FAILURE OF ENTROPIC SELECTION FOR DIAGONAL MATRIX COMPLETION IN DEEP LINEAR NETWORKS

RODRIGO TREVIÑO

Abstract. We study the free-energy gradient flow associated with the depth-N deep linear network on the space of invertible real $d \times d$ matrices. The loss observes only the diagonal,

$$E_d(W) = \frac{1}{2}\sum_{i=1}^{d}(W_{ii} - 1)^2,$$

and the regularizer is the Boltzmann entropy of the balanced factorization fiber computed by Menon and Yu. This is a natural higher-dimensional version of a matrix-completion problem posed by Menon as a test of whether entropy selects among a noncompact family of minimizers.

We first isolate the general obstruction to such a selection principle. Up to a uniformly bounded term, the entropy is a logarithmic sum of the maximal exterior-volume expansions of W. This yields an exact criterion for boundedness of $E - \beta^{-1}S_N$ in terms of the maximum entropy on energy sublevel sets, as well as practical sufficient conditions based on logarithmic, superlogarithmic, and polynomial growth of the energy. We prove that the entropy tends to $+\infty$ along every nontrivial affine ray through a full-rank matrix. Consequently, for a squared linear matrix-sensing loss, the free energy is bounded below if and only if the observation operator is injective; in the injective case it is coercive in norm. We also distinguish norm coercivity from attainment on the open full-rank manifold and give a compactness criterion under which low-temperature minimizers select entropy maximizers on the minimizing set of the energy.

For diagonal completion, every width $d \geq 2$, depth $N > 2$, and inverse temperature $\beta > 0$, we prove that the free energy is unbounded below and has exactly 2^d full-rank critical points, one in each diagonal sign chamber. Every critical point is diagonal and hyperbolic. Its unstable dimension is $\binom{d}{2}$ and its stable dimension is $d(d+1)/2$. We also prove that a full-rank trajectory cannot converge to a finite rank-deficient matrix. Corank at least two is excluded by an infinite free-energy barrier, while the corank-one boundary is repelling because the entropy produces a positive linear drift in the smallest singular value. The flow is forward complete. Every bounded orbit converges to one of the 2^d saddles, and the union of their stable sets has Lebesgue measure zero. Consequently, almost every full-rank initial condition has an unbounded forward orbit.

To separate this failure from the low-rank selection displayed by ordinary training, we also analyze the unregularized flow at width two. A full-rank trajectory cannot converge exactly to a nonzero rank-one completion, and the completion manifold is normally attracting at its full-rank points, so low-rank selection arises only as a singular limit of the endpoint map as the initialization scale ε tends to zero.

Thus finite-temperature fiber entropy does not provide an equilibrium selection principle for diagonal completion, or more generally for any genuinely underdetermined linear matrix-sensing problem on the unrestricted matrix space. The remaining questions concern escape rates, asymptotic directions, possible singular escape at infinity, the metastable regime as $\beta \to \infty$, and the full law of the zero-temperature endpoint under random initialization.

Date: September 2026.

2020 *Mathematics Subject Classification.* 37N40, 53C20, 15A18, 68T07.

Key words and phrases. Deep linear networks, matrix completion, implicit regularization, Boltzmann entropy, Riemannian gradient flow, stable manifolds.

Contents

1. Introduction

A deep linear network represents a matrix W as a product of N trainable factors. On balanced factorizations, Euclidean gradient flow in factor space projects to a Riemannian gradient flow for the end-to-end matrix. The metric depends on the depth and can therefore alter which minimizer is reached when a loss has a degenerate minimizing set. This provides a finite-dimensional model for implicit regularization in overparametrized learning; see [ACH18, BRTW22, CMV23, Men25].

Menon and Yu introduced a Boltzmann entropy S_N for the deep linear network by taking the logarithm of the volume of the balanced parameter-space fiber over an end-to-end matrix [MY25]. Given a loss (or energy) E and inverse temperature $\beta > 0$, this leads to the free energy

$$F_\beta(W) = E(W) - \frac{1}{\beta} S_N(W),$$

where $S_N(W)$ is the entropy, and to the Riemannian gradient flow

$$\dot{W} = -\operatorname{grad}_{g^N} F_\beta(W). \tag{1}$$

For spectral confining losses, Chen, Kotwal, and Menon characterize the equilibria and dynamics of this flow and prove strict concavity of the entropy in singular-value coordinates [CKM25]. The nonspectral case, including matrix completion, is explicitly left open there.

The motivating completion loss is

$$E_d(W) = \frac{1}{2} \sum_{i=1}^{d} (W_{ii} - 1)^2. \tag{2}$$

Its minimizers form the noncompact affine space of matrices with diagonal equal to 1. Menon posed the 2×2 version both as a low-rank selection problem for the ordinary deep-linear flow and as a test of whether the free energy selects among degenerate minimizers [Men25, Sections 14.1–14.2]. Numerical work on the unregularized flow shows attraction toward low-rank completions in small examples [CMV23].

The main conclusion of this paper is that the free-energy minimization proposal fails structurally for (2). The free energy has no minimizer, all of its finite full-rank equilibria are saddles, finite rank-deficient limits are impossible, and generic trajectories are unbounded.

To separate this finite-temperature failure from the low-rank behavior of ordinary training, Section 11 analyzes the unregularized flow at width two. We prove that a full-rank trajectory cannot converge exactly to a nonzero rank-one completion. Low-rank selection instead appears as a singular limit of the endpoint map as the initialization scale tends to zero. On a symmetric invariant plane the endpoint can be computed exactly: depth two retains a nondegenerate rank-two limit for positive determinant, whereas every depth $N > 2$ sends the smaller terminal singular value to zero at order ε. We also prove persistence under a nonlinear nonsymmetric equal-diagonal perturbation and obtain an endpoint asymmetry of order $\varepsilon^{2/N}$. The same-sign symmetric-plane endpoints recover, in width two, a recent result of Shin and Yun [SY26]; see Remark 11.6.

The failure is not caused merely by degeneracy of the minimizing set. A compact family of minimizers can support a genuine low-temperature selection principle: provided minimizers of F_β exist and remain precompact, their limit points maximize the entropy on the set of energy minimizers. What is fatal here is the combination of noncompact low-energy directions and unbounded entropy along those directions.

The entropy growth is weak in one sense and strong in another. It is only logarithmic in the size of the matrix, so any superlogarithmically confining energy dominates it. But it diverges along every nontrivial affine ray through a full-rank matrix. Therefore any loss that is bounded on such a ray has free energy unbounded below. This applies in particular to every noninjective squared linear matrix-sensing problem on the unrestricted matrix space. In Section 4 we prove the sharp dichotomy

$$\frac{1}{2}\|\mathcal{L}(W) - y\|_2^2 - \frac{1}{\beta} S_N(W) \text{ is bounded below if and only if } \mathcal{L} \text{ is injective.}$$

When $\mathcal{L}$ is injective, the quadratic sensing loss dominates the logarithmic entropy and the free energy is coercive in norm. We also formulate an exact criterion in terms of the entropy available on energy sublevel sets and explain what additional control near the rank boundary is needed for a minimizer to be attained inside $\mathrm{GL}_d(\mathbb{R})$.

1.1. Main results.

We reserve d for the matrix width and N for the network depth. The entropy S_N is defined precisely in Section 2. The following are the main results of the paper.

Theorem 1.1. *Fix $d \geq 2$, an integer depth $N > 2$, and $\beta > 0$. Let*

$$F_\beta(W) = \frac{1}{2}\sum_{i=1}^{d}(W_{ii} - 1)^2 - \frac{1}{\beta} S_N(W), \qquad W \in \mathrm{GL}_d(\mathbb{R}),$$

and let $W(t)$ solve the depth-N free-energy flow (1). Then the following statements hold.

(i) *The free energy is unbounded below:*

$$\inf_{W \in \mathrm{GL}_d(\mathbb{R})} F_\beta(W) = -\infty.$$

(ii) *There are exactly 2^d critical points in $\mathrm{GL}_d(\mathbb{R})$. Every critical point is diagonal. More precisely, for each sign vector*

$$\epsilon = (\epsilon_1, \dots, \epsilon_d) \in \{-1, 1\}^d$$

there is a unique $x^\epsilon \in (0, \infty)^d$ such that

$$W_\epsilon = \operatorname{diag}(\epsilon_1 x_1^\epsilon, \dots, \epsilon_d x_d^\epsilon)$$

is critical, and these are all the critical points.

(iii) *Every W_ϵ is a nondegenerate saddle. In addition, the stable and unstable spaces of the linearized free-energy flow satisfy*

$$\dim E^s = \frac{d(d+1)}{2}, \qquad \dim E^u = \binom{d}{2}.$$

(iv) *No solution starting in $\mathrm{GL}_d(\mathbb{R})$ can converge to a finite rank-deficient matrix.*

(v) *Every full-rank solution is defined for all $t \geq 0$ and remains in its initial connected component of $\mathrm{GL}_d(\mathbb{R})$.*

(vi) *Every bounded forward orbit converges to one of the 2^d critical points. The set of initial conditions with bounded forward orbit has Lebesgue measure zero in $\mathbb{M}_d$. Consequently, for Lebesgue-almost every $W_0 \in \mathrm{GL}_d(\mathbb{R})$,*

$$\sup_{t \geq 0} \|W(t)\|_F = \infty.$$

The theorem does not assert that $\|W(t)\|_F$ has a limit or tends monotonically to infinity. It also does not exclude an unbounded trajectory for which some singular values tend to zero while others diverge. We call this possible behavior *singular escape at infinity*. The theorem excludes only convergence to a finite singular matrix.

1.2. **Mechanism and proof strategy.** Four elementary mechanisms drive the proof.

First, a zero-loss shear has one singular value tending to infinity. The entropy grows logarithmically along this family, so the free energy tends to $-\infty$.

Second, strict concavity of the entropy in singular-value coordinates forces the polar factor of a critical matrix to align with the coordinate axes. This makes every critical point diagonal. On each diagonal sign chamber the free energy becomes a strongly convex function of the positive diagonal magnitudes, giving exactly one point per sign vector.

Third, the full matrix Hessian is indefinite at each of these diagonal points. Every unordered coordinate pair contributes one stable and one unstable off-diagonal mode. The d diagonal modes are stable. This produces a Morse index $\binom{d}{2}$.

Fourth, the entropy has two different effects near the rank boundary. If at least two singular values vanish, the entropy tends to $-\infty$, so the free energy tends to $+\infty$ and cannot be approached by a descending trajectory. If exactly one singular value vanishes, the entropy instead creates a positive drift of order σ_d, while the energy contribution is of order $\sigma_d^{2-2/N} = o(\sigma_d)$ when $N > 2$. The corank-one boundary is therefore repelling.

The remainder of the paper makes these points precise and then uses the Lyapunov identity, compactness, LaSalle's principle, and the stable-manifold theorem to obtain generic unboundedness.

1.3. **Relation to prior work and scope.** Implicit regularization in matrix factorization has been studied from several complementary viewpoints. At depth two, nuclear-norm behavior was identified under important symmetry and commuting-measurement assumptions by Gunasekar et al. [GWB+17]. Arora et al. studied the additional bias created by deeper factorizations and emphasized that it is not generally captured by a fixed norm [ACHL19]. Cohen, Menon, and Veraszto developed the Riemannian description of deep-linear matrix completion and supplied rigorous and numerical evidence relating its geometry to low-rank outcomes [CMV23]. More recently, Shin and Yun studied the factor gradient flow started from a balanced deterministic family of initializations under block-diagonal observations, derived exact equations for the limiting singular values, and showed that at depth at least three the stable rank of the limit tends to one as the initialization scale tends to zero, in contrast with depth two [SY26]. In width two their diagonal case coincides with the symmetric-plane analysis of Section 11; see Remark 11.6.

All of those low-rank statements concern the ordinary, unregularized training flow, which is the formal $\beta = \infty$ limit of the system studied here. Menon and Yu subsequently computed the balanced-fiber entropy [MY25], and Chen, Kotwal, and Menon analyzed the resulting free energy for spectral losses, proving in particular the strict concavity theorem used below [CKM25]. They explicitly left nonspectral losses such as matrix completion open.

Our principal contribution concerns the finite-β free-energy flow for the simplest diagonal observation pattern. Its conclusions do not contradict low-rank selection by the ordinary flow. Rather, they show that an arbitrarily small but fixed entropic drift changes the global long-time problem. The width-two analysis in Section 11 makes the contrast explicit: at $\beta = \infty$, full-rank endpoints can converge to rank-one completions as the initialization scale tends to zero, while at every fixed finite β the generic orbit is unbounded. We work throughout with the reduced balanced dynamics and use the full-rank fiber entropy only in the finite-temperature problem.

Tool and computational resource disclosure: Following the Leiden Declaration's recommendations, I am disclosing that ChatGPT Sol and Claude Fable were used in the preparation of this paper.

Acknowledgements: This work was supported by a Career Transition Grant from Coefficient Giving.

2. Deep-linear geometry and the entropy

Most of the material in this section can be gathered from the great survey [Men25].

2.1. The balanced end-to-end flow.

Let $\mathbb{M}_d$ denote the real $d \times d$ matrices, equipped with the Frobenius inner product

$$\langle P, Q \rangle_F = \mathrm{Tr}(P^T Q).$$

A depth-N deep linear network is an N-tuple $(W_N, \ldots, W_1)$ representing the end-to-end matrix

$$W = W_N W_{N-1} \cdots W_1.$$

Balanced factorizations satisfy

$$W_{p+1}^T W_{p+1} = W_p W_p^T, \qquad p = 1, \ldots, N-1.$$

They form invariant manifolds for factor-space gradient flow. On the full-rank balanced manifold, the end-to-end dynamics close on $\mathrm{GL}_d(\mathbb{R})$ and are determined by the positive self-adjoint operator

$$\mathcal{A}_{N,W}(P) = \sum_{k=1}^{N} (WW^T)^{\frac{N-k}{N}} P (W^T W)^{\frac{k-1}{N}}.$$

For a smooth function $H : \mathrm{GL}_d(\mathbb{R}) \to \mathbb{R}$, the corresponding Riemannian gradient is

$$\mathrm{grad}_{g^N} H(W) = \mathcal{A}_{N,W}(H'(W)),$$

where $H'(W)$ is the Euclidean gradient. These facts follow from the balanced reduction and Riemannian-submersion results in [BRTW22, Men25, MY25].

If

$$W = U\Sigma V^T, \qquad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d), \qquad \sigma_i > 0,$$

then the singular frame $u_i v_j^T$ diagonalizes $\mathcal{A}_{N,W}$:

$$\mathcal{A}_{N,W}(u_i v_j^T) = a_{ij}(\sigma) u_i v_j^T, \tag{3}$$

where

$$a_{ii}(\sigma) = N\sigma_i^{2-2/N}$$

and, for $i \neq j$,

$$a_{ij}(\sigma) = \frac{\sigma_i^2 - \sigma_j^2}{\sigma_i^{2/N} - \sigma_j^{2/N}}. \tag{4}$$

The quotient in (4) is interpreted by continuity when $\sigma_i = \sigma_j$. It is positive. Thus $\mathcal{A}_{N,W}$ is positive definite on $\mathrm{GL}_d(\mathbb{R})$. The Riemannian metric is

$$g_W^N(P, Q) = \left\langle P, \mathcal{A}_{N,W}^{-1}(Q) \right\rangle_F.$$

Given the free energy F_β, the flow studied in this paper is therefore

$$\dot{W} = -\mathcal{A}_{N,W}\big(F'_\beta(W)\big) = -\mathcal{A}_{N,W}\big(E'_d(W)\big) + \frac{1}{\beta}\mathcal{A}_{N,W}\big(S'_N(W)\big). \tag{5}$$

2.2. **The Boltzmann entropy.** Menon and Yu compute the volume of the balanced factorization fiber over a full-rank matrix [MY25]. Up to an additive constant depending only on N and d, the resulting entropy is

$$S_N(W) = \mathcal{S}_N(\sigma(W)), \tag{6}$$

where

$$\mathcal{S}_N(\sigma_1, \ldots, \sigma_d) = \frac{1}{2} \sum_{1 \le i < j \le d} \log \frac{\sigma_i^2 - \sigma_j^2}{\sigma_i^{2/N} - \sigma_j^{2/N}}.$$

The expression extends real-analytically across positive repeated singular values [CKM25, Theorem 1.8]. Since it is symmetric in the singular values, it defines a smooth orthogonally invariant function on $\mathrm{GL}_d(\mathbb{R})$.

The diagonal-completion energy and its Euclidean gradient are

$$E_d(W) = \frac{1}{2}\sum_{i=1}^{d} (W_{ii} - 1)^2, \qquad E'_d(W) = \mathrm{diag}(W_{11} - 1, \ldots, W_{dd} - 1).$$

The free energy is

$$F_\beta(W) = E_d(W) - \frac{1}{\beta} S_N(W).$$

Remark 2.1 (Rank-deficient strata). The entropy formula (6) is the full-rank fiber entropy. When we take limits toward a rank-deficient matrix, we study the boundary behavior of this full-rank function and of the full-rank flow. We do not identify that boundary value with the volume of a fiber in a lower-rank balanced stratum, whose dimension and normalization differ.

The following entropy result of Chen, Kotwal, and Menon will be used in the equilibrium analysis below.

Theorem 2.2 (Strict concavity in singular-value coordinates [CKM25])**.** *If $N > 2$ and $d \geq 2$, then*

$$\nabla^2 \mathcal{S}_N(\sigma) \prec 0 \qquad \text{for every } \sigma \in (0,\infty)^d.$$

More generally, strict concavity holds for every $(N,d) \neq (2,2)$.

We also use the standard differential calculus of singular-value functions. If f is a smooth symmetric function on $(0,\infty)^d$ and

$$\Phi(W) = f(\sigma(W)),$$

then at an SVD $W = U\Sigma V^T$,

$$\Phi'(W) = U \operatorname{diag}(\partial_1 f, \ldots, \partial_d f) V^T. \tag{7}$$

At repeated singular values, symmetry makes the right-hand side independent of the chosen singular basis. We give the second-order formula needed later in Section 7. General treatments of this calculus include [Lew96, LS01, DK15].

3. Pairwise entropy calculus

The entropy is a sum of two-variable terms. It is useful to isolate them.

Definition 3.1. For $a, b > 0$, define

$$q_N(a,b) = \frac{1}{2} \log \frac{a^2 - b^2}{a^{2/N} - b^{2/N}}, \tag{8}$$

with the analytic continuation at $a = b$. Put

$$c_N = 1 - \frac{1}{N}.$$

Then

$$\mathcal{S}_N(\sigma) = \sum_{i<j} q_N(\sigma_i, \sigma_j).$$

Lemma 3.2 (Basic identities)**.** *For every integer $N \geq 2$:*

(i) *q_N is symmetric and real analytic on $(0,\infty)^2$, including the diagonal, and*

$$q_N(a,a) = \frac{1}{2} \log N + c_N \log a.$$

(ii) *For $t > 0$,*

$$q_N(ta, tb) = q_N(a,b) + c_N \log t. \tag{9}$$

(iii) *The function q_N is strictly increasing in each argument.*

(iv) *If $a \geq b > 0$ and $r = b/a$, then*

$$q_N(a,b) = c_N \log a + h_N(r), \qquad h_N(r) = \frac{1}{2}\log\frac{1-r^2}{1-r^{2/N}}. \tag{10}$$

The function h_N extends continuously to $[0,1]$, with

$$h_N(0) = 0, \qquad h_N(1) = \frac{1}{2}\log N.$$

In particular, h_N is bounded.

Proof. Set $x = a^{2/N}$ and $y = b^{2/N}$. The divided-difference identity

$$\frac{a^2-b^2}{a^{2/N}-b^{2/N}} = \frac{x^N-y^N}{x-y} = \sum_{m=0}^{N-1} x^{N-1-m}y^m \tag{11}$$

shows symmetry, positivity, analyticity across $x = y$, and strict increase in each positive variable. At $a = b$, the sum is $Na^{2-2/N}$. Homogeneity and (10) follow by factoring out $a^{2-2/N}$. The endpoint values of h_N follow directly, using l'Hospital's rule at $r = 1$. $\square$

Write

$$g_i(\sigma) = \partial_i \mathcal{S}_N(\sigma).$$

Corollary 3.3 (Positivity of the entropy gradient)**.** *For $d \geq 2$ and every $\sigma \in (0,\infty)^d$,*

$$g_i(\sigma) > 0, \qquad i = 1,\dots,d.$$

Proof. Every g_i is the sum of the partial derivatives with respect to the ith variable of the $d-1$ pair terms involving σ_i. Each derivative is positive by Lemma 3.2. $\square$

Lemma 3.4 (Strict reverse ordering of the gradient)**.** *Assume $N > 2$. If $\sigma_i \neq \sigma_j$, then*

$$(\sigma_i - \sigma_j)(g_i(\sigma) - g_j(\sigma)) < 0. \tag{12}$$

Consequently,

$$g_i(\sigma) = g_j(\sigma) \iff \sigma_i = \sigma_j.$$

Proof. Let τ exchange coordinates i and j, and put $\widetilde{\sigma} = \tau\sigma$. Symmetry gives $\nabla\mathcal{S}_N(\widetilde{\sigma}) = \tau\nabla\mathcal{S}_N(\sigma)$. By Theorem 2.2,

$$\big(\nabla\mathcal{S}_N(\sigma) - \nabla\mathcal{S}_N(\widetilde{\sigma})\big)\cdot(\sigma - \widetilde{\sigma}) < 0$$

when $\sigma \neq \widetilde{\sigma}$. The left-hand side is

$$2(\sigma_i - \sigma_j)(g_i - g_j).$$

This proves (12). The converse implication follows from symmetry when $\sigma_i = \sigma_j$. $\square$

Lemma 3.5 (Boundary derivative)**.** *Fix $0 < m < M < \infty$. Uniformly for $a \in [m,M]$, as $b \downarrow 0$,*

$$\partial_b q_N(a,b) = \frac{1}{N}a^{-2/N}b^{2/N-1}\big(1+o(1)\big).$$

Equivalently, with $p = 2 - 2/N$,

$$Nb^p\partial_b q_N(a,b) = a^{-2/N}b + o(b). \tag{13}$$

The exact identity is

$$Nb^p\partial_b q_N(a,b) = \frac{b}{a^{2/N}-b^{2/N}} - \frac{Nb^{3-2/N}}{a^2-b^2}. \tag{14}$$

Proof. Differentiate (8) with respect to the smaller argument:

$$\partial_b q_N(a,b) = -\frac{b}{a^2-b^2} + \frac{1}{N}\frac{b^{2/N-1}}{a^{2/N}-b^{2/N}}.$$

The second term is dominant when $N>2$, and its expansion is uniform for $a\in[m,M]$. Multiplication by Nb^p gives (14) and (13). □

Proposition 3.6 (Exterior-volume representation and logarithmic bounds). *Order the singular values as $\sigma_1 \geq \cdots \geq \sigma_d > 0$. Then*

$$S_N(W) = c_N\sum_{i=1}^{d-1}(d-i)\log\sigma_i + R_N(W), \tag{15}$$

where

$$0 \leq R_N(W) \leq \frac{1}{2}\binom{d}{2}\log N. \tag{16}$$

Equivalently,

$$S_N(W) = c_N\sum_{k=1}^{d-1}\log\left\|\bigwedge^k W\right\|_{\mathrm{op}} + R_N(W). \tag{17}$$

Consequently,

$$S_N(W) \leq C_{N,d} + c_N\sum_{k=1}^{d-1}\log^+\left\|\bigwedge^k W\right\|_{\mathrm{op}}, \tag{18}$$

and

$$S_N(W) \leq C_{N,d} + \binom{d}{2}c_N\log(1+\sigma_1(W)). \tag{19}$$

Proof. For $i<j$, write $r_{ij} = \sigma_j/\sigma_i \in (0,1]$. By (10),

$$q_N(\sigma_i,\sigma_j) = c_N\log\sigma_i + h_N(r_{ij}),$$

where $0 \leq h_N(r) \leq \frac{1}{2}\log N$ by (11). Summing over $i<j$ gives (15), with

$$R_N(W) = \sum_{i<j} h_N(r_{ij}),$$

and proves (16).

The singular values of the induced map $\bigwedge^k W$ are all products $\sigma_{i_1}\cdots\sigma_{i_k}$, so

$$\left\|\bigwedge^k W\right\|_{\mathrm{op}} = \sigma_1\cdots\sigma_k.$$

Therefore

$$\sum_{k=1}^{d-1}\log\left\|\bigwedge^k W\right\|_{\mathrm{op}} = \sum_{k=1}^{d-1}\sum_{i=1}^{k}\log\sigma_i = \sum_{i=1}^{d-1}(d-i)\log\sigma_i,$$

which proves (17). Replacing each logarithm by its positive part yields (18). Finally,

$$\left\| \bigwedge^k W \right\|_{\mathrm{op}} \le \sigma_1(W)^k,$$

and $\sum_{k=1}^{d-1} k = \binom{d}{2}$, giving (19) after an adjustment of the constant for $\sigma_1 \le 1$. $\square$

4. When is the free energy bounded below?

The diagonal-completion loss is one example of a broader obstruction. In this section we separate three issues that are sometimes conflated:

(i) whether F_β is bounded from below;
(ii) whether $F_\beta(W) \to +\infty$ as $\|W\|_F \to \infty$;
(iii) whether the sublevel sets of F_β are compact in the full-rank state space $\mathrm{GL}_d(\mathbb{R})$.

The first property rules out arbitrarily negative values, the second gives confinement in norm, and the third is the natural condition for an equilibrium selection principle on the open manifold $\mathrm{GL}_d(\mathbb{R})$. Boundedness from below alone does not prevent divergent gradient trajectories: the function $f(x) = e^{-x}$ is bounded below on $\mathbb{R}$, but its negative-gradient flow satisfies $\dot{x} = e^{-x}$ and has $x(t) \to +\infty$.

4.1. The logarithmic entropy scale.

The exterior-volume formula in Proposition 3.6 shows that the fiber entropy grows at most logarithmically. It is convenient to write

$$\Psi(W) = \sum_{k=1}^{d-1} \log^+ \left\| \bigwedge^k W \right\|_{\mathrm{op}}, \qquad \log^+ x = \max\{\log x, 0\}. \tag{20}$$

Then Proposition 3.6 gives

$$S_N(W) \le C_{N,d} + c_N \Psi(W), \tag{21}$$

and, more crudely,

$$S_N(W) \le C_{N,d} + \alpha_{N,d} \log(1 + \|W\|_F), \qquad \alpha_{N,d} = c_N \binom{d}{2}. \tag{22}$$

The coefficient in the norm estimate is sharp at the level of a uniform radial bound: along $W = tI_d$,

$$S_N(tI_d) = \alpha_{N,d} \log t + C_{N,d}.$$

Thus polynomial growth of the energy is far more than is needed to dominate the entropy. The natural threshold is logarithmic.

4.2. An exact criterion from energy sublevel sets.

Let $E : \mathrm{GL}_d(\mathbb{R}) \to \mathbb{R}$ be bounded below. Adding a constant to E changes F_β by the same constant and has no effect on boundedness or on the gradient flow, so we normalize $E \ge 0$. Define the *entropy profile of the energy* by

$$\Theta_E(r) = \sup\{S_N(W) : W \in \mathrm{GL}_d(\mathbb{R}),\ E(W) \le r\}, \qquad r \ge 0,$$

with values in $\mathbb{R} \cup \{+\infty\}$.

Proposition 4.1 (Exact sublevel criterion). *For every $\beta > 0$,*

$$\sup_{W\in \mathrm{GL}_d(\mathbb{R})} \big(S_N(W) - \beta E(W)\big) = \sup_{r\geq 0}\big(\Theta_E(r) - \beta r\big) \tag{23}$$

in the extended real numbers. Consequently,

$$F_\beta \text{ is bounded below} \quad\Longleftrightarrow\quad \sup_{r\geq 0}\big(\Theta_E(r) - \beta r\big) < \infty. \tag{24}$$

Proof. For any $W \in \mathrm{GL}_d(\mathbb{R})$, setting $r = E(W)$ gives

$$S_N(W) - \beta E(W) \leq \Theta_E(r) - \beta r,$$

so the left-hand side of (23) is at most the right-hand side. Conversely, if $E(W) \leq r$, then

$$S_N(W) - \beta r \leq S_N(W) - \beta E(W).$$

Taking the supremum over such W gives

$$\Theta_E(r) - \beta r \leq \sup_{W\in \mathrm{GL}_d(\mathbb{R})} \big(S_N(W) - \beta E(W)\big).$$

Now take the supremum over r. Since

$$\inf F_\beta = -\frac{1}{\beta}\sup_W\big(S_N(W) - \beta E(W)\big),$$

(24) follows. ☐

Corollary 4.2. *The following statements hold.*

(i) *If $\Theta_E(r_0) = +\infty$ for some finite r_0, then F_β is unbounded below for every finite $\beta > 0$.*
(ii) *If $\Theta_E(r) \leq ar + C$ for all $r \geq 0$, then F_β is bounded below for every $\beta \geq a$.*
(iii) *If $\Theta_E(r) < \infty$ for finite r and $\Theta_E(r) = o(r)$ as $r \to \infty$, then F_β is bounded below for every $\beta > 0$.*

Proof. Each statement follows immediately from (24). For part (iii), choose R so that $\Theta_E(r) \leq \beta r/2$ for $r \geq R$ and use finiteness of Θ_E on the compact interval $[0, R]$. ☐

The criterion pinpoints the obstruction in diagonal completion: there is unlimited entropy already at zero loss. In the notation above,

$$\Theta_{E_d}(0) = +\infty.$$

4.3. **Affine flat directions.** The preceding failure is structural for any loss with a nontrivial affine direction on which the energy stays bounded.

Lemma 4.3 (Entropy diverges on every nontrivial affine ray). *Let $W_0 \in \mathrm{GL}_d(\mathbb{R})$ and $0 \neq H \in \mathbb{M}_d$. Then $W_0 + tH$ is invertible for all but finitely many $t \in \mathbb{R}$, and*

$$S_N(W_0 + tH) \longrightarrow +\infty \qquad \text{as } |t| \to \infty \tag{25}$$

through those values for which $W_0 + tH \in \mathrm{GL}_d(\mathbb{R})$.

Proof. The function $t \mapsto \det(W_0 + tH)$ is a polynomial that is nonzero at $t = 0$, so it is not identically zero and has only finitely many real roots.

For $1 \leq k \leq d-1$, the matrix of $\bigwedge^k(W_0 + tH)$ in the standard exterior basis consists of the $k \times k$ minors of $W_0 + tH$, each of which is a polynomial in t. Since W_0 is invertible, at least one

$k \times k$ minor is nonzero at $t = 0$. The corresponding polynomial is not identically zero and hence, for all sufficiently large $|t|$, its absolute value is bounded below by a positive constant. Therefore

$$\left\| \bigwedge^k (W_0 + tH) \right\|_{\mathrm{op}} \geq c_k > 0 \tag{26}$$

for large $|t|$.

For $k = 1$, choose an entry $H_{ij} \neq 0$. Then

$$|(W_0 + tH)_{ij}| \longrightarrow \infty,$$

so $\|W_0 + tH\|_{\mathrm{op}} \to \infty$. The exterior-volume representation (17) now has one term tending to $+\infty$, while all remaining terms are bounded below by (26); its remainder is uniformly bounded. This proves (25). $\square$

Corollary 4.4 (Affine-ray obstruction). *Suppose that for some $W_0 \in \mathrm{GL}_d(\mathbb{R})$ and $0 \neq H \in \mathbb{M}_d$,*

$$\sup_{|t| \geq t_0} E(W_0 + tH) < \infty.$$

Then F_β is unbounded below for every finite $\beta > 0$. In particular, this holds if E is constant on a nontrivial affine ray through a full-rank matrix.

Proof. Combine (25) with the assumed upper bound for the energy. $\square$

4.4. **A dichotomy for linear matrix sensing.** Let $\mathcal{L} : \mathbb{M}_d \to \mathbb{R}^m$ be linear, let $y \in \mathbb{R}^m$, and consider the squared sensing loss

$$E_{\mathcal{L},y}(W) = \frac{1}{2}\|\mathcal{L}(W) - y\|_2^2.$$

Theorem 4.5 (Linear-sensing dichotomy). *For every $d \geq 2$, $N \geq 2$, and $\beta > 0$, the free energy*

$$F_{\beta,\mathcal{L},y}(W) = E_{\mathcal{L},y}(W) - \frac{1}{\beta} S_N(W)$$

is bounded below on $\mathrm{GL}_d(\mathbb{R})$ if and only if $\mathcal{L}$ is injective. If $\mathcal{L}$ is injective, then

$$F_{\beta,\mathcal{L},y}(W) \longrightarrow +\infty \qquad \textit{as } \|W\|_F \to \infty.$$

Proof. Suppose first that $\mathcal{L}$ is not injective. Choose $0 \neq H \in \ker \mathcal{L}$ and any $W_0 \in \mathrm{GL}_d(\mathbb{R})$. Then

$$E_{\mathcal{L},y}(W_0 + tH) = E_{\mathcal{L},y}(W_0)$$

for every t. The conclusion follows from Corollary 4.4.

Now suppose that $\mathcal{L}$ is injective. By finite dimensionality, there is $s_{\mathcal{L}} > 0$ such that

$$\|\mathcal{L}(W)\|_2 \geq s_{\mathcal{L}} \|W\|_F \qquad \text{for all } W \in \mathbb{M}_d.$$

The elementary inequality $\|u - v\|_2^2 \geq \frac{1}{2}\|u\|_2^2 - \|v\|_2^2$ gives

$$E_{\mathcal{L},y}(W) \geq \frac{s_{\mathcal{L}}^2}{4} \|W\|_F^2 - \frac{1}{2}\|y\|_2^2.$$

Combining this with (22),

$$F_{\beta,\mathcal{L},y}(W) \geq \frac{s_{\mathcal{L}}^2}{4} \|W\|_F^2 - \frac{\alpha_{N,d}}{\beta} \log(1 + \|W\|_F) - C,$$

which tends to $+\infty$ with $\|W\|_F$. The same inequality gives a global lower bound. $\square$

Corollary 4.6 (Underdetermined matrix sensing)**.** *On the unrestricted matrix space, every genuinely underdetermined linear sensing loss has unbounded DLN free energy. In particular, if P_Ω observes a proper subset Ω of the d^2 entries, then*

$$\frac{1}{2}\|P_\Omega(W) - y\|_F^2 - \frac{1}{\beta}S_N(W)$$

is unbounded below for every finite $\beta > 0$.

Remark 4.7. If the admissible matrices are restricted to a linear subspace $V \subset \mathbb{M}_d$, injectivity in Theorem 4.5 should be understood for $\mathcal{L}|_V$. The affine-ray argument applies whenever $V \cap \mathrm{GL}_d(\mathbb{R})$ contains a point and $\ker(\mathcal{L}|_V)$ contains a nonzero direction through which one can remain in $V \cap \mathrm{GL}_d(\mathbb{R})$ for arbitrarily large parameter values.

4.5. **Practical sufficient conditions.** The entropy estimate gives conditions that apply beyond quadratic sensing losses.

Proposition 4.8 (Exterior-volume domination)**.** *Let $E : \mathrm{GL}_d(\mathbb{R}) \to \mathbb{R}$ be bounded below and let Ψ be defined by* (20)*.*

(i) *If*

$$E(W) \geq \frac{c_N}{\beta}\Psi(W) - C,$$

then F_β is bounded below.

(ii) *If there is $\delta > 0$ such that*

$$E(W) \geq \left(\frac{c_N}{\beta} + \delta\right)\Psi(W) - C, \tag{27}$$

then $F_\beta(W) \to +\infty$ as $\|W\|_F \to \infty$.

Proof. The first claim follows immediately from (21). Under (27),

$$F_\beta(W) \geq \delta\Psi(W) - C'.$$

Since the $k = 1$ summand in Ψ is $\log^+\|W\|_{\mathrm{op}}$, the right-hand side tends to $+\infty$ when $\|W\|_F \to \infty$. □

Corollary 4.9 (Norm-growth criteria)**.** *Let $\alpha_{N,d} = c_N\binom{d}{2}$.*

(i) *If*

$$E(W) \geq \frac{\alpha_{N,d}}{\beta}\log(1 + \|W\|_F) - C,$$

then F_β is bounded below.

(ii) *If for some $\delta > 0$,*

$$E(W) \geq \left(\frac{\alpha_{N,d}}{\beta} + \delta\right)\log(1 + \|W\|_F) - C,$$

then F_β is coercive in norm.

(iii) *If*

$$\frac{E(W)}{\log(1 + \|W\|_F)} \longrightarrow +\infty \qquad \text{as } \|W\|_F \to \infty, \tag{28}$$

then F_β is coercive in norm for every finite $\beta > 0$.

(iv) *In particular, any estimate*

$$E(W) \geq a\,\|W\|_F^p - C, \qquad a > 0, \quad p > 0,$$

implies norm coercivity of F_β for every finite $\beta > 0$.

Proof. Use (22). Parts (iii) and (iv) imply the strict logarithmic domination in part (ii) outside a sufficiently large ball. □

The adjective "superlogarithmic" in (28) is essential. Ordinary coercivity of E is not enough: a coercive energy that grows more slowly than $\log \|W\|_F$ can still be dominated by S_N along scalar rays.

A sublevel-set version is sometimes more convenient. Define

$$R_E(r) = \sup\{\|W\|_F : W \in \mathrm{GL}_d(\mathbb{R}),\ E(W) \leq r\} \in [0, +\infty].$$

Whenever $R_E(r) < \infty$, (22) implies

$$\Theta_E(r) \leq C_{N,d} + \alpha_{N,d} \log(1 + R_E(r)).$$

Thus the condition

$$\sup_{r \geq 0} \left[\alpha_{N,d} \log(1 + R_E(r)) - \beta r\right] < \infty$$

is sufficient for boundedness below. In particular,

$$\log(1 + R_E(r)) = o(r)$$

is sufficient for every $\beta > 0$.

4.6. **Boundedness, coercivity, and attainment.** Even norm coercivity does not automatically yield a minimizer on $\mathrm{GL}_d(\mathbb{R})$, because $\mathrm{GL}_d(\mathbb{R})$ is open and a minimizing sequence may approach the rank boundary. The boundary behavior of the entropy explains exactly where an additional hypothesis is needed.

Proposition 4.10 (Entropy at the finite rank boundary)**.** *Let $W_m \in \mathrm{GL}_d(\mathbb{R})$ be bounded and suppose $W_m \to W_*$.*

(i) *If $\operatorname{rank} W_* \leq d - 2$, then $S_N(W_m) \to -\infty$.*

(ii) *If $\operatorname{rank} W_* = d - 1$, then $S_N(W_m)$ has a finite limit determined by the $d - 1$ positive limiting singular values. More precisely, if these are $\tau_1, \ldots, \tau_{d-1} > 0$, then*

$$\lim_{m \to \infty} S_N(W_m) = \sum_{1 \leq i < j \leq d-1} q_N(\tau_i, \tau_j) + c_N \sum_{i=1}^{d-1} \log \tau_i. \tag{29}$$

Proof. If at least two singular values tend to zero, the pair term involving those two singular values tends to $-\infty$ by the homogeneity relation (9); all pair terms are bounded above on a bounded set by Proposition 3.6. This proves part (i). If only $\sigma_d \to 0$, then for each $i < d$, (10) and $h_N(0) = 0$ give

$$q_N(\sigma_i, \sigma_d) \longrightarrow c_N \log \tau_i,$$

while all pairs among the first $d - 1$ singular values converge continuously. Summing proves (29). □

Corollary 4.11 (A sufficient condition for attainment). *Assume that $E : \mathrm{GL}_d(\mathbb{R}) \to \mathbb{R}$ is continuous and bounded below, that F_β is coercive in norm, and that*

$$E(W_m) \longrightarrow +\infty \tag{30}$$

for every bounded sequence $W_m \in \mathrm{GL}_d(\mathbb{R})$ converging to a matrix of rank $d-1$. Then every sublevel set of F_β is compact in $\mathrm{GL}_d(\mathbb{R})$, and F_β attains a global minimum.

Proof. Norm coercivity bounds every sublevel set in $\mathbb{M}_d$. A sequence in such a set cannot approach rank at most $d-2$, because Proposition 4.10 and the lower bound on E would force $F_\beta \to +\infty$. It cannot approach rank $d-1$ because of (30) and the finite entropy limit in (29). Hence the sublevel set is closed in a bounded subset of $\mathbb{M}_d$ and stays a positive distance from the rank boundary. It is compact in $\mathrm{GL}_d(\mathbb{R})$, and the direct method gives a minimizer. $\square$

For example, ridge-regularized completion,

$$E_{\Omega,\lambda}(W) = \frac{1}{2}\|P_\Omega(W) - Y\|_F^2 + \frac{\lambda}{2}\|W\|_F^2, \qquad \lambda > 0,$$

is norm-coercive and therefore makes F_β bounded below. The ridge term does not create a corank-one barrier, however, so additional treatment of lower-rank strata or a barrier such as $-\mu \log\det(W^T W)$ is still needed if one insists on a minimizer inside $\mathrm{GL}_d(\mathbb{R})$.

4.7. **When entropy does define a selection principle.**

The usual low-temperature selection statement becomes valid once compactness and attainment are supplied.

Proposition 4.12 (Low-temperature entropy selection). *Let $\mathcal{X}$ be a closed subset of a finite-dimensional Euclidean space, and let $E, S : \mathcal{X} \to \mathbb{R}$ be continuous. Put*

$$m = \min_{\mathcal{X}} E, \qquad \mathcal{M} = \operatorname*{argmin}_{\mathcal{X}} E,$$

and assume $\mathcal{M} \neq \varnothing$. Suppose that for every sufficiently large β, the function

$$F_\beta = E - \beta^{-1} S$$

has a minimizer W_β, and that the family $\{W_\beta\}$ is precompact. Then every limit point W_* of W_β as $\beta \to \infty$ satisfies

$$W_* \in \operatorname*{argmax}_{W \in \mathcal{M}} S(W). \tag{31}$$

If the entropy has a unique maximizer on $\mathcal{M}$, then W_β converges to it.

Proof. Fix $V \in \mathcal{M}$. Minimality of W_β gives

$$E(W_\beta) - \frac{1}{\beta} S(W_\beta) \le m - \frac{1}{\beta} S(V),$$

so

$$0 \le E(W_\beta) - m \le \frac{S(W_\beta) - S(V)}{\beta}. \tag{32}$$

Precompactness and continuity make $S(W_\beta)$ bounded along every sequence $\beta \to \infty$, so (32) implies $E(W_\beta) \to m$. Every limit point therefore belongs to $\mathcal{M}$. The same inequality also gives

$$S(W_\beta) \ge S(V) \qquad \text{for every } V \in \mathcal{M}.$$

Passing to a convergent subsequence yields $S(W_*) \ge S(V)$ for every $V \in \mathcal{M}$, proving (31). Uniqueness of the entropy maximizer then forces the whole precompact family to converge. $\square$

Thus a natural setting for entropic selection is a compact energy-minimizing set on which the entropy is continuous and has a distinguished maximum. Diagonal completion violates this picture in the strongest possible way: its minimizing set is noncompact and its entropy is unbounded above there. Instead of choosing a finite minimizer, the entropic term drives the system toward larger and larger unobserved directions.

4.8. **Applied energies: how to read the examples.** The preceding results separate three questions that should be kept distinct in applications. First, does the energy E attain its infimum on the natural closed parameter space? Second, is

$$F_\beta(W) = E(W) - \frac{1}{\beta} S_N(W)$$

bounded below, or even coercive in norm? Third, if the state space is the open manifold $\mathrm{GL}_d(\mathbb{R})$, is the infimum of F_β attained away from the rank boundary? An energy may have many minimizers while F_β is unbounded below, as happens for underdetermined least squares. Conversely, F_β may be norm-coercive even when the minimizer of E lies on the excluded rank boundary. The examples below record the energy, its elementary variational properties, and the conclusion supplied by the criteria in this section. No proofs are repeated.

The universal estimate

$$S_N(W) \le C_{N,d} + \alpha_{N,d} \log(1 + \|W\|_F), \qquad \alpha_{N,d} = \left(1 - \frac{1}{N}\right)\binom{d}{2},$$

should be kept in mind throughout. Positive polynomial growth of E dominates the entropy at infinity, whereas any nontrivial affine direction on which E stays bounded forces F_β to be unbounded below. Logarithmically growing energies lie at the threshold and must be tested direction by direction. Finally, norm coercivity does not by itself settle attainment in $\mathrm{GL}_d(\mathbb{R})$, because the entropy has a finite limit at a generic corank-one boundary point; see Proposition 4.10 and Corollary 4.11.

4.9. **Quadratic losses, inverse problems, and low-rank regularization.**

4.9.1. *Full quadratic loss and Gaussian matrix denoising.* The basic full-observation energy is

$$E_A(W) = \frac{1}{2\sigma^2} \|W - A\|_F^2, \qquad A \in \mathbb{M}_d,$$

where σ^2 may be interpreted as a noise variance. This is the negative log-likelihood, up to an additive constant, for observing A in isotropic Gaussian matrix noise. Such losses also underlie singular-value shrinkage and matrix denoising [GD17].

Properties of the energy. On the closed space $\mathbb{M}_d$, the unique minimizer is A, and the energy has quadratic growth. If $A \in \mathrm{GL}_d(\mathbb{R})$, the energy minimum is attained in the full-rank state space. If A is singular, the infimum of $E_A|_{\mathrm{GL}_d(\mathbb{R})}$ equals zero but is not attained.

Conclusion for the free energy. By Corollary 4.9, F_β is bounded below and norm-coercive for every depth and every finite $\beta > 0$. If A is singular, or more generally if a minimizing sequence can approach corank one, norm coercivity alone does not decide whether the free-energy infimum is attained in $\mathrm{GL}_d(\mathbb{R})$. An explicit rank-boundary barrier gives attainment by Corollary 4.11.

4.9.2. *Injective squared sensing and identifiable least squares.* Let $\mathcal{L} : \mathbb{M}_d \to \mathbb{R}^m$ be linear and consider

$$E_{\mathcal{L},y}(W) = \frac{1}{2}\|\mathcal{L}(W) - y\|_2^2. \tag{33}$$

This form includes multivariate linear regression, linear system identification, and matrix sensing. Low-rank matrix-sensing theory often assumes a restricted isometry on low-rank matrices rather than injectivity on all of $\mathbb{M}_d$ [CP11].

Properties of the energy. If $\mathcal{L}$ is injective on the actual state space, then $\mathcal{L}^*\mathcal{L}$ is positive definite there. The least-squares energy has a unique minimizer and satisfies a quadratic lower bound

$$E_{\mathcal{L},y}(W) \geq c\,\|W\|_F^2 - C.$$

If the state space is a proper linear model $V \subset \mathbb{M}_d$, the relevant hypothesis is injectivity of $\mathcal{L}|_V$.

Conclusion for the free energy. The linear-sensing dichotomy, Theorem 4.5, shows that F_β is norm-coercive and bounded below for every N and β. As before, attainment inside $\mathrm{GL}_d(\mathbb{R})$ requires either a separate corank-one argument, a closed stratified state space, or a boundary barrier.

4.9.3. *Underdetermined least squares and matrix completion.* The same energy (33) behaves completely differently when $\ker \mathcal{L} \neq \{0\}$. Entrywise matrix completion is the special case

$$E_\Omega(W) = \frac{1}{2}\|P_\Omega(W - Y)\|_F^2,$$

where P_Ω retains the observed entries. Matrix completion and underdetermined matrix sensing are central models in low-rank recovery [CR09, CP11, RFP10].

Properties of the energy. A nonzero $H \in \ker \mathcal{L}$ gives

$$E_{\mathcal{L},y}(W_0 + tH) = E_{\mathcal{L},y}(W_0)$$

for every t. The ordinary energy may nevertheless attain its minimum, often on a noncompact affine set. For incomplete entry observations on unrestricted matrices, P_Ω is noninjective whenever at least one entry is unobserved.

Conclusion for the free energy. By Corollary 4.4 and Theorem 4.5, F_β is unbounded below for every finite $\beta > 0$. Thus the existence of exact or least-squares minimizers of E does not produce an entropic selection principle. The conclusion can change after restricting to a smaller model class on which the observation map is injective, or after introducing a confining penalty or proper prior.

4.9.4. *Ridge and Tikhonov regularization.* A standard confining modification is

$$E_\lambda(W) = \frac{1}{2}\|\mathcal{L}(W) - y\|_2^2 + \frac{\lambda}{2}\,\|W\|_F^2\,, \qquad \lambda > 0.$$

This is ridge or Tikhonov regularization, and in a Bayesian interpretation it corresponds to a centered Gaussian prior. Ridge regression was introduced in [HK70]; modern generalized-linear-model implementations are discussed in [FHT10].

Properties of the energy. The energy is strongly convex on $\mathbb{M}_d$, has a unique minimizer, and satisfies

$$E_\lambda(W) \geq \frac{\lambda}{2}\,\|W\|_F^2\,.$$

This remains true whether or not $\mathcal{L}$ is injective.

Conclusion for the free energy. The free energy is bounded below and norm-coercive for every N and β. Ridge regularization therefore repairs the escape-to-infinity defect of matrix completion. It does not itself diverge at rank loss, so attainment in $\mathrm{GL}_d(\mathbb{R})$ is still a separate boundary question. Moreover, the ridge term supplies its own explicit selection bias, so any resulting selection is not attributable to entropy alone.

4.9.5. *Nuclear-norm regularization.* A common convex surrogate for low rank is

$$E_{*,\lambda}(W) = \frac{1}{2}\|\mathcal{L}(W) - y\|_2^2 + \lambda\|W\|_*, \qquad \|W\|_* = \sum_{i=1}^{d} \sigma_i(W).$$

Nuclear-norm penalties are used throughout matrix completion, trace regression, and low-rank recovery [RFP10, KLT11].

Properties of the energy. The energy is convex and coercive on $\mathbb{M}_d$ for $\lambda > 0$, hence it attains a minimum. Its minimizers are frequently rank deficient, which is the intended regularizing effect. The nuclear norm has linear rather than quadratic growth.

Conclusion for the free energy. Since $\|W\|_* \geq \|W\|_F$, the free energy is norm-coercive for every N and β. The rank boundary remains important: both $\|W\|_*$ and S_N have finite limits at a generic corank-one point. If a zero singular value is opened by an amount s, the nuclear penalty changes at order s, while for $N > 2$ the leading entropy gain is of order $s^{2/N}$ and for $N = 2$ it is also linear. Thus depth two is a coefficient-sensitive borderline case, and a global attainment statement requires a separate rank-boundary analysis.

4.9.6. *Layerwise quadratic weight decay and factor-induced Schatten penalties.* For a depth-N factorization $W = W_N \cdots W_1$, the usual factor-space penalty is

$$\mathcal{R}(W_1, \ldots, W_N) = \frac{\lambda}{2}\sum_{j=1}^{N} \|W_j\|_F^2 .$$

When the intermediate widths are large enough, minimizing this penalty over factorizations with fixed product gives

$$\inf_{W_N \cdots W_1 = W} \mathcal{R}(W_1, \ldots, W_N) = \frac{\lambda N}{2}\sum_{i=1}^{d} \sigma_i(W)^{2/N}.$$

The variational relation between factor penalties and Schatten-type penalties is used throughout deep matrix factorization and low-rank recovery [GWB$^+$17, ACHL19, GVRH20]; for two interacting matrices, the connection between weight decay, nuclear-norm regularization, and low-rank attention products is studied in [KAvO24].

Properties of the energy. The induced end-to-end penalty has positive-power growth at infinity, so it is confining. For $N > 2$ it is nonconvex and non-Lipschitz at a zero singular value. Its small-singular-value scale is exactly $s^{2/N}$.

Conclusion for the free energy. The positive-power growth dominates the logarithmic entropy at infinity, so the effective free energy is norm-coercive. Near a corank-one boundary point, however, the induced penalty and the entropy change on the same $s^{2/N}$ scale. Consequently, whether a minimizer remains on the boundary or moves into $\mathrm{GL}_d(\mathbb{R})$ can depend on λ, β, the nonzero singular values, and the depth. Factor-space weight decay therefore guarantees confinement but does not automatically settle full-rank attainment.

4.10. **Classification and margin-based energies.**

4.10.1. *Separable logistic and exponential losses.* For matrix-valued features $X_1, \dots, X_M$ and labels $y_j \in \{-1, 1\}$, consider

$$E_{\log}(W) = \sum_{j=1}^{M} \log\left(1 + \exp\bigl(-y_j \langle W, X_j \rangle_F\bigr)\right), \tag{34}$$
$$E_{\exp}(W) = \sum_{j=1}^{M} \exp\bigl(-y_j \langle W, X_j \rangle_F\bigr).$$

The existence of finite logistic maximum-likelihood estimators is governed by separation and overlap conditions [AA84]; the diverging-gradient-descent regime on separable data is analyzed in [SHN+18].

Properties of the energy. If there is a separating matrix H satisfying

$$y_j \langle H, X_j \rangle_F > 0 \qquad \text{for every } j,$$

then $E_{\log}(W_0 + tH) \to 0$ and $E_{\exp}(W_0 + tH) \to 0$ as $t \to \infty$. Their infimum is generally not attained at a finite parameter.

Conclusion for the free energy. The energy stays bounded along a nontrivial affine ray, so Corollary 4.4 gives

$$\inf_{\mathrm{GL}_d(\mathbb{R})} F_\beta = -\infty$$

for every finite β. The entropy does not cure the classical optimum-at-infinity phenomenon; it reinforces the failure of confinement.

4.10.2. *Identifiable nonseparable logistic loss.* The energy is again $E_{\log}$ in (34), but the data are assumed to satisfy the usual no-separation and no-quasi-separation conditions, and the feature map is identifiable.

Properties of the energy. Under these hypotheses, logistic regression has a finite maximum-likelihood estimator [AA84]. If the design spans the parameter space and there is positive recession in every nonzero direction, compactness of the unit sphere gives a linear lower bound

$$E_{\log}(W) \geq c \,\|W\|_F - C.$$

Gauge symmetries or unobserved directions must first be removed; otherwise the energy is constant on affine fibers despite the absence of statistical separation.

Conclusion for the free energy. Linear growth is superlogarithmic, so F_β is bounded below and norm-coercive for every finite β. If the parametrization retains a nontrivial gauge direction, the affine-ray obstruction applies instead. Thus identifiability is as important here as nonseparability.

4.10.3. *Multiclass softmax loss and gauge invariance.* For class-weight rows $w_1, \dots, w_C$ and examples (x_j, y_j), the multiclass cross-entropy is

$$E_{\text{soft}}(W) = -\sum_{j=1}^{M} \log \frac{\exp(w_{y_j}^T x_j)}{\sum_{c=1}^{C} \exp(w_c^T x_j)}.$$

Multinomial logistic models and their regularization are discussed in [FHT10].

Properties of the energy. For every a, adding the same score $a^T x$ to each class leaves all softmax probabilities unchanged. In matrix notation,

$$E_{\text{soft}}(W + \mathbf{1}a^T) = E_{\text{soft}}(W).$$

This is an exact affine gauge symmetry, independent of whether the data are separable.

Conclusion for the free energy. On an unrestricted noncompact parametrization, the gauge direction forces F_β to be unbounded below. One must fix a reference class, impose a sum-to-zero gauge, quotient by the symmetry, or add a proper confining prior before applying an entropic selection argument.

4.10.4. *Separable unregularized hinge loss.* The empirical hinge energy is

$$E_{\text{hinge}}(W) = \sum_{j=1}^{M} \max\{0, 1 - y_j \langle W, X_j \rangle_F\}.$$

It is the basic loss in support-vector classification [CV95].

Properties of the energy. For separable data, E_{hinge} attains its minimum value zero, but it does so on an unbounded cone: once all margins exceed one, scaling the separator preserves zero loss.

Conclusion for the free energy. The free energy is unbounded below. This is a useful contrast with separable logistic regression: hinge loss has genuine finite minimizers, but the minimizing set is noncompact and supports unbounded entropy. Existence of minimizers for E is therefore not enough for entropic selection.

4.10.5. *Quadratically regularized hinge loss.* The soft-margin support-vector energy is

$$E_{\text{SVM},\lambda}(W) = E_{\text{hinge}}(W) + \frac{\lambda}{2} \|W\|_F^2, \qquad \lambda > 0.$$

Quadratic margin regularization is part of the classical support-vector formulation [CV95].

Properties of the energy. The energy is strongly coercive and, on the ambient Euclidean space, has a unique minimizer whenever the quadratic term acts on every parameter direction. It is nonsmooth only on the margin hyperplanes.

Conclusion for the free energy. Quadratic growth makes F_β norm-coercive for every N and β. If the minimization is required to occur in $\mathrm{GL}_d(\mathbb{R})$, nonsmooth behavior at the rank boundary still deserves separate treatment, but entropy cannot cause escape to infinity.

4.11. **Robust-statistics energies.**

4.11.1. *Huber regression.* Let $r_j(W)$ be affine residuals and define

$$E_{\text{H}}(W) = \sum_{j=1}^{M} \rho_\delta(r_j(W)), \qquad \rho_\delta(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le \delta, \\ \delta|u| - \frac{1}{2}\delta^2, & |u| > \delta. \end{cases}$$

Huber's loss is a foundational robust-estimation criterion [Hub64].

Properties of the energy. The loss is quadratic near zero and linear in the tails. If the residual operator is injective on the state space, then

$$E_{\text{H}}(W) \ge c\,\|W\|_F - C,$$

so E_{H} is coercive and attains a minimum. If the residual operator has a kernel, the energy is constant on the corresponding affine directions.

Conclusion for the free energy. In the identifiable case, linear growth dominates the logarithmic entropy and F_β is norm-coercive. In the nonidentifiable case, the affine-ray obstruction makes F_β unbounded below. Thus replacing squared loss by Huber loss changes the growth from quadratic to linear but does not change the decisive role of identifiability.

4.11.2. *Cauchy and Student-t negative log-likelihoods.* A representative heavy-tailed energy is

$$E_{\nu,s}(W) = \frac{\nu+1}{2}\sum_{j=1}^{M} \log\left(1 + \frac{r_j(W)^2}{\nu s^2}\right),$$

with the Cauchy model obtained at $\nu = 1$. Student-t likelihoods are widely used for robust statistical modeling [LLT89].

Properties of the energy. The loss is generally nonconvex in nonlinear models and has logarithmic tails. Along a ray $W_0 + tH$, each residual that grows linearly contributes a multiple of $\log|t|$, while residuals annihilated by H contribute only a bounded amount. If every residual is unchanged in some nonzero direction, the energy is exactly flat there.

Conclusion for the free energy. This is the borderline regime. Both $E_{\nu,s}$ and S_N may grow like $\log|t|$, so boundedness depends on the directional coefficients, not merely on whether $E_{\nu,s} \to \infty$. The exact test is the entropy-profile criterion in Proposition 4.1. A kernel direction is still fatal, while sufficiently large logarithmic coefficients can make F_β bounded below at a given temperature.

4.11.3. *Bounded redescending losses.* Tukey's biweight is a standard example:

$$\rho_c(u) = \begin{cases} \dfrac{c^2}{6}\left[1 - (1-(u/c)^2)^3\right], & |u| \le c, \\ \dfrac{c^2}{6}, & |u| > c, \end{cases} \qquad E_{\mathrm{T}}(W) = \sum_{j=1}^{M} \rho_c(r_j(W)).$$

Redescending robust procedures originate in work such as [BT74].

Properties of the energy. For a finite dataset, E_{T} is bounded above on the entire parameter space. It may possess useful finite local or global minimizers, but it supplies no confinement at infinity.

Conclusion for the free energy. Along every nontrivial affine ray through a full-rank matrix, the energy stays bounded while the fiber entropy diverges. Hence F_β is unbounded below for every finite β. A bounded redescending loss must be accompanied by an explicit confining penalty or a proper prior if one wants a well-posed free-energy minimization problem.

4.12. **Log-determinant energies on the positive-definite cone.**

4.12.1. *Gaussian covariance and precision estimation.* For a positive-definite precision matrix Θ, the Gaussian negative log-likelihood and its graphical-lasso modification are

$$E_{\mathrm{glasso}}(\Theta) = \mathrm{Tr}(\widehat{\Sigma}\Theta) - \log\det\Theta + \lambda\|\Theta\|_{1,\mathrm{off}}, \qquad \Theta \succ 0.$$

This energy is used to estimate sparse conditional-dependence graphs [FHT08].

Properties of the energy. If $\widehat{\Sigma} \succ 0$, then $\mathrm{Tr}(\widehat{\Sigma}\Theta)$ grows at least linearly as $\|\Theta\| \to \infty$, while $-\log\det\Theta \to +\infty$ as Θ approaches the singular boundary. Thus the energy has compact sublevel sets in the positive-definite cone and attains a positive-definite minimum. If $\widehat{\Sigma}$ is singular, coercivity can fail in null directions and must be restored by the penalty or an additional hypothesis.

Conclusion for the free energy. When $\widehat{\Sigma} \succ 0$, the linear confinement dominates the logarithmic fiber entropy at infinity, and the log-determinant term supplies precisely the rank-boundary barrier required by Corollary 4.11. The free energy therefore has compact sublevel sets and an attained full-rank minimum. This is a model example in which both possible escape mechanisms are excluded.

4.13. **A practical taxonomy.** The examples fall into four broad classes.

4.13.1. *Superlogarithmically confining energies.* Full least squares, injective matrix sensing, ridge regularization, identifiable Huber regression, quadratically regularized margin losses, and positive-definite covariance likelihoods grow linearly or quadratically in every unbounded direction. They satisfy

$$\frac{E(W)}{\log(1+\|W\|_F)} \longrightarrow +\infty,$$

so F_β is norm-coercive for every finite temperature. Actual attainment in $\mathrm{GL}_d(\mathbb{R})$ may still require a rank-boundary argument unless the energy already contains a barrier.

4.13.2. *Underdetermined or gauge-invariant energies.* Matrix completion, noninjective matrix sensing, rank-deficient regression, and unfixed softmax parametrizations have nontrivial affine directions on which the loss is constant. Their ordinary energies may have excellent finite minimizers, but Corollary 4.4 forces the corresponding free energy to be unbounded below. Restricting the model, quotienting the symmetry, or adding a confining prior is essential.

4.13.3. *Losses whose optima occur at infinity or on unbounded minimizing sets.* Separable logistic and exponential losses approach their infimum along separating rays, while separable hinge loss vanishes on an unbounded cone. The former may have no finite minimizer and the latter may have many, but in either case the energy remains bounded along an escape direction and the entropy sends F_β to $-\infty$.

4.13.4. *Borderline logarithmic and rank-sensitive energies.* Student-t and Cauchy negative log-likelihoods can grow on the same logarithmic scale as S_N. Nuclear-norm penalties, factor-induced Schatten penalties, and other low-rank regularizers are confining at infinity but can compete with the entropy near the rank boundary. Here boundedness and attainment depend on temperature, depth, coefficients, singular-value geometry, and the observation map. The entropy profile Θ_E from Proposition 4.1 is the appropriate global test.

The practical rule is

$$\begin{aligned}
\text{flat or bounded-energy escape directions} &\implies F_\beta \text{ is unbounded below},\\
\text{superlogarithmic growth of } E &\implies F_\beta \text{ is norm-coercive},\\
\text{logarithmic growth of } E &\implies \text{compare directional coefficients},\\
\text{a minimum in } \mathrm{GL}_d(\mathbb{R}) &\implies \text{also control the rank boundary}
\end{aligned}$$

This taxonomy separates two failures of deterministic entropic selection. The energy may fail to confine the system at infinity, or it may be norm-coercive while still allowing minimizing sequences to approach corank one. Entropic selection is well posed only after these global defects have been excluded.

5. The free energy is unbounded below

The diagonal observation operator is noninjective, so unboundedness already follows abstractly from Theorem 4.5. We nevertheless give the explicit zero-loss shear because it identifies the precise logarithmic escape rate of the entropy and will be useful in interpreting the dynamics.

The noncompact zero-loss set already prevents free-energy minimization.

Proposition 5.1 (Zero-loss shear)**.** *For $a \in \mathbb{R}$, let*

$$W_a = I_d + a e_1 e_2^T.$$

Then $E_d(W_a) = 0$, and the singular values of W_a are

$$s_a, \underbrace{1, \ldots, 1}_{d-2 \text{ times}}, s_a^{-1},$$

where

$$s_a = \frac{\sqrt{a^2+4}+|a|}{2}. \tag{35}$$

In particular, $s_a \sim |a|$ as $|a| \to \infty$.

Proof. Only the upper-left 2×2 block differs from the identity. That block is

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}.$$

The product of its singular values is the absolute determinant, equal to 1, while the sum of their squares is its squared Frobenius norm, equal to a^2+2. Solving these two equations gives s_a and s_a^{-1} with (35). All remaining singular values are 1. ☐

Theorem 5.2 (Failure of free-energy minimization)**.** *For every $d \geq 2$, every integer $N \geq 2$, and every finite $\beta > 0$,*

$$\inf_{W \in \mathrm{GL}_d(\mathbb{R})} F_\beta(W) = -\infty.$$

More precisely,

$$S_N(W_a) = (d-1)c_N \log s_a + O_{N,d}(1) \qquad \text{as } |a| \to \infty. \tag{36}$$

Proof. There are $d-2$ pairs $(s_a, 1)$ and one pair (s_a, s_a^{-1}). By (10), each contributes $c_N \log s_a + O(1)$. The $d-2$ pairs $(1, s_a^{-1})$ and all pairs of unit singular values remain bounded. This proves (36). Since $E_d(W_a) = 0$,

$$F_\beta(W_a) = -\beta^{-1} S_N(W_a) \longrightarrow -\infty.$$

☐

Remark 5.3. The argument is not tied to the target diagonal $(1, \ldots, 1)$. If every prescribed diagonal entry is nonzero, adding a large off-diagonal shear preserves the observed entries and invertibility while producing the same logarithmic entropy growth.

6. Classification of all full-rank critical points

Because $\mathcal{A}_{N,W}$ is invertible on $\mathrm{GL}_d(\mathbb{R})$, equilibria of (5) are exactly Euclidean critical points of F_β.

6.1. **The critical-point equation.** At a critical point,

$$E_d'(W) = \frac{1}{\beta} S_N'(W). \tag{37}$$

The left-hand side is diagonal. The right-hand side is spectral. Positivity and strict concavity force these structures to align.

Lemma 6.1 (Shared polar factor)**.** *Let $W = U\Sigma V^T \in \mathrm{GL}_d(\mathbb{R})$, and let $g_i = \partial_i \mathcal{S}_N(\sigma)$. Then*

$$S_N'(W) = U \operatorname{diag}(g_1, \ldots, g_d) V^T.$$

Moreover, $S_N'(W)$ is invertible and has the same orthogonal polar factor UV^T as W.

Proof. The gradient formula is (7). By Corollary 3.3, every g_i is positive. Thus the displayed decomposition is an SVD of $S_N'(W)$ with the same left and right singular frames as W, and its orthogonal polar factor is UV^T. □

Theorem 6.2 (Critical matrices are diagonal)**.** *Assume $N > 2$. Every critical point of F_β in $\mathrm{GL}_d(\mathbb{R})$ is diagonal.*

Proof. Let

$$P = E_d'(W) = \operatorname{diag}(W_{11} - 1, \ldots, W_{dd} - 1).$$

At a critical point, (37) and Lemma 6.1 show that P is invertible. Write the right polar decomposition of W as

$$W = QH, \qquad Q = UV^T, \qquad H = V\Sigma V^T > 0.$$

The entropy gradient has the corresponding polar decomposition

$$S_N'(W) = QG, \qquad G = V \operatorname{diag}(g_1, \ldots, g_d) V^T > 0.$$

Equation (37) gives

$$P = \frac{1}{\beta} QG.$$

The polar decomposition of an invertible matrix is unique. Since P is diagonal, its orthogonal polar factor is the diagonal sign matrix

$$D_\epsilon = \operatorname{diag}(\operatorname{sgn} P_{11}, \ldots, \operatorname{sgn} P_{dd}).$$

Therefore $Q = D_\epsilon$, and

$$G = \beta D_\epsilon P$$

is diagonal and positive.

The positive matrices G and H commute because they are simultaneously diagonalized by V. Partition the standard coordinate space into eigenspaces of the diagonal matrix G. The commutation relation makes H block diagonal with respect to this partition. Within one such block, all eigenvalues g_i of G agree. By Lemma 3.4, the corresponding singular values σ_i agree. Hence every eigenvalue of the restriction of H to the block is the same positive number, so that restriction is a scalar multiple of the identity. It follows that H is diagonal in the standard basis. Since $Q = D_\epsilon$ is also diagonal, $W = QH$ is diagonal. □

6.2. **One critical point in each sign chamber.** Write a diagonal invertible matrix as

$$W = \operatorname{diag}(\epsilon_1 x_1, \ldots, \epsilon_d x_d), \qquad x_i > 0, \quad \epsilon_i \in \{-1, 1\}.$$

On this chamber,

$$S'_N(W) = \operatorname{diag}(\epsilon_1 g_1(x), \ldots, \epsilon_d g_d(x)),$$

so (37) is equivalent to

$$x_i - \epsilon_i = \frac{1}{\beta} g_i(x), \qquad i = 1, \ldots, d. \tag{38}$$

These are the Euler equations for

$$\Phi_\epsilon(x) = \frac{1}{2}\sum_{i=1}^{d}(x_i - \epsilon_i)^2 - \frac{1}{\beta}\mathcal{S}_N(x), \qquad x \in (0, \infty)^d. \tag{39}$$

Lemma 6.3 (Boundary behavior of Φ_ϵ)**.** *Assume $N > 2$.*

(i) *$\Phi_\epsilon(x) \to +\infty$ as $\|x\| \to \infty$.*

(ii) *If at least two coordinates tend to zero while x remains bounded, then $\Phi_\epsilon(x) \to +\infty$.*

(iii) *At a boundary point with exactly one zero coordinate and all others positive, the inward one-sided derivative of Φ_ϵ in that coordinate is $-\infty$.*

Proof. The logarithmic upper bound in Proposition 3.6, applied to a diagonal matrix, shows that the quadratic term in (39) dominates at infinity.

If $x_i, x_j \to 0$, use (10) with $a = \max\{x_i, x_j\}$ to obtain

$$q_N(x_i, x_j) = c_N \log a + O(1) \longrightarrow -\infty.$$

All other pair terms are bounded above on a bounded set. Thus $\mathcal{S}_N(x) \to -\infty$ and $\Phi_\epsilon(x) \to +\infty$.

Finally, suppose $x_i \downarrow 0$ while x_j remains in a compact subset of $(0, \infty)$ for $j \neq i$. By Lemma 3.5,

$$g_i(x) \longrightarrow +\infty.$$

Hence

$$\partial_i \Phi_\epsilon(x) = x_i - \epsilon_i - \frac{1}{\beta} g_i(x) \longrightarrow -\infty.$$

□

Theorem 6.4 (Exactly 2^d critical points)**.** *Assume $N > 2$. For every $\epsilon \in \{-1, 1\}^d$, the function Φ_ϵ has a unique critical point $x^\epsilon \in (0, \infty)^d$. Consequently,*

$$\#\operatorname{Crit}_{\mathrm{GL}_d(\mathbb{R})}(F_\beta) = 2^d,$$

and the critical points are exactly

$$W_\epsilon = \operatorname{diag}(\epsilon_1 x_1^\epsilon, \ldots, \epsilon_d x_d^\epsilon).$$

Proof. By Theorem 2.2,

$$\nabla^2 \Phi_\epsilon(x) = I - \frac{1}{\beta}\nabla^2 \mathcal{S}_N(x) \succ 0.$$

Thus Φ_ϵ is strongly convex and has at most one critical point.

To prove existence, extend Φ_ϵ to the closed positive orthant by taking the finite limit at boundary points with exactly one zero coordinate and setting the value equal to $+\infty$ when at least two coordinates vanish. By Lemma 6.3, this lower-semicontinuous extension is coercive, so it attains a minimum. A boundary point with at least two zeros has infinite value. A boundary point with exactly one zero cannot minimize because moving slightly inward lowers Φ_ϵ. Hence the minimizer lies in $(0,\infty)^d$ and is the unique critical point.

The diagonality theorem now shows that the resulting 2^d points exhaust the full critical set. $\square$

Remark 6.5 (Permutation types). If a sign vector has k positive entries, uniqueness and permutation symmetry imply that all positive coordinates of x^ϵ have one common magnitude and all negative coordinates have another. Hence there are only $d+1$ critical-point types up to coordinate permutation, although there are 2^d distinct points.

Proposition 6.6 (The two scalar equilibria)**.** *Let*

$$C_{d,N,\beta} = \frac{(d-1)c_N}{2\beta}.$$

The all-positive and all-negative critical points are

$$W_+ = w_+ I_d, \qquad W_- = -w_- I_d,$$

where

$$w_+(w_+-1) = C_{d,N,\beta}, \qquad w_-(w_-+1) = C_{d,N,\beta}.$$

Thus

$$w_+ = \frac{1+\sqrt{1+4C_{d,N,\beta}}}{2}, \qquad w_- = \frac{-1+\sqrt{1+4C_{d,N,\beta}}}{2}.$$

Proof. At $x = w\mathbf{1}$, symmetry and the homogeneity

$$\mathcal{S}_N(t\sigma) = \mathcal{S}_N(\sigma) + \binom{d}{2} c_N \log t$$

give

$$g_i(w\mathbf{1}) = \frac{(d-1)c_N}{2w}.$$

Substitute this into (38) for the all-positive and all-negative sign vectors. $\square$

Proposition 6.7 (Zero-temperature location of the saddles)**.** *For a fixed sign vector ϵ, as $\beta\to\infty$,*

$$x_i^\epsilon(\beta) \longrightarrow \begin{cases} 1, & \epsilon_i = 1, \\ 0, & \epsilon_i = -1. \end{cases}$$

Equivalently,

$$W_\epsilon(\beta) \longrightarrow \operatorname{diag}\left(\frac{1+\epsilon_1}{2}, \dots, \frac{1+\epsilon_d}{2}\right).$$

Proof. Let m be the number of negative entries of ϵ, and let x^0 denote the boundary vector with coordinates 1 at positive signs and 0 at negative signs. The quadratic part of (39) has the unique minimum $m/2$ on the closed positive orthant, attained at x^0.

Choose a competitor $x^{(\beta)}$ with positive coordinates equal to 1 and negative coordinates equal to β^{-1}. Its quadratic energy tends to $m/2$, while Lemma 3.2 gives $|\mathcal{S}_N(x^{(\beta)})| = O(\log\beta)$. Hence

$$\limsup_{\beta\to\infty} \Phi_\epsilon(x^\epsilon(\beta)) \le \frac{m}{2}.$$

The logarithmic upper bound for $\mathcal{S}_N$ makes the family $x^\epsilon(\beta)$ bounded. On a fixed bounded set, $\mathcal{S}_N$ is bounded above, so

$$\Phi_\epsilon(x) \geq \frac{1}{2}\sum_i (x_i - \epsilon_i)^2 - O(\beta^{-1}).$$

It follows that the quadratic term at $x^\epsilon(\beta)$ tends to its minimum $m/2$. Strict convexity of that quadratic function on the closed orthant forces $x^\epsilon(\beta) \to x^0$. $\square$

7. The full Hessian and the Morse index

The restriction of F_β to a diagonal sign chamber is strongly convex, but each critical point becomes a saddle when off-diagonal perturbations are allowed. We first record the relevant second-order formula.

7.1. **A spectral Hessian formula.** Let $f : (0,\infty)^d \to \mathbb{R}$ be a symmetric C^2 function and define $\Phi(W) = f(\sigma(W))$. Fix

$$W_* = D_\epsilon X, \qquad D_\epsilon = \operatorname{diag}(\epsilon_1, \dots, \epsilon_d), \qquad X = \operatorname{diag}(x_1, \dots, x_d) > 0.$$

Put $g_i = \partial_i f(x)$.

Lemma 7.1 (Second variation at a signed diagonal matrix)**.** *The matrix space splits into the Frobenius-orthogonal sum*

$$\mathbb{M}_d = \mathcal{D} \oplus \bigoplus_{i<j} \mathcal{V}_{ij}, \qquad \mathcal{D} = \{\textit{diagonal matrices}\}, \quad \mathcal{V}_{ij} = \operatorname{span}\{E_{ij}, E_{ji}\}, \tag{40}$$

and the Hessian of Φ respects this splitting.

For $H = \operatorname{diag}(h_1, \dots, h_d)$,

$$D^2\Phi(W_*)[H,H] = (D_\epsilon h)^T \nabla^2 f(x)(D_\epsilon h). \tag{41}$$

For $H = aE_{ij} + bE_{ji}$,

$$D^2\Phi(W_*)[H,H] = \frac{A_{ij}}{2}(\epsilon_i a + \epsilon_j b)^2 + \frac{B_{ij}}{2}(\epsilon_i a - \epsilon_j b)^2, \tag{42}$$

where

$$A_{ij} = \frac{g_i - g_j}{x_i - x_j}, \qquad B_{ij} = \frac{g_i + g_j}{x_i + x_j}.$$

At $x_i = x_j$, A_{ij} is interpreted by its continuous limit.

Proof. The diagonal formula follows because, for sufficiently small t, the singular values of

$$W_* + tH = \operatorname{diag}(\epsilon_i x_i + t h_i)$$

are $x_i + t\epsilon_i h_i$.

For the off-diagonal formula, pass to the singular frame of W_*, where the perturbation has entries

$$k_{ij} = \epsilon_i a, \qquad k_{ji} = \epsilon_j b.$$

It is enough to work in the (i,j) block. Assume first that $x_i \neq x_j$. For

$$X_{ij}(t) = \begin{pmatrix} x_i & tk_{ij} \\ tk_{ji} & x_j \end{pmatrix},$$

the eigenvalues of $X_{ij}(t)^T X_{ij}(t)$ give the second derivatives of the two singular values. A direct 2×2 perturbation calculation yields

$$D^2\Phi(W_*)[H,H] = \frac{1}{2}\frac{g_i - g_j}{x_i - x_j}(k_{ij} + k_{ji})^2 + \frac{1}{2}\frac{g_i + g_j}{x_i + x_j}(k_{ij} - k_{ji})^2.$$

Substitution of k_{ij}, k_{ji} gives (42). Orthogonal invariance, or polarization of the same calculation, shows that distinct spaces in (40) do not couple. The repeated-value formula follows by continuity of the Hessian of a smooth symmetric singular-value function; see also [LS01, DK15]. ☐

Apply Lemma 7.1 to $f = \mathcal{S}_N$ at a critical point W_ϵ.

Lemma 7.2 (Signs of the off-diagonal entropy modes)**.** *Assume $N > 2$. At every positive x,*

$$A_{ij} < 0, \qquad B_{ij} > 0.$$

Proof. If $x_i \neq x_j$, the first inequality is Lemma 3.4. If $x_i = x_j$, it follows from negative definiteness of $\nabla^2 \mathcal{S}_N$ by taking the second derivative in the $e_i - e_j$ direction. The second inequality follows from $g_i, g_j > 0$ and $x_i, x_j > 0$. ☐

Theorem 7.3 (Universal Morse index)**.** *At every critical point W_ϵ, the Hessian of F_β has*

$$\frac{d(d+1)}{2}$$

positive eigenvalues and

$$\binom{d}{2}$$

negative eigenvalues. It has no zero eigenvalues.

Proof. On the diagonal subspace $\mathcal{D}$, the energy Hessian is the identity. By (41) and Theorem 2.2, the entropy Hessian is negative definite after a diagonal sign congruence. Hence

$$D^2 F_\beta|_{\mathcal{D}} = D^2 E_d|_{\mathcal{D}} - \frac{1}{\beta} D^2 S_N|_{\mathcal{D}}$$

is positive definite. This gives d positive directions.

On $\mathcal{V}_{ij}$, the energy Hessian vanishes. By (42) and Lemma 7.2, the entropy Hessian has one negative and one positive eigenvalue. Therefore $-\beta^{-1} D^2 S_N$ has one positive and one negative eigenvalue on each $\mathcal{V}_{ij}$. There are $\binom{d}{2}$ such pair spaces. Adding the counts gives

$$d + \binom{d}{2} = \frac{d(d+1)}{2}$$

positive directions and $\binom{d}{2}$ negative directions, with no kernel. ☐

The unstable off-diagonal line associated with the pair (i,j) is particularly explicit. It is the line on which

$$\epsilon_i a + \epsilon_j b = 0,$$

namely

$$\mathrm{span}\{E_{ij} - \epsilon_i \epsilon_j E_{ji}\}.$$

The complementary off-diagonal line

$$\operatorname{span}\{E_{ij} + \epsilon_i \epsilon_j E_{ji}\}$$

is stable at the Hessian level.

7.2. **Linearization of the Riemannian flow.** Let

$$\mathcal{X}(W) = -\mathcal{A}_{N,W}(F'_\beta(W))$$

denote the vector field. At a critical point, the derivative of $\mathcal{A}_{N,W}$ is multiplied by $F'_\beta(W_\epsilon) = 0$. Hence

$$D\mathcal{X}(W_\epsilon) = -\mathcal{A}_{N,W_\epsilon} \circ D^2 F_\beta(W_\epsilon). \tag{43}$$

Corollary 7.4 (Hyperbolicity and stable dimensions)**.** *Every W_ϵ is a hyperbolic equilibrium of the free-energy flow. Its stable and unstable dimensions are*

$$\dim E^s = \frac{d(d+1)}{2}, \qquad \dim E^u = \binom{d}{2}.$$

Proof. The operator $\mathcal{A}_{N,W_\epsilon}$ is Frobenius self-adjoint and positive definite. Therefore

$$\mathcal{A}_{N,W_\epsilon} D^2 F_\beta(W_\epsilon)$$

is similar to the symmetric matrix

$$\mathcal{A}_{N,W_\epsilon}^{1/2} D^2 F_\beta(W_\epsilon) \mathcal{A}_{N,W_\epsilon}^{1/2}.$$

By Sylvester's law of inertia, the latter has the same inertia as $D^2 F_\beta(W_\epsilon)$. The minus sign in (43) makes positive Hessian directions dynamically stable and negative Hessian directions unstable. The result follows from Theorem 7.3. ☐

8. The rank boundary cannot be a finite limit

8.1. **The Lyapunov identity.** Along the free-energy flow,

$$\begin{aligned}
\frac{d}{dt} F_\beta(W(t)) &= \left\langle F'_\beta(W), \dot{W} \right\rangle_F \\
&= -\left\langle F'_\beta(W), \mathcal{A}_{N,W} F'_\beta(W) \right\rangle_F \\
&= -\left\| \operatorname{grad}_{g^N} F_\beta(W) \right\|_{g^N}^2 \leq 0.
\end{aligned} \tag{44}$$

Thus F_β is a strict Lyapunov function on $\mathrm{GL}_d(\mathbb{R})$.

8.2. **An infinite barrier at corank at least two.**

Proposition 8.1 (Corank at least two)**.** *Let $W(t)$ be a full-rank solution. It cannot converge to a finite matrix W_* with*

$$\operatorname{rank} W_* \leq d - 2.$$

Proof. If $W(t) \to W_*$ with rank at most $d-2$, then

$$\sigma_{d-1}(t) \to 0, \qquad \sigma_d(t) \to 0,$$

while all singular values remain bounded. By (10),

$$q_N(\sigma_{d-1}, \sigma_d) \leq c_N \log \sigma_{d-1} + C \longrightarrow -\infty.$$

Every other pair term is bounded above along the bounded trajectory. Hence

$$S_N(W(t)) \longrightarrow -\infty.$$

The energy has a finite limit, so $F_\beta(W(t)) \to +\infty$, contradicting the monotonicity (44). □

Remark 8.2. This argument is stronger than a local vector-field calculation. The full-rank free energy has an infinite barrier at every finite matrix of corank at least two.

8.3. **Repulsion from the corank-one boundary.** Suppose the smallest singular value is simple. Let u_i, v_i be corresponding singular vectors and set

$$e_i(W) = u_i^T E_d'(W) v_i.$$

By (3) and the standard singular-value derivative formula,

$$\dot{\sigma}_i = -N\sigma_i^{2-2/N} e_i + \frac{N}{\beta}\sigma_i^{2-2/N} g_i. \tag{45}$$

Proposition 8.3 (Corank-one repulsion). *Let $N > 2$. A full-rank solution cannot converge to a finite matrix of rank $d-1$.*

Proof. Suppose that $W(t) \to W_*$ and $\operatorname{rank} W_* = d-1$. Then

$$\sigma_d(t) \to 0,$$

while $\sigma_1(t), \ldots, \sigma_{d-1}(t)$ converge to positive numbers. For large t, σ_d is isolated and differentiable. Put $b = \sigma_d$ and $a_j = \sigma_j$ for $j < d$. By Lemma 3.5,

$$g_d = \frac{1}{N} b^{2/N-1} \sum_{j<d} a_j^{-2/N} \big(1 + o(1)\big).$$

The quantity e_d is bounded because $W(t)$ is bounded. Substitution into (45) yields

$$\dot{b} = \frac{1}{\beta}\left(\sum_{j<d} a_j^{-2/N}\right) b - N e_d b^{2-2/N} + o(b). \tag{46}$$

Since $N > 2$,

$$2 - \frac{2}{N} > 1,$$

so $b^{2-2/N} = o(b)$. The positive linear term in (46) dominates. Hence there are $c > 0$ and T such that

$$\dot{\sigma}_d(t) \ge c\sigma_d(t) > 0 \qquad \text{for } t \ge T.$$

This is incompatible with $\sigma_d(t) \to 0$. □

Corollary 8.4 (No finite singular limit). *For $N > 2$, no solution starting in $\mathrm{GL}_d(\mathbb{R})$ converges to a finite rank-deficient matrix.*

Remark 8.5 (Why $N = 2$ is different). At depth $N = 2$, the energy term in (46) is also linear in σ_d. The entropy term no longer dominates automatically, and the sign of the total coefficient depends on the state. This is one of the two places where the proof of Theorem 1.1 genuinely uses $N > 2$.

9. Forward completeness

The vector field is smooth on $\mathrm{GL}_d(\mathbb{R})$, but the metric is incomplete at the rank boundary and the free energy is unbounded below at infinity. We therefore prove forward completeness directly.

9.1. **A metric lower bound for radial escape.**

Lemma 9.1 (Radial length estimate)**.** *Let $W : [a, b] \to \mathrm{GL}_d(\mathbb{R})$ be an absolutely continuous curve. Then*

$$\mathrm{Length}_{g^N}(W|_{[a,b]}) \geq \sqrt{N} \left| \sigma_1(W(b))^{1/N} - \sigma_1(W(a))^{1/N} \right|. \tag{47}$$

Proof. The function $t \mapsto \sigma_1(W(t))$ is absolutely continuous and differentiable almost everywhere. At a time when the top singular value is simple, the usual first-variation formula and (3) give

$$|\dot{\sigma}_1|^2 \leq N\sigma_1^{2-2/N} \left\|\dot{W}\right\|_{g^N}^2.$$

If the top singular value has multiplicity $m > 1$, its directional derivatives are eigenvalues of the symmetric part of the corresponding $m \times m$ singular-frame block of $\dot{W}$. Their absolute values are bounded by the Frobenius norm of that block. On the whole top block, including its off-diagonal entries, the continuous value of the coefficient in (3) is $N\sigma_1^{2-2/N}$. The same inequality therefore holds at every differentiability time; see also the standard singular-value differential calculus in [Lew96, LS01].

It follows that

$$\left|\frac{d}{dt}\sigma_1^{1/N}\right| = \frac{1}{N}\sigma_1^{1/N-1}|\dot{\sigma}_1| \leq \frac{1}{\sqrt{N}} \left\|\dot{W}\right\|_{g^N}$$

for almost every t. Integration proves (47). □

Lemma 9.2 (Length-energy estimate)**.** *Along a solution of* (5),

$$\mathrm{Length}_{g^N}(W|_{[0,T]})^2 \leq T\big(F_\beta(W(0)) - F_\beta(W(T))\big).$$

Proof. The speed is

$$\left\|\dot{W}\right\|_{g^N} = \left\|\mathrm{grad}_{g^N} F_\beta\right\|_{g^N}.$$

Cauchy-Schwarz and (44) give

$$\begin{aligned}\mathrm{Length}_{g^N}(W|_{[0,T]})^2 &\leq T \int_0^T \left\|\mathrm{grad}_{g^N} F_\beta\right\|_{g^N}^2 \, dt \\ &= T\big(F_\beta(W(0)) - F_\beta(W(T))\big).\end{aligned}$$

□

Proposition 9.3 (No finite-time escape to infinity)**.** *A solution of* (5) *satisfies*

$$\sup_{t<T} \|W(t)\| < \infty$$

for all $T > 0$.

Proof. Suppose a maximal solution has finite terminal time $T_{\max}$ and unbounded norm. Since $\|W\|_F$ is equivalent to $\sigma_1(W)$ up to a dimension-dependent factor, for arbitrarily large R there is a first time $t_R < T_{\max}$ with $\sigma_1(W(t_R)) = R$.

By Proposition 3.6 and $E_d \geq 0$,

$$\begin{aligned}F_\beta(W(0)) - F_\beta(W(t_R)) &\leq F_\beta(W(0)) + \frac{1}{\beta}S_N(W(t_R)) \\ &\leq C_0 + C_1 \log(1 + R)\end{aligned}$$

for constants independent of R. Combining Lemmas 9.1 and 9.2 gives

$$\sqrt{N}\left(R^{1/N}-\sigma_1(W(0))^{1/N}\right)\le\sqrt{T_{\max}\big(C_0+C_1\log(1+R)\big)}.$$

The left side grows like $R^{1/N}$ and the right side like $\sqrt{\log R}$, a contradiction. □

9.2. **No finite-time collision with the rank boundary.**

Lemma 9.4 (A lower entropy bound on trajectories)**.** *Along every solution,*

$$S_N(W(t))\ge-\beta F_\beta(W(0)).$$

Proof. Since $F_\beta(W(t))\le F_\beta(W(0))$ and $E_d\ge 0$,

$$S_N(W(t))=\beta\big(E_d(W(t))-F_\beta(W(t))\big)\ge-\beta F_\beta(W(0)).$$

□

Lemma 9.5 (Two small singular values are impossible on a bounded sublevel)**.** *Fix $M,L>0$. There exists $\delta>0$ such that*

$$\sigma_1(W)\le M,\qquad S_N(W)\ge -L$$

imply

$$\sigma_{d-1}(W)\ge\delta.$$

Proof. Otherwise there is a bounded sequence with $\sigma_{d-1}\to 0$. Then also $\sigma_d\to 0$, and the pair term $q_N(\sigma_{d-1},\sigma_d)$ tends to $-\infty$ by (10). All remaining pair terms are bounded above because $\sigma_1\le M$. Thus $S_N\to-\infty$, contradicting $S_N\ge -L$. □

Proposition 9.6 (No finite-time rank loss)**.** *A solution of* (5) *cannot reach the rank boundary in finite time.*

Proof. Suppose a maximal solution remains bounded on $[0,T_{\max})$ with $T_{\max}<\infty$. By Lemmas 9.4 and 9.5, there is $m>0$ such that

$$\sigma_{d-1}(W(t))\ge m$$

throughout the interval. If σ_d were to approach zero, it would eventually be simple. Since $g_d>0$, equation (45) gives

$$\dot\sigma_d\ge -N|e_d|\sigma_d^p,\qquad p=2-\frac{2}{N}>1.$$

Boundedness of W makes $|e_d|$ uniformly bounded by a constant C. Hence, after some time t_0,

$$\dot\sigma_d\ge -NC\sigma_d^p.$$

Since $p>1$, this differential inequality yields

$$\sigma_d(t)\ge\big(\sigma_d(t_0)^{1-p}+NC(p-1)(t-t_0)\big)^{-1/(p-1)}>0$$

for every finite $t\ge t_0$. Thus a positive singular value cannot reach zero in finite time. The trajectory remains in a compact subset of $\mathrm{GL}_d(\mathbb{R})$, so the smooth ODE extends past $T_{\max}$, a contradiction. □

Theorem 9.7 (Forward completeness)**.** *Every solution of* (5) *with initial condition in* $\mathrm{GL}_d(\mathbb{R})$ *exists for all $t\ge 0$ and remains full rank. In particular, the sign of* $\det W(t)$ *is constant.*

Proof. A finite maximal time would force either unbounded norm or approach to the boundary of $\mathrm{GL}_d(\mathbb{R})$. These alternatives are excluded by Propositions 9.3 and 9.6. Continuity of the nonzero determinant then preserves its sign. □

10. Bounded orbits and generic unboundedness

10.1. A bounded orbit is precompact in $\mathrm{GL}_d(\mathbb{R})$.

Proposition 10.1 (Uniform separation from the rank boundary)**.** *If a forward orbit is bounded in $\mathbb{M}_d$, then there is $\delta > 0$ such that*

$$\sigma_d(W(t)) \geq \delta \qquad \textit{for all } t \geq 0.$$

Consequently, the orbit has compact closure in $\mathrm{GL}_d(\mathbb{R})$.

Proof. Let $\sigma_1(W(t)) \leq M$. By Lemmas 9.4 and 9.5,

$$\sigma_{d-1}(W(t)) \geq m > 0.$$

It remains to control $b = \sigma_d$.

When $0 < b < m/2$, the smallest singular value is simple. For $a_j = \sigma_j \in [m, M]$, use the exact formula (14) in (45). Uniformly as $b \downarrow 0$,

$$\frac{N}{\beta} b^{2-2/N} g_d = \frac{1}{\beta} \left(\sum_{j<d} a_j^{-2/N} \right) b + o(b).$$

The energy contribution is bounded below by

$$-NCb^{2-2/N} = o(b).$$

Thus there are $\delta \in (0, m/2)$ and $c > 0$ such that

$$0 < b \leq \delta \quad \Longrightarrow \quad \dot{b} \geq cb > 0. \tag{48}$$

A trajectory cannot cross the level $b = \delta$ downward, and if it starts below δ it initially moves upward. Hence b has a positive lower bound depending on the initial point. Together with boundedness of σ_1, this gives compact closure in $\mathrm{GL}_d(\mathbb{R})$. □

10.2. Convergence of every bounded orbit.

We use only elementary properties of omega-limit sets and the Lyapunov function.

Lemma 10.2 (Bounded orbits converge to equilibria)**.** *Every bounded forward orbit converges to one of the 2^d critical points W_ϵ.*

Proof. By Proposition 10.1, the orbit has compact closure in $\mathrm{GL}_d(\mathbb{R})$. Its omega-limit set

$$\omega(W_0) = \bigcap_{T>0} \overline{\{W(t) : t \geq T\}}$$

is nonempty, compact, invariant, and connected. Connectedness follows because it is the nested intersection of closures of connected tails of the trajectory.

The decreasing function $F_\beta(W(t))$ has a finite limit on the compact orbit closure. It is therefore constant on $\omega(W_0)$. Invariance and the dissipation identity (44) imply

$$\operatorname{grad}_{g^N} F_\beta = 0$$

at every point of the omega-limit set. Thus

$$\omega(W_0) \subset \mathrm{Crit}(F_\beta).$$

By Theorem 6.4, the critical set is finite. A connected subset of a finite set is a singleton, say $\omega(W_0) = \{W_\epsilon\}$. A precompact trajectory with a singleton omega-limit set converges to that point. □

10.3. **Measure-zero bounded dynamics.**

Theorem 10.3 (Generic unboundedness)**.** *The set*

$$\mathcal{B} = \left\{ W_0 \in \mathrm{GL}_d(\mathbb{R}) : \sup_{t \geq 0} \|W(t; W_0)\|_F < \infty \right\}$$

has Lebesgue measure zero. Hence almost every full-rank initial condition has an unbounded forward orbit.

Proof. By Lemma 10.2,

$$\mathcal{B} \subset \bigcup_{\epsilon \in \{-1,1\}^d} W^s(W_\epsilon),$$

where $W^s(W_\epsilon)$ is the global stable set. By Corollary 7.4 and the stable-manifold theorem [HPS77], the local stable manifold of W_ϵ is an embedded submanifold of dimension

$$\frac{d(d+1)}{2} < d^2.$$

It therefore has Lebesgue measure zero. The result follows. ☐

Corollary 10.4 (Absolutely continuous random initialization)**.** *Let W_0 have a distribution absolutely continuous with respect to Lebesgue measure on $\mathbb{M}_d$. Conditional on $W_0 \in \mathrm{GL}_d(\mathbb{R})$, the free-energy orbit is unbounded with probability one. In particular, if G has independent standard Gaussian entries and $\varepsilon > 0$, then*

$$\mathbb{P}\left(\sup_{t \geq 0} \|W(t; \varepsilon G)\|_F = \infty \right) = 1.$$

Proof. A Gaussian matrix is full rank almost surely, and the exceptional set in Theorem 10.3 is Lebesgue null. ☐

11. The zero-temperature comparison in width two

The results above concern the finite-temperature free-energy flow. We now compare them with the ordinary deep-linear training flow, obtained formally by setting $\beta = \infty$. This comparison is useful for two reasons. First, it clarifies the sense in which numerical trajectories can display low-rank selection even though a full-rank trajectory does not converge to a nonzero rank-one completion at a fixed initialization scale. Second, it gives an explicit width-two model for the singular limit in which small-initialization training selects rank one while every fixed finite-temperature problem has generic unbounded dynamics.

Throughout this section $d = 2$ and

$$E_2(W) = \frac{1}{2}\big((W_{11} - 1)^2 + (W_{22} - 1)^2\big).$$

The unregularized reduced balanced flow is

$$\dot{W} = -\mathcal{A}_{N,W}\big(E_2'(W)\big). \tag{49}$$

Its zero-loss set and its rank-one part are

$$\mathcal{Z} = \left\{ \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix} : a, b \in \mathbb{R} \right\}, \qquad \mathcal{H} = \left\{ \begin{pmatrix} 1 & \gamma \\ \gamma^{-1} & 1 \end{pmatrix} : \gamma \neq 0 \right\}.$$

Every point of $\mathcal{Z}$ is an equilibrium of (49).

11.1. **Attraction to the completion manifold and exact rank-one limits.** The completion manifold is normally attracting at each of its full-rank points. The rank-one subset has a different status: it is seen in the limit of small initialization, but it cannot be the exact limit of a trajectory that starts full rank.

Proposition 11.1 (Completion attraction and exclusion of exact rank-one limits)**.** *The following statements hold for every $N \geq 2$.*

(i) *The equilibrium manifold $\mathcal{Z} \cap \mathrm{GL}_2(\mathbb{R})$ is normally attracting at every point, and its normal contraction is uniform on compact subsets.*
(ii) *If $W(0)$ is full rank, then $W(t)$ cannot converge to a point of $\mathcal{H}$ as $t \to \infty$.*

Consequently, low-rank selection from full-rank data can occur only as a singular limit of the endpoint map, for example as the initialization scale tends to zero.

Proof. Let $\Pi : \mathbb{M}_2 \to \mathbb{M}_2$ denote projection onto the diagonal entries. At $W_* \in \mathcal{Z}$, the residual $E_2'(W_*)$ vanishes, and hence the derivative of the vector field in (49) is

$$DX_0(W_*)[H] = -\mathcal{A}_{N,W_*}\big(\Pi H\big).$$

If W_* is full rank, $\mathcal{A}_{N,W_*}$ is positive definite. The operator $-\mathcal{A}_{N,W_*}\Pi$ is similar to the negative semidefinite self-adjoint operator

$$-\mathcal{A}_{N,W_*}^{1/2}\Pi\mathcal{A}_{N,W_*}^{1/2}.$$

Its kernel has dimension two and corresponds to $T_{W_*}\mathcal{Z}$, while its two remaining eigenvalues are strictly negative. The spectral gap is uniform on compact subsets of $\mathcal{Z} \cap \mathrm{GL}_2(\mathbb{R})$, which proves part (i).

For part (ii), suppose that a full-rank trajectory converges to

$$W_\gamma = \begin{pmatrix} 1 & \gamma \\ \gamma^{-1} & 1 \end{pmatrix} \in \mathcal{H}.$$

Write

$$p(t) = \begin{pmatrix} W_{11}(t) - 1 \\ W_{22}(t) - 1 \end{pmatrix}.$$

Because the energy gradient is diagonal, the residual satisfies an exact linear equation with time-dependent coefficients,

$$\dot{p}(t) = -M_N(W(t))p(t), \qquad (M_N(W))_{ij} = \langle E_{ii}, \mathcal{A}_{N,W}(E_{jj})\rangle_F\,.$$

Put

$$\sigma_\gamma = \left|\gamma + \gamma^{-1}\right|, \qquad \alpha_N = 2 - \frac{2}{N}, \qquad \eta_\gamma = \frac{N-2}{\sigma_\gamma^2}.$$

Using the rank-one formula for $\mathcal{A}_{N,W_\gamma}$ gives

$$M_N(W_\gamma) = \sigma_\gamma^{\alpha_N}\begin{pmatrix} 1+\eta_\gamma & \eta_\gamma \\ \eta_\gamma & 1+\eta_\gamma \end{pmatrix}.$$

Its eigenvalues are

$$\sigma_\gamma^{\alpha_N}, \qquad \sigma_\gamma^{\alpha_N}(1+2\eta_\gamma),$$

and are strictly positive. Continuity therefore gives constants $c, C > 0$ such that

$$\|p(t)\| \leq Ce^{-ct}$$

for all sufficiently large t.

Let $\sigma_2(t) > 0$ be the smaller singular value of $W(t)$. For large t it is simple, and the singular-value equation associated with (49) is

$$\dot{\sigma}_2 = -N\sigma_2^{2-2/N} e_2(t), \qquad e_2(t) = u_2(t)^T E_2'(W(t))v_2(t).$$

The preceding exponential estimate implies

$$|e_2(t)| \le \|E_2'(W(t))\|_F \le Ce^{-ct},$$

so e_2 is integrable on a terminal time interval. If $N = 2$, then

$$\frac{d}{dt} \log \sigma_2(t) = -2e_2(t),$$

which implies that $\sigma_2(t)$ has a positive limit. If $N > 2$, set $p_N = (N-2)/N$. Then

$$\frac{d}{dt}\sigma_2(t)^{-p_N} = (N-2)e_2(t),$$

so $\sigma_2(t)^{-p_N}$ has a finite limit and $\sigma_2(t)$ again cannot tend to zero. Both conclusions contradict $W(t) \to W_\gamma$. $\square$

Remark 11.2. The first part of the proposition concerns attraction to the two-dimensional completion manifold $\mathcal{Z}$, not attraction to the one-dimensional hyperbola $\mathcal{H}$ in the full four-dimensional phase space. Inside the rank-one stratum, the diagonal-residual map has rank two and the same positive normal matrix computed in the proof, so $\mathcal{H}$ is locally normally attracting there. In the full matrix space, however, the additional direction tangent to $\mathcal{Z}$ and transverse to $\mathcal{H}$ is neutral.

Remark 11.3 (Exact rank collapse under block observations). Part (ii) is a statement about the diagonal observation pattern, not a general rank rigidity of the reduced flow. For block-diagonal observations with block size $s \ge 2$ and n blocks, Shin and Yun prove that the balanced factor flow started from a full-rank deterministic initialization converges to a zero-loss limit whose singular values vanish exactly beyond the nth [SY26, Theorem 3.3]. There is no conflict. In the block case the collapsing directions are the within-block traceless modes, which are observed; on the invariant matrix family used there, the residual component in such a direction equals the vanishing eigenvalue itself, so the singular-value equation degenerates to $\dot{\lambda} = -N\lambda^{3-2/N}$ and the decay is only algebraic. The uniform spectral gap of $M_N(W_\gamma)$, which produces the exponential residual decay driving the proof of part (ii), is therefore absent. Whether infinite-time rank collapse can occur is thus governed by whether the collapsing directions are seen by the observation pattern.

11.2. **Exact singular-value and angular coordinates.** We next derive an exact four-dimensional coordinate system in the positive-determinant component. Let

$$R(\vartheta) = \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix},$$

and write, locally away from repeated singular values,

$$W = R(\theta+\delta)\begin{pmatrix} s & 0 \\ 0 & t \end{pmatrix} R(\theta-\delta)^T, \qquad s > t > 0.$$

Set

$$u = \frac{s+t}{2}, \qquad v = \frac{s-t}{2}.$$

Then

$$W = \begin{pmatrix} u\cos 2\delta + v\cos 2\theta & v\sin 2\theta - u\sin 2\delta \\ v\sin 2\theta + u\sin 2\delta & u\cos 2\delta - v\cos 2\theta \end{pmatrix}. \tag{50}$$

Thus θ records the common orientation of the two singular frames, while δ records their mismatch. In these coordinates,

$$E_2(W) = \big(u\cos 2\delta - 1\big)^2 + v^2\cos^2 2\theta. \tag{51}$$

Define

$$A_N(s,t) = \sum_{j=0}^{N-1} s^{2(N-1-j)/N} t^{2j/N} = \frac{s^2 - t^2}{s^{2/N} - t^{2/N}},$$

with the continuous value $A_N(s,s) = Ns^{2-2/N}$.

Proposition 11.4 (Exact width-two system)**.** *In the preceding chart, the unregularized flow* (49) *is*

$$\begin{aligned} \dot{s} &= Ns^{2-2/N}\left[(1 - u\cos 2\delta)\cos 2\delta - v\cos^2 2\theta\right], \\ \dot{t} &= Nt^{2-2/N}\left[(1 - u\cos 2\delta)\cos 2\delta + v\cos^2 2\theta\right], \\ \dot{\theta} &= \frac{A_N(s,t)}{4}\sin 4\theta, \\ \dot{\delta} &= \frac{A_N(s,t)}{2u}\big(u\cos 2\delta - 1\big)\sin 2\delta. \end{aligned} \tag{52}$$

In particular, $\theta = \pi/4$ and $\delta = 0$ define invariant hypersurfaces, and their intersection is the symmetric equal-diagonal plane.

Proof. A variation of the singular-value decomposition satisfies

$$U^T(dW)V = \begin{pmatrix} ds & 2v\,d\theta - 2u\,d\delta \\ 2v\,d\theta + 2u\,d\delta & dt \end{pmatrix}.$$

The eigenvalues of $\mathcal{A}_{N,W}$ in the SVD frame are

$$Ns^{2-2/N}, \qquad Nt^{2-2/N}, \qquad A_N(s,t), \qquad A_N(s,t).$$

Consequently, the Riemannian metric is

$$g^N = \frac{ds^2}{Ns^{2-2/N}} + \frac{dt^2}{Nt^{2-2/N}} + \frac{8v^2}{A_N(s,t)}\,d\theta^2 + \frac{8u^2}{A_N(s,t)}\,d\delta^2.$$

Differentiating (51) and taking its negative gradient with respect to this diagonal metric gives (52). ∎

The angular equations in (52) separate much of the geometry. If $0 < \theta < \pi/2$, then θ moves monotonically toward $\pi/4$. If $u\cos 2\delta < 1$ and $|\delta| < \pi/4$, then $|\delta|$ decreases. Thus the two singular frames align even though, as shown below, the antisymmetric matrix coordinate can be transiently amplified.

11.3. **The symmetric invariant plane.** Let

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathcal{P} = \{xI + zQ : x, z \in \mathbb{R}\}.$$

The plane $\mathcal{P}$ is invariant under (49). If

$$\lambda_+ = x + z, \qquad \lambda_- = x - z$$

are the two signed eigenvalues, then

$$\dot{\lambda}_\pm = N(1-x)|\lambda_\pm|^{2-2/N}, \qquad x = \frac{\lambda_+ + \lambda_-}{2}. \tag{53}$$

This system can be integrated exactly up to the terminal zero-loss condition $x_\infty = 1$.

Proposition 11.5 (Exact depth transition on the symmetric plane)**.** *Assume the initialization scale $\varepsilon > 0$ is sufficiently small.*

(i) *Let $N = 2$ and*

$$\lambda_+(0) = \varepsilon a, \qquad \lambda_-(0) = \varepsilon b, \qquad a, b > 0.$$

Then

$$W_\infty = \begin{pmatrix} 1 & \dfrac{a-b}{a+b} \\ \dfrac{a-b}{a+b} & 1 \end{pmatrix}.$$

Thus the endpoint is independent of ε and fills the interior of the minimal-nuclear-norm segment as a/b varies.

(ii) *Let $N = 2$ and*

$$\lambda_+(0) = \varepsilon a > 0, \qquad \lambda_-(0) = -\varepsilon b < 0.$$

Then

$$\lambda_-(\infty) = -\delta_\varepsilon, \qquad \lambda_+(\infty) = 2 + \delta_\varepsilon,$$

where

$$\delta_\varepsilon = \sqrt{1 + \varepsilon^2 ab} - 1 = \frac{ab}{2}\varepsilon^2 + O(\varepsilon^4).$$

In particular, the smaller singular value is of order ε^2 and $W_\infty \to J$.

(iii) *Let $N > 2$, put $p_N = (N-2)/N$, and suppose*

$$\lambda_+(0) = \varepsilon a, \qquad \lambda_-(0) = \varepsilon b, \qquad a > b > 0.$$

Then

$$\lambda_-(\infty) \sim \left(b^{-p_N} - a^{-p_N}\right)^{-1/p_N} \varepsilon, \qquad W_\infty \longrightarrow J.$$

If instead

$$\lambda_+(0) = \varepsilon a > 0, \qquad \lambda_-(0) = -\varepsilon b < 0,$$

then

$$|\lambda_-(\infty)| \sim \left(a^{-p_N} + b^{-p_N}\right)^{-1/p_N} \varepsilon, \qquad W_\infty \longrightarrow J.$$

Exchanging the two eigenlines gives the corresponding limits toward DJD.

For every fixed $\varepsilon > 0$ the endpoints in all of these cases are full rank.

Proof. Equation (53) shows that the plane is invariant. In every case under consideration $x(0) < 1$. Since $E_2 = (x-1)^2$ on $\mathcal{P}$, energy dissipation implies that $x(t)$ is nondecreasing and cannot cross 1. In the same-sign cases the eigenvalues are bounded because their sum is $2x \le 2$. In the opposite-sign cases the conserved quantities derived below rule out simultaneous divergence of the two eigenvalue magnitudes. Hence the trajectory is bounded. If $x_\infty < 1$, the positive eigenvalue in (53) would keep increasing at a rate bounded away from zero, a contradiction. Thus $x_\infty = 1$.

When $N = 2$ and both eigenvalues are positive,

$$\frac{d}{dt}\log\lambda_+ = \frac{d}{dt}\log\lambda_- = 2(1-x),$$

so their ratio is conserved. The identity $\lambda_+(\infty) + \lambda_-(\infty) = 2$ gives part (i).

If $N = 2$ and the signs are opposite, then

$$\frac{d}{dt}\log\lambda_+ = 2(1-x), \qquad \frac{d}{dt}\log|\lambda_-| = -2(1-x).$$

Hence $\lambda_+|\lambda_-| = \varepsilon^2 ab$. Writing the terminal values as $2+\delta_\varepsilon$ and $-\delta_\varepsilon$ gives

$$\delta_\varepsilon(2+\delta_\varepsilon) = \varepsilon^2 ab,$$

which proves part (ii).

Suppose now that $N > 2$ and define

$$\psi_N(\lambda) = \operatorname{sgn}(\lambda)|\lambda|^{-p_N}.$$

A direct differentiation of (53) gives

$$\frac{d}{dt}\psi_N(\lambda_\pm) = -(N-2)(1-x).$$

Thus

$$\psi_N(\lambda_+) - \psi_N(\lambda_-)$$

is conserved. In the positive-positive case, if $t_\varepsilon = \lambda_-(\infty)$, then

$$t_\varepsilon^{-p_N} - (2-t_\varepsilon)^{-p_N} = \varepsilon^{-p_N}\left(b^{-p_N} - a^{-p_N}\right).$$

This implies the first asymptotic in part (iii). In the positive-negative case, if $\delta_\varepsilon = |\lambda_-(\infty)|$, then

$$(2+\delta_\varepsilon)^{-p_N} + \delta_\varepsilon^{-p_N} = \varepsilon^{-p_N}\left(a^{-p_N} + b^{-p_N}\right),$$

which gives the second asymptotic. The matrix limits follow from the fixed eigenvectors of $\mathcal{P}$. $\square$

Remark 11.6 (Relation to the results of Shin and Yun). The same-sign cases of Proposition 11.5 recover, in width two, the diagonal-observation case of [SY26, Theorem 3.3], which precedes this paper. Shin and Yun analyze the factor gradient flow with every layer initialized to the symmetric matrix $(\alpha - \alpha/m)I_d + (\alpha/m)J_d$ for $\alpha > 0$ and $m > 1$. This initialization is balanced, so the end-to-end trajectory solves the reduced flow (49); at $d = 2$ with unit diagonal targets it lies on the plane $\mathcal{P}$ with

$$\lambda_+(0) = \left(\frac{\alpha(m+1)}{m}\right)^N, \qquad \lambda_-(0) = \left(\frac{\alpha(m-1)}{m}\right)^N.$$

Their implicit equations for the limiting singular values [SY26, equations (8) and (9)] are precisely the conservation of $\psi_N(\lambda_+) - \psi_N(\lambda_-)$ used in the proof above, their depth-two closed form is the endpoint of part (i), and their decoupled case $m = \infty$ is the degenerate ray $a = b$, whose endpoint is the identity. Their Corollary 3.4, in which the stable rank of the limit tends to one as $\alpha \to 0$, is the singular limit of the endpoint map and is consistent with Proposition 11.1: at every fixed α the width-two endpoint is full rank. Since the ratio $\lambda_+(0)/\lambda_-(0) = \big((m+1)/(m-1)\big)^N$ covers all of $(1, \infty)$, their family realizes every same-sign initialization on $\mathcal{P}$ up to scale. The opposite-sign cases, in which the determinant is negative, and Theorem 11.7 have no analogue in [SY26]. The derivation here also gives their algebraic computation a geometric interpretation: the implicit equations are conservation laws of the Riemannian system (53). At depth two, the positive-determinant segment law of part (i) should also be compared with the connectivity framework of Bai, Zhao, and Zhang, in which depth-two completion with disconnected observations whose components are complete bipartite selects minimal-nuclear-norm solutions under additional assumptions [BZZ24].

The proposition gives a rigorous form of the depth-two-versus-deeper transition observed numerically. At depth two, positive determinant can retain two singular values of order one as $\varepsilon \to 0$. At every depth $N > 2$, the smaller terminal singular value is of order ε on this plane. For negative determinant, depth two instead produces the smaller scale ε^2.

11.4. A nonlinear nonsymmetric perturbation.

The symmetric plane sets both angular variables to their invariant values. The next result allows a fixed mismatch between the left and right singular frames. It is therefore a genuinely nonsymmetric perturbation, although the two observed diagonal entries remain equal.

Theorem 11.7 (Equal-diagonal perturbations). *Fix $N > 2$, $a > b > 0$, and $\delta_0 \in (-\pi/4, \pi/4)$. Consider the solution of (52) with*

$$s(0) = \varepsilon a, \qquad t(0) = \varepsilon b, \qquad \theta(0) = \frac{\pi}{4}, \qquad \delta(0) = \delta_0.$$

For all sufficiently small $\varepsilon > 0$, the solution converges to a full-rank zero-loss completion $W_{\varepsilon,\infty}$. If $p_N = (N-2)/N$, then

$$t_\infty \sim \big(b^{-p_N} - a^{-p_N}\big)^{-1/p_N} \varepsilon.$$

Moreover, there is a constant $C_{N,a,b} > 0$ such that

$$\sin 2\delta_\infty = C_{N,a,b} \sin 2\delta_0 \, \varepsilon^{2/N} \big(1 + o(1)\big).$$

Consequently,

$$W_{\varepsilon,\infty} \longrightarrow J \qquad \text{as } \varepsilon \to 0,$$

and

$$(W_{\varepsilon,\infty})_{21} - (W_{\varepsilon,\infty})_{12} = O(\varepsilon^{2/N}).$$

If $\sin 2\delta_0 \neq 0$, the last quantity has exact order $\varepsilon^{2/N}$.

Proof. The hypersurface $\theta = \pi/4$ is invariant by (52). On this hypersurface, put $x = u \cos 2\delta$. For sufficiently small ε, one has $x(0) < 1$. Since $E_2 = (x-1)^2$ there, energy dissipation implies that $x(t)$ increases and cannot cross 1. Also $|\delta(t)|$ decreases, so $\cos 2\delta(t) \geq \cos 2\delta_0 > 0$. It follows that s and t increase, while

$$u(t) = \frac{x(t)}{\cos 2\delta(t)} \leq \frac{1}{\cos 2\delta_0}.$$

Thus the trajectory is bounded and remains full rank for every finite time. Its monotone coordinates have limits. If $x_\infty < 1$, the equation for $\dot{s}$ would remain bounded below by a positive constant, contradicting boundedness. Hence $x_\infty = 1$, so the terminal point has zero loss.

On $\theta = \pi/4$, the s and t equations imply the exact conservation law

$$t^{-p_N} - s^{-p_N} = \varepsilon^{-p_N}\left(b^{-p_N} - a^{-p_N}\right). \tag{54}$$

Since $x_\infty = 1$ implies $u_\infty \geq 1$, one has $s_\infty \geq 1$. It follows from (54) that

$$t_\infty \sim \left(b^{-p_N} - a^{-p_N}\right)^{-1/p_N}\varepsilon.$$

For $\delta_0 \neq 0$, the angular equation and the s equation give

$$\frac{d}{ds}\log|\sin 2\delta| = -\frac{2A_N(s,t)}{N(s+t)s^{2-2/N}}. \tag{55}$$

The case $\delta_0 = 0$ follows by invariance. Put

$$c_0 = b^{-p_N} - a^{-p_N}, \qquad h(q) = \left(q^{-p_N} + c_0\right)^{-1/p_N}.$$

Equation (54) gives $t = \varepsilon h(s/\varepsilon)$. By the homogeneity of A_N, integration of (55) yields

$$\log\frac{|\sin 2\delta_0|}{|\sin 2\delta_\infty|} = \int_a^{s_\infty/\varepsilon} \frac{2A_N(q,h(q))}{N(q+h(q))q^{2-2/N}}\,dq.$$

The integrand satisfies

$$\frac{2A_N(q,h(q))}{N(q+h(q))q^{2-2/N}} = \frac{2}{Nq} + O(q^{-1-2/N}) \qquad (q \to \infty).$$

Consequently, the integral equals

$$\frac{2}{N}\log\frac{1}{\varepsilon} + O(1).$$

It follows that $\delta_\infty \to 0$. Since $x_\infty = u_\infty \cos 2\delta_\infty = 1$, we obtain $u_\infty \to 1$, and the preceding asymptotic for t_∞ gives $s_\infty \to 2$. Subtracting $2/(Nq)$ from the integrand shows that the $O(1)$ term has a finite limit. More precisely,

$$L_{N,a,b} = \int_a^\infty \left[\frac{2A_N(q,h(q))}{N(q+h(q))q^{2-2/N}} - \frac{2}{Nq}\right]dq$$

converges, and

$$C_{N,a,b} = \exp(-L_{N,a,b})\left(\frac{a}{2}\right)^{2/N} > 0.$$

This proves the asserted asymptotic for $\sin 2\delta_\infty$.

Finally, (50) with $\theta = \pi/4$ gives

$$(W_{\varepsilon,\infty})_{21} - (W_{\varepsilon,\infty})_{12} = 2u_\infty \sin 2\delta_\infty.$$

Since $u_\infty \to 1$, the asymmetry estimate and the convergence to J follow. □

At depth three, when $\sin 2\delta_0 \neq 0$, the theorem gives the two distinct endpoint scales

$$\sigma_2(W_{\varepsilon,\infty}) \asymp \varepsilon, \qquad |(W_{\varepsilon,\infty})_{21} - (W_{\varepsilon,\infty})_{12}| \asymp \varepsilon^{2/3}.$$

Thus the endpoint can be much less symmetric than its smallest singular value would suggest, while still converging to the symmetric rank-one matrix J.

11.5. **Transverse linearization and the remaining width-two problem.** For completeness, linearize (52) along a trajectory in the symmetric slice

$$\theta = \frac{\pi}{4}, \qquad \delta = 0.$$

Writing $\eta = \theta - \pi/4$, one obtains

$$\dot{\eta} = -A_N(s,t)\eta, \qquad \dot{\delta} = -\frac{A_N(s,t)}{u}(1-u)\delta.$$

Both angular perturbations contract while $u < 1$. In raw matrix coordinates, however, the conclusion is subtler. Let

$$y = v\cos 2\theta, \qquad r = -u\sin 2\delta.$$

To first order,

$$\dot{y} = \left[N(1-u)\frac{s^{2-2/N} - t^{2-2/N}}{s-t} - A_N(s,t)\right] y,$$

and

$$\dot{r} = (1-u)\frac{B_N(s,t)}{s+t}\, r,$$

where

$$B_N(s,t) = N\big(s^{2-2/N} + t^{2-2/N}\big) - 2A_N(s,t).$$

Weighted arithmetic-geometric mean gives $B_N(s,t) \geq 0$, with equality identically when $N = 2$ and strict inequality for $N > 2$ when $s \neq t$. Thus the left-right angular mismatch contracts, but the antisymmetric matrix entry may be transiently amplified. Theorem 11.7 shows that this amplification is compatible with an endpoint asymmetry that still vanishes as $\varepsilon \to 0$.

The exact Gaussian limit law remains open. The system (52) shows that unequal diagonal entries are driven toward $\theta = \pi/4$, and their feedback into the singular-value equations is quadratic in $\theta - \pi/4$. A complete proof for random initialization would nevertheless require a global analysis of the angular phase portrait, a signed-SVD treatment of the negative-determinant component, and control of the separatrices between the branches converging toward J and DJD.

12. Interpretation and open problems

12.1. **Failure of an equilibrium selection principle.** The original thermodynamic proposal was that the minimizing set

$$\operatorname*{argmin}_{W\in \mathrm{GL}_d(\mathbb{R})} F_\beta(W)$$

might select among a degenerate family of energy minimizers, especially as $\beta \to \infty$ [Men25, MY25]. For the diagonal-completion loss, the proposed set is empty for every finite β. The stronger dynamical conclusions are:

the only finite full-rank equilibria are saddles,

finite rank-deficient limits are impossible,

and generic trajectories are unbounded.

Thus fiber entropy does not select one completion from the affine zero-loss set. It creates drift along noncompact directions that the loss does not control.

The obstruction is not peculiar to the diagonal observation pattern. By Theorem 4.5, every non-injective squared linear matrix-sensing loss on the unrestricted matrix space has unbounded free energy. The criterion in Proposition 4.1 gives the more intrinsic statement: entropic minimization fails whenever a finite energy sublevel carries arbitrarily large fiber entropy. Diagonal completion realizes the strongest version of this pathology, since the entropy is already unbounded on the exact minimizing set.

This does not make the entropy irrelevant. It identifies a different mechanism: after the energy has nearly minimized the observed coordinates, the entropy acts on the flat unobserved degrees of freedom. In this example, that action is destabilizing rather than selecting.

12.2. **Noncommuting limits and metastability.** The zero-temperature analysis in Section 11 makes one side of the comparison rigorous. At every fixed positive initialization scale, a full-rank unregularized trajectory cannot converge exactly to a nonzero rank-one completion. Nevertheless, on the symmetric invariant plane and on the nonsymmetric equal-diagonal family of Theorem 11.7, the full-rank endpoint converges to J or DJD as $\varepsilon \to 0$. For $N > 2$ the smaller terminal singular value is of order ε, while at $N = 2$ the positive-determinant symmetric family retains a nondegenerate rank-two limit and the negative-determinant family has a smaller singular value of order ε^2.

This is compatible with the finite-temperature theorem because the two procedures ask different long-time questions. Taking the zero-temperature system first and then sending $t \to \infty$ produces a finite completion whose subsequent small-initialization limit may be rank one. At any fixed finite β, by contrast, almost every orbit is unbounded. Thus low-rank selection by ordinary training and eventual entropy-driven escape are not competing descriptions of the same fixed dynamical system.

A genuine metastability theorem remains open. For large finite β, one expects a trajectory first to shadow the unregularized flow, enter a neighborhood of the completion selected at $\beta = \infty$, and only later depart along an entropy-driven direction. Such a theorem should quantify

$$T_{\text{escape}}(\beta, W_0, N, d)$$

and relate it to the initialization scale. The rank-boundary proof suggests that the closest approach to a corank-one state should depend sensitively on β. At corank at least two, the logarithmic barrier may permit exponentially small singular values before the entropy term becomes dynamically decisive.

12.3. **The remaining zero-temperature width-two problem.** The exact calculations in Section 11 do not yet determine the full law of the endpoint under Gaussian end-to-end initialization. The positive-determinant SVD system identifies two stable angular mechanisms: unequal diagonal entries are driven toward $\theta = \pi/4$, and the mismatch of the left and right singular frames is damped while the observed loss is decreasing. The feedback of the first angular displacement into the singular-value competition begins quadratically. These facts explain why the rank-one limit persists on the invariant families treated above, but they do not yet give a global basin theorem.

A complete solution requires a global analysis of the angular phase portrait, a signed-SVD chart for the negative-determinant component, and a description of the separatrices between the branches tending toward J and DJD. The final probabilistic step is then to push the Gaussian initialization law through the resulting endpoint map. At depth two and positive determinant, this should identify the nondegenerate law on the minimal-nuclear-norm segment. At depths $N > 2$, it should determine the probabilities and rates of concentration near the two rank-one completions.

12.4. **What generic unboundedness does not yet say.** The conclusion

$$\sup_{t\geq 0} \|W(t)\|_F = \infty$$

does not determine an escape rate or asymptotic direction. It also does not prove that the norm tends to infinity without large returns. Natural next questions are:

(1) Does $\|W(t)\|_F \to \infty$ for almost every initial condition?

(2) Does the completion loss satisfy $E_d(W(t)) \to 0$ along generic escaping trajectories?

(3) Does $W(t)/\|W(t)\|_F$ converge, and if so, what ranks and sign patterns can occur?

(4) Can one classify singular escape at infinity, where some singular values diverge and others vanish?

12.5. **The finite-temperature depth-two problem.** The depth-two results in Section 11 concern the unregularized system. At finite temperature, unboundedness of the free energy remains true in every width, but the equilibrium and rank-boundary arguments used in the main theorem change. For $d \geq 3$, the entropy is still strictly concave in singular-value coordinates, while its derivative at a one-zero boundary point is finite and the energy and entropy terms have the same order in the smallest singular value. Consequently, the proof of existence in every sign chamber and the universal boundary repulsion both require modification. Width two is additionally exceptional because the singular-value Hessian has a kernel and critical manifolds appear.

A complete finite-temperature depth-two classification in arbitrary width is therefore a separate problem. It may involve temperature-dependent loss of interior equilibria through the corank-one boundary.

12.6. **Adding confinement and recovering a selection problem.** The general criteria of Section 4 identify a hierarchy of hypotheses. To make F_β merely bounded below, the energy must dominate the logarithmic entropy growth. To prevent escape in norm, strict domination is needed. To guarantee an actual minimizer inside the open manifold $\mathrm{GL}_d(\mathbb{R})$, one must additionally prevent minimizing sequences from approaching the corank-one boundary, where the full-rank entropy has a finite limit.

A coercive penalty such as

$$E_{d,\lambda}(W) = E_d(W) + \frac{\lambda}{2}\|W\|_F^2, \qquad \lambda > 0,$$

is more than sufficient at infinity: the quadratic term dominates the logarithmic entropy for every finite β. More generally, Corollary 4.9 shows that any superlogarithmically coercive energy confines the free energy in norm. The ridge term does not itself create a barrier at rank loss, however, so it does not by itself guarantee that the infimum is attained in $\mathrm{GL}_d(\mathbb{R})$. One may instead work on a closed full-rank region such as $\{\sigma_d \geq \eta\}$, add a barrier at corank one, or formulate compatible entropies on all rank strata.

Once compactness and attainment are available, the usual low-temperature mechanism becomes rigorous. Proposition 4.12 shows that every limit point of minimizers of $E - \beta^{-1}S_N$ as $\beta \to \infty$ maximizes S_N on the minimizing set of E. Thus entropy can genuinely select among degenerate minimizers when that set is compact, the entropy is continuous there, and the free-energy minimizers remain precompact. The failure for diagonal completion is not a failure of this variational argument. It is the failure of its compactness hypotheses: the minimizing set is noncompact and S_N has no finite maximizer on it.

13. Proof of the main theorem

Proof of Theorem 1.1. Part (i) is Theorem 5.2. Part (ii) is Theorems 6.2 and 6.4. Part (iii) is Theorem 7.3 and Corollary 7.4. Part (iv) is Corollary 8.4. Part (v) is Theorem 9.7. Part (vi) follows from Lemma 10.2 and Theorem 10.3. ☐

Appendix A. A direct derivation of the off-diagonal Hessian block

For completeness, we give the 2×2 calculation behind (42). Let

$$X = \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix}, \qquad H = \begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}, \qquad x \neq y > 0.$$

Then

$$(X + tH)^T (X + tH) = \begin{pmatrix} x^2 + t^2 b^2 & t(xa + yb) \\ t(xa + yb) & y^2 + t^2 a^2 \end{pmatrix}.$$

Second-order eigenvalue perturbation gives

$$\sigma_x''(0) = \frac{1}{x}\left(b^2 + \frac{(xa + yb)^2}{x^2 - y^2}\right),$$
$$\sigma_y''(0) = \frac{1}{y}\left(a^2 - \frac{(xa + yb)^2}{x^2 - y^2}\right).$$

If $g_x = \partial_x f$ and $g_y = \partial_y f$, the first singular-value derivatives vanish, so

$$D^2(f \circ \sigma)(X)[H, H] = g_x \sigma_x''(0) + g_y \sigma_y''(0).$$

Algebra gives

$$D^2(f \circ \sigma)(X)[H, H] = \frac{1}{2}\frac{g_x - g_y}{x - y}(a + b)^2 + \frac{1}{2}\frac{g_x + g_y}{x + y}(a - b)^2.$$

Passing from X to a signed diagonal matrix replaces (a, b) by $(\epsilon_i a, \epsilon_j b)$, which is (42).

Appendix B. Uniform corank-one repulsion on bounded sets

We record the quantitative estimate used in Proposition 10.1. Suppose

$$m \leq \sigma_j \leq M, \qquad j < d, \qquad 0 < b = \sigma_d < m/2.$$

From (14),

$$N b^{2-2/N} g_d = \sum_{j<d} \left[\frac{b}{\sigma_j^{2/N} - b^{2/N}} - \frac{N b^{3-2/N}}{\sigma_j^2 - b^2}\right] \geq \frac{d-1}{M^{2/N}} b - C_1 b^{1+2/N} - C_2 b^{3-2/N}.$$

Also,

$$-N e_d b^{2-2/N} \geq -C_3 b^{2-2/N}$$

on a bounded set. Since every exponent in the three error terms is strictly greater than 1 when $N > 2$, there is $\delta > 0$ such that

$$\dot{b} \geq \frac{d-1}{2\beta M^{2/N}} b \qquad \text{whenever } 0 < b \leq \delta.$$

This is the outward barrier (48).

References

[AA84] A. Albert and J. A. Anderson, *On the existence of maximum likelihood estimates in logistic regression models*, Biometrika **71** (1984), no. 1, 1–10.

[ACH18] Sanjeev Arora, Nadav Cohen, and Elad Hazan, *On the optimization of deep networks: Implicit acceleration by overparameterization*, Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 80, 2018, pp. 244–253.

[ACHL19] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo, *Implicit regularization in deep matrix factorization*, Advances in Neural Information Processing Systems 32, 2019, pp. 7411–7422.

[BRTW22] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg, *Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers*, Information and Inference: A Journal of the IMA **11** (2022), no. 1, 307–353.

[BT74] Albert E. Beaton and John W. Tukey, *The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data*, Technometrics **16** (1974), no. 2, 147–185.

[BZZ24] Zhiwei Bai, Jiajie Zhao, and Yaoyu Zhang, *Connectivity shapes implicit regularization in matrix factorization models for matrix completion*, Advances in Neural Information Processing Systems 37, 2024.

[CKM25] Alan Chen, Tejas S. Kotwal, and Govind Menon, *Entropic regularization in the deep linear network*, 2025, arxiv: 2512.06137.

[CMV23] Nadav Cohen, Govind Menon, and Zsolt Veraszto, *Deep linear networks for matrix completion—an infinite depth limit*, SIAM Journal on Applied Dynamical Systems **22** (2023), no. 4, 3208–3232.

[CP11] Emmanuel J. Candès and Yaniv Plan, *Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements*, IEEE Transactions on Information Theory **57** (2011), no. 4, 2342–2359.

[CR09] Emmanuel J. Candès and Benjamin Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics **9** (2009), no. 6, 717–772.

[CV95] Corinna Cortes and Vladimir Vapnik, *Support-vector networks*, Machine Learning **20** (1995), no. 3, 273–297.

[DK15] Dmitriy Drusvyatskiy and Courtney Kempton, *Variational analysis of spectral functions simplified*, 2015, arxiv:1506.05170.

[FHT08] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics **9** (2008), no. 3, 432–441.

[FHT10] ———, *Regularization paths for generalized linear models via coordinate descent*, Journal of Statistical Software **33** (2010), no. 1, 1–22.

[GD17] Matan Gavish and David L. Donoho, *Optimal shrinkage of singular values*, IEEE Transactions on Information Theory **63** (2017), no. 4, 2137–2152.

[GVRH20] Paris Giampouras, René Vidal, Athanasios A. Rontogiannis, and Benjamin D. Haeffele, *A novel variational form of the Schatten-p quasi-norm*, Advances in Neural Information Processing Systems 33, 2020, pp. 21453–21463.

[GWB$^+$17] Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro, *Implicit regularization in matrix factorization*, Advances in Neural Information Processing Systems 30, 2017.

[HK70] Arthur E. Hoerl and Robert W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), no. 1, 55–67.

[HPS77] Morris W. Hirsch, Charles C. Pugh, and Michael Shub, *Invariant manifolds*, Lecture Notes in Mathematics, vol. 583, Springer-Verlag, Berlin, 1977.

[Hub64] Peter J. Huber, *Robust estimation of a location parameter*, The Annals of Mathematical Statistics **35** (1964), no. 1, 73–101.

[KAvO24] Seijin Kobayashi, Yassir Akram, and Johannes von Oswald, *Weight decay induces low-rank attention layers*, Advances in Neural Information Processing Systems 37, 2024, pp. 4481–4510.

[KLT11] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov, *Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion*, The Annals of Statistics **39** (2011), no. 5, 2302–2329.

[Lew96] Adrian S. Lewis, *Derivatives of spectral functions*, Mathematics of Operations Research **21** (1996), no. 3, 576–588.

[LLT89] Kenneth L. Lange, Roderick J. A. Little, and Jeremy M. G. Taylor, *Robust statistical modeling using the t distribution*, Journal of the American Statistical Association **84** (1989), no. 408, 881–896.

[LS01] Adrian S. Lewis and Hristo S. Sendov, *Twice differentiable spectral functions*, SIAM Journal on Matrix Analysis and Applications **23** (2001), no. 2, 368–386.

[Men25] Govind Menon, *The geometry of the deep linear network*, XIV Symposium on Probability and Stochastic Processes, Progress in Probability, Birkhäuser, 2025, arXiv:2411.09004, pp. 1–47.

[MY25] Govind Menon and Tianmin Yu, *An entropy formula for the deep linear network*, 2025, to appear in SIAM Journal on Mathematical Analysis.

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Review **52** (2010), no. 3, 471–501.

[SHN+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro, *The implicit bias of gradient descent on separable data*, Journal of Machine Learning Research **19** (2018), no. 70, 1–57.

[SY26] Baekrok Shin and Chulhee Yun, *Implicit bias and loss of plasticity in matrix completion: Depth promotes low-rankness*, International Conference on Learning Representations, 2026, arXiv:2603.04703.