

# Tail-Weight Control and Localized Generalization in Nearly Low-Rank Adversarial Classification

Kunyu Wang<sup>1</sup>, Dehan Wang<sup>1</sup>, and Wenjun Chen<sup>2</sup>

<sup>1</sup> The Chinese University of Hong Kong  
kunyuwang@link.cuhk.edu.hk, dhwang@link.cuhk.edu.hk

<sup>2</sup> Columbia University  
wc2741@columbia.edu

**Abstract.** We study norm-constrained linear classification under Euclidean adversarial perturbations in a Gaussian model with a low-dimensional informative subspace and an independent noise tail. For bounded ramp loss, we prove that a principal-space witness with risk below one half forces every near-optimal predictor to have small tail weight. A path-specific density bound yields constants without requiring positive tail variance. Under isotropic principal covariance, we establish a unique population minimizer and joint local growth. Boundary normalization then removes the common attack penalty from centered margins, giving localized finite-sample guarantees governed by principal dimension and total tail energy. Globalized growth removes the entrance condition at weaker constants; a model-aware comparison retains local guarantees. Experiments with twenty paired repetitions show decreasing excess risk and tail use with sample size, and nearly unchanged behavior when tail dimension grows at fixed total energy. Pure-noise controls and optimizer diagnostics clarify the scope and limitations of these conclusions.

**Keywords:** Adversarial classification, spectral tails, local complexity, excess risk, Gaussian mixtures

## 1 Introduction

Many observed coordinates may contain little label information, yet a learner trained in the full space can use them to fit a finite sample. Under adversarial perturbations, these coefficients also affect the worst-case margin. We study whether full-space training controls the weight assigned to an independent, uninformative tail, and how that structural control affects generalization.

We study norm-constrained linear scores, a full-space Euclidean adversary and bounded ramp loss. Signal lies in a principal Gaussian block; the orthogonal tail is independent noise. The principal subspace defines an analytical oracle but is not supplied to full-space fitting.

Deleting a noise coefficient need not improve a nonconvex bounded surrogate. A principal-space witness with risk below  $1/2$  places near-optimal predictors in a favorable margin region, where tail deletion reduces the attack penalty without

a harmful variance change. We prove joint growth under isotropy and obtain a fast-rate bound with explicit local constants after an entrance check. A second version globalizes the growth inequality and removes the entrance condition at the cost of weaker constants.

*Related Work.* Adversarial Rademacher bounds for linear and nonlinear classes were developed in [2,12]; robust sample complexity can differ substantially from ordinary learning [10]. Dan et al. [6] establish sharp statistical guarantees for robust Gaussian classification. Javanmard and Soltanolkotabi [8] characterize classification accuracies in a proportional high-dimensional limit. Their estimator, loss or asymptotic regime differs from our finite-sample, fixed-radius ramp analysis.

Chen et al. [5] study benign overfitting under adversarial training with an exponential objective. Tanner et al. [11] analyze feature blocks with heterogeneous usefulness and robustness. Their predictive but vulnerable blocks should not be identified with our independent zero-signal tail. Robust low-dimensional representations and latent models are studied in [1,7]; here the learner performs no preliminary projection. Our statistical argument follows local empirical-process methods [4]. Adversarial fast rates already appear in [9]; our contribution is to verify the centered-loss variance and geometry conditions for this nonsmooth ramp model. Finally, surrogate calibration requires separate assumptions [3]; the bounds below do not identify ramp and robust classification excess risks.

Our contribution is explicit deletion growth, including zero tail variance, and its connection to finite-sample bounds through isotropic joint geometry and normalization. We claim neither the first adversarial fast rate nor a global nonconvex solver.

## 2 Model and Gaussian Risk

Let  $Y$  be uniform on  $\{-1, +1\}$ . For  $r, k \geq 1$ , observe

$$X = (Z, \tau G), \quad Z = Y\mu + \xi, \quad \xi \sim N(0, \Lambda), \quad G \sim N(0, I_k), \quad (1)$$

where  $Y, \xi, G$  are independent. Write  $w = (a, b) \in \mathbb{R}^{r+k}$  and

$$\mathcal{W}_B = \{w : \|w\|_2 \leq B\}, \quad \Gamma_z = \Lambda + \mu\mu^\top, \quad T = \mathbb{E}\|X\|_2^2 = \text{tr } \Gamma_z + \tau^2 k.$$

The decomposition is an analytical coordinate system, justified by orthogonal invariance of the attack and constraint. There is no intercept. Here *principal* means signal-bearing, not necessarily the leading covariance eigenspace. The theorem covers a broader block model; the nearly low-rank subregime has  $\tau^2 \ll \lambda_{\min}(\Gamma_z)$ . Small individual tail eigenvalues and small aggregate energy  $\tau^2 k$  are distinct properties.

For a full-space attack of radius  $\varepsilon$ , the exact worst-case margin and ramp loss are

$$M_w(x, y) = yw^\top x - \varepsilon\|w\|_2, \quad \phi_\gamma(t) = \min\{0, \max\{0, 1 - t/\gamma\}\}. \quad (2)$$

Let  $S = ((X_i, Y_i))_{i=1}^n$  be an independent and identically distributed sample from (1). Set  $\ell_w = \phi_\gamma(M_w)$ ,  $R(w) = \mathbb{E}\ell_w$ ,  $R_n(w) = n^{-1} \sum_i \ell_w(X_i, Y_i)$ ,  $R^* = \min_{\mathcal{W}_B} R$  and  $\mathcal{E}(w) = R(w) - R^*$ . Compactness and continuity give existence of a population minimizer. All risks being compared use the same  $B, \gamma, \varepsilon$ . Counting zero margins as errors gives  $\mathbb{P}(M_w \leq 0) \leq R(w)$ ; this pointwise domination does not compare the two excess risks.

**Assumption 1.** *The parameters  $B, \gamma, \varepsilon$  are positive,  $0 \leq \tau \leq \tau_{\max} < \infty$ , and  $0 < \lambda_- I_r \preceq \Lambda \preceq \lambda_+ I_r$ .*

**Assumption 2.** *For some  $\Delta > 0$ , a feasible  $w_0 = (a_0, 0)$  satisfies  $R(w_0) \leq 1/2 - 2\Delta$ .*

The witness exists exactly when  $B(\|\mu\|_2 - \varepsilon) > \gamma/2$ : this condition makes the aligned boundary vector a witness, and any witness must have mean margin above  $\gamma/2$ . The maximum feasible mean margin is  $B(\|\mu\|_2 - \varepsilon)_+$ , which exceeds this positive threshold exactly under the stated condition.

The signed margin is Gaussian with

$$\nu_w = a^\top \mu - \varepsilon \|w\|_2, \quad v_w = a^\top \Lambda a + \tau^2 \|b\|_2^2. \quad (3)$$

Indeed, conditional on either value of  $Y$ ,  $(Y\xi, YG)$  has the same centered joint Gaussian law. Thus, for  $v > 0$ ,

$$R(w) = F_\gamma(\nu_w, v_w), \quad F_\gamma(\nu, v) = \frac{1}{\gamma} \int_0^\gamma \Phi\left(\frac{t - \nu}{\sqrt{v}}\right) dt. \quad (4)$$

For  $v = 0$ , use  $F_\gamma(\nu, 0) = \phi_\gamma(\nu)$ . Formula (4) follows from  $\phi_\gamma(z) = \gamma^{-1} \int_0^\gamma \mathbf{1}\{z \leq t\} dt$  and Tonelli's theorem.

**Lemma 1.** *For  $v > 0$ , writing  $p(\nu, v) = \Phi((\gamma - \nu)/\sqrt{v}) - \Phi(-\nu/\sqrt{v})$ ,*

$$\partial_\nu F_\gamma = -p(\nu, v)/\gamma < 0, \quad (5)$$

$$\partial_v F_\gamma = \frac{\varphi((\gamma - \nu)/\sqrt{v}) - \varphi(\nu/\sqrt{v})}{2\gamma\sqrt{v}}. \quad (6)$$

Moreover,  $F_\gamma(\nu, v) < 1/2$  iff  $\nu > \gamma/2$ , and  $\partial_v F_\gamma \geq 0$  iff  $\nu \geq \gamma/2$ . The risk threshold also holds at  $v = 0$ .

*Proof.* Differentiate (4) under the integral on compact sets with positive variance. The variance derivative integrates the identity

$$\partial_v \Phi\left(\frac{t - \nu}{\sqrt{v}}\right) = \frac{1}{2\sqrt{v}} \partial_t \varphi\left(\frac{t - \nu}{\sqrt{v}}\right).$$

Symmetry gives  $F_\gamma(\gamma/2, v) = 1/2$  and (5) gives the threshold. Since the normal density decreases with absolute argument, the sign in (6) follows from  $(\gamma - \nu)^2 \leq \nu^2$  iff  $\nu \geq \gamma/2$ . The deterministic case follows directly from the ramp definition.  $\square$

### 3 Tail Localization and a Global Learning Bound

Let  $m = \|\mu\|_2$  and define the following constants under the witness condition:

$$\begin{aligned} v_{\max} &= B^2 \max\{\lambda_+, \tau_{\max}^2\}, \\ \kappa &= \frac{\gamma}{\sqrt{2\pi v_{\max}}} \exp\left(-\frac{(m-\varepsilon)^2}{2\lambda_-}\right), \quad c_b = \frac{\varepsilon\kappa}{2\gamma B} > 0. \end{aligned} \quad (7)$$

This is a bound on actual deletion paths. It exploits coupling between mean and variance, rather than minimizing normal interval mass over a rectangle containing unattainable parameter pairs. It is independent of  $k$  for fixed structural parameters, including a fixed  $\tau_{\max}$ .

**Theorem 1 (Tail-Weight Localization).** *Under Assumptions 1 and 2, every feasible  $w = (a, b)$  with  $\mathcal{E}(w) \leq \Delta$  satisfies*

$$R(a, b) - R(a, 0) \geq c_b \|b\|_2^2, \quad \|b\|_2^2 \leq \mathcal{E}(w)/c_b. \quad (8)$$

*Every population minimizer has  $b = 0$ . For a fixed principal distribution,  $R^* = \min_{\|a\|_2 \leq B} R(a, 0)$  is independent of  $k$  and  $\tau$ .*

*Proof.* Near-optimality gives  $R(w) \leq R(w_0) + \Delta \leq 1/2 - \Delta$ , hence  $\nu_w > \gamma/2$ . Let  $q = \|a\|_2$  and  $\beta = \|b\|_2$ . Then  $q > \gamma/(2m) > 0$ . Along  $w_u = (a, \sqrt{u}b)$ ,  $0 \leq u \leq 1$ ,

$$\nu_u = a^\top \mu - \varepsilon \sqrt{q^2 + u\beta^2}, \quad v_u = a^\top \Lambda a + u\tau^2 \beta^2.$$

Every path point is feasible and  $\gamma/2 < \nu_u \leq (m - \varepsilon)q$ ,  $\lambda_- q^2 \leq v_u \leq v_{\max}$ . For  $t \in [0, \gamma]$ ,  $|t - \nu_u| \leq \nu_u$ , so  $(t - \nu_u)^2/v_u \leq (m - \varepsilon)^2/\lambda_-$ . The normal density on  $[0, \gamma]$  is therefore at least  $\kappa/\gamma$ , giving  $p(\nu_u, v_u) \geq \kappa$ . By the chain rule and Lemma 1,

$$\begin{aligned} \frac{d}{du} F_\gamma(\nu_u, v_u) &= \frac{\varepsilon \beta^2 p(\nu_u, v_u)}{2\gamma \sqrt{q^2 + u\beta^2}} + \tau^2 \beta^2 \partial_v F_\gamma(\nu_u, v_u) \\ &\geq \frac{\varepsilon \kappa}{2\gamma B} \beta^2 = c_b \beta^2. \end{aligned}$$

Integrating proves the deletion bound. Feasibility of  $(a, 0)$  gives  $R(a, 0) \geq R^*$ , proving the second inequality. Applying it to a minimizer yields  $b = 0$ . No division by zero occurs, even when  $\tau = 0$ , because  $q > 0$  and the principal covariance is nondegenerate.  $\square$

*Global Concentration.* For  $0 < \eta < 1$ , define

$$U_n(\eta) = \frac{2B(\sqrt{T} + \varepsilon)}{\gamma\sqrt{n}} + \sqrt{\frac{\log(2/\eta)}{2n}}. \quad (9)$$

With probability at least  $1 - \eta$ , simultaneously over  $\mathcal{W}_B$ ,

$$|R(w) - R_n(w)| \leq U_n(\eta). \quad (10)$$

To verify this, use the no-outer-absolute-value Rademacher convention. For signs  $\sigma_i$ , the margin process is bounded by

$$\frac{B}{n} \mathbb{E} \left\| \sum_i \sigma_i Y_i X_i \right\|_2 + \frac{\varepsilon B}{n} \mathbb{E} \left| \sum_i \sigma_i \right| \leq \frac{B(\sqrt{T} + \varepsilon)}{\sqrt{n}}.$$

Scalar contraction contributes  $1/\gamma$ , and symmetrization contributes 2. Each one-sided loss deviation changes by at most  $1/n$  when one observation changes. Applying bounded differences to the two sides with failure probabilities  $\eta/2$  proves (10). Gaussian inputs need not be bounded because the loss is bounded.

If a feasible output satisfies

$$R_n(\hat{w}) \leq \inf_{w \in \mathcal{W}_B} R_n(w) + \rho, \quad \rho \geq 0, \quad (11)$$

then on this event  $\mathcal{E}(\hat{w}) \leq \min\{1, 2U_n(\eta) + \rho\}$ . When  $2U_n(\eta) + \rho \leq \Delta$ , Theorem 1 controls its tail weight. The tolerance in (11) concerns the global empirical minimum, not a stationary point.

*Growing Dimension.* Suppose  $B, \gamma, \varepsilon, \lambda_-, \lambda_+, \tau_{\max}$  are fixed,  $\|\mu_n\|_2$  is uniformly bounded, all problems have a common positive witness gap, and  $\rho_n \rightarrow 0$ . If  $r_n + \tau_n^2 k_n = o(n)$ , then  $\mathcal{E}_n(\hat{w}_n) \rightarrow 0$  and  $\|\hat{b}_n\|_2 \rightarrow 0$  in probability. Indeed, set  $\eta_n = n^{-2}$  in (9); then  $U_n(\eta_n) \rightarrow 0$ . Formula (7) provides a common positive  $c_b$ , so eventual entry and (8) prove both conclusions. For example, fixed  $r_n, k_n = n^2$  and  $\tau_n = 1/n$  are allowed. Failure of this sufficient energy condition does not establish impossibility for every learner.

## 4 Isotropic Geometry and Boundary Normalization

Retain Assumption 1 and now take  $\sigma_z > 0$  with

$$\Lambda = \sigma_z^2 I_r, \quad \mu = m u_\star, \quad \|u_\star\|_2 = 1, \quad B(m - \varepsilon) > \gamma/2. \quad (12)$$

Use  $\lambda_- = \lambda_+ = \sigma_z^2$  in (7). Let  $g(q) = R(qu_\star, 0)$  and  $q_{\min} = \gamma/(2m)$ . Define

$$c_{\text{rad}} = -g'(B) > 0, \quad c_a = \min \left\{ \frac{c_{\text{rad}}}{B}, \frac{\kappa m}{2\gamma B} \right\}. \quad (13)$$

**Theorem 2 (Joint Local Growth).** *Suppose Assumption 1 and (12) hold. Then  $w^\star = (Bu_\star, 0)$  is the unique minimizer. Choose  $\Delta > 0$  with  $R(w^\star) \leq 1/2 - 2\Delta$ . Every feasible  $w = (a, b)$  with  $\mathcal{E}(w) \leq \Delta$  satisfies*

$$\mathcal{E}(w) \geq c_a \|a - Bu_\star\|_2^2 + c_b \|b\|_2^2. \quad (14)$$

For  $r = 1$ , the stronger choice  $c_a = c_{\text{rad}}/B$  is valid.

*Proof.* Theorem 1 removes the tail of any optimum. For  $a \neq 0$ , write  $q = \|a\|_2$  and  $c = a^\top u_\star/q$ . Isotropy gives

$$R(a, 0) = \frac{1}{\gamma} \int_0^\gamma \Phi \left( \frac{t/q - mc + \varepsilon}{\sigma_z} \right) dt,$$

which is strictly decreasing in  $c$  for fixed  $q$ . Along the aligned direction,

$$-g'(q) = \frac{1}{\gamma\sigma_z} \int_0^{\gamma/q} z\varphi\left(\frac{z - (m - \varepsilon)}{\sigma_z}\right) dz > 0. \quad (15)$$

Thus  $q = B$  is optimal, proving uniqueness. The integral in (15) decreases with  $q$ , so  $-g'(q) \geq -g'(B) = c_{\text{rad}}$  on  $[q_{\min}, B]$ .

For a near-optimal  $w$ ,  $q > q_{\min}$  by the previous proof. Decompose exactly

$$\begin{aligned} \mathcal{E}(a, b) &= [R(a, b) - R(a, 0)] \\ &\quad + [R(a, 0) - R(qu_*, 0)] + [g(q) - g(B)]. \end{aligned}$$

The first term is at least  $c_b\|b\|_2^2$ . For the second, the variance remains  $\sigma_z^2 q^2$  during alignment, while the mean increases from  $ma^\top u_* - \varepsilon q$  to  $(m - \varepsilon)q$ . Each intermediate mean lies above  $\gamma/2$  and below  $(m - \varepsilon)q$ ; the path density bound in the previous proof therefore applies. Integrating (5) and then (15) gives

$$\mathcal{E}(a, b) \geq c_b\|b\|_2^2 + c_{\text{rad}}(B - q) + \frac{\kappa m}{\gamma}(q - a^\top u_*). \quad (16)$$

Finally,  $\|a - Bu_*\|_2^2 = (B - q)^2 + 2B(q - a^\top u_*)$  and  $(B - q)^2 \leq B(B - q)$  prove (14). If  $r = 1$ , the local region forces positive alignment, so the angular term is unnecessary.  $\square$

Isotropy keeps variance fixed during alignment; general anisotropic principal growth is not verified here.

*Shared Classification Optimum.* Under (12),  $w^*$  also minimizes robust classification error. Indeed, for  $q = \|a\|_2$ ,  $s = \|w\|_2$  and positive mean margin,

$$\frac{a^\top \mu - \varepsilon s}{\sqrt{\sigma_z^2 q^2 + \tau^2 \|b\|_2^2}} \leq \frac{mq - \varepsilon s}{\sigma_z q} \leq \frac{m - \varepsilon}{\sigma_z}.$$

Nonpositive mean margin gives error at least  $1/2$ , including degenerate margins. Hence the minimum is  $\Phi(-(m - \varepsilon)/\sigma_z)$ , attained by every  $(cu_*, 0)$  with  $0 < c \leq B$ . The two risks share an optimizer but not their optimal values or scaling behavior; pointwise loss domination alone supplies no excess-risk conversion.

**Corollary 1 (One-Dimensional Sphere Tail Bound).** *Under Theorem 2, let  $r = 1$ ,  $\|w\|_2 = B$  and  $\mathcal{E}(w) \leq \Delta$ . Then*

$$\mathcal{E}(w) \geq c_{b,\text{tail}}\|b\|_2^2, \quad c_{b,\text{tail}} = c_b + \frac{c_{\text{rad}}}{2B}. \quad (17)$$

*Proof.* Orient the scalar coordinate along  $u_*$ . The local region gives  $a > 0$ , and (16) gives  $\mathcal{E}(a, b) \geq c_b\|b\|_2^2 + c_{\text{rad}}(B - a)$ . On the sphere,  $B - a = \|b\|_2^2/(B + a) \geq \|b\|_2^2/(2B)$ .  $\square$

This is a separate tail inequality. Substituting  $c_{b,\text{tail}}$  into the joint-growth ellipsoid while keeping  $c_a$  unchanged would reuse the radial contribution without justification. The local complexity below retains (14) and its original constants.

**Proposition 1 (Boundary Normalization).** *Fix a unit vector  $v_0$  independently of the sample. Define  $S_B(w) = Bw/\|w\|_2$  for  $w \neq 0$  and  $S_B(0) = Bv_0$ . For every feasible  $w$  and every  $(x, y)$ ,  $\ell_{S_B(w)}(x, y) \leq \ell_w(x, y)$ . The ball and sphere  $\mathcal{W}_B^\partial = \{w : \|w\|_2 = B\}$  have the same empirical and population infima. Normalizing an output satisfying (11) preserves its tolerance  $\rho$ .*

*Proof.* For nonzero  $w$ , let  $c = B/\|w\|_2 \geq 1$ . Homogeneity gives  $M_{cw} = cM_w$ . For all real  $t$ ,  $\phi_\gamma(ct) \leq \phi_\gamma(t)$ : nonpositive margins stay in the loss-one region, and positive margins cannot incur greater loss. At  $w = 0$ , the original loss equals one. Equality of infima and preservation of the tolerance follow from sphere inclusion and this pointwise inequality.  $\square$

On the sphere the penalty is the common constant  $-\varepsilon B$ . Consequently (10) holds there with the improved quantity

$$U_n^\partial(\eta) = \frac{2B\sqrt{T}}{\gamma\sqrt{n}} + \sqrt{\frac{\log(2/\eta)}{2n}}. \quad (18)$$

To see this, enclose the sphere class by  $\phi_\gamma(yw^\top x - \varepsilon B)$  indexed over the ball, retaining the constant shift even for interior parameters. Contraction now applies to the linear score class. This refinement does not remove attack dependence from the risk or geometric constants.

## 5 Localized Risk and a Post-Fit Comparison

This section assumes the isotropic conditions and witness gap of Theorem 2 and gives the sharper sphere result. Proposition 1 makes it applicable to normalized approximate empirical minimizers over the ball.

Let  $h_w = \ell_w - \ell_{w^*}$  and  $\mathcal{H}(s) = \{h_w : w \in \mathcal{W}_B^\partial, \mathcal{E}(w) \leq s\}$  for  $0 \leq s \leq \Delta$ . Use  $\mathfrak{R}_n(\mathcal{F}) = \mathbb{E}_{S, \sigma} \sup_{f \in \mathcal{F}} n^{-1} \sum_i \sigma_i f(X_i, Y_i)$ . Define

$$K^2 = \frac{1}{\gamma^2} \left( \frac{\text{tr } \Gamma_z}{c_a} + \frac{\tau^2 k}{c_b} \right), \quad V = \frac{1}{\gamma^2} \max \left\{ \frac{\|\Gamma_z\|_{\text{op}}}{c_a}, \frac{\tau^2}{c_b} \right\}. \quad (19)$$

**Lemma 2 (Local Statistical Inputs).** *For  $0 \leq s \leq \Delta$ ,  $\mathfrak{R}_n(\mathcal{H}(s)) \leq K\sqrt{s/n}$ . Also  $\mathbb{E}h_w^2 \leq V\mathcal{E}(w)$  whenever  $\mathcal{E}(w) \leq \Delta$ .*

*Proof.* Put  $\delta w = w - w^*$  and  $D = \text{diag}(c_a I_r, c_b I_k)$ . By (14),  $\delta w^\top D \delta w \leq \mathcal{E}(w)$ . Both parameter norms are  $B$ , so the centered margin is exactly  $Y \delta w^\top X$ . Enlarge the local difference set to the ellipsoid  $\delta^\top D \delta \leq s$ . Its signed linear supremum is  $(\sqrt{s}/n) \|D^{-1/2} \sum_i \sigma_i Y_i X_i\|_2$ . Its expected square inside the norm is  $n \text{tr}(D^{-1} \Gamma)$ , with  $\Gamma = \text{diag}(\Gamma_z, \tau^2 I_k)$ . Jensen's inequality and sample-wise Lipschitz contraction around the fixed margin  $M_{w^*}$  give the complexity bound. For the second moment,

$$\mathbb{E}h_w^2 \leq \gamma^{-2} \delta w^\top \Gamma \delta w \leq V \delta w^\top D \delta w \leq V \mathcal{E}(w).$$

The variance bound follows from this second-moment bound, not from boundedness alone.  $\square$

For  $n \geq 2$  and  $0 < \eta < 1$ , let

$$\begin{aligned} J_n &= \lceil \log_2 n \rceil, & t_n &= \log \frac{2(J_n + 1)}{\eta}, \\ A_n &= \frac{1 + 64K^2 + (8V + 16/3)t_n}{n}. \end{aligned} \quad (20)$$

**Lemma 3 (Uniform Localized Deviation).** *With probability at least  $1 - \eta/2$ , simultaneously for every sphere predictor with  $e = \mathcal{E}(w) \leq \Delta$ ,*

$$(P - P_n)h_w \leq e/2 + A_n/2. \quad (21)$$

Here  $P$  and  $P_n$  denote population and empirical averaging.

*Proof.* For  $s > 0$ , consider  $\mathcal{H}(\min\{s, \Delta\})$ . Its functions take values in  $[-1, 1]$ , have variance at most  $Vs$  and expected complexity at most  $K\sqrt{s/n}$ . The first inequality of [4, Theorem 2.1], with its tuning parameter equal to one, gives

$$\sup_h (P - P_n)h \leq 4K\sqrt{s/n} + \sqrt{2Vst/n} + \frac{8t}{3n} \quad (22)$$

with failure probability at most  $e^{-t}$ . All suprema are measurable: the relevant sets are compact and the losses continuously parameterized.

Apply (22) at  $s_j = 2^j/n$ ,  $j = 0, \dots, J_n$ , with  $t = t_n$ . The union has failure probability at most  $\eta/2$ . For any local predictor, select  $j = 0$  if  $e \leq 1/n$ , otherwise the smallest  $j$  with  $s_j \geq e$ . Such a  $j$  exists since  $s_{J_n} \geq 1 \geq e$ . With  $x = \max\{e, 1/n\}$ ,  $s_j \leq 2x$ , hence

$$(P - P_n)h_w \leq (4\sqrt{2}K + 2\sqrt{Vt_n})\sqrt{x/n} + 8t_n/(3n).$$

Write  $C = 4\sqrt{2}K + 2\sqrt{Vt_n}$ . Young's inequality gives  $C\sqrt{x/n} \leq x/2 + C^2/(2n)$  and  $C^2 \leq 64K^2 + 8Vt_n$ . Since  $x \leq e + 1/n$ , this is (21).  $\square$

**Theorem 3 (Post-Fit Localized Risk).** *With probability at least  $1 - \eta$ , simultaneously for every  $w \in \mathcal{W}_B^\partial$ , define*

$$g_S^*(w) = [R_n(w) - R_n(w^*)]_+. \quad (23)$$

*If  $2U_n^\partial(\eta/2) + g_S^*(w) \leq \Delta$ , then*

$$\mathcal{E}(w) \leq \min\{1, 2U_n^\partial(\eta/2) + g_S^*(w), A_n + 2g_S^*(w)\}. \quad (24)$$

*The joint parameter bound (14) holds on the same event. In particular, for  $\tilde{w} = S_B(\hat{w})$  satisfying (11), one may replace  $g_S^*$  by  $\rho$  in the condition and bound.*

*Proof.* Intersect the sphere global deviation event with failure probability  $\eta/2$  and the event of Lemma 3. On their intersection, for every sphere predictor,

$$\mathcal{E}(w) \leq R_n(w) - R_n(w^*) + 2U_n^\partial(\eta/2) \leq g_S^*(w) + 2U_n^\partial(\eta/2).$$

The entrance condition places that same predictor in the local region. Using  $P_n h_w \leq g_S^*(w)$  in (21) gives  $\mathcal{E}(w) \leq g_S^*(w) + \mathcal{E}(w)/2 + A_n/2$ . Rearranging proves (24). The event is uniform before  $w$  or its gap is chosen, so a data-dependent gap is allowed and the two events need not be independent. For an approximate empirical minimizer, normalization preserves the tolerance and  $g_S^*(\tilde{w}) \leq \rho$ .  $\square$

**Corollary 2 (Entry-Free Risk Bound).** *Under the same assumptions, set  $\tilde{c}_a = \min\{c_a, \Delta/(4B^2)\}$  and  $\tilde{c}_b = \min\{c_b, \Delta/(4B^2)\}$ . Define  $\tilde{K}, \tilde{V}$  by (19) with these constants, and*

$$\tilde{t}_n = \log \frac{J_n + 1}{\eta}, \quad \tilde{A}_n = \frac{1 + 64\tilde{K}^2 + (8\tilde{V} + 16/3)\tilde{t}_n}{n}.$$

*With probability at least  $1 - \eta$ , simultaneously on the sphere,*

$$\mathcal{E}(w) \leq \min\{1, \tilde{A}_n + 2g_S^*(w)\}. \quad (25)$$

*No entrance check is required; normalized approximate empirical minimizers may replace  $g_S^*$  by  $\rho$ .*

*Proof.* The reduced-constant growth inequality holds on the whole ball. For  $\mathcal{E}(w) \leq \Delta$ , apply Theorem 2; otherwise its proposed right-hand side is at most  $\Delta\|w - w^*\|_2^2/(4B^2) \leq \Delta < \mathcal{E}(w)$ . Lemma 2 therefore holds for  $0 \leq s \leq 1$  with the reduced constants. Repeat the proof of Lemma 3, replacing the cutoff  $\Delta$  by 1 and  $t_n$  by  $\tilde{t}_n$ . Its union failure probability is now  $\eta$ , with no global event needed. Thus  $(P - P_n)h_w \leq \mathcal{E}(w)/2 + \tilde{A}_n/2$ ; use  $P_n h_w \leq g_S^*(w)$  and rearrange.  $\square$

With fixed  $m, \sigma_z, B, \gamma, \varepsilon, \tau_{\max}, \Delta$ , both sets of geometric constants are independent of  $r, k$ . Since  $\text{tr } \Gamma_z = r\sigma_z^2 + m^2$  and  $\|\Gamma_z\|_{\text{op}} = \sigma_z^2 + m^2$ , both statistical terms have order  $(r + \tau^2 k + \log((1 + \log n)/\eta))/n$ , plus the comparison gap or optimization tolerance. Entry enables the original local constants, not this rate order itself. The iterated logarithm comes from the finite grid. The comparison is model-aware:  $w^*, T, \Delta$  and the constants are supplied by the synthetic model, not generally known from real data. Neither fitting nor initialization uses  $w^*$ . A zero gap means only  $R_n(w) \leq R_n(w^*)$ , not global empirical optimality.

## 6 Controlled Experiments

The baseline is  $r = 1, k = 32, m = 2, \sigma_z = B = \gamma = 1, \tau = 0.15, n = 512$  and evaluation attack  $\varepsilon = 0.25$ . Each of 22 settings uses 20 independent samples, paired across full-space robust ramp, principal-space oracle robust ramp, full-space ordinary ramp, normalized robust logistic, and the normalized signed mean  $B \sum_i Y_i X_i / \|\sum_i Y_i X_i\|_2$ . The last is both a standalone baseline and a shared initialization. Except for ordinary ramp and the training-budget sweep, training uses attack 0.25.

*Optimization and Evaluation.* We use SciPy 1.17.1 SLSQP with supplied gradients, `ftol` =  $10^{-10}$ , at most 1000 iterations, and 12 starts: the signed-mean direction at radius  $B$  plus 11 random directions with radii uniform on  $[B/4, B]$ . A signed mean of norm at most  $10^{-14}$  is replaced by a random start. The piecewise ramp gradient averages over all  $n$  samples, with contributions only for  $0 < M < \gamma$  (zero at either kink). The norm-penalty gradient is set to zero at  $\|w\|_2 \leq 10^{-14}$  and to  $\varepsilon w / \|w\|_2$  otherwise. This is an implementation convention, not differentiability at the kinks.

Every terminal candidate is projected to feasibility; the smallest training objective is selected regardless of optimizer status, then normalized. All 1,760 selected main-grid fits succeeded. Across robust-ramp starts, the median objective range was 0.8661; 9.63% of random starts ended at loss one, versus no signed-mean starts. Success does not certify (11). For zero-based setting  $j$  and repetition  $q$ , the sample seed is  $20260919 + 100000j + q$ , and optimizer seeds add 50000 plus the method index in the order above (excluding signed mean). Settings are ordered by the sample-size, fixed- $\tau$  dimension, fixed-energy dimension, signal, training-attack and principal-dimension sweeps; values increase within each sweep. Thus methods share samples, whereas separate sweeps use independent draws even at identical parameters.

Population ramp risks and robust margin errors are evaluated from (3)–(4); the latter is  $\Phi(-\nu_w/\sqrt{v_w})$  for  $v_w > 0$ . The population reference is analytic in the executed grid. Witness settings use Theorem 2; the pure-noise and threshold controls are treated separately. We save every raw and normalized coefficient vector, all optimizer starts, and the source hash, seed and dependency versions. Independent Gaussian quadrature and reconstruction of all 3,520 coefficient vectors agreed with recorded metrics to within  $2.23 \times 10^{-16}$ .

Normalization never increases ramp loss but increased the logistic training objective by more than  $10^{-9}$  in 29 of 440 fits (maximum 0.19508); this method is labelled *normalized robust logistic*. Positive scaling preserves robust margin signs, not surrogate risks. Only exact zero invokes  $v_0 = e_1$ ; otherwise max-absolute-coordinate scaling avoids underflow, without a positive norm threshold.

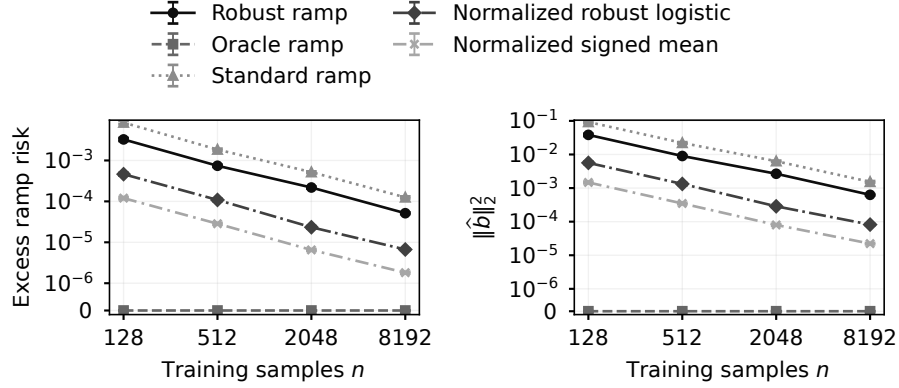
## 6.1 Sample Size and Tail Energy

Excess risk and tail weight decrease with  $n$  (Fig. 1, Table 1), approximately fourfold per quadrupling over most of this grid. Four sample sizes do not identify an exponent or validate bound constants.

**Table 1.** Normalized robust ramp, 20 repetitions.  $G_n = \min\{1, 2U_n^\partial(\eta/2) + g_S^*\}$  at  $\eta = 0.05$ ; all recorded gaps are zero.  $A_n$  is shown only after entry; a dash means inapplicable

| $n$  | Mean $\mathcal{E}$ | Mean $\ \hat{b}\ _2^2$ | $G_n$    | $A_n$    |
|------|--------------------|------------------------|----------|----------|
| 128  | 0.003268           | 0.038356               | 1.000000 | —        |
| 512  | 0.000742           | 0.009011               | 0.553621 | —        |
| 2048 | 0.000218           | 0.002669               | 0.276811 | —        |
| 8192 | 0.000051           | 0.000630               | 0.138405 | 1.152511 |

At fixed  $\tau = 0.15$ , increasing  $k$  from 8 to 512 increases total tail energy from 0.18 to 11.52, excess risk from 0.000233 to 0.011574, and squared tail norm from 0.002851 to 0.124352. In contrast, setting  $\tau = \sqrt{0.72/k}$  keeps energy fixed. At  $k = 8, 32, 128, 512$ , the excess risks are 0.000905, 0.000851, 0.000832 and 0.000841; squared tail norms remain between 0.01007 and 0.01050 (Fig. 2). These controlled



**Fig. 1.** Sample-size sweep, including the signed-mean baseline: means and standard errors over 20 paired samples. Vertical axes are logarithmic above  $10^{-6}$  and linear near zero, retaining exact oracle zeros without an added offset

comparisons support aggregate tail energy as a more informative quantity than coordinate count alone. The robust classification error in the sample-size sweep changes only from 0.040557 to 0.040067, illustrating why risk and coefficient diagnostics are reported separately.

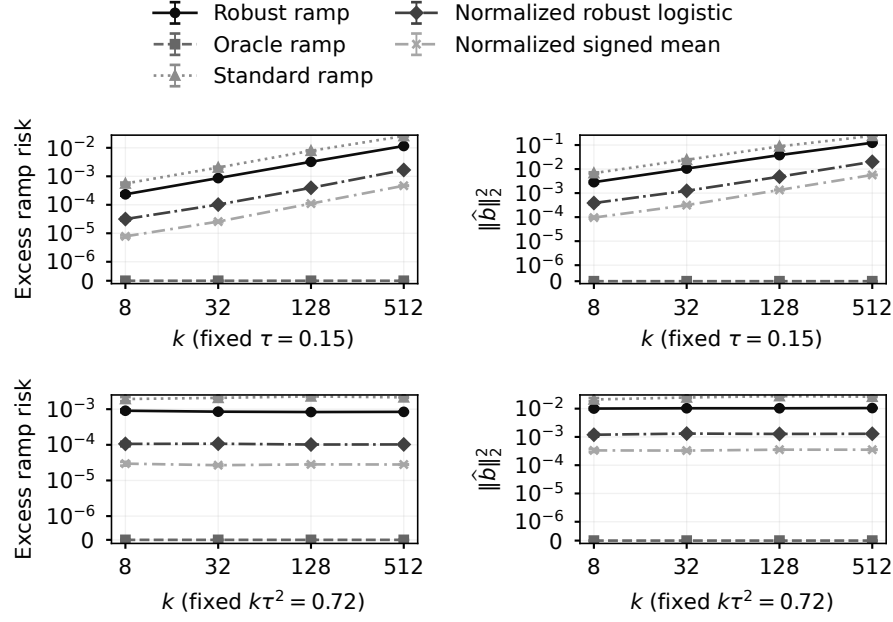
## 6.2 Entry, Constants and Boundary Controls

All 80 robust-ramp fits in the sample-size sweep have  $g_S^* = 0$ . Only  $n = 8192$  satisfies the sufficient sphere entrance condition:  $2U_n^\partial(\eta/2) = 0.138405 < \Delta = 0.192503$ . The confidence statement is for one sample and fixed configuration, not joint 95% coverage of the grid.

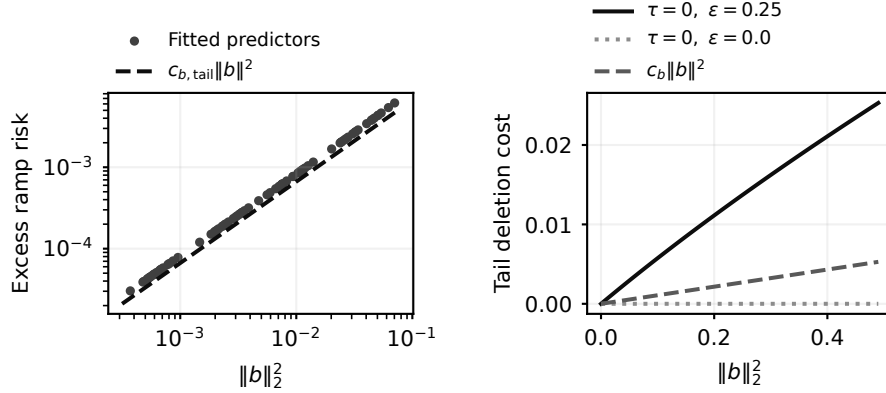
The path constant is  $\kappa = 0.0862773$ , giving  $c_b = 0.0107847$ . At  $n = 8192$ ,  $\eta = 0.05$ , the local term is 1.33670 with the general isotropic  $c_a$ , or 1.15251 with the sharper scalar  $c_a = 0.111634$ . The entry-free term is  $\bar{A}_n = 1.90874$ : removing entry does not make the bound numerically useful here. The operative Theorem 3 bound remains 0.138405. Dividing by  $c_b$  or the stronger  $c_{b,\text{tail}} = 0.0666018$  gives 12.8335 or 2.0781, both worse than the feasibility cap  $B^2 = 1$ .

Structural tightness is different. Using the actual unrounded population excesses at  $n = 8192$  gives  $\bar{\mathcal{E}}/c_{b,\text{tail}} = 0.000772$ , versus observed  $\|\hat{b}\|_2^2 = 0.000630$ . Figure 3 checks the inequality for all 80 fitted sample-size predictors. In this configuration, the statistical risk bound is much looser than the structural conversion. This uses known population risk after fitting, not a training-time certificate.

For a paired  $\tau = 0$  control, we reuse the 20  $n = 512$  labels, principal features and optimizer seeds, changing only tail variance. All 80 selected control fits succeeded; normalized full-space robust ramp has exactly zero tail weight and excess risk in all 20 samples. Independent quadrature and reconstruction of its



**Fig. 2.** Dimension and energy controls, with the same five methods and vertical scales as Fig. 1. Top: fixed  $\tau = 0.15$ . Bottom: fixed  $k\tau^2 = 0.72$ . Bars are standard errors over 20 paired samples



**Fig. 3.** Direct structural checks. Left: 80 fitted predictors and the scalar sphere lower bound (logarithmic axes); all satisfy  $\mathcal{E} \leq \Delta$ . Right: zero-tail-variance deletion paths at fixed  $a = 0.7$ ,  $0 \leq b \leq 0.7$ , with  $m = 2$ ,  $\sigma_z = B = \gamma = 1$ . Positive attack alone creates a tail cost; without attack it vanishes

160 raw/normalized method vectors agree within  $1.39 \times 10^{-17}$ . The fixed- $a$  paths in Fig. 3 isolate the attack penalty from the sphere’s principal-weight reallocation.

With common evaluation budget 0.25, training budgets 0, 0.25, 0.5 yield mean squared tail norms 0.024422, 0.009204, 0.004544. Increasing  $r$  from 1 to 4 to 16 raises excess risk from 0.000780 to 0.001771 to 0.007604, while tail weight stays near 0.009. These sweeps separate the training penalty and principal estimation cost from tail energy; they do not establish a universal advantage of larger attacks.

The signal sweep uses  $m \in \{0, 0.75, 1.25, 2\}$ . The first two settings lack a positive witness: at  $m = 0.75$ ,  $B(m - \varepsilon) = \gamma/2$  and the aligned ramp risk is  $1/2$ , although its robust classification error is  $\Phi(-0.5) = 0.30854 < 1/2$ . The witness condition concerns the favorable ramp region, not the possibility of nontrivial robust classification. At  $m = 0$ , ordinary binary classification error is  $1/2$ , while the fitted robust margin error is approximately 0.599.

*Initialization Versus the Fitted Predictor.* The normalized empirical signed mean is already one of the initializations. Comparing it with the selected normalized robust-ramp output on the same samples shows lower training risk but higher population risk and tail weight in all 80 sample-size trials. At  $n = 512$ , mean training risk changes from 0.118693 to 0.118030, while excess risk rises from 0.0000284 to 0.0007420 and squared tail norm from 0.0003484 to 0.0090111. This directly exposes the empirical fitting incentive to use tail noise. In this grid, the simple initialization generalizes better; the theory characterizes full-space empirical fitting rather than asserting its superiority.

*Why the Witness Matters.* For pure noise  $X \sim N(0, \Sigma)$ , (4) along  $w = cu$  has integrand  $\Phi((t/c + \varepsilon)/\sqrt{u^\top \Sigma u})$ , decreasing in  $c > 0$  and positive directional variance. Thus, when  $\lambda_{\max}(\Sigma) > 0$ , minimizers have norm  $B$  in a top-eigenvalue direction. Unconditional tail deletion is false: for  $\mu = 0$ ,  $\Lambda = 1$ ,  $a = 0.1$ ,  $\tau = 2$ ,  $\varepsilon = 0.2$ ,  $B = \gamma = 1$ , changing  $b$  from 0 to 0.5 lowers risk from 0.969311 to 0.717434. The witness is absent.

## 7 Discussion

A tail-free population optimum does not prevent finite-sample tail use. The witness-based deletion inequality and isotropic geometry yield statistical bounds, while the experiments expose empirical fitting incentives relative to a simple initialization.

The finite-sample constants remain conservative; general anisotropic growth and global optimization guarantees are unresolved. No same-task minimax lower bound is established: a principal-subspace oracle can remove the tail entirely. A vanishing attack, informative tail or non-Euclidean geometry requires separate arguments. Our bounds concern ramp excess risk. Although the isotropic model shares an optimal direction with robust classification, a quantitative excess-risk conversion has not been established here.

## References

1. Awasthi, P., Chatziafratis, V., Chen, X., Vijayaraghavan, A.: Adversarially robust low dimensional representations. In: Proceedings of the 34th Conference on Learning Theory, *Proceedings of Machine Learning Research*, vol. 134, pp. 237–325. PMLR (2021). URL <https://proceedings.mlr.press/v134/awasthi21a.html>
2. Awasthi, P., Frank, N., Mohri, M.: Adversarial learning guarantees for linear hypotheses and neural networks. In: Proceedings of the 37th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 119, pp. 431–441. PMLR (2020). URL <https://proceedings.mlr.press/v119/awasthi20a.html>
3. Bao, H., Scott, C., Sugiyama, M.: Calibrated surrogate losses for adversarially robust classification. In: Proceedings of the Thirty Third Conference on Learning Theory, *Proceedings of Machine Learning Research*, vol. 125, pp. 408–451. PMLR (2020). URL <https://proceedings.mlr.press/v125/bao20a.html>
4. Bartlett, P.L., Bousquet, O., Mendelson, S.: Local Rademacher complexities. *The Annals of Statistics* **33**(4), 1497–1537 (2005). DOI 10.1214/009053605000000282. URL <https://arxiv.org/abs/math/0508275>
5. Chen, J., Cao, Y., Gu, Q.: Benign overfitting in adversarially robust linear classification. In: Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, *Proceedings of Machine Learning Research*, vol. 216, pp. 313–323. PMLR (2023). URL <https://proceedings.mlr.press/v216/chen23b.html>
6. Dan, C., Wei, Y., Ravikumar, P.: Sharp statistical guarantees for adversarially robust Gaussian classification. In: Proceedings of the 37th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 119, pp. 2345–2355. PMLR (2020). URL <https://arxiv.org/abs/2006.16384>
7. Javanmard, A., Mehrabi, M.: Adversarial robustness for latent models: Revisiting the robust-standard accuracies tradeoff. *Operations Research* **72**(3), 1016–1030 (2024). DOI 10.1287/opre.2022.0162. URL <https://arxiv.org/abs/2110.11950>. First published online in 2023
8. Javanmard, A., Soltanolkotabi, M.: Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics* **50**(4), 2127–2156 (2022). DOI 10.1214/22-AOS2180. URL <https://arxiv.org/abs/2010.11213>
9. Mustafa, W., Lei, Y., Kloft, M.: On the generalization analysis of adversarial learning. In: Proceedings of the 39th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 162, pp. 16,174–16,196. PMLR (2022). URL <https://proceedings.mlr.press/v162/mustafa22a.html>
10. Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., Madry, A.: Adversarially robust generalization requires more data. In: *Advances in Neural Information Processing Systems*, vol. 31 (2018). URL <https://arxiv.org/abs/1804.11285>
11. Tanner, K., Vilucchio, M., Loureiro, B., Krzakala, F.: A high dimensional statistical model for adversarial training: Geometry and trade-offs. In: Proceedings of the 28th International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, vol. 258, pp. 2530–2538. PMLR (2025). URL <https://proceedings.mlr.press/v258/tanner25a.html>
12. Yin, D., Ramchandran, K., Bartlett, P.L.: Rademacher complexity for adversarially robust generalization. In: Proceedings of the 36th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 97, pp. 7085–7094. PMLR (2019). URL <https://arxiv.org/abs/1810.11914>