

On the Existence of Consistent Adversarial Attacks in High-Dimensional Linear Classification

Matteo Vilucchio¹ **Lenka Zdeborová**² **Bruno Loureiro**³

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 captures 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 overparameterization affects vulnerability to label-preserving perturbations in a metric- and regime-dependent manner, 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

¹Information, Learning & Physics Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
²Statistical Physics of Computation Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
³Département d'Informatique de l'École Normale Supérieure (DI ENS), CNRS, PSL University, Paris, France. Correspondence to: Matteo Vilucchio <matteo.vilucchio@epfl.ch>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

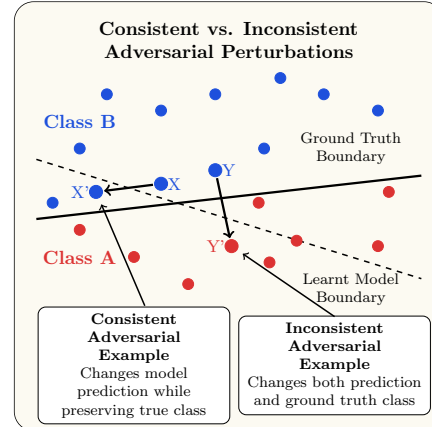


Figure 1. Illustration of the difference between consistent Adversarial Perturbation and inconsistent Adversarial Perturbation.

ground-truth label despite perturbations, whereas inconsistent ones change the true classification. Consider the classic panda example from [Goodfellow et al. \(2015\)](#): an image that, after subtle perturbations, is misclassified by a neural network but still depicts a panda to humans (consistent attack). If perturbations transformed the image to genuinely resemble a different animal, it would be inconsistent. This distinction is crucial: vulnerability to consistent attacks represents genuine model failure to capture invariant features.

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 (Madry et al., 2018). While this assumption has been explored in theoretical works on robust generalization (Donhauser et al., 2021; Raghunathan et al., 2020), a mathematical understanding of such attacks, including their existence and effectiveness in tricking even simple linear classifiers, remains elusive.

Following a large body of work in high-dimensional statistics (Krogh & Hertz, 1991; Seung et al., 1992; Bean et al., 2013a; Thrampoulidis et al., 2018; Aubin et al., 2020; Vilucchio et al., 2024), we analyze this problem through the lens of exact asymptotics of linear classifiers. We develop novel metrics that precisely quantify vulnerability to both con-

sistent and inconsistent adversarial attacks. We 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 (Goodfellow et al., 2015; Madry et al., 2018), the most widely adopted method for 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. Ribeiro & Schön (2023) consider the regression case for squared loss, but in classification the question is still open. While overparameterization generally improves standard generalization (Goldt et al., 2022b; Hastie et al., 2022), 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. However, our analysis reveals a more nuanced picture: the effect of overparameterization is both metric- and regime-dependent, and the finite-range trends of consistent robust and boundary errors can turn over in the extreme overparameterized limit. See Section A for a further discussion of related works.

Our **main contributions** can be summarized as¹

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 overparameterization 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 asymptotic formulas that exactly characterize their high-dimensional limits under a Gaussian design assumption.
3. We study how robust empirical risk minimization can mitigate consistent attacks in this high-dimensional limit, for both the well-specified and latent variable model. For the latter, this requires a novel exact asymptotic characterization of the robust ERM estimator under misspecification,

¹Code reproducing all results is available at <https://github.com/IdePHICS/adv-consistent>.

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.

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

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. We formalize this notion in the context of binary classifiers.

Consider a binary classification task defined by a covariate distribution μ over $\mathbb{R}^d$, a target score $g_\star : \mathbb{R}^d \rightarrow \mathbb{R}$, and a link $\varphi_\star$ such that $f_\star = \varphi_\star \circ g_\star$ and $f_\star(\mathbf{x}) = \mathbb{P}(y = +1|\mathbf{x})$.

Definition 2.1 (Consistent attack). Let $f_\star = \varphi_\star \circ g_\star$ and $\hat{f} : \mathbb{R}^d \rightarrow [0, 1]$ denote two binary classifiers, referred to as the *target* and the *model*, $\mathbf{x} \in \mathbb{R}^d$ a covariate and $\hat{y} : [0, 1] \rightarrow \{\pm 1\}$ a decision rule associated to $\hat{f}$. We say a perturbation $\delta \in B_q(\varepsilon)$ of the model $\hat{f}$ is a *consistent adversarial attack* with respect to the target $f_\star$, the covariate $\mathbf{x} \in \mathbb{R}^d$ and the decision rule $\hat{y}$ if the following conditions hold:

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

If target-score invariance fails the attack is *inconsistent*.

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

Remark 2.2. For deterministic classifiers, target-score invariance implies target-label invariance, but not conversely: $\{\delta : g_\star(\mathbf{x}) = g_\star(\mathbf{x} + \delta)\} \subseteq \{\delta : \hat{y}(f_\star(\mathbf{x})) = \hat{y}(f_\star(\mathbf{x} + \delta))\}$.

For probabilistic classifiers, it instead preserves the conditional label distribution $\mathbb{P}(y|x)$, not individual random label realizations—the natural statistical analogue. Hence our consistent metrics lower-bound label-preserving vulnerability, while the standard robust error upper-bounds it.

We introduce metrics to quantify these properties.

Definition 2.3 (Adversarial errors). Let (μ, g_*, φ_*) denote a binary classification task. Given a classifier $\hat{f} : \mathbb{R}^d \rightarrow [0, 1]$ and its associated predictor $\hat{y}(x) = \text{sign}(2\hat{f}(x) - 1)$, we define the following three metrics

- **Robust error:** This is the standard notion of robust generalization error in the adversarial literature (Clarysse et al., 2022; Dohmatob, 2024; Dohmatob & Scetbon, 2023), and simply quantifies how vulnerable $\hat{f}$ is to arbitrary perturbations in a ℓ^q -ball:

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

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

$$E_{\text{rob}}^{\text{cns}} = \mathbb{E}_{(x,y)} \left[\max_{\substack{\delta \in B_q(\varepsilon): \\ g_*(x) = g_*(x+\delta)}} \mathbb{1}\{y \neq \hat{y}(x + \delta)\} \right], \quad (2)$$

Note that the critical difference between E_{rob} and $E_{\text{rob}}^{\text{cns}}$ is the target-score invariance constraint from Theorem 2.1 (cf. Theorem 2.2).

- **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:

$$E_{\text{bnd}}^{\text{cns}} = \mathbb{E}_{(x,y)} \left[\max_{\substack{\delta \in B_q(\varepsilon): \\ g_*(x) = g_*(x+\delta)}} \mathbb{1}\left\{ \begin{array}{l} y \neq \hat{y}(x + \delta) \\ y = \hat{y}(x) \end{array} \right\} \right]. \quad (3)$$

Remark 2.4. For each (x, y) , the feasible set of consistent perturbations is a subset of $B_q(\varepsilon)$, while the boundary-error integrand additionally requires $y = \hat{y}(x)$. The corresponding maxima are therefore ordered pointwise, and hence $0 \leq E_{\text{bnd}}^{\text{cns}} \leq E_{\text{rob}}^{\text{cns}} \leq E_{\text{rob}}$.

3. Consistent Attacks In Well-Specified Linear Classification

Despite an 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 3.1 (Linear classifiers). A linear binary classifier in $\mathbb{R}^d$ is a function

$$f_w(x) = \mathbb{P}_w(y = +1|x) = \varphi(g_w(x)), \quad (4)$$

defined by the *weight vector* $w \in \mathbb{R}^d$, the score $g_w(x) = \langle w, x \rangle$, and a monotonic *link function* $\varphi : \mathbb{R} \rightarrow [0, 1]$.

The class of linear binary classifiers encompasses 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 $\varphi(t) = \mathbb{1}\{t \geq 0\}$ 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_{w_*} denote a reference linear classifier with weights $w_* \in \mathbb{S}^{d-1}(\sqrt{d})$ and link function φ_* , which we refer to as the *ground-truth*. For the linear target score $g_{w_*}(x) = \langle w_*, x \rangle$, target-score invariance is equivalent to $\langle \delta, w_* \rangle = 0$, i.e. the attack must be orthogonal to w_* . Therefore, the set of admissible consistent adversarial attacks with respect to the target classifier defines a hyperplane:

$$H_q(\varepsilon) := \{\delta \in B_q(\varepsilon) : \langle w_*, \delta \rangle = 0\}. \quad (5)$$

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

$$|\langle \hat{w}, \delta \rangle| > |\langle \hat{w}, x \rangle|, \quad \text{sign}(\langle \hat{w}, \delta \rangle) = -\text{sign}(\langle \hat{w}, x \rangle). \quad (6)$$

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

Proposition 3.2 (Existence of consistent attack). *Consider two linear classifiers defined by the weights $w_* \in \mathbb{S}^{d-1}(\sqrt{d})$ and $\hat{w} \in \mathbb{R}^d$. Let $x \in \mathbb{R}^d$ denote a covariate, and assume $\langle \hat{w}, x \rangle \neq 0$. Then, a consistent attack $\delta \in B_q(\varepsilon)$ with respect to $w_*, x \in \mathbb{R}^d$ and the decision function $\hat{y}(x) = \text{sign}(2f_{\hat{w}}(x) - 1)$ exists if and only if:*

$$\varepsilon d_{q^*}^*(\hat{w}_\perp) \geq |\langle \hat{w}, x \rangle|, \quad (7)$$

where $\hat{w}_\perp = \hat{w} - 1/d\langle w_*, \hat{w} \rangle w_*$ is the predictor components orthogonal to the target weights, $d_{q^*}^*(\hat{w}_\perp) = \inf_{\mu \in \mathbb{R}} \|\hat{w}_\perp - \mu w_*\|_{q^*}$ is the ℓ^{q^*} distance to the span(w_*) and q^* is the dual of q .

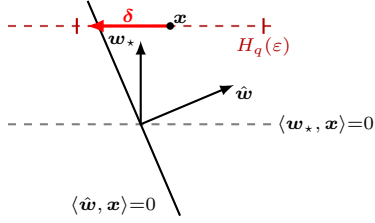


Figure 2. Geometry of consistent attacks (Section 3.1). Target-score invariance forces the perturbation to lie in the hyperplane $H_q(\varepsilon) = \{\delta \in B_q(\varepsilon) : \langle w_*, \delta \rangle = 0\}$ orthogonal to the target weights w_* (here a segment, delimited by the two red ticks), so δ stays parallel to the target boundary (upper dashed line). A consistent attack succeeds when it pushes x across the model boundary $\langle \hat{w}, x \rangle = 0$ without leaving $H_q(\varepsilon)$.

Remark 3.3. For $q = 2$, the infimum in the definition of $d_{q^*}^*$ is achieved at $\mu = 0$:

$$d_2^*(\hat{w}_\perp) = \inf_{\mu \in \mathbb{R}} \|\hat{w}_\perp - \mu w_*\|_2 = \|\hat{w}_\perp\|_2. \quad (8)$$

While $d_{q^*}^*(\hat{w}_\perp) \leq \|\hat{w}_\perp\|_{q^*}$ is always an upper bound, it is not tight for $q \neq 2$, except for particular choices of $w_* \in \mathbb{S}^{d-1}(\sqrt{d})$, for instance $w_* = \sqrt{d}e_1$. This highlights how the existence of consistent attacks depends 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. (7) holds for an inconsistent attack, but without the constraint $\langle w_*, \delta \rangle = 0$. The direct comparison is

$$d_{q^*}^*(\hat{w}_\perp) = \inf_{\nu \in \mathbb{R}} \|\hat{w}_\perp - \nu w_*\|_{q^*} \leq \|\hat{w}_\perp\|_{q^*}, \quad (9)$$

this loosens the existence condition, as expected. In particular, the stronger the overlap between the ground-truth and the model $\langle \hat{w}, w_* \rangle$, 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 w_* . This leads to the following corollary.

Corollary 3.4 (Existence of inconsistent attack). *Under the same setting as Theorem 3.2, an inconsistent adversarial attack exists if and only if:*

$$\epsilon \|\hat{w}\|_{q^*} \geq |\langle \hat{w}, x \rangle|. \quad (10)$$

Moreover, every consistent perturbation is feasible for the unrestricted adversary, and hence $E_{\text{rob}}^{\text{cns}} \leq E_{\text{rob}}$.

Theorem 3.2 allows us to identify the region in $\mathbb{R}^d$ that is vulnerable to consistent attacks. Indeed, defining the ground-truth orthogonal margin

$$\kappa_q(x) = \frac{|\langle \hat{w}, x \rangle|}{d_{q^*}^*(\hat{w}_\perp)}, \quad (11)$$

a covariate $x \in \mathbb{R}^d$ is vulnerable to a consistent $\delta \in H_q(\varepsilon)$ attack if and only if $\varepsilon > \kappa_q(x)$, and the *vulnerable region* is given by $\{x \in \mathbb{R}^d : \kappa_q(x) < \varepsilon\} \subset \mathbb{R}^d$ — a tube around the decision hyperplane. Note that this reveals two strategies for mitigating consistent adversarial attacks by increasing κ_q : (a) aligning with the target weights; (b) reducing $d_{q^*}^*(\hat{w}_\perp)$. While the first option is typically beyond the statistician’s control, the second can be achieved by explicitly regularizing training with the norm dual to the attack, which upper-bounds $d_{q^*}^*(\hat{w}_\perp)$, see Eq. (9). This is consistent with previous theoretical results suggesting the use of the dual norm in ERM (Yin et al., 2019; Awasthi et al., 2020; Tsilivis et al., 2024; Vilucchio et al., 2026), and is 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 best illustrated in the Gaussian case.

Corollary 3.5 (Existence for Gaussian Covariates with General Covariance). *In the same setting as Theorem 3.2 with Gaussian covariates $x \sim \mathcal{N}(\mathbf{0}, \Sigma)$, the probability that a consistent attack exists is:*

$$\mathbb{P}[\exists \text{ consistent } \delta \in H_q(\varepsilon)] = 2\Phi\left(\frac{\varepsilon d_{q^*}^*(\hat{w}_\perp)}{\sqrt{\hat{w}^\top \Sigma \hat{w}}}\right) - 1, \quad (12)$$

where Φ is the standard normal c.d.f. For the noiseless link and nondegenerate target and model fields, let $\rho = d_{q^*}^*(\hat{w}_\perp) / \|\hat{w}\|_{q^*}$. The finite-dimensional errors additionally satisfy $\rho E_{\text{rob}}(\varepsilon) \leq E_{\text{rob}}^{\text{cns}}(\varepsilon) = E_{\text{rob}}(\rho\varepsilon) \leq E_{\text{rob}}(\varepsilon)$.

Remark 3.6. The probability in Eq. (12) is non-decreasing in q (non-increasing in q^*) and d . To build intuition, we first consider the *isotropic case* $\Sigma = \frac{1}{d}\text{Id}_d$ with a random predictor $\hat{w} = m w_* + \sqrt{1 - m^2} \xi$, where $\xi \sim \mathcal{N}(0, \text{Id}_d)$ and $m = \frac{1}{d} \langle w_*, \hat{w} \rangle$ is the correlation with the target. Then:

$$\begin{aligned} \mathbb{P}[\exists \text{ consis. } \delta \in H_q(\varepsilon)] \\ = 2\Phi\left(\varepsilon \sqrt{d} \frac{\sqrt{1 - m^2} d_{q^*}^*(\xi)}{\|\hat{w}\|_2}\right) - 1. \end{aligned}$$

Since $d_{q^*}^*(\xi) = \Theta(d^{1/q^*})$ as $d \rightarrow \infty$, for fixed $\varepsilon = \Theta(1)$ and $q > 1$ the probability approaches one unless $m^2 = 1$ (perfect alignment). This highlights the vulnerability of high-dimensional predictors, though sparse predictors can exhibit different behavior since $d_{q^*}^*(\hat{w}_\perp)$ depends on the part of the predictor support that does not overlap with the target. We report how the existence probability depends on the attack geometry q and the dimension d in Figure 3 (Left, Center).

Remark 3.7 (Algorithmic vs. statistical irreducibility). The probability in Theorem 3.6 is strictly positive whenever $m^2 < 1$, with two contributing sources. The *algorithmic gap*, the suboptimality of a given estimator relative to the best

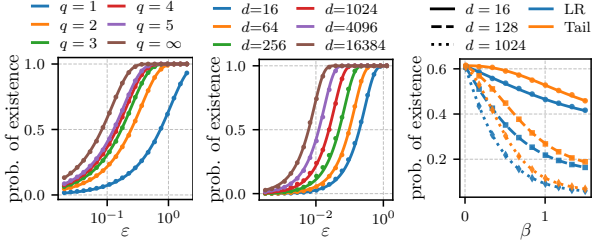


Figure 3. Probability of existence of consistent adversarial attacks for Gaussian covariates via Eq. (12), with $\mathbf{w}_* = \sqrt{d}\mathbf{e}_1$ and $\hat{\mathbf{w}} \in \mathbb{S}^{d-1}(\sqrt{d})$ having correlation $m = 0.5$ with $\mathbf{w}_*$. (Left) Isotropic case $\Sigma = 1/d \text{Id}_d$: probability vs ε for different attack geometries q with $d = 10$. (Center) Dimensional scaling with $q = 2$, showing probability vs ε for varying d for $\Sigma = 1/d \text{Id}_d$. (Right) Power law covariance $\lambda_i \sim i^{-\beta}$ for several d : probability vs β at fixed $\varepsilon = 1$ for two predictor types (LR uses top $k=10$ modes; Tail uses bottom modes $(k+2):d$). Solid curves show theory; dots show empirical frequencies from $n = 10^4$ samples satisfying Eq. (7).

achievable one, can be reduced by tuning the estimation of $\hat{\mathbf{w}}$ to maximize m . The *statistical gap* is the residual misalignment of the *best possible* estimator at finite α : for the continuous weight priors considered here, even the Bayes-optimal estimator achieves only $m^2 < 1$, with generalization error decaying as $\Theta(1/\alpha)$ but never vanishing (Aubin et al., 2020). This contrasts with discrete or sparse priors, where a phase transition can yield exact recovery above a critical α (Barbier et al., 2019). Consistent attacks are thus an *irreducible* statistical phenomenon under our assumptions.

Remark 3.8. Consider Σ with *power law spectrum* with eigenvalues $\lambda_i \propto i^{-\beta}$ (for $\beta > 0$) normalized to $\sum_i \lambda_i = d$ and let $\{\mathbf{v}_i\}_{i=1}^d$ denote the corresponding orthonormal eigenbasis. Suppose $\mathbf{w}_* = \sqrt{d}\mathbf{v}_1$ and that the predictor has a fixed correlation m with the target. As β increases, the spectrum concentrates on λ_1 (which grows toward d), reducing the attack probability for all predictors (Figure 3, Right). However, vulnerability also depends on where $\hat{\mathbf{w}}_\perp$ distributes its mass in the eigenbasis, the two cases are:

- **Low-rank (LR) alignment:** When $\hat{\mathbf{w}}_\perp \propto \sum_{i=2}^{k+1} \mathbf{v}_i$ concentrates on the top-2:($k+1$) eigenmodes and it benefits from eigenvalues $\lambda_2, \dots, \lambda_{k+1}$ being large.
- **Tail alignment:** When $\hat{\mathbf{w}}_\perp \propto \sum_{i=k+2}^d \mathbf{v}_i$ concentrates on tail modes, it places norm on eigenmodes $\lambda_{k+2}, \dots, \lambda_d$.

Figure 3 (Right) shows how the tail predictor is *more vulnerable* than the LR predictor by a constant logit-level factor.

3.2. Robust Empirical Risk Minimization

A natural question is whether robust training can effectively mitigate consistent attacks. Robust empirical risk minimization (ERM) learns robust classifier rules from a dataset

$\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{-1, +1\} : i = 1, \dots, n\}$. From $\mathcal{D}$, the statistician estimates a classifier by optimizing the robust *empirical* (regularized) risk, defined

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^n \max_{\|\mathbf{v}_i\|_s \leq R} \ell(y_i \langle \mathbf{w}, \mathbf{x}_i + \mathbf{v}_i \rangle) + \lambda \|\mathbf{w}\|_2^2, \quad (13)$$

where $\ell : \mathbb{R} \rightarrow \mathbb{R}_+$ is a non-increasing convex loss function and $\lambda \geq 0$ is a regularization parameter, while $R \geq 0$ is the training attack radius. The case $R = 0$ corresponds to standard ERM, while $R > 0$ corresponds to robust training. Given the dataset $\mathcal{D}$, we estimate the model’s weights as

$$\hat{\mathbf{w}} \in \arg \min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}). \quad (14)$$

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

3.3. High-Dimensional Asymptotic Analysis

Motivated by Theorem 3.6, 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 $\hat{\mathbf{w}}$ is a trained predictor under the Gaussian design assumption, and discuss the benefits of robust ERM in mitigating consistent adversarial attacks. We will work under the following assumptions.

Assumption 3.9 (Data distribution). We assume the covariates are isotropic Gaussian $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, 1/d \text{Id}_d)$ and that labels are generated from a ground-truth linear classifier $y \sim \text{Rad}(f_{\mathbf{w}_*}(\mathbf{x}))$ where $f_{\mathbf{w}_*}(\mathbf{x}) = \mathbb{P}(y = +1 | \mathbf{x}) = \varphi(\langle \mathbf{w}_*, \mathbf{x} \rangle)$ with monotonic link function φ and ground-truth weights $\mathbf{w}_* \in \mathbb{S}^{d-1}(\sqrt{d})$.

Assumption 3.10 (Scaling of the Adversarial Strength). For a given perturbation geometry $\delta \in B_q(\varepsilon)$ with $q > 1$, we assume that $\varepsilon = O_d(d^{-1/q^*})$ as $d \rightarrow \infty$, where q^* is the dual. We define the rescaled radius as $\tilde{\varepsilon} = \varepsilon d^{1/q^*}$. Likewise, for the ℓ^s inner maximization in Eq. (13), we take $R = r d^{-1/s^*}$, where the dimension-free training budget $r = O(1)$ is the parameter entering the asymptotic formulas.

Remark 3.11. As briefly discussed in Theorem 3.6, Theorem 3.10 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 (Taheri et al., 2023; Javanmard et al., 2020; Tanner et al., 2025).

Robust adversarial training for well-specified linear classifiers on Gaussian covariates (Theorem 3.9) has been studied by (Vilucchio et al., 2026) in the proportional high-dimensional asymptotics where n diverges with d at constant ratio $\alpha = n/d = \Theta(1)$. They characterize the limiting

distribution of the robust estimator $\hat{w}$ for general ℓ^s -robust training through a nonlinear system of equations.

For clarity of exposition, we present our result about the limiting behavior of consistent metrics for $s = 2$ and noiseless channel (for the generic form see Section C.3).

Theorem 3.12 (Consistent metrics for well-specified model). *Under Theorems 3.9 and 3.10, with a noiseless link, the metrics defined in Eqs. (2) and (3) with decision rule $\hat{y}(x) = \text{sign}(2f_{\hat{w}}(x) - 1)$, where $\hat{w}$ comes from Eq. (14), concentrate in the high-dimensional limit where $n, d \rightarrow \infty$ with $n/d = \Theta(1)$ to the following two dimensional integrals*

$$E_{\text{rob}}^{\text{cns}} = \int dp(z_1, z_2) \mathbb{I}\{\text{sign}(z_1)z_2 - \tilde{\varepsilon}\mathcal{A} < 0\}, \quad (15)$$

$$E_{\text{bnd}}^{\text{cns}} = \int dp(z_1, z_2) \mathbb{I}\{0 < \text{sign}(z_1)z_2 < \tilde{\varepsilon}\mathcal{A}\}, \quad (16)$$

where $\mathcal{A} = \sqrt{2(\tau - m^2)} \sqrt[q]{\Gamma(\frac{q^*+1}{2})}/\sqrt{\pi}$ and $(z_1, z_2) \sim \mathcal{N}(0, (\frac{1}{m} \frac{m}{\tau}))$ with $\tau = 1/d \|\hat{w}\|_2^2$ the predictor's ℓ_2 -norm and $m = 1/d \langle w_*, \hat{w} \rangle$ its scalar product with the target.

The proof proceeds by explicitly solving the inner maximization and then showing concentration to the previous expression in the high-dimensional limit; see Section C.3. The values of τ and m can be obtained from the characterization of $\hat{w}$ in (Vilucchio et al., 2026).

Remark 3.13 (Intuition via local fields). Equations (15) and (16) admit a *local fields* reading (Clarté et al., 2023): under the normalized design $x \sim \mathcal{N}(0, \text{Id}_d/d)$, the Gaussian pair $(z_1, z_2) = (\langle w_*, x \rangle, \langle \hat{w}, x \rangle)$ encodes the signed distances of x to the target and model boundaries. The attack subtracts $\tilde{\varepsilon}\mathcal{A}$ from the signed model margin $\text{sign}(z_1)z_2$; hence the boundary error is precisely the probability that this initially positive margin lies between 0 and $\tilde{\varepsilon}\mathcal{A}$. The attack surface $\mathcal{A} \propto \|\hat{w}_\perp\|_2$ only sees the predictor component orthogonal to w_* (Theorem 3.2).

For completeness, we report the limiting value of E_{rob} already derived in (Tanner et al., 2025) as being

$$E_{\text{rob}} = \int dp(z_1, z_2) \mathbb{I}\{\text{sign}(z_1)z_2 - \tilde{\varepsilon}\mathcal{B} < 0\},$$

$$\mathcal{B} = \sqrt{2\tau} \sqrt[q]{\frac{\Gamma(\frac{q^*+1}{2})}{\sqrt{\pi}}}.$$

The consistent versions of the errors (Eqs. (15) and (16)) depend on the quantity $\sqrt{\tau - m^2}$ while E_{rob} depends on $\sqrt{\tau}$. Since $\sqrt{\tau - m^2} < \sqrt{\tau}$ (as $\tau \geq m^2 > 0$), this explains why consistent attacks are less effective than inconsistent ones for the same attack strength ε , confirming our earlier result from Theorem 3.4 and as illustrated in Figure 4 (Left). The former quantity is precisely the length of $P_\perp \hat{w}$ appearing in Theorem 3.2. Section B.2 presents additional experiments comparing consistent and inconsistent boundary errors.

Figure 4 (Left) shows the asymptotic dependence of the metrics in Theorem 2.3 with the rescaled perturbation strength $\tilde{\varepsilon}$ 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_2 -attack one needs a strength of $\tilde{\varepsilon} \approx 1$ ($\varepsilon \approx d^{-1/2}$).

Using the results from (Vilucchio et al., 2026) to find the values of τ and m we can use Theorem 3.12 to analyze the consistent errors of the robustly trained $\hat{w}$ from Eq. (14) trained on n samples in dimension d , where both n and d tend to infinity with fixed ratio $\alpha = n/d$.

Figure 4 (Right) shows the performance of robustly trained $\hat{w}$ as a function of $\alpha = n/d$, demonstrating a monotonic decrease of all metrics with the sample complexity α for two different attack geometries. While the errors E_{rob} and $E_{\text{rob}}^{\text{cns}}$ start from the same values, $E_{\text{rob}}^{\text{cns}}$ decreases faster with α , indicating that with more samples the model learns representations that are more effective against consistent adversarial attacks.

We observe that ℓ^2 -attacks ($q=2$) yield higher robust errors than ℓ^∞ -attacks ($q=\infty$), demonstrating that attack geometry affects vulnerability beyond the difference in perturbation budget available across norms: while Theorem 3.10 ensures attacks are non-trivial, the geometric properties of each norm determine how effectively that budget translates into successful adversarial perturbations. Additional experiments are presented in Section B.2.

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 (Chen et al., 2024).

In this section, we investigate this question on a popular mathematical testbed for studying the role of overparameterization, the *latent variable model* (Hastie et al., 2022). In this model, the ground-truth classifier $f_{w_*}(z) = \varphi(\langle w_*, z \rangle)$ is defined in a latent space with latent covariates $z \in \mathbb{R}^d$ and weights $w_* \in \mathbb{S}^{d-1}(\sqrt{d})$. Labels are generated according to the latent rule $y \sim \text{Rad}(f_{w_*}(z))$ as in Eq. (4). The statistician does not observe the latent covariates $z \in \mathbb{R}^d$ directly but instead has access to a transformed representation $x \in \mathbb{R}^p$ defined as $x = Fz + u$, where $F \in \mathbb{R}^{p \times d}$ is the *feature matrix* and $u \sim \mathcal{N}(0, 1/p \text{Id}_p)$ is an independent noise term.

While this model seems simplistic, recent Gaussian universality results have shown that in the proportional limit,

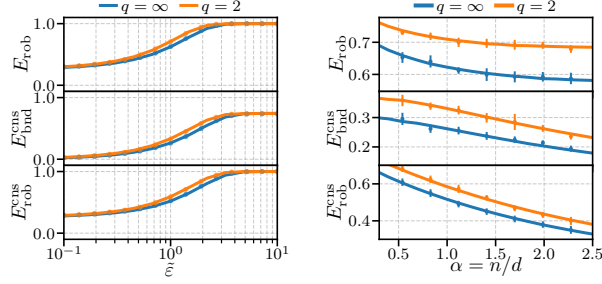


Figure 4. (Left) Dependence of the metrics in Eqs. (1) to (3) for the well-specified model as a function of the attackers norm. Points are 10 different simulations for $d = 500$, showing good agreement already at finite dimension. (Right) Performance difference for optimally regularized non-robust training under attacks constrained with different norms. The points show 10 different simulations with $d = 500$ and n scaled such that $n = \alpha d$.

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*) (Mei & Montanari, 2022; Goldt et al., 2022a; Gerace et al., 2020; Hu & Lu, 2023; Montanari & Saeed, 2022). 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 . When $p > d$, we will say the model is *overparameterized*, and when $p < d$, the model is *underparameterized*.

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 ($\delta \in \mathbb{R}^d$) or feature space ($\delta \in \mathbb{R}^p$). Considering perturbations in feature space, *i.e.* perturbations to $\mathbf{x}$, 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 former, where the conditions in Theorem 2.1 translate to:

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

where the model weights are denoted by $\hat{\theta} \in \mathbb{R}^p$ 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 4.1 (Existence of consistent latent space attacks). *Consider the setting of binary classification in the latent space model: a linear classifier with weights $\mathbf{w}_* \in \mathbb{S}^{d-1}(\sqrt{d})$ assigns labels according to $y \sim \text{Rad}(f_{\mathbf{w}_*}(\mathbf{z}))$,*

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

$$\varepsilon d_{q^*}^*(P_{\perp} F^{\top} \hat{\theta}) \geq |\langle \hat{\theta}, F\mathbf{z} + \mathbf{u} \rangle| \quad (17)$$

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

Remark 4.2. While in the well-specified case the vulnerable region is determined by the margin $\kappa_q(\mathbf{x})$ in Eq. (11), in the latent model this is defined by the latent margin:

$$\eta_q(\mathbf{z}) := \frac{|\langle \hat{\theta}, F\mathbf{z} + \mathbf{u} \rangle|}{d_{q^*}^*(P_{\perp} F^{\top} \hat{\theta})}. \quad (18)$$

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

Corollary 4.3 (Existence for Gaussian latent variables). *In the case of i.i.d. Gaussian latent variables $\mathbf{z} \sim \mathcal{N}(0, \frac{1}{d} \text{Id}_d)$ and $\mathbf{u} \sim \mathcal{N}(0, \frac{1}{p} \text{Id}_p)$, the probability a consistent attack exists is given by:*

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

where Φ is the standard normal c.d.f.

Remark 4.4. In the latent variable model, the probability of existence depends on the predictor’s latent-space projection $F^{\top} \hat{\theta}$ rather than on the predictor alone. In particular, components of $\hat{\theta} \in \mathbb{R}^p$ lying in $\text{Ker}(F^{\top})$ contribute only to the ℓ^2 -norm in the denominator. Thus, in the overparameterized setting $p > d$, one can reduce the probability of a consistent attack either by increasing alignment with the target $\langle \mathbf{w}_*, F^{\top} \hat{\theta} \rangle$ or by concentrating most of the norm in the $p - d$ directions in $\text{Ker}(F^{\top})$. This is closely related to the conditions for benign overfitting in (Bartlett et al., 2020).

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

$$E_{\text{rob}}^{\text{cns}} = \mathbb{E} \left[\max_{\substack{\delta \in B_q(\varepsilon): \\ \langle \mathbf{w}_*, \mathbf{z} \rangle = \langle \mathbf{w}_*, \mathbf{z} + \delta \rangle}} \mathbb{1}\{y \neq \hat{y}(F(\mathbf{z} + \delta) + \mathbf{u})\} \right], \quad (19)$$

$$E_{\text{bnd}}^{\text{cns}} = \mathbb{E} \left[\max_{\substack{\delta \in B_q(\varepsilon): \\ \langle \mathbf{w}_*, \mathbf{z} \rangle = \langle \mathbf{w}_*, \mathbf{z} + \delta \rangle}} \mathbb{1}\left\{ \begin{array}{l} y \neq \hat{y}(F(\mathbf{z} + \delta) + \mathbf{u}) \\ y = \hat{y}(F\mathbf{z} + \mathbf{u}) \end{array} \right\} \right], \quad (20)$$

$$E_{\text{rob}} = \mathbb{E} \left[\max_{\delta \in B_q(\varepsilon)} \mathbb{1}\{y \neq \hat{y}(F(\mathbf{z} + \delta) + \mathbf{u})\} \right]. \quad (21)$$

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 $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{-1, +1\} : i = 1, \dots, n\}$ 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 $\hat{\theta}$ estimated from $\mathcal{D}$ using Eqs. (13) and (14).² Our results will hold under the following assumptions.

Assumption 4.5 (High-dimensional limit). We consider the proportional high-dimensional limit where $n, p, d \rightarrow \infty$ at fixed ratios $\alpha := n/d$ and $\psi := p/n$. For convenience, we also define $\gamma := (\alpha\psi)^{-1} = d/p$.

Assumption 4.6 (Data Distribution Latent Space Model). We assume that data $(\mathbf{x}, y) \in \mathbb{R}^p \times \{-1, +1\}$ is drawn from a latent variable model with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \text{Id}_d/d)$ and ground-truth linear classifier $f_{\mathbf{w}_*}(\mathbf{z}) = \varphi(\langle \mathbf{w}_*, \mathbf{z} \rangle)$ with $\mathbf{w}_* \in \mathbb{S}^{d-1}(\sqrt{d})$. The observed features $\mathbf{x} \in \mathbb{R}^p$ are generated as $\mathbf{x} = \mathbf{F}\mathbf{z} + \mathbf{u}$ with independent noise $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \text{Id}_p/p)$ and

$$\mathbf{F} = \begin{cases} \begin{bmatrix} \text{Id}_d \\ \mathbf{0}_{(p-d) \times d} \end{bmatrix} & \text{if } p \geq d \\ \sqrt{\frac{d}{p}} \begin{bmatrix} \text{Id}_p & \mathbf{0}_{p \times (d-p)} \end{bmatrix} & \text{if } p < d \end{cases}. \quad (22)$$

This normalization keeps both fields dimension-free: $\text{Var}(\langle \mathbf{w}_*, \mathbf{z} \rangle) = 1$ and, conditionally on $\hat{\theta}$, $\text{Var}(\langle \hat{\theta}, \mathbf{x} \rangle) = \|\mathbf{F}^\top \hat{\theta}\|_2^2/d + \|\hat{\theta}\|_2^2/p$.

Remark 4.7. All the phenomenology that follows also holds for a random Gaussian feature matrix $F \in \mathbb{R}^{p \times d}$. The choice of feature matrix in Eq. (22) was previously considered in (Hastie et al., 2022) in the context of ridge regression. Theorem 4.8 extends the discussion to binary classification.

The main technical result for this part consists in characterizing the high-dimensional behavior of the robust empirical risk minimizer $\hat{\theta}$ as the solution of a system of self-consistent equations. For clarity of exposition we state here the result for $s = 2$ in Eqs. (13) and (14) and leave the generic statement to Section C.2.

Theorem 4.8 (Self-Consistent Equations for the Latent Space Model). *Under Theorems 3.10, 4.5 and 4.6 the values of $m = \frac{\mathbf{w}_*^\top \mathbf{F}^\top \hat{\theta}}{d}$, $\tau = \frac{\|\mathbf{F}^\top \hat{\theta}\|_2^2}{d} + \frac{\|\hat{\theta}\|_2^2}{p}$ and $P = \frac{\|\hat{\theta}\|_2^2}{p}$ 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*

$$\begin{cases} \bar{m} = \alpha \sqrt{\gamma} \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\int_{\mathbb{R}} dy \partial_{\omega} \mathcal{Z}_0 f_{\ell} \right] \\ \bar{\tau} = \alpha \gamma \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\int_{\mathbb{R}} dy \mathcal{Z}_0 f_{\ell}^2 \right] \\ \bar{V} = -\alpha \gamma \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\int_{\mathbb{R}} dy \mathcal{Z}_0 \partial_{\omega} f_{\ell} \right] \\ \bar{P} = 2r P^{1/s} \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\int_{\mathbb{R}} dy y \mathcal{Z}_0 f_{\ell} \right] \end{cases}, \quad (23)$$

that depend on the loss function ℓ through $\mathcal{Z}_0 \equiv \mathcal{Z}_0(y, m/\sqrt{\tau}\xi, 1 - m^2/\tau)$ and $f_{\ell} \equiv f_{\ell}(y, \sqrt{\tau}\xi, V, P)$, where $\mathcal{Z}_0(y, \omega, V) = \mathbb{E}_{z \sim \mathcal{N}(\omega, V)}[\mathbb{P}(y | z)]$ and $f_{\ell}(y, \omega, V, P) = (\mathcal{P}_{V\ell(y, -y r^{s^*} \sqrt{P})}(\omega) - \omega)/V$.

²Here the observed dimension is p , so Eqs. (13) and (14) are read with $\mathbf{w} \in \mathbb{R}^p$ and $R = r p^{-1/s^*}$.

The second set of equations for $\gamma \leq 1$ is

$$\begin{cases} m = \frac{\bar{m}}{2\lambda + \bar{P} + \bar{V}(1 + \frac{1}{\gamma})} \\ \tau = \frac{1}{(2\lambda + \bar{P} + \bar{V}(1 + \frac{1}{\gamma}))^2} \left[\bar{m}^2 \frac{1-\gamma}{\gamma} + \bar{\tau} \frac{(1+\gamma)^2 + \gamma - \gamma^2}{\gamma} \right] \\ V = \frac{1}{2\lambda + \bar{P} + \bar{V}(1 + \frac{1}{\gamma})} \\ P = \frac{1}{(2\lambda + \bar{P} + \bar{V}(1 + \frac{1}{\gamma}))^2} [\bar{m}^2 + 2\bar{\tau}] \end{cases} \quad (24)$$

and for $\gamma > 1$ is

$$\begin{cases} m = \frac{\bar{m}}{2\lambda + \bar{P} + 2\bar{V}} \\ \tau = \frac{\bar{m}^2 + 2\bar{\tau}}{(2\lambda + \bar{P} + 2\bar{V})^2} \end{cases} \quad \begin{cases} V = \frac{1}{2\lambda + \bar{P} + 2\bar{V}} \\ P = \frac{\bar{m}^2 + 2\bar{\tau}}{(2\lambda + \bar{P} + 2\bar{V})^2} \end{cases} \quad (25)$$

The proof (see Section C.2) uses the Gordon Min-Max theorem to characterize the minimizer of the robust risk in Eq. (13) via a low-dimensional set of self-consistent equations. Such asymptotic characterizations via implicit equation systems are standard in the analysis of high-dimensional systems (Goldt et al., 2022b; Gerace et al., 2024; Cui et al., 2023; Vilucchio et al., 2025).

With the previous theorem, we can characterize the behavior of the proper adversarial errors in this data model.

Theorem 4.9 (Proper Metrics for Latent Space Model). *Under the same setting as Theorem 4.8, with a noiseless link, the metrics defined in Eqs. (19) and (20) evaluated for $\hat{\theta}$ from Eq. (14) and decision rule $\hat{y} = \text{sign}(2f_{\hat{\theta}}(\mathbf{x}) - 1)$ concentrate to*

$$E_{\text{rob}}^{\text{cns}} = \int dp(z_1, z_2) \mathbb{I}\{\text{sign}(z_1)z_2 - \tilde{\varepsilon}\mathcal{C} < 0\}, \quad (26)$$

$$E_{\text{bnd}}^{\text{cns}} = \int dp(z_1, z_2) \mathbb{I}\{0 < \text{sign}(z_1)z_2 < \tilde{\varepsilon}\mathcal{C}\}, \quad (27)$$

where $\mathcal{C}$ is a complicated expression presented in Section C.3 and $(z_1, z_2) \sim \mathcal{N}(0, (\frac{1}{m} \ m; \tau))$ with τ, m the solutions of the system of equations in Theorem 4.8.

Remark 4.10. The difference between the two settings appears in the constant $\mathcal{C}$, which depends on the distribution of $\hat{\theta}$ through the parameters found by the equations in Theorem 4.8. Consequently, the local-fields interpretation of Theorem 3.13 carries over, with $\mathcal{C}$ now playing the role of the available attack surface $\mathcal{A}$.

4.3. The Interplay Between Overparameterization and Consistent Attacks

Theorems 4.8 and 4.9 allow us to investigate the efficacy of consistent attacks on overparameterized models. We start by considering robust training with optimally tuned λ and r . All three metrics considered in this work decrease as a function of the amount of training data α ; see Figure 5 (Left). Thus, more data are always beneficial. Interestingly, the curves cross for different overparameterization levels γ :

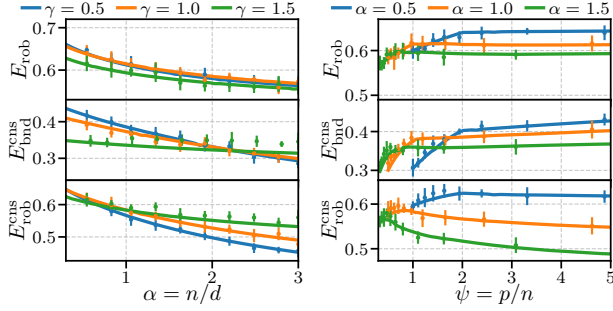


Figure 5. Error as a function of α and ψ for the latent space model defined in Section 4. For both panels the lines are the exact asymptotic solution of Eqs. (23) to (25) and the error bars are average and std over 10 realizations with $d = 500$ and p, n scaled accordingly. (Left) Robust error as a function of the number of training samples. All metrics decrease as the amount of training data increases. (Right) Robust errors as a function of the number of latent-space parameters. We see that E_{rob} and $E_{\text{bnd}}^{\text{cns}}$ increase with the number of features, while $E_{\text{rob}}^{\text{cns}}$ decreases.

no single level of overparameterization is optimal for every amount of available data.

In Figure 5 (Right) we consider the role of optimal robust training as a function of different overparameterization levels $\psi = p/n$. The three metrics present a different behavior for large ψ . The first two stay approximately constant, with at most a mild increase, while the third decreases with overparameterization, and this decrease becomes faster as more data are given to the model (larger α).

We study the behavior of the novel metrics $E_{\text{bnd}}^{\text{cns}}$ and $E_{\text{rob}}^{\text{cns}}$ in the regime of extreme overparameterization.

Proposition 4.11 (Extreme overparameterization). *Consider the setting of Theorems 4.8 and 4.9 with logistic loss, a noiseless link, and $s=q=2$. As $\psi = p/n \rightarrow \infty$ (equivalently $\gamma = d/p \rightarrow 0^+$ at fixed α), $m = m_0\sqrt{\gamma} + O(\gamma)$, $\tau = \tau_0 + O(\sqrt{\gamma})$, $P = O(\gamma)$, where m_0 and $\tau_0 > 0$ are determined by the limiting saddle-point system. Then*

$$\partial_\psi E_{\text{rob}}^{\text{cns}} = \frac{\alpha m_0}{2\pi\sqrt{\tau_0}} e^{-\tilde{\varepsilon}^2/2} \gamma^{3/2} + O(\gamma^2) > 0, \quad (28)$$

$$\partial_\psi E_{\text{bnd}}^{\text{cns}} = -\frac{\alpha m_0}{2\pi\sqrt{\tau_0}} (1 - e^{-\tilde{\varepsilon}^2/2}) \gamma^{3/2} + O(\gamma^2), \quad (29)$$

so the leading coefficient in Eq. (28) (resp. Eq. (29)) is non-negative (resp. non-positive) for $\tilde{\varepsilon}, m_0 \geq 0$.

Thus the finite-range trends in Figure 5 (Right) eventually turn over: in the extreme regime $E_{\text{rob}}^{\text{cns}}$ increases, whereas $E_{\text{bnd}}^{\text{cns}}$ decreases with the explicit coefficient in Eq. (29). This regime dependence may help explain apparently contradictory observations (Chen et al., 2024); see Section C.4 for the proof. We further verify that this dual effect is not an artifact of the linear estimator: replacing it with a trained two-layer ReLU network on the same data-generating process reproduces the same qualitative trends (Section B.3). Additional experiments in Section B.2 include ablations of other latent space models (Goldt et al., 2022a) and results

on MNIST/FashionMNIST (Deng, 2012; Xiao et al., 2017).

5. Conclusions

We investigated the fundamental distinction between consistent and inconsistent adversarial attacks in high-dimensional binary classification, introducing the consistent robust error $E_{\text{rob}}^{\text{cns}}$ and the consistent boundary error $E_{\text{bnd}}^{\text{cns}}$ as refined metrics, and deriving their exact asymptotics for robust empirical risk minimization in both the well-specified and the latent variable settings. Curiously, overparameterization has a regime-dependent effect on consistent robustness. Over the finite range in Figure 5, the boundary and overall errors follow opposite trends, while the extreme-limit expansion predicts an eventual turnover of both. This non-monotonicity may help explain contradictory empirical observations (Chen et al., 2024).

We hope these findings, though derived from exactly solvable models, prove valuable for the broader robust machine learning community. Rather than viewing overparameterization as inherently detrimental to adversarial robustness, one should account for the specific robustness objective at hand, as overparameterization can improve overall performance. Moreover, our characterizations provide theoretical guidance for selecting regularization parameters and attack budgets during robust training.

Limitations and future work. Our analysis is set in the proportional high-dimensional regime with Gaussian covariates; while this regime belongs to a broad universality class, relaxing it is an interesting direction. The latent variable model captures the random-features/NTK regime but not feature learning; extending the analysis beyond it is an important next step. Finally, target-score invariance ($g_*(\mathbf{x} + \boldsymbol{\delta}) = g_*(\mathbf{x})$) is stronger than label preservation for non-injective links, so our metrics provide conservative lower bounds on label-preserving vulnerability; extending to approximate score invariance ($|g_*(\mathbf{x} + \boldsymbol{\delta}) - g_*(\mathbf{x})| < \eta$) is a natural generalization.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback, and Fanny Yang and Antonio Ribeiro for inspiring discussions. This work was supported by the Swiss National Science Foundation (SMArtNet 212049; OperaGOST 200390) and the French ANR France 2030 program (ANR-23-IACL-0008), under the Choose France–CNRS AI Rising Talents program.

Impact Statement. This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

- Adomaityte, U., Defilippis, L., Loureiro, B., and Sicuro, G. High-dimensional robust regression under heavy-tailed data: Asymptotics and universality. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(11):114002, 2024a.
- Adomaityte, U., Sicuro, G., and Vivo, P. Classification of heavy-tailed features in high dimensions: a superstatistical approach. *Advances in Neural Information Processing Systems*, 36, 2024b.
- Aubin, B., Krzakala, F., Lu, Y., and Zdeborová, L. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12199–12210. Curran Associates, Inc., 2020.
- Awasthi, P., Frank, N., and Mohri, M. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pp. 431–441. PMLR, 2020.
- Balcan, M.-F. F., Hanneke, S., Pukdee, R., and Sharma, D. Reliable learning in challenging environments. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 48035–48050. Curran Associates, Inc., 2023.
- Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborova, L. 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>.
- Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.
- Bean, D., Bickel, P. J., El Karoui, N., and Yu, B. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013a.
- Bean, D., Bickel, P. J., Karoui, N. E., and Yu, B. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013b. doi: [10.1073/pnas.1307845110](https://doi.org/10.1073/pnas.1307845110).
- Biau, G. and Mason, D. M. High-dimensional p p-norms. *Mathematical Statistics and Limit Theorems: Festschrift in Honour of Paul Deheuvels*, pp. 21–40, 2015.
- Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum dependent learning curves in kernel regression and wide neural networks. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034. PMLR, 13–18 Jul 2020.
- Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004. doi: [10.1017/CBO9780511804441](https://doi.org/10.1017/CBO9780511804441).
- Briceño-Arias, L. M., Chierchia, G., Chouzenoux, E., and Pesquet, J.-C. A random block-coordinate douglas-rachford splitting method with low computational complexity for binary logistic regression. *Computational Optimization and Applications*, 72(3):707–726, Apr 2019. ISSN 1573-2894. doi: [10.1007/s10589-019-00060-6](https://doi.org/10.1007/s10589-019-00060-6).
- Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. Robust overfitting may be mitigated by properly learned smoothening. In *International conference on learning representations*, 2020.
- Chen, Z., Demetrio, L., Gupta, S., Feng, X., Xia, Z., Cinà, A. E., Pintor, M., Oneto, L., Demontis, A., Biggio, B., et al. Over-parameterization and adversarial robustness in neural networks: An overview and empirical analysis. *arXiv preprint arXiv:2406.10090*, 2024.
- Clarté, L., Loureiro, B., Krzakala, F., and Zdeborová, L. On double-descent in uncertainty quantification in overparametrized models. volume 206, pp. 7089–7125. PMLR Proceedings of Machine Learning Research, 2023.
- Clarysse, J., Hörmann, J., and Yang, F. Why adversarial training can hurt robust accuracy. *arXiv preprint arXiv:2203.02006*, 2022.
- Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Error scaling laws for kernel classification under source and capacity conditions. *Machine Learning: Science and Technology*, 4(3):035033, 2023.
- Dan, C., Wei, Y., and Ravikumar, P. Sharp statistical guarantees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pp. 2345–2355. PMLR, 2020.
- Dandi, Y., Stephan, L., Krzakala, F., Loureiro, B., and Zdeborová, L. Universality laws for gaussian mixtures in generalized linear models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 54754–54768. Curran Associates, Inc., 2023.

- Defilippis, L., Loureiro, B., and Misiakiewicz, T. Dimension-free deterministic equivalents and scaling laws for random feature regression. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 104630–104693. Curran Associates, Inc., 2024.
- Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.
- Dohmatob, E. Consistent adversarially robust linear classification: non-parametric setting. In *Forty-first International Conference on Machine Learning*, 2024.
- Dohmatob, E. and Scetbon, M. Robust linear regression: Phase-transitions and precise tradeoffs for general norms. *arXiv preprint arXiv:2308.00556*, 2023.
- Dohmatob, E. and Scetbon, M. Precise accuracy/robustness tradeoffs in regression: Case of general norms. In *Forty-first International Conference on Machine Learning*, 2024.
- Donhauser, K., Tifrea, A., Aerni, M., Heckel, R., and Yang, F. Interpolation can hurt robust generalization even when there is no noise. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23465–23477. Curran Associates, Inc., 2021.
- Donoho, D. and Tanner, J. 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.
- Dudeja, R., M. Lu, Y., and Sen, S. Universality of approximate message passing with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683, 2023.
- Fort, S. and Lakshminarayanan, B. Ensemble everything everywhere: Multi-scale aggregation for adversarial robustness. *arXiv preprint arXiv:2408.05446*, 2024.
- Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR, 2020.
- Gerace, F., Krzakala, F., Loureiro, B., Stephan, L., and Zdeborová, L. Gaussian universality of perceptrons with random labels. *Phys. Rev. E*, 109:034305, Mar 2024. doi: 10.1103/PhysRevE.109.034305.
- Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pp. 426–471. PMLR, 2022a.
- Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mezard, M., and Zdeborova, L. The gaussian equivalence of generative models for learning with shallow neural networks. In Bruna, J., Hesthaven, J., and Zdeborova, L. (eds.), *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pp. 426–471. PMLR, 16–19 Aug 2022b.
- Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- Gordon, Y. On milman’s inequality and random subspaces which escape through a mesh in R^n . In Lindenstrauss, J. and Milman, V. D. (eds.), *Geometric Aspects of Functional Analysis*, pp. 84–106, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg. ISBN 978-3-540-39235-4.
- Hassani, H. and Javanmard, A. The curse of over-parametrization in adversarial training: Precise analysis of robust generalization for random features regression. *The Annals of Statistics*, 52(2):441–465, 2024.
- Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.
- Hu, H. and Lu, Y. M. 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.
- Isserlis, L. 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.
- Javanmard, A. and Soltanolkotabi, M. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, 2022.
- Javanmard, A., Soltanolkotabi, M., and Hassani, H. Precise tradeoffs in adversarial training for linear regression. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2034–2078. PMLR, 09–12 Jul 2020.
- Krogh, A. and Hertz, J. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

- Lee, H. and Chung, H. W. Toward understanding adversarial distillation: Why robust teachers fail. *arXiv preprint arXiv:2605.21999*, 2026.
- Liao, Z., Couillet, R., and Mahoney, M. W. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13939–13950. Curran Associates, Inc., 2020.
- Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, 2018.
- Mai, X., Liao, Z., and Couillet, R. 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)*, pp. 3357–3361, 2019. doi: 10.1109/ICASSP.2019.8683376.
- Mei, S. and Montanari, A. 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.
- Mezard, M., Parisi, G., and Virasoro, M. *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.
- Mignacco, F., Krzakala, F., Lu, Y., Urbani, P., and Zdeborova, L. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International conference on machine learning*, pp. 6874–6883. PMLR, 2020.
- Montanari, A. and Saeed, B. N. Universality of empirical risk minimization. In Loh, P.-L. and Raginsky, M. (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4310–4312. PMLR, 02–05 Jul 2022.
- Okajima, K., Meng, X., Takahashi, T., and Kabashima, Y. Average case analysis of lasso under ultra sparse conditions. In Ruiz, F., Dy, J., and van de Meent, J.-W. (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11317–11330. PMLR, 25–27 Apr 2023.
- Parikh, N. and Boyd, S. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, jan 2014. ISSN 2167-3888. doi: 10.1561/24000000003.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Peng, J. and Yang, Y. On damage of interpolation to adversarial robustness in regression. *arXiv preprint arXiv:2601.16070*, 2026.
- Raghuathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7909–7919. PMLR, 13–18 Jul 2020.
- Ribeiro, A., Zachariah, D., Bach, F., and Schön, T. Regularization properties of adversarially-trained linear regression. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 23658–23670. Curran Associates, Inc., 2023.
- Ribeiro, A. H. and Schön, T. B. Overparameterized linear regression under adversarial attacks. *IEEE Transactions on Signal Processing*, 71:601–614, 2023.
- Ribeiro, A. H., Schön, T. B., Zachariah, D., and Bach, F. Efficient optimization algorithms for linear adversarial training. In Li, Y., Mandt, S., Agrawal, S., and Khan, E. (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1207–1215. PMLR, 03–05 May 2025.
- Rice, L., Wong, E., and Kolter, J. Z. 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*, pp. 8093–8104. PMLR, 2020.
- Schröder, D., Cui, H., Dmitriev, D., and Loureiro, B. Deterministic equivalent and error universality of deep random features learning. In Krause, A., Brunskill, E., Cho, K.,

- Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30285–30320. PMLR, 23–29 Jul 2023.
- Schröder, D., Dmitriev, D., Cui, H., and Loureiro, B. Asymptotics of learning with deep structured (Random) features. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43862–43894. PMLR, 21–27 Jul 2024.
- Seung, H. S., Sompolinsky, H., and Tishby, N. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45: 6056–6091, Apr 1992.
- Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J. P., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! In Wallach, H. M., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3353–3364, 2019.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- Tabanelli, H., Mergny, P., Zdeborova, L., and Krzakala, F. Computational thresholds in multi-modal learning via the spiked matrix-tensor model. *arXiv preprint arXiv:2506.02664*, 2025.
- Taheri, H., Pedarsani, R., and Thrampoulidis, C. Asymptotic behavior of adversarial training in binary linear classification. *IEEE Trans. Neural Netw. Learn. Syst.*, PP, July 2023.
- Tanner, K., Vilucchio, M., Loureiro, B., and Krzakala, F. A high dimensional statistical model for adversarial training: Geometry and trade-offs. In Li, Y., Mandt, S., Agrawal, S., and Khan, E. (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2530–2538. PMLR, 03–05 May 2025.
- Tao, T. and Vu, V. Random matrices: Universality of local eigenvalue statistics up to the edge. *Communications in Mathematical Physics*, 298:549–572, 2010.
- Thrampoulidis, C., Oymak, S., and Hassibi, B. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.
- Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized Linear Regression: A precise analysis of the estimation error. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1683–1709, Paris, France, 03–06 Jul 2015. PMLR.
- Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.
- Tsilivis, N., Frank, N., Srebro, N., and Kempe, J. The price of implicit bias in adversarially robust generalization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 58023–58057, 2024.
- Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Vilucchio, M., Troiani, E., Erba, V., and Krzakala, F. Asymptotic characterisation of the performance of robust linear regression in the presence of outliers. In *International Conference on Artificial Intelligence and Statistics*, pp. 811–819. PMLR, 2024.
- Vilucchio, M., Dandi, Y., Rossignol, M. P., Gerbelot, C., and Krzakala, F. Asymptotics of non-convex generalized linear models in high-dimensions: A proof of the replica formula. *arXiv preprint arXiv:2502.20003*, 2025.
- Vilucchio, M., Tsilivis, N., Loureiro, B., and Kempe, J. On the geometry of regularization in adversarial training: High-dimensional asymptotics and generalization bounds. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
- Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. Better diffusion models further improve adversarial training. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B.,

- Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36246–36263. PMLR, 2023.
- Wei, A., Hu, W., and Steinhardt, J. More than a toy: Random matrix models predict how real-world neural representations generalize. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23549–23588. PMLR, 17–23 Jul 2022.
- Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- Xiao, L., Hu, H., Misiakiewicz, T., Lu, Y., and Pennington, J. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35:4558–4570, 2022.
- Xing, Y., Aghasi, A., and Ghadimi, S. High-dimensional regression with missing data: An asymptotic study. *Stat*, 14(4):e70125, 2025. doi: 10.1002/sta4.70125.
- Yin, D., Ramchandran, K., and Bartlett, P. L. Rademacher complexity for adversarially robust generalization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *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*, pp. 7085–7094. PMLR, 2019.
- Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J., and Wang, L. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

A. 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 (Thrapoulidis et al., 2014; 2015; Vilucchio et al., 2025), statistical physics approaches (Mignacco et al., 2020; Gerace et al., 2020; Bordelon et al., 2020; Loureiro et al., 2021; Okajima et al., 2023; Adomaityte et al., 2024b;a), and random matrix theory (Bean et al., 2013b; Mai et al., 2019; Liao et al., 2020; Mei & Montanari, 2022; Xiao et al., 2022; Schröder et al., 2023; 2024; Defilippis et al., 2024; Tabanelli et al., 2025). Our work is particularly motivated by recent advances in Gaussian universality (Goldt et al., 2022a; Montanari & Saeed, 2022; Dandi et al., 2023), 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 (Tao & Vu, 2010; Donoho & Tanner, 2009; Wei et al., 2022; Dudeja et al., 2023). Our technical contribution extends these frameworks to characterize consistent adversarial attacks: while existing asymptotic analyses focus on standard generalization or robust error, we derive the first closed-form high-dimensional limits for metrics that isolate label-preserving perturbations, revealing geometric dependencies on $\sqrt{\tau - m^2}$ rather than $\sqrt{\tau}$ alone.

Adversarial Robustness: Robust empirical risk minimization, commonly known as adversarial training, was pioneered in computer vision (Goodfellow et al., 2015; Madry et al., 2018) and has since evolved into a primary defense against adversarial attacks. Researchers have developed numerous approaches to improve its computational efficiency (Shafahi et al., 2019; Rice et al., 2020) and statistical properties (Zhai et al., 2019; Chen et al., 2020; Wang et al., 2023). On the theoretical front, several works have investigated the properties of robust empirical risk minimization for linear models (Raghuathan et al., 2020; Dan et al., 2020; Ribeiro et al., 2023; 2025), including sharp proportional asymptotics under different data designs (Javanmard et al., 2020; Hassani & Javanmard, 2024; Donhauser et al., 2021; Dohmatob & Scetbon, 2024; Ribeiro & Schön, 2023; Tanner et al., 2025; Peng & Yang, 2026). A common thread in most of the previous analyses is the characterization of the standard robust error E_{rob} , treating all perturbations in $B_q(\varepsilon)$ uniformly without distinguishing whether they preserve or alter the ground-truth label. For instance, Taheri et al. (2023) and Javanmard & Soltanolkotabi (2022) derive sharp asymptotics for E_{rob} under adversarial training in well-specified linear models, while Clarysse et al. (2022) study adversarial perturbations directed along the signal in a noiseless setting, and Tanner et al. (2025) extend the analysis to Mahalanobis-norm attacks. In high-dimensional linear regression with missing data, Xing et al. (2025) analyze ridge and ℓ_2 adversarial training and develop a fast leave-one-out cross-validation approximation. Relatedly, Tsilivis et al. (2024) study how the implicit bias induced by optimization algorithms and architectures in robust ERM affects standard robust generalization. Their focus is on matching implicit regularization to the threat model, whereas we isolate label-preserving attacks and characterize their high-dimensional behavior. Our work departs from this line by considering the consistency constraint $\langle \mathbf{w}_*, \boldsymbol{\delta} \rangle = 0$, which requires a fundamentally different geometric analysis: the existence and strength of consistent attacks depend on the predictor component orthogonal to the target (Theorem 3.2 and Theorem C.10), a quantity that plays no role in unconstrained robust error. Of particular relevance is Vilucchio et al. (2026), which derives high-dimensional asymptotics for binary classification in the well-specified model and which we build upon in Section 3. Our contributions beyond their work are the geometric existence conditions (Theorem 3.2), the entire latent variable model analysis (Section 4), generalizing previous results (Mei & Montanari, 2022; Goldt et al., 2022a), and a characterization of how feature dimension p affects consistent robustness differently than standard robust error (Section 4.3). Similarly, Donhauser et al. (2021) study the robust error of non-robustly trained ℓ^2 -regularized estimators in well-specified models, but do not consider adversarial training or the latent variable model. Beyond the proportional regime, Lee & Chung (2026) draw a related distinction in a feature-learning analysis of adversarial distillation, separating failures on intrinsically hard-to-learn samples from robustness loss on otherwise correctly classified points, mirroring our distinction between failures due to label-preserving perturbations and those due to limited expressivity.

Consistent Attacks: 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 (Szegedy et al., 2014; Goodfellow et al., 2015; Fort & Lakshminarayanan, 2024). The notion of a consistent attack in the theoretical literature was explored in Raghuathan et al. (2020); Donhauser et al. (2021); Balcan et al. (2023). These works establish that proper adversarial examples—perturbations preserving the Bayes-optimal label—can hurt robust generalization and that interpolating models may be vulnerable even without label noise. However, they do not provide: (i) geometric characterizations of when consistent attacks *exist* in high dimensions as a function of model-target alignment and attack geometry (our Theorems 3.2, 3.4, 4.1 and 4.3); (ii) separate quantification of vulnerability on correctly vs. incorrectly classified points (our $E_{\text{bnd}}^{\text{cns}}$ metric, Eq. (3)); or (iii) exact asymptotic

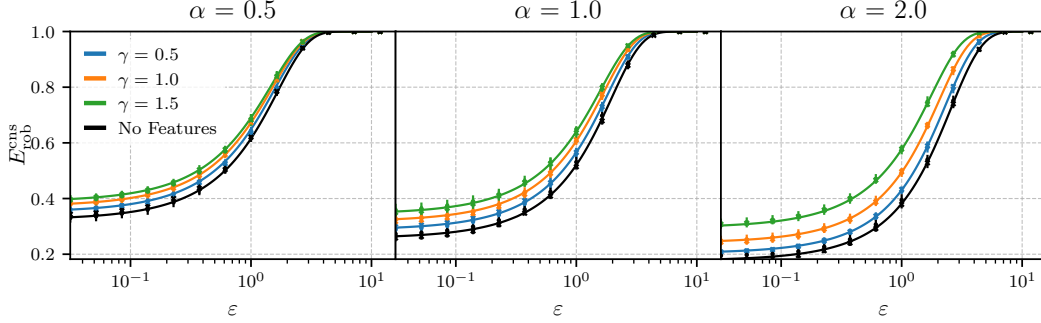


Figure 6. Behavior of $E_{\text{rob}}^{\text{cns}}$ for Gaussian features as a function of the attack strength. The results use a model trained as in Eqs. (13) and (14), with α and γ specified in the figure. The error bars refer to 10 repetitions of the experiments for $d = 1024$. The metrics consider the $\hat{\mathbf{w}}$ trained with $\lambda = 10^{-3}$, $r = 0.0$ and $s = \infty$.

formulas under Gaussian designs revealing how overparameterization affects consistent robustness (our Theorems 3.12, 4.8 and 4.9). Our metrics thus enable fine-grained analysis distinguishing model failures due to label-preserving perturbations from those due to limited expressivity or sample size, providing precise predictions for the vulnerability landscape in high-dimensional linear classification.

B. Additional Experimental Details

B.1. Setting of the Figures in the Main Text

We jointly optimize the dimension-free training budget $r \geq 0$ and ridge parameter $\lambda \geq 0$ using the asymptotic theory, which gives a deterministic objective for each pair (r, λ) . We use the gradient-free Nelder–Mead algorithm implemented in `scipy.optimize.minimize`.

Figure 3 (Left) For each value of q , the empirical points are obtained by generating $n = 1000$ samples in $d = 10$ dimensions, with the model parameter set to $m = 0.5$. The probability is the fraction of samples for which an adversarial perturbation of norm ε exists, while the solid theory lines are computed using Eq. (12).

Figure 3 (Center) Now we fix $q = 2$ and vary the data dimension d . The empirical points are again obtained by generating $n = 1000$ samples for each d , with $m = 0.5$; the probabilities and theoretical lines are computed as before.

Figure 3 (Right) We consider Gaussian covariates with diagonal covariance $\Sigma = \text{diag}(\lambda_1, \dots, \lambda_d)$ and power-law spectrum $\lambda_i \propto i^{-\beta}$, normalized so that $\sum_{i=1}^d \lambda_i = d$, and sweep β . We fix the target $\mathbf{w}_* = \sqrt{d} \mathbf{e}_1$ (top eigenvector) and correlation $m = 0.5$. For each β , we construct two predictors: low-rank alignment with $\hat{\mathbf{w}}_{\perp}$ supported on the modes from 2 to $k+1$ and tail alignment with support on modes $k+2:d$. We set $\varepsilon = 1$, $q = 2$, $k = 10$, and evaluate $d \in \{16, 128, 1024\}$. Solid curves plot Eq. (12); empirical points report the fraction of $n = 10^4$ samples from $\mathcal{N}(\mathbf{0}, \Sigma)$ satisfying Eq. (7).

Figure 4 (Left) The curves use $\mathbf{w}$ obtained from standard training, i.e. minimization of Eq. (14) with $r = 0$, $\lambda = 10^{-3}$, and training norm $s = 2$. The sample complexity is fixed at $\alpha = 1$. The points are produced as 10 different realizations with $d = 500$ fixed.

Figure 4 (Right) Here we show performance under evaluation attacks constrained in either ℓ^∞ or ℓ^2 . In both cases, r and λ are jointly tuned to minimize the corresponding error; the regularizer is ridge and the adversarial-training norm is $s = 2$. The points are produced as 10 different realizations with $d = 500$ fixed.

Figure 5 (Left, Right) We consider robust ERM with training norm $s = \infty$, jointly tuning r and λ for each setting. The points are produced from 10 realizations with $d = 500$ and n, p scaled accordingly.

B.2. Additional Experiments

Alternative latent-space model To test whether our findings depend on the feature map in Theorem 4.6, which generates the observed input $\mathbf{x}$ from the latent variable $\mathbf{z}$, we consider an alternative latent-space model. Another model used to characterize overparameterization is the hidden manifold model (Goldt et al., 2022b; Gerace et al., 2020), where the latent-space covariates are Gaussian, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \text{Id}_d)$, and the feature-space covariates are a linear transformation $\mathbf{x} = \mathbf{F} \mathbf{z}$ with independent entries $F_{ij} \sim \mathcal{N}(0, 1)$.

We explore the behavior of the error metrics defined in Eqs. (19) and (20) in Figures 6 and 7. We see that the behavior is

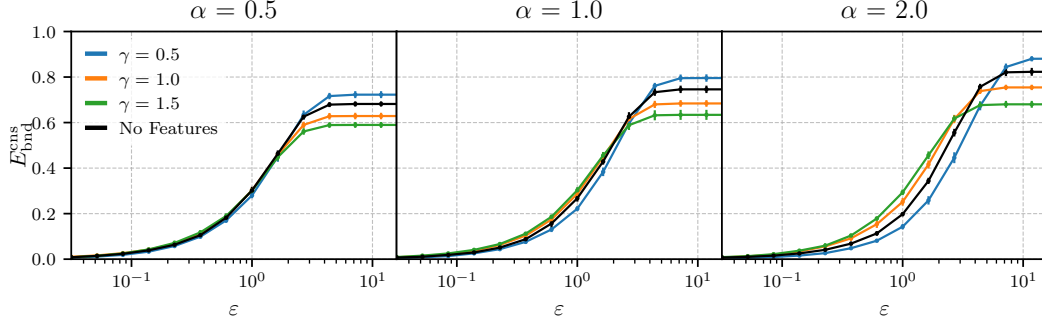


Figure 7. Behavior of $E_{\text{bnd}}^{\text{cns}}$ for Gaussian features as a function of the attack strength. The results use a model trained as in Eqs. (13) and (14), with α and γ specified in the figure. The error bars refer to 10 repetitions with $d = 1024$. The metrics consider the $\hat{\mathbf{w}}$ trained with $\lambda = 10^{-3}$, $r = 0.0$ and $s = \infty$.

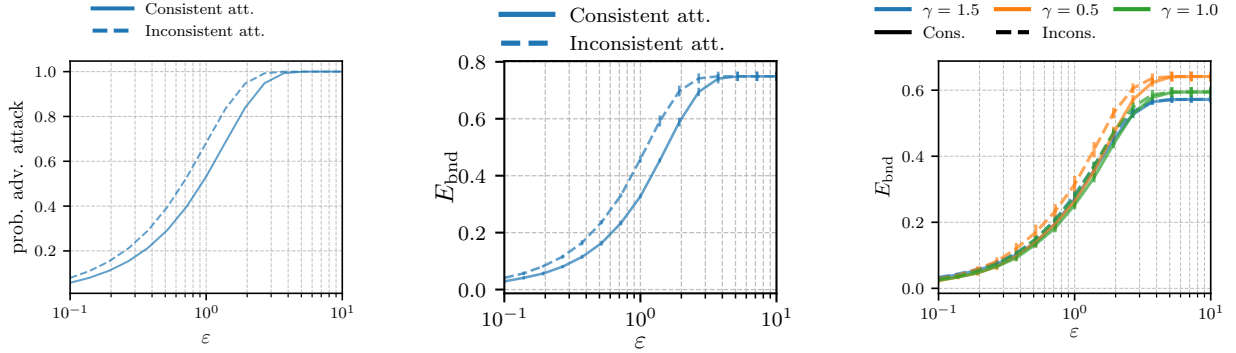


Figure 8. Comparison of consistent and inconsistent cases as a function of the attack strength. (Left, Center) Case of the well-specified model presented in Section 3. We show both the probability of existence as per Theorem 3.2 and the value of the boundary error. The error bars refer to 10 repetitions of the experiments for $d = 500$. The metrics consider the $\hat{\mathbf{w}}$ 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 $\hat{\mathbf{w}}$ trained with $\alpha = 1.0$, $\lambda = 10^{-2}$, $r = 0.0$ and $s = 2$.

similar to the one of the model defined in the main text. The black line in the same figure shows the performance of the well-specified model from the main text. Crucially we see also in this case the metric $E_{\text{rob}}^{\text{cns}}$ equals the value of the clean generalization error in the $\epsilon \rightarrow 0^+$ limit and it reaches one in the $\epsilon \rightarrow \infty$ limit. We have that $E_{\text{bnd}}^{\text{cns}}$ is zero in the $\epsilon \rightarrow 0^+$ limit.

Errors as a function of attack strength We additionally compare the consistent and inconsistent formulation of the boundary error as a function of the attack strength in Figure 8. The consistent boundary error is defined in Eq. (3) while the inconsistent one is defined as

$$E_{\text{bnd}} = \mathbb{E}_{(\mathbf{x}, y)} \left[\max_{\delta \in B_q(\epsilon)} \mathbb{I} \left\{ \begin{array}{l} y \neq \hat{y}(\mathbf{x} + \delta) \\ y = \hat{y}(\mathbf{x}) \end{array} \right\} \right]. \quad (30)$$

where the crucial difference is the removal of the target-score invariance condition $g_*(\mathbf{x}) = g_*(\mathbf{x} + \delta)$. We have that also in this case consistent attacks produce a milder increase in the boundary error but the qualitative behavior is the same.

In Figure 9 we show the behavior of the different attacks considered in the main text for the latent space model. The qualitative behavior of the different models (different choice of γ) is the same. As in the well-specified model, under the scaling of ϵ in Theorem 3.10, the ℓ_2 -bounded attack is more effective at increasing the error than the ℓ_∞ -attack.

Experiments on MNIST (Deng, 2012) and Fashion-MNIST (Xiao et al., 2017) datasets In Figure 10 we show the empirical results on MNIST binary classification (digits 0 vs 1) and the Fashion-MNIST dataset for binary classification (Sandals vs Ankle Boots) using a teacher-student random features model. We load MNIST images, center them by subtracting the dataset mean, and denote the centered projection images as $\tilde{\mathbf{x}}_\mu \in \mathbb{R}^{784}$. In the complex Random Fourier Features (RFF) setting, we generate a random Gaussian projection matrix $\Omega \in \mathbb{R}^{D_{\text{rff}} \times 784}$ with $D_{\text{rff}} = 392$ and construct latent

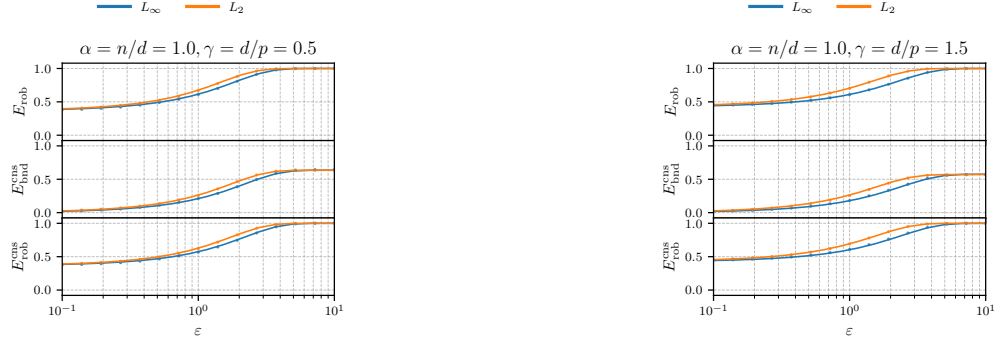


Figure 9. Behavior of the different adversarial errors for the Latent Space Model of Section 4. Here we show the behavior of the errors considered in the main text for different choices of $\gamma = d/p$ with $\gamma = 0.5$ for (Left) and $\gamma = 1.5$ for (Right). We see that qualitatively the behaviors are the same in the two regimes.

features $z_\mu = [\text{Re}(\exp(-i\Omega\tilde{x}_\mu)); \text{Im}(\exp(-i\Omega\tilde{x}_\mu))] \in \mathbb{R}^{2D_{\text{eff}}}$, so that the latent dimension is $d = 2D_{\text{eff}} = 784$. A teacher classifier $w_\star \in \mathbb{R}^d$ is trained via LinearSVC (no intercept) (Pedregosa et al., 2011) on these latent features, and labels are generated through a noiseless channel. The student observes transformed features $x_\mu = Fz_\mu + u_\mu$ as per Theorem 4.6. We train the student via ℓ^∞ ($s = \infty$) adversarial training Eqs. (13) and (14) with ridge regularization $\lambda = 10^{-2}$ and training attack budget $r = 1.0$. We evaluate the consistent adversarial metrics defined in Theorem 2.3: consistent robust error $E_{\text{rob}}^{\text{cns}}$, and consistent boundary error $E_{\text{bnd}}^{\text{cns}}$ using $n_{\text{gen}} = 1000$ test samples with attack budget $\varepsilon = 1.0$. Results are averaged over 10 random seed repetitions with $d = 784$ fixed, varying the sample complexity $\alpha = n/d$ and overparameterization level $\gamma = d/p$.

The results are presented in Figure 10. The qualitative behavior is consistent with the Gaussian theory in Sections 4 and 4.3. The consistent boundary error $E_{\text{bnd}}^{\text{cns}}$ increases with overparameterization, while the consistent robust error $E_{\text{rob}}^{\text{cns}}$ increases or decreases depending on the number of data points (the value of $\alpha = n/d$).

B.3. Neural network experiment

To test whether the dual role of overparameterization predicted by our theory (Theorem 4.11) persists beyond the linear estimator, we repeat the analysis replacing the linear model by a nonlinear neural network while keeping the data-generating process identical to Theorem 4.6. This isolates the effect of the model class: the latent ground truth $w_\star$, the labels, and the attack geometry are unchanged, and only the estimator differs.

The estimator is a two-layer ReLU multilayer perceptron with hidden dimension 256, trained to minimize the binary cross-entropy loss with Adam (learning rate 10^{-3} , weight decay 10^{-4}) and early stopping (patience 50 on a 20% validation split). We fix the latent dimension $d = 1000$ and vary the feature dimension p from 1000 to 16000, so that the overparameterization ratio $\psi = p/n$ spans more than one order of magnitude. Adversarial perturbations are computed in the latent space with

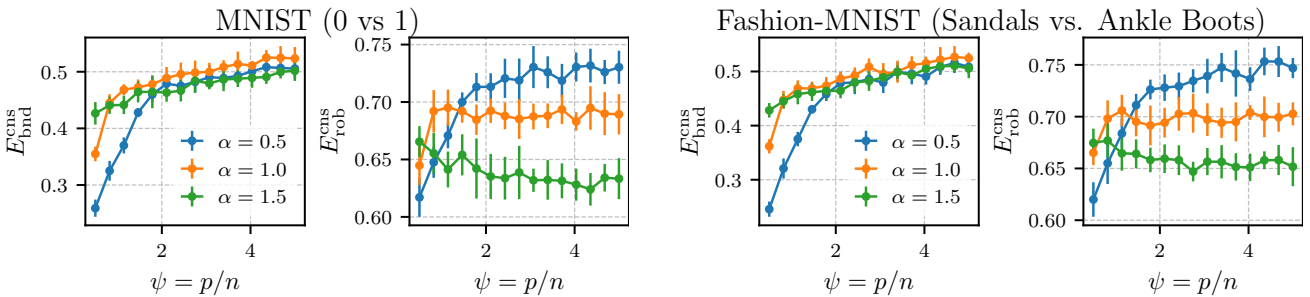


Figure 10. Empirical validation of the latent variable model predictions on real image datasets using complex Random Fourier Features. (Left) Results on MNIST binary classification (digits 0 vs 1) with $d = 784$ features. (Right) Results on Fashion-MNIST binary classification (Sandals vs. Ankle Boots) with $d = 784$ features. Points represent empirical averages over 10 repetitions with error bars indicating standard deviation. The qualitative behavior matches the one seen for the Gaussian model in Figure 5.

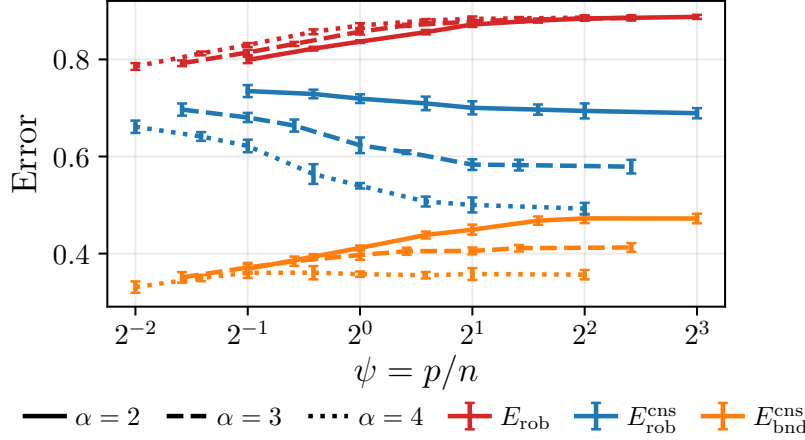


Figure 11. Consistent metrics for a two-layer ReLU MLP trained on the latent variable model of Theorem 4.6, as a function of the overparameterization $\psi = p/n$ ($d = 1000$, ℓ_2 attack with $\varepsilon = 0.05$), for three sample complexities $\alpha = n/d \in \{2, 3, 4\}$. Replacing the linear estimator by a nonlinear network preserves the dual behavior predicted by the theory: E_{rob} and $E_{\text{bnd}}^{\text{cns}}$ increase with ψ , while $E_{\text{rob}}^{\text{cns}}$ decreases. Averages over 10 seeds; error bars are one standard deviation.

ℓ_2 -PGD (50 steps, 5 random restarts, $\varepsilon = 0.05$), enforcing the consistency constraint $\langle \mathbf{w}_*, \delta \rangle = 0$ by projection at each step. All metrics are evaluated on a held-out test set and averaged over 10 independent seeds.

Figure 11 reports the three metrics as a function of p . The qualitative predictions of the linear theory carry over to the nonlinear estimator: as overparameterization increases, the standard robust error E_{rob} and the consistent boundary error $E_{\text{bnd}}^{\text{cns}}$ grow, while the overall consistent robust error $E_{\text{rob}}^{\text{cns}}$ decreases, driven by the improvement in clean generalization. This confirms that the dual effect of overparameterization is not an artifact of the linear model class.

C. Proofs of the Results in the Main Text

The proofs in this appendix are organized according to the structure of the main results. We first present the finite-dimensional proofs in Section C.1, next, in Section C.2, we provide proofs for the high-dimensional asymptotic results, including the derivation of the self-consistent equations stated in Theorem 4.8. Finally, in Section C.3 we detail the derivation of the limiting error metrics for both the well-specified and latent-space models.

C.1. Proofs for the finite-dimensional results

In this section we gather the proofs of the finite-dimensional results presented in the main text.

Proof of Theorem 3.2. As discussed above, a consistent attack must satisfy the three conditions in Eqs. (5) and (6). First, note that this is only possible if $\hat{\mathbf{w}}_{\perp} \neq 0$, otherwise any admissible perturbation would *a fortiori* violate Eq. (6). Therefore, from now on we assume $\hat{\mathbf{w}}_{\perp} \neq 0$. Consider an admissible attack $\delta \in H_q(\varepsilon)$. Since $-\delta \in H_q(\varepsilon)$, we can always fix the sign of δ to satisfy the constraint $\text{sign}(\langle \hat{\mathbf{w}}, \delta \rangle) = -\text{sign}(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle)$. The restrictive condition is the margin $|\langle \hat{\mathbf{w}}, \delta \rangle| > |\langle \hat{\mathbf{w}}, \mathbf{x} \rangle|$. Since $\langle \delta, \mathbf{w}_* \rangle = 0$, we have $\langle \hat{\mathbf{w}}, \delta \rangle = \langle \hat{\mathbf{w}}_{\perp}, \delta \rangle$, and hence the margin condition is equivalent to $|\langle \hat{\mathbf{w}}_{\perp}, \delta \rangle| > |\langle \hat{\mathbf{w}}, \mathbf{x} \rangle|$. This condition is satisfied if and only if it is satisfied by the supremum:

$$\sup_{\delta \in H_q(\varepsilon)} |\langle \hat{\mathbf{w}}_{\perp}, \delta \rangle| > |\langle \hat{\mathbf{w}}, \mathbf{x} \rangle|. \quad (31)$$

Standard results from convex analysis imply that this supremum equals ε times the ℓ^{q^*} -distance from $\hat{\mathbf{w}}_{\perp}$ to $\text{span}(\mathbf{w}_*)$ (the metric projector):

$$\sup_{\delta \in H_q(\varepsilon)} |\langle \hat{\mathbf{w}}_{\perp}, \delta \rangle| = \varepsilon \inf_{\mu \in \mathbb{R}} \|\hat{\mathbf{w}}_{\perp} - \mu \mathbf{w}_*\|_{q^*} =: \varepsilon d_{q^*}^*(\hat{\mathbf{w}}_{\perp}), \quad (32)$$

where q^* is the Hölder conjugate of q : $1/q + 1/q^* = 1$. □

Proof of Theorem 3.4. The existence part follows the same proof as in Theorem 3.2, but without the orthogonality constraint. We then have:

$$\sup_{\delta \in B_q(\varepsilon)} |\langle \hat{\mathbf{w}}, \delta \rangle| = \varepsilon \|\hat{\mathbf{w}}\|_{q^*}. \quad (33)$$

Finally, $H_q(\varepsilon) \subseteq B_q(\varepsilon)$, so pointwise maximization over the two feasible sets gives $E_{\text{rob}}^{\text{cns}} \leq E_{\text{rob}}$. $\square$

Proof of Theorem 3.5. Conditionally on $\hat{\mathbf{w}}$, we have $\langle \hat{\mathbf{w}}, \mathbf{x} \rangle \sim \mathcal{N}(0, \hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}})$, so $\langle \hat{\mathbf{w}}, \mathbf{x} \rangle \stackrel{d}{=} \sqrt{\hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}}} \cdot Z$ with $Z \sim \mathcal{N}(0, 1)$. The condition $\varepsilon d_{q^*}^*(\hat{\mathbf{w}}_\perp) \geq |\langle \hat{\mathbf{w}}, \mathbf{x} \rangle|$ from Equation (7) yields $\mathbb{P}[|Z| \leq \varepsilon d_{q^*}^*(\hat{\mathbf{w}}_\perp) / \sqrt{\hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}}}] = 2\Phi(\varepsilon d_{q^*}^*(\hat{\mathbf{w}}_\perp) / \sqrt{\hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}}}) - 1$. We next prove the error comparison, still at fixed dimension and conditionally on $\hat{\mathbf{w}}$. Write $b = \|\hat{\mathbf{w}}\|_{q^*}$, $a = d_{q^*}^*(\hat{\mathbf{w}}_\perp)$, and $\rho = a/b \in [0, 1]$. Let Z_1 and Z_2 be the standardized target and model fields, respectively, with correlation $c \in [-1, 1]$. Under the noiseless link the standardized signed margin is $M = \text{sign}(Z_1)Z_2$. If F_c denotes its c.d.f., the support functions of $H_q(\varepsilon)$ and $B_q(\varepsilon)$ give

$$E_{\text{rob}}^{\text{cns}}(\varepsilon) = F_c\left(\frac{\varepsilon a}{\sqrt{\hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}}}}\right) = E_{\text{rob}}(\rho\varepsilon), \quad E_{\text{rob}}(\varepsilon) = F_c\left(\frac{\varepsilon b}{\sqrt{\hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}}}}\right). \quad (34)$$

It remains to compare the c.d.f. at the two radii. Let ϕ and Φ denote the standard normal density and c.d.f. For $t \geq 0$,

$$F_c(t) = 2\mathbb{P}(Z_1 > 0, Z_2 \leq t), \quad f_c(t) = F'_c(t) = 2\phi(t)\Phi\left(\frac{ct}{\sqrt{1-c^2}}\right) \leq 2\phi(t),$$

where the cases $c = \pm 1$ follow by continuity. If $\Phi_2(\cdot, \cdot; c)$ is the standard bivariate Gaussian c.d.f. with correlation c , then $F_c(t) = 2[\Phi(t) - \Phi_2(0, t; c)]$. Since $\partial_c \Phi_2(0, t; c)$ is the positive bivariate Gaussian density at $(0, t)$, $F_c(t)$ is non-increasing in c . Hence $F_c(t) \geq F_1(t) = 2\Phi(t) - 1$. Consequently,

$$F_c(t) - tf_c(t) \geq 2\Phi(t) - 1 - 2t\phi(t) \geq 0,$$

because the last expression vanishes at $t = 0$ and has derivative $2t^2\phi(t) \geq 0$. Thus $t \mapsto F_c(t)/t$ is non-increasing on $(0, \infty)$, and $F_c(\rho t) \geq \rho F_c(t)$. Combining this with Eq. (34) and monotonicity of F_c proves $\rho E_{\text{rob}}(\varepsilon) \leq E_{\text{rob}}^{\text{cns}}(\varepsilon) \leq E_{\text{rob}}(\varepsilon)$. The case $\varepsilon = 0$ follows directly. $\square$

Proof of Theorem 3.8. Assume $q = 2$ so that $q^* = 2$ and $d_{q^*}^*(\cdot) = \|\cdot\|_2$. Consider the power-law spectrum

$$\lambda_i = \frac{d i^{-\beta}}{Z_\beta(d)}, \quad Z_\beta(d) = \sum_{j=1}^d j^{-\beta}, \quad (\beta > 0),$$

with orthonormal eigenbasis $\{v_i\}_{i=1}^d$, and let $\mathbf{w}_* = \sqrt{d}v_1$. Fix a correlation level $m \in [0, 1]$ and consider predictors of the form

$$\hat{\mathbf{w}} = m\sqrt{d}\mathbf{v}_1 + \hat{\mathbf{w}}_\perp, \quad \|\hat{\mathbf{w}}_\perp\|_2^2 = d(1 - m^2).$$

Write $\hat{\mathbf{w}}_\perp = \sum_{i=2}^d c_i v_i$ with $\sum_{i=2}^d c_i^2 = d(1 - m^2)$. The Gaussian logit appearing in Eq. (12) equals

$$\gamma(\hat{\mathbf{w}}) := \frac{\varepsilon \|\hat{\mathbf{w}}_\perp\|_2}{\sqrt{\hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}}}} = \frac{\varepsilon \sqrt{d(1 - m^2)}}{\sqrt{\hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}}}}.$$

Thus, for $q = 2$, comparing predictors reduces to comparing the denominators $\sqrt{\hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}}}$.

Using $\Sigma v_i = \lambda_i v_i$ and orthonormality,

$$\hat{\mathbf{w}}^\top \Sigma \hat{\mathbf{w}} = m^2 d \lambda_1 + \sum_{i=2}^d c_i^2 \lambda_i = \frac{d^2}{Z_\beta(d)} \left(m^2 + (1 - m^2) \sum_{i=2}^d p_i i^{-\beta} \right), \quad (35)$$

where we introduced the weights

$$p_i := \frac{c_i^2}{d(1-m^2)} \geq 0, \quad \sum_{i=2}^d p_i = 1.$$

Fix $k \in \{1, \dots, d-2\}$ and define two predictors with the same m and the same $\|\hat{\mathbf{w}}_{\perp}\|_2^2 = d(1-m^2)$ but different allocation of $\hat{\mathbf{w}}_{\perp}$:

$$\hat{\mathbf{w}}_{\perp}^{\text{LR}} := \sqrt{\frac{d(1-m^2)}{k}} \sum_{i=2}^{k+1} \mathbf{v}_i, \quad \hat{\mathbf{w}}_{\perp}^{\text{Tail}} := \sqrt{\frac{d(1-m^2)}{d-k-1}} \sum_{i=k+2}^d \mathbf{v}_i.$$

Equivalently, p_i is uniform on $S_{\text{LR}} = \{2, \dots, k+1\}$ (resp. on $S_{\text{Tail}} = \{k+2, \dots, d\}$). Define the corresponding averages

$$\mu_{\text{LR}} := \frac{1}{k} \sum_{i=2}^{k+1} i^{-\beta}, \quad \mu_{\text{Tail}} := \frac{1}{d-k-1} \sum_{i=k+2}^d i^{-\beta}.$$

Then Eq. (35) yields

$$\hat{\mathbf{w}}^{\top} \Sigma \hat{\mathbf{w}} \Big|_{\text{LR}} = \frac{d^2}{Z_{\beta}(d)} \left(m^2 + (1-m^2) \mu_{\text{LR}} \right), \quad \hat{\mathbf{w}}^{\top} \Sigma \hat{\mathbf{w}} \Big|_{\text{Tail}} = \frac{d^2}{Z_{\beta}(d)} \left(m^2 + (1-m^2) \mu_{\text{Tail}} \right).$$

Since $\beta > 0$, the sequence $a_i := i^{-\beta}$ is strictly decreasing. For every $(i, j) \in S_{\text{LR}} \times S_{\text{Tail}}$ we have $i < j$, hence $a_i > a_j$. Summing over all $k(d-k-1)$ such pairs gives

$$(d-k-1) \sum_{i=2}^{k+1} i^{-\beta} > k \sum_{j=k+2}^d j^{-\beta}.$$

Dividing by $k(d-k-1)$ yields $\mu_{\text{LR}} > \mu_{\text{Tail}}$.

Because $m^2 + (1-m^2)\mu$ is strictly increasing in μ , the inequality $\mu_{\text{LR}} > \mu_{\text{Tail}}$ implies

$$\hat{\mathbf{w}}^{\top} \Sigma \hat{\mathbf{w}} \Big|_{\text{LR}} > \hat{\mathbf{w}}^{\top} \Sigma \hat{\mathbf{w}} \Big|_{\text{Tail}} \implies \gamma_{\text{Tail}} > \gamma_{\text{LR}}.$$

Moreover, the multiplicative separation between logits is

$$\frac{\gamma_{\text{Tail}}}{\gamma_{\text{LR}}} = \sqrt{\frac{m^2 + (1-m^2)\mu_{\text{LR}}}{m^2 + (1-m^2)\mu_{\text{Tail}}}} = \sqrt{\frac{m^2 + (1-m^2) \frac{1}{k} \sum_{i=2}^{k+1} i^{-\beta}}{m^2 + (1-m^2) \frac{1}{d-k-1} \sum_{i=k+2}^d i^{-\beta}}} > 1.$$

Since Φ is increasing, this implies $2\Phi(\gamma_{\text{Tail}}) - 1 > 2\Phi(\gamma_{\text{LR}}) - 1$, i.e. tail-aligned predictors are more vulnerable than LR-aligned predictors by the explicit constant factor above (at the logit level). $\square$

The proofs of Theorems 4.1 and 4.3 follow the same lines as the proofs above.

C.2. Proof of the high-dimensional asymptotic results

In this section, we provide rigorous proofs for the theoretical results presented in the main paper, focusing on Theorem 4.8.

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 (23) to (25).

We begin by stating the CGMT in its general form.

Theorem C.1 (CGMT (Gordon, 1988; Thrampoulidis et al., 2014)). *Let $\mathbf{G} \in \mathbb{R}^{m \times n}$ 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 $\mathcal{S}_w \subset \mathbb{R}^n$ and $\mathcal{S}_u \subset \mathbb{R}^m$, consider the following optimization problems with continuous function ψ on $\mathcal{S}_w \times \mathcal{S}_u$:*

$$\mathbf{C}(\mathbf{G}) := \min_{\mathbf{w} \in \mathcal{S}_w} \max_{\mathbf{u} \in \mathcal{S}_u} \mathbf{u}^\top \mathbf{G} \mathbf{w} + \psi(\mathbf{w}, \mathbf{u}) \quad (36)$$

$$\mathcal{C}(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{w} \in \mathcal{S}_w} \max_{\mathbf{u} \in \mathcal{S}_u} \|\mathbf{w}\|_2 \mathbf{g}^\top \mathbf{u} + \|\mathbf{u}\|_2 \mathbf{h}^\top \mathbf{w} + \psi(\mathbf{w}, \mathbf{u}) \quad (37)$$

The following statements hold:

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

We will utilize a specialized version of the CGMT developed by (Loureiro et al., 2021) for generalized linear models.

C.2.1. 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 (Boyd & Vandenberghe, 2004; Parikh & Boyd, 2014). We provide key definitions below.

Definition C.2 (Moreau Envelope). For a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, its Moreau envelope is defined as:

$$\mathcal{M}_{Vf(\cdot)}(\omega) = \min_{\mathbf{x}} \left[\frac{1}{2V} \|\mathbf{x} - \omega\|_2^2 + f(\mathbf{x}) \right] \quad (38)$$

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

Definition C.3 (Proximal Operator). For a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, its Proximal operator is defined as:

$$\mathcal{P}_{Vf(\cdot)}(\omega) = \arg \min_{\mathbf{x}} \left[\frac{1}{2V} \|\mathbf{x} - \omega\|_2^2 + f(\mathbf{x}) \right] \quad (39)$$

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

Theorem C.4 (Gradient of Moreau Envelope (Thrampoulidis et al., 2018), Lemma D1). *For a convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with Moreau envelope $\mathcal{M}_{Vf(\cdot)}$ and Proximal operator $\mathcal{P}_{Vf(\cdot)}$:*

$$\nabla_{\omega} \mathcal{M}_{Vf(\cdot)}(\omega) = \frac{1}{V} (\omega - \mathcal{P}_{Vf(\cdot)}(\omega)) \quad (40)$$

Additionally, we will use these important properties:

$$\mathcal{M}_{Vf(\cdot+\mathbf{u})}(\omega) = \mathcal{M}_{Vf(\cdot)}(\omega + \mathbf{u}), \quad \mathcal{P}_{Vf(\cdot+\mathbf{u})}(\omega) = \mathcal{P}_{Vf(\cdot)}(\omega + \mathbf{u}) - \mathbf{u} \quad (41)$$

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

Definition C.5 (Dual of a number). For $a \in (1, \infty]$, we define its dual $a^* \in [1, \infty)$ as the unique number satisfying $1/a + 1/a^* = 1$, with the conventions $1^* = \infty$ and $\infty^* = 1$.

C.2.2. ASSUMPTIONS AND PRELIMINARY DISCUSSION

The main text uses normalized covariates, $\mathbf{z} \sim \mathcal{N}(0, \text{Id}_d/d)$ and $\mathbf{u} \sim \mathcal{N}(0, \text{Id}_p/p)$, with unscaled prediction fields. To match the standard CGMT normalization, in this subsection only we set $\tilde{\mathbf{z}} = \sqrt{d}\mathbf{z}$, $\tilde{\mathbf{u}} = \sqrt{p}\mathbf{u}$, $\tilde{\mathbf{x}} = \sqrt{p}\mathbf{x}$, and $\tilde{\mathbf{F}} = \sqrt{p/d}\mathbf{F}$, and then drop the tildes. Thus the fields become $\mathbf{w}_*^\top \mathbf{z}/\sqrt{d}$ and $\mathbf{w}^\top \mathbf{x}/\sqrt{p}$. This exact change of variables leaves the order parameters in Theorem 4.8 unchanged.

Assumption C.6 (Estimation from the dataset). Given a dataset $\mathcal{D}$ made of n input-output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, we estimate

$$\hat{\mathbf{w}} \in \arg \min_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^n \max_{\|\mathbf{v}_i\|_s \leq r p^{1/2-1/s^*}} \ell \left(y_i \frac{\mathbf{w}^\top (\mathbf{x}_i + \mathbf{v}_i)}{\sqrt{p}} \right) + \lambda \|\mathbf{w}\|_2^2, \quad (42)$$

where $\ell : \mathbb{R} \rightarrow \mathbb{R}$ 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 C.7 (Data Distribution). In this rescaled convention, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \text{Id}_d)$, $f_{\mathbf{w}_*}(\mathbf{z}) = \varphi(\langle \mathbf{w}_*, \mathbf{z} \rangle / \sqrt{d})$ with $\mathbf{w}_* \in \mathbb{S}^{d-1}(\sqrt{d})$, and $\mathbf{x} = \mathbf{F} \mathbf{z} + \mathbf{u} \in \mathbb{R}^p$ with independent $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \text{Id}_p)$ and

$$\mathbf{F} = \begin{cases} \begin{bmatrix} \sqrt{\frac{p}{d}} \text{Id}_d \\ \mathbf{0}_{(p-d) \times d} \end{bmatrix} & \text{if } p \geq d \\ \begin{bmatrix} \text{Id}_p & \mathbf{0}_{p \times (d-p)} \end{bmatrix} & \text{if } p < d \end{cases}. \quad (43)$$

Assumption C.8 (High-Dimensional Limit). We consider the proportional high-dimensional regime where both the number of training data and input dimension $n, d, p \rightarrow \infty$ 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 C.9 (Scaling of Adversarial Norm Constraint). For observed dimension p and training norm $s > 1$, the dimension-free budget r in the main text corresponds here to physical radius $rp^{1/2-1/s^*}$. Hence $p^{-1/2} \sup_{\|\mathbf{v}\|_s \leq rp^{1/2-1/s^*}} \mathbf{w}^\top \mathbf{v} = r(\|\mathbf{w}\|_{s^*}^s / p)^{1/s^*} = O(1)$.

C.2.3. PROBLEM SIMPLIFICATION

Recall that we start from the following optimization problem:

$$\Phi_p = \min_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^n \max_{\|\mathbf{v}_i\|_s \leq rp^{1/2-1/s^*}} \ell \left(y_i \frac{\mathbf{w}^\top (\mathbf{x}_i + \mathbf{v}_i)}{\sqrt{p}} \right) + \lambda \|\mathbf{w}\|_2^2. \quad (44)$$

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

$$\Phi_p = \min_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^n \ell \left(y_i \frac{\mathbf{w}^\top \mathbf{x}_i}{\sqrt{p}} - \frac{r}{p^{1/s^*}} \|\mathbf{w}\|_{s^*} \right) + \lambda \|\mathbf{w}\|_2^2. \quad (45)$$

To facilitate our analysis, we introduce $P = \|\mathbf{w}\|_{s^*}^s / p$ and its Lagrange multiplier $\hat{P}$, which decouple the norm constraint. This leads to

$$\Phi_p = \min_{\mathbf{w} \in \mathbb{R}^p, P} \max_{\hat{P}} \sum_{i=1}^n \ell \left(y_i \frac{\mathbf{w}^\top \mathbf{x}_i}{\sqrt{p}} - r \sqrt[s^*]{\hat{P}} \right) + \lambda \|\mathbf{w}\|_2^2 + \hat{P} \|\mathbf{w}\|_{s^*}^s - p P \hat{P}, \quad (46)$$

This reformulation exposes the dimension-free adversarial shift rP^{1/s^*} and allows us to apply the CGMT.

C.2.4. SCALARIZATION AND APPLICATION OF CGMT

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

$$\tilde{\ell}(\mathbf{y}, \mathbf{z}) = \sum_{i=1}^n \ell \left(y_i \mathbf{z}_i - r \sqrt[s^*]{\hat{P}} \right), \quad \tilde{r}(\mathbf{w}) = \|\mathbf{w}\|_2^2 + \hat{P} \|\mathbf{w}\|_{s^*}^s. \quad (47)$$

A crucial step in our analysis involves inverting the order of the min-max optimization. We justify this operation by considering minimization with respect to $\mathbf{w} \in \mathbb{R}^p$ at fixed $\hat{P}$ and P . The objective is convex in $\mathbf{w}$, and the relevant saddle formulation has convex constraint sets. 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 Theorem C.7 fits the framework of (Loureiro et al., 2021). Specifically, it can be seen as the case treated in Section 3.1 with the non-linearity reducing to additive Gaussian noise.

This reformulation enables us to directly apply (Loureiro et al., 2021, Lemma 11), a careful application of Theorem C.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 captures the limiting behavior of the optimum.

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

$$\tilde{\Phi} = \min_{P, m, \eta, \tau_1} \max_{\hat{P}, \kappa, \tau_2, \nu} \left[\frac{\kappa \tau_1}{2} - \alpha \mathcal{L}_\ell - \frac{\eta}{2 \tau_2} (\nu^2 + \kappa^2) - \frac{\eta \tau_2}{2} - \mathcal{L}_r + m \nu - P \hat{P} \right] \quad (48)$$

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

In this expression, $\mathbf{g}$ and $\mathbf{h}$ 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. (38))

$$\mathcal{L}_\ell = \frac{1}{n} \mathbb{E} \left[\mathcal{M}_{\frac{\tau_1}{\kappa} \tilde{\ell}(\mathbf{y}, \cdot)}(m \mathbf{s} + \eta \mathbf{h}) \right] \quad (49)$$

$$\mathcal{L}_r = \frac{1}{d} \mathbb{E} \left[\mathcal{M}_{\frac{\eta}{\tau_2} \tilde{r}(\cdot)} \left(\frac{\eta}{\tau_2} (\kappa \mathbf{g} + \nu \mathbf{w}_*) \right) \right] \quad (50)$$

The extremization problem in Eq. (48) captures the leading-order behavior of the original optimization problem in Eq. (44) as $n, d, p \rightarrow \infty$.

This dimensional reduction replaces the original high-dimensional problem with a tractable, dimension-free set of scalar equations.

At finite size, the optimization problem $\tilde{\Phi}$ still depends implicitly on n, d , and p . We introduce two variables

$$\mathbf{w}_{\text{eq}} = \mathcal{P}_{\frac{\eta^*}{\tau_2^*} \tilde{r}(\cdot)} \left(\frac{\eta^*}{\tau_2^*} (\nu^* \mathbf{t} + \kappa^* \mathbf{g}) \right), \quad \mathbf{z}_{\text{eq}} = \mathcal{P}_{\frac{\tau_1^*}{\kappa^*} \tilde{\ell}(\cdot, \mathbf{y})} (m^* \mathbf{s} + \eta^* \mathbf{h}) \quad (51)$$

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

Building upon (Loureiro et al., 2021, Theorem 5), we can establish a convergence result. Let $\hat{\mathbf{w}}$ be an optimal solution of the problem defined in Eq. (44), and let $\hat{\mathbf{z}} = \frac{1}{\sqrt{p}} \mathbf{X} \hat{\mathbf{w}}$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$ stacks the data points by row. For any Lipschitz function $\varphi_1 : \mathbb{R}^p \rightarrow \mathbb{R}$ and any separable, pseudo-Lipschitz function $\varphi_2 : \mathbb{R}^n \rightarrow \mathbb{R}$, there exist constants $\epsilon, C, c > 0$ such that

$$\mathbb{P} \left(\left| \varphi_1 \left(\frac{\hat{\mathbf{w}}}{\sqrt{p}} \right) - \mathbb{E} \left[\varphi_1 \left(\frac{\mathbf{w}_{\text{eq}}}{\sqrt{p}} \right) \right] \right| \geq \epsilon \right) \leq \frac{C}{\epsilon^2} e^{-c n \epsilon^4} \quad (52)$$

$$\mathbb{P} \left(\left| \varphi_2 \left(\frac{\hat{\mathbf{z}}}{\sqrt{n}} \right) - \mathbb{E} \left[\varphi_2 \left(\frac{\mathbf{z}_{\text{eq}}}{\sqrt{n}} \right) \right] \right| \geq \epsilon \right) \leq \frac{C}{\epsilon^2} e^{-c n \epsilon^4} \quad (53)$$

It demonstrates that the limiting values of any function depending on $\hat{\mathbf{w}}$ and $\hat{\mathbf{z}}$ can be computed by taking the expectation of the same function evaluated at $\mathbf{w}_{\text{eq}}$ or $\mathbf{z}_{\text{eq}}$, 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.

C.2.5. DERIVATION OF SADDLE POINT EQUATIONS

We next show that extremizing over $m, \eta, \tau_1, P, \hat{P}, \nu, \tau_2, \kappa$ yields the optimal value $\tilde{\Phi}$ in Eq. (48). We derive the saddle-point equations directly and then show that their high-dimensional limits coincide with those in the main text.

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

$$\begin{aligned} \frac{\partial}{\partial m} : \nu &= \alpha \frac{\kappa}{n \tau_1} \mathbb{E} \left[\left(\frac{m}{\eta} \mathbf{h} - \mathbf{s} \right)^\top \mathcal{P}_{\frac{\tau_1}{\kappa} \tilde{\ell}(\cdot, \mathbf{y})} (m \mathbf{s} + \eta \mathbf{h}) \right] \\ \frac{\partial}{\partial \eta} : \tau_2 &= \alpha \frac{\kappa}{\tau_1} \eta - \frac{\kappa \alpha}{\tau_1 n} \mathbb{E} \left[\mathbf{h}^\top \mathcal{P}_{\frac{\tau_1}{\kappa} \tilde{\ell}(\cdot, \mathbf{y})} (m \mathbf{s} + \eta \mathbf{h}) \right] \\ \frac{\partial}{\partial \tau_1} : \frac{\tau_1^2}{2} &= \frac{1}{2} \alpha \frac{1}{n} \mathbb{E} \left[\left\| m \mathbf{s} + \eta \mathbf{h} - \mathcal{P}_{\frac{\tau_1}{\kappa} \tilde{\ell}(\cdot, \mathbf{y})} (m \mathbf{s} + \eta \mathbf{h}) \right\|_2^2 \right] \\ \frac{\partial}{\partial P} : \hat{P} &= \frac{\alpha}{n} \partial_P \mathbb{E} \left[\mathcal{M}_{\frac{\tau_1}{\kappa} \tilde{\ell}(\mathbf{y}, \cdot)} (m \mathbf{s} + \eta \mathbf{h}) \right] \end{aligned} \quad (54)$$

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

$$\begin{aligned} \frac{\partial}{\partial \kappa} : \tau_1 &= \frac{1}{d} \mathbb{E} \left[\mathbf{g}^\top \mathcal{P}_{\frac{\eta}{\tau_2} \tilde{r}(\cdot)} \left(\frac{\eta}{\tau_2} (\nu \mathbf{w}_* + \kappa \mathbf{g}) \right) \right] \\ \frac{\partial}{\partial \nu} : m &= \frac{1}{d} \mathbb{E} \left[\mathbf{w}_*^\top \mathcal{P}_{\frac{\eta}{\tau_2} \tilde{r}(\cdot)} \left(\frac{\eta}{\tau_2} (\nu \mathbf{w}_* + \kappa \mathbf{g}) \right) \right] \\ \frac{\partial}{\partial \tau_2} : \frac{1}{2d} \frac{\tau_2}{\eta} \mathbb{E} \left[\left\| \frac{\eta}{\tau_2} (\nu \mathbf{w}_* + \kappa \mathbf{g}) - \mathcal{P}_{\frac{\eta}{\tau_2} \tilde{r}(\cdot)} \left(\frac{\eta}{\tau_2} (\nu \mathbf{w}_* + \kappa \mathbf{g}) \right) \right\|_2^2 \right] &= \frac{\eta}{2\tau_2} (\nu^2 + \kappa^2) - m\nu - \kappa\tau_1 + \frac{\eta\tau_2}{2} + \frac{\tau_2}{2\eta} m^2 \\ \frac{\partial}{\partial \hat{P}} : P &= \frac{1}{d} \partial_{\hat{P}} \mathbb{E} \left[\mathcal{M}_{\frac{\eta}{\tau_2} \tilde{r}(\cdot)} \left(\frac{\eta}{\tau_2} (\kappa \mathbf{g} + \nu \mathbf{w}_*) \right) \right] \end{aligned} \quad (55)$$

Rewriting these equations in the form of Theorem 4.8 follows (Loureiro et al., 2021, Appendix C.2): it requires two changes of variables and an integration by parts.

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

$$\begin{aligned} m &\leftarrow m, & \tau &\leftarrow \eta^2 + m^2, & V &\leftarrow \frac{\tau_1}{\kappa}, & P &\leftarrow P, \\ \bar{V} &\leftarrow \frac{\tau_2}{\eta}, & \bar{\tau} &\leftarrow \kappa^2, & \bar{m} &\leftarrow \nu, & \bar{P} &\leftarrow \hat{P}. \end{aligned} \quad (56)$$

and the use of Isserlis' theorem (Isserlis, 1918) to simplify the expectations involving the Gaussian vectors $\mathbf{g}$ and $\mathbf{h}$.

C.2.6. REWRITING OF THE SADDLE POINT EQUATIONS

We now rewrite the zero-gradient equations in a form more amenable to numerical solution. Recently, Vilucchio et al. (2025) showed the relationship of these equations with the so-called ‘‘state evolution’’ for the AMP algorithm. This rewriting brings a computational advantage, since their numerical stability is improved.

To simplify the equations we first notice that, due to the separability, all terms depending on the proximal of either $\tilde{\ell}$ or $\tilde{r}$ 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.

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

$$\mathcal{Z}_0(y, \omega, V) = \int \frac{dx}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} \delta(y - f^0(x)). \quad (57)$$

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. This brings the first set of equations to be

$$\begin{cases} \bar{m} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} dy \partial_\omega \mathcal{Z}_0(y, \sqrt{\frac{m^2}{\tau}} \xi, 1 - \frac{m^2}{\tau}) f_\ell(\sqrt{\tau} \xi, P, r) \right] \\ \bar{\tau} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} dy \mathcal{Z}_0(y, \sqrt{\frac{m^2}{\tau}} \xi, 1 - \frac{m^2}{\tau}) f_\ell^2(\sqrt{\tau} \xi, P, r) \right] \\ \bar{V} = -\alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} dy \mathcal{Z}_0(y, \sqrt{\frac{m^2}{\tau}} \xi, 1 - \frac{m^2}{\tau}) \partial_\omega f_\ell(\sqrt{\tau} \xi, P, r) \right] \\ \bar{P} = r P^{1/s} \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} dy \mathcal{Z}_0(y, \sqrt{\frac{m^2}{\tau}} \xi, 1 - \frac{m^2}{\tau}) y f_\ell(\sqrt{\tau} \xi, P, r) \right] \end{cases} \quad (58)$$

where $\xi \sim \mathcal{N}(0, 1)$ and we defined for notational convenience the following

$$f_\ell(\omega, P, r) = \frac{1}{V} \left(\mathcal{P}_{V\ell(y \cdot -r^* \sqrt{P})}(\omega) - \omega \right). \quad (59)$$

Next, we simplify the remaining set of equations by introducing

$$\mathcal{Z}_{w_*}(\varrho, \Lambda) = \int dw e^{-\frac{1}{2}w^2} e^{-\frac{\Lambda}{2}w^2 + \varrho w}, \quad (60)$$

which, in turn, leads to the second set of equations

$$\begin{cases} \bar{m} = \frac{1}{\sqrt{\gamma}} \mathbb{E}_\xi [\partial_\varrho \mathcal{Z}_{w_*} f_w^1] \\ \bar{\tau} = (1 + \gamma) \mathbb{E}_\xi [\mathcal{Z}_{w_*} (f_w^1)^2] + (1 - \gamma) \mathbb{E}_\xi [(f_w^2)^2] \\ \bar{V} = (1 + \gamma) \mathbb{E}_\xi [\mathcal{Z}_{w_*} \partial_\varrho f_w^1] + (1 - \gamma) \mathbb{E}_\xi [\partial_\varrho f_w^2] \\ \bar{P} = \gamma \mathbb{E}_\xi [\mathcal{Z}_{w_*} |f_w^1|^{s^*}] + (1 - \gamma) \mathbb{E}_\xi [|f_w^2|^{s^*}] \end{cases} \quad \text{for } \gamma \leq 1, \quad (61)$$

$$\begin{cases} \bar{m} = \frac{1}{\sqrt{\gamma}} \mathbb{E}_\xi [\partial_\varrho \mathcal{Z}_{w_*} f_w] \\ \bar{\tau} = 2 \mathbb{E}_\xi [\mathcal{Z}_{w_*} f_w] \\ \bar{V} = 2 \mathbb{E}_\xi [\mathcal{Z}_{w_*} \partial_\varrho f_w] \\ \bar{P} = \mathbb{E}_\xi [\mathcal{Z}_{w_*} |f_w|^{s^*}] \end{cases} \quad \text{for } \gamma > 1, \quad (62)$$

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

$$f_w(\varrho, \Lambda, \pi) = \arg \min_z \left[\pi |z|^{s^*} + \lambda z^2 + \frac{\Lambda}{2} z^2 - \varrho z \right]. \quad (63)$$

C.2.7. SPECIALIZATION FOR $s = 2$

To obtain the specific form of the saddle-point equations presented in Theorem 4.8, one applies the closed-form proximal operator for $s = 2$,

$$f_w(\varrho, \Lambda, \pi) = \frac{\varrho}{\lambda + \pi + \Lambda}. \quad (64)$$

This renders Equations (24) and (25) amenable to explicit integration: they become Gaussian integrals over the variable ξ . Performing these Gaussian integrals yields the form shown in the main text.

C.3. Limiting Values for the Error Metrics

We start by proving the following lemma that will be useful in the following.

Lemma C.10 (Consistent perturbation minimization). *For any decreasing function $g: \mathbb{R} \rightarrow \mathbb{R}$, any $y \in \{\pm 1\}$, and any vectors $\mathbf{x}, \mathbf{w} \in \mathbb{R}^d$,*

$$\delta: \|\delta\|_q \leq \varepsilon, \langle \mathbf{w}_*, \delta \rangle = 0 \quad \max_{\delta} g(y \langle \mathbf{w}, \mathbf{x} + \delta \rangle) = \inf_{\kappa \in \mathbb{R}} g\left(y \langle \mathbf{w}, \mathbf{x} \rangle - \varepsilon \|\mathbf{w} - \kappa \mathbf{w}_*\|_{q^*}\right) \quad (65)$$

Proof of Theorem C.10. Since $g: \mathbb{R} \rightarrow \mathbb{R}$ in Eq. (65) is decreasing (but not necessarily continuous), it suffices to minimize its argument and then evaluate g at the minimizer. Splitting the argument,

$$\delta: \|\delta\|_q \leq \varepsilon, \langle \mathbf{w}_*, \delta \rangle = 0 \quad \min_{\delta} y \langle \mathbf{w}, \mathbf{x} + \delta \rangle = y \langle \mathbf{w}, \mathbf{x} \rangle + \min_{\delta: \|\delta\|_q \leq \varepsilon, \langle \mathbf{w}_*, \delta \rangle = 0} y \langle \mathbf{w}, \delta \rangle, \quad (66)$$

the first term does not depend on δ , so we focus on the second. Since $y \in \{+1, -1\}$, the change of variables $\delta \rightarrow y\delta$ leaves the constraints invariant (the q -norm is even and the orthogonality constraint is linear) and removes the y factor from the objective. We can therefore write

$$\delta: \|\delta\|_q \leq \varepsilon, \langle \mathbf{w}_*, \delta \rangle = 0 \quad \min_{\delta} \langle \mathbf{w}, \delta \rangle = \min_{\delta: \|\delta\|_q \leq \varepsilon} \sup_{\kappa \in \mathbb{R}} \langle \mathbf{w}, \delta \rangle + \kappa \langle \mathbf{w}_*, \delta \rangle = \min_{\delta: \|\delta\|_q \leq \varepsilon} \sup_{\kappa \in \mathbb{R}} \langle \mathbf{w} + \kappa \mathbf{w}_*, \delta \rangle, \quad (67)$$

where we introduced a Lagrange multiplier κ for the linear constraint. Strong duality holds for this min-sup, so we may interchange the order:

$$\sup_{\kappa \in \mathbb{R}} \min_{\delta: \|\delta\|_q \leq \varepsilon} \langle \mathbf{w} + \kappa \mathbf{w}_*, \delta \rangle = \sup_{\kappa \in \mathbb{R}} (-\varepsilon \|\mathbf{w} + \kappa \mathbf{w}_*\|_{q^*}) = -\varepsilon \inf_{\kappa \in \mathbb{R}} \|\mathbf{w} + \kappa \mathbf{w}_*\|_{q^*}, \quad (68)$$

using the duality $1/q + 1/q^* = 1$. Hence the optimal attack contribution is

$$-\varepsilon \inf_{\kappa \in \mathbb{R}} \|\mathbf{w} + \kappa \mathbf{w}_*\|_{q^*}, \quad (69)$$

which is the term that appears inside g on the right-hand side of Eq. (65). $\square$

C.3.1. PROPER ERROR METRICS

We now compute the limiting values of the proper error metrics. Theorem C.10 is the basis of the proofs of Theorems 3.12 and 4.9: it tells us that, given the joint limiting distribution of $\mathbf{w}_*$ and $\mathbf{w}$, one can evaluate the limiting form of the attack term as a function of κ and then take the extremization over κ .

Proof of Theorem 3.12. We first recall that the limiting distribution of the trained estimator is

$$\hat{\mathbf{w}} \sim \mathcal{P}_{\bar{V}\bar{\tau}(\cdot, \bar{P})}(\bar{m}\mathbf{w}_* + \sqrt{\bar{\tau} - \bar{m}^2}\boldsymbol{\xi}), \quad (70)$$

where $\bar{m}$, $\bar{\tau}$, $\bar{V}$ and $\bar{P}$ are the solution of the saddle-point equations of Vilucchio et al. (2026). In the case $s = 2$ the proximal can be evaluated in closed form, yielding $\hat{\mathbf{w}}_i = m(\mathbf{w}_*)_i + \sqrt{\tau - m^2}\xi_i$ with $\xi \sim \mathcal{N}(\mathbf{0}, \text{Id}_d)$, where m and τ are the (unbarred) order parameters determined by the saddle point.

We now evaluate the limit of the term in Eq. (69), which is the ℓ^{q^*} -norm of a Gaussian vector with a specific covariance. For any $p \in [1, \infty]$ the ℓ^p -norm of a Gaussian vector concentrates around its expectation: this is an application of (Vershynin, 2018, Theorem 5.5) to the Lipschitz function $\|M \cdot\|_p$, where M is the square root of the covariance of the Gaussian vector. The explicit limit for our case can be found as (Biau & Mason, 2015, Corollary 1). Under the scaling $\tilde{\varepsilon} := \varepsilon d^{1/q^*}$ of the adversarial budget, we obtain

$$\varepsilon \inf_{\kappa} \|\tilde{\mathbf{w}} + \kappa \mathbf{w}_*\|_{q^*} \xrightarrow{d \rightarrow \infty} \tilde{\varepsilon} \inf_{\kappa} \frac{\sqrt{2}}{\pi^{(2q^*)^{-1}}} \sqrt{(m + \kappa)^2 + \tau - m^2}^{q^*} \sqrt{\Gamma\left(\frac{q^* + 1}{2}\right)} \quad (71)$$

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

$$\tilde{\varepsilon} \sqrt{2} \sqrt{\tau - m^2}^{q^*} \sqrt{\frac{\Gamma((q^* + 1)/2)}{\sqrt{\pi}}}. \quad (72)$$

The existence of such a closed form is a peculiarity of the case $s = 2$. In the general case the limiting value still exists because the proximal of a convex function is itself Lipschitz; applying Eq. (70) then yields

$$\mathcal{A} = \inf_{\kappa \in \mathbb{R}} \mathbb{E}_{\xi} \left[\int d\mathbf{w}_* |f_{\mathbf{w}} - \kappa \mathbf{w}_*|^{q^*} \right]^{1/q^*}. \quad (73)$$

where $f_{\mathbf{w}} = \mathcal{P}_{\bar{V}\bar{\tau}(\cdot, \bar{P})}(\bar{m}\mathbf{w}_* + \sqrt{\bar{\tau} - \bar{m}^2}\boldsymbol{\xi})$ is the limiting distribution of the entries of $\hat{\mathbf{w}}$.

Once the attack-surface term concentrates to its limit, the limiting distribution of the metrics Eqs. (2) and (3) can be obtained by the *local fields* method (Clarté et al., 2023): in the well-specified model the relevant pair of random variables is

$$\begin{pmatrix} \langle \mathbf{w}_*, \mathbf{x} \rangle \\ \langle \hat{\mathbf{w}}, \mathbf{x} \rangle \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & m \\ m & \tau \end{pmatrix}\right), \quad (74)$$

where m and τ are the order parameters obtained from the saddle-point equations. For completeness, let $P(dy | z_1)$ denote the conditional label law on the label space $\mathcal{Y}$, and set $a = \tilde{\varepsilon}\mathcal{A} \geq 0$. If δ_0 minimizes the raw model field, the change of variables in the proof of Theorem C.10 shows that a sample with label y receives $y\delta_0$. Consequently, $z_2^{\text{adv}} = z_2 - ya$ and its signed margin is $yz_2^{\text{adv}} = yz_2 - a$. In particular, the attack subtracts a from the *signed margin*; the expression $y(z_2 - a)$ would instead apply the same raw-field shift to both labels and is not equivalent. Applying the lemma to the indicator loss therefore gives the generic-channel formulas

$$E_{\text{rob}}^{\text{cns}} = \int d\mathbf{p}(z_1, z_2) \int_{\mathcal{Y}} P(dy | z_1) \mathbb{1}\{yz_2 < a\}, \quad (75)$$

$$E_{\text{bnd}}^{\text{cns}} = \int d\mathbf{p}(z_1, z_2) \int_{\mathcal{Y}} P(dy | z_1) \mathbb{1}\{0 < yz_2 < a\}. \quad (76)$$

Since $a \geq 0$, these indicators also obey, away from the zero-probability Gaussian boundaries,

$$\mathbb{1}\{yz_2 < a\} = \mathbb{1}\{yz_2 < 0\} + \mathbb{1}\{0 < yz_2 < a\}. \quad (77)$$

Thus $E_{\text{rob}}^{\text{cns}}$ is the clean error plus $E_{\text{bnd}}^{\text{cns}}$, providing a direct consistency check on the formulas. For binary labels the inner integral is the corresponding sum over $y \in \{\pm 1\}$. For a noiseless link, the conditional law is the point mass at $y = \text{sign}(z_1)$, and these expressions reduce to Eqs. (15) and (16). The same signed-margin calculation gives the standard robust error after replacing $\mathcal{A}$ by $\mathcal{B}$, and the latent-space formulas after replacing $\mathcal{A}$ by $\mathcal{C}$. $\square$

Proof of Theorem 4.9. The argument mirrors the proof of Theorem 3.12; only the limiting form of the attack-surface constant changes, and that change is governed by the block structure of the feature matrix F (Theorem 4.6).

Step 1: Apply Theorem C.10 in the latent attack form. By Theorem 4.1, a consistent latent-space perturbation $\delta \in \mathbb{R}^d$ with $\langle w_*, \delta \rangle = 0$ acts on the model field through $\langle \hat{\theta}, F \delta \rangle$. Applying Theorem C.10 with $w \leftarrow F^\top \hat{\theta}$ and $x \leftarrow z$, for any non-increasing g ,

$$\max_{\delta \in H_q(\varepsilon)} g(y \langle \hat{\theta}, F(z + \delta) + u \rangle) = \inf_{\kappa \in \mathbb{R}} g\left(y \langle \hat{\theta}, Fz + u \rangle - \varepsilon \|F^\top \hat{\theta} - \kappa w_*\|_{q^*}\right). \quad (78)$$

Under Theorem 3.10, the attack contribution concentrates to $\tilde{\mathcal{C}}$, with

$$\mathcal{C} = \lim_{d,p,n \rightarrow \infty} \frac{1}{d^{1/q^*}} \inf_{\kappa \in \mathbb{R}} \|F^\top \hat{\theta} - \kappa w_*\|_{q^*}.$$

Step 2: Block decomposition of $F^\top \hat{\theta}$. Theorem 4.6 prescribes

$$F = \begin{cases} [\text{Id}_d; \mathbf{0}_{(p-d) \times d}] & p \geq d, \\ \sqrt{\frac{d}{p}} [\text{Id}_p, \mathbf{0}_{p \times (d-p)}] & p < d. \end{cases}$$

Hence, writing $\hat{\theta} = (\hat{\theta}^s, \hat{\theta}^k)$ for the block decomposition induced by F :

- For $\gamma \leq 1$ ($p \geq d$): $\hat{\theta}^s = \hat{\theta}_{1:d} \in \mathbb{R}^d$ couples to w_* through $F^\top \hat{\theta} = \hat{\theta}^s$; the kernel block $\hat{\theta}_{d+1:p}$ does not contribute to the latent attack surface.
- For $\gamma > 1$ ($p < d$): $F^\top \hat{\theta} = \sqrt{\gamma}(\hat{\theta}^\top, \mathbf{0}_{d-p}^\top)^\top$, so the first p coordinates carry the signal and the remaining $d-p$ coordinates are zero.

The rewriting of the saddle-point equations in Section C.2 (Equations (24) and (25)) implements precisely this decomposition: it yields the limiting laws f_w^1 (signal block, $\gamma \leq 1$), f_w^2 (kernel block, $\gamma \leq 1$) and f_w^3 (single block, $\gamma > 1$) for the empirical distribution of the entries of $\hat{\theta}$, expressed through the proximal of the regularizer.

Step 3: Limit of the dual norm. The ℓ^{q^*} -norm of a vector with i.i.d. asymptotically Gaussian entries concentrates by the Lipschitz-Gaussian inequality (Vershynin (2018, Theorem 5.5); cf. the well-specified argument). Splitting the sum that defines $\|F^\top \hat{\theta} - \kappa w_*\|_{q^*}^{q^*}$ across the blocks identified in Step 2 and using the empirical-distribution limits from Loureiro et al. (2021, Lemma 5) yields

$$\mathcal{C} = \inf_{\kappa \in \mathbb{R}} \begin{cases} \left(\mathbb{E}_\xi \int dw_* h^1(w_*) |f_w^1 - \kappa w_*|^{q^*} \right)^{1/q^*} & \gamma \leq 1, \\ \left(\frac{1}{\gamma} \mathbb{E}_\xi \int dw_* h^3(w_*) |\sqrt{\gamma} f_w^3 - \kappa w_*|^{q^*} + \left(1 - \frac{1}{\gamma}\right) \mathbb{E}_\xi \int dw_* h^3(w_*) |\kappa w_*|^{q^*} \right)^{1/q^*} & \gamma > 1, \end{cases} \quad (79)$$

where the densities $h^i(w_*)$ encode the Gaussian "channel" prior on w_* at the relevant saddle point and the f_w^i are the limiting proximals of Step 2. Explicitly, with

$$h(w_*; \varrho, \Lambda) = e^{-w_*^2/2} e^{-\Lambda w_*^2/2 + \varrho w_*},$$

one has $h^1 = h(w_*; \bar{m}\xi/\sqrt{(1+\gamma)\bar{\tau}}, \bar{m}^2/(1+\gamma)\bar{\tau})$ and $h^3 = h(w_*; \bar{m}\xi/\sqrt{2\bar{\tau}}, \bar{m}^2/2\bar{\tau})$. In the $\gamma > 1$ branch, the weights $1/\gamma = p/d$ and $1 - 1/\gamma$ are the fractions of nonzero and zero coordinates of $F^\top \hat{\theta}$, respectively.

Step 4: Local-fields conclusion. Substituting $\mathcal{C}$ into Eq. (78) and applying the generic-channel formulas Eqs. (75) and (76) with $\mathcal{A}$ replaced by $\mathcal{C}$ with $z_1 = \langle \mathbf{w}_*, \mathbf{z} \rangle$ and $z_2 = \langle \hat{\boldsymbol{\theta}}, \mathbf{x} \rangle$ (jointly Gaussian with the covariance stated in Theorem 4.9, by Loureiro et al. (2021, Lemma 5)) yields the generic noisy-channel result. For the noiseless link $y = \text{sign}(z_1)$, it reduces to Eq. (26) and Eq. (27). $\square$

Corollary C.11 (Euclidean simplification of $\mathcal{C}$). *Under the setting of Theorem 4.9, when the attack norm is Euclidean ($q = 2$, hence $q^* = 2$), the attack-surface constant in Equation (79) simplifies to*

$$\mathcal{C} = \sqrt{\tau_\ell - m^2}, \quad (80)$$

where $\tau_\ell := \|\mathbf{F}^\top \hat{\boldsymbol{\theta}}\|_2^2 / d$ is the latent-space self-overlap and $m := \langle \mathbf{F}^\top \hat{\boldsymbol{\theta}}, \mathbf{w}_* \rangle / d$ is the target alignment, both as in Theorem 4.8.

Proof. For $q^* = 2$, the dual norm appearing in Eq. (78) is Euclidean, and the infimum admits a one-dimensional Hilbert-space projection in closed form. For any $v \in \mathbb{R}^d$ and any nonzero $\mathbf{w}_* \in \mathbb{R}^d$,

$$\inf_{\kappa \in \mathbb{R}} \|v - \kappa \mathbf{w}_*\|_2^2 = \|v\|_2^2 - \frac{\langle v, \mathbf{w}_* \rangle^2}{\|\mathbf{w}_*\|_2^2}, \quad \text{attained at } \kappa^* = \frac{\langle v, \mathbf{w}_* \rangle}{\|\mathbf{w}_*\|_2^2}, \quad (81)$$

which is Pythagoras applied to the orthogonal decomposition $v = \frac{\langle v, \mathbf{w}_* \rangle}{\|\mathbf{w}_*\|_2^2} \mathbf{w}_* + P_\perp v$ with $P_\perp = \text{Id}_d - \mathbf{w}_* \mathbf{w}_*^\top / \|\mathbf{w}_*\|^2$.

Specializing to $v = \mathbf{F}^\top \hat{\boldsymbol{\theta}}$, and using $\|\mathbf{w}_*\|_2^2 = d$ from $\mathbf{w}_* \in \mathbb{S}^{d-1}(\sqrt{d})$,

$$\inf_{\kappa \in \mathbb{R}} \|\mathbf{F}^\top \hat{\boldsymbol{\theta}} - \kappa \mathbf{w}_*\|_2^2 = \|\mathbf{F}^\top \hat{\boldsymbol{\theta}}\|_2^2 - \frac{(\mathbf{F}^\top \hat{\boldsymbol{\theta}}, \mathbf{w}_*)^2}{d} = d\tau_\ell - \frac{(dm)^2}{d} = d(\tau_\ell - m^2),$$

attained at $\kappa^* = m$. Substituting into the definition of $\mathcal{C}$ in Step 3 of the proof of Theorem 4.9 — for $q^* = 2$ the dimensional prefactor exactly cancels the $\sqrt{d}$ inside the square root — gives Eq. (80).

This Euclidean projection does not require the case split in Eq. (79): by the definitions of τ_ℓ , m , and $\|\mathbf{w}_*\|_2^2 / d = 1$, expanding the quadratic in either branch reduces it to

$$\mathcal{C}^2 = \inf_{\kappa \in \mathbb{R}} [\tau_\ell - 2\kappa m + \kappa^2] = \tau_\ell - m^2,$$

confirming Eq. (80) directly from Eq. (79). $\square$

C.4. Asymptotic Behavior Under Extreme Overparameterization

This subsection is devoted to the proof of Proposition 4.11. We analyze the behavior of the consistent errors in the limit $\psi \rightarrow \infty$, which corresponds to $\gamma = d/p \rightarrow 0^+$ at fixed sample complexity α and adversarial budget ε .

Proof of Proposition 4.11. Put $t = \sqrt{\gamma}$ and define, for later use, the standard Gaussian density and distribution function by

$$\phi(u) = \frac{e^{-u^2/2}}{\sqrt{2\pi}}, \quad \Phi(v) = \int_{-\infty}^v \phi(u) du.$$

Step 1: Scaling of the saddle-point solution. We first justify the expansion of the order parameters from Eqs. (23) and (24). For the logistic loss $\ell(y, z) = \log(1 + e^{-yz})$, the scalar proximal has the representation (Briceño-Arias et al., 2019, Proposition 2)

$$\mathcal{P}_{V\ell(y, \cdot)}(\omega) = \omega + yW_{e^{-y\omega}}(Ve^{-y\omega}), \quad (82)$$

where the generalized Lambert function W_r is defined by $u(e^u + r) = v$ if and only if $u = W_r(v)$. Together with the proximal shift identity Eq. (41), this also covers the shifted loss in the robust-training residual $f_\ell(y, \omega, V, P)$. Equivalently, the proximal optimality condition gives

$$|f_\ell(y, \omega, V, P)| < 1, \quad \partial_\omega f_\ell(y, \omega, V, P) < 0,$$

and both quantities are smooth in $(\omega, V, \sqrt{P})$ for $V > 0$. We therefore expand the fixed-point equations in t , following the asymptotic fixed-point methodology used for high-dimensional logistic ERM (Aubin et al., 2020, Supplementary Sec. V).

For a noiseless link, if $\eta = m/\sqrt{\tau}$, the channel appearing in Eq. (23) is explicitly

$$\mathcal{Z}_0(y, \eta\xi, 1 - \eta^2) = \Phi\left(\frac{y\eta\xi}{\sqrt{1 - \eta^2}}\right). \quad (83)$$

In particular, at $\eta = 0$, $\mathcal{Z}_0(y, 0, 1) = 1/2$ and $\partial_\omega \mathcal{Z}_0(y, 0, 1) = y\phi(0)$. The explicit powers of t in the channel equations and the boundedness above first give

$$\bar{m} = O(t), \quad \bar{\tau} = O(t^2), \quad \bar{V} = O(t^2), \quad \bar{P} = O(\sqrt{P}). \quad (84)$$

Let $D_\gamma = 2\lambda + \bar{P} + \bar{V}(1 + \gamma^{-1})$ be the denominator in Eq. (24). Here regularity means that D_γ stays bounded away from zero, the bounded hyperparameters are C^2 in t with $\lambda = \lambda_0 + O(t)$, and the limiting saddle system has a nonsingular Jacobian. The prior equation for P and Eq. (84) then imply $P = O(t^2)$, so $\bar{P} = O(t)$. The remaining prior equations consequently give $m = O(t)$, $\tau = O(1)$, and $V = O(1)$. The leading coefficients can also be read directly from the equations. Define

$$\begin{aligned} f_0(y, \xi) &= f_\ell(y, \sqrt{\tau_0}\xi, V_0, 0), \\ M_0 &= \mathbb{E}_\xi \left[\sum_{y=\pm 1} y\phi(0)f_0(y, \xi) \right], \\ T_0 &= \mathbb{E}_\xi \left[\frac{1}{2} \sum_{y=\pm 1} f_0(y, \xi)^2 \right], \\ S_0 &= -\mathbb{E}_\xi \left[\frac{1}{2} \sum_{y=\pm 1} \partial_\omega f_\ell(y, \sqrt{\tau_0}\xi, V_0, 0) \right]. \end{aligned}$$

Since $yf_0(y, \xi) > 0$ and $\partial_\omega f_\ell < 0$ for logistic loss, $M_0, T_0, S_0 > 0$. Expanding Eq. (83) and the smooth proximal residual in t yields

$$\bar{m} = \alpha M_0 t + O(t^2), \quad \bar{\tau} = \alpha T_0 t^2 + O(t^3), \quad \bar{V} = \alpha S_0 t^2 + O(t^3), \quad \bar{P} = O(t). \quad (85)$$

Thus the limiting saddle system is

$$\begin{aligned} D_0 &= 2\lambda_0 + \alpha S_0, & V_0 &= D_0^{-1}, \\ m_0 &= \frac{\alpha M_0}{D_0}, & \tau_0 &= \frac{\alpha^2 M_0^2 + \alpha T_0}{D_0^2}, & P_0 &= \frac{\alpha^2 M_0^2 + 2\alpha T_0}{D_0^2}. \end{aligned} \quad (86)$$

These relations are self-consistent because M_0, T_0, S_0 depend on (τ_0, V_0) . The robust-training budget $r = O(1)$ enters through $\bar{P} = O(t)$ and hence only at the next order. The implicit-function theorem applied to Eq. (86) gives

$$m = m_0 t + O(t^2), \quad \tau = \tau_0 + O(t), \quad P = P_0 t^2 + O(t^3), \quad V = V_0 + O(t). \quad (87)$$

In particular, $m_0 > 0$ on the positively aligned branch and $\tau_0 > 0$.

Step 2: Expansion of the attack threshold. For $q = 2$, Corollary C.11 and the definition $\tau = \tau_\ell + P$ give the exact identity

$$\mathcal{C}^2 = \tau - P - m^2. \quad (88)$$

Define the normalized alignment and attack threshold

$$\eta_\gamma = \frac{m}{\sqrt{\tau}}, \quad b_\gamma = \frac{\tilde{\varepsilon}\mathcal{C}}{\sqrt{\tau}}, \quad k_\gamma = \frac{\eta_\gamma}{\sqrt{1 - \eta_\gamma^2}}.$$

Equations Eq. (87)–Eq. (88) imply

$$k_\gamma = \frac{m_0}{\sqrt{\tau_0}} \sqrt{\gamma} + O(\gamma), \quad b_\gamma = \tilde{\varepsilon} + O(\gamma). \quad (89)$$

Step 3: Expansion of the error metrics. Let $U = z_2/\sqrt{\tau}$. Since (z_1, U) is standard bivariate Gaussian with correlation η_γ , conditioning on $U = u$ gives $\mathbb{P}(z_1 > 0 \mid U = u) = \Phi(k_\gamma u)$. The noiseless formulas in Theorem 4.9 and the central symmetry of (z_1, z_2) therefore give

$$E_{\text{rob}}^{\text{cns}} = 2\mathbb{P}(z_1 > 0, z_2 < \tilde{\varepsilon}\mathcal{C}) = 2 \int_{-\infty}^{b_\gamma} \phi(u) \Phi(k_\gamma u) \, du, \quad (90)$$

$$E_{\text{bnd}}^{\text{cns}} = 2\mathbb{P}(z_1 > 0, 0 < z_2 < \tilde{\varepsilon}\mathcal{C}) = 2 \int_0^{b_\gamma} \phi(u) \Phi(k_\gamma u) \, du. \quad (91)$$

The Taylor expansion $\Phi(ku) = \frac{1}{2} + \phi(0)ku + O(k^3|u|^3)$ is integrable against $\phi(u)du$. Using

$$\int_{-\infty}^b u\phi(u) \, du = -\phi(b), \quad \int_0^b u\phi(u) \, du = \phi(0) - \phi(b),$$

together with Eq. (89) gives

$$E_{\text{rob}}^{\text{cns}} = \Phi(\tilde{\varepsilon}) - \frac{m_0}{\pi\sqrt{\tau_0}} e^{-\tilde{\varepsilon}^2/2} \sqrt{\gamma} + O(\gamma), \quad (92)$$

$$E_{\text{bnd}}^{\text{cns}} = \Phi(\tilde{\varepsilon}) - \frac{1}{2} + \frac{m_0}{\pi\sqrt{\tau_0}} \left(1 - e^{-\tilde{\varepsilon}^2/2}\right) \sqrt{\gamma} + O(\gamma). \quad (93)$$

Regularity of the saddle branch makes these remainders differentiable. Finally, $\gamma = (\alpha\psi)^{-1}$ implies $\partial_\psi = -\alpha\gamma^2\partial_\gamma$; differentiating Eqs. (92) and (93) yields

$$\begin{aligned} \partial_\psi E_{\text{rob}}^{\text{cns}} &= \frac{\alpha m_0}{2\pi\sqrt{\tau_0}} e^{-\tilde{\varepsilon}^2/2} \gamma^{3/2} + O(\gamma^2), \\ \partial_\psi E_{\text{bnd}}^{\text{cns}} &= -\frac{\alpha m_0}{2\pi\sqrt{\tau_0}} \left(1 - e^{-\tilde{\varepsilon}^2/2}\right) \gamma^{3/2} + O(\gamma^2), \end{aligned}$$

which are Eqs. (28) and (29). The first leading coefficient is positive. The second is non-positive, and it is strictly negative for $\tilde{\varepsilon} > 0$; at $\tilde{\varepsilon} = 0$, the boundary error is identically zero. $\square$

D. Replica derivation of the self-consistent equations

In this appendix we give an alternative derivation of the asymptotic equations of Theorem 4.8 using the (heuristic) replica method from statistical physics (Mezard et al., 1987). The replica computation reproduces *exactly* the self-consistent equations Eqs. (23) to (25) obtained from the rigorous CGMT-based derivation of Section C.2, providing an independent cross-check. We sketch the main steps and refer to (Loureiro et al., 2021; Vilucchio et al., 2025) for the detailed manipulations, which are standard in the high-dimensional asymptotics literature.

D.1. Gibbs measure and free energy

To the adversarial training problem Eqs. (13) and (14) we associate the Gibbs measure at inverse temperature $\beta > 0$,

$$\mu_\beta(d\mathbf{w}) = \frac{1}{\mathcal{Z}_\beta} \exp \left\{ -\beta \sum_{\mu=1}^n \ell \left(y_\mu \frac{\mathbf{w}^\top \mathbf{x}_\mu}{\sqrt{p}} - \frac{\varepsilon_t}{\sqrt{p}} \|\mathbf{w}\|_{q^*} \right) - \frac{\beta\lambda}{2} \|\mathbf{w}\|_2^2 \right\} d\mathbf{w}, \quad (94)$$

whose partition function $\mathcal{Z}_\beta$ concentrates on the minimizers of the adversarially trained risk as $\beta \rightarrow \infty$. The quantity of interest is the dataset-averaged free-energy density

$$f := - \lim_{\beta \rightarrow \infty} \lim_{d \rightarrow \infty} \frac{1}{\beta d} \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_\beta, \quad (95)$$

which we compute via the replica trick $\mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_\beta = \lim_{r \rightarrow 0^+} \partial_r \mathbb{E}_{\mathcal{D}} \mathcal{Z}_\beta^r$.

D.2. Replicated partition function and order parameters

After averaging over the data, the r -replicated partition function depends on the linear combinations $\nu_\mu := \langle \mathbf{w}_*, \mathbf{z}_\mu \rangle / \sqrt{d}$ and $\lambda_\mu^a := \langle \mathbf{w}^a, \mathbf{x}_\mu \rangle / \sqrt{p}$ (for $a = 1, \dots, r$), which are jointly Gaussian with the covariances

$$\mathbb{E} \nu_\mu^2 = 1, \quad \mathbb{E} \lambda_\mu^a \nu_\mu = \frac{\sqrt{\gamma}}{d} \mathbf{w}_*^\top \mathbf{F}^\top \mathbf{w}^a, \quad \mathbb{E} \lambda_\mu^a \lambda_\mu^b = \frac{1}{p} (\mathbf{w}^a)^\top (\mathbf{F} \mathbf{F}^\top + \text{Id}_p) \mathbf{w}^b. \quad (96)$$

Introducing the order parameters m^a , Q^{ab} , P^a for these covariances together with the associated Lagrange multipliers $\hat{m}^a$, $\hat{Q}^{ab}$, $\hat{P}^a$ via Fourier representations of the constraining δ -functions, the replicated partition function takes the saddle-point form

$$\mathbb{E}_{\mathcal{D}} \mathcal{Z}_\beta^r \propto \int d\Theta d\hat{\Theta} \exp\{p \Phi^{(r)}(\Theta, \hat{\Theta})\}, \quad \Phi^{(r)} = \Psi_t + \alpha\gamma \Psi_y^{(r)} + \Psi_w^{(r)}, \quad (97)$$

where Ψ_t is the Legendre-transform (trace) term, and $\Psi_y^{(r)}$, $\Psi_w^{(r)}$ are the channel and prior contributions, whose explicit forms follow standard manipulations (Mezard et al., 1987; Loureiro et al., 2021) and are omitted here.

D.3. Replica-symmetric ansatz and zero-temperature limit

We adopt the replica-symmetric (RS) ansatz $m^a = m$, $Q^{aa} = V + q$, $Q^{ab} = q$ for $a < b$, $P^a = P$, together with matching scalings for the hatted variables, and take the zero-temperature limit through the standard substitutions $V \rightarrow V/\beta$, $\hat{V} \rightarrow \beta\hat{V}$, $\hat{q} \rightarrow \beta^2\hat{q}$, $\hat{m} \rightarrow \beta\hat{m}$, $\hat{P} \rightarrow \beta\hat{P}$. The RS channel term becomes

$$\tilde{\Psi}_y = -\mathbb{E}_\xi \left[\int dy \mathcal{Z}_0 \left(y, \frac{m}{\sqrt{q}} \xi, 1 - \frac{m^2}{q} \right) \mathcal{M}_{V\ell(y, \cdot - \varepsilon_t q^* \sqrt{P})}(\sqrt{q} \xi) \right], \quad (98)$$

with $\mathcal{Z}_0(y, \omega, V) = \int (2\pi V)^{-1/2} e^{-(x-\omega)^2/(2V)} P_0(y | x) dx$ the conditional-channel partition function, and $\mathcal{M}_f$ the Moreau envelope of f . The RS prior term splits according to the block structure of $\mathbf{F}$ from Theorem C.7, yielding, for $\gamma \leq 1$,

$$\tilde{\Psi}_w^{(\gamma \leq 1)} = \gamma \mathbb{E}_\xi \left[\mathcal{Z}_{w_*}^{(1)} \log \mathcal{Z}_w^{(1)} \right] + (1 - \gamma) \mathbb{E}_\xi \left[\log \mathcal{Z}_w^{(2)} \right], \quad (99)$$

and, for $\gamma > 1$,

$$\tilde{\Psi}_w^{(\gamma > 1)} = \mathbb{E}_\xi \left[\mathcal{Z}_{w_*}^{(3)} \log \mathcal{Z}_w^{(3)} \right], \quad (100)$$

where the partition functions $\mathcal{Z}_{w_*}(\varrho, \Lambda) = \int P_*(d\mathbf{w}_*) e^{-\Lambda \mathbf{w}_*^2/2 + \varrho \mathbf{w}_*}$ and $\mathcal{Z}_w(\varrho, \Lambda, \pi) = \int d\mathbf{w} e^{-\lambda \mathbf{w}^2/2 - \Lambda \mathbf{w}^2/2 - \pi |\mathbf{w}|^{q^*} + \varrho \mathbf{w}}$ are evaluated at the block-specific arguments

$$\begin{aligned} \mathcal{Z}_{w_*}^{(1)} &= \mathcal{Z}_{w_*} \left(\frac{\hat{m} \xi}{\sqrt{\hat{q}(1+\gamma)}}, \frac{\hat{m}^2}{\hat{q}(1+\gamma)} \right), & \mathcal{Z}_w^{(2)} &= \mathcal{Z}_w(\sqrt{\hat{q}} \xi, \hat{V}, \frac{\hat{P}}{2}), \\ \mathcal{Z}_w^{(1)} &= \mathcal{Z}_w \left(\sqrt{\hat{q}(1+\frac{1}{\gamma})} \xi, \hat{V}(1+\frac{1}{\gamma}), \frac{\hat{P}}{2} \right), & \mathcal{Z}_{w_*}^{(3)} &= \mathcal{Z}_{w_*} \left(\frac{\hat{m} \xi}{\sqrt{2\hat{q}}}, \frac{\hat{m}^2}{2\hat{q}} \right), \\ & & \mathcal{Z}_w^{(3)} &= \mathcal{Z}_w(\sqrt{2\hat{q}} \xi, 2\hat{V}, \frac{\hat{P}}{2}). \end{aligned}$$

The block superscripts (1), (2), (3) refer respectively to the signal block ($\gamma \leq 1$, dimension d), the kernel block ($\gamma \leq 1$, dimension $p - d$), and the single block ($\gamma > 1$, dimension p) of $\mathbf{F}$.

D.4. Saddle-point equations and agreement with Theorem 4.8

Extremizing $\Phi^{(0)} := \lim_{r \rightarrow 0^+} r^{-1} \Phi^{(r)}$ in the eight RS variables $(m, V, q, P, \hat{m}, \hat{V}, \hat{q}, \hat{P})$ gives the saddle-point system. The channel block, using the proximal-shifted residual $f_\ell(y, \omega, V, P) := V^{-1}(\mathcal{P}_{V\ell(y, \cdot - y_{\varepsilon_t} q^* \sqrt{P})}(\omega) - \omega)$, reads

$$\begin{aligned} \hat{m} &= \alpha \sqrt{\gamma} \mathbb{E}_\xi \int dy \partial_\omega \mathcal{Z}_0 f_\ell, & \hat{q} &= \alpha \gamma \mathbb{E}_\xi \int dy \mathcal{Z}_0 f_\ell^2, \\ \hat{V} &= -\alpha \gamma \mathbb{E}_\xi \int dy \mathcal{Z}_0 \partial_\omega f_\ell, & \hat{P} &= \alpha \gamma \varepsilon_t q^* P^{q^*-1} \mathbb{E}_\xi \int dy y \mathcal{Z}_0 f_\ell. \end{aligned} \quad (101)$$

The prior block, computed from $\partial_{\hat{m}, \hat{q}, \hat{V}, \hat{P}} \tilde{\Psi}_w$ via the identities $\partial_1 \mathcal{Z}_w = \mathcal{Z}_w f_w$ and $\partial_2 \mathcal{Z}_w = -\frac{1}{2}(\partial_\varrho f_w + f_w^2)$, with the proximal $f_w(\varrho, \Lambda, \pi) = \arg \min_z [\pi |z|^{q^*} + \lambda z^2 + \frac{\Lambda}{2} z^2 - \varrho z]$, reproduces the prior equations Eqs. (24) and (25) with the same case split $\gamma \leq 1$ versus $\gamma > 1$ inherited from the block structure of Eq. (43).

Under the change of variables $(m, \tau, V, P, \bar{m}, \bar{\tau}, \bar{V}, \bar{P}) = (m, q, V, P, \hat{m}/\sqrt{\gamma}, \hat{q}/\gamma, \hat{V}, \hat{P}/(\alpha\gamma))$, the channel system Eq. (101) coincides with Eq. (23), and the prior system matches Eqs. (24) and (25). The two derivations — CGMT (rigorous, Section C.2) and replica (heuristic, this appendix) — therefore agree, providing an independent statistical-physics check of Theorem 4.8.

E. Numerical Details

The self-consistent equations from Theorem 4.8 are amenable to 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. Denoting by x_i^{new} the value of an order parameter freshly computed at iteration i and by x_{i-1} its value at the previous iteration, the update reads $x_i := \mu x_i^{\text{new}} + (1 - \mu) x_{i-1}$, where $\mu \in (0, 1]$ is the damping parameter.

Once convergence is achieved for fixed (r, λ) , we jointly optimize both hyperparameters with a two-dimensional, gradient-free Nelder–Mead search over $r, \lambda \geq 0$.

For each iteration we evaluate the proximal operator numerically using SciPy’s (Virtanen et al., 2020) Brent algorithm for univariate minimization (`scipy.optimize.minimize_scalar`). The numerical integration is handled with SciPy’s adaptive quadrature (`scipy.integrate.quad`). These routines let us evaluate the equations and the integrations to the desired accuracy.

All experiments were run on consumer-grade hardware (Mac Studio M2 Ultra, 2022), and no individual run took more than one day of CPU time.