Toward a capacity upper bound of 1.9

A self-contained improvement to $\alpha_- \le 2 - 10^{-29}$

Symmetric storage of random binary patterns with free biases

Research continuation: derivations, correctness checks, and the remaining target

Abstract

For optimal strict storage of iid unbiased binary patterns by real symmetric zero-diagonal weights with arbitrary biases, the research argument developed here yields $\alpha_- \le 2 - 10^{-29}$. It improves the previous manuscript's $2 - 10^{-89}$ bound, while remaining far from the target 1.9. The main proof is self-contained: its new ingredients are an explicit positive-map gap from projector incoherence, a one-feature lifting bound for the hard margin, and an elementary ordinary singular-value estimate. Sharper Gaussian aperture and normalization estimates complete the numerical certificate. The dependent full-network transfer and its probability inputs are included. A separate optional result at load 1.9 invokes a published random-matrix theorem and concerns selected centers only. The final part gives an exact global-dual target, rigorous finite certificate checks, and a critical assessment of what must change to reach a macroscopic improvement. This is a working research proof, not a claimed result from the literature.

Contents

1 Result, exact target, and reading guide **3**

I The new estimates and the explicit upper bound **4**

2 A row estimate with explicit smoothing error **5**

3 Fourier projection and the exact reciprocity loss **7**

4 A population kernel gap of 0.032 **8**

5 Asymptotic interpretation and remaining bottleneck **8**

6 Positive selectors, margins, and leakage **9**

7 From a population gap to a full-network gap **10**

8 Sharper Gaussian aperture and internal normalization **10**

9 An explicit improved scalar certificate **13**

10 A self-contained margin bound by adding one feature **15**

II Self-contained probability and transfer proofs **16**

11 Definitions and Gaussian comparison facts **16**

12 Gaussian comparison and concentration, with proofs **16**

13 Concentration and partial Gaussian replacement **18**

14 Uniform exclusion of heavy coefficient mass **20**

15 Positive-orthant coercivity **21**

16 Head–tail dissection and a replacement lemma **22**

17 The Gaussian comparison and its explicit auxiliary lower bound **23**

18 One-sided concentration and the support union **25**

19 An elementary ordinary singular-value bound of $1/25$ **26**

20 Regularity of the exterior projection **29**

21 Technical proof of the full-network transfer **30**

22 Two-sided kernel sampling and graph testability **31**

23 Smooth positive selectors and feature response **32**

24 The actual full-row center matrix **33**

III Progress at 1.9 and the remaining research problem **34**

25 An auxiliary center incompatibility result at load 1.9 **34**

26 A global positive-dual criterion and exact finite certificates **38**

27 What prevents this argument from reaching 1.9? **41**

28 Literature audit and model distinctions **43**

1 Result, exact target, and reading guide

The improved conclusion of the research argument is

$$\boxed{\alpha_- \le 2 - 10^{-29}.}$$

The earlier explicit conclusion was $2 - 10^{-89}$, so the deficit from two increases by a factor 10^{60}. This remains extremely small: the present argument does *not* prove $\alpha_- \le 1.9$.

Proof status. This document consolidates the new estimates and the probabilistic mechanisms they use. The new positive-map inequality, sharper aperture constants, internal normalization, margin lifting, ordinary singular-value estimate, and scalar certificate have been checked independently within this investigation. The full-network transfer was rechecked as well. This is a research proof offered for further mathematical scrutiny, not a bound attributed to the published literature. All probabilistic estimates used for the displayed upper bound are proved in the document. Adding one auxiliary feature gives a direct mean-margin bound, and a stronger elementary singular-value estimate makes it sufficient. The earlier Oymak–Tropp universality input is no longer needed for this bound.

An original 1990 paper by Nardulli and Pasquariello asserts the usual Gardner saturation formula in the symmetric model. Its cited derivation was not available in full for this investigation. A 2022 treatment emphasizes that the fully connected many-pattern problem is not settled by the usual diluted replica calculation. This historical conflict remains unresolved here; it is a reason to audit a new sub-two proof carefully, not permission to silently discard either statement. Precise references appear at the end.

1.1 What the mathematical problem asks

Let $S = (s_i^\mu) \in \{-1, 1\}^{n \times m}$ have independent uniform entries. Its columns are the memories. We may choose real interaction strengths J_{ij} and real biases h_i, subject to

$$J = J^\top, \qquad J_{ii} = 0.$$

Strict storage means

$$s_i^\mu \left(\sum_{j \ne i} J_{ij} s_j^\mu + h_i \right) > 0 \quad \text{for every neuron } i \text{ and memory } \mu. \tag{1}$$

The expression in parentheses is the field at neuron i. Multiplying by its desired sign makes a correctly stored bit have positive margin. The symmetry restriction couples two classifiers: the coefficient of neuron j in classifier i must equal the coefficient of neuron i in classifier j.

For fixed load u, put $m = \lfloor un \rfloor$ and define

$$\alpha_- := \sup\{u : \mathbb{P}(\text{strict storage at load } u) \to 1\}.$$

We prove a limiting storage probability at most 10^{-3} at $u = 2 - 10^{-29}$. This suffices for the bound: storage is monotone in the number of memories. No existence of a sharp threshold is assumed. The probability of repeated columns is at most $\binom{m}{2} 2^{-n} = o(1)$, so conditioning on distinct memories does not change the conclusion.

1.2 Linear programming and the exact impossibility certificate

Collect the $n(n-1)/2$ free edges and n biases in a vector θ. There is a matrix $A = A(S)$ with nm rows such that the margins are $A\theta$. The row indexed by (i, μ) has coefficient $s_i^\mu s_j^\mu$ on edge $\{i, j\}$, coefficient s_i^μ on bias i, and zero elsewhere. Thus storage is the linear inequality problem $A\theta > 0$. Because there are finitely many inequalities and they are homogeneous, this is equivalent to $A\theta \geq \mathbf{1}$ after scaling θ.

Exactly one of the following alternatives holds:

$$\exists \theta : A\theta > 0; qquad \exists \lambda \geq 0, \lambda \neq 0 : A^\top \lambda = 0.$$

For completeness, let a_r be the rows of A. If zero belongs to their convex hull, its convex-combination coefficients give the second alternative. If it does not, the point v of that compact convex hull closest to zero satisfies $v \neq 0$ and $a_r^\top v \geq \|v\|^2 > 0$ for every r, giving the first alternative. They cannot both hold because $\lambda^\top A\theta$ would be strictly positive and zero. Written out, the dual equations are

$$\sum_\mu \lambda_{i\mu} s_i^\mu = 0, \qquad \sum_\mu (\lambda_{i\mu} + \lambda_{j\mu}) s_i^\mu s_j^\mu = 0 \quad (i < j). \tag{2}$$

There is no diagonal equation, because self-interactions are forbidden.

1.3 Why the improved proof works

For one neuron, all coefficient vectors satisfying its inequalities form a convex cone. Near load two, its unit directions are close to one another. Choose a positive, smoothly regularized simplex optimizer for each neuron; its residual points close to that neuron's maximum-margin direction. The new lemma shows that these positive selectors necessarily contain enough nonlinear dependence on their random labels to produce a uniform mismatch between reciprocal coefficients. It improves a previously tiny population constant to a gap of at least 0.032.

The full-network passage retains all dependencies between neurons. It uses conditioning on an exterior set, a controlled response to finitely many added features, and uniform sampling of a symmetric kernel. Finally, if the original symmetric network existed, its rows would be close enough to the selected rows that this mismatch could not occur. The final strict comparison is $0.0294 < 0.03$.

The essential limitation is that the proof still requires the *entire* one-neuron cone to be narrow. The final part develops a different global dual criterion that does not impose that requirement and is the preferred next route toward 1.9.

Notation and how to read. $\|\cdot\|$ is Euclidean norm; $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_F$ are operator and Frobenius norms. A Rademacher variable is a uniform sign. $x_- := \max\{-x, 0\}$. A projector is an orthogonal projection. The main derivation and all numerical gates appear before the technical appendices. Readers new to the methods can first read this section, the row negative-energy proof, the final symmetry comparison, and the global dual section; the appendices then supply each probabilistic input.

Part I

The new estimates and the explicit upper bound

For a real number x, write $x_- = \max\{-x, 0\}$. Let Y be uniform on $\{-1, 1\}^m$, and let P be an orthogonal projector on $\mathbb{R}^m$. Throughout the quantified statements assume

$$\max_i P_{ii} \leq p_0 := 0.500001, \qquad \max_{i \neq j} |P_{ij}| \leq \mu_0 := 10^{-16}. \tag{3}$$

No Gaussian assumption on P or on its range is made.

2 A row estimate with explicit smoothing error

Lemma 2.1 (Rademacher-to-Gaussian comparison)**.** *Let $X = \sum_{j=1}^{r} b_j \varepsilon_j$, where the ε_j are independent uniform signs, $\sum_j b_j^2 = q \le 1$, and $\max_j |b_j| \le \eta$. For every real a and every $\tau > 0$, with G a standard Gaussian,*

$$\mathbb{E}(a + X)_-^2 \ge \mathbb{E}(a + \sqrt{q}G)_-^2 - \tau^2 - \frac{0.35\eta}{\tau}.$$

The estimate is uniform in a, in the number of summands, and in their coefficients.

Proof. Put $F(x) = x_-^2$ and $F_\tau(x) = \mathbb{E}F(x + \tau G)$. Convexity and the fact that F' is 2-Lipschitz give

$$0 \le F_\tau(x) - F(x) \le \tau^2.$$

Direct differentiation of the Gaussian convolution gives

$$\sup_x |F_\tau'''(x)| \le \frac{\sqrt{2/\pi}}{\tau}.$$

Replace each $b_j\varepsilon_j$ by b_jG_j, one at a time. Taylor's theorem through second order applies conditionally on all other summands. The means and variances agree, and therefore the sum of the absolute expectation errors is at most

$$\frac{\sqrt{2/\pi}}{6\tau}(1 + \mathbb{E}|G|^3) \sum_j |b_j|^3 < \frac{0.35}{\tau}\eta \sum_j b_j^2 \le \frac{0.35\eta}{\tau}.$$

Here $\mathbb{E}|G|^3 = 2\sqrt{2/\pi}$. Undo smoothing on the Rademacher side only; on the Gaussian side, $F_\tau \ge F$. This proves the stated one-sided estimate. □

Lemma 2.2 (Restricted coordinate mass)**.** *If $v = Pv$ and $H \subseteq [m]$, then*

$$\sum_{j \in H} v_j^2 \le \left(\max_j P_{jj} + (|H| - 1) \max_{j \ne k} |P_{jk}|\right) \|v\|^2 .$$

In particular $|v_i| \le \sqrt{P_{ii}}\, \|v\|$.

Proof. Let E_H restrict a vector to its coordinates in H. Since $v = Pv$,

$$\|E_H v\|^2 \le \|E_H P\|_{\mathrm{op}}^2 \|v\|^2 = \|P_{HH}\|_{\mathrm{op}} \|v\|^2 .$$

The operator norm of the positive semidefinite matrix P_{HH} is bounded by its maximum absolute row sum. The last assertion is the case $H = \{i\}$. □

Theorem 2.3 (Uniform row negative energy)**.** *If (3) holds, then, for every $i \in [m]$ and every $v = Pv$,*

$$\mathbb{E}\left(Y_i \sum_j v_j Y_j\right)_-^2 \ge \frac{3}{80} \|v\|^2 . \tag{4}$$

Consequently, every real matrix N satisfying $N = NP$ obeys

$$\mathbb{E} \sum_i (Y_i (NY)_i)_-^2 \ge \frac{3}{80} \|N\|_{\mathrm{F}}^2 . \tag{5}$$

The matrix N need not be symmetric.

Proof. By homogeneity assume $\|v\| = 1$. The variables Y_iY_j, $j \neq i$, are mutually independent uniform signs. Set $\eta = 10^{-6}$ and let

$$H = \{i\} \cup \{j : |v_j| > \eta\}.$$

Then $|H| \leq 1 + \eta^{-2}$. Lemma 2.2 gives

$$\sum_{j \in H} v_j^2 \leq 0.500001 + 10^{12} \cdot 10^{-16} = 0.500101.$$

Thus the tail $X = \sum_{j \notin H} v_j Y_i Y_j$ has variance $q \geq 0.499899$ and all its coefficients have absolute value at most η. The remaining nonconstant head sum is centered and independent of X. Jensen's inequality, conditional on X, drops this sum and gives

$$\mathbb{E}\left(v_i + \sum_{j \neq i} v_j Y_i Y_j\right)_-^2 \geq \mathbb{E}(v_i + X)_-^2.$$

Apply Lemma 2.1 with $\tau = 0.01$. Its error is at most

$$10^{-4} + \frac{0.35 \cdot 10^{-6}}{0.01} = 0.000135.$$

The Gaussian expectation $\mathbb{E}(a + \sigma G)_-^2$ decreases with a and increases with $\sigma \geq 0$; the latter follows either by differentiation or by adding an independent Gaussian and using convexity. Since $v_i \leq \sqrt{0.500001}$, the remaining Gaussian expectation is at least

$$B := \mathbb{E}\left(\sqrt{0.500001} + \sqrt{0.499899}G\right)_-^2.$$

For clarity, the numerical comparison proving $B - 0.000135 > 3/80$ is recorded explicitly. At $a_* = \sigma_* = 1/\sqrt{2}$,

$$B_* := \mathbb{E}(a_* + \sigma_* G)_-^2 = \Phi(-1) - \tfrac{1}{2}\phi(1) > 0.03766989, \qquad \sqrt{B_*} < 0.195.$$

The difference of the two Gaussian random variables has L^2 norm less than 0.000073, because

$$0 < \sqrt{0.500001} - 1/\sqrt{2} < 0.000001, \quad 0 < 1/\sqrt{2} - \sqrt{0.499899} < 0.000072.$$

The negative-part function is 1-Lipschitz. The reverse triangle inequality in L^2 therefore yields

$$B \geq (\sqrt{B_*} - 0.000073)^2 > 0.03766989 - 2(0.195)(0.000073) = 0.03764142.$$

Subtracting 0.000135 gives $0.03750642 > 0.0375 = 3/80$. For example, the elementary numerical enclosures

$$0.2419707245 < \phi(1) < 0.2419707246, \qquad 0.1586552539 < \Phi(-1) < 0.1586552540$$

justify the displayed bounds on B_*. They follow by applying the alternating series to $e^{-1/2}$ and to $\int_0^1 e^{-x^2/2}\,dx = \sum_{k \geq 0} (-1)^k/[2^k k!(2k+1)]$, with the standard alternating remainder bound; π can likewise be enclosed by Machin's identity $\pi = 16\arctan(1/5) - 4\arctan(1/239)$ and its alternating series. This proves (4). Apply it separately to every row of $N = NP$ and sum to obtain (5); zero rows contribute zero. There is no exceptional-row removal or unproved uniformity step. □

3 Fourier projection and the exact reciprocity loss

Theorem 3.1 (Positive-map inequality)**.** *Assume* (3)*. Let* $f : \{-1,1\}^m \to \mathbb{R}^m$ *satisfy* $Y_i f_i(Y) \geq 0$ *for every* i *and every* Y*. Define*

$$F = \mathbb{E}\,\|f\|^2\,, \quad M = \mathbb{E}[f(Y)Y^\top], \quad R = \mathbb{E}\,\|f - MY\|^2\,, \quad s = \|\operatorname{skew} M\|_{\mathrm{F}}^2\,, \quad e = \|(I-P)M\|_{\mathrm{F}}^2\,.$$

Then, with $d = 3/80$*,*

$$R + \|M(I-P)\|_{\mathrm{F}}^2 \geq dF, \tag{6}$$

$$F - \operatorname{tr}(M^2) = R + 2s \geq \left(\sqrt{dF} - \sqrt{e}\right)_+^2, \tag{7}$$

$$R + s + e \geq \frac{1}{80}F \geq \frac{1}{80}\,\|M\|_{\mathrm{F}}^2\,. \tag{8}$$

Proof. The coordinates Y_j form an orthonormal family in L^2. Hence MY is the orthogonal projection of f onto vector-valued homogeneous linear functions, and

$$R = F - \|M\|_{\mathrm{F}}^2\,.$$

Put $Q = I - P$ and $N = MP$. Let $a_N(Y)$ be the coordinatewise projection of NY onto the orthant with signs Y, so that $[a_N(Y)]_i = Y_i(Y_i(NY)_i)_+$. Pointwise sign alignment implies

$$f(Y)^\top NY \leq f(Y)^\top a_N(Y).$$

Since $N = MP$ is the Frobenius-orthogonal projection of M onto matrices whose rows belong to the range of P, Fourier orthogonality, Cauchy–Schwarz, and Theorem 2.3 give

$$\|N\|_{\mathrm{F}}^2 = \mathbb{E}[f(Y)^\top NY] \leq \sqrt{F}\sqrt{\mathbb{E}\,\|a_N(Y)\|^2} \leq \sqrt{F}\sqrt{\left(1 - \frac{3}{80}\right)\|N\|_{\mathrm{F}}^2}.$$

Thus $\|N\|_{\mathrm{F}}^2 \leq (1 - 3/80)F$, including the case $N = 0$. The identity $R + \|MQ\|_{\mathrm{F}}^2 = F - \|N\|_{\mathrm{F}}^2$ proves (6). We next record carefully the required comparison between row and column leakage:

$$\|MQ\|_{\mathrm{F}} \leq \sqrt{2s} + \sqrt{e}. \tag{9}$$

Choose orthonormal coordinates adapted to the ranges of P, Q and write

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

Then $e = \|C\|_{\mathrm{F}}^2 + \|D\|_{\mathrm{F}}^2$ and $2s \geq \left\|B - C^\top\right\|_{\mathrm{F}}^2$. Consequently

$$\begin{aligned} \|MQ\|_{\mathrm{F}} &= \sqrt{\|B\|_{\mathrm{F}}^2 + \|D\|_{\mathrm{F}}^2} \\ &\leq \left\|B - C^\top\right\|_{\mathrm{F}} + \sqrt{\|C\|_{\mathrm{F}}^2 + \|D\|_{\mathrm{F}}^2} \leq \sqrt{2s} + \sqrt{e}. \end{aligned}$$

Let $g = R + 2s$. From (6) and (9),

$$dF \leq R + (\sqrt{2s} + \sqrt{e})^2 \leq g + 2\sqrt{ge} + e = (\sqrt{g} + \sqrt{e})^2.$$

This proves (7), including the case $e \geq dF$. The identity $F - \operatorname{tr}(M^2) = R + 2s$ follows by decomposing M into its symmetric and skew-symmetric parts. Finally,

$$dF \leq R + (\sqrt{2s} + \sqrt{e})^2 \leq R + 3(s+e) \leq 3(R + s + e),$$

which proves (8). □

4 A population kernel gap of 0.032

Corollary 4.1 (Mean leakage suffices)**.** *Assume* (3)*. Let $u(Y)$ be a unit vector for every Y and satisfy $Y_i u_i(Y) \geq 0$. Suppose*

$$\mathbb{E}\,\|(I-P)u(Y)\|^2 \leq 10^{-12}.$$

Define the symmetric kernel

$$h(x,y) = (u(x)^\top y)(u(y)^\top x)$$

on the uniform cube. Then its integral operator obeys

$$\|T_h\|_{\mathrm{op}} \leq 1, \qquad \mathbb{E}|h(X,Y)|^4 \leq 105, \qquad \sup_{g\geq 0,\ \mathbb{E}g^2=1} \langle g, T_h g\rangle \leq 0.968.$$

Proof. For arbitrary real g, set $f = gu$ and $M_g = \mathbb{E}[guY^\top]$. Expanding the kernel gives $\langle g, T_h g\rangle = \mathrm{tr}(M_g^2)$. Fourier orthogonality implies $|\,\mathrm{tr}(M_g^2)| \leq \|M_g\|_{\mathrm{F}}^2 \leq \mathbb{E}g^2$, proving the operator norm bound. For a fixed unit vector v, expansion into even monomials gives $\mathbb{E}(v^\top Y)^8 \leq 105$. Cauchy–Schwarz therefore gives

$$\mathbb{E}|h(X,Y)|^4 \leq \sqrt{\mathbb{E}(u(X)^\top Y)^8\,\mathbb{E}(u(Y)^\top X)^8} \leq 105.$$

Let

$$G_0 = \{y : \|(I-P)u(y)\|^2 \leq 10^{-4}\}, \qquad B_0 = G_0^c.$$

Markov's inequality gives $p := \mathbb{P}(B_0) \leq 10^{-8}$. For $g \geq 0$ supported on G_0, Fourier contraction gives

$$\|(I-P)M_g\|_{\mathrm{F}}^2 \leq \mathbb{E}\,\|(I-P)g(Y)u(Y)\|^2 \leq 10^{-4}\mathbb{E}g^2.$$

Theorem 3.1, and $\sqrt{3/80} > 0.1936$, now yield

$$\langle g, T_h g\rangle \leq [1 - (\sqrt{3/80} - 0.01)^2]\mathbb{E}g^2 \leq 0.9663\mathbb{E}g^2,$$

because $(0.1936 - 0.01)^2 = 0.03370896 > 0.0337$. The Hilbert–Schmidt bounds for the other blocks are

$$v^2 := \|T_{G_0B_0}\|_{\mathrm{op}}^2 \leq \sqrt{105}\sqrt{p} \leq 0.0011, \qquad w := \|T_{B_0B_0}\|_{\mathrm{op}} \leq 105^{1/4}\sqrt{p} \leq 0.0004.$$

For an arbitrary nonnegative g, put $a = \|g1_{G_0}\|_2$ and $b = \|g1_{B_0}\|_2$. Its quadratic form is at most

$$0.9663a^2 + 2vab + 0.0004b^2 \leq 0.968(a^2+b^2).$$

The last inequality holds because the symmetric 2×2 matrix with diagonal entries $0.968 - 0.9663 = 0.0017$ and $0.968 - 0.0004 = 0.9676$ and off-diagonal entries $-v$ is positive semidefinite: its determinant is at least $0.0017 \cdot 0.9676 - 0.0011 = 0.00054492 > 0$. □

5 Asymptotic interpretation and remaining bottleneck

The explicit constants are deliberately conservative. If projectors P_m have $\max_i (P_m)_{ii} \leq p + o(1)$ for fixed $p < 1$ and $\max_{i\neq j} |(P_m)_{ij}| = o(1)$, the same proof, taking the fixed head cutoff η small and then m large, gives uniformly over $N = NP_m$:

$$\mathbb{E}\sum_i (Y_i(NY)_i)_-^2 \geq [c(p) - o(1)]\,\|N\|_{\mathrm{F}}^2\,,$$

where

$$c(p) = \mathbb{E}(\sqrt{p} + \sqrt{1-p}G)_-^2 = \Phi\Big(-\sqrt{\frac{p}{1-p}}\Big) - \sqrt{p(1-p)}\,\phi\Big(\sqrt{\frac{p}{1-p}}\Big)\,.$$

The $o(1)$ is uniform in N; the quantified row argument proves this directly, without a matrix net. For $p = 1/2$, $c(p) = 0.0376698916\ldots$.

The preceding result supplies a much larger population reciprocity gap than the previous diagonal-plus-matching argument. To turn it into a capacity upper bound, one must still establish the full-center transfer and a sufficiently small whole-cone radius at the selected load. A population gap alone is not a proof of the target $\alpha_- \le 1.9$.

6 Positive selectors, margins, and leakage

For a signed feature matrix $B \in \{-1,1\}^{d\times m}$ let $\Delta_m = \{\pi \ge 0 : \mathbf{1}^\top \pi = 1\}$ and

$$\kappa = \min_{\pi\in\Delta_m} \|B\pi\|, \qquad g_* = B\pi_*.$$

The point g_* is unique even when the minimizing weights are not. The projection inequality $B^\top g_* \ge \kappa^2 \mathbf{1}$ proves that, when $\kappa > 0$, g_*/κ is a feasible maximum-margin unit vector. Conversely every unit w has minimum margin at most $w^\top B\pi_* \le \kappa$.

Choose fixed $a, \eta, \tau > 0$ and define the unique positive selector

$$\pi^{\mathrm{reg}} = \arg\min_{\pi>0,\ \mathbf{1}^\top\pi=1} \left\{ \|B\pi\|^2 + \eta m\|\pi\|^2 - \frac{\tau}{m}\sum_\mu \log(m\pi_\mu) \right\}.$$

The barrier diverges at the simplex boundary, ensuring existence in its interior; the ridge term is strictly convex, ensuring uniqueness. The barrier term is nonnegative by Jensen's inequality.

Positive-orthant coercivity in the appendix gives $\|B\pi\| \ge a_0\sqrt{d}\|\pi\|$ for every $\pi \ge 0$, with exponentially high probability. The uniform comparator has residual at most $L_0 \le 10$. For $m/d < 2$ this yields

$$\kappa \ge a_0/\sqrt{2}, qquad m\|\pi_*\|^2 \le U := 200/a_0^2.$$

Mix the hard optimizer with the uniform vector: $\pi_a = (1-a)\pi_* + a\mathbf{1}/m$. Then $m(\pi_a)_\mu \ge a$ and $m\|\pi_a\|^2 \le U$. Objective comparison and the projection inequality give

$$\begin{aligned} \|B\pi^{\mathrm{reg}} - g_*\|^2 &\le \|B\pi^{\mathrm{reg}}\|^2 - \kappa^2 \\ &\le 2a\kappa L_0 + a^2L_0^2 + \eta U + \tau\log(1/a) =: D. \end{aligned} \tag{10}$$

The normalized full coefficient directions are consequently within $r_{\mathrm{reg}} \le 2\sqrt{D}/\kappa_{\min}$.

Now take an exterior feature matrix X, including the constant bias row, and an independent sign label vector y. Write $D_y = \mathrm{diag}(y)$, $B = XD_y$, $f_X(y) = D_y\pi^{\mathrm{reg}}$, and $u_X(y) = f_X(y)/\|f_X(y)\|$. This map is sign aligned: $y_\mu(u_X(y))_\mu > 0$. The kernel projector P_X retains exactly the directions annihilated by X. If $s_{\min}(X) \ge \sigma_0\sqrt{m}$, then

$$\|(I-P_X)u_X(y)\|^2 \le \frac{\|Xf_X(y)\|^2}{\sigma_0^2 m\|f_X(y)\|^2} \le \sigma_0^{-2}(\kappa^2 + D),$$

because $\|f_X(y)\|^2 \ge 1/m$. All failures of the coercivity, operator-norm and singular-value events have vanishing joint probability; normalized leakage is always at most one. Thus, if Λ bounds $\limsup \mathbb{E}\kappa^2$, then

$$\limsup \mathbb{E}_X e_X \le \sigma_0^{-2}(\Lambda + D), qquad e_X := \mathbb{E}_y\|(I-P_X)u_X(y)\|^2. \tag{11}$$

Markov's inequality yields good exterior cores with probability at least $1-p_0$ whenever $\sigma_0^{-2}(\Lambda+D) < p_0 e_*$. There is no requirement that every label or every exterior have small leakage.

7 From a population gap to a full-network gap

For a symmetric kernel h on a finite probability space, set

$$(T_h g)(x) = \mathbb{E}_y[h(x,y)g(y)], \qquad \lambda_+(T_h) = \sup_{g \geq 0,\ \mathbb{E}g^2 \leq 1} \langle g, T_h g \rangle.$$

Use the analogous definition for a finite symmetric matrix. The unit-ball convention includes the value zero. The new positive-map section proves that the kernel

$$h_X(x,y) = (u_X(x)^\top y)(u_X(y)^\top x)$$

has $\lambda_+(T_{h_X}) \leq 1 - \gamma_{\mathrm{pop}}$ and fourth moment at most 105 on every good exterior, with $\gamma_{\mathrm{pop}} = 0.032$ allowed.

Here is the precise transfer statement used in the argument. Its finite-feature and uniform sampling inputs are proved in the transfer appendix below.

Theorem 7.1 (Full-center transfer)**.** *For each neuron let W_i be the normalized internal part of its full regularized coefficient vector, with zero diagonal. Suppose that for every fixed number of omitted neurons the common exterior is good with limiting probability at least $1 - p_0$. Then, for every fixed $0 < \gamma_{\mathrm{full}} < \gamma_{\mathrm{pop}}$,*

$$\liminf_n \mathbb{P}\left\{ \sum_{i<j} (t_i W_{ij} - t_j W_{ji})^2 \geq \gamma_{\mathrm{full}} \|t\|^2 \text{ for every } t \geq 0 \right\} \geq 1 - p_0.$$

Proof. Put $L_{ij} = \sqrt{n} W_{ij}$, $C_{ij} = L_{ij} L_{ji}$, $C_{ii} = 0$. The transfer appendix proves three facts. First, finite arrays of L_{ij}, conditioned on their common exterior, converge to $u_X(y_i)^\top y_j$. Second, the empirical fourth moment of C is tight. Third, both population kernels and finite graphs have their positive-cone quadratic maxima approximated uniformly by sufficiently large fixed samples when the fourth moment is bounded.

For clarity, the order of limits is essential. Set $\xi = (\gamma_{\mathrm{pop}} - \gamma_{\mathrm{full}})/4$ and fix a probability tolerance. Choose a finite fourth-moment cutoff, then a finite sample size large enough for both sampling statements with error at most ξ. Hold that sample size fixed while $n \to \infty$. Finite-array convergence costs another ξ. On a good core the induced array has value at most $1 - \gamma_{\mathrm{pop}} + 2\xi$. If the full graph had $\lambda_+(C/n) > 1 - \gamma_{\mathrm{full}}$, graph sampling would make the same induced value greater than $1 - \gamma_{\mathrm{full}} - \xi$, except with the prescribed small error. These bounds are incompatible because $\gamma_{\mathrm{pop}} - \gamma_{\mathrm{full}} > 3\xi$. Finally let the probability tolerance tend to zero; the core failure budget remains p_0.

Every row of W has norm one and $W_{ii} = 0$, so exactly

$$\sum_{i<j} (t_i W_{ij} - t_j W_{ji})^2 = \|t\|^2 - t^\top (C/n) t.$$

The asserted bound follows. Dependence between different rows has never been replaced by independence; independence is used only for the omitted labels conditional on the common exterior. □

8 Sharper Gaussian aperture and internal normalization

This section combines the replacement and heavy-coordinate theorems proved in the technical appendices with sharper Gaussian estimates. All parameters below are fixed before the dimension tends to infinity.

8.1 A Gaussian operator norm estimate with its comparison proof

If G is an $m \times d$ standard Gaussian matrix, then

$$\mathbb{E}\|G\|_{\mathrm{op}} \le \sqrt{m} + \sqrt{d}.$$

Here is a short proof of the comparison needed for this assertion. Index by unit vectors $(x, y) \in \mathbb{S}^{d-1} \times \mathbb{S}^{m-1}$ and compare $X_{xy} = y^\top G x$ with $Y_{xy} = g^\top y + h^\top x$, for independent standard Gaussian g, h. For two indices, writing $a = x^\top x'$ and $b = y^\top y'$,

$$\mathbb{E}(X_{xy} - X_{x'y'})^2 = 2(1 - ab) \le 4 - 2(a + b) = \mathbb{E}(Y_{xy} - Y_{x'y'})^2,$$

since $(1 - a)(1 - b) \ge 0$. To justify the resulting maximum comparison, first use finite index sets and the smooth maximum $F_\beta(z) = \beta^{-1} \log \sum_j e^{\beta z_j}$. Gaussian integration by parts along the interpolation from X to Y gives

$$\frac{d}{dt}\mathbb{E}F_\beta(\sqrt{t}Y + \sqrt{1-t}X) = \frac{\beta}{4}\sum_{j,k}\mathbb{E}(p_j p_k)\{\mathbb{E}(Y_j - Y_k)^2 - \mathbb{E}(X_j - X_k)^2\} \ge 0.$$

Send $\beta \to \infty$ and then take finite nets. Thus $\mathbb{E}\sup X \le \mathbb{E}\sup Y = \mathbb{E}\|g\| + \mathbb{E}\|h\| \le \sqrt{m} + \sqrt{d}$. The Gaussian Lipschitz concentration lemma and the existing lower singular value comparison now imply, whenever $m/d \to 2 - 10^{-29}$,

$$0.28\sqrt{m} \le s_{\min}(G) \le \|G\|_{\mathrm{op}} \le 1.72\sqrt{m} \tag{12}$$

with probability at least $1 - e^{-cm}$ for some fixed $c > 0$ and all sufficiently large m. Indeed the limiting lower and upper comparison constants are $1 - 1/\sqrt{2 - 10^{-29}} > 0.2928$ and $1 + 1/\sqrt{2 - 10^{-29}} < 1.7072$.

8.2 Removing an unnecessary loss in the width estimate

Let $c_0 = 1/\sqrt{\pi}$ and use the previous bad shell

$$E_{\delta,t} = \{x \in \mathbb{S}^{m-1} : \|x_-\|_1/\sqrt{m} \le t, |\mathbf{1}^\top x/\sqrt{m} - c_0| \ge \delta\}.$$

Its Gaussian width satisfies the sharper estimate

$$\frac{w(E_{\delta,t})}{\sqrt{m}} \le \frac{1 - \delta^2/2}{\sqrt{2}} + Lt + \sqrt{\psi(L)} + o(1), \qquad \psi(L) = \mathbb{E}(G_0 - L)_+^2, \tag{13}$$

where G_0 is scalar standard Gaussian. We will also use $\psi(L) \le 2e^{-L^2/2}$: bound $(G_0 - L)_+^2$ by $(G_0^2 - L^2)_+$ and integrate $\int_L^\infty 2s\mathbb{P}(G_0 > s)\, ds \le 2e^{-L^2/2}$. To prove the width estimate, write $v = g_+/\|g_+\|$ and $a_m = |\mathbf{1}^\top v/\sqrt{m} - c_0|$. Every $x \in E_{\delta,t}$ obeys $\|x - v\| \ge (\delta - a_m)_+$. Consequently

$$g_+^\top x \le \|g_+\|\{1 - (\delta - a_m)_+^2/2\}.$$

The law of large numbers gives $a_m \to 0$ in probability and $\|g_+\|/\sqrt{m} \to 1/\sqrt{2}$ in L^1. The multiplier is bounded. The remaining negative-coordinate term is at most $L\|x_-\|_1 + \|(g_- - L)_+\|\|x_-\|$ as before. Taking suprema and expectations proves (13). The event $g_+ = 0$ has probability 2^{-m} and contributes $o(\sqrt{m})$.

8.3 Margin-space diameter to coefficient-space diameter

For nonzero vectors a, b, the following exact identity will be useful twice:

$$\left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\|^2 = \frac{\|a-b\|^2 - (\|a\| - \|b\|)^2}{\|a\|\|b\|}. \tag{14}$$

In particular, for unit vectors x, y and an injective linear map T,

$$\left\| \frac{Tx}{\|Tx\|} - \frac{Ty}{\|Ty\|} \right\| \le \frac{\|T\|_{\mathrm{op}}}{s_{\min}(T)} \|x - y\|.$$

Apply this with $T = G^\dagger$ restricted to the range of G.

Suppose the range of G avoids $E_{\delta,t}$, the event (12) holds, and two unit coefficient vectors w_1, w_2 have mean negative hinge at most h. Put $x_\nu = Gw_\nu/\|Gw_\nu\|$. Then $\|(x_\nu)_-\|_1/\sqrt{m} \le h/0.28$ and $c_0 - \delta \le \mathbf{1}^\top x_\nu/\sqrt{m} \le c_0 + \delta$. For $\delta \le 10^{-5}$, the normalized sum has negative-part ratio at most

$$\frac{2h/0.28}{2(c_0 - \delta)} < 8h.$$

Thus $8h < t$ ensures that the same shell estimate applies to the normalized sum. It follows that

$$\|x_1 + x_2\| \ge \frac{2(c_0 - \delta)}{c_0 + \delta}, \qquad \|x_1 - x_2\|^2 \le \frac{16 c_0 \delta}{(c_0 + \delta)^2} \le \frac{16\delta}{c_0}.$$

Using $1.72/0.28 = 43/7$, $c_0 > 0.56$ and $\sqrt{0.56} > 0.745$ gives

$$\|w_1 - w_2\| \le \frac{43}{7} \frac{4}{\sqrt{c_0}} \sqrt{\delta} < 33\sqrt{\delta}. \tag{15}$$

This replaces the earlier coefficient 1600.

8.4 A smaller loss when deleting the heavy coordinates

Let A be an $m \times d$ Rademacher matrix with $d/m \to 1/(2 - 10^{-29})$. Two 1/4-nets have at most 9^{m+d} pairs, and each fixed bilinear form has tail at most $2e^{-z^2/2}$. The usual net approximation therefore gives

$$\mathbb{P}\{\|A\|_{\mathrm{op}} > 6\sqrt{m}\} \le 2\exp\{(m+d)\log 9 - 4.5m\} \le 2e^{-m}$$

for all sufficiently large m. If deleting coordinates of a unit vector w removes squared mass $q < 1/100$, and $\widetilde{w}$ is the normalized remainder, then exactly

$$\|w - \widetilde{w}\|^2 = 2 - 2\sqrt{1-q} = \frac{2q}{1 + \sqrt{1-q}} < 1.01^2 q.$$

Consequently a pair of feasible vectors loses less than $2.02\sqrt{q_0}$ in separation and creates total mean hinge at most

$$6(2.02)\sqrt{q_0} < 13\sqrt{q_0}. \tag{16}$$

The remainder still has coordinate maximum at most $2m^{-2/5}$. Thus the flat-coefficient replacement theorem applies.

More explicitly, take the deterministic set of pairs of unit vectors with coordinate maximum at most $2m^{-2/5}$ and separation at least $\rho/2$. The Gaussian argument gives its minimum total mean hinge at least h with probability tending to one. Nonnegativity and expectation replacement give a Rademacher mean

at least $(h+g)/2$ eventually, for any fixed $g < h$. The mean-hinge concentration lemma with two vectors therefore gives

$$\mathbb{P}\{F_T(A) \le g\} \le \exp\{-m(h-g)^2/128\}.$$

If two feasible Rademacher directions had separation at least ρ, their normalized remainders would belong to this set as soon as $2.02\sqrt{q_0} < \rho/2$. Their total hinge would be less than $13\sqrt{q_0}$. Thus the strict inequality $13\sqrt{q_0} < g$ is sufficient for a contradiction; an extra factor of two is not needed.

8.5 A sharper final symmetry contradiction

Let z_i be any unit feasible full row, including bias. Let z_i^* be the unit hard max-margin center, and z_i^{reg} the unit regularized full center. The aperture event gives $\|z_i - z_i^*\| \le \rho$, while regularization gives $\|z_i^* - z_i^{\mathrm{reg}}\| \le r_{\mathrm{reg}}$. The largest-coordinate theorem implies that both feasible unit vectors have internal norm at least $\sqrt{1-q_0}$; hence the regularized vector has internal norm at least $\sqrt{1-q_0} - r_{\mathrm{reg}}$. Choose $b_0 \le \sqrt{1-q_0} - r_{\mathrm{reg}}$. Equation (14), applied to their internal parts, proves

$$\|R_i - W_i\| \le \frac{\rho + r_{\mathrm{reg}}}{b_0},$$

where R_i is the feasible unit internal row and W_i the normalized regularized internal row. In the full-center construction the positive cavity factor cancels in this normalization, so this is exactly its W_i.

If symmetric nonzero rows J_i exist, put $t_i = \|J_i\|$ and $R_i = J_i/t_i$. The nonzero condition follows because both labels occur at neuron i and the margins are strict. Symmetry gives $t_i R_{ij} = t_j R_{ji}$. Therefore

$$\sum_{i<j}(t_i W_{ij} - t_j W_{ji})^2 \le 2\sum_i t_i^2 \|W_i - R_i\|^2 \le \frac{2(\rho + r_{\mathrm{reg}})^2}{b_0^2}\|t\|^2.$$

The contradiction with full-center incompatibility now needs only

$$\boxed{2(\rho + r_{\mathrm{reg}})^2/b_0^2 < \gamma_{\mathrm{full}}.} \tag{17}$$

9 An explicit improved scalar certificate

The following choice satisfies all the refined gates:

$$\epsilon = 10^{-29}, \quad u = 2-\epsilon, \quad \rho = 0.12, \quad \delta = 3.2\cdot 10^{-6}, \quad L = 12,$$
$$t = 2.85\cdot 10^{-13}, \quad h = 3.5\cdot 10^{-14}, \quad g = 3.2\cdot 10^{-14},$$
$$\Delta = \epsilon/100, \quad q_0 = 0.51\epsilon, \quad M_{\mathrm{head}} = 13, \quad \sigma_0 = 1/25,$$
$$a_0 = 10^{-34}, \quad a_{\mathrm{mix}} = \eta = \tau = 10^{-232}, \quad e_* = 10^{-12}, \quad p_0 = 10^{-3},$$
$$\gamma_{\mathrm{pop}} = 0.032, \quad \gamma_{\mathrm{full}} = 0.03, \quad b_0 = 0.99.$$

The projector requirements are $\max_i P_{ii} \le 0.500001$ and $\max_{i\ne j}|P_{ij}| \le 10^{-16}$, both supplied by the existing entrywise $o(1)$ projection-regularity theorem.

Gaussian geometry. Here $\delta^2 = 1.024\cdot 10^{-11}$ and, using $1/\sqrt{2} > 0.7$,

$$\frac{\delta^2}{2\sqrt{2}} > 3.584\cdot 10^{-12}, \qquad Lt = 3.42\cdot 10^{-12}, \qquad \sqrt{\psi(12)} \le \sqrt{2}e^{-36} < 4\cdot 10^{-16}.$$

The load correction is below 10^{-28} and the eventual width $o(1)$ can be made below 10^{-14}. The width reserve therefore exceeds 10^{-13} for all sufficiently large dimensions, which is a fixed positive reserve for Gaussian escape. Moreover,

$$8h = 2.8 \cdot 10^{-13} < t, \quad h > g, \quad 33\sqrt{\delta} < 33(0.0018) = 0.0594 < \rho/2 = 0.06.$$

Thus the Gaussian robust pair gap is at least $h > g$.

Heavy coordinates and Bernoulli transfer. Since $\sqrt{5.1} < 2.26$,

$$13\sqrt{q_0} < 2.938 \cdot 10^{-14} < g = 3.2 \cdot 10^{-14}, \qquad 2.02\sqrt{q_0} < \rho/2.$$

The clipping requirement holds:

$$2e^{-13^2/2} < 5 \cdot 10^{-37} < \Delta/(400u), \qquad \gamma_0 = \Delta/(4uM_{\mathrm{head}}) > \epsilon/10400.$$

Since $10400^2 \cdot 1024 < 1.2 \cdot 10^{11}$, $\gamma_0^2/1024 > 8 \cdot 10^{-70}$. The flat-pair concentration exponent is $(h-g)^2/128 > 7 \cdot 10^{-32}$. The expectation replacement and lower-tail argument just proved therefore gives aperture and heavy-coordinate control with failure probability at most $4e^{-10^{-71}m}$ for all sufficiently large m. A union over all n neurons tends to zero.

Positive coercivity and regularization. The existing positive-coercivity coefficient is

$$a_+ = \frac{\epsilon}{2000\sqrt{\log(2000/\epsilon)}} > \epsilon/18000 > 10^{-34} = a_0,$$

because $\log(2 \cdot 10^{32}) < 81$. Consequently $\kappa_{\min} > a_0/\sqrt{2} > 10^{-35}$ and $U = 200/a_0^2 = 2 \cdot 10^{70}$. With $L_0 \le 10$ and the hard center's residual at most L_0, the mixed competitor excess obeys

$$D \le 2a_{\mathrm{mix}}\kappa L_0 + a_{\mathrm{mix}}^2 L_0^2 + \eta U + \tau \log(1/a_{\mathrm{mix}}) < 3 \cdot 10^{-162}.$$

It follows that $r_{\mathrm{reg}} \le 2\sqrt{D}/\kappa_{\min} < 4 \cdot 10^{-46}$.

Mean leakage and population-to-network passage. The one-feature lifting lemma proved below gives $\limsup \mathbb{E}\kappa^2 \le \epsilon/(2-\epsilon) < 0.51\epsilon$. The ordinary coercivity appendix supplies $\sigma_0 = 1/25$ with exponential failure probability. Both inputs are proved here. Thus

$$\limsup \mathbb{E}_X e_X \le 625\{\epsilon/(2-\epsilon) + D\} < 3.2 \cdot 10^{-27}, \qquad \limsup \mathbb{P}(e_X > 10^{-12}) < 3.2 \cdot 10^{-15} < p_0.$$

The positive-map theorem gives population gap 0.032 on these good cores. The full-center transfer proved in this document permits any smaller fixed gap, in particular $\gamma_{\mathrm{full}} = 0.03$.

The final strict inequality. Since $r_{\mathrm{reg}} < 10^{-6}$,

$$\frac{2(\rho + r_{\mathrm{reg}})^2}{b_0^2} < \frac{2(0.120001)^2}{0.99^2} < 0.0294 < 0.03.$$

The deterministic contradiction therefore applies on an event whose limiting probability is at least $1 - p_0$. In particular the same proof framework yields

$$\limsup_{n\to\infty} \mathbb{P}(\text{strict storage at } m = \lfloor (2 - 10^{-29})n \rfloor) \le 10^{-3}, \qquad \boxed{\alpha_- \le 2 - 10^{-29}.}$$

This is the improved conclusion of the research manuscript. It is not a proof of $\alpha_- \le 1.9$. The reduction in the upper bound from $2 - 10^{-89}$ is $10^{-29} - 10^{-89}$; the deficit from two grows by a factor 10^{60}, while remaining macroscopically tiny. The appendix proofs and this certificate use only elementary probability, Gaussian comparison proved by interpolation, and finite-dimensional spectral arguments.

10 A self-contained margin bound by adding one feature

The heavy-coordinate theorem in the appendix also bounds the ordinary hard margin through a simple augmentation. Combined with the 1/25 singular-value estimate, it removes the previously used external universality theorem from the capacity proof.

Lemma 10.1 (One-feature lifting)**.** *Let $B \in \{-1,1\}^{d\times m}$ be iid Rademacher with $m/d \to u \in (1,2)$, and define*

$$\kappa(B) = \min_{\pi\geq 0,\, \mathbf{1}^\top\pi=1} \|B\pi\|\,.$$

Suppose the established heavy-coordinate theorem is applied to iid designs with $d+1$ coefficient coordinates and the same limiting load u. Then

$$\limsup_{m,d\to\infty} \mathbb{E}\kappa(B)^2 \leq \frac{2-u}{u}. \tag{18}$$

More explicitly, for every fixed $q \in (1-u/2, 1)$ and sufficiently large dimensions, there is a constant $c(u,q) > 0$ such that

$$\mathbb{P}\left\{\kappa(B)^2 > \frac{q}{1-q}\right\} \leq e^{-c(u,q)m}.$$

Proof. Append to B an independent Rademacher row $r^\top$, obtaining $\widetilde{B} \in \{-1,1\}^{(d+1)\times m}$. Its aspect ratio satisfies $m/(d+1) \to u$, so the heavy-coordinate theorem applies at the same limiting load. On its good event, every unit vector z with $\widetilde{B}^\top z \geq 0$ has every coordinate of squared magnitude at most q: each coordinate is bounded by the sum over the largest coordinates.

If $\kappa(B) > 0$, let g_* be the closest point to zero in the convex hull of the columns of B and set $w = g_*/\kappa(B)$. Projection geometry gives $B^\top w \geq \kappa(B)\mathbf{1}$. Therefore, for every choice of the appended sign row r,

$$z = \frac{(w, \kappa(B))}{\sqrt{1+\kappa(B)^2}}$$

is a feasible unit vector for $\widetilde{B}$, because

$$B^\top w + \kappa(B) r \geq \kappa(B)(\mathbf{1} + r) \geq 0.$$

Its last squared coordinate is $\kappa(B)^2/(1+\kappa(B)^2)$. Hence the good heavy-coordinate event forces $\kappa(B)^2 \leq q/(1-q)$. If $\kappa(B) = 0$, the same conclusion is immediate. This implication holds for every appended row; averaging over that row therefore gives the asserted probability bound for the original matrix B.

Deterministically $\kappa(B)^2 \leq d$, by choosing any vertex of the simplex. Thus

$$\mathbb{E}\kappa(B)^2 \leq \frac{q}{1-q} + de^{-c(u,q)m}.$$

First take the dimension limit for each fixed $q > 1-u/2$, and then let $q \downarrow 1-u/2$. The bound tends to $(2-u)/u$, proving (18). □

The bias and common cores. For X containing the constant bias row and an independent label y, the matrix XD_y is iid in its joint marginal law. The lemma therefore bounds the exact unconditional mean used by the common-core leakage argument. For every fixed number k of omitted neurons, its coefficient dimension $d = n-k+1$ has $m/d \to u$. The additional independent row in the proof is a mathematical device; it does not alter the network being studied.

Part II
Self-contained probability and transfer proofs

11 Definitions and Gaussian comparison facts

For $A \in \mathbb{R}^{m\times d}$ and $w \in \mathbb{R}^d$, define the mean negative-margin loss

$$\ell_A(w) = \frac{1}{m}\sum_{i=1}^{m}(-(Aw)_i)_+.$$

Thus $\ell_A(w) = 0$ exactly when $Aw \geq 0$ entrywise. All vector norms without a subscript are Euclidean. A Rademacher matrix means a matrix of independent uniform signs.

12 Gaussian comparison and concentration, with proofs

This section supplies the Gaussian facts used later. The comparison is needed only in expectation; no unproved universality or comparison theorem is used. All Euclidean norms in this section are denoted by $\|\cdot\|$.

Lemma 12.1 (One-sided Gaussian comparison)**.** *Let T, V be compact sets, let $b : T \to \mathbb{R}^q$ be continuous, and let $\psi : T \times V \to \mathbb{R}$ be continuous, where $V \subset \mathbb{R}^p$. Let $G \in \mathbb{R}^{p\times q}$, $g \in \mathbb{R}^p$, $h \in \mathbb{R}^q$, and $\xi \in \mathbb{R}$ be mutually independent standard Gaussian objects. Then*

$$\begin{aligned}\mathbb{E}\min_{x\in T}\max_{v\in V}\{v^TGb(x) + \xi\|b(x)\|\|v\| + \psi(x,v)\}\\ \geq \mathbb{E}\min_{x\in T}\max_{v\in V}\{\|b(x)\|g^Tv + \|v\|h^Tb(x) + \psi(x,v)\}.\end{aligned}$$

Consequently the same inequality with the ξ term removed from the left holds with error at most $\mathbb{E}|\xi|\sup_T\|b\|\sup_V\|v\|$. Neither set needs to be convex. The index x may contain coordinates other than $b(x)$.

Proof. First take finite index sets $T = \{x_i\}$ and $V = \{v_j\}$, and write $b_i = b(x_i)$. Let $X_{ij} = v_j^TGb_i + \xi\|b_i\|\|v_j\|$ and $Y_{ij} = \|b_i\|g^Tv_j + \|v_j\|h^Tb_i$. They have the same individual variances. Their covariance difference is

$$\mathrm{Cov}(X_{ij}, X_{i'j'}) - \mathrm{Cov}(Y_{ij}, Y_{i'j'}) = (\|b_i\|\|b_{i'}\| - b_i^Tb_{i'})(\|v_j\|\|v_{j'}\| - v_j^Tv_{j'}) \geq 0.$$

This difference vanishes when $i = i'$. For $\beta > 0$ smooth the minimum and maximum by

$$F_\beta(z) = -\beta^{-1}\log\sum_i \exp\left[-\log\sum_j \exp\{\beta(z_{ij} + \psi(x_i, v_j))\}\right].$$

Writing p_i for the outer soft-min weights and q_{ij} for the inner soft-max weights, its mixed second derivative for $i \neq i'$ is $\beta p_ip_{i'}q_{ij}q_{i'j'} \geq 0$. For independent copies of X, Y, Gaussian integration by parts gives

$$\frac{d}{dt}\mathbb{E}F_\beta(\sqrt{t}X + \sqrt{1-t}Y) = \frac{1}{2}\sum_{ij,i'j'}(\mathrm{Cov}\,X - \mathrm{Cov}\,Y)_{ij,i'j'}\mathbb{E}\partial_{ij}\partial_{i'j'}F_\beta \geq 0.$$

The integration-by-parts identity follows by integrating the derivative of the Gaussian density in each coordinate. The same-minimum-index terms vanish and all other terms have the indicated sign. Let

$\beta \to \infty$; the error is at most $(\log|T| + \log|V|)/\beta$. For compact sets, use finite nets, continuity, and the integrable Gaussian operator-norm bound to pass to the limit. Finally, removing ξ changes the value pointwise by at most $|\xi| \sup_T \|b\| \sup_V \|v\|$. □

Lemma 12.2 (Gaussian Lipschitz concentration)**.** *If Z is a standard Gaussian vector and F is L-Lipschitz in Euclidean distance, then*

$$\mathbb{P}\{F(Z) \le \mathbb{E}F(Z) - t\}, \quad \mathbb{P}\{F(Z) \ge \mathbb{E}F(Z) + t\} \le \exp\{-t^2/(2L^2)\} \qquad (t > 0).$$

Proof. Here is a short derivation, including the needed logarithmic Sobolev inequality. Write γ for standard Gaussian measure and $P_s f(x) = \mathbb{E}f(e^{-s}x + \sqrt{1-e^{-2s}}Z)$. For smooth positive bounded f, differentiation of this formula and Gaussian integration by parts give

$$-\frac{d}{ds}\int (P_s f)\log(P_s f)\, d\gamma = \int \frac{\|\nabla P_s f\|^2}{P_s f}\, d\gamma.$$

Also $\nabla P_s f = e^{-s} P_s \nabla f$. Cauchy–Schwarz applied with weight f yields

$$\frac{\|\nabla P_s f\|^2}{P_s f} \le e^{-2s} P_s\left(\frac{\|\nabla f\|^2}{f}\right).$$

Integrating in s from zero to infinity and using the invariance of γ under P_s proves

$$\mathrm{Ent}_\gamma(f) \le \frac{1}{2}\int \frac{\|\nabla f\|^2}{f}\, d\gamma.$$

Substitute $f = e^{\lambda F}$, first after smooth truncation. Since $\|\nabla F\| \le L$, the logarithmic moment-generating function $H(\lambda)$ obeys $\lambda H'(\lambda) - H(\lambda) \le \lambda^2 L^2/2$. Integration gives $\log \mathbb{E}e^{\lambda(F-\mathbb{E}F)} \le \lambda^2 L^2/2$. Exponential Markov and optimization over λ give the tails. Smooth approximation and truncation extend the result to all Lipschitz F; Gaussian tails supply domination for the exponential moments. □

Lemma 12.3 (Escape and singular values)**.** *For a compact $E \subset \mathbb{S}^{r-1}$ define its Gaussian width $w(E) = \mathbb{E}\sup_{x\in E} g^T x$, with $g \sim N(0, I_r)$. For an iid Gaussian $H \in \mathbb{R}^{k\times r}$,*

$$\mathbb{E}\inf_{x\in E}\|Hx\| \ge \mathbb{E}\|g_k\| - w(E) \ge \sqrt{k} - \sqrt{2} - w(E),$$

and the infimum is below its expectation by t with probability at most $e^{-t^2/2}$. In particular, for r/d sufficiently close to 2 and $G \in \mathbb{R}^{r\times d}$ Gaussian, with probability $1 - e^{-cr}$,

$$\sigma_{\min}(G) \ge 0.1\sqrt{r}, \qquad \|G\|_{\mathrm{op}} \le 10\sqrt{r}.$$

Proof. Use Lemma 12.1 with $x \in E$ and v on the unit sphere of $\mathbb{R}^k$. Both norms equal one, so the added scalar ξ is a common mean-zero offset and disappears in expectation. The auxiliary value is $\|g_k\| + \inf_{x\in E} h^T x$. Furthermore $\mathbb{E}\|g_k\| \ge \sqrt{k} - \sqrt{2}$, since $|\|g_k\| - \sqrt{k}| \le |\|g_k\|^2 - k|/\sqrt{k}$ and $\mathrm{Var}(\|g_k\|^2) = 2k$. The infimum is 1-Lipschitz in the Frobenius norm of H, so the preceding concentration lemma applies. For the smallest singular value take $E = \mathbb{S}^{d-1}$ and use $w(E) \le \sqrt{d}$. For the operator norm use 1/4-nets of both unit spheres, of sizes at most $9^r, 9^d$. Each fixed bilinear form is $N(0,1)$, and the net approximation gives $\|G\|_{\mathrm{op}} \le 2\max|v^T G w|$. Thus $\mathbb{P}(\|G\|_{\mathrm{op}} > 10\sqrt{r}) \le 2e^{-8r}$ for $d \le r$. □

Remark 12.4 (Range and kernel for Gaussian matrices). The column space of an iid Gaussian $r \times d$ matrix has the same law as the kernel of an independent iid Gaussian $(r-d)\times r$ matrix. Indeed both distributions are invariant under every orthogonal change of coordinates in $\mathbb{R}^r$ and have dimension d almost surely. Such an invariant distribution is unique. To see this without assuming uniqueness, let

$R = Z(Z^TZ)^{-1/2}$ for a square iid Gaussian matrix Z. Almost surely Z is invertible. The distribution of R is invariant under both left and right multiplication by any fixed orthogonal matrix, as follows directly from the same invariance of Z and the displayed polar-factor formula. Thus RV has a distribution independent of the fixed d-plane V. If V itself has any orthogonally invariant law and is independent of R, conditioning first on R and then on V proves that its law must equal this one. This argument is used only for Gaussian matrices. The binary-matrix transfer is proved separately by replacement and coordinate truncation.

13 Concentration and partial Gaussian replacement

Lemma 13.1 (A one-sided entropy estimate)**.** *Let Z be a real function of independent uniform signs, and write $x^{(j)}$ for the sign vector with coordinate j flipped. If, pointwise,*

$$\sum_j (Z(x) - Z(x^{(j)}))_+^2 \le v,$$

then

$$\mathbb{P}\{Z \ge \mathbb{E}Z + t\} \le e^{-t^2/(2v)} \qquad (t > 0).$$

Proof. For a function f of one uniform sign, write its values as $a \ge b$. Jensen's inequality gives

$$\mathrm{Ent}(e^f) \le \mathrm{Cov}(e^f, f) = \frac{1}{4}(e^a - e^b)(a-b) \le \frac{1}{4}e^a(a-b)^2 = \frac{1}{2}\mathbb{E}[e^f (f - f^{\mathrm{flip}})_+^2].$$

Entropy tensorization, proved by conditional Jensen and induction on the number of independent coordinates, therefore implies

$$\mathrm{Ent}(e^{sZ}) \le \frac{s^2}{2}\mathbb{E}\left[e^{sZ}\sum_j (Z - Z^{(j)})_+^2\right] \le \frac{s^2 v}{2}\mathbb{E}e^{sZ}.$$

For $\psi(s) = \log \mathbb{E}e^{s(Z-\mathbb{E}Z)}$, this is $s\psi'(s) - \psi(s) \le s^2 v/2$. Integrating the derivative of $\psi(s)/s$ from zero and applying exponential Markov with $s = t/v$ proves the claim. □

Lemma 13.2 (Mean-hinge concentration on arbitrary compact domains)**.** *Let $r \in \{1, 2\}$ and let T be a nonempty compact subset of $(B^d)^r$. Define*

$$F_T(A) = \min_{(w_1,\dots,w_r)\in T} \sum_{\nu=1}^{r} \ell_A(w_\nu).$$

For a Rademacher A,

$$\mathbb{P}\{F_T(A) \le \mathbb{E}F_T(A) - t\} \le \exp\left(-\frac{mt^2}{8r^2}\right).$$

The domain need not be convex.

Proof. At a matrix A, choose a minimizing tuple $w(A)$. A change in entry A_{ij} of magnitude at most 2 increases the value by at most

$$\frac{2}{m}\sum_{\nu=1}^{r} |w_{\nu j}(A)|,$$

because the old minimizing tuple remains an admissible competitor and $x \mapsto (-x)_+$ is 1-Lipschitz. Thus, pointwise,

$$\sum_{i,j}(F_T(A^{(ij)}) - F_T(A))_+^2 \le \frac{4}{m}\sum_j \left(\sum_\nu |w_{\nu j}|\right)^2 \le \frac{4r^2}{m}.$$

Apply Lemma 13.1 to $Z = -F_T$. This choice of sign is essential: the current minimizer controls upward changes of F_T, and hence the lower tail of F_T. □

Lemma 13.3 (Uniform partial replacement for flat coefficients)**.** *Assume m/d stays in a fixed compact subinterval of $(1,\infty)$. Let $r \in \{1,2\}$, let $J \subset \{1,\dots,d\}$ be fixed, and let $T_m \subset (B^d)^r$ be a nonempty deterministic compact set such that*

$$|w_{\nu j}| \le K m^{-2/5} \quad \text{for all } (w_1,\dots,w_r) \in T_m, \quad j \notin J, \quad \nu \le r,$$

where K is fixed. Let A be Rademacher, and obtain $\widetilde{A}$ by retaining A_J and replacing A_{J^c} by independent standard Gaussian entries. Uniformly over these choices of J, T_m,

$$|\mathbb{E}F_{T_m}(A) - \mathbb{E}F_{T_m}(\widetilde{A})| \le C\{m^{-1/40}\log m + m^{-3/20}\} = o(1).$$

The constant depends on r, K and bounds on the aspect ratio.

Proof. Smooth the hinge using

$$\varphi_\eta(x) = \eta\log(1 + e^{-x/\eta}), \qquad \eta = m^{-1/40}.$$

It differs from $(-x)_+$ by a quantity in $[0, \eta\log 2]$, and

$$|\varphi_\eta'| \le 1, \quad |\varphi_\eta''| \le (4\eta)^{-1}, \quad |\varphi_\eta'''| \le (4\eta^2)^{-1}.$$

Choose a Euclidean m^{-3}-net inside T_m. Its logarithmic cardinality is $O(m\log m)$. The mean-hinge objective, and its smoothed version, are Lipschitz in the tuple of coefficients with constant at most $\sqrt{r}\,\|A\|_{\mathrm{op}}/\sqrt{m}$. For matrices with independent sign or Gaussian entries, a 1/4-net argument and the fixed-bilinear-form subgaussian bound give $\mathbb{E}\,\|A\|_{\mathrm{op}}/\sqrt{m} = O(1)$. Thus the expected discretization error is $O(m^{-3})$.

For a net point w, let $H_w(A) = m^{-1}\sum_{i,\nu}\varphi_\eta((Aw_\nu)_i)$. Replace the finite minimum by

$$\Psi(A) = -\frac{1}{\beta}\log\sum_w e^{-\beta H_w(A)}, \qquad \beta = m^{41/40}.$$

Its softening error is at most $O(m\log m)/\beta = O(m^{-1/40}\log m)$. For one entry in a replaced column, put $b = Km^{-2/5}$. The first three entry derivatives of H_w are bounded by constants times b/m, $b^2/(m\eta)$, and $b^3/(m\eta^2)$. Direct differentiation of the soft minimum gives

$$\Psi''' = \langle H'''\rangle - 3\beta\,\mathrm{Cov}(H', H'') + \beta^2\langle (H' - \langle H'\rangle)^3\rangle.$$

Consequently, globally in all matrix entries,

$$|\partial_{ij}^3\Psi| \le Cb^3\left\{\frac{1}{m\eta^2} + \frac{\beta}{m^2\eta} + \frac{\beta^2}{m^3}\right\}.$$

Taylor expansion through degree two compares a sign and a standard Gaussian: their first two moments match and their third absolute moments are bounded. Replacing at most $md = O(m^2)$ entries therefore costs at most

$$Cm^{-1/5}\left\{\eta^{-2} + \frac{\beta}{m\eta} + \frac{\beta^2}{m^2}\right\} = O(m^{-3/20}).$$

Combining the smoothing, net and replacement errors proves the result. Nothing in the derivative bounds depends on the unreplaced entries in A_J, or on the particular compact index set. □

14 Uniform exclusion of heavy coefficient mass

Lemma 14.1 (A clipped-square empirical estimate)**.** *Let $C \in \mathbb{R}^{m\times k}$ be Rademacher and $g \in \mathbb{R}^m$ be independent standard Gaussian, with $k = o(m)$. Fix $\tau \in (0,1)$ and $M > 0$ such that $2e^{-M^2/2} \le \tau/4$. Then, with probability at least $1 - e^{-c_{\tau,M}m}$ for all sufficiently large m, simultaneously for every $a \in \mathbb{R}^k$, $r \in \mathbb{R}$ satisfying $\|a\|^2 + r^2 = 1$,*

$$\frac{1}{m}\sum_i \min\{(-C_i a + r g_i)_+, M\}^2 \in [1/2 - \tau, 1/2 + \tau].$$

Proof. For a fixed unit (a, r), $Z = -C_i a + r g_i$ is symmetric, has variance one, and satisfies $\mathbb{E}e^{tZ} \le e^{t^2/2}$. Therefore $\mathbb{E}Z_+^2 = 1/2$ and

$$0 \le \mathbb{E}Z_+^2 - \mathbb{E}\min(Z_+, M)^2 = \int_M^\infty 2t\,\mathbb{P}\{Z > t\}\,dt \le 2e^{-M^2/2} \le \tau/4.$$

The clipped square is bounded by M^2, so Hoeffding gives deviation probability at most $2e^{-m\tau^2/(8M^4)}$ at tolerance $\tau/4$. The function $x \mapsto \min(x_+, M)^2$ is $2M$-Lipschitz. On $\|[-C, g]\|_{\mathrm{op}} \le 10\sqrt{m}$, its empirical average is $20M$-Lipschitz on the unit sphere in $\mathbb{R}^{k+1}$. This operator-norm event has failure probability at most $2e^{-8m}$ by the elementary sphere-net argument. Use a $\tau/(80M)$-net, of size at most $(1+160M/\tau)^{k+1}$. Its logarithm is $o(m)$, so the union bound and net extension prove the claim. □

Theorem 14.2 (Heavy mass is uniformly small)**.** *Let $A \in \mathbb{R}^{m\times d}$ be Rademacher, with $m/d \to u \in (1,2)$. Put $k = \lceil m^{4/5} \rceil$. Fix $q_0 \in (1 - u/2, 1)$ and set*

$$\Delta = q_0 - (1 - u/2) > 0, \qquad \tau = \frac{\Delta}{100u}.$$

Choose a fixed M with $2e^{-M^2/2} \le \tau/4$, and put $\gamma_0 = \Delta/(4uM)$. For all sufficiently large m, except with probability at most $e^{-m\gamma_0^2/1024}$, every unit vector whose k largest coordinates carry squared mass at least q_0 has

$$\ell_A(w) > \gamma_0/8.$$

In particular, every feasible unit vector $Aw \ge 0$ satisfies

$$\sum_{j:\,|w_j| > m^{-2/5}} w_j^2 < q_0.$$

Proof. For each J of size k, consider the deterministic compact set

$$T_J = \{w \in S^{d-1} : \|w_J\|^2 \ge q_0,\ \|w_{J^c}\|_\infty \le k^{-1/2}\}.$$

Every vector in the theorem belongs to some T_J: take the indices of its k largest coordinates. There are at most $(ed/k)^k = \exp(o(m))$ such sets.

Fix J and apply Lemma 13.3 to $F_J(A) = \min_{w\in T_J} \ell_A(w)$, replacing only columns in J^c. For the hybrid matrix write $w = (a, b)$, $C = A_J$, and $G = \widetilde{A}_{J^c}$. Its loss has the exact min–max representation

$$F_J(\widetilde{A}) = \min_{(a,b)\in T_J} \max_{p\in[0,1]^m} -\frac{1}{m} p^T(Ca + Gb).$$

Conditional on C, Lemma 12.1 bounds its expectation below by that of the Gaussian auxiliary problem, with error at most $\mathbb{E}|\xi|/\sqrt{m}$. Indeed $\|b\| \le 1$, $\|p\| \le \sqrt{m}$, and the objective is divided by m. Writing $r = \|b\|$ and using $h^T b \ge -r\,\|h\|$, the auxiliary value is bounded below by

$$\min_{\substack{\|a\|^2 + r^2 = 1 \\ \|a\|^2 \ge q_0,\ r \ge 0}} \max_{p\in[0,1]^m} \frac{1}{m}\{z^T p - r\,\|h\|\,\|p\|\}, \qquad z = -Ca + rg,$$

where $h \in \mathbb{R}^{d-k}$ is standard Gaussian. Removing the tail-coordinate restriction enlarged the minimum domain, and hence gave a valid lower bound.

Let $z_M = \min(z_+, M)$ entrywise, and choose $p = z_M/M$. Since $z^T z_M \geq \|z_M\|^2$, the displayed objective is at least

$$\frac{1}{M}\left\{a_2 - r\frac{\|h\|}{\sqrt{m}}\sqrt{a_2}\right\}, \qquad a_2 = \frac{1}{m}\|z_M\|^2.$$

By Lemma 14.1, uniformly over all (a, r) in question, $a_2 \in [1/2 - \tau, 1/2 + \tau]$, except with exponentially small probability. Also $\|h\|/\sqrt{m} \leq u^{-1/2} + \tau$ except with exponentially small probability for all sufficiently large m. Therefore the auxiliary value is at least

$$\frac{1}{M}\left\{\frac{1}{2} - \tau - \sqrt{1-q_0}(u^{-1/2} + \tau)\sqrt{\frac{1}{2} + \tau}\right\}.$$

At $\tau = 0$, the expression in braces equals

$$\frac{1}{2}\left(1 - \sqrt{1 - 2\Delta/u}\right) \geq \frac{\Delta}{2u}.$$

Changing τ from zero changes this lower bound by at most 3τ: use $\sqrt{1/2+\tau} - \sqrt{1/2} \leq \tau/\sqrt{2}$, $\sqrt{1-q_0} \leq 1$, and $\tau < 1/100$. Thus the auxiliary value is at least $\gamma_0 = \Delta/(4uM)$.

The auxiliary objective is nonnegative because $p = 0$ is allowed. Its high-probability lower bound therefore gives expectation at least $\gamma_0 - o(1)$. The expectation comparison, whose additional scalar error is $O(m^{-1/2})$, shows that the hybrid primary expectation is at least $\gamma_0/3$ for all sufficiently large m. Partial replacement then gives $\mathbb{E}F_J(A) \geq \gamma_0/4$. All constants and thresholds are uniform in J. Lemma 13.2, with $r = 1$, now yields

$$\mathbb{P}\{F_J(A) \leq \gamma_0/8\} \leq e^{-m\gamma_0^2/512}.$$

The union over $\binom{d}{k} = \exp(o(m))$ supports gives the stated probability. Finally, at most $m^{4/5}$ coordinates of a unit vector can exceed $m^{-2/5}$ in absolute value, so their mass is bounded by that of the largest k coordinates. □

15 Positive-orthant coercivity

Fix $1 < u < 2$ and consider dimensions satisfying $d \leq m \leq ud$. Write

$$\varepsilon = 2 - u, \qquad M = 4\sqrt{\log(1000/\varepsilon)}, \qquad a_u = \frac{\varepsilon}{250M} = \frac{\varepsilon}{1000\sqrt{\log(1000/\varepsilon)}}.$$

Theorem 15.1. *There is $d_0(u)$ such that, for $d \geq d_0(u)$ and an iid Rademacher matrix $B \in \{-1,1\}^{d\times m}$,*

$$\mathbb{P}\left\{\inf_{\lambda\geq 0,\,\|\lambda\|_2=1} \frac{\|B\lambda\|_1}{d} < a_u\right\} \leq e^{-c_u d}, \qquad c_u = \frac{\varepsilon^2}{10000M^2}. \tag{19}$$

Consequently, on the same event of high probability,

$$\|B\lambda\|_2 \geq a_u\sqrt{d}\,\|\lambda\|_2 \quad \textit{for every } \lambda \geq 0. \tag{20}$$

16 Head–tail dissection and a replacement lemma

Let $k = \lceil m^{4/5} \rceil$. For a set $J \subseteq \{1, \dots, m\}$ with $|J| = k$, define the compact set

$$T_J = \{\lambda \geq 0 : \|\lambda\|_2 = 1, \|\lambda_{J^c}\|_\infty \leq k^{-1/2}\},$$

and the functional

$$F_J(B) = \min_{\lambda \in T_J} \frac{1}{d} \|B\lambda\|_1 .$$

Every nonnegative unit vector belongs to at least one T_J: choose J to contain its k largest coordinates. Also

$$\log \binom{m}{k} \leq k \log(em/k) = o(d). \tag{21}$$

Let $\widetilde{B}_J$ have Rademacher columns on J and independent standard Gaussian columns on J^c, with all entries independent.

Lemma 16.1 (Uniform expectation replacement)**.** *Uniformly over $|J| = k$,*

$$|\mathbb{E}F_J(B) - \mathbb{E}F_J(\widetilde{B}_J)| \leq C \left(m^{-1/40} \log m + m^{-3/20} \right) = o(1), \tag{22}$$

where C is numerical when $d \leq m \leq 2d$.

Proof. Smooth the absolute value using

$$\ell_h(t) = h \log(e^{t/h} + e^{-t/h}), \qquad h = m^{-1/40}.$$

Then $0 \leq \ell_h(t) - |t| \leq h \log 2$ and

$$|\ell_h'| \leq 1, \qquad |\ell_h''| \leq h^{-1}, \qquad |\ell_h'''| \leq 2h^{-2}.$$

Choose a deterministic m^{-3}-net inside T_J, whose logarithmic cardinality is $O(m \log m)$. The empirical loss $d^{-1} \sum_i \ell_h((B\lambda)_i)$ is Lipschitz in λ with constant $\|B\|_{\mathrm{op}} / \sqrt{d}$. For every mixture of the two matrix distributions, this operator norm divided by $\sqrt{d}$ has bounded expectation. A sphere-net argument proves the assertion: fixed bilinear forms have subgaussian moment-generating functions, and the two unit-sphere nets have total logarithmic cardinality $O(d + m)$. Thus the expected discretization error is $O(m^{-3})$.

Replace the minimum over the finite net by a soft minimum with inverse temperature $\beta = m^{41/40}$. The softening error is $O(m^{-1/40} \log m)$. Denote the resulting smooth matrix functional by Φ_J.

Consider an entry B_{ij} in a tail column $j \in J^c$. For every net vector, $|\lambda_j| \leq k^{-1/2}$. Its empirical smoothed loss L_λ satisfies

$$|\partial_{ij} L_\lambda| \leq k^{-1/2}/d, \qquad |\partial_{ij}^2 L_\lambda| \leq k^{-1}/(dh), \qquad |\partial_{ij}^3 L_\lambda| \leq 2k^{-3/2}/(dh^2).$$

For a soft minimum, the third derivative is

$$\partial^3 \Phi_J = \mathbb{E}_{\mathrm{Gibbs}} \partial^3 L - 3\beta \operatorname{Cov}_{\mathrm{Gibbs}}(\partial^2 L, \partial L) + \beta^2 \mathbb{E}_{\mathrm{Gibbs}}[(\partial L - \mathbb{E}_{\mathrm{Gibbs}} \partial L)^3].$$

It follows, uniformly over all values of all matrix entries, that

$$|\partial_{ij}^3 \Phi_J| \leq C k^{-3/2} \left(\frac{1}{dh^2} + \frac{\beta}{d^2 h} + \frac{\beta^2}{d^3} \right). \tag{23}$$

Rademacher and Gaussian variables have matching means and variances and bounded third absolute moments. Taylor expansion through degree two therefore bounds the expectation error from replacing all $d(m-k)$ tail entries by

$$\begin{aligned} Cdmk^{-3/2}\left(\frac{1}{dh^2}+\frac{\beta}{d^2h}+\frac{\beta^2}{d^3}\right) &\le Cmk^{-3/2}\left(h^{-2}+\frac{\beta}{dh}+\frac{\beta^2}{d^2}\right) \\ &\le Cm^{-3/20}. \end{aligned}$$

The head columns are never replaced. Adding the absolute-value smoothing, net and soft-minimum errors proves (22). □

17 The Gaussian comparison and its explicit auxiliary lower bound

Lemma 17.1 (Gaussian expectation comparison)**.** *Let G be an iid Gaussian matrix, let g, h be independent standard Gaussian vectors of the corresponding dimensions, and let ξ be an independent standard Gaussian scalar. Let x and v range over compact sets, let $b(x)$ be continuous, and let $\psi(x,v)$ be continuous. Then*

$$\begin{aligned} &\mathbb{E}\min_x \max_v \{v^T G b(x) + \psi(x,v)\} \\ &\quad \ge \mathbb{E}\min_x \max_v \{\|b(x)\|\, g^T v + \|v\|\, h^T b(x) + \psi(x,v)\} - \mathbb{E}|\xi| \sup_{x,v} \|b(x)\|\,\|v\|\,. \end{aligned}$$

Neither index set needs to be convex.

Proof. First take finite index sets, writing i for a minimizing index and v for a maximizing index. Compare the Gaussian processes

$$X_{i,v} = v^T G b_i + \xi\,\|b_i\|\,\|v\|\,, \qquad Y_{i,v} = \|b_i\|\, g^T v + \|v\|\, h^T b_i.$$

Their covariance difference is

$$\mathrm{Cov}(X_{i,v}, X_{j,v'}) - \mathrm{Cov}(Y_{i,v}, Y_{j,v'}) = (\|b_i\|\,\|b_j\| - b_i^T b_j)(\|v\|\,\|v'\| - v^T v') \ge 0.$$

It vanishes whenever $i = j$. Add the deterministic offsets $\psi(i,v)$, and soften the outer minimum and inner maximum with inverse temperature β. If π_i are the outer soft-minimum weights and $q_{v|i}$ the inner soft-maximum weights, then for $i \ne j$ the mixed second derivative of this nested soft extremum is

$$\partial_{i,v}\partial_{j,v'}\Phi = \beta \pi_i \pi_j q_{v|i} q_{v'|j} \ge 0.$$

For the Gaussian interpolation $\sqrt{t}X + \sqrt{1-t}Y$, integration by parts gives

$$\frac{d}{dt}\mathbb{E}\Phi(\sqrt{t}X + \sqrt{1-t}Y + \psi) = \frac{1}{2}\sum_{i,v,j,v'} (\mathrm{Cov}\, X - \mathrm{Cov}\, Y)_{i,v;j,v'}\, \mathbb{E}[\partial_{i,v}\partial_{j,v'}\Phi] \ge 0.$$

The covariance identity from Gaussian integration by parts follows directly by integrating the Gaussian density; a degenerate covariance can be treated by adding an arbitrarily small independent Gaussian vector and taking a limit. Terms with the same minimizing index have zero covariance difference, and all other terms are nonnegative. Let $\beta \to \infty$ to get the expected min–max comparison for the augmented process X. Removing its scalar term changes a min–max value by at most $|\xi| \sup \|b_i\|\,\|v\|$, proving the finite-set assertion.

For compact sets, take finer finite nets. Continuity makes the finite min–max values converge pointwise to the compact-set values. They are dominated by an integrable multiple of $\|G\|_{\mathrm{op}} + \|g\| + \|h\| + |\xi| + 1$, because the index sets and the deterministic offsets are bounded. Dominated convergence completes the proof. □

Write $\lambda = (a, b)$ according to J, J^c, with $a, b \geq 0$, and $r = \|b\|$. Conditional on the Rademacher head B_J, the hybrid primary problem is

$$F_J(\widetilde{B}_J) = \min_{(a,b)\in T_J} \max_{v\in[-1,1]^d} \frac{1}{d}\{v^T G b + v^T B_J a\}.$$

Its Gaussian auxiliary is

$$\mathcal{A}_J = \min_{(a,b)\in T_J} \max_{v\in[-1,1]^d} \frac{1}{d}\{r g^T v + \|v\|\ h^T b + v^T B_J a\}, \tag{24}$$

where $g \in \mathbb{R}^d$, $h \in \mathbb{R}^{m-k}$ are independent standard Gaussian vectors, also independent of B_J.

Lemma 17.2 (Uniform auxiliary bound)**.** *For the fixed $u < 2$ above, there is $c'(u) > 0$ such that*

$$\mathbb{P}\{\mathcal{A}_J < \varepsilon/(5M)\} \leq e^{-c'(u)d} \tag{25}$$

for all sufficiently large dimensions. The bound is uniform over J.

Proof. Set $\sigma = \varepsilon/100$. For fixed $a \geq 0, r \geq 0$ with $\|a\|^2 + r^2 = 1$, put

$$z = B_J a + rg, \qquad \phi_M(t) = t\,\mathrm{clip}(t, -M, M), \qquad \psi_M(t) = \mathrm{clip}(t, -M, M)^2.$$

The coordinates of z are independent, symmetric, have variance one, and satisfy $\mathbb{E}e^{sz_i} \leq e^{s^2/2}$. The two functions are globally $2M$-Lipschitz. Also $\psi_M(t) \leq t^2$ and

$$0 \leq \mathbb{E}z_i^2 - \mathbb{E}\phi_M(z_i) \leq \mathbb{E}[z_i^2 \mathbf{1}_{\{|z_i|>M\}}] \leq 2(M^2+2)e^{-M^2/2} \leq \sigma/4.$$

The last inequality follows from the stated choice $M = 4\sqrt{\log(1000/\varepsilon)}$. Therefore, for every admissible (a, r),

$$\mathbb{E}\phi_M(z_i) \geq 1 - \sigma/4, \qquad \mathbb{E}\psi_M(z_i) \leq 1.$$

We detail why the empirical bounds can hold simultaneously for all these head parameters. Uniform subgaussian tails imply moment bounds for $\phi_M(z_i)$ and $\psi_M(z_i)$, and hence a Bernstein bound for their empirical averages. For example, $|\phi_M(z_i)| \leq M|z_i|$ and $|\psi_M(z_i)| \leq z_i^2$; integrating the subgaussian tail gives centered moments bounded by $r!C_M^r$ for all integers $r \geq 2$. Expanding the moment-generating function yields a deviation probability at most $2e^{-c(M,\sigma)d}$ at the fixed tolerance $\sigma/4$.

The parameter set $\{(a, r) : a \geq 0, r \geq 0, \|a\|^2 + r^2 = 1\}$ has dimension $k + 1 = o(d)$. Choose a fixed-resolution net, of radius $\sigma/(100M)$, inside this set. Its size is at most $(CM/\sigma)^{k+1} = e^{o(d)}$. The matrix $Z = [B_J \;\; g]$ has operator norm at most $10\sqrt{d}$ except on an event of probability e^{-cd}, by the elementary subgaussian sphere-net argument. Consequently

$$\left|\frac{1}{d}\sum_i \phi_M((Z\theta)_i) - \frac{1}{d}\sum_i \phi_M((Z\theta')_i)\right| \leq 20M\,\|\theta - \theta'\|,$$

and the same estimate holds for ψ_M. Applying the fixed-parameter concentration on the net and extending by this Lipschitz bound proves, simultaneously for every admissible (a, r),

$$\frac{1}{d}\sum_i \phi_M(z_i) \geq 1 - \sigma, \qquad \frac{1}{d}\sum_i \psi_M(z_i) \leq 1 + \sigma \tag{26}$$

with probability at least $1 - e^{-c(u)d}$.

Furthermore,

$$\|h_-\|/\sqrt{d} \leq \sqrt{u/2} + \sigma \tag{27}$$

with probability at least $1 - e^{-c(u)d}$. One can obtain this directly from the moment-generating function of the independent variables $(h_j)_-^2$, whose mean is $1/2$ and whose moment-generating function is finite for parameters below $1/2$; here $(m-k)/d \le u$.

Since $b \ge 0$, $h^T b \ge -r\,\|h_-\|$. For each (a, b) in (24), choose the admissible dual vector

$$v_i = \mathrm{clip}(z_i, -M, M)/M.$$

On (26)–(27), the auxiliary payoff is at least

$$\frac{1}{M}\left[\frac{1}{d}\sum_i \phi_M(z_i) - r\frac{\|h_-\|}{\sqrt{d}}\sqrt{\frac{1}{d}\sum_i \psi_M(z_i)}\right] \ge \frac{1}{M}[1 - \sigma - (\sqrt{u/2} + \sigma)\sqrt{1+\sigma}].$$

Because $\sqrt{u/2} = \sqrt{1 - \varepsilon/2} \le 1 - \varepsilon/4$ and $\sqrt{1+\sigma} \le 1 + \sigma/2$, the bracket is at least

$$\varepsilon/4 - \tfrac{5}{2}\sigma - \tfrac{1}{2}\sigma^2 \ge \varepsilon/5.$$

This bound is uniform over the minimizing domain, proving the claim after adjusting the exponential constant to absorb the finite union of bad events. □

The auxiliary is nonnegative even outside its good event, since its maximizing cube contains $v = 0$. Apply Lemma 17.1 conditional on the head, with the normalization $1/d$. Here $\|b\| \le 1$ and $\|v\| \le \sqrt{d}$, so removing the scalar augmentation costs at most $\mathbb{E}|\xi|/\sqrt{d}$. Lemma 17.2 therefore yields, for all sufficiently large dimensions,

$$\mathbb{E}F_J(\widetilde{B}_J) \ge \frac{\varepsilon}{5M}(1 - e^{-c'(u)d}) - \frac{\mathbb{E}|\xi|}{\sqrt{d}} \ge \frac{\varepsilon}{6M}.$$

Lemma 16.1 then implies, uniformly in J, that for all sufficiently large dimensions,

$$\mathbb{E}F_J(B) \ge \varepsilon/(20M). \tag{28}$$

18 One-sided concentration and the support union

Lemma 18.1 (Concentration of the minimum)**.** *For every fixed J and $t > 0$,*

$$\mathbb{P}\{F_J(B) \le \mathbb{E}F_J(B) - t\} \le e^{-dt^2/8}. \tag{29}$$

Proof. At a matrix B, choose a minimizing vector $\lambda(B) \in T_J$. If one sign entry B_{ij} is flipped, the same vector is a competitor for the new matrix. Since the absolute value is 1-Lipschitz,

$$F_J(B^{(ij)}) - F_J(B) \le 2|\lambda_j(B)|/d.$$

It follows pointwise that

$$\sum_{i,j}(F_J(B^{(ij)}) - F_J(B))_+^2 \le \frac{4}{d}\,\|\lambda(B)\|^2 = \frac{4}{d}.$$

For completeness, if a function Z of independent uniform signs obeys $\sum_s (Z - Z^{(s)})_+^2 \le v$, the two-point entropy inequality and tensorization give

$$\mathrm{Ent}(e^{tZ}) \le \tfrac{1}{2}t^2 v\mathbb{E}e^{tZ}.$$

Indeed, for two values $a \ge b$ of tZ, Jensen gives

$$\mathrm{Ent}(e^{tZ}) \le \tfrac{1}{4}(e^a - e^b)(a-b) \le \tfrac{1}{4}e^a(a-b)^2.$$

The usual integration of the logarithmic moment-generating function yields $\mathbb{P}\{Z \ge \mathbb{E}Z + s\} \le e^{-s^2/(2v)}$. Apply this to $Z = -F_J$ and $v = 4/d$. □

Proof of Theorem 15.1. By (28), $\mathbb{E}F_J(B) - a_u \geq \varepsilon/(25M)$ for all sufficiently large dimensions. Lemma 18.1 therefore gives

$$\mathbb{P}\{F_J(B) < a_u\} \leq \exp\left(-\frac{d\varepsilon^2}{5000M^2}\right).$$

Union over all J, using (21). For large enough dimensions the support-count exponent is at most $d\varepsilon^2/(10000M^2)$, leaving the bound (19). Every nonnegative unit vector belongs to a slice, so the resulting inequality holds on the entire positive unit sphere. Homogeneity and $\|v\|_2 \geq \|v\|_1/\sqrt{d}$ give (20). □

19 An elementary ordinary singular-value bound of $1/25$

The previous elementary net bound e^{-35} can be replaced by $1/25$. The argument retains the large coordinates as Rademacher variables, replaces only the small coordinates by Gaussians, and uses a clipped quadratic witness in the auxiliary problem. It applies to all coefficient vectors, with no nonnegativity assumption and with no deletion of their large coordinates.

Theorem 19.1 (Ordinary Rademacher coercivity)**.** *There is an absolute m_0 such that, whenever $m \geq m_0$ and $0.49m \leq d \leq 0.51m$, an iid sign matrix $A \in \{-1,1\}^{m\times d}$ satisfies*

$$\mathbb{P}\left\{\inf_{\|w\|=1} \frac{1}{m}\|Aw\|_1 < \frac{1}{25}\right\} \leq e^{-m/40000}.$$

In particular, outside this event,

$$s_{\min}(A) \geq \frac{1}{25}\sqrt{m}.$$

The same statement applies to a transposed exterior design with one constant bias feature, after the exact independent-column-sign transformation.

19.1 Dissection and expectation replacement

Set $k = \lceil d^{4/5}\rceil$. For each $J \subseteq [d]$ with $|J| = k$, let

$$T_J = \{w \in \mathbb{R}^d : \|w\| = 1,\ \|w_{J^c}\|_\infty \leq k^{-1/2}\}, \qquad F_J(A) = \min_{w\in T_J} \frac{1}{m}\|Aw\|_1.$$

Every unit vector belongs to one of these compact sets: take J to contain its k largest coordinates. Moreover

$$\log\binom{d}{k} \leq k\log(ed/k) = o(m)$$

uniformly over the stated aspect ratios. Let $\widetilde{A}_J$ agree with A on J and have independent standard Gaussian entries on J^c.

Lemma 19.2 (Small-coordinate replacement)**.** *Uniformly in J and in $0.49m \leq d \leq 0.51m$,*

$$|\mathbb{E}F_J(A) - \mathbb{E}F_J(\widetilde{A}_J)| \leq C\{m^{-1/40}\log m + m^{-3/20}\} = o(1).$$

Proof. Replace absolute value by $\ell_a(x) = a\log(e^{x/a} + e^{-x/a})$, where $a = m^{-1/40}$. Then $0 \leq \ell_a(x) - |x| \leq a\log 2$ and

$$|\ell_a'| \leq 1, \qquad |\ell_a''| \leq a^{-1}, \qquad |\ell_a'''| \leq 2a^{-2}.$$

Take a deterministic m^{-3}-net inside T_J, with logarithmic cardinality $O(d\log m)$. The empirical loss is Lipschitz in w with constant $\|A\|_{\mathrm{op}}/\sqrt{m}$. For every entrywise mixture of signs and standard Gaussians,

$\mathbb{E}\,\|A\|_{\mathrm{op}} \le C\sqrt{m}$: two 1/4-nets give this by the subgaussian moment-generating-function bound for each fixed bilinear form, followed by integration of its tail. Thus the expected discretization cost is $O(m^{-3})$.

Replace the minimum over this finite net by the soft minimum with inverse temperature $\beta = m^{41/40}$. Its error is $O(d \log m/\beta) = O(m^{-1/40} \log m)$. Write Φ_J for the smoothed matrix functional. For an entry (i,j) with $j \notin J$, derivatives of an individual net loss L_w satisfy

$$|\partial_{ij} L_w| \le \frac{k^{-1/2}}{m}, \quad |\partial_{ij}^2 L_w| \le \frac{k^{-1}}{ma}, \quad |\partial_{ij}^3 L_w| \le \frac{2k^{-3/2}}{ma^2}.$$

Differentiating the soft minimum three times expresses its third derivative as the Gibbs average of the third loss derivative, minus 3β times a covariance of the first two derivatives, plus β^2 times a centered third moment of the first derivative. Consequently, uniformly in all matrix entries,

$$|\partial_{ij}^3 \Phi_J| \le Ck^{-3/2}\left(\frac{1}{ma^2} + \frac{\beta}{m^2 a} + \frac{\beta^2}{m^3}\right).$$

Signs and standard Gaussians have the same first two moments and uniformly bounded third absolute moments. One-at-a-time Taylor replacement therefore costs, over at most md tail entries,

$$Cdk^{-3/2}\left(a^{-2} + \frac{\beta}{ma} + \frac{\beta^2}{m^2}\right) \le Cm^{-3/20}.$$

The head entries are never replaced. Adding the smoothing, net, and soft minimum costs proves the lemma. Convexity of T_J was not used. □

19.2 The Gaussian comparison and its auxiliary problem

We record the Gaussian comparison needed here. If G is iid Gaussian, g, h are independent standard Gaussian vectors, (a,b) ranges over a compact set, v ranges over a compact set, and $\psi(a,b,v)$ is deterministic and continuous, then

$$\begin{aligned}&\mathbb{E}\min_{(a,b)}\max_{v}\{v^\top G b + \psi(a,b,v)\}\\ &\quad\ge \mathbb{E}\min_{(a,b)}\max_{v}\{\|b\|\, g^\top v + \|v\|\, h^\top b + \psi(a,b,v)\} - \mathbb{E}|\xi| \sup_{b,v}\|b\|\,\|v\|,\end{aligned}$$

where ξ is an independent standard Gaussian scalar. Here no convexity is required. For completeness, augment the primary process by $\xi\,\|b\|\,\|v\|$. For two indices its covariance minus the auxiliary covariance equals

$$(\|b\|\,\|b'\| - b^\top b')(\|v\|\,\|v'\| - v^\top v') \ge 0;$$

this difference is zero when the minimizing index (a,b) is the same. For finite index sets, smooth the inner maximum and outer minimum. The mixed second derivative of this nested functional at two distinct minimizing indices is $\beta p_i p_j q_{v|i} q_{v'|j} \ge 0$. Gaussian interpolation and integration by parts therefore give the comparison with the scalar term. Take the smoothing parameter to infinity, then finer finite nets; all terms have integrable Gaussian norm bounds. Removing the scalar term costs at most the quantity displayed above. This proves the comparison directly.

Condition on the Rademacher head $C = A_J$, and write $w = (a,b)$, $r = \|b\|$. Since the dual unit ball of ℓ_1 is $[-1,1]^m$, the hybrid minimum has the auxiliary quantity

$$\mathcal{A}_J = \min_{(a,b)\in T_J}\max_{v\in[-1,1]^m}\frac{1}{m}\{v^\top(Ca + rg) + \|v\|\, h^\top b\},$$

where $g \in \mathbb{R}^m$, $h \in \mathbb{R}^{d-k}$ are standard Gaussian and independent of C. The comparison's scalar-removal cost is at most $\mathbb{E}|\xi|/\sqrt{m}$.

Lemma 19.3 (Uniform auxiliary lower bound)**.** *For all sufficiently large m, uniformly over J and the stated aspect ratios,*

$$\mathbb{P}\{\mathcal{A}_J > 0.0666\} \geq 1 - e^{-cm}$$

for an absolute $c > 0$.

Proof. For every (a, r) with $r \geq 0$ and $\|a\|^2 + r^2 = 1$, set $z = Ca + rg$ and

$$\psi(z_i) = \min\{z_i^2, 16\}, \qquad U(a, r) = \frac{1}{m} \sum_i \psi(z_i).$$

Each z_i is centered, symmetric, has variance one, and has moment-generating function bounded by $e^{t^2/2}$. The tail inequality $\mathbb{P}(|z_i| > s) \leq 2e^{-s^2/2}$ implies

$$0 \leq 1 - \mathbb{E}\psi(z_i) = \mathbb{E}(z_i^2 - 16)_+ \leq \int_{16}^{\infty} 2e^{-t/2}\, dt = 4e^{-8} < 0.0014.$$

Thus $0.9986 < \mathbb{E}U(a, r) \leq 1$ uniformly in the parameters.

Here is a complete uniformity argument. The variable $\psi(z_i)$ belongs to $[0, 16]$. For a variable V in this interval, put $H(t) = \log \mathbb{E} \exp\{t(V - \mathbb{E}V)\}$. Then $H''(t)$ is a variance under the exponentially tilted law and is at most $16^2/4$: centering at the midpoint 8 bounds this variance by $\mathbb{E}(V - 8)^2 \leq 64$. Since $H(0) = H'(0) = 0$, integration twice gives $H(t) \leq t^2 16^2/8$. Independence and exponential Markov, optimized in t, give for an average of m such variables a two-sided deviation bound at tolerance 0.005 of $2\exp(-2m(0.005)^2/16^2) = 2\exp(-m/5{,}120{,}000)$. Choose a 10^{-5}-net inside the parameter half-sphere $\{(a, r) : r \geq 0, \|a\|^2 + r^2 = 1\}$. It has at most $(1 + 2 \cdot 10^5)^{k+1} = e^{o(m)}$ points.

The matrix $Z = [C\ g]$ has $\|Z\|_{\mathrm{op}} \leq 10\sqrt{m}$ except with probability $e^{-c'm}$, by the same two-net subgaussian argument; eventually $k + 1 \leq m$. Since $x \mapsto \min\{x^2, 16\}$ is 8-Lipschitz,

$$|U(\theta) - U(\theta')| \leq \frac{8}{\sqrt{m}} \|Z(\theta - \theta')\| \leq 80 \|\theta - \theta'\|.$$

The net extension therefore costs at most 0.0008. A union bound proves, simultaneously for all admissible (a, r),

$$0.99 \leq U(a, r) \leq 1.01$$

outside a probability $e^{-c''m}$ for sufficiently large m.

Also $\|h\| / \sqrt{m} \leq 0.72$ with exponentially high probability, uniformly in $d - k \leq 0.51m$: this follows from $\mathbb{E}e^{t\|h\|^2} = (1 - 2t)^{-(d-k)/2}$ for $0 < t < 1/2$, since $0.72^2 = 0.5184 > 0.51$.

For each (a, b) choose $v_i = \mathrm{clip}(z_i, -4, 4)/4$. Since $h^\top b \geq -r\|h\|$ and $z_i \,\mathrm{clip}(z_i, -4, 4) \geq \psi(z_i)$, the auxiliary payoff is at least

$$\frac{1}{4}\left[U(a, r) - r\frac{\|h\|}{\sqrt{m}}\sqrt{U(a, r)}\right] \geq \frac{0.99 - 0.72\sqrt{1.01}}{4} > 0.0666.$$

The last inequality uses $\sqrt{1.01} < 1.005$. The bound holds simultaneously over the minimizing domain. Combining the exponential failure probabilities proves the lemma. □

The auxiliary value is nonnegative everywhere because $v = 0$ is admissible. Taking expectations, applying the conditional Gaussian comparison, and then Lemma 19.2, gives for all sufficiently large dimensions

$$\mathbb{E}F_J(A) \geq 0.06$$

uniformly in J and the aspect ratio.

19.3 Lower-tail concentration and the union of slices

At a sign matrix A, choose a minimizing vector $w(A) \in T_J$. If one entry A_{ij} is flipped, the same vector is a competitor for the new matrix. Since absolute value is 1-Lipschitz,

$$F_J(A^{(ij)}) - F_J(A) \le \frac{2|w_j(A)|}{m}.$$

Hence, pointwise,

$$\sum_{i,j} \big(F_J(A^{(ij)}) - F_J(A)\big)_+^2 \le \frac{4}{m}.$$

We include the concentration implication. If a function W of independent uniform signs satisfies $\sum_s (W - W^{(s)})_+^2 \le v$, two-point entropy and tensorization give, for $t > 0$,

$$\mathrm{Ent}(e^{tW}) \le \frac{1}{2} t^2 v \mathbb{E} e^{tW}.$$

Indeed, for two values $a \ge b$ of tW, Jensen's inequality bounds the two-point entropy by $\frac{1}{4}(e^a - e^b)(a - b) \le \frac{1}{4} e^a (a-b)^2$; averaging over the coordinate and tensorizing yields the displayed inequality. If $H(t) = \log \mathbb{E} e^{tW}$, it implies $tH'(t) - H(t) \le t^2 v/2$. Integrating the derivative of $H(t)/t$ and applying exponential Markov gives $\mathbb{P}(W \ge \mathbb{E}W + s) \le e^{-s^2/(2v)}$.

Apply this to $W = -F_J$ and $v = 4/m$. Since $\mathbb{E}F_J(A) - 1/25 \ge 0.02$,

$$\mathbb{P}\{F_J(A) < 1/25\} \le \exp\{-m(0.02)^2/8\} = e^{-m/20000}.$$

For sufficiently large m, $\log \binom{d}{k} \le m/40000$. Union over all J gives the theorem's bound $e^{-m/40000}$. Every unit vector belongs to a slice, so the bound holds for the full unit sphere. Finally $\|Aw\|_2 \ge \|Aw\|_1 / \sqrt{m}$ proves the singular-value conclusion. A matrix with one constant feature becomes iid after multiplication of its rows by independent signs, which preserves these image norms and its singular values.

20 Regularity of the exterior projection

Lemma 20.1 (Fixed high moments of sign quadratic forms)**.** *For each positive integer q there is a finite numerical C_q such that, for independent uniform sign vectors ϵ, η and any real matrix K,*

$$\mathbb{E}|\epsilon^T K \eta|^{2q} \le C_q \|K\|_F^{2q}.$$

For a square matrix K the same bound holds for $\epsilon^T K \epsilon - \operatorname{tr} K$, with $\|K\|_F$ on the right.

Proof. Expansion into even monomials gives $\mathbb{E}|\sum a_i \epsilon_i|^{2q} \le (2q-1)!! \|a\|^{2q}$. Condition on η and apply this bound to the sum over ϵ. Minkowski's inequality in L^q gives

$$\big(\mathbb{E}\|K\eta\|^{2q}\big)^{1/q} \le \sum_i \big(\mathbb{E}|(K\eta)_i|^{2q}\big)^{1/q} \le \big((2q-1)!!\big)^{1/q} \|K\|_F^2.$$

This proves the bilinear assertion. For the quadratic assertion remove the diagonal and symmetrize K. Introduce an independent random subset I by including each index with probability $1/2$. The off-diagonal quadratic form is four times the conditional expectation, over I, of $\sum_{i \in I, j \notin I} K_{ij} \epsilon_i \epsilon_j$. Jensen's inequality and the bilinear bound on the two independent groups of signs finish the proof, with an additional factor 4^{2q}. □

Lemma 20.2 (Projection regularity)**.** *Let X be a $d \times m$ iid sign matrix with $d/m \to 1/u$ for fixed $u = 2 - 10^{-29}$, and let P be the orthogonal projection onto* $\ker X$*. Then, with probability tending to one,*

$$\sigma_{\min}(X) \geq (1/25)\sqrt{m}, \qquad \max_i P_{ii} \leq 0.500001, \qquad \max_{i \neq j} |P_{ij}| \leq 10^{-16}.$$

The statement remains true when one row of X is the constant row $\mathbf{1}^T$ and its other rows are iid signs.

Proof. The ordinary coercivity theorem proved above implies $\|X^T v\|_2 \geq \|X^T v\|_1/\sqrt{m} \geq (1/25)\sqrt{m}\|v\|$ because d/m lies in $[0.49, 0.51]$ eventually. Its exponentially small failure probability permits a union over all matrices obtained by deleting at most two columns of X. For these matrices the harmless replacement of m by $m-2$ only changes the constant by a factor tending to one.

Write the columns as x_i, put $A = XX^T$, and let $K_i = (A - x_i x_i^T)^{-1}$. On the preceding event, $\|K_i\|_{\mathrm{op}} = O(1/m)$ and $\|K_i\|_F = O(m^{-1/2})$, where the constants may be large but are fixed. Conditional on all other columns, the quadratic-form bound with $q = 2$ gives

$$\mathbb{E}|x_i^T K_i x_i - \operatorname{tr} K_i|^4 = O(m^{-2}).$$

For rigor, truncate K_i to zero when its singular-value event fails; that event depends only on the other columns. Markov and a union over i therefore show that the displayed deviations are $o(1)$ uniformly. Sherman–Morrison also gives

$$0 \leq \operatorname{tr} K_i - \operatorname{tr} A^{-1} = \frac{x_i^T K_i^2 x_i}{1 + x_i^T K_i x_i} \leq \|K_i\|_{\mathrm{op}} = O(1/m).$$

For $Q = I - P = X^T A^{-1} X$ it follows that

$$Q_{ii} = \frac{x_i^T K_i x_i}{1 + x_i^T K_i x_i} = \frac{\operatorname{tr} A^{-1}}{1 + \operatorname{tr} A^{-1}} + o(1)$$

uniformly. Since $\operatorname{tr} Q = d$, every diagonal is $d/m + o(1)$.

For $i \neq j$ condition on the remaining columns and write $K_{ij} = (A - x_i x_i^T - x_j x_j^T)^{-1}$. Woodbury gives

$$Q_{ij} = \frac{x_i^T K_{ij} x_j}{(1 + x_i^T K_{ij} x_i)(1 + x_j^T K_{ij} x_j) - (x_i^T K_{ij} x_j)^2}.$$

The denominator is $\det(I + U^T K_{ij} U) \geq 1$, with $U = [x_i, x_j]$. The bilinear sixth-moment bound gives $\mathbb{E}|x_i^T K_{ij} x_j|^6 = O(m^{-3})$ after the same truncation. A union over the $O(m^2)$ pairs proves $\max_{i \neq j} |Q_{ij}| = o(1)$. This proves the entrywise assertions, since $1 - 1/u < 1/2$.

Finally, multiply each column of a matrix with one constant row by an independent uniform sign. The resulting matrix is fully iid sign: the map on each column is a bijection of the sign cube. Column sign changes preserve singular values, projection diagonals, and absolute off-diagonal entries. The assertion therefore transfers exactly. □

21 Technical proof of the full-network transfer

Write the neuron rows of S as $y_i^\top$, and let $X_i = (S_{-i}; \mathbf{1}^\top)$. The signed design is $X_i D_{y_i}$. For each fixed number of omitted neurons the exterior and bias design has the same limiting load u. In this appendix, "RM1" denotes the proved positive-orthant coercivity estimate with a fixed uniform coefficient a_0; it applies simultaneously to full and bias-deleted neuron designs with exponential failure probability. The operator norm is at most $10\sqrt{m}$ with exponential failure probability. These estimates supply the uniform moments and prevent a vanishing internal row normalization. They do not require the cone-aperture conclusion. The notation λ_+ uses the nonnegative unit ball as defined earlier.

22 Two-sided kernel sampling and graph testability

Theorem 22.1 (Uniform fourth-moment sampling)**.** *Let h_ℓ be symmetric kernels on possibly changing finite probability spaces with $\sup_\ell \mathbb{E}|h_\ell(X,Y)|^4 \le C$. For $r_\ell \to \infty$, sample independently and set $H_{ij} = h_\ell(X_i, X_j)$ for $i \neq j$, $H_{ii} = 0$. Then*

$$|\lambda_+(H/r_\ell) - \lambda_+(T_{h_\ell})| \xrightarrow{\mathbb{P}} 0. \tag{T1}$$

The convergence is uniform over all such kernels. There is no condition relating r_ℓ to the dimension of the space.

Proof. Fix a spectral cutoff $\zeta > 0$. Retain eigenvalues with absolute value greater than ζ, writing $h = h_L + h_R$. The retained rank is $N \le \sqrt{C}/\zeta^2$, and $\|T_{h_R}\|_{\mathrm{op}} \le \zeta$. Each retained eigenfunction obeys

$$|\phi_j(x)|^2 \le \zeta^{-2} \int h(x,y)^2 \, d\mu(y), \qquad \mathbb{E}\phi_j^4 \le C/\zeta^4. \tag{T2}$$

Thus both pieces have fourth moments bounded by constants depending only on C, ζ; for instance $\|h_L\|_4 \le NC^{3/4}/\zeta^2$.

For $\Phi = (\phi_1, \dots, \phi_N)$, define compact convex bodies

$$K = \{\mathbb{E}[g\Phi] : g \ge 0,\ \mathbb{E}g^2 \le 1\}, \quad K_r = \{r^{-1/2} \sum_i z_i \Phi(X_i) : z \ge 0,\ \|z\| \le 1\}. \tag{T3}$$

Their support functions squared are $\mathbb{E}(a^\top \Phi)_+^2$ and $r^{-1}\sum_i (a^\top \Phi(X_i))_+^2$. These converge uniformly over unit a: use a finite net, Chebyshev, $\mathbb{E}\|\Phi\|^4 \le N^2 C/\zeta^4$, and

$$|(a^\top \Phi)_+^2 - (b^\top \Phi)_+^2| \le 2\|a-b\|\|\Phi\|^2. \tag{T4}$$

Hence the bodies converge in Hausdorff distance. By Bessel, K lies in the unit ball; empirical covariance convergence bounds the radius of K_r. Maximizing $\sum_j \lambda_j v_j^2$ over the bodies therefore gives two-sided convergence for h_L. Removing its sampled diagonal costs $r^{-1}\max_i |h_L(X_i, X_i)| = o_{\mathbb{P}}(1)$, because its diagonal has a uniformly bounded second moment.

For the zero-diagonal residual matrix, expanding the fourth trace according to the number of distinct sample indices gives

$$\mathbb{E}\,\mathrm{tr}(H_R/r)^4 \le \mathrm{tr}(T_{h_R}^4) + (2/r + 1/r^2)\mathbb{E}h_R^4 \le \zeta^2\sqrt{C} + O_{C,\zeta}(r^{-1}). \tag{T5}$$

Four distinct indices give $\mathrm{tr}(T_{h_R}^4)$; three give $\mathbb{E}h_R(X,Y)^2 h_R(X,Z)^2 \le \mathbb{E}h_R^4$; two give $\mathbb{E}h_R^4$. Therefore, for fixed $\epsilon > 0$,

$$\limsup_\ell \mathbb{P}(\|H_R/r_\ell\|_{\mathrm{op}} > \epsilon) \le \zeta^2\sqrt{C}/\epsilon^4. \tag{T6}$$

Take $r_\ell \to \infty$ first and then $\zeta \downarrow 0$. The population residual also has operator norm at most ζ. This proves the assertion, including its uniformity. □

Corollary 22.2 (Finite-graph testability)**.** *View a symmetric zero-diagonal matrix C_n as a kernel on the uniform space of its vertices. If $n^{-2}\sum_{ij}|(C_n)_{ij}|^4 \le M$, a random sample of k vertices with replacement satisfies*

$$\lambda_+(C_n[I]/k) = \lambda_+(C_n/n) + o_{\mathbb{P}}(1) \quad (k \to \infty), \tag{T7}$$

uniformly in n and in these matrices. For fixed k and $n \to \infty$, sampling without replacement has the same limit.

23 Smooth positive selectors and feature response

All Hessian norms and gradient dual norms below are restricted to the tangent space $\mathbf{1}^\perp$.

For fixed $\eta, \tau > 0$, define $z = z(X, y) > 0$, $\mathbf{1}^\top z = m$, as the unique minimizer of

$$\mathcal{F}_{X,y}(z) = m^{-1}\|XD_y z\|^2 + \eta\|z\|^2 - \tau \sum_\mu \log z_\mu. \tag{T8}$$

Set $f = D_y z/m$. Then $y^\top f = 1$ and $y_\mu f_\mu > 0$. If $\|X\|_{\mathrm{op}} \le C_X\sqrt{m}$, comparison with $z = \mathbf{1}$ and Jensen give

$$\|z\|^2/m \le (C_X^2 + \eta)/\eta =: M_0. \tag{T9}$$

Define

$$H = 2D_y X^\top X D_y/m + 2\eta I + \tau \operatorname{diag}(z_\mu^{-2}), \qquad K = H^{-1} - \frac{H^{-1}\mathbf{1}\mathbf{1}^\top H^{-1}}{\mathbf{1}^\top H^{-1}\mathbf{1}}. \tag{T10}$$

Thus K is the tangent inverse on $\mathbf{1}^\perp$, and

$$K \preceq (2\eta)^{-1} I, \quad K \preceq \tau^{-1}\operatorname{diag}(z_\mu^2), \quad \theta = (1 + 2\operatorname{tr} K/m)^{-1} \in [\eta/(1+\eta), 1]. \tag{T11}$$

Lemma 23.1 (Finite-feature response)**.** *Let a be an independent uniform sign vector, $v = D_y a/\sqrt{m}$, $t = v^\top z$, and let z_1 minimize $\mathcal{F}_{X,y}(z) + (v^\top z)^2$. Uniformly over $\|X\|_{\mathrm{op}} \le C_X\sqrt{m}$, for each fixed finite p,*

$$v^\top z_1 = \theta t + o_{L^p}(1), \quad \|z_1 - z\| = O_{L^p}(1), \quad \max_\mu |z_{1,\mu}/z_\mu - 1| = o_{\mathbb{P}}(1). \tag{T12}$$

Adding this feature changes $2\operatorname{tr} K/m$ by $o_{\mathbb{P}}(1)$.

Proof. For $q = v^\top K v$, the constrained Newton step is

$$\Delta = -2tKv/(1 + 2q). \tag{T13}$$

Since $(K^2)_{\mu\mu} \le z_\mu^2/(2\eta\tau)$, conditional Rademacher concentration gives

$$\max_\mu |(Kv)_\mu|/z_\mu = O_{\mathbb{P}}(\sqrt{\log m/m}).$$

Also t has uniformly subgaussian moments by (T9). Consequently $\rho = \max|\Delta_\mu|/z_\mu = o_{\mathbb{P}}(1)$. For $H' = H + 2vv^\top$,

$$\|\Delta\|_{H'}^2 = 4t^2 q/(1 + 2q) \le 2t^2. \tag{T14}$$

Within a relative coordinate neighborhood of radius ρ, the barrier Hessian, and hence the full Hessian, is bounded multiplicatively between $(1+\rho)^{-2}$ and $(1-\rho)^{-2}$ times its baseline value. The tangent gradient residual at $z + \Delta$ therefore has H'^{-1} norm $e = o_{\mathbb{P}}(1)$.

An H'-ball of radius $8e$ about $z + \Delta$ stays positive and in a relative neighborhood of radius $\rho + 8e/\sqrt{\tau} = o_{\mathbb{P}}(1)$, since $H' \succeq \tau \operatorname{diag}(z_\mu^{-2})$. Within the ball the Hessian is at least $H'/2$, eventually. The objective on the boundary exceeds its central value by at least $-8e^2 + (8e)^2/4 > 0$. Convexity traps the exact minimizer inside the ball. Thus

$$v^\top z_1 = t/(1 + 2v^\top K v) + o_{\mathbb{P}}(1). \tag{T15}$$

The conditional variance estimate $\operatorname{Var}(v^\top K v \mid K) \le 2\|K\|_{\mathrm{F}}^2/m^2 = O(m^{-1})$ gives $v^\top K v - \operatorname{tr} K/m = o_{\mathbb{P}}(1)$. Exact convex optimality gives sign-preserving shrinkage $|v^\top z_1| \le |t|$ and $\|z_1 - z\| \le |t|/(2\sqrt{\eta})$. These and the subgaussian moments upgrade the prediction to every fixed L^p. Relative coordinate convergence follows from the Newton trap. It makes the old-core Hessian multiplicatively close to H. Adding the rank-one Hessian $2vv^\top$ changes its tangent inverse trace by at most $1/(2\eta)$. Division by m proves trace stability. □

24 The actual full-row center matrix

Take $X_i = (S_{-i}; \mathbf{1}^\top)$ in (T8) and denote its selector quantities by f_i, z_i, K_i, θ_i. For $i \neq j$ set

$$B_{ij} = \frac{y_j^\top f_i}{\theta_i \|f_i\|}, \quad D_i = \sum_{j \neq i} B_{ij}^2, \quad W_{ij} = B_{ij}/\sqrt{D_i}, \quad W_{ii} = 0. \tag{T16}$$

Thus each row of W is the internal residual coefficient vector, normalized to norm one. Define it arbitrarily with norm one if $D_i = 0$; RM1 makes that event negligible at any required polynomial rate.

Lemma 24.1 (Finite arrays, degrees, and moments)**.** *Fix k sampled vertices, condition on the other rows, and use the common core $X_0 = (S_{\mathrm{ext}}; \mathbf{1}^\top)$. Let $u_i^0 = f(X_0, y_i)/\|f(X_0, y_i)\|$. Jointly on this fixed vertex set, for every fixed finite p,*

$$B_{ij} = (u_i^0)^\top y_j + o_{L^p}(1). \tag{T17}$$

All fixed moments of B_{ij} are uniformly bounded, and

$$D_i/(n-1) \longrightarrow 1 \quad \text{in } L^2 \quad \text{for any fixed } i. \tag{T18}$$

Proof. For edge i, j, first add the other $k-2$ features and add target j last. Each finite update changes f by $O_{L^p}(m^{-1})$. The omitted target is independent of earlier changes, so its score changes by $O_{L^p}(m^{-1})$. Lemma 23.1 handles the last addition. Trace stability identifies its row factor with θ_i, and $\|f\| \geq m^{-1/2}$ makes the relative change in its norm vanish. This proves (T17).

For moments, omit only target j. Exact rank-one shrinkage gives $|y_j^\top f_i| \leq |y_j^\top f_i^{-j}|$. Conditional on omitted-target data, the right side is a Rademacher linear form with norm at most $\sqrt{M_0/m}$. Use $\theta_i \geq \eta/(1+\eta)$ and $\|f_i\| \geq m^{-1/2}$. On operator-norm exceptional events, the deterministic bound $|B_{ij}| \leq [(1+\eta)/\eta]\sqrt{m}$ suffices.

With two vertices, (T17) gives $\mathbb{E}B_{12}^2 \to 1$. With three, conditional on X_0, y_1, the two targets are independent and each squared score has mean one; hence $\mathbb{E}B_{12}^2 B_{13}^2 \to 1$. The moment bounds and exchangeability give (T18). □

Lemma 24.2 (Internal normalization)**.** *Set $L_{ij} = \sqrt{n}\, W_{ij}$. For every fixed sampled array,*

$$L_{ij} = (u_i^0)^\top y_j + o_{L^p}(1). \tag{T19}$$

Its fixed moments are bounded uniformly. If $C_{ij} = L_{ij} L_{ji}$, $C_{ii} = 0$, then

$$\sup_n \mathbb{E}\left[n^{-2} \sum_{ij} |C_{ij}|^4 \right] < \infty. \tag{T20}$$

Proof. Use RM1 on the signed row design *without bias*: simultaneously for all rows,

$$\sum_{j \neq i} (y_j^\top f_i)^2 \geq a_0^2 (n-1) \|f_i\|^2. \tag{T21}$$

Since $\theta_i \leq 1$, this gives $D_i/n \geq d_0(u) > 0$. Together with (T18) it gives (T19). On the exceptional event, $|L_{ij}| \leq \sqrt{n}$, since the normalized row has norm one; the exponential RM1 failure probability dominates every moment used. Finally apply Cauchy–Schwarz to the eighth moments of L_{ij}, L_{ji}. □

Part III

Progress at 1.9 and the remaining research problem

25 An auxiliary center incompatibility result at load 1.9

This section proves a statement at the actual target load $u = 19/10$. It concerns the regularized center rows only. It is not a storage-capacity bound, because the required comparison with every feasible row in the classification cones remains unproved at this load. Unlike the elementary $2 - 10^{-29}$ capacity branch, this auxiliary result uses the published Oymak–Tropp theorem.

The precise external input. Write $\sigma_T(A) = \inf_{x\in T} \|Ax\|$ and

$$\mathcal{E}_r(T) = \mathbb{E} \inf_{x\in T} \{\sqrt{r}\,\|x\| + G^\top x\}, \qquad G \sim N(0, I_N).$$

Oymak–Tropp, Theorem 9.1(1)–(3), for an $r \times N$ iid sign matrix A, $r \le N^{6/5}$, and nonempty closed $T \subseteq B_2^N$, gives universal constants C such that, for $z \ge 0$,

$$\begin{aligned}
\mathbb{P}\{\sigma_T(A)^2 \le \mathbb{E}\sigma_T(A)^2 - Cz\} &\le e^{-z^2/r},\\
\mathbb{P}\{\sigma_T(A)^2 \ge \mathbb{E}\sigma_T(A)^2 + Cz\} &\le C\exp\{-z^2/(r + z\sqrt{N})\},\\
\mathbb{E}\sigma_T(A)^2 &\ge (\mathcal{E}_r(T)_+)^2 - C(r+N)^{0.92}.
\end{aligned}$$

When T is convex it also gives

$$\mathbb{E}\sigma_T(A)^2 \le (\mathcal{E}_r(T)_+)^2 + C(r+N)^{0.94}.$$

Constants may be enlarged between lines. These are applications to iid Rademacher matrices, which meet the paper's bounded matrix model with bound one. Neither the ordinary unit sphere nor any other nonconvex set is used in the convex upper-bound statement.

Lemma 25.1 (Two quantitative probability estimates at $u = 1.9$)**.** *Let $B \in \{-1,1\}^{d\times m}$ be iid and $m/d \to 19/10$. Then*

$$\mathbb{P}\{\kappa(B)^2 > 0.0011\} \longrightarrow 0, \qquad \mathbb{P}\{s_{\min}(B) < 0.27\sqrt{m}\} \longrightarrow 0.$$

The second failure probability is exponentially small in m; the first can be bounded by $Ce^{-c\sqrt{m}}$ for sufficiently large dimensions. The constants are fixed and need not be uniform as the target load changes.

Proof. For the singular-value assertion apply the lower expectation and lower-tail statements to $A = B^\top$, $T = \mathbb{S}^{d-1}$. Here

$$\mathcal{E}_m(T) = \sqrt{m} - \mathbb{E}\,\|G_d\| \ge \sqrt{m} - \sqrt{d},$$

and consequently

$$\liminf \frac{1}{m}\mathbb{E}s_{\min}(B)^2 \ge (1 - \sqrt{10/19})^2 > 0.074.$$

For example $\sqrt{10/19} < 0.726$ gives a square exceeding $0.274^2 = 0.075076 > 0.074$. Thus the expectation is eventually at least $0.074m$. Choose $z = 0.001m/C$ in the lower-tail statement. Since $0.27^2 = 0.0729 < 0.074 - 0.001$, the asserted exponential estimate follows. The dimension condition $m \le d^{6/5}$ holds eventually.

For the margin, fix the deterministic cap $M = 100$ and let

$$\Delta_{m,M} = \{\pi \geq 0 : \mathbf{1}^\top \pi = 1,\ \pi_j \leq M/m\}, \qquad T = \sqrt{m/M}\,\Delta_{m,M}, \qquad \kappa_M(B) = \min_{\pi \in \Delta_{m,M}} \|B\pi\|\,.$$

This T is compact, convex and contained in the unit ball. Use the Gaussian trial weights $Z_j = \min\{(-G_j)_+, 5\}$, divided by their sum if this sum is at least half its mean, and uniform weights otherwise. Let $\mu_1 = \mathbb{E}Z_j$, $\mu_2 = \mathbb{E}Z_j^2$ and $b = \mathbb{E}[(-G_j)_+ Z_j]$. They satisfy

$$\mu_1 > 0.3988, \qquad \mu_2 \leq 1/2, \qquad b > 0.499999, \qquad M > 2(5)/\mu_1.$$

These bounds follow directly from $\mu_1 = \phi(0) - \phi(5) + 5\overline{\Phi}(5)$, $b = 1/2 - \overline{\Phi}(5)$, and the Gaussian tail estimate $\overline{\Phi}(5) \leq \phi(5)/5 < 3 \cdot 10^{-7}$. The fixed-cap law-of-large-numbers argument, with its uniformly bounded second moment, gives

$$\limsup \mathcal{E}_d(\Delta_{m,M}) \leq \frac{\sqrt{\mu_2/(19/10)} - b}{\mu_1} < \frac{0.513 - 0.499999}{0.3988} < 0.0327.$$

Here $1/\sqrt{3.8} < 0.513$ is an exact decimal check. The convex upper expectation statement, rescaled by M/m, therefore gives

$$\limsup \mathbb{E}\kappa_M(B)^2 \leq 0.0327^2 = 0.00106929 < 0.00108.$$

Eventually this expectation is at most 0.00108.

Let $Q = \sigma_T(B)^2 = (m/100)\kappa_M(B)^2$. The event $\kappa_M(B)^2 > 0.0011$ makes $Q - \mathbb{E}Q > 2 \cdot 10^{-7} m$. In the upper-tail statement choose $z = 10^{-7} m/C$. Its exponent is of order $\sqrt{m}$, because its denominator is $d + z\sqrt{m}$ and z is a fixed positive multiple of m. Thus its probability is at most $Ce^{-c\sqrt{m}}$ for sufficiently large m. Finally $\kappa(B) \leq \kappa_M(B)$. □

Fixed regularization parameters. For definiteness, use the positive-coercivity theorem at upper load 1.95. Its coefficient is greater than 10^{-6}, so take

$$a_0 = 10^{-6}, \quad U = 200/a_0^2 = 2 \cdot 10^{14}, \qquad a_{\mathrm{mix}} = \eta = \tau = 10^{-64}.$$

The uniform comparator and positive coercivity give $\kappa_{\min} > 10^{-7}$, and the existing mixed-simplex comparison gives

$$D \leq 2a_{\mathrm{mix}}\kappa L_0 + a_{\mathrm{mix}}^2 L_0^2 + \eta U + \tau \log(1/a_{\mathrm{mix}}) < 3 \cdot 10^{-50}, \qquad \kappa, L_0 \leq 10.$$

In particular $D < 10^{-12}$ and $r_{\mathrm{reg}} < 4 \cdot 10^{-18}$. These are fixed constants before the dimension limit.

Proposition 25.2 (Center reciprocity at load 1.9)**.** *At $m = \lfloor 1.9n \rfloor$, form the regularized full-row center matrix W with the preceding fixed parameters, normalizing each internal row to norm one as in the full-center construction. The established finite-feature and sampling transfer, together with the external theorem stated above, gives*

$$\mathbb{P}\left\{ \inf_{t \geq 0,\, \|t\| = 1} \sum_{i<j} (t_i W_{ij} - t_j W_{ji})^2 \geq 0.003 \right\} \longrightarrow 1.$$

This statement does not assert that the full feasible cones have diameter small enough to imply nonstorage.

Proof. Consider a common exterior X with any fixed number of target neurons removed, including the constant bias row, and an independent uniform label y. Its signed design XD_y is iid in its marginal law. Thus the preceding lemma, positive coercivity and the comparator event imply, outside a joint probability tending to zero,

$$\ell_X(y) := \left\| (I - P_X) \frac{f(X,y)}{\|f(X,y)\|} \right\|^2 \le \frac{0.0011 + D}{0.27^2} < 0.016.$$

This uses $\|f\| \ge m^{-1/2}$. No assertion uniform in all labels is being made. Define the conditional bad-label mass

$$p_X = \Pr_y\{\ell_X(y) > 0.016 \mid X\}.$$

The joint estimate gives $\mathbb{E}_X p_X \to 0$. Therefore $\Pr_X(p_X > 10^{-10}) \le 10^{10}\mathbb{E}_X p_X \to 0$. The projector regularity argument also gives, with probability tending to one, $\max_i (P_X)_{ii} \le 0.500001$ and $\max_{i \ne j} |(P_X)_{ij}| \le 10^{-16}$; its limiting diagonal is $9/19$. Call cores satisfying these two conditions and $p_X \le 10^{-10}$ good.

On a good core set $u(y) = f(X,y)/\|f(X,y)\|$. This is unit and sign aligned. For a nonnegative test function supported on $G_X = \{y : \ell_X(y) \le 0.016\}$, the new positive-map inequality gives a reciprocity loss at least

$$\left(\sqrt{3/80} - \sqrt{0.016}\right)^2 > 0.0045.$$

The last strict inequality follows from $3/80 + 0.016 = 0.0535$ and $\sqrt{(3/80)(0.016)} = \sqrt{0.0006} < 0.0245$. Thus the positive-cone quadratic form on G_X is at most 0.9955 times its squared norm. For its complement, the kernel's fourth moment is at most 105, so the cross-block squared operator norm is at most $11\sqrt{p_X} \le 0.00011$, and the bad-block operator norm is at most $4\sqrt{p_X} \le 0.00004$.

The two-block comparison matrix is bounded above by $0.996I$: the difference has positive diagonal entries 0.0005 and 0.99596, and determinant at least $0.0005(0.99596) - 0.00011 > 0$. Hence every good core has population positive-cone kernel value at most 0.996, a gap of 0.004.

The full-center transfer requires fixed regularization, positive coercivity, bounded fixed kernel moments, the finite-feature response, and this good-core population gap. Those inputs now hold with good-core probability tending to one. Its proof does not require a cone-aperture estimate once the good-core leakage argument has been replaced. It preserves every smaller fixed gap, in particular 0.003. Applying it proves the proposition. □

What is still needed for $\alpha_- \le 1.9$. The proposition controls one selected center in each classification cone. The network, if it exists, may select other vectors in those cones. An upper-bound proof still needs a simultaneous bound on the entire cone, or a direct argument incorporating all its vectors and all nonnegative row scales. The center result therefore closes one quantitative part of that program at 1.9 while leaving the decisive full-cone step open.

25.1 A barrier to using the certified 0.003 gap with a diameter bound

The missing diameter estimate is not merely an unchecked numerical detail. At load 1.9, the following lower bound shows that the diameter-based comparison cannot close with only the presently certified center gap.

Lemma 25.3 (A lower bound for the cone diameter at load 1.9)**.** *Under the same published Oymak–Tropp input, an iid sign design $B \in \{-1,1\}^{d\times m}$ with $m/d \to 19/10$ satisfies*

$$\mathbb{P}\{\kappa(B) > 0.030\} \longrightarrow 1.$$

Consequently its unit feasible cone has diameter greater than 0.057 with probability tending to one.

Proof. On the positive-coercivity and uniform-comparator event, every hard simplex optimizer obeys $m\,\|\pi_*\|^2 \le U$, with the same fixed $U = 2 \cdot 10^{14}$ as above. Let

$$K_{m,U} = \{\pi \ge 0 : \mathbf{1}^\top \pi = 1,\ \|\pi\| \le \sqrt{U/m}\}, \quad T = \sqrt{m/U}\, K_{m,U}, \quad \kappa_U(B) = \min_{\pi \in K_{m,U}} \|B\pi\|\,.$$

The set T is compact and contained in the unit ball. On the stated event, $\kappa_U(B) = \kappa(B)$.

Fix $k = 0.031$ and a standard Gaussian vector $G \in \mathbb{R}^m$. For every simplex vector π, Cauchy–Schwarz gives

$$\sqrt{d}\,\|\pi\| + G^\top \pi \ge k + \left(\sqrt{d} - \|(k\mathbf{1} - G)_+\|\right)\|\pi\|\,.$$

The scalar expectation of the squared positive part is

$$A(k) = \mathbb{E}(k - G_1)_+^2 = (1 + k^2)\Phi(k) + k\phi(k).$$

Since $\phi(k) < 0.4$ and $\Phi(k) \le 1/2 + 0.4k$,

$$A(0.031) < (1 + 0.031^2)(1/2 + 0.4 \cdot 0.031) + 0.4 \cdot 0.031 = 0.5252924164 < 0.5253 < 10/19.$$

The law of large numbers therefore shows that $E_m = \{\|(k\mathbf{1} - G)_+\| \le \sqrt{d}\}$ has probability tending to one. On E_m the preceding comparison objective is at least k for every $\pi \in K_{m,U}$. On its complement it is at least $-\sqrt{U/m}\,\|G\|$. If $p_m = \mathbb{P}(E_m^c)$, Cauchy–Schwarz yields

$$\mathcal{E}_d(K_{m,U}) \ge k(1 - p_m) - \sqrt{U/m}\,\mathbb{E}[\|G\|\,\mathbf{1}_{E_m^c}] \ge k(1 - p_m) - \sqrt{U}\sqrt{p_m} = k - o(1).$$

The large constant U is fixed before the dimension limit.

Apply the lower expectation statement of Oymak–Tropp to T and rescale. Its error is $CU(d+m)^{0.92}/m = o(1)$, so

$$\liminf \mathbb{E}\kappa_U(B)^2 \ge k^2 = 0.000961.$$

Eventually the expectation is at least 0.00095. For $Q = \sigma_T(B)^2 = (m/U)\kappa_U(B)^2$, choose $z = 0.000025m/(CU)$ in the lower-tail concentration statement. Then $\kappa_U(B)^2 \le 0.0009$ implies $Q \le \mathbb{E}Q - Cz$, because $0.0009 < 0.00095 - 0.000025$. The resulting probability is at most e^{-cm} for a fixed positive, possibly very small, constant c depending on U. Intersecting with the optimizer norm-cap event proves $\mathbb{P}\{\kappa(B) > 0.030\} \to 1$.

Let w_* be a unit maximum-margin vector on this event. Choose a coordinate j with $|(w_*)_j| \le d^{-1/2}$ and put $a = (w_*)_j$, $t = 0.029$. The two vectors $w_* \pm te_j$ are strictly feasible, since all design entries have magnitude one and $\kappa(B) - t > 0$. Their normalized squared distance is

$$2 - \frac{2(1 - t^2)}{\sqrt{(1 + t^2)^2 - 4t^2a^2}} \longrightarrow \frac{4t^2}{1 + t^2}$$

uniformly over $|a| \le d^{-1/2}$. Thus their distance tends to $2t/\sqrt{1 + t^2} > 0.057$, proving the diameter assertion. □

Using only $\gamma_{\text{full}} = 0.003$, the diameter-based final comparison requires

$$\frac{2(\rho + r_{\text{reg}})^2}{b_0^2} < 0.003.$$

Since $b_0 \le 1$ and $r_{\text{reg}} \ge 0$, this forces $\rho < \sqrt{0.003/2} < 0.039$, whereas the feasible-cone diameter is greater than 0.057 with high probability. Those two sufficient-certificate requirements are incompatible at load 1.9.

This is a barrier for combining the currently certified gap with the all-pairs diameter comparison. It is not a proof that the true center gap equals 0.003, or that capacity is at least 1.9. A stronger center-gap estimate, a direct radius bound about the selected centers, or an argument optimizing over entire cones could avoid this particular loss. In particular, a diameter lower bound is only twice a lower bound on the best possible covering radius; the two quantities must not be interchanged.

26 A global positive-dual criterion and exact finite certificates

This section supplies a deterministic criterion and finite computational certificates. It makes no asymptotic extrapolation and does not prove a capacity upper bound at load 1.9.

26.1 The margin matrix and the two different alternatives

Fix a sign matrix $S = (s_i^\mu) \in \{-1, 1\}^{n\times m}$ and set $N = nm$ and $d = n(n+1)/2$. A parameter vector $\theta \in \mathbb{R}^d$ consists of J_{ij} for $i < j$ and h_i for $i \le n$; extend J by symmetry and set $J_{ii} = 0$. Define $A = A(S) \in \mathbb{R}^{N\times d}$ by

$$(A\theta)_{i\mu} = s_i^\mu \left(\sum_{j\neq i} J_{ij} s_j^\mu + h_i \right).$$

Thus strict storage is $A\theta > 0$ in every coordinate. Write $B = A^\top$. The equation $B\lambda = 0$ is precisely

$$\sum_\mu \lambda_{i\mu} s_i^\mu = 0, \qquad \sum_\mu (\lambda_{i\mu} + \lambda_{j\mu}) s_i^\mu s_j^\mu = 0 \quad (i < j).$$

There are two relevant, distinct alternatives:

1. A nonzero certificate $\lambda \ge 0$, $B\lambda = 0$ exists exactly when strict storage is impossible. Indeed such a certificate contradicts $A\theta > 0$ by taking an inner product. Conversely, if zero is outside the convex hull of the rows of A, its nearest point to zero supplies a strictly positive separating direction. Failure of strict separation therefore puts zero in that convex hull and supplies the certificate.

2. An everywhere-positive certificate $\lambda > 0$, $B\lambda = 0$ rules out the stronger closed-cone condition

$$A\theta \ge 0, \qquad A\theta \neq 0.$$

   The equivalence in the converse direction is proved below. The requirement $A\theta \neq 0$ matters if A has a nontrivial nullspace.

The first conclusion does not automatically imply the second. For example, the matrix with rows $(1, 0), (-1, 0), (0, 1)$ has no strictly positive margin vector, but $\theta = (0, 1)$ gives nonnegative margins with one positive coordinate. Its nonnegative dual certificates necessarily have third coordinate zero. Thus a proof about positive certificates must not be silently substituted for the ordinary nonnegative Farkas alternative.

26.2 An exact uniform-box criterion

Define

$$R(S) = \sup_{\theta:\, A\theta\neq 0} \frac{|\mathbf{1}^\top A\theta|}{\|A\theta\|_1},$$

with the convention $R(S) = 0$ when $A = 0$. The triangle inequality gives $0 \le R(S) \le 1$. In the original parameters this is

$$R(S) = \sup_{J,h:\, A\theta \neq 0} \frac{|\operatorname{tr}(JSS^\top) + h^\top S\mathbf{1}|}{\sum_{i,\mu} |(Js^\mu)_i + h_i|}, \qquad J = J^\top, \quad J_{ii} = 0.$$

Proposition 26.1 (Uniform-box equivalence)**.** *For every $q \ge 0$,*

$$R(S) \le q \quad \Longleftrightarrow \quad \exists\, \delta \in \mathbb{R}^N : \quad B(\mathbf{1} + \delta) = 0, \qquad \|\delta\|_\infty \le q.$$

Consequently the following three conditions are equivalent:

$$R(S) < 1; \qquad \exists\, \lambda > 0 :\ B\lambda = 0; \qquad \nexists\, \theta :\ A\theta \ge 0,\ A\theta \neq 0.$$

Proof. Let $Z = B[-1, 1]^N$. This is a compact, centrally symmetric convex set. Its support function is

$$h_Z(\theta) = \sup_{\|z\|_\infty \le 1} \theta^\top B z = \sum_{a=1}^{N} |(A\theta)_a|.$$

By the separating-hyperplane characterization of a compact convex set, $-B\mathbf{1} \in qZ$ if and only if

$$|\theta^\top B\mathbf{1}| \le q \sum_{a=1}^{N} |(A\theta)_a| \quad \text{for every } \theta.$$

Vectors with $A\theta = 0$ impose only $0 \le 0$. This proves the displayed equivalence, including $q = 0$ and the possibly lower-dimensional case.

If $R(S) < 1$, choose $q \in [R(S), 1)$ and put $\lambda = \mathbf{1} + \delta$. Then $\lambda_a \ge 1 - q > 0$. Conversely, if $B\lambda = 0$ with $\ell = \min_a \lambda_a > 0$ and $v = \max_a \lambda_a$, rescale by $2/(\ell + v)$ to obtain a certificate in the box

$$\left[1 - \frac{v - \ell}{v + \ell},\, 1 + \frac{v - \ell}{v + \ell}\right]^N.$$

Its box radius is strictly below one.

For the last equivalence, the set $\{z \in \operatorname{range} A : \|z\|_1 = 1\}$ is compact whenever it is nonempty. Hence $R(S)$ is attained there. Equality $|\mathbf{1}^\top z| = \|z\|_1$ holds exactly when all nonzero coordinates of z have the same sign. Changing θ to $-\theta$ if necessary therefore shows that $R(S) = 1$ is equivalent to a nonzero nonnegative margin vector. The case $A = 0$ is immediate. □

Proving, at $m = \lfloor 1.9n \rfloor$, that $\mathbb{P}\{R(S) \le 1 - \delta\}$ stays bounded away from zero for some fixed $\delta > 0$ would therefore prove $\alpha_- \le 1.9$. No such probabilistic estimate is established in this section. The criterion is a precise target for a global argument; it does not assume independence between the margin rows.

26.3 A numerical construction with a rigorous stopping check

Consider the convex potential

$$\Phi(\theta) = \sum_{a=1}^{N} \exp(-(A\theta)_a) - \sum_{a=1}^{N} (A\theta)_a.$$

Its gradient is

$$\nabla\Phi(\theta) = -B\big(\mathbf{1} + \exp(-A\theta)\big).$$

Thus an exact stationary point produces an everywhere-positive dual. Minimizing this potential is only a method for proposing finite certificates. A small floating-point gradient or an optimization solver's success message is not a proof. The next lemma supplies an exact check.

Proposition 26.2 (Residual correction)**.** *Put*

$$U = \begin{pmatrix} S \\ \mathbf{1}^\top \end{pmatrix}, \qquad G = UU^\top.$$

Suppose $G \succeq sI$ for some $s > 0$. If $x \in \mathbb{R}^N$ has $t = \min_a x_a > 0$ and

$$c = \frac{\sqrt{n}\,\|Bx\|_2}{s} < t,$$

then an exact vector $\lambda > 0$ with $B\lambda = 0$ exists, and one may choose it with $\|\lambda - x\|_\infty \le c$.

Proof. For every parameter vector $\theta = (J, h)$,

$$\begin{aligned} \|A\theta\|_2^2 &= \|JS + h\mathbf{1}^\top\|_F^2 \\ &\ge s(\|J\|_F^2 + \|h\|_2^2) = s\left(2\sum_{i<j} J_{ij}^2 + \sum_i h_i^2\right) \ge s\|\theta\|_2^2. \end{aligned}$$

Hence $A^\top A \succeq sI$ is invertible. Define

$$\Delta = -A(A^\top A)^{-1}Bx, \qquad \lambda = x + \Delta.$$

Then $B\lambda = 0$. Each row of A has $n-1$ edge coefficients and one bias coefficient, all of magnitude one, so its norm is $\sqrt{n}$. It follows that

$$\|\Delta\|_\infty \le \sqrt{n}\,\|(A^\top A)^{-1}Bx\|_2 \le \frac{\sqrt{n}\,\|Bx\|_2}{s} = c < t.$$

Thus every coordinate of λ is positive. When x is rational, this formula also defines a rational certificate; no large rational linear system needs to be explicitly solved to prove its existence. □

An entirely integer verification of the sufficient inequality. Let L be a positive integer, choose a symmetric integer matrix K, and set

$$e = \max\{\|LI - KG\|_1, \|LI - KG\|_\infty\}, \qquad k = \|K\|_\infty.$$

Here the matrix 1- and ∞-norms are maximum absolute column and row sums. If $e < L$ and $k > 0$, then

$$G \succeq \frac{L-e}{k} I.$$

Indeed G is positive semidefinite, $\|I - (K/L)G\|_{\mathrm{op}} \le e/L < 1$ forces invertibility, and an eigenvector of its smallest eigenvalue gives $1 - e/L \le \lambda_{\min}(G)\|K/L\|_{\mathrm{op}}$. Symmetry of K gives $\|K/L\|_{\mathrm{op}} \le k/L$.

Write the candidate as $x = x^{\mathrm{int}}/L$ with integer coordinates. Compute $r = Bx^{\mathrm{int}}$ directly from the integer bias and edge equations. Let r_2 be an integer with $r_2^2 \ge \sum_a r_a^2$, and let n_2 be an integer with $n_2^2 \ge n$. Define

$$C = n_2 r_2 k, \qquad x_{\min}^{\mathrm{int}} = \min_a x_a^{\mathrm{int}}.$$

The single integer inequality

$$x_{\min}^{\mathrm{int}}(L-e) > C$$

proves existence of the exact positive dual, with correction bound $c \le C/[L(L-e)]$. If $T = \sum_a x_a^{\mathrm{int}}$, its subsequent normalization to mean one has minimum coordinate at least the exact rational number

$$\frac{N\{x_{\min}^{\mathrm{int}}(L-e) - C\}}{T(L-e) + NC}.$$

Likewise Proposition 26.1 gives the exact bound

$$R(S) \le \frac{(x^{\mathrm{int}}_{\max} - x^{\mathrm{int}}_{\min})(L-e) + 2C}{(x^{\mathrm{int}}_{\max} + x^{\mathrm{int}}_{\min})(L-e)} < 1.$$

Floating point may be used to propose K or x^{int}; the certifying comparisons themselves use only integers.

26.4 Three finite certificates near load 1.9

For these finite diagnostics the generator used $m = \mathrm{round}(1.9n)$, not $\lfloor 1.9n \rfloor$. There are no half-integer rounding cases in the table. The actual loads are displayed explicitly. Candidate weights were first normalized to mean one and then rounded with $L = 10^9$. The quoted exact floors include both residual correction and a final normalization to mean one.

n	m	actual m/n	candidate floor	exact floor $>$	$R(S) <$
32	61	$61/32 = 1.90625$	0.206407749	0.20628	0.963951
64	122	$61/32 = 1.90625$	0.162303678	0.15781	0.984543
96	182	$91/48 \simeq 1.895833$	0.159890902	0.15574	0.984614

For transparency, the integer verification quantities are:

n	e	k	r_2	n_2
32	1010	246986542	70449	6
64	3084	329379897	1472117	8
96	6548	174344094	2056837	10

They give correction bounds respectively below 0.000104400, 0.003879098, and 0.003585998. The reported strict inequalities for the exact floors and for $R(S)$ were checked by integer cross multiplication against the displayed decimal rationals.

These examples have everywhere-positive certificates, which exclude even a nonzero weak margin vector. They are not examples conditioned on every neuron being separately feasible. Independent numerical row LPs reported 12, 12, and 7 individually infeasible neurons respectively; one additional row in the last case was unresolved within its time limit. Those row-LP diagnoses are not used by the exact positive-dual proof.

The final PDF embeds the reproducibility files as attachments:

- `verify_target19_certificates.py`: exact verifier.
- `target19_verified_certificates.npz`: sign matrices, rounded candidate weights, and integer approximate inverses.
- `target19_certificate_verification.json`: exact bounds and integer comparison data.
- `toward_1_9_capacity.tex`: the consolidated proof source.

The finite statements prove neither a positive limiting probability nor a capacity threshold. Their role is to provide verified examples and a quantitative target for an eventual probabilistic argument.

27 What prevents this argument from reaching 1.9?

The improvement is rigorous within the displayed research argument, but its size must be described honestly. Increasing the deficit from 10^{-89} to 10^{-29} does not make the bound numerically close to 1.9. The distance to that target remains $0.1 - 10^{-29}$.

There is a useful diagnostic showing that merely tuning the present scalar gates cannot solve the target. Even allowing an ideal population gap at most one and setting every regularization loss to zero, the final gate requires $2\rho^2 < 1$, because $b_0 \le 1$ and a full mismatch constant cannot exceed one (test a coordinate unit vector t). The Gaussian conversion gate requires $33\sqrt{\delta} < \rho/2$. Thus

$$\delta < \frac{1}{8712}.$$

Even ignoring the positive thickening errors, the width gate requires

$$\sqrt{1 - 1/u} > \frac{1 - \delta^2/2}{\sqrt{2}}, \qquad u = 2 - \epsilon.$$

Squaring and subtracting $1/2$ gives

$$\frac{\epsilon}{2(2-\epsilon)} < \frac{\delta^2}{2} - \frac{\delta^4}{8}.$$

Since the left side is at least $\epsilon/4$, necessarily

$$\epsilon < 2\delta^2 < \frac{2}{8712^2} < 3 \cdot 10^{-8}.$$

This is only a limitation of these particular sufficient gates, not a lower bound on the true storage capacity. The actual head-deletion and width errors impose substantially stronger losses. A theorem at 1.9 must therefore change the geometric mechanism, rather than optimize this certificate's constants again.

27.1 Three concrete programs, in order of priority

1. A global positive-dual bound. Use the exact ratio $R(S)$ in the preceding section. Prove at $m = \lfloor 1.9n \rfloor$ that $R(S) \le 1 - \delta$ with probability bounded away from zero, for fixed $\delta > 0$. This would immediately produce a positive dual and prove $\alpha_- \le 1.9$. The missing inequality couples $\operatorname{tr}(JSS^\top) + h^\top S\mathbf{1}$ to the entire absolute-field sum, uniformly over symmetric hollow J and arbitrary h. Separate crude norm bounds discard the dependence that matters. A successful proof must exploit symmetry inside that uniform comparison.

2. Positive duals with a controlled correction. Construct a positive approximate certificate, prove a lower bound on its smallest coordinate, and control an exact correction solving the bias and edge equations. The finite residual-correction proposition gives the precise deterministic inequality to aim for. For the asymptotic problem, Euclidean correction bounds alone may be too weak: a useful theorem should bound the maximum-coordinate correction on the actual structured residual. The exponential potential is a way to propose candidates, but proving existence of its minimizer is part of the unresolved problem, not an automatic step.

3. A comparison for all feasible row choices. The current proof compares every row to one selected center using the worst possible cone diameter. A more direct approach would bound the smallest reciprocal mismatch over the product of all feasible cones, allowing row scales simultaneously. The target is a strictly positive lower bound for

$$\inf_{R_i \text{ feasible unit internal rows } t \ge 0,\ \|t\|=1} \sum_{i<j} (t_i R_{ij} - t_j R_{ji})^2.$$

Its exact bias constraints must remain in the feasible-row definition. A positive bound rules out symmetry immediately and avoids the narrow-cone step. Population information about a single selector is insufficient: one needs a uniform theorem over the choices of all rows, or a valid dual certificate for this infimum. This is the precise place where a macroscopic improvement would enter.

28 Literature audit and model distinctions

The present capacity concerns optimally chosen real symmetric hollow weights and strict storage of iid unbiased binary memories. A result for Hebbian weights, a prescribed learning algorithm, nonzero self-couplings, specially designed memories, or a retrieval basin is not automatically a bound for this problem.

Gardner, Gutfreund and Yekutieli analyze interactions with prescribed symmetry, but their analytic calculation is explicitly for strong dilution, C of order $\log N$ [1]. Aguirre-López, Pastore and Franz revisit that program in 2022 and explicitly identify the fully connected many-pattern case as a problem requiring further work [2]. These sources help explain why substituting independent row orientations or a diluted-network answer is not justified.

The official 1990 Nardulli–Pasquariello preview is a genuine conflicting historical claim [3]. It states symmetric zero-diagonal interactions and attributes the Gardner saturation formula to their symmetric analysis. The complete 1991 derivation was not obtained for an assumption-by-assumption audit. Thus this investigation neither treats the original claim as a rigorously verified theorem for all our quantifiers nor claims to have resolved why it differs from the present research argument.

Hillar, Sohl-Dickstein and Koepsell connect strict storage with a convex minimum-probability-flow optimization [4]. That is relevant to the global exponential-potential route, although an optimization formulation does not by itself give the random-pattern threshold.

Modern dependent Gaussian comparison tools are relevant, but their hypotheses must be checked. Mallory, Huang and Austern require a suitable blockwise Gaussian approximation for their universality result and a structured covariance decomposition for their dependent CGMT [6]. The repeated products $s_i^\mu s_j^\mu$ in this problem do not become independent merely because their pairwise covariances vanish. No such substitution is used in the proof above.

The light-coordinate comparison in the new positive-map lemma is an elementary degree-one instance of the broader invariance-principle method [7]. The sampling argument is related to classical random approximations of integral-operator spectra [8], but its needed uniform statement is proved here. There is no claim that all auxiliary ideas, or the numerical constant as a literature result, are new.

References

[1] E. Gardner, H. Gutfreund, and I. Yekutieli. *The phase space of interactions in neural networks with definite symmetry*. J. Phys. A: Math. Gen. 22 (1989), 1995–2008. https://doi.org/10.1088/0305-4470/22/12/005.

[2] F. Aguirre-López, M. Pastore, and S. Franz. *Satisfiability transition in asymmetric neural networks*. J. Phys. A: Math. Theor. 55 (2022), 305001. https://arxiv.org/abs/2204.03577.

[3] G. Nardulli and G. Pasquariello. *Independence from temperature of domains of attraction in neural networks*. International Neural Network Conference (1990), starting at p. 973. Official two-page preview: https://page-one.springer.com/pdf/preview/10.1007/978-94-009-0643-3_149.

[4] C. Hillar, J. Sohl-Dickstein, and K. Koepsell. *Efficient and optimal binary Hopfield associative memory storage using minimum probability flow* (2012). https://arxiv.org/abs/1204.2916.

[5] S. Oymak and J. A. Tropp. *Universality laws for randomized dimension reduction, with applications*. Information and Inference 7 (2018), 337–446. https://arxiv.org/abs/1511.09433.

[6] M. E. Mallory, K. H. Huang, and M. Austern. *Universality of High-Dimensional Logistic Regression and a Novel CGMT under Dependence with Applications to Data Augmentation*. Proceedings of Machine Learning Research 291 (2025); revised arXiv version 2026. https://arxiv.org/abs/2502.15752.

[7] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. *Noise stability of functions with low influences: invariance and optimality*. Annals of Mathematics 171 (2010), 295–341. https://annals.math.princeton.edu/2010/171-1/p05.

[8] V. Koltchinskii and E. Giné. *Random matrix approximation of spectra of integral operators*. Bernoulli 6 (2000), 113–167. https://projecteuclid.org/journals/bernoulli/volume-6/issue-1/Random-matrix-approximation-of-spectra-of-integral-operators/bj/1082665383.full.