, \boldsymbol{\delta}\rangle) \leq 0$. This is the case if and only if:

$$|\langle \hat{\boldsymbol{w}}, \boldsymbol{\delta}\rangle| > |\langle \hat{\boldsymbol{w}}, \boldsymbol{x}\rangle|, \qquad \text{and} \qquad \text{sign}(\langle \hat{\boldsymbol{w}}, \boldsymbol{\delta}\rangle) = -\text{sign}(\langle \hat{\boldsymbol{w}}, \boldsymbol{x}\rangle). \tag{8}$$

In other words, in order to flip the model prediction, an attacker must have an anti-parallel component to the predictor weights and exceed the prediction margin $|\langle \hat{\boldsymbol{w}}, \boldsymbol{x}\rangle|$. Putting together, we can derive the following geometrical characterization for the existence of consistent perturbations.

Proposition 1 (Existence of consistent attack)**.** *Consider two linear classifiers defined by the weights $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$ and $\hat{\boldsymbol{w}} \in \mathbb{R}^d$. Let $\boldsymbol{x} \in \mathbb{R}^d$ denote a covariate, and assume $\langle \hat{\boldsymbol{w}}, \boldsymbol{x} \rangle \neq 0$. Then, a consistent attack $\boldsymbol{\delta} \in B_q(\varepsilon)$ with respect to $\boldsymbol{w}_\star, \boldsymbol{x} \in \mathbb{R}^d$ and the decision function $\hat{y}(\boldsymbol{x}) = \operatorname{sign}(2 f_{\hat{\boldsymbol{w}}}(\boldsymbol{x}) - 1)$ exists if and only if:*

$$\varepsilon \, d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp) \geq |\langle \hat{\boldsymbol{w}}, \boldsymbol{x} \rangle| \tag{9}$$

where $\hat{\boldsymbol{w}}_\perp = \hat{\boldsymbol{w}} - {}^{\langle \boldsymbol{w}_\star, \hat{\boldsymbol{w}} \rangle}/_{d} \boldsymbol{w}_\star$ is the predictor components orthogonal the target weights, $d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp) = \inf_{\mu \in \mathbb{R}} ||\hat{\boldsymbol{w}}_\perp - \mu \boldsymbol{w}_\star||_{q^\star}$ is the $\ell^{q^\star}$ distance to the $\operatorname{span}(\boldsymbol{w}_\star)$ and $q^\star$ is the dual of q.

Proof. As discussed above, a consistent attack must satisfy the three conditions in eqs. (7) and (8). First, note that this is only possible if $\hat{\boldsymbol{w}}_\perp \neq 0$, otherwise any admissible perturbation would *a fortiori* violate eq. (8). Therefore, from now on we assume $\hat{\boldsymbol{w}}_\perp \neq 0$. Consider an admissible attack $\boldsymbol{\delta} \in H_q(\varepsilon)$. Since $-\boldsymbol{\delta} \in H_q$, we can always fix the sign of $\boldsymbol{\delta}$ to satisfy the constraint $\operatorname{sign}(\langle \hat{\boldsymbol{w}}, \boldsymbol{\delta} \rangle) = -\operatorname{sign}(\langle \hat{\boldsymbol{w}}, \boldsymbol{x} \rangle)$. The restrictive condition is the margin $|\langle \hat{\boldsymbol{w}}, \boldsymbol{\delta} \rangle| > |\langle \hat{\boldsymbol{w}}, \boldsymbol{x} \rangle|$. Since $\langle \boldsymbol{\delta}, \boldsymbol{w}_\star \rangle = 0$, have $\langle \hat{\boldsymbol{w}}, \boldsymbol{\delta} \rangle = \langle \hat{\boldsymbol{w}}_\perp, \boldsymbol{\delta} \rangle$, and hence the margin condition is equivalent to $|\langle \hat{\boldsymbol{w}}_\perp, \boldsymbol{\delta} \rangle| > |\langle \hat{\boldsymbol{w}}, \boldsymbol{x} \rangle|$. This condition is satisfied if and only if it is satisfied by the supremum:

$$\sup_{\boldsymbol{\delta} \in H_q(\varepsilon)} |\langle \hat{\boldsymbol{w}}_\perp, \boldsymbol{\delta} \rangle| > |\langle \hat{\boldsymbol{w}}, \boldsymbol{x} \rangle| \tag{10}$$

Standard results from convex analysis implies that the supremum is achieved by the $\ell^{q^\star}$ distance to $\operatorname{span}(\boldsymbol{w}_\star)$, also known as the metric projector:

$$\sup_{\boldsymbol{\delta} \in H_q(\varepsilon)} |\langle \hat{\boldsymbol{w}}_\perp, \boldsymbol{\delta} \rangle| = \varepsilon \inf_{\mu \in \mathbb{R}} ||\hat{\boldsymbol{w}}_\perp - \mu \boldsymbol{w}_\star||_{q^\star} \coloneqq \varepsilon \, d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp) \tag{11}$$

where $q^\star$ is the dual in the ℓ^q sense: ${}^1/_q + {}^1/_{q^\star} = 1$. □

Remark 3. For $q = 2$, the infimum in eq. (11) is achieved at $\mu = 0$:

$$d^\star_2(\hat{\boldsymbol{w}}_\perp) = \inf_{\mu \in \mathbb{R}} ||\hat{\boldsymbol{w}}_\perp - \mu \boldsymbol{w}_\star||_2 = ||\hat{\boldsymbol{w}}_\perp||_2. \tag{12}$$

While $d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp) \leq ||\hat{\boldsymbol{w}}_\perp||_q$ is always an upper bound, it is not tight for $q \neq 2$, except for particular choices of $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$, for instance $\boldsymbol{w}_\star = \sqrt{d} \boldsymbol{e}_1$. This highlights how the existence of consistent attacks crucially depend on an interplay between the Euclidean geometry of the constraint set and the ℓ^q geometry of the adversarial attack.

A similar condition to eq. (9) holds for an inconsistent attack, but without the orthogonality constraint $\langle \boldsymbol{w}_\star, \boldsymbol{\delta} \rangle = 0$. Since:

$$||\hat{\boldsymbol{w}}||_{q^\star} \geq ||\hat{\boldsymbol{w}}_\perp||_{q^\star} \geq d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp) \tag{13}$$

this provides a less strict existence condition, as expected. In particular, the stronger the overlap between the ground-truth and the model $\langle \hat{\boldsymbol{w}}, \boldsymbol{w}_\star \rangle$, the stronger the attack needs to be in order to flip the model prediction, in contrast to inconsistent perturbations which are independent of the ground-truth weights $\boldsymbol{w}_\star$. This leads to the following corollary.

Corollary 1 (Existence of inconsistent attack)**.** *Under the same setting as Proposition 1, an inconsistent adversarial attack exists if and only if:*

$$\epsilon ||\hat{\boldsymbol{w}}||_{q^\star} \geq |\langle \hat{\boldsymbol{w}}, x \rangle|. \tag{14}$$

Since $\rho = {}^{d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp)}/_{||\hat{\boldsymbol{w}}||_{q^\star}} \in [0, 1]$, this further implies the following bounds:

$$\rho E_{\text{rob}} \leq E^{\text{cns}}_{\text{rob}} \leq E_{\text{rob}} \, . \tag{15}$$

Proof. The existence part follows the same proof as in Proposition 1, but without the orthogonality constraint. We then have:

$$\sup_{\boldsymbol{\delta} \in B_q(\varepsilon)} |\langle \hat{\boldsymbol{w}}, \boldsymbol{\delta} \rangle| = \varepsilon ||\hat{\boldsymbol{w}}||_{q^\star}. \tag{16}$$

The upper-bound is immediate from Remark 2. The lower-bound follows from noting that $E^{\text{cns}}_{\text{rob}}(\varepsilon) = E_{\text{rob}}(\rho\varepsilon)$ and that both errors are non-decreasing functions of ε. □

Proposition 1 allow us to identify the region in $\mathbb{R}^d$ which is vulnerable to consistent attacks. Indeed, defining the ground-truth orthogonal margin

$$\kappa_q(\boldsymbol{x}) = \frac{|\langle \hat{\boldsymbol{w}}, \boldsymbol{x} \rangle|}{d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp)} \tag{17}$$

a covariate $\boldsymbol{x} \in \mathbb{R}^d$ is vulnerable to a consistent $\boldsymbol{\delta} \in H_q(\varepsilon)$ attack if and only if $\varepsilon > \kappa_q(\boldsymbol{x})$, and the *vulnerable region* is given by $\{\boldsymbol{x} \in \mathbb{R}^d : \kappa_q(\boldsymbol{x}) < \epsilon\} \subset \mathbb{R}^d$ — a tube around the decision hyperplane. Note that this implies two ways of mitigating consistent adversarial attacks (i.e. increase κ_q): (a) To align with the target weights m; (b) To reduce $d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp)$. While the first option is typically out of the control of the statistician, the second option can be achieved by explicitly regularizing the training with respect to the norm dual to the attack, which is an upper-bound to $d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp)$ — see eq. (13). This is consistent with previous theoretical results suggesting the use of the dual norm in ERM [47, 59–61]. This will be the subject of Section 3.3.

Another important factor in the margin $\kappa_q(\varepsilon)$ is the interplay between the underlying Euclidean (ℓ^2) geometry defining the classifier and the ℓ^q geometry of the adversarial attack. This interplay is better illustrated in the Gaussian case.

Corollary 2 (Existence for Gaussian covariate)**.** *In the case of i.i.d. Gaussian covariates $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, {}^1\!/_d \operatorname{Id}_d)$, the probability a consistent attack exists is given by:*

$$\mathbb{P}\left[\exists \textit{ consistent } \boldsymbol{\delta} \in H_q(\varepsilon)\right] = 2\Phi\left(\varepsilon\sqrt{d}\frac{d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp)}{||\hat{\boldsymbol{w}}||_2}\right) - 1\,, \tag{18}$$

where Φ is the standard normal c.d.f.

Proof. Conditionally on the predictor, we have $\langle \hat{\boldsymbol{w}}, \boldsymbol{x} \rangle \overset{d}{=} {}^{||\hat{\boldsymbol{w}}||_2}\!/\sqrt{d}\; Z$ with $Z \sim \mathcal{N}(0,1)$. It is immediate to see that the condition in eq. (9) implies the result. □

Remark 4. The function in eq. (18) is non-decreasing in q (non-increasing in $q^\star$) and d. To get some intuition, consider the case of a random predictor, of unit norm, with a correlation $m = {}^{\langle \boldsymbol{w}_\star, \hat{\boldsymbol{w}} \rangle}\!/_d$ (in expectation) with the target:

$$\hat{\boldsymbol{w}} = m\boldsymbol{w}_\star + \sqrt{1-m^2}\boldsymbol{\xi} \tag{19}$$

with $\boldsymbol{\xi} \sim \mathcal{N}(0, \operatorname{Id}_d)$. It is immediate to show that:

$$\mathbb{P}\left[\exists \text{ consistent } \boldsymbol{\delta} \in H_q(\varepsilon)\right] = 2\Phi\left(\varepsilon\sqrt{d}\sqrt{1-m^2}\frac{d^\star_{q^\star}(\boldsymbol{\xi})}{||\hat{\boldsymbol{w}}||_2}\right) - 1 \tag{20}$$

since $d^\star_{q^\star}(\boldsymbol{\xi}) = \Theta(d^{1/q^\star})$ for $q \geq 1$ when $d \to \infty$, for $\varepsilon = \Theta(1)$ and $q > 1$ the probability of existence of a consistent attack is almost surely one unless the predictor achieves perfect alignment $m^2 = 1$ with the target. This highlights the susceptibility of high-dimensional predictors to adversarial attacks. However, the situation can be quite different for a sparse predictor, since the enumerator $d^\star_{q^\star}(\hat{\boldsymbol{w}}_\perp)$ only penalize the part of the support which does not overlap with the target target. We report how the existence probability depends on the attack geometry q and the dimension d in Figure 2.

3.2 Robust Empirical Risk Minimization

A natural question is whether robust training can effectively mitigate consistent attacks. Robust empirical risk minimization emerged as a principled way to learn classifier rules from data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{-1, +1\} : i = 1, \ldots, n\}$ that are inherently robust to adversarial perturbations. From the dataset $\mathcal{D}$, the statistician estimates a classifier by optimizing the robust *empirical* (regularized) risk, defined as

$$\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{n} \max_{\|\boldsymbol{v}_i\|_s \leq r} \ell(y_i\langle \boldsymbol{w}, \boldsymbol{x}_i + \boldsymbol{v}_i \rangle) + \lambda\|\boldsymbol{w}\|_2^2\,, \tag{21}$$

where $\ell : \mathbb{R} \to \mathbb{R}_+$ is a non-increasing convex loss function, the term $\|\boldsymbol{w}\|_2^2$ is a convex regularization term, and $\lambda \geq 0$ is a regularization parameter. The inner maximization over $\boldsymbol{v}_i$ models the worst-case perturbation for each data

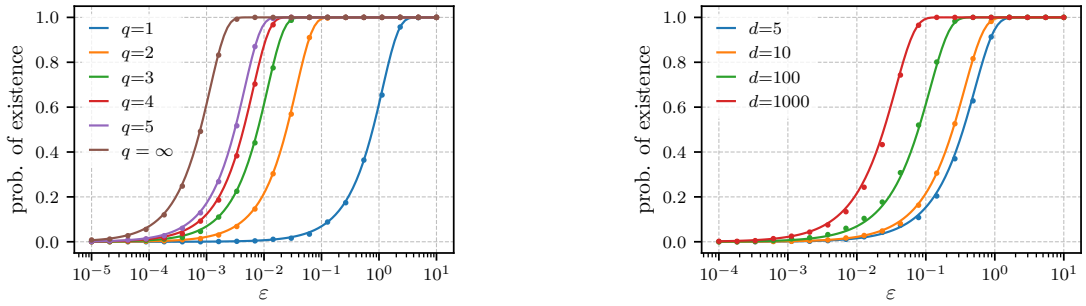

Figure 2: Probability of existence of a consistent adversarial attack for Gaussian covariates in eq. (18) as a function of the radius $\varepsilon > 0$, with $\boldsymbol{w}_\star = \sqrt{d}\boldsymbol{e}_1 \in \mathbb{R}^d$ and $\hat{\boldsymbol{w}} \in \mathbb{S}^{d-1}(\sqrt{d})$ with correlation $m = \langle \boldsymbol{w}_\star, \hat{\boldsymbol{w}} \rangle / d = 0.5$. *(Left)* Different curves show different choices of attack geometry q with $d = 10$. *(Right)* different curves show different covariate dimension d, for fixed $q = 2$. Solid curves were computed from the theoretical expression, while dots are computed by drawing $n = 10^3$ and estimating the frequency of times the constraints in eq. (9) is satisfied.

point, constrained by the attack budget r during training. The case with $r = 0$ corresponds to standard ERM while any case with $r > 0$ corresponds to robust training. Given the dataset $\mathcal{D}$, we estimate the model binary classifier as

$$\hat{\boldsymbol{w}} \in \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{w}) \,. \tag{22}$$

While robust training has proven effective in practice, understanding its properties for protection to consistent attacks still requires analysis.

3.3 High-dimensional asymptotic analysis

Motivated by Remark 4, we now investigate the behavior of both standard and robustly trained predictors in the high-dimensional limit where consistent adversarial attacks proliferate. More concretely, we will derive sharp asymptotic results for the case where $\hat{\boldsymbol{w}}$ is a trained predictor under the Gaussian design assumption, and discuss the benefits of robust empirical risk minimization in mitigating consistent adversarial attacks. We will work under the following assumptions.

Assumption 3.1 (Data distribution)**.** We assume the covariates are isotropic Gaussian $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, {}^1\!/_d \, \mathrm{Id}_d)$ and that labels are generated from a ground-truth linear classifier $y \sim \mathrm{Rad}(f_{\boldsymbol{w}_\star}(\boldsymbol{x}))$ where $f_{\boldsymbol{w}_\star}(\boldsymbol{x}) = \mathbb{P}(y = +1|\boldsymbol{x}) = \varphi(\langle \boldsymbol{w}_\star, \boldsymbol{x} \rangle)$ with monotonic link function φ and ground-truth weights $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$.

Assumption 3.2 (Scaling of the Adversarial Strength)**.** For a given perturbation geometry $\boldsymbol{\delta} \in B_q(\varepsilon)$ with $q > 1$, we assume that $\varepsilon = O_d(d^{-1/q^\star})$ as $d \to \infty$, where $q^\star$ is the dual. We define the rescaled radius as $\widetilde{\varepsilon} = \varepsilon \, d^{1/q^\star}$.

Remark 5. As briefly discussed in Remark 4, Assumption 3.2 provides the right scaling for non-trivial attacks in the high-dimensional limit considered in this work: a slower scaling would result in a perturbation strength which is too weak and any faster scaling would result in a perturbation that flips any label. The same scaling was considered in previous asymptotic analyses of robust training in [47, 49, 52, 54].

Assumption 3.3 (Asymptotic Gaussianity)**.** We consider the high-dimensional limit for which $d \to \infty$. Suppose that $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$ and that $\boldsymbol{w}$ is a d dimensional Gaussian such that the following quantities concentrate to

$$\frac{1}{d}\|\boldsymbol{w}\|_2^2 \xrightarrow[d\to\infty]{} q \,, \quad \frac{1}{d}\langle \boldsymbol{w}_\star, \boldsymbol{w} \rangle \xrightarrow[d\to\infty]{} m \,, \tag{23}$$

where $q, m \in \mathbb{R}$ such that $q \geq m^2$.

Remark 6. Even though this assumption may appear restrictive at first sight, it is asymptotically satisfied in the high-dimensional limit of interest in this work by different estimators, such as the minimizer of a convex ERM [9, 20, 62, 63], robust ERM [10] and Bayesian estimation [64–66]. This will be precisely the case we study for robust training in the following analysis.

The performance of robust adversarial training for well-specified linear classifiers on Gaussian covariates (Assumption 3.1) has been studied by [47] in the proportional high-dimensional asymptotics where n diverges with d at constant ratio $\alpha = n/d = \Theta(1)$. In particular, the authors characterize the limiting distribution of the entries of $\hat{\boldsymbol{w}}$ as a function of the parameters (q, m) satisfying Assumption 3.3 (c.f. Remark 6).

We are now ready to state our main result about the limiting behavior of the consistent metrics in the high-dimensional limit.

Theorem 3.1 (Consistent metrics for well-specified model)**.** *Under Assumptions 3.1 to 3.3 the metrics defined in eqs. (2) and (3) with decision rule $\hat{y}(\boldsymbol{x}) = \operatorname{sign}(2f_{\boldsymbol{w}}(\boldsymbol{x}) - 1)$ concentrate in high dimension to the following two dimensional integrals*

$$E_{\rm rob}^{\rm cns} = \int \mathrm{d}(\nu, \mu)\, \mathbb{1}\{\nu(\mu - \widetilde{\varepsilon}\mathcal{A}) < 0\}\,, \tag{24}$$

$$E_{\rm bnd}^{\rm cns} = \int \mathrm{d}(\nu, \mu)\, \mathbb{1}\{\nu(\mu - \widetilde{\varepsilon}\mathcal{A}) < 0\}\mathbb{1}\{\mu\nu > 0\}\,, \tag{25}$$

where

$$\mathcal{A} = \sqrt{q - m^2}\sqrt{2}\,\sqrt[q^\star]{\frac{\Gamma\left(\frac{q^\star+1}{2}\right)}{\sqrt{\pi}}}\,, \tag{26}$$

and the pair ν, μ is jointly Gaussian with zero mean and covariance $\left(\begin{smallmatrix} 1 & m \\ m & q \end{smallmatrix}\right)$ where the values are taken from eq. (23).

The proof of this statement can be found in Appendix B. The argument is based on the explicit solution of the inner maximization and subsequent evaluation of the resulting expression in the high dimensional limit.

For completeness we report here the limiting value of the $E_{\rm rob}$ as being

$$E_{\rm rob} = \int \mathrm{d}(\nu, \mu)\, \mathbb{1}\{\nu(\mu - \widetilde{\varepsilon}\mathcal{B}) < 0\}\,, \quad \mathcal{B} = \sqrt{2q}\,\sqrt[q^\star]{\frac{\Gamma\left(\frac{q^\star+1}{2}\right)}{\sqrt{\pi}}}\,, \tag{27}$$

and we note that this limiting form was already established in previous work [47, 49, 54].

An important observation is that the consistent version of the errors (eqs. (24) and (25)) depend on the quantity $\sqrt{q - m^2}$ while the inconsistent version (eq. (27)) depends on $\sqrt{q}$. Since $\sqrt{q - m^2} < \sqrt{q}$ (as $q \geq m^2 > 0$), this mathematical distinction explains why consistent attacks are less effective than inconsistent ones for the same attack strength ε, confirming our earlier result from Corollary 1 and as illustrated in Figure 3 *(Center)*. The former quantity is precisely the length of $P_\perp \boldsymbol{w}$ appearing in Proposition 1.

Some additional experiments that compare the consistent and inconsistent version of the boundary error are provided in Appendix A.

Figure 3 *(Center)* shows the asymptotic dependence of the metrics in Definition 2 with the rescaled perturbation strength $\widetilde{\varepsilon}_g$ in the high-dimensional limit. This provides a quantitative measure of how strong a consistent adversarial attack needs to be to flip a certain percentage of the classifier labels: for instance, to flip 50% of the labels with an L_∞ attack one needs $\widetilde{\varepsilon}_g \approx 1$ ($\varepsilon_g \approx d^{-1/2}$).

Figure 3 *(Right)* shows the performance of robustly trained $\hat{\boldsymbol{w}}$ as a function of $\alpha = n/d$, demonstrating a monotonic decrease of all the metrics defined above with the sample complexity α for two different attack geometries. While the errors $E_{\rm rob}$ and $E_{\rm rob}^{\rm cns}$ start from the same values, the value $E_{\rm rob}^{\rm cns}$ decreases faster with α, indicating that with more samples the model learns more robust representations that are particularly effective against proper adversarial attacks.

Together with Theorem 3.1, we can leverage the results from [47] to study the consistent errors of $\hat{\boldsymbol{w}}$ from eq. (22) trained from a dataset of n input-output pairs and dimension d where both n and d diverge with $\alpha = n/d$ fixed. Additional details are discussed in Appendix B.

4 The role of overparameterization: Latent Variable Model

Despite many empirical works on the subject, the interplay between adversarial attacks and overparameterization remains poorly understood, with contradictory evidence on the susceptibility of large neural networks to adversarial attacks [67].

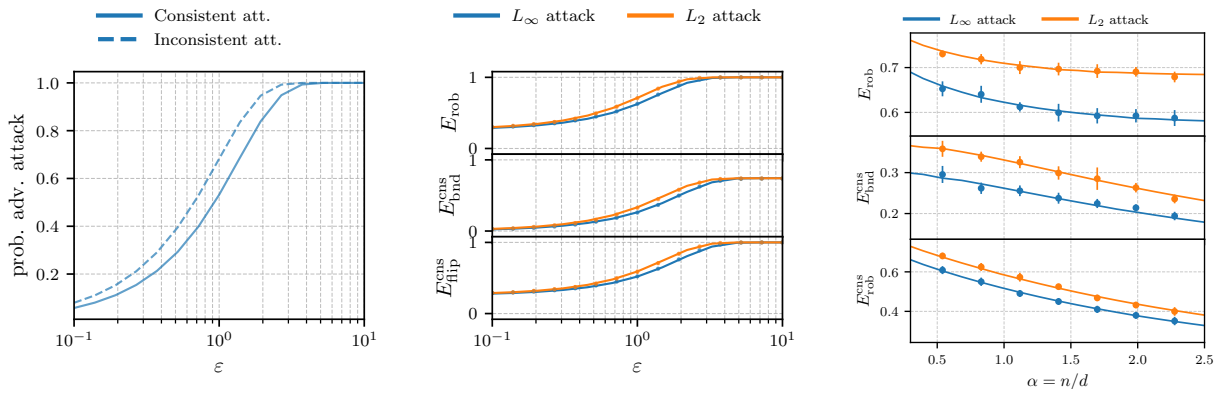

Figure 3: *(Left)* Probability of existence of adversarial perturbations under Gaussian data for both cases of consistent and inconsistent attacks. Here we consider the predictor trained with non-robust training and vanishing regularization $\lambda = 10^{-3}$. *(Center)* Dependence of the metrics in eqs. (1) to (3) for the well-specified model as a function of the attackers norm. Here we have taken $q = 31.786$ and $m = 3.879$. The points are simulations for $d = 500$. We see good agreement already at finite dimension. *(Right)* Performance difference for optimally regularized non robust training under attacks constrained with different norms. The points show the average and std of 10 different realizations with $d = 500$.

In this section, we investigate this question on a popular mathematical testbed for studying the role of overparameterization, the *latent variable model* [13]. In this model, the ground-truth classifier $f_{\boldsymbol{w}_\star}(\boldsymbol{z}) = \varphi(\langle \boldsymbol{w}_\star, \boldsymbol{z}\rangle)$ is defined in a latent space with latent covariates $\boldsymbol{z} \in \mathbb{R}^d$ and weights $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$. Labels are generated according to the latent rule $y \sim \mathrm{Rad}(f_{\boldsymbol{w}_\star}(\boldsymbol{z}))$ as in eq. (6). The statistician does not observe the latent covariates $\boldsymbol{z}$ directly but instead has access to a transformed representation $\boldsymbol{x} \in \mathbb{R}^p$ defined as

$$\boldsymbol{x} = \mathrm{F}\,\boldsymbol{z} + \boldsymbol{u}\,, \tag{28}$$

where $\mathrm{F} \in \mathbb{R}^{p\times d}$ is the *feature matrix* and $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, {}^1\!/_p\,\mathrm{Id}_p)$ is an independent covariate noise term.

While this model seems simplistic, recent Gaussian universality results have shown that in the proportional limit, ERM on this latent variable model is equivalent to ERM on a two-layer neural network with frozen first-layer weights (a.k.a. *random features model*) [18, 27, 29, 32–34, 68]. This places the latent variable model as a convenient testbed to study the phenomenology associated to overparameterized networks — such as benign interpolation — in a mathematically tractable setting. In this mapping, the level of overparameterization is given by the features dimension p. For this reason, we will often switch between the latent space and the random features when discussing the model, for instance referring to $p > d$ as the *overparameterized* case.

4.1 Geometry of consistent attacks on the latent space

We now discuss the geometrical properties of consistent attacks in the latent variable model. Note that in this context an adversary could either attack the latent space ($\boldsymbol{\delta} \in \mathbb{R}^d$) or feature space ($\boldsymbol{\delta} \in \mathbb{R}^p$). Considering perturbations in feature space, i.e. perturbations to $\boldsymbol{x}$, will result in a similar analysis as the one carried out for the model of Section 2. Therefore in the following we focus on the latter, where the conditions in Definition 1 translate to:

- **Target invariance:** $\boldsymbol{\delta} \in H_q(\varepsilon) = \{\boldsymbol{\delta} \in B_q(\varepsilon) : \langle \boldsymbol{w}_\star, \boldsymbol{\delta}\rangle = 0\}$.
- **Model deception:** $|\langle \hat{\boldsymbol{\theta}}, \mathrm{F}\,\boldsymbol{\delta}\rangle| > |\langle \hat{\boldsymbol{\theta}}, \mathrm{F}\,\boldsymbol{z} + \boldsymbol{u}\rangle|$ and $\mathrm{sign}(\hat{\boldsymbol{\theta}}, \mathrm{F}\,\boldsymbol{\delta}) = -\,\mathrm{sign}(\langle \hat{\boldsymbol{\theta}}, \mathrm{F}\,\boldsymbol{z} + \boldsymbol{u}\rangle)$

where the model weights are denoted by $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$ to avoid confusion. Adapting the argument in Section 3.1 to this case is straightforward, leading to the following characterization of consistent latent space attacks.

Proposition 2 (Existence of consistent latent space attacks)**.** *Consider the setting of binary classification in the latent space model: a linear classifier defined by the weights $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$ assign labels according to $y \sim \mathrm{Rad}(f_{\boldsymbol{w}_\star}(\boldsymbol{z}))$ where $f_{\boldsymbol{w}_\star}(\boldsymbol{z}) = \varphi_\star(\langle \boldsymbol{w}_\star, \boldsymbol{z}\rangle)$, while the statistician observes only the pairs $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \{-1, +1\}$ with $\boldsymbol{x} = F\boldsymbol{z} + \boldsymbol{u} \in \mathbb{R}^d$,*

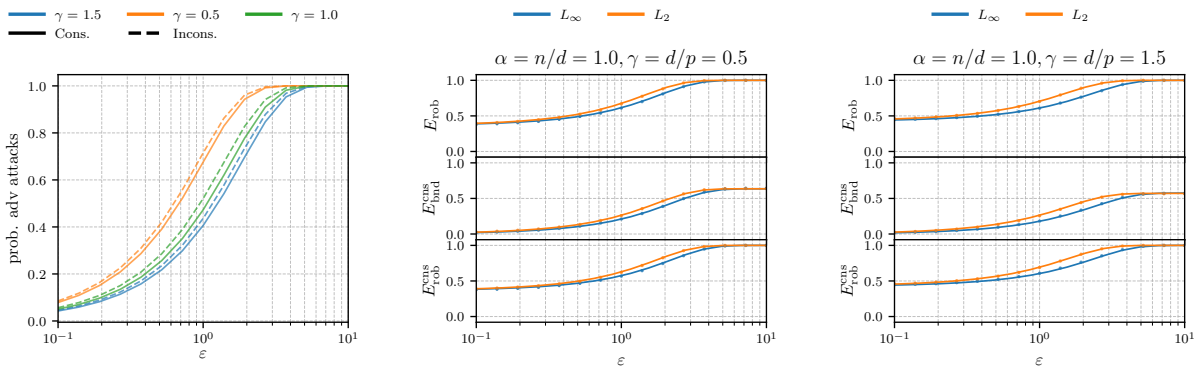

Figure 4: Dependence of the different perturbation metrics for the Latent Space model of Section 4. Here we consider the case of non-robust empirical risk minimizer in the min norm limit ($\lambda = 10^{-3}, r = 0$). We see that the consistent perturbations are less effective than the standard one both in the underparameterized ($\gamma > 1$) and in the overparameterized ($\gamma < 1$).

fitting a linear classifier $f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}) = \varphi(\langle \hat{\boldsymbol{\theta}}, \boldsymbol{x} \rangle)$ *with weights* $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$. *Then, a consistent attack* $\boldsymbol{\delta} \in B_q(\varepsilon)$ *with respect to* $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$, $\boldsymbol{z} \in \mathbb{R}^d$ *and the decision function* $\hat{y}(\boldsymbol{x}) = \text{sign}(2 f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}) - 1)$ *exists if and only if:*

$$\varepsilon d_{q^\star}^\star (P_\perp F^\top \hat{\boldsymbol{\theta}}) \geq |\langle \hat{\boldsymbol{\theta}}, F\boldsymbol{z} + \boldsymbol{u} \rangle| \tag{29}$$

where $P_\perp = \text{Id}_d - \boldsymbol{w}_\star \boldsymbol{w}_\star^\top / d$ *is the projector in the space orthogonal target weights and* $q^\star$ *is the dual of* q.

Remark 7. While in the well-specified case the vulnerable region is determined by the margin $\kappa_q(\boldsymbol{x})$ defined in eq. (17), in the latent model this is defined by the latent margin:

$$\eta_q(\boldsymbol{z}) := \frac{|\langle \hat{\boldsymbol{\theta}}, \text{F}\, \boldsymbol{z} + \boldsymbol{u} \rangle|}{d_{q^\star}^\star (P_\perp F^\top \hat{\boldsymbol{\theta}})} \tag{30}$$

Note that this can be larger or smaller than $\kappa_q(\boldsymbol{x})$, depending on the details of the problem.

Corollary 3 (Existence for Gaussian latent variables)**.** *In the case of i.i.d. Gaussian latent variables* $\boldsymbol{z} \sim \mathcal{N}(0, {}^1\!/_d \, \text{Id}_d)$ *and* $\boldsymbol{u} \sim \mathcal{N}(0, {}^1\!/_p \, \text{Id}_p)$, *the probability a consistent attack exists is given by:*

$$\mathbb{P}\left[\exists \text{ consistent } \boldsymbol{\delta} \in H_q(\varepsilon)\right] = 2\Phi\left(\frac{d_{q^\star}^\star (P_\perp F^\top \hat{\boldsymbol{\theta}})}{\sqrt{||\hat{\boldsymbol{\theta}}||_2^2 + {}^p\!/_d ||\, \text{F}^\top \hat{\boldsymbol{\theta}}||_2^2}} \sqrt{p} \varepsilon \right) - 1 \tag{31}$$

where Φ *is the standard normal c.d.f.*

Remark 8. In the latent variable model, it is the projection of the predictor in latent space $\text{F}^\top \hat{\boldsymbol{\theta}}$ and not the predictor itself that counts for the probability of existence. In particular, the energy of $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$ which is part of $\text{Ker}(\text{F}^\top)$ only contributes to the ℓ^2 norm in the denominator. In other words: in the overparametrized setting $p > d$ one can reduce the probability of existence of consistent attacks by both having high alignment with the target $\langle \boldsymbol{w}_\star, \text{F}^\top \hat{\boldsymbol{\theta}} \rangle$ or by placing a lot of energy in the $p - d$ directions in $\text{Ker}(\text{F}^\top)$. This is closely related to the conditions for benign overfitting in [69].

4.2 High-dimensional asymptotics

We now move to the analysis of trained predictors in the context of the latent variable model.

Consider training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \{-1, +1\} : i = 1, \cdots, n\}$ independently drawn from the latent variable model. Our goal in this section is to characterize the asymptotic behavior of the estimated binary linear classifier defined by the vector $\hat{\boldsymbol{\theta}}$ estimated from $\mathcal{D}$ using eqs. (21) and (22). Our results will hold under the following assumptions.

Assumption 4.1 (High-dimensional limit)**.** We consider the proportional high-dimensional limit where $n, p, d \to \infty$ at fixed ratios $\alpha := {}^{n}/_{d}$ and $\psi := {}^{p}/_{n}$. For convenience, we also define $\gamma := (\alpha\psi)^{-1} = {}^{d}/_{p}$.

Assumption 4.2 (Data Distribution Latent Space Model)**.** We assume that data $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \{-1, +1\}$ is drawn from a latent variable model with $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, {}^{1}/_{d}\,\mathrm{Id}_d)$ and ground-truth linear classifier $f_{\boldsymbol{w}_\star}(\boldsymbol{z}) = \varphi(\langle \boldsymbol{w}_\star, \boldsymbol{z}\rangle)$ with $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$. The observed features $\boldsymbol{x} \in \mathbb{R}^p$ are generated as $\boldsymbol{x} = \mathrm{F}\,\boldsymbol{z} + \boldsymbol{u}$ with $\boldsymbol{u} \sim \mathcal{N}(0, \mathrm{Id}_p)$ independent of the other quantities and

$$\mathrm{F} = \begin{cases} \begin{bmatrix} \sqrt{\frac{p}{d}}\,\mathrm{Id}_d \\ \mathbf{0}_{(p-d)\times d} \end{bmatrix} & \text{if } p \geq d \\ \begin{bmatrix} \mathrm{Id}_p & \mathbf{0}_{p\times(d-p)} \end{bmatrix} & \text{if } p < d \end{cases} . \tag{32}$$

Remark 9. All the phenomenology that follows also hold for a random Gaussian feature matrix $F \in \mathbb{R}^{p\times d}$. The choice of feature matrix in eq. (53) was previously considered in [13] in the context of ridge regression. Our results also extended this discussion to binary classification.

In the following we will investigate the impact of adversarial attacks in the latent variable model. Note that in this context an adversary could either attack the latent space ($\boldsymbol{\delta} \in \mathbb{R}^d$) or feature space ($\boldsymbol{\delta} \in \mathbb{R}^p$). Considering perturbations in feature space, i.e. perturbations to $\boldsymbol{x}$, will result in a similar analysis as the one carried out for the model of Section 2. Therefore in the following we focus on the second case.

The consistent and inconsistent adversarial errors associated to a predictor $\hat{y}$ are defined in the latent space model as

$$E_{\mathrm{rob}}^{\mathrm{cns}} = \mathbb{E}\left[\max_{\boldsymbol{\delta}\in B_q(\varepsilon): f_\star(\boldsymbol{x}) = f_\star(\boldsymbol{x}+\boldsymbol{\delta})} \mathbb{1}\{y \neq \hat{y}(\mathrm{F}(\boldsymbol{z}+\boldsymbol{\delta}) + \boldsymbol{u})\}\right], \tag{33}$$

$$E_{\mathrm{bnd}}^{\mathrm{cns}} = \mathbb{E}\left[\max_{\boldsymbol{\delta}\in B_q(\varepsilon): f_\star(\boldsymbol{x}) = f_\star(\boldsymbol{x}+\boldsymbol{\delta})} \mathbb{1}\{y \neq \hat{y}(\mathrm{F}(\boldsymbol{z}+\boldsymbol{\delta}) + \boldsymbol{u})\}\mathbb{1}\{y = \hat{y}(\mathrm{F}\,\boldsymbol{z} + \boldsymbol{u})\}\right], \tag{34}$$

$$E_{\mathrm{rob}} = \mathbb{E}\left[\max_{\boldsymbol{\delta}\in B_q(\varepsilon)} \mathbb{1}\{y \neq \hat{y}(\mathrm{F}(\boldsymbol{z}+\boldsymbol{\delta}) + \boldsymbol{u})\}\right]. \tag{35}$$

The main technical result for this part consists in characterizing the high dimensional behavior of the robust empirical risk minimizer $\hat{\boldsymbol{\theta}}$ as the solution of a system of self-consistent equations. We state here the result for $s = \infty$ in eqs. (21) and (22) and leave the case for generic s to the appendix.

Theorem 4.1 (Self Consistent equations for Latent Space Model)**.** *Under Assumptions 3.2, 4.1 and 4.2 the values of the following sufficient statistics*

$$m = \frac{1}{d}\boldsymbol{w}_\star^\top \mathrm{F}^\top \hat{\boldsymbol{\theta}}\,, \quad q_f = \frac{1}{p}\|\hat{\boldsymbol{\theta}}\|_2^2\,, \quad q_\ell = \frac{1}{d}\hat{\boldsymbol{\theta}}^\top \mathrm{F}\,\mathrm{F}^\top \hat{\boldsymbol{\theta}}\,, \quad q = q_\ell + q_f\,, \quad P = \frac{1}{p}\|\hat{\boldsymbol{\theta}}\|_{q^\star}^{q^\star}\,, \tag{36}$$

concentrate in high dimension to the solution of the following system of self consistent equations. The self-consistent equations are made of a first set

$$\begin{cases} \hat{m} = \alpha\sqrt{\gamma}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \partial_\omega \mathcal{Z}_0 f_\ell\right] \\ \hat{q} = \alpha\gamma\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \partial_\omega \mathcal{Z}_0 f_\ell\right] \\ \hat{V} = -\alpha\gamma\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \partial_\omega \mathcal{Z}_0 f_\ell\right] \\ \hat{P} = 2rP^{1/s}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \partial_\omega \mathcal{Z}_0 f_\ell\right] \end{cases}, \tag{37}$$

that depend on the loss function g and binary linear classifier link function φ through $\mathcal{Z}_0 \equiv \mathcal{Z}_0(y, {}^{m}/_{\sqrt{q}}\xi, 1 - {}^{m^2}/_{q})$ and $f_\ell \equiv f_\ell(y, \sqrt{q}\xi, V, P)$ defined as

$$\mathcal{Z}_0(y, \omega, V) = \mathbb{E}_{z\sim\mathcal{N}(\omega,V)}[\mathbb{P}(y \mid z)]\,, \quad f_\ell(y, \omega, V, P) = \left(\mathcal{P}_{V\ell(y,\cdot,-y\varepsilon\sqrt[s^\star]{P})}(\omega) - \omega\right)/V\,, \tag{38}$$

where $\mathbb{P}(y \mid z)$ is different from zero only for $y = \pm 1$ and it is equal to $\mathbb{P}(y = +1 \mid z) = \varphi(z)$ from eq. (6). Also with $\mathcal{P}_{f(\cdot)}$ we indicate the proximal operator of a function f and $\xi \sim \mathcal{N}(0, 1)$. The second set of equations

$$\begin{cases} \hat{m} = \frac{1}{\sqrt{\gamma}}\mathbb{E}_\xi\left[\partial_\varrho \mathcal{Z}_{w_\star} f_w^1\right] \\ \hat{q} = (1+\gamma)\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}(f_w^1)^2\right] + (1-\gamma)\mathbb{E}_\xi\left[(f_w^2)^2\right] \\ \hat{V} = (1+\gamma)\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\partial_\varrho f_w^1\right] + (1-\gamma)\mathbb{E}_\xi\left[\partial_\varrho f_w^2\right] \\ \hat{P} = \gamma\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\left|f_w^1\right|^{s^\star}\right] + (1-\gamma)\mathbb{E}_\xi\left[\left|f_w^2\right|^{s^\star}\right] \end{cases} \quad \text{for } \gamma \leq 1\,, \tag{39}$$

$$\begin{cases} \hat{m} = \frac{1}{\sqrt{\gamma}}\mathbb{E}_\xi\left[\partial_\varrho \mathcal{Z}_{w_\star} f_w\right] \\ \hat{q} = 2\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star} f_w\right] \\ \hat{V} = 2\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star} \partial_\varrho f_w\right] \\ \hat{P} = \mathbb{E}_\xi\left[\mathcal{Z}_{w_\star} |f_w|^{s^\star}\right] \end{cases} \qquad \text{for } \gamma > 1\,, \tag{40}$$

depends on the regularization norm and the prior over the ground truth weights through $\mathcal{Z}_{w_\star} \equiv \mathcal{Z}_{w_\star}(\hat{m}\xi/\sqrt{(1+\gamma)\hat{q}}, \hat{m}^2/(1+\gamma)\hat{q})$, $f_w^1 \equiv f_w(\sqrt{\hat{q}}\sqrt{1+\frac{1}{\gamma}}\xi, \hat{V}(1+\frac{1}{\gamma}), \hat{P}/2)$ *and* $f_w^2 \equiv f_w(\sqrt{\hat{q}}\xi, \hat{V}, \hat{P}/2)$ *for* $\gamma \le 1$ *and* $\mathcal{Z}_{w_\star} \equiv \mathcal{Z}_{w_\star}(\hat{m}\xi/\sqrt{2\hat{q}}, \hat{m}^2/2\hat{q})$ *and* $f_w \equiv f_w(\sqrt{2\hat{q}}\xi, 2\hat{V}, \hat{P}/2)$ *for* $\gamma > 1$ *where*

$$\mathcal{Z}_{w_\star}(\varrho, \Lambda) = \frac{1}{\sqrt{\Lambda+1}} e^{\frac{\varrho^2}{2(\Lambda+1)}}\,, \quad f_w(\varrho, \Lambda, \pi) = \frac{\operatorname{sign}(\varrho/\Lambda)\cdot \max(|\varrho/\Lambda| - \pi/\Lambda, 0)}{2\lambda/\Lambda + 1}\,. \tag{41}$$

The values of q_ℓ *and* q_f *are obtained from the solution of the previous equations as*

$$q_\ell = \frac{1}{\gamma}\begin{cases} \mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}(f_w^1)^2\right] & \gamma \le 1 \\ \mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}(f_w)^2\right] & \gamma > 1 \end{cases}, \quad q_f = \begin{cases} \gamma\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}(f_w^1)^2\right] + (1-\gamma)\mathbb{E}_\xi\left[(f_w^2)^2\right] & \gamma \le 1 \\ \mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}(f_w)^2\right] & \gamma > 1 \end{cases}, \tag{42}$$

with the previous definitions of $\mathcal{Z}_{w_\star}$ *and* f_w.

The proof of the previous statement can be found in Appendix B. It is based on the use of Gordon Min-Max theorem to characterize the minimizer of the robust risk in eq. (21) through a low dimensional set of self consistent equation. This kind of asymptotic characterization is fairly common in the study of high dimensional system [12, 16, 70, 71].

With the previous result we can characterize the high-dimensional behavior of the proper adversarial errors in this data model.

Theorem 4.2 (Proper Metrics for Latent Space Model)**.** *Under the same setting of Theorem 4.1 the metrics defined in eqs. (33) and (34) evaluated for* $\hat{\boldsymbol{\theta}}$ *from eq. (22) and decision rule* $\hat{y} = \operatorname{sign}(2f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x}) - 1)$ *concentrate to the following values*

$$E_{\text{rob}}^{\text{cns}} = \int \mathrm{d}(\nu,\mu)\, \mathbb{1}\left\{\nu\left(\mu - \widetilde{\varepsilon}\sqrt[s^\star]{\mathcal{A}}\right) < 0\right\}, \tag{43}$$

$$E_{\text{bnd}}^{\text{cns}} = \int \mathrm{d}(\nu,\mu)\, \mathbb{1}\left\{\nu\left(\mu - \widetilde{\varepsilon}\sqrt[s^\star]{\mathcal{A}}\right) < 0\right\}\mathbb{1}\{\mu\nu > 0\}, \tag{44}$$

where

$$\mathcal{A} = \inf_{\kappa\in\mathbb{R}} \begin{cases} \gamma\mathbb{E}_\xi\left[\int \mathrm{d}w_\star\, h^1 |f_w^1 - \kappa w_\star|^{s^\star}\right] + \gamma\mathbb{E}_\xi\left[\int \mathrm{d}w_\star\, h^1 |\kappa w_\star|^{s^\star}\right] & \\ \quad + (1-\gamma)\mathbb{E}_\xi\left[\int \mathrm{d}w_\star\, h^2 |f_w^2 - \kappa w_\star|^{s^\star}\right] + (1-\gamma)\mathbb{E}_\xi\left[\int \mathrm{d}w_\star\, h^2 |\kappa w_\star|^{s^\star}\right] & \text{for } \gamma \le 1 \\ \mathbb{E}_\xi\left[\int \mathrm{d}w_\star\, h^3 |f_w - \kappa w_\star|^{s^\star}\right] + \mathbb{E}_\xi\left[\int \mathrm{d}w_\star\, h^3 |\kappa w_\star|^{s^\star}\right] & \text{for } \gamma > 1 \end{cases}, \tag{45}$$

where $h^1 \equiv h(\hat{m}\xi/\sqrt{(1+\gamma)\hat{q}}, \hat{m}^2/(1+\gamma)\hat{q})$, $h^2 \equiv h(0,0)$ *and* $h^3 \equiv h(\hat{m}\xi/\sqrt{2\hat{q}}, \hat{m}^2/2\hat{q})$ *with* $h(\varrho,\Lambda) = e^{-\frac{1}{2}w_\star^2} e^{-\frac{\Lambda}{2}w_\star^2 + \varrho w_\star}$. *Additionally the pair* ν, μ *is jointly Gaussian with zero mean and covariance* $\left(\begin{smallmatrix} 1 & m \\ m & q \end{smallmatrix}\right)$ *where the values of* q, m *are the ones obtained by the solution of the system of equations in Theorem 4.1.*

4.3 The interplay between overparameterization and consistent attacks

Theorem 4.1 and 4.2 allow us to investigate the efficacy of consistent adversarial attacks on overparameterized models.

We start by considering robust training with optimally tuned regularization parameter and r. We see that all the three metrics considered in this work decrease with a function of the amount of data used in training, as shown in Figure 5 *(Left)*, meaning that the more data is always beneficial no matter the metric considered. Interestingly there is a crossing for the different lines for different overparameterization level γ, meaning that the same level of overeparameterization is not optimal for any amount of data availability.

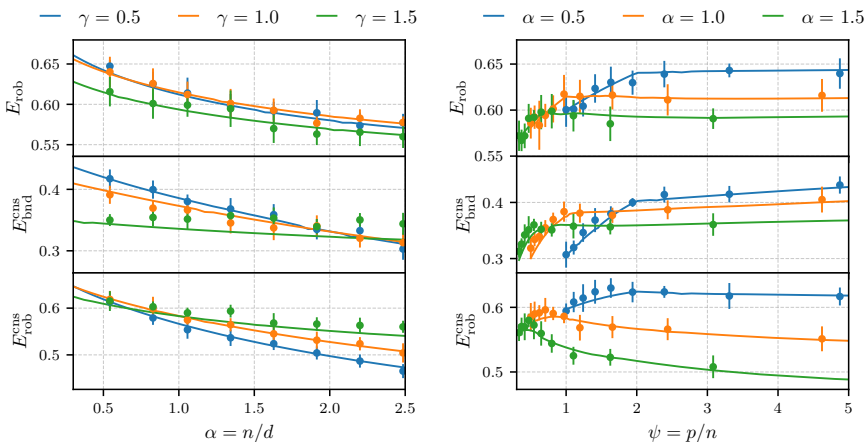

Figure 5: Dependence of error as a function of α and ψ for the latent space model defined in Section 4. For both panels the lines are the exact asymptotic solution of eqs. (37), (39) and (40) and the error bars are average and std over 10 realizations with $d = 500$ and p, n scaled accordingly. *(Left)* Robust error as a function of the number of data available during training. We see that all the metrics decrease as a function of the number of training data. *(Right)* Robust errors as a function of the number of latent space parameters. We see that while E_{rob} and $E^{\text{cns}}_{\text{bnd}}$ increase with the number of features while $E^{\text{cns}}_{\text{rob}}$ decreases.

In Figure 5 *(Right)* we consider the role of optimal robust training as a function of $\psi = p/n$. The metrics E_{rob} and $E^{\text{cns}}_{\text{rob}}$ present a different behavior in the large ψ regime. The first one stays approximately constant while the second one decreases with overparmeterization and this decrease is faster and faster the more data is given to the model (greater α). On the other hand we have that $E^{\text{cns}}_{\text{bnd}}$ in the same region is increasing.

In conclusion, although overparameterized models are more vulnerable to consistent adversarial attacks, this does not *a fortiori* imply a detriment in the overall model performance, as measured for instance by $E^{\text{cns}}_{\text{rob}}$, since improvement of previously badly classified points can have a compensatory effect. This might provide an explanation for the contradictory observations in the empirical literature [67]. Additional experiments are presented in Appendix A.

5 Conclusions

In this work, we investigated the fundamental distinction between consistent and inconsistent adversarial attacks in high-dimensional binary classification. We introduced novel metrics for consistent adversarial attacks and studied the robustness of the robust empirical risk minimization estimators, both in a well-specified and in a latent space setting. Curiously we found that overparameterization has a dual effect on consistent adversarial robustness depending on the error metric considered. Specifically: while the boundary error increases with overparameterization—indicating heightened vulnerability for correctly classified examples—the overall consistent robust error decreases. This counter-intuitive result stems from the beneficial role overparameterization plays in improving clean generalization, which mitigates the increased vulnerability of decision boundaries.

We hope that these findings will revel of value for the broader robust machine learning community. Rather than viewing overparameterization as detrimental to adversarial robustness, one should consider their specific robustness objectives and take into consideration that overparameterization could improve overall performances. Moreover, our exact characterizations provide theoretical guidance for selecting optimal regularization parameters and attack budgets during robust training.

Several limitations and directions for future work emerge from our study. Specifically exploring the connection between consistent adversarial robustness and other notions of robustness, such as distributional robustness or robustness to natural perturbations.

Acknowledgment

We thank Fanny Yang and Antonio Ribeiro for the inspiring discussions. This work was supported by the Swiss National Science Foundation under grants SNSF SMArtNet (grant number 212049) and SNSF OperaGOST (grant

number 200390) and by the French government, managed by the National Research Agency (ANR), under the France 2030 program with the reference "ANR-23-IACL-0008" and the Choose France - CNRS AI Rising Talents program.

References

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, 2018.

[3] Konstantin Donhauser, Alexandru Tifrea, Michael Aerni, Reinhard Heckel, and Fanny Yang. Interpolation can hurt robust generalization even when there is no noise. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23465–23477. Curran Associates, Inc., 2021.

[4] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7909–7919. PMLR, 13–18 Jul 2020.

[5] Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

[6] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45: 6056–6091, Apr 1992.

[7] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013.

[8] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[9] Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12199–12210. Curran Associates, Inc., 2020.

[10] Matteo Vilucchio, Emanuele Troiani, Vittorio Erba, and Florent Krzakala. Asymptotic characterisation of the performance of robust linear regression in the presence of outliers. In *International Conference on Artificial Intelligence and Statistics*, pages 811–819. PMLR, 2024.

[11] Antônio H Ribeiro and Thomas B Schön. Overparameterized linear regression under adversarial attacks. *IEEE Transactions on Signal Processing*, 71:601–614, 2023.

[12] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. The gaussian equivalence of generative models for learning with shallow neural networks. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova, editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 426–471. PMLR, 16–19 Aug 2022.

[13] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

[14] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

[15] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized Linear Regression: A precise analysis of the estimation error. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1683–1709, Paris, France, 03–06 Jul 2015. PMLR.

[16] Matteo Vilucchio, Yatin Dandi, Cedric Gerbelot, and Florent Krzakala. Asymptotics of non-convex generalized linear models in high-dimensions: A proof of the replica formula. *arXiv preprint arXiv:2502.20003*, 2025.

[17] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International conference on machine learning*, pages 6874–6883. PMLR, 2020.

[18] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

[19] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1024–1034. PMLR, 13–18 Jul 2020.

[20] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

[21] Koki Okajima, Xiangming Meng, Takashi Takahashi, and Yoshiyuki Kabashima. Average case analysis of lasso under ultra sparse conditions. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 11317–11330. PMLR, 25–27 Apr 2023.

[22] Urte Adomaityte, Gabriele Sicuro, and Pierpaolo Vivo. Classification of heavy-tailed features in high dimensions: a superstatistical approach. *Advances in Neural Information Processing Systems*, 36, 2024.

[23] Urte Adomaityte, Leonardo Defilippis, Bruno Loureiro, and Gabriele Sicuro. High-dimensional robust regression under heavy-tailed data: Asymptotics and universality. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(11):114002, 2024.

[24] Derek Bean, Peter J. Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013. doi: 10.1073/pnas.1307845110.

[25] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361, 2019. doi: 10.1109/ICASSP.2019.8683376.

[26] Zhenyu Liao, Romain Couillet, and Michael W Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13939–13950. Curran Associates, Inc., 2020.

[27] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[28] Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35: 4558–4570, 2022.

[29] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30285–30320. PMLR, 23–29 Jul 2023.

[30] Dominik Schröder, Daniil Dmitriev, Hugo Cui, and Bruno Loureiro. Asymptotics of learning with deep structured (Random) features. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 43862–43894. PMLR, 21–27 Jul 2024.

[31] Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 104630–104693. Curran Associates, Inc., 2024.

[32] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

[33] Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4310–4312. PMLR, 02–05 Jul 2022.

[34] Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Universality laws for gaussian mixtures in generalized linear models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 54754–54768. Curran Associates, Inc., 2023.

[35] Terence Tao and Van Vu. Random matrices: Universality of local eigenvalue statistics up to the edge. *Communications in Mathematical Physics*, 298:549–572, 2010.

[36] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[37] Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23549–23588. PMLR, 17–23 Jul 2022.

[38] Rishabh Dudeja, Yue M. Lu, and Subhabrata Sen. Universality of approximate message passing with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683, 2023.

[39] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3353–3364, 2019.

[40] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8093–8104. PMLR, 2020.

[41] Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

[42] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[43] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 36246–36263. PMLR, 2023.

[44] Chen Dan, Yuting Wei, and Pradeep Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pages 2345–2355. PMLR, 2020.

[45] Jacob Clarysse, Julia Hörrmann, and Fanny Yang. Why adversarial training can hurt robust accuracy. *arXiv preprint arXiv:2203.02006*, 2022.

[46] Antonio Ribeiro, Dave Zachariah, Francis Bach, and Thomas Schön. Regularization properties of adversarially-trained linear regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 23658–23670. Curran Associates, Inc., 2023.

[47] Matteo Vilucchio, Nikolaos Tsilivis, Bruno Loureiro, and Julia Kempe. On the geometry of regularization in adversarial training: High-dimensional asymptotics and generalization bounds. *arXiv preprint arXiv:2410.16073*, 2024.

[48] Antonio H. Ribeiro, Thomas B. Schön, Dave Zachariah, and Francis Bach. Efficient optimization algorithms for linear adversarial training. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1207–1215. PMLR, 03–05 May 2025.

[49] Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2034–2078. PMLR, 09–12 Jul 2020.

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

[51] Hamed Hassani and Adel Javanmard. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *The Annals of Statistics*, 52(2):441–465, 2024.

[52] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Asymptotic behavior of adversarial training in binary linear classification. *IEEE Trans. Neural Netw. Learn. Syst.*, PP, July 2023.

[53] Elvis Dohmatob and Meyer Scetbon. Precise accuracy/robustness tradeoffs in regression: Case of general norms. In *Forty-first International Conference on Machine Learning*, 2024.

[54] Kasimir Tanner, Matteo Vilucchio, Bruno Loureiro, and Florent Krzakala. A high dimensional statistical model for adversarial training: Geometry and trade-offs. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2530–2538. PMLR, 03–05 May 2025.

[55] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[56] Stanislav Fort and Balaji Lakshminarayanan. Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness. *arXiv preprint arXiv:2408.05446*, 2024.

[57] Elvis Dohmatob. Consistent adversarially robust linear classification: non-parametric setting. In *Forty-first International Conference on Machine Learning*, 2024.

[58] Elvis Dohmatob and Meyer Scetbon. Robust linear regression: Phase-transitions and precise tradeoffs for general norms. *arXiv preprint arXiv:2308.00556*, 2023.

[59] Dong Yin, Kannan Ramchandran, and Peter L. Bartlett. Rademacher complexity for adversarially robust generalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7085–7094. PMLR, 2019.

[60] Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441. PMLR, 2020.

[61] Nikolaos Tsilivis, Natalie Frank, Nathan Srebro, and Julia Kempe. The price of implicit bias in adversarially robust generalization. *arXiv preprint arXiv:2406.04981*, 2024.

[62] Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

[63] Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6):2879–2912, 2024.

[64] Jean Barbier, Florent Krzakala, Nicolas Macris, Leo Miolane, and Lenka Zdeborova. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116 (12):5451–5460, 2019. doi: https://doi.org/10.1073/pnas.1802705116.

[65] Lucas Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Theoretical characterization of uncertainty in high-dimensional linear classification. *Machine Learning: Science and Technology*, 4(2):025029, 2023.

[66] Lucas Andry Clarté, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. On double-descent in uncertainty quantification in overparametrized models. volume 206, pages 7089–7125. PMLR Proceedings of Machine Learning Research, 2023.

[67] Zhang Chen, Luca Demetrio, Srishti Gupta, Xiaoyi Feng, Zhaoqiang Xia, Antonio Emanuele Cinà, Maura Pintor, Luca Oneto, Ambra Demontis, Battista Biggio, et al. Over-parameterization and adversarial robustness in neural networks: An overview and empirical analysis. *arXiv preprint arXiv:2406.10090*, 2024.

[68] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023. doi: 10.1109/TIT.2022.3217698.

[69] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[70] Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of perceptrons with random labels. *Phys. Rev. E*, 109:034305, Mar 2024. doi: 10.1103/PhysRevE.109.034305.

[71] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Error scaling laws for kernel classification under source and capacity conditions. *Machine Learning: Science and Technology*, 4(3):035033, 2023.

[72] Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in R^n. In Joram Lindenstrauss and Vitali D. Milman, editors, *Geometric Aspects of Functional Analysis*, pages 84–106, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg. ISBN 978-3-540-39235-4.

[73] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

[74] Neal Parikh and Stephen Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, jan 2014. ISSN 2167-3888. doi: 10.1561/2400000003.

[75] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

[76] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[77] Gérard Biau and David M Mason. High-dimensional p p-norms. *Mathematical Statistics and Limit Theorems: Festschrift in Honour of Paul Deheuvels*, pages 21–40, 2015.

[78] M. Mezard, G. Parisi, and M.A. Virasoro. *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications*. World Scientific Lecture Notes In Physics. World Scientific Publishing Company, 1987. ISBN 9789813103917.

[79] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

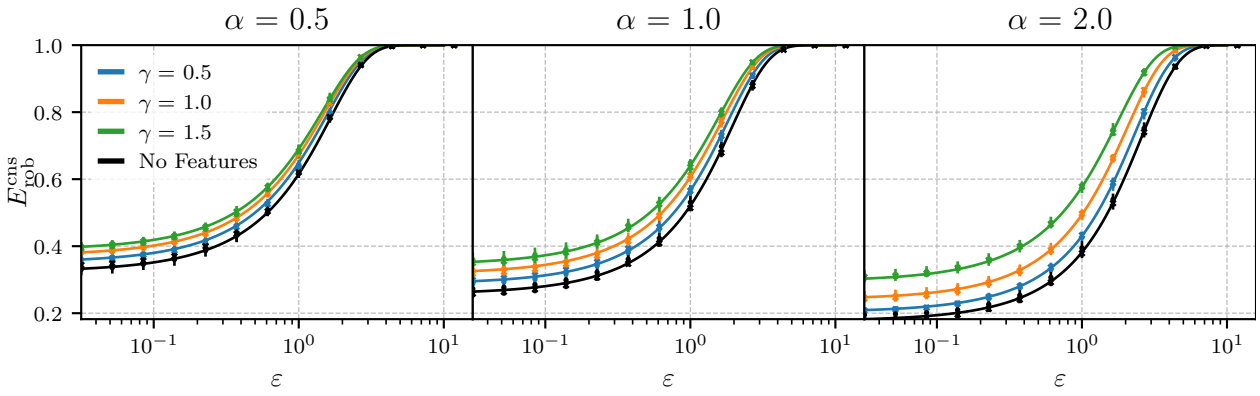

Figure 6: Behavior of $E_{\text{rob}}^{\text{cns}}$ for the Gaussian Features case as a function of the attack strength perturbation. The performances are for model trained as per eqs. (21) and (22) and with the α and γ specified in the figure. The error bars refer to 10 repetitions of the experiments for $d = 1024$. The metrics consider the $\hat{\boldsymbol{w}}$ trained with $\lambda = 10^{-3}$, $r = 0.0$ and $s = \infty$.

Appendix

A Additional Experiments and Figures Setting

A.1 Setting of the Figures in the Main Text

We note that the optimization over the hyperparameters r and λ are performed with the use of the theory. In the asymptotic limit the self consistent equation give a deterministic function of the model's parameter. With gradient free optimization techniques[1] we find the minimal values.

Figure 3 ***(Left)*** The curves are realized for a $\boldsymbol{w}$ obtained from standard training, i.e. minimization of eq. (22) with $r = 0$, $\lambda = 10^{-3}$ and $r = 2$. We have that the number of data is fixed at $\alpha = 1.0$. The points are produced as 10 different realizations with $d = 500$ fixed.

Figure 3 ***(Right)*** Here we show the performances of different types of attack metrics, either L_∞ or L_2 constrained. We have that in both cases the errors correspond to optimally tuned robust estimation, r and λ chosen to have minimum errors. We have that the regularization geometry is $r = 2$ and that the geometry in adversarial training is $s = 2$. The points are produced as 10 different realizations with $d = 500$ fixed.

Figure 4 In this case we consider vanishing regularization non robust trained empirical risk minimization of eqs. (21) and (22). We consider the values of α and γ as per the figure. The value of $\lambda = 10^{-3}$. The points are produced as 10 different realizations with $d = 500$ fixed.

Figure 5 ***(Left,Right)*** We consider optimally tuned robust empirical risk minimization with $s = \infty$ and $r = 2$. The points are produced as 10 different realizations with $d = 500$ fixed and the values of n, p scaled accordingly.

A.2 Additional Experiments

To test the robustness of our findings with respect to the choice of the feature matrix procedure chosen in Assumption 4.2, specifically the generation of the input data $\boldsymbol{x}$ as a function of the latent variable $\boldsymbol{z}$ we consider a different kind of latent space model.

Another model used to characterize overparameterization is the hidden manifold model [12, 18] where the latent space covariates are still drawn from a gaussian $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \mathrm{Id}_d)$ but the feature space covariates are a linear transformation $\boldsymbol{x} = \mathrm{F}\,\boldsymbol{z}$ where $\mathrm{F}_{ij} \sim \mathcal{N}(0, 1)$ each component independently.

[1] Specifically we use `np.minimize` with `Nelder-Mead` algorithm.

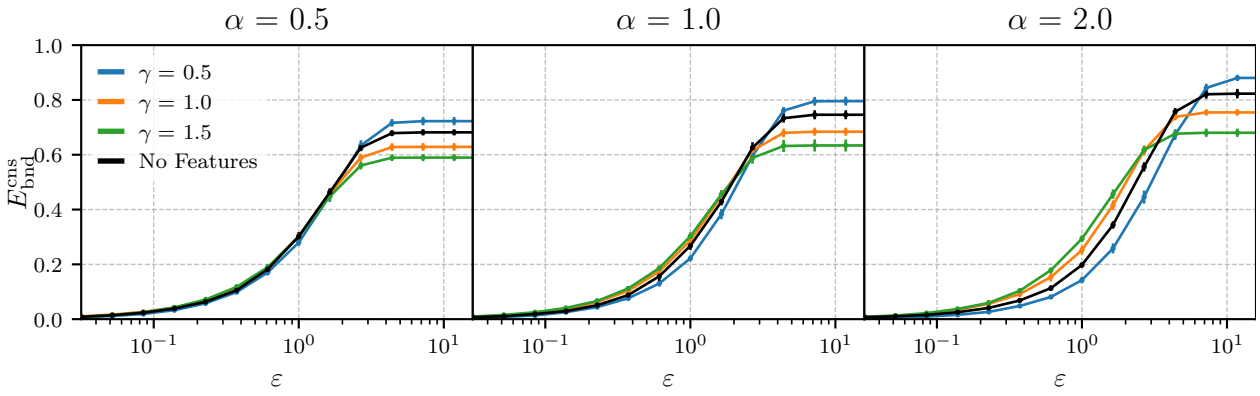

Figure 7: Behavior of $E^{\text{cns}}_{\text{bnd}}$ for the Gaussian Features case as a function of the attack strength perturbation. The performances are for model trained as per eqs. (21) and (22) and with the α and γ. The error bars refer to 10 repetitions of the experiments for $d = 1024$. The metrics consider the $\hat{\boldsymbol{w}}$ trained with $\lambda = 10^{-3}$, $r = 0.0$ and $s = \inf$.

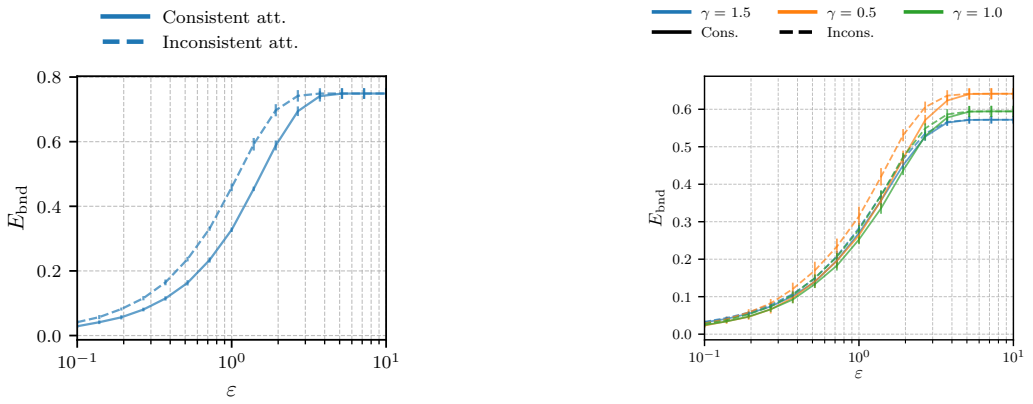

Figure 8: Comparison of consistent boundary error and inconsistent boundary error as a function of the attack strength. *(Left)* Case of the well-specified model presented in Section 3. The error bars refer to 10 repetitions of the experiments for $d = 500$. The metrics consider the $\hat{\boldsymbol{w}}$ trained with $\alpha = 1.0$, $\lambda = 10^{-2}$, $r = 0.0$ and $s = 2$. *(Right)* Case of the latent space model presented in Section 4. The error bars refer to 10 repetitions of the experiments for $d = 512$. The metrics consider the $\hat{\boldsymbol{w}}$ trained with $\alpha = 1.0$, $\lambda = 10^{-2}$, $r = 0.0$ and $s = 2$.

We explore the behavior of the error metrics defined in eqs. (33) and (34) in Figures 6 and 7. We see that the behavior is similar to the one of the model defined in the main text. The black line presented in the same figure is the performances of the well specified model in the main text.

Crucially we see also in this case the metric $E^{\text{cns}}_{\text{rob}}$ equals the value of the clean generalization error in the $\varepsilon_g \to 0^+$ limit and it reaches one in the $\varepsilon_g \to \infty$ limit. We have that $E^{\text{cns}}_{\text{bnd}}$ is zero in the $\varepsilon_g \to 0^+$ limit.

We additionally compare the consistent and inconsistent formulation of the boundary error as a function of the attack strength in Figure 8. The inconsistent boundary error is defined from the same formula as eq. (3) with the removal of the consistent condition $f_\star(\boldsymbol{x}) = f_\star(\boldsymbol{x} + \boldsymbol{\delta})$. We have that also in this case consistent attacks produce a milder increase in the boundary error but the qualitative behavior is the same.

B Proof of the Results in the Main Text

In this section, we provide rigorous proofs for the theoretical results presented in the main paper, focusing on Theorem 4.1.

Central to our analysis is the Convex Gaussian MinMax Theorem (CGMT), a fundamental tool that bridges complex high-dimensional optimization problems with simpler low-dimensional counterparts. The CGMT enables

us to transform our challenging primary optimization problem into a more tractable auxiliary problem, ultimately leading to the self-consistent equations presented in Equations (37), (39) and (40).

We begin by stating the CGMT in its general form.

Theorem B.1 (CGMT [14, 72]). *Let $\boldsymbol{G} \in \mathbb{R}^{m\times n}$ be an i.i.d. standard normal matrix and $\mathbf{g} \in \mathbb{R}^m$, $\mathbf{h} \in \mathbb{R}^n$ two i.i.d. standard normal vectors independent of each other. For compact sets $\mathcal{S}_{\boldsymbol{w}} \subset \mathbb{R}^n$ and $\mathcal{S}_{\boldsymbol{u}} \subset \mathbb{R}^n$, consider the following optimization problems with continuous function ψ on $\mathcal{S}_{\boldsymbol{w}} \times \mathcal{S}_{\boldsymbol{u}}$:*

$$\mathbf{C}(\boldsymbol{G}) := \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \boldsymbol{u}^\top \boldsymbol{G}\boldsymbol{w} + \psi(\boldsymbol{w},\boldsymbol{u}) \tag{46}$$

$$\mathcal{C}(\mathbf{g},\mathbf{h}) := \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \|\boldsymbol{w}\|_2 \mathbf{g}^\top \boldsymbol{u} + \|\boldsymbol{u}\|_2 \mathbf{h}^\top \boldsymbol{w} + \psi(\boldsymbol{w},\boldsymbol{u}) \tag{47}$$

The following statements hold:

1. *For all $c \in \mathbb{R}$: $\mathbb{P}(\mathbf{C}(\boldsymbol{G}) < c) \le 2\mathbb{P}(\mathcal{C}(\mathbf{g},\mathbf{h}) \le c)$*
2. *When $\mathcal{S}_{\boldsymbol{w}}$ and $\mathcal{S}_{\boldsymbol{u}}$ are convex sets and ψ is convex-concave on $\mathcal{S}_{\boldsymbol{w}} \times \mathcal{S}_{\boldsymbol{u}}$, for all $c \in \mathbb{R}$: $\mathbb{P}(\mathbf{C}(\boldsymbol{G}) > c) \le 2\mathbb{P}(\mathcal{C}(\mathbf{g},\mathbf{h}) \ge c)$*
3. *Consequently, for all $\mu \in \mathbb{R}$, $t > 0$: $\mathbb{P}(|\mathbf{C}(\boldsymbol{G}) - \mu| > t) \le 2\mathbb{P}(|\mathcal{C}(\mathbf{g},\mathbf{h}) - \mu| \ge t)$*

We will utilize a specialized version of the CGMT developed by [20] for generalized linear models.

B.1 Mathematical Preliminaries

Our analysis relies heavily on Moreau envelopes and proximal operators from convex analysis. These concepts have become essential tools in the asymptotic analysis of high-dimensional convex problems [73, 74]. We provide key definitions below.

Definition 4 (Moreau Envelope). For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, its Moreau envelope is defined as:

$$\mathcal{M}_{Vf(\cdot)}(\boldsymbol{\omega}) = \min_{\boldsymbol{x}} \left[\frac{1}{2V}\|\boldsymbol{x}-\boldsymbol{\omega}\|_2^2 + f(\boldsymbol{x}) \right] \tag{48}$$

where $\mathcal{M}_{Vf(\cdot)} : \mathbb{R}^n \to \mathbb{R}$.

Definition 5 (Proximal Operator). For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, its Proximal operator is defined as:

$$\mathcal{P}_{Vf(\cdot)}(\boldsymbol{\omega}) = \arg\min_{\boldsymbol{x}} \left[\frac{1}{2V}\|\boldsymbol{x}-\boldsymbol{\omega}\|_2^2 + f(\boldsymbol{x}) \right] \tag{49}$$

where $\mathcal{P}_{Vf(\cdot)} : \mathbb{R}^n \to \mathbb{R}^n$.

Theorem B.2 (Gradient of Moreau Envelope [8], Lemma D1). *For a convex function $f : \mathbb{R}^n \to \mathbb{R}$ with Moreau envelope $\mathcal{M}_{Vf(\cdot)}$ and Proximal operator $\mathcal{P}_{Vf(\cdot)}$:*

$$\boldsymbol{\nabla}_{\boldsymbol{\omega}} \mathcal{M}_{Vf(\cdot)}(\boldsymbol{\omega}) = \frac{1}{V}\left(\boldsymbol{\omega} - \mathcal{P}_{Vf(\cdot)}(\boldsymbol{\omega})\right) \tag{50}$$

Additionally, we will use these important properties:

$$\mathcal{M}_{Vf(\cdot+\boldsymbol{u})}(\boldsymbol{\omega}) = \mathcal{M}_{Vf(\cdot)}(\boldsymbol{\omega}+\boldsymbol{u}), \quad \mathcal{P}_{Vf(\cdot+\boldsymbol{u})}(\boldsymbol{\omega}) = \boldsymbol{u} + \mathcal{P}_{Vf(\cdot)}(\boldsymbol{\omega}+\boldsymbol{u}) \tag{51}$$

which follow directly from a change of variables in the minimization.

Definition 6 (Dual of a Number). We define the the dual of a number $a \ge 0$ as being $a^\star$ as the only number such that $1/a + 1/a^\star = 1$.

B.2 Assumptions and Preliminary Discussion

We restate here all the assumptions that we make for the problem.

Assumption B.1 (Estimation from the dataset)**.** Given a dataset $\mathcal{D}$ made of n pairs of input outputs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ we estimate the vector $\hat{\boldsymbol{w}}$ as being

$$\hat{\boldsymbol{w}} \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^n \max_{\|\boldsymbol{v}_i\|_s \le r} \ell\left(y_i \frac{\boldsymbol{w}^\top(\boldsymbol{x}_i + \boldsymbol{v}_i)}{\sqrt{d}}\right) + \lambda\|\boldsymbol{w}\|_2^2\,, \tag{52}$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is a convex non-increasing function and where the second term is a convex regularization function whose strength can be tuned with $\lambda \in [0, \infty)$.

Assumption B.2 (Data Distribution)**.** We assume that data $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \{-1, +1\}$ is drawn from a latent variable model with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, {}^1\!/_d \operatorname{Id}_d)$ and ground-truth linear classifier $f_{\boldsymbol{w}_\star}(\boldsymbol{z}) = \varphi(\langle \boldsymbol{w}_\star, \boldsymbol{z}\rangle)$ with $\boldsymbol{w}_\star \in \mathbb{S}^{d-1}(\sqrt{d})$. The observed features $\boldsymbol{x} \in \mathbb{R}^p$ are generated as $\boldsymbol{x} = \mathrm{F}\,\boldsymbol{z} + \boldsymbol{u}$ with $\boldsymbol{u} \sim \mathcal{N}(0, \operatorname{Id}_p)$ independent of the other quantities and

$$\mathrm{F} = \begin{cases} \begin{bmatrix} \sqrt{\frac{p}{d}}\operatorname{Id}_d \\ \boldsymbol{0}_{(p-d)\times d} \end{bmatrix} & \text{if } p \ge d \\ \begin{bmatrix} \operatorname{Id}_p & \boldsymbol{0}_{p\times(d-p)} \end{bmatrix} & \text{if } p < d \end{cases}. \tag{53}$$

Assumption B.3 (High-Dimensional Limit)**.** We consider the proportional high-dimensional regime where both the number of training data and input dimension $n, d, p \to \infty$ at a fixed ratio $\alpha := {}^n\!/_d$ and $\psi = {}^p\!/_n$.

This setting considers most of the losses used in machine learning setups for binary classification, *e.g.* logistic, hinge, exponential losses. We additionally remark that with the given choice of regularization the whole cost function is coercive.

Assumption B.4 (Scaling of Adversarial Norm Constraint)**.** For a given perturbation geometry $\boldsymbol{\delta} \in B_q(r)$ with $q > 1$, we assume that $r = O_d(d^{1/q^\star + 1/2})$ as $d \to \infty$, where $q^\star$ is the dual. We define the rescaled radius as $\widetilde{\varepsilon}_t = {}^{\varepsilon}\!/_{d^{1/q^\star+1/2}}$.

B.3 Problem Simplification

Recall that we start from the following optimization problem:

$$\Phi_d = \min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^n \max_{\|\boldsymbol{v}_i\|_s \le r} \ell\left(y_i \frac{\boldsymbol{w}^\top(\boldsymbol{x}_i + \boldsymbol{v}_i)}{\sqrt{d}}\right) + \lambda\|\boldsymbol{w}\|_2^2\,. \tag{54}$$

The non-increasing property of ℓ allows us to simplify the inner maximization, leading to an equivalent formulation

$$\Phi_d = \min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{i=1}^n \ell\left(y_i \frac{\boldsymbol{w}^\top \boldsymbol{x}_i}{\sqrt{d}} - \frac{r}{\sqrt{d}}\|\boldsymbol{w}\|_{s^\star}\right) + \lambda\|\boldsymbol{w}\|_2^2\,. \tag{55}$$

To facilitate our analysis, we introduce auxiliary variables $P = \|\boldsymbol{w}\|_\star^{p^\star}/d$ and $\hat{P}$ (the Lagrange parameter relative to this variable), which allow us to decouple the norm constraints. This leads to a min-max formulation

$$\Phi_d = \min_{\boldsymbol{w}\in\mathbb{R}^d, P} \max_{\hat{P}} \sum_{i=1}^n \ell\left(y_i \frac{\boldsymbol{w}^\top \boldsymbol{x}_i}{\sqrt{d}} - \frac{r}{\sqrt[s^\star]{d}}\sqrt[s^\star]{P}\right) + \lambda\|\boldsymbol{w}\|_2^2 + \hat{P}\|\boldsymbol{w}\|_{s^\star}^{s^\star} - dP\hat{P}, \tag{56}$$

where we switched the value of r for its value without the scaling in d. This reformulation is what will allow us to apply the CGMT in subsequent steps.

It's worth noting the significance of the scaling for r as detailed in Assumption B.4. In the high-dimensional limit $d \to \infty$, it's essential that all terms in Φ_d exhibit the same scaling with respect to d. This careful scaling ensures that our asymptotic analysis remains well-behaved and meaningful in the high-dimensional regime.

B.4 Scalarization and Application of CGMT

To facilitate our analysis, we further introduce effective regularization and loss functions, $\widetilde{r}$ and $\widetilde{\ell}$, respectively. These functions are defined as

$$\widetilde{\ell}(\boldsymbol{y},\boldsymbol{z}) = \sum_{i=1}^{n} \ell\left(y_i \boldsymbol{z}_i - \frac{r}{\sqrt[s^\star]{d}} \sqrt[s^\star]{P}\right), \quad \widetilde{r}(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2 + \hat{P}\|\boldsymbol{w}\|_{s^\star}^{s^\star}. \tag{57}$$

A crucial step in our analysis involves inverting the order of the min-max optimization. We can justify this operation by considering the minimization with respect to $\boldsymbol{w} \in \mathbb{R}^d$ at fixed values of $\hat{P}$ and P. This reordering is valid due to the convexity of our original problem. Specifically, the objective function is convex in $\boldsymbol{w}$ and concave in $\hat{P}$ and P, and the constraint sets are convex. Under these conditions, we apply Sion's minimax theorem, which guarantees the existence of a saddle point and allows us to interchange the order of minimization and maximization without affecting the optimal value.

We additionally notice that the data distribution defined in Assumption B.2 lies under the same framework as the one presented in [20]. Specifically can be seen as the case treated in Section 3.1 with the choice of non linearity just adding Gaussian noise.

This reformulation enables us to directly apply [20, Lemma 11]. This lemma represents a meticulous application of Theorem B.1 to scenarios involving non-separable convex regularization and loss functions. The result is a lower-dimensional equivalent of our original high-dimensional minimization problem that represent the limiting behavior of the solution of the high-dimensional problem.

Consequently, our analysis now focuses on a low-dimensional functional, which takes the form

$$\widetilde{\Phi} = \min_{P,m,\eta,\tau_1} \max_{\hat{P},\kappa,\tau_2,\nu} \left[\frac{\kappa\tau_1}{2} - \alpha\mathcal{L}_\ell - \frac{\eta}{2\tau_2}\left(\nu^2\rho + \kappa^2\right) - \frac{\eta\tau_2}{2} - \mathcal{L}_r + m\nu - P\hat{P}\right] \tag{58}$$

where we have restored the min max order of the problem.

In this expression, $\boldsymbol{g}$ and $\boldsymbol{h}$ are independent Gaussian vectors with i.i.d. standard normal components. The terms $\mathcal{L}_\ell$ and $\mathcal{L}_r$ represent the scaled averages of Moreau Envelopes (eq. (48))

$$\mathcal{L}_\ell = \frac{1}{n}\mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\kappa}\widetilde{\ell}(\boldsymbol{y},\cdot)}\left(\frac{m}{\sqrt{\rho}}\boldsymbol{s} + \eta\boldsymbol{h}\right)\right] \tag{59}$$

$$\mathcal{L}_r = \frac{1}{d}\mathbb{E}\left[\mathcal{M}_{\frac{\eta}{\tau_2}\widetilde{r}(\cdot)}\left(\frac{\eta}{\tau_2}(\kappa\boldsymbol{g} + \nu\boldsymbol{w}_\star)\right)\right] \tag{60}$$

The extremization problem in eq. (58) is related to the original optimization problem in eq. (54) as it can be thought as the leading part in the limit $n, d \to \infty$.

This dimensional reduction is the step that allows us to study the asymptotic properties of our original high-dimensional problem through a more tractable low-dimensional optimization and thus have in the end a low dimensional set of equations to study.

It's important to note that the optimization problem $\widetilde{\Phi}$ is still implicitly defined in terms of the dimension d and, consequently, as a function of the sample size n. We introduce two variables

$$\boldsymbol{w}_{\text{eq}} = \mathcal{P}_{\frac{\eta^*}{\tau_2^*}\widetilde{r}(.)}\left(\frac{\eta^*}{\tau_2^*}\left(\nu^*\mathbf{t} + \kappa^*\mathbf{g}\right)\right), \quad \boldsymbol{z}_{\text{eq}} = \mathcal{P}_{\frac{\tau_1^*}{\kappa^*}\widetilde{\ell}(,,\mathbf{y})}\left(\frac{m^*}{\sqrt{\rho}}\mathbf{s} + \eta^*\mathbf{h}\right) \tag{61}$$

where $(\eta^\star, \tau_2^\star, P^\star, \hat{P}^\star, \kappa^\star, \nu^\star, m^\star, \tau_1^\star)$ are the extremizer points of $\widetilde{\Phi}$.

Building upon [20, Theorem 5], we can establish a convergence result. Let $\hat{\boldsymbol{w}}$ be an optimal solution of the problem defined in eq. (54), and let $\hat{\boldsymbol{z}} = \frac{1}{\sqrt{d}}\boldsymbol{X}\hat{\boldsymbol{w}}$. For any Lipschitz function $\varphi_1 : \mathbb{R}^d \to \mathbb{R}$, and any separable, pseudo-Lipschitz function $\varphi_2 : \mathbb{R}^n \to \mathbb{R}$, there exist constants $\epsilon, C, c > 0$ such that

$$\mathbb{P}\left(\left|\phi_1\left(\frac{\hat{\mathbf{w}}}{\sqrt{d}}\right) - \mathbb{E}\left[\phi_1\left(\frac{\boldsymbol{w}_{\text{eq}}}{\sqrt{d}}\right)\right]\right| \geq \epsilon\right) \leq \frac{C}{\epsilon^2}e^{-cn\epsilon^4} \tag{62}$$

$$\mathbb{P}\left(\left|\phi_2\left(\frac{\hat{\mathbf{z}}}{\sqrt{n}}\right) - \mathbb{E}\left[\phi_2\left(\frac{\boldsymbol{z}_{\text{eq}}}{\sqrt{n}}\right)\right]\right| \geq \epsilon\right) \leq \frac{C}{\epsilon^2}e^{-cn\epsilon^4} \tag{63}$$

It demonstrates that the limiting values of any function depending on $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{z}}$ can be computed by taking the expectation of the same function evaluated at $\boldsymbol{w}_{\rm eq}$ or $\boldsymbol{z}_{\rm eq}$, respectively. This convergence property allows us to translate results from our low-dimensional proxy problem back to the original high-dimensional setting with high probability.

B.5 Derivation of Saddle Point equations

We now want to show that extremizing the values of $m, \eta, \tau_1, P, \hat{P}, \nu, \tau_2, \kappa$ lead to the optimal value $\widetilde{\Phi}$ of eq. (58). We are going to directly derive the saddle point equations and then argue that in the high-dimensional limit they become exactly the ones reported in the main text.

We obtain the first set of derivatives that depend only on the loss function and the channel part by taking the derivatives with respect to m, η, τ_1, P to obatin

$$\begin{aligned}
&\frac{\partial}{\partial m} : \nu = \alpha \frac{\kappa}{n\tau_1} \mathbb{E}\left[\left(\frac{m}{\eta\rho}\mathbf{h} - \frac{\mathbf{s}}{\sqrt{\rho}}\right)^{\top} \mathcal{P}_{\frac{\tau_1}{\kappa}\widetilde{\ell}(\cdot,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right] \\
&\frac{\partial}{\partial \eta} : \tau_2 = \alpha\frac{\kappa}{\tau_1}\eta - \frac{\kappa\alpha}{\tau_1 n}\mathbb{E}\left[\mathbf{h}^{\top}\mathcal{P}_{\frac{\tau_1}{\kappa}\widetilde{\ell}(\cdot,\mathbf{y})}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right] \\
&\frac{\partial}{\partial \tau_1} : \frac{\tau_1^2}{2} = \frac{1}{2}\alpha\frac{1}{n}\mathbb{E}\left[\left\|\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h} - \mathcal{P}_{\frac{\tau_1}{\kappa}\widetilde{\ell}(\cdot,y)}\left(\frac{m}{\sqrt{\rho}}\mathbf{s} + \eta\mathbf{h}\right)\right\|_2^2\right] \\
&\frac{\partial}{\partial P} : \hat{P} = \frac{\alpha}{n}\partial_P \mathbb{E}\left[\mathcal{M}_{\frac{\tau_1}{\kappa}\widetilde{\ell}(\boldsymbol{y},\cdot)}\left(\frac{m}{\sqrt{\rho}}\boldsymbol{s} + \eta\boldsymbol{h}\right)\right]
\end{aligned} \tag{64}$$

By taking the derivatives with respect to the remaining variables $\kappa, \nu, \tau_2, \hat{P}$ we obtain a set of equations depending on regularization and prior over the teacher weights

$$\begin{aligned}
&\frac{\partial}{\partial \kappa} : \tau_1 = \frac{1}{d}\mathbb{E}\left[\mathbf{g}^{\top}\mathcal{P}_{\frac{\eta}{\tau_2}\widetilde{r}(\cdot)}\left(\frac{\eta}{\tau_2}\left(\nu\boldsymbol{w}_\star + \kappa\mathbf{g}\right)\right)\right] \\
&\frac{\partial}{\partial \nu} : m = \frac{1}{d}\mathbb{E}\left[\boldsymbol{w}_\star^{\top}\mathcal{P}_{\frac{\eta}{\tau_2}\widetilde{r}(\cdot)}\left(\frac{\eta}{\tau_2}\left(\nu\boldsymbol{w}_\star + \kappa\mathbf{g}\right)\right)\right] \\
&\frac{\partial}{\partial \tau_2} : \frac{1}{2d}\frac{\tau_2}{\eta}\mathbb{E}\left[\left\|\frac{\eta}{\tau_2}\left(\nu\boldsymbol{w}_\star + \kappa\mathbf{g}\right) - \mathcal{P}_{\frac{\eta}{\tau_2}\widetilde{r}(\cdot)}\left(\frac{\eta}{\tau_2}\left(\nu\boldsymbol{w}_\star + \kappa\mathbf{g}\right)\right)\right\|_2^2\right] = \frac{\eta}{2\tau_2}\left(\nu^2\rho + \kappa^2\right) - m\nu - \kappa\tau_1 + \frac{\eta\tau_2}{2} + \frac{\tau_2}{2\eta}\frac{m^2}{\rho} \\
&\frac{\partial}{\partial \hat{P}} : P = \frac{1}{d}\partial_{\hat{P}}\mathbb{E}\left[\mathcal{M}_{\frac{\eta}{\tau_2}\widetilde{r}(\cdot)}\left(\frac{\eta}{\tau_2}(\kappa\boldsymbol{g} + \nu\boldsymbol{w}_\star)\right)\right]
\end{aligned} \tag{65}$$

The rewriting of these equations in the desired form in Theorem 4.1 follows from the same considerations as in [20, Appendix C.2], specifically two changes of variables and a integration by parts.

To perform this rewriting the first ingredient we need is the following change of variables

$$\begin{aligned}
&m \leftarrow m\,, \qquad q \leftarrow \eta^2 + \frac{m^2}{\rho}\,, \qquad V \leftarrow \frac{\tau_1}{\kappa}\,, \qquad P \leftarrow P\,, \\
&\hat{V} \leftarrow \frac{\tau_2}{\eta}\,, \qquad \hat{q} \leftarrow \kappa^2\,, \qquad \hat{m} \leftarrow \nu\,, \qquad \hat{P} \leftarrow \hat{P}\,.
\end{aligned} \tag{66}$$

ant the use of Isserlis' theorem [75] to simplify the expectation where Gaussian $\boldsymbol{g}, \boldsymbol{h}$ vectors are present.

B.5.1 Rewriting of the Saddle Point Equations

To obtain specifically the form implied in the main text we introduce

$$\mathcal{Z}_0(y, \omega, V) = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2}\delta\left(y - f^0(x)\right)\,. \tag{67}$$

The function $\mathcal{Z}_0$ can be interpreted as a partition function of the conditional distribution $\mathbb{P}_{\text{out}}$ and contains all of the information about the label generating process.

In the case of s_p norms, we can leverage the separable nature of the regularization to simplify our equations. The key insight here is that the proximal operator of a separable regularization is itself separable. This property allows us to treat each dimension independently, leading to a significant simplification of our high-dimensional problem.

First, due to the separability, all terms depending on the proximal of either $\widetilde{\ell}$ or $\widetilde{r}$ simplify the n or d at the denominator. This cancellation is crucial as it eliminates the explicit dependence on the problem dimension, allowing us to derive dimension-independent equations.

Next, we introduce

$$\mathcal{Z}_{\text{w}}(\varrho, \Lambda) = \int \mathrm{d}w \, e^{-\frac{1}{2}w^2} e^{-\frac{\Lambda}{2}w^2 + \varrho w}, \tag{68}$$

which, in turn, leads in the form shown in the main text.

We note additionally that to obtain the specific form of the saddle point equations presented in Theorem 4.1 one needs to apply the specific form for the proximal operator of the Elastic-Net, specifically that

$$\mathcal{P}_{V(\lambda_1|\cdot| + \lambda_2|\cdot|^2)}(v) = \frac{\text{sign}(v) \cdot \max(|v| - \lambda_1 V, 0)}{2V\lambda_2 + 1}\,. \tag{69}$$

B.6 Preliminaries Calculations For The Error Functions

We start by proving the following lemma that will be useful in the following. The following lemma will be specific for the case of

Lemma 1 (Concentration of adversarial perturbations)**.** *Given a decreasing function g, $y \in \{\pm 1\}$. For $\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\delta} \in \mathbb{R}^d$ we have that*

$$\max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_q \leq \varepsilon, \langle \boldsymbol{w}_\star, \boldsymbol{\delta}\rangle = 0} g\left(y\frac{\langle \boldsymbol{w}, \boldsymbol{x} + \boldsymbol{\delta}\rangle}{\sqrt{d}}\right) = \sup_{\kappa \in \mathbb{R}} g\left(y\frac{\langle \boldsymbol{w}, \boldsymbol{x}\rangle}{\sqrt{d}} - \frac{\varepsilon}{\sqrt{d}}\|\boldsymbol{w} - \kappa \boldsymbol{w}_\star\|_{q^\star}\right) \tag{70}$$

Lemma 1. Since $g : \mathbb{R} \to \mathbb{R}$ in eq. (70) is a decreasing, non necessary continuous, function, one simply minimize the argument of the function and then pass it through the original function. We can analyze the following

$$\min_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_q \leq \varepsilon, \langle \boldsymbol{w}_\star, \boldsymbol{\delta}\rangle = 0} y\,\langle \boldsymbol{w}, \boldsymbol{x} + \boldsymbol{\delta}\rangle = y\,\langle \boldsymbol{w}, \boldsymbol{x}\rangle + \min_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_q \leq \varepsilon, \langle \boldsymbol{w}_\star, \boldsymbol{\delta}\rangle = 0} y\,\langle \boldsymbol{w}, \boldsymbol{\delta}\rangle \tag{71}$$

we thus focus now on the second part only as the first part can be considered afterwards and separately from the minimization. Since we consider a binary classification problem $y \in \{+1, -1\}$ we can perform the change of variables $\boldsymbol{\delta} \to y\boldsymbol{\delta}$ and we see that the constraints do not depend on y. We can write

$$\min_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_q \leq \varepsilon, \langle \boldsymbol{w}_\star, \boldsymbol{\delta}\rangle = 0} \langle \boldsymbol{w}, \boldsymbol{\delta}\rangle = \min_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_q \leq \varepsilon} \sup_{\kappa} \langle \boldsymbol{w}, \boldsymbol{\delta}\rangle + \kappa \langle \boldsymbol{w}_\star, \boldsymbol{\delta}\rangle = \min_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_q \leq \varepsilon} \sup_{\kappa} \langle \boldsymbol{w} + \kappa \boldsymbol{w}_\star, \boldsymbol{\delta}\rangle \tag{72}$$

Now we want to use the fact that strong duality holds for the primal and dual problem then by interchanging the order we obtain

$$\sup_{\kappa} \min_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_q \leq \varepsilon} \langle \boldsymbol{w} + \kappa \boldsymbol{w}_\star, \boldsymbol{\delta}\rangle = \sup_{\kappa} -\varepsilon \|\boldsymbol{w} + \kappa \boldsymbol{w}_\star\|_{q^\star} = -\varepsilon \inf_{\kappa} \|\boldsymbol{w} + \kappa \boldsymbol{w}_\star\|_{q^\star} \tag{73}$$

where we have used the definition of the dual norm and $1/q + 1/q^\star = 1$. By reintroducing the factors in front we have

$$-\frac{\varepsilon}{\sqrt{d}} \inf_{\kappa} \|\boldsymbol{w} + \kappa \boldsymbol{w}_\star\|_{q^\star} \tag{74}$$

and thus the form we are interested in. □

The previous Lemma is interesting as it is the basis of the proof of Theorems 3.1 and 4.2. Additionally it tells us that given the distribution of both $\boldsymbol{w}_\star$ and $\boldsymbol{w}$ one can evaluate the limiting form for the term appearing as a function of κ and then take the extremization over κ.

Now we would like to evaluate the limiting value of the previous equation under the condition that it is the norm of a Gaussian vector with a specific covariance. Firstly we acknowledge that any L_p norm of a Gaussian vector

concentrates. The proof of this is an application of [76, Theorem 5.5] applied to the Lipschitz function $\|M\cdot\|_p$ with M being the square root of the covariance or the Gaussian vector.

Specifically for the case considered the result can be found stated as [77, Corollary 1]. Specifically we have that by considering a scaling for ε such that $\varepsilon\sqrt[q^\star]{d}/\sqrt{d} = \widetilde{\varepsilon}$. If we have that the two variables are correlated element to element as $\boldsymbol{w}_i = m(\boldsymbol{w}_\star)_i + q\xi$ where $\boldsymbol{w}_\star, \xi \sim \mathcal{N}(\mathbf{0}, \mathrm{Id}_d)$ independently we have that

$$\frac{\varepsilon}{\sqrt{d}} \inf_{\kappa} \|\widetilde{\boldsymbol{w}} + \kappa \boldsymbol{w}_\star\|_{q^\star} \xrightarrow[d\to\infty]{} \widetilde{\varepsilon} \inf_{\kappa} \frac{\sqrt{2}}{\pi^{(2q^\star)^{-1}}} \sqrt{\left(\frac{m}{\rho} + \kappa\right)^2 + q - \frac{m^2}{\rho}} \sqrt[q^\star]{\Gamma\left(\frac{q^\star + 1}{2}\right)} \tag{75}$$

the previous equation is always minimized for $\kappa = -m/\rho$ and thus it leads to

$$\widetilde{\varepsilon}\sqrt{2}\sqrt{q - \frac{m^2}{\rho}} \sqrt[q^\star]{\frac{\Gamma\left((q^\star + 1)/2\right)}{\sqrt{\pi}}}\,. \tag{76}$$

This is the case of Theorem 3.1.

To study the limiting value of eq. (74) for the trained predictor one should know the limiting joint distribution of $\hat{\boldsymbol{w}}$ and $\boldsymbol{w}_\star$. For the case of Theorem 4.2 one can apply [20, Lemma 5], which is a more complete version of eq. (61), to characterize the probability distribution of the trained vector and obtain the form in Theorem 4.2.

B.7 Proper Error Metrics

Once one has that the perturbation due to the adversarial attack concentrates to some limitng value one can also find the limiting distribution of the metrics eqs. (2) and (3). The results can be derived with the local fields method [66]. We are expressing it for the more difficult case of the latent space model. Specifically we have that

$$\begin{pmatrix} \frac{\langle \boldsymbol{w}_\star, \boldsymbol{z}\rangle}{\sqrt{d}} \\ \frac{\langle \hat{\boldsymbol{w}}, \boldsymbol{x}\rangle}{\sqrt{p}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & m \\ m & q \end{pmatrix}\right) \tag{77}$$

where the values of m, q are the ones that can be found from teh solution of the set of self consistent equations in Theorem 4.1.

C Statistical Physics Derivation of the Main Result

Here we present an alternative derivation of our main result using statistical physics methods, specifically the replica technique [78]. We begin by formulating a Gibbs measure from the empirical risk with an inverse temperature parameter β, then analyze the zero-temperature limit to characterize the optimal solution space.

C.1 Gibbs Measure Formulation

We define a Gibbs probability measure over the weight space $\boldsymbol{w} \in \mathbb{R}^d$. This measure is constructed so that weights minimizing the empirical risk have the highest probability. By taking the zero temperature limit ($\beta \to \infty$), we can focus exclusively on these optimal solutions.

The Gibbs measure is defined as:

$$\mu_\beta(\mathrm{d}\boldsymbol{w}) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\left[\sum_{\mu=1}^n g\left(y^\mu, \boldsymbol{w}^\top \boldsymbol{x}_\mu, \boldsymbol{w}, \varepsilon_t\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right]} \mathrm{d}\boldsymbol{w} \tag{78}$$

$$= \frac{1}{\mathcal{Z}_\beta} \underbrace{\prod_{\mu=1}^n e^{-\beta g\left(y^\mu, \boldsymbol{w}^\top \boldsymbol{x}_\mu, \boldsymbol{w}, \varepsilon_t\right)}}_{P_{g,\varepsilon_t}} \underbrace{e^{-\frac{\beta\lambda}{2}\|\boldsymbol{w}\|_2^2}}_{P_w} \mathrm{d}\boldsymbol{w} \tag{79}$$

Here, P_{g,ε_t} represents the probability distribution associated with the channel, while P_w denotes the prior probability distribution on weights.

The partition function $\mathcal{Z}_\beta$ normalizes this measure:

$$\mathcal{Z}_\beta = \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{w} e^{-\frac{\beta\lambda}{2}\|\boldsymbol{w}\|_2^2} \prod_{\mu=1}^{n} e^{-\beta g\left(y^\mu, \boldsymbol{w}^\top \boldsymbol{x}_\mu, \boldsymbol{w}, \varepsilon_t\right)} \tag{80}$$

As $\beta \to \infty$, the measure concentrates around solutions that minimize the empirical risk. The free energy density, our primary quantity of interest, is given by:

$$\beta f_\beta = -\lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_\beta \tag{81}$$

C.2 Replica Technique Application

To compute the average of the free energy, we employ the replica trick:

$$\lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_\beta = \lim_{r\to 0} \lim_{d\to\infty} \frac{1}{d} \frac{\partial_r \mathbb{E}_{\mathcal{D}} \mathcal{Z}^r}{1} \tag{82}$$

This approach involves three key limits: 1. The zero temperature limit ($\beta \to \infty$) to identify the global minimum of our optimization problem 2. The thermodynamic limit of large dimension ($d \to \infty$) with fixed sampling ratio 3. The replica limit ($r \to 0$) enabling the logarithm computation

We begin with the replicated partition function, noting our case includes a dependency on ε_t in the output probability:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}} \mathcal{Z}_\beta^r &= \prod_{\mu=1}^{n} \mathbb{E}_{\boldsymbol{x}_\mu} \prod_{a=1}^{r} \int_{\mathbb{R}^p} P_w\left(\mathrm{d}\boldsymbol{w}^a\right) P_{g,\varepsilon_t}\left(y^\mu \mid \frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}^a}{\sqrt{p}}\right) \\
&= \prod_{\mu=1}^{n} \int_{\mathbb{R}} \mathrm{d}y^\mu \int_{\mathbb{R}^d} P_{\boldsymbol{w}_\star}\left(\mathrm{d}\boldsymbol{w}_\star\right) \int_{\mathbb{R}^{p\times r}} \prod_{a=1}^{r} P_w\left(\mathrm{d}\boldsymbol{w}^a\right) \mathbb{E}_{\boldsymbol{x}_\mu}\left[P_0\left(y^\mu \mid \frac{\boldsymbol{z}_\mu^\top \boldsymbol{w}_\star}{\sqrt{d}}\right) \prod_{a=1}^{r} P_{g,\varepsilon_t}\left(y^\mu \mid \frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}^a}{\sqrt{d}}, \boldsymbol{s}^a\right)\right]
\end{aligned} \tag{83}$$

where P_{g,ε_t} is explicitly defined as:

$$P_{g,\varepsilon_t}\left(y^\mu \mid \frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}^a}{\sqrt{p}}, \boldsymbol{w}^a\right) = \frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{-\beta g\left(y \frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}^a}{\sqrt{p}} - \frac{\varepsilon_t}{\sqrt{p}} \|\boldsymbol{w}^a\|_{q^\star}\right)}, \tag{84}$$

and P_0 can represent any general noisy channel distribution.

The expectation term can be rewritten as:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_\mu}\left[P_0\left(y^\mu \mid \frac{\boldsymbol{z}_\mu^\top \boldsymbol{w}_\star}{\sqrt{d}}\right) \prod_{a=1}^{r} P_{g,\varepsilon_t}\left(y^\mu \mid \frac{\boldsymbol{x}_\mu \boldsymbol{w}^a}{\sqrt{p}}, \boldsymbol{w}^a\right)\right] \\
&= \int_{\mathbb{R}} \mathrm{d}\nu_\mu P_0\left(y \mid \nu_\mu\right) \int_{\mathbb{R}^r} \left[\prod_{a=1}^{r} \mathrm{d}\lambda_\mu^a P_{g,\varepsilon_t}\left(y^\mu \mid \lambda_\mu^a, \boldsymbol{w}^a\right)\right] \mathbb{E}_{\boldsymbol{x}_\mu}\left[\delta\left(\nu_\mu - \frac{\boldsymbol{z}_\mu^\top \boldsymbol{w}_\star}{\sqrt{d}}\right) \prod_{a=1}^{r} \delta\left(\lambda_\mu^a - \frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}^a}{\sqrt{p}}\right)\right]
\end{aligned} \tag{85}$$

When averaging over the dataset, the new variables follow a Gaussian distribution with covariances:

$$\rho \equiv \mathbb{E}\left[\nu_\mu^2\right] = \frac{1}{d}\|\boldsymbol{w}_\star\|_2^2, \tag{86}$$

$$m^a \equiv \mathbb{E}\left[\lambda_\mu^a \nu_\mu\right] = \frac{\sqrt{\gamma}}{d} \boldsymbol{w}_\star^\top \mathrm{F}^\top \boldsymbol{w}^a, \tag{87}$$

$$Q^{ab} \equiv \mathbb{E}\left[\lambda_\mu^a \lambda_\mu^b\right] = \frac{1}{p} \boldsymbol{w}^{a\top}\left(\mathrm{F}\,\mathrm{F}^\top + \mathrm{Id}_p\right) \boldsymbol{w}^b, \tag{88}$$

which can be organized into a single covariance matrix for the Gaussian pair (ν_μ, λ_μ).

We perform the following change of variables for the overlap matrix:

$$
\begin{aligned}
1 \propto & \int_{\mathbb{R}} \mathrm{d}\rho \, \delta\left(\gamma p \rho - \|\boldsymbol{w}_\star\|_2^2\right) \int_{\mathbb{R}^r} \prod_{a=1}^{r} \mathrm{d}m^a \, \delta\left(p\sqrt{\gamma} m^a - \boldsymbol{w}_\star^\top \mathrm{F}^\top \boldsymbol{w}^a\right) \\
& \int_{\mathbb{R}^{r\times r}} \prod_{1\le a\le b\le r} \mathrm{d}Q^{ab} \, \delta\left(pQ^{ab} - \boldsymbol{w}^{a\top}(\mathrm{F}\,\mathrm{F}^\top + \mathrm{Id}_p)\boldsymbol{w}^b\right) \\
= & \int_{\mathbb{R}} \frac{\mathrm{d}\rho \, \mathrm{d}\hat{\rho}}{2\pi} e^{-i\hat{\rho}\left(p\gamma\rho - \|\boldsymbol{w}_\star\|_2^2\right)} \int_{\mathbb{R}^r} \prod_{a=1}^{r} \frac{\mathrm{d}m^a \, \mathrm{d}\hat{m}^a}{2\pi} e^{-i\sum_{a=1}^r \hat{m}^a\left(p\sqrt{\gamma}m^a - \boldsymbol{w}_\star^\top \boldsymbol{s}^a\right)} \\
& \int_{\mathbb{R}^{r\times r}} \prod_{1\le a\le b\le r} \frac{\mathrm{d}Q^{ab} \, \mathrm{d}\hat{Q}^{ab}}{2\pi} e^{-i\hat{Q}^{ab}\left(pQ^{ab} - \boldsymbol{w}^{a\top}(\mathrm{F}\,\mathrm{F}^\top + \mathrm{Id}_p)\boldsymbol{w}^b\right)}
\end{aligned}
\tag{89}
$$

We define additional overlaps:

$$
P^a = \frac{1}{p}\|\boldsymbol{w}^a\|_{q^\star}^{q^\star}\,, \tag{90}
$$

which enter our computation as:

$$
1 \propto \int \prod_{a=1}^{r} \mathrm{d}P^a \, \delta\Big(pP^a - \|\boldsymbol{w}^a\|_{q^\star}^{q^\star}\Big) = \int \prod_{a=1}^{r} \frac{\mathrm{d}P^a \, \mathrm{d}\hat{P}^a}{2\pi} e^{-i\hat{P}^a\left(pP^a - \|\boldsymbol{w}^a\|_{q^\star}^{q^\star}\right)} \tag{91}
$$

The replicated partition function can now be written as:

$$
\mathbb{E}_{\mathcal{D}} \mathcal{Z}_\beta^r = \int \frac{\mathrm{d}\rho \, \mathrm{d}\hat{\rho}}{2\pi} \prod_{a=1}^{r} \frac{\mathrm{d}m^a \, \mathrm{d}\hat{m}^a}{2\pi} \frac{\mathrm{d}P^a \, \mathrm{d}\hat{P}^a}{2\pi} \prod_{1\le a\le b\le r} \frac{\mathrm{d}Q^{ab} \, \mathrm{d}\hat{Q}^{ab}}{2\pi} e^{p\Phi^{(r)}} \tag{92}
$$

where the r-replicated functional $\Phi^{(r)}$ is:

$$
\Phi^{(r)} = \Psi_t + \alpha\gamma\Psi_y^{(r)}\left(\rho, m^a, Q^{ab}, P^a\right) + \Psi_w^{(r)}\left(\hat{\rho}, \hat{m}^a, \hat{Q}^{ab}, \hat{P}^a\right) \tag{93}
$$

We have defined the trace term Ψ_t as:

$$
\Psi_t^{(r)} = -\gamma\rho\hat{\rho} - \sqrt{\gamma}\sum_{a=1}^{r} m^a\hat{m}^a - \sum_{1\le a\le b\le r} Q^{ab}\hat{Q}^{ab} - \sum_{a=1}^{r} P^a\hat{P}^a \tag{94}
$$

The prior part of the replicated free energy $\Psi_w^{(r)}$ is:

$$
\begin{aligned}
\Psi_w^{(r)} = & \frac{1}{p}\log\Bigg[\int_{\mathbb{R}^d} P_{\boldsymbol{w}_\star}\left(\mathrm{d}\boldsymbol{w}_\star\right) e^{\hat{\rho}\|\boldsymbol{w}_\star\|_2^2} \\
& \int_{\mathbb{R}^{p\times r}} \prod_{a=1}^{r} P_w\left(\mathrm{d}\boldsymbol{w}^a\right) e^{\sum_{a=1}^r\left(\hat{m}^a\boldsymbol{w}_\star^\top \mathrm{F}^\top \boldsymbol{w}^a + \hat{P}^a\|\boldsymbol{w}^a\|_{q^\star}^{q^\star}\right) + \sum_{1\le a\le b\le r}\left(\hat{Q}^{ab}\boldsymbol{w}^a(\mathrm{F}\,\mathrm{F}^\top + \mathrm{Id}_p)\boldsymbol{w}^b\right)}\Bigg]
\end{aligned}
\tag{95}
$$

And the channel part $\Psi_y^{(r)}$ is:

$$
\Psi_y^{(r)} = \log\left[\int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}} \mathrm{d}\nu P_0(y \mid \nu) \int \prod_{a=1}^{r} \mathrm{d}\lambda^a P_{g,\varepsilon_t}\left(y \mid \lambda^a, P^a\right) \mathcal{N}\left(\nu, \lambda^a; \mathbf{0}, \Sigma^{ab}\right)\right] \tag{96}
$$

where we've used the fact that (ν_μ, λ_μ) $\mu = 1, \dots n$ factors over all data points.

In the thermodynamic limit ($d \to \infty$ with fixed γ and α), the integral in eq. (92) concentrates around values that extremize $\Phi^{(r)}$, giving the free energy density:

$$
\beta f_\beta = -\lim_{r\to 0^+} \frac{1}{r} \operatorname{extr} \Phi^{(r)} = -\lim_{r\to 0^+} \partial_r \operatorname{extr} \Phi^{(r)} \tag{97}
$$

C.3 Replica Symmetric Ansatz

We propose the following replica symmetric ansatz for our variables:

$$\begin{array}{llll} m^a = m & \hat{m}^a = \hat{m} & & \text{for } a = 1, \dots r \\ q^{aa} = Q & \hat{q}^{aa} = -\frac{1}{2}\hat{Q} & & \text{for } a = 1, \dots r \\ q^{ab} = q & \hat{q}^{ab} = \hat{q} & & \text{for } 1 \le a < b \le r \\ P^a = P & \hat{P}^a = -\frac{1}{2}\hat{P} & & \text{for } a = 1, \dots r \end{array} \tag{98}$$

The trace term becomes:

$$\Psi_t = \frac{1}{2}q\hat{q} + \frac{1}{2}Q\hat{Q} + \frac{1}{2}P\hat{P} - \sqrt{\gamma}m\hat{m} = \frac{1}{2}(q\hat{q} + (V+q)(\hat{V}-\hat{q})) + \frac{1}{2}P\hat{P} - \sqrt{\gamma}m\hat{m} \tag{99}$$

And the channel term becomes:

$$\Psi_y = \mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right) \log \mathcal{Z}_y(y, \sqrt{q}\xi, V, P)\right] \tag{100}$$

with definitions:

$$\mathcal{Z}_0(y, \omega, V) = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} P_0(y \mid x)\,, \tag{101}$$

$$\mathcal{Z}_y(y, \omega, V, P) = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} \frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{-\beta g\left(yx - \varepsilon_t\, {}^{q^\star}\sqrt{P}\right)}\,. \tag{102}$$

C.3.1 Prior Term for Separable Regularization

Applying our ansatz to eq. (103), we can take the zero replica limit on the prior term:

$$\Psi_w = \frac{1}{p}\mathbb{E}_{\xi, \boldsymbol{w}_\star}\left[\log \int_{\mathbb{R}^p} P_w\left(\mathrm{d}\boldsymbol{w}\right) e^{-\frac{\hat{V}}{2}\boldsymbol{w}^\top(\mathrm{F}\,\mathrm{F}^\top + \mathrm{Id}_p)\boldsymbol{w} - \frac{\hat{P}}{2}\|\boldsymbol{w}\|_{q^\star}^{q^\star} - \boldsymbol{w}^\top\left(\hat{m}\,\mathrm{F}\,\boldsymbol{w}_\star - \sqrt{\hat{q}}(\mathrm{F}\,\mathrm{F}^\top + \mathrm{Id}_p)^{1/2}\mathbf{1}\xi\right)}\right] \tag{103}$$

Using the specific form of the feature matrix from eq. (53), we can simplify:

$$\mathrm{F}\,\mathrm{F}^\top = \begin{cases} \begin{bmatrix} \frac{p}{d}\,\mathrm{Id}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} & \text{if } p \ge d \\ \mathrm{Id}_p & \text{if } p < d \end{cases}, \quad \mathrm{F}\,\mathrm{F}^\top + \mathrm{Id}_p = \begin{cases} \begin{bmatrix} \left(1 + \frac{p}{d}\right)\mathrm{Id}_d & \mathbf{0} \\ \mathbf{0} & \mathrm{Id}_{(p-d)} \end{bmatrix} & \text{if } p \ge d \\ 2\,\mathrm{Id}_p & \text{if } p < d \end{cases}. \tag{104}$$

For $\gamma = {}^d\!/_p < 1$ (when $p \ge d$), we have:

$$\begin{aligned} \Psi_w^{(\gamma \le 1)} = \gamma \int \frac{e^{-\frac{1}{2}\xi^2}}{\sqrt{2\pi}} P(\mathrm{d}w_\star) \log \int P(dw) e^{-\frac{\hat{V}}{2}(1+\frac{1}{\gamma})w^2 - \frac{\hat{P}}{2}|w|^{q^\star} + \frac{\hat{m}}{\sqrt{\gamma}}ww_\star + \sqrt{\hat{q}}\sqrt{1+\frac{1}{\gamma}}w\xi} \\ + (1-\gamma)\int \frac{e^{-\frac{1}{2}\xi^2}}{\sqrt{2\pi}} P(\mathrm{d}w_\star) \log \int P(\mathrm{d}w) e^{-\frac{\hat{V}}{2}w^2 - \frac{\hat{P}}{2}|w|^{q^\star} + \sqrt{\hat{q}}w\xi}\,, \end{aligned} \tag{105}$$

And for $\gamma = {}^p\!/_d > 1$ (when $p < d$):

$$\Psi_w^{(\gamma > 1)} = \int \frac{e^{-\frac{1}{2}\xi^2}}{\sqrt{2\pi}} P(\mathrm{d}w_\star) \log \int P(\mathrm{d}w) e^{-\hat{V}w^2 - \frac{\hat{P}}{2}|w|^{q^\star} + \hat{m}w_* w + \sqrt{2\hat{q}}w\xi}\,, \tag{106}$$

After variable changes, we can express:

$$\Psi_w^{(\gamma \le 1)} = \gamma \mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}}{\sqrt{\hat{q}}}\frac{\xi}{\sqrt{1+\gamma}}, \frac{\hat{m}^2}{\hat{q}}\frac{1}{1+\gamma}\right) \log \mathcal{Z}_w\left(\sqrt{\hat{q}}\sqrt{1+\frac{1}{\gamma}}\xi, \hat{V}\left(1+\frac{1}{\gamma}\right), \frac{\hat{P}}{2}\right)\right] \tag{107}$$

$$+ (1-\gamma)\mathbb{E}_\xi\left[\log \mathcal{Z}_w\left(\sqrt{\hat{q}}\xi, \hat{V}, \frac{\hat{P}}{2}\right)\right], \tag{108}$$

$$\Psi_w^{(\gamma>1)} = \mathbb{E}_\xi \left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}}{\sqrt{2\hat{q}}}\xi, \frac{\hat{m}^2}{2\hat{q}}\right) \log \mathcal{Z}_w\left(\sqrt{2\hat{q}}\xi, 2\hat{V}, \frac{\hat{P}}{2}\right)\right], \tag{109}$$

where we define:

$$\mathcal{Z}_{w_\star}(\zeta, \Lambda) = \int P_\star(\mathrm{d}w_\star) e^{-\frac{1}{2}\Lambda\omega^2 + \zeta\omega} \tag{110}$$

$$\mathcal{Z}_w^\lambda(\zeta, \Lambda, \phi) = \int \mathrm{d}w e^{-\frac{\beta\lambda}{2}w^2} e^{-\frac{\Lambda}{2}w^2 - \phi|w|^{q^\star} + \zeta w}. \tag{111}$$

We also define:

$$f_w(\gamma, \Lambda, \lambda_2, \lambda_{q^\star}) = \arg\min_z \left[\lambda_2 z^2 + \lambda_{q^\star}|z|^{q^\star} + \frac{\Lambda}{2}z^2 - \gamma z\right] \tag{112}$$

C.4 Zero Temperature Limit

For the zero temperature limit, we apply the following parameter scalings:

$$\begin{aligned} V &\to \beta^{-1}V \qquad q \to q \qquad m \to m \qquad P \to P \\ \hat{V} &\to \beta\hat{V} \qquad \hat{q} \to \beta^2\hat{q} \qquad \hat{m} \to \beta\hat{m} \qquad \hat{P} \to \beta\hat{P} \end{aligned} \tag{113}$$

The channel term limit becomes:

$$\widetilde{\Psi}_y = \lim_{\beta\to\infty} \frac{1}{\beta}\Psi_y = -\mathbb{E}_\xi \left[\int \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right) \mathcal{M}_{Vg(y,\cdot;P,\varepsilon_t)}(\sqrt{q}\xi)\right] \tag{114}$$

where $\mathcal{M}_{Vg(y,\cdot;P,\varepsilon_t)}$ is the Moreau envelope:

$$\mathcal{M}_{Vg(y,\cdot;P,\varepsilon_t)}(\omega) = \min_z \left[\ell\left(yx - \varepsilon_t \sqrt[q^\star]{P}\right) + \frac{1}{2V}(z-\omega)^2\right]. \tag{115}$$

The zero temperature limit of the prior term gives:

$$\begin{aligned} \widetilde{\Psi}_w^{(\gamma\le 1)} &= \gamma\mathbb{E}_\xi \left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}}{\sqrt{\hat{q}}}\frac{\xi}{\sqrt{1+\gamma}}, \frac{\hat{m}^2}{\hat{q}}\frac{1}{1+\gamma}\right) \log \mathcal{Z}_w\left(\sqrt{\hat{q}}\sqrt{1+\frac{1}{\gamma}}\xi, \hat{V}\left(1+\frac{1}{\gamma}\right), \frac{\hat{P}}{2}\right)\right] \\ &\quad + (1-\gamma)\mathbb{E}_\xi \left[\log \mathcal{Z}_w\left(\sqrt{\hat{q}}\xi, \hat{V}, \frac{\hat{P}}{2}\right)\right]. \\ \widetilde{\Psi}_w^{(\gamma>1)} &= \mathbb{E}_\xi \left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}}{\sqrt{2\hat{q}}}\xi, \frac{\hat{m}^2}{2\hat{q}}\right) \log \mathcal{Z}_w\left(\sqrt{2\hat{q}}\xi, 2\hat{V}, \frac{\hat{P}}{2}\right)\right]. \end{aligned} \tag{116}$$

After taking the zero temperature limit, our free energy density is:

$$\lim_{\beta\to\infty} f_\beta = \operatorname*{extr}_{\substack{V,q,m,P \\ \hat{V},\hat{q},\hat{m},\hat{P}}} \left\{-\frac{1}{2}(q\hat{V} - \hat{q}V) - \frac{1}{2}P\hat{P} + \sqrt{\gamma}m\hat{m} + \alpha\gamma\widetilde{\Psi}_y + \widetilde{\Psi}_w\right\}. \tag{117}$$

C.5 Saddle-Point Equations

The extremization condition in eq. (117) yields the following relation for overlaps:

$$\begin{aligned} \hat{q} &= -2\alpha\gamma\partial_q\widetilde{\Psi}_y, & q &= -2\partial_{\hat{q}}\widetilde{\Psi}_w \\ \hat{Q} &= -2\alpha\gamma\partial_Q\widetilde{\Psi}_y, & Q &= -2\partial_{\hat{Q}}\widetilde{\Psi}_w, \\ \hat{P} &= -2\alpha\gamma\partial_P\widetilde{\Psi}_y, & P &= -2\partial_{\hat{P}}\widetilde{\Psi}_w \\ \hat{m} &= \alpha\sqrt{\gamma}\partial_m\widetilde{\Psi}_y, & m &= \frac{1}{\sqrt{\gamma}}\partial_{\hat{m}}\widetilde{\Psi}_w. \end{aligned} \tag{118}$$

The saddle-point equations for the channel part are:

$$\begin{aligned}
\hat{P} &= \alpha\gamma\varepsilon_t q^\star P^{q^\star-1}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, y\mathcal{Z}_0 f_{\ell,\varepsilon_t}\right], \\
\hat{V} &= -\alpha\gamma\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \mathcal{Z}_0\partial_\omega f_{\ell,\varepsilon_t}\right], \\
\hat{q} &= \alpha\gamma\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \mathcal{Z}_0 f^2_{\ell,\varepsilon_t}\right], \\
\hat{m} &= \alpha\sqrt{\gamma}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \partial_\omega\mathcal{Z}_0 f_{\ell,\varepsilon_t}\right],
\end{aligned} \tag{119}$$

For the prior term derivatives, we use the identities:

$$\partial_1\mathcal{Z}_w(\gamma,\Lambda) = \mathcal{Z}_w(\gamma,\Lambda) f_w(\gamma,\Lambda), \tag{120}$$

$$\partial_2\mathcal{Z}_w(\gamma,\Lambda) = -\frac{1}{2}\left(\partial_\gamma f_w(\gamma,\Lambda) + f_w^2(\gamma,\Lambda)\right), \tag{121}$$

The derivative with respect to $\hat{m}$ gives:

$$\partial_{\hat{m}}\widetilde{\Psi}_w^{(\gamma\le 1)} = \sqrt{\gamma}\mathbb{E}_\xi\left[\partial_1\mathcal{Z}_{w_\star}\left(\frac{\hat{m}\xi}{\sqrt{\hat{q}(1+\gamma)}}, \frac{\hat{m}^2}{\hat{q}(1+\gamma)}\right) f_w\left(\sqrt{\hat{q}}\sqrt{1+\frac{1}{\gamma}}\xi, \hat{V}\left(1+\frac{1}{\gamma}\right)\right)\right], \tag{122}$$

$$\partial_{\hat{m}}\Psi_w^{(\gamma>1)} = \mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}\xi}{\sqrt{2\hat{q}}}, \frac{\hat{m}^2}{2\hat{q}}\right) f_{w_\star}\left(\frac{\hat{m}\xi}{\sqrt{2\hat{q}}}, \frac{\hat{m}^2}{2\hat{q}}\right) f_w\left(\sqrt{2\hat{q}}\xi, 2\hat{V}\right)\right], \tag{123}$$

The derivative with respect to $\hat{q}$ gives:

$$\partial_{\hat{q}}\widetilde{\Psi}_w^{(\gamma\le 1)} = -\frac{1}{2}(1+\gamma)\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}\xi}{\sqrt{\hat{q}(1+\gamma)}}, \frac{\hat{m}^2}{\hat{q}(1+\gamma)}\right) f_w\left(\sqrt{\hat{q}}\sqrt{1+\frac{1}{\gamma}}\xi, \hat{V}\left(1+\frac{1}{\gamma}\right)\right)^2\right] \tag{124}$$

$$-\frac{1}{2}(1-\gamma)\mathbb{E}_\xi\left[f_w\left(\sqrt{\hat{q}}\xi, \hat{V}\right)^2\right], \tag{125}$$

$$\partial_{\hat{q}}\widetilde{\Psi}_w^{(\gamma<1)} = -\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}\xi}{\sqrt{2\hat{q}}}, \frac{\hat{m}^2}{2\hat{q}}\right) f_w\left(\sqrt{2\hat{q}}\xi, 2\hat{V}\right)^2\right]. \tag{126}$$

The derivative with respect to $\hat{Q}$ gives:

$$\partial_{\hat{Q}}\widetilde{\Psi}_w^{(\gamma\le 1)} = -\frac{1}{2}(1+\gamma)\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}\xi}{\sqrt{\hat{q}(1+\gamma)}}, \frac{\hat{m}^2}{\hat{q}(1+\gamma)}\right)\partial_1 f_w\left(\sqrt{\hat{q}}\sqrt{1+\frac{1}{\gamma}}\xi, \hat{V}\left(1+\frac{1}{\gamma}\right)\right)\right] \tag{127}$$

$$-\frac{1}{2}(1-\gamma)\mathbb{E}_\xi\left[\partial_1 f_w\left(\sqrt{\hat{q}}\xi, \hat{V}\right)\right] \tag{128}$$

$$\partial_{\hat{Q}}\widetilde{\Psi}_w^{(\gamma>1)} = -\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}\xi}{\sqrt{2\hat{q}}}, \frac{\hat{m}^2}{2\hat{q}}\right)\partial_1 f_w\left(\sqrt{2\hat{q}}\xi, 2\hat{V}\right)\right] \tag{129}$$

And finally, the derivative with respect to $\hat{P}$ gives:

$$\partial_{\hat{P}}\widetilde{\Psi}_w^{(\gamma\le 1)} = -\frac{1}{2}\gamma\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}}{\sqrt{\hat{q}}}\frac{\xi}{\sqrt{1+\gamma}}, \frac{\hat{m}^2}{\hat{q}}\frac{1}{1+\gamma}\right)\left|f_w\left(\sqrt{\hat{q}}\sqrt{1+\frac{1}{\gamma}}\xi, \hat{V}\left(1+\frac{1}{\gamma}\right)\right)\right|^{q^\star}\right] \tag{130}$$

$$-\frac{1}{2}(1-\gamma)\mathbb{E}_\xi\left[\left|f_w\left(\sqrt{\hat{q}}\xi, \hat{V}\right)\right|^{q^\star}\right]. \tag{131}$$

$$\partial_{\hat{P}}\widetilde{\Psi}_w^{(\gamma>1)} = -\mathbb{E}_\xi\left[\mathcal{Z}_{w_\star}\left(\frac{\hat{m}}{\sqrt{2\hat{q}}}\xi, \frac{\hat{m}^2}{2\hat{q}}\right)\left|f_w\left(\sqrt{2\hat{q}}\xi, 2\hat{V}\right)\right|^{q^\star}\right]. \tag{132}$$

Combining these derivatives with eq. (118) yields the final self-consistent equations.

The values of the feature space and latent space norms are

$$q_\ell = \frac{1}{d}\left\|\mathrm{F}^\top \boldsymbol{w}\right\|_2^2, \quad q_f = \frac{1}{p}\|\boldsymbol{w}\|_2^2, \tag{133}$$

and we have that

$$q_\ell = \frac{1}{\gamma}\begin{cases}\mathbb{E}\big[\mathcal{Z}_{w_\star}()f_w^2\big] & \gamma \le 1\\ \mathbb{E}\big[\mathcal{Z}_{w_\star}()f_w^2\big] & \gamma > 1\end{cases}, \tag{134}$$

$$q_f = \begin{cases}\gamma\mathbb{E}\big[\mathcal{Z}_{w_\star}()f_w^2\big] + (1-\gamma)\mathbb{E}\big[f_w^2\big] & \gamma \le 1\\ \mathbb{E}\big[\mathcal{Z}_{w_\star}()f_w^2\big] & \gamma > 1\end{cases} \tag{135}$$

D Numerical Details

The self-consistent equations from Theorem 4.1 are written in a way amenable to be solved via fixed-point iteration. Starting from a random initialization, we iterate through both the hat and non-hat variable equations until the maximum absolute difference between the order parameters in two successive iterations falls below a tolerance of 10^{-5}.

To speed-up convergence we use a damping scheme, updating each order parameter at iteration i, designated as x_i, using $x_i := x_i\mu + x_{i-1}(1-\mu)$, with μ as the damping parameter.

Once convergence is achieved for fixed λ, hyper-parameters are optimized using a gradient-free numerical minimization procedure for a one dimensional minimization.

For each iteration, we evaluate the proximal operator numerically using SciPy's [79] Brent's algorithm for root finding (`scipy.optimize.minimize_scalar`). The numerical integration is handled with SciPy's quad method (`scipy.integrate.quad`), which provides adaptive quadrature of a given function over a specified interval. These numerical techniques allow us to evaluate the equations and perform the necessary integrations with the desired accuracy.

Regarding the computer hardware all the experiments have been run on consumer grade hardware, specifically MacStudio M2 Ultra 2022, and none of the run took more than 1 day of CPU time.