NOISE IMPROVES IMPLICIT REGULARIZATION IN LINEAR NETWORKS FOR MATRIX SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how stochasticity affects implicit regularization in two-layer linear networks for matrix sensing. We compare full-batch gradient descent with stochastic gradient descent with additional label noise, using the same data and large initialization. For a low rank target, we prove that deterministic gradient descent does not recover the true matrix, while noisy SGD achieves recovery to arbitrarily small error. The proof relies on the over-parameterization of two-layer networks, and we provide numerical evidence that the effect strengthens with depth.

1 INTRODUCTION

Stochastic gradient descent (SGD) and its variants are the fundamental algorithms underlying the successes of modern machine learning. The standard motivation is computational: when fitting a model over a large dataset, a full gradient is computationally intractable. From this point of view, the randomness of SGD is a price one pays for scalability. However, SGD also has a major effect on model generalization through the phenomenon of *implicit regularization*: when training over-parameterized models with many interpolating solutions, SGD can select solutions which generalize well, even without explicit regularization Nakkiran et al. (2021).

Although implicit regularization is empirically well-documented, a complete theoretical account is still lacking. We don't yet understand the precise relationship between model architecture, loss, data, training algorithm, and the resulting implicit regularization. This theoretical gap is arguably largely due to the technical difficulty of analytically characterizing the output of SGD training, but even once that task has been accomplished, one still must solve the statistical problem of going from a regularized solution to provable generalization, which can be highly non-trivial in general. A prominent line of work has thus instead studied *deterministic* gradient descent, linking small initialization with implicit regularization by matrix norms Gunasekar et al. (2017).

In this work, we consider rather the role of stochasticity itself in implicit regularization. In particular, we are interested in the following question:

Does noise in gradient descent strengthen implicit regularization?

There is theoretical evidence for an affirmative answer to this question in the study of *diagonal* linear networks, where a linear predictor is represented as $\theta = u_+^{\odot 2} - u_-^{\odot 2}$ Pesme et al. (2021); Li et al. (2022). In this setting, Pesme et al. (2021) show that stochasticity changes the implicit bias and can improve generalization. Li et al. (2022) gives a particularly sharp separation for this diagonal model: even from an arbitrarily large initialization in the kernel regime, label-noise SGD can move toward a sparse solution, whereas gradient descent stays close to its initialization and retains the behavior of the kernel regime. Despite the relative simplicity of the diagonal linear network model, the resulting dynamics are complex, and these works leave open the question of whether the same phenomenon persists in more realistic models.

The goal of this work is to give an affirmative answer to this question in the setting of two-layer linear networks, where we represent a matrix predictor as $M = UV^\top$ and train the factors U and V rather than M directly. Compared to diagonal networks, this model incorporates full interactions between neurons, while still avoiding non-linear activations. Our main result shows that SGD with label noise on two-layer linear networks can recover an underlying low-rank matrix with arbitrarily small error, while deterministic gradient descent run from the same initialization achieves trivial recovery.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 MAIN CONTRIBUTIONS

We study low-rank matrix estimation from linear observations. The unknown parameter is a fixed rank-r matrix $M_\star \in \mathbb{R}^{d_1 \times d_2}$, and the training data consists of noise-less observations $(y_i)_{i=1}^n = (\langle A_i, M_\star \rangle)_{i=1}^n$ on sensing matrices $A_1, \dots, A_n \in \mathbb{R}^{d_1 \times d_2}$, with iid standard Gaussian entries. The statistical task is to estimate $M_\star$ from these n observations in the over-parametrized regime where $n < d_1 d_2$ there are many interpolating solutions.

Because rank minimization is computationally intractable, the classic and extensively studied approach to low-rank matrix recovery is to solve a convex relaxation where one adds a nuclear norm penalty as a proxy for low rank solutions Wainwright (2019). The purpose of this work is rather to understand how we can recover $M_\star$ *without* explicit regularization of this nature.

In particular, we consider estimating $M_\star$ through the linear neural network parametrization $M = UV^\mathsf{T}$ and compare two training procedures on the *same dataset* and from the *same initialization*. The first is full-batch gradient descent on the empirical squared loss $L(U, V) = \frac{1}{2n} \sum_{i=1}^n (\langle UV^\top, A_i \rangle - y_i)^2$. The second is stochastic gradient descent with label noise: at every iteration it forms a gradient based on a uniform random observation along with a fresh, independent perturbation to the corresponding label.

We introduce the model and training dynamics formally in Section 2. Our main theorem gives a direct separation between deterministic and noisy training while holding the dataset, architecture, initialization, and step size fixed. Although it is known that with sufficiently small initialization gradient descent can exactly recover a low-rank matrix Jin et al. (2023), to isolate the effect of noise, we study the complementary large-initialization regime, where gradient descent does not meaningfully recover the underlying matrix. Our main finding is that, although gradient descent does not achieve recovery, stochastic gradient descent with label-noise *does*. Informally, our result states the following.

Theorem 1 (informal). *Let $M_\star \in \mathbb{R}^{d_1 \times d_2}$ have rank r, and draw an iid sample of $n \asymp r(d_1 + d_2)$ Gaussian linear observation matrices $(A_i)_{i=1}^n$ of size $d_1 \times d_2$. Start full-batch gradient descent and label-noise SGD from the same sufficiently large initialization. For a sufficiently small step size, classical gradient descent converges to an exact empirical interpolator whose Frobenius error remains bounded away from zero. In contrast, for every target accuracy $\varepsilon > 0$, label-noise SGD run sufficiently long produces an estimator whose Frobenius error is at most ε with high probability.*

This result provides strong evidence that stochasticity can improve the implicit regularization of an over-parameterized model. Strict improvements in generalization due to stochasticity have been demonstrated for diagonal neural networks Li et al. (2022), but, to the best of our knowledge, ours is the first work to prove a strict benefit to generalization due to stochasticity for two-layer linear networks. Beyond the theoretical evidence for the benefit of stochasticity, our analysis of label-noise SGD is also the first to show that an algorithm other than GD can achieve full recovery. In particular, while GD requires small initialization Jin et al. (2023), we show that label-noise SGD works with essentially arbitrary initial conditions. As we discuss more below, this separation necessarily uses the over-parameterization associated to the two-layer network; in particular, it does not hold for one-layer networks. Furthermore, our empirical results for linear networks with more than two layers suggest that this implicit regularization effect seems to strengthen with depth.

1.2 RELATED WORK

Noise and implicit regularization. A substantial line of work studies how stochasticity changes the implicit bias of overparametrized models. The study of label noise, in particular, was popularized by the works Blanc et al. (2020); Damian et al. (2021), which connect persistent label noise to flatness and Hessian-trace regularization around manifolds of interpolating solutions. Concrete benefits for generalization performance were first shown in the diagonal linear network model by Pesme et al. (2021) through an explicit characterization of the implicit bias of stochastic gradient flow. Li et al. (2022) develops the slow-time framework that we use and applies it to the diagonal linear network model; specifically, they show that label-noise SGD can escape a large-initialization regime and recover a sparse predictor with $O(\kappa \log d)$ samples, whereas gradient descent in that regime requires $\Omega(d)$ samples. Related work shows that structured or label noise can favor sparse or low-complexity

solutions in similar parametrizations HaoChen et al. (2021); Pillaud-Vivien et al. (2022). Beyond diagonal parametrizations, Varre et al. (2024) establishes noise-induced preference for low rank solutions in two-layer linear networks via decreasing determinant. Zhang et al. (2026) establish a two-phase training dynamics for overparameterized two-layer linear networks. First, label-noise SGD progressively shrinks the first-layer weights, driving all neurons out of the lazy regime. Once the weights are small, they show that gradient descent aligns the neuron directions with a ground-truth interpolator, after which the end-to-end predictor converges to that interpolator. Corlouer et al. (2026) analyze SGD in deep linear networks in the small-initialization saddle-to-saddle regime, characterizing the effects of sampling and label noise on stagewise learning dynamics..

Implicit regularization for matrix factorization. Low-rank matrix sensing is classically recovered by nuclear-norm minimization assuming certain sufficient conditions on the observation matrices Recht et al. (2010); Candès & Plan (2011). For the factorized parameterization, Gunasekar et al. (2017) proposed the well-known conjecture that deterministic gradient descent with vanishing initialization selects a minimum-nuclear-norm solution, and significant effort has subsequently gone into studying this conjecture. The simple nuclear-norm minimization framework was put into question by subsequent works Arora et al. (2019); Razin & Cohen (2020), which promoted instead a refinement of the conjecture to direct rank minimization. Notable works in this direction include Li et al. (2018; 2021); Jin et al. (2023), that proved recovery from small initialization and described an incremental low-rank learning dynamics. Large initialization has received significantly less attention; Chen et al. (2024) analyzes a population matrix-factorization problem and shows that incremental learning can persist in that regime. For ℓ_2-regularized deep linear networks trained with SGD, a form of implicit regularization towards low rank solutions was established by Wang & Jacot (2024). Our result deliberately works at large initialization, where deterministic gradient descent remains close to a fixed linearized dynamics. This removes the usual small-initialization low-rank bias and makes it possible to isolate the additional regularization created by noise.

Flatness and low-rank structure. There is also a line of work connecting the so-called flatness of the loss landscape to the generalization properties of the solution. Ding et al. (2024) shows that Hessian-trace-flat interpolators recover the ground-truth matrix in matrix and bilinear sensing. Gatmiry et al. (2023) characterizes the corresponding flatness regularizer for matrix factorization; at depth two it is a weighted nuclear norm. Similar identities appear in our proofs, see section 3.1.

2 GD FAILS WHILE LABEL-NOISE SGD RECOVERS

2.1 MATRIX SENSING AND TRAINING DYNAMICS

Measurements. Let $M_\star \in \mathbb{R}^{d_1 \times d_2}$ be a fixed nonzero matrix of rank r. We observe

$$y_i = \langle A_i, M_\star \rangle, \qquad i = 1, \dots, n,$$

where $A_i \in \mathbb{R}^{d_1 \times d_2}$ has independent standard Gaussian entries. Define the sensing map

$$\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^n, \qquad [\mathcal{A}(M)]_i = \langle A_i, M \rangle .$$

Factorized model. We represent the predicting matrix as

$$M(U, V) = UV^\mathsf{T}, \qquad U \in \mathbb{R}^{d_1 \times m}, \quad V \in \mathbb{R}^{d_2 \times m},$$

with width $m \geq d_1 + d_2$. The empirical loss is

$$L(U, V) = \frac{1}{2n} \left\| \mathcal{A}(UV^\mathsf{T}) - y \right\|_2^2 . \tag{1}$$

We will measure statistical recovery by relative Frobenius error: $\|M - M_\star\|_\mathrm{F} / \|M_\star\|_\mathrm{F}$.

A common initialization. Fix any matrix $M_0 \in \mathbb{R}^{d_1 \times d_2}$ and factors

$$U_0 \in \mathbb{R}^{d_1 \times m}, \qquad V_0 \in \mathbb{R}^{d_2 \times m},$$

such that

$$U_0 V_0^\mathsf{T} = M_0, \qquad W_0 := \begin{pmatrix} U_0 \\ V_0 \end{pmatrix} \text{ has full row rank.} \tag{2}$$

Both algorithms start from the same initialization (U_0, V_0). In particular, U_0 and V_0 are each full row rank. Write

$$\alpha_{\text{init}} = \min\{\sigma_{\min}(U_0), \sigma_{\min}(V_0)\}, \tag{3}$$

where $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ denote the smallest and largest singular values of A, respectively. The deterministic lower bound below will additionally require α_{init} to be sufficiently large. The noisy-recovery branch only uses the full-row-rank condition on W_0. The dependence of the deterministic lower bound on the spectral shape is measured by

$$\kappa_{\text{init}} = \frac{\sigma_{\max}(U_0)^2 + \sigma_{\max}(V_0)^2}{\sigma_{\min}(U_0)^2 + \sigma_{\min}(V_0)^2}. \tag{4}$$

Deterministic gradient descent. The deterministic branch applies full-batch gradient descent to (1) with step size η and is run to convergence. The factor updates are

$$(U_{k+1}, V_{k+1}) = (U_k, V_k) - \eta \nabla L(U_k, V_k). \tag{5}$$

Label-noise SGD. At iteration k, sample i_k uniformly from $\{1, \dots, n\}$ and draw an independent Rademacher perturbation $\xi_k \in \{-\sigma, +\sigma\}$. Define

$$\ell_{i,\xi}(U, V) = \frac{1}{2}\big(\langle A_i, UV^\mathsf{T}\rangle - y_i - \xi\big)^2.$$

The factor updates are

$$(U_{k+1}, V_{k+1}) = (U_k, V_k) - \eta \nabla \ell_{i_k,\xi_k}(U_k, V_k). \tag{6}$$

The underlying measurements are noiseless; only the labels used for the stochastic updates are perturbed.

2.2 Main separation theorem

The following theorem formalizes the resulting separation between deterministic and noisy training from the same initialization.

Theorem 1. *There exist $c, C > 0$ such that if $Cr(d_1 + d_2) \le n \le \frac{d_1 d_2}{2}$, then with probability at least $1 - Ce^{-cn}$ over the sensing matrices, for every $\varepsilon, \delta, \sigma \in (0, 1)$, the following holds from the common initialization (2) provided $\eta > 0$ is sufficiently small:*

(a) If α_{init} is sufficiently large relative to κ_{init} and the fixed problem parameters, then deterministic full-batch gradient descent converges to an exact interpolator $\widehat{M}_{\text{GD}}$ satisfying

$$\left\|\widehat{M}_{\text{GD}} - M_\star\right\|_{\text{F}} \ge \frac{1}{4\sqrt{2}\,\kappa_{\text{init}}} \left\|M_\star - M_0\right\|_{\text{F}}.$$

(b) There exists $N_\eta < \infty$ such that label-noise SGD run for N_η iterations has terminal product $\widehat{M}_{\text{noise}} = U_{N_\eta} V_{N_\eta}^\mathsf{T}$ satisfying

$$\mathbb{P}\left(\frac{\left\|\widehat{M}_{\text{noise}} - M_\star\right\|_{\text{F}}}{\left\|M_\star\right\|_{\text{F}}} \le \varepsilon\right) \ge 1 - \delta.$$

The probability in this display is conditional on the sensing matrices and is over the sampled indices and label perturbations.

This result shows that, in the regime where there are many interpolating matrices, label-noise SGD will effectively select the lowest rank interpolating matrix, while GD will generally not achieve any meaningful recovery. While the conclusion for GD is specific to large initialization, the conclusion for label-noise SGD does not require large initialization: it holds rather for any fixed initialization whose stacked factor matrix W_0 has full row rank. We emphasize that, to the best of our knowledge, the only prior generalization guarantees for over-parameterized matrix sensing were for GD with vanishing initialization Jin et al. (2023), and thus our generalization guarantees demonstrate for the

first time that another algorithm, label-noise SGD, can achieve generalization in this regime and in fact for any initialization with full-rank stacked matrix W_0.

We also note that, although the deterministic lower bound additionally requires α_{init} to be sufficiently large relative to κ_{init} and the fixed problem parameters, this additional requirement places no restriction on the prescribed product M_0: in Lemma 13, we show that every fixed M_0 admits full-stack-rank factorizations with arbitrarily large α_{init}.

We finally remark that the separation of Theorem 1 strongly relies on the fact that we take an over-parameterized representation $M = UV^T$ of the matrix M. Indeed, as we verify in Proposition 11 in Appendix A.5, if we instead consider a single layer model where we train on M directly, label-noise SGD *does not* achieve meaningful recovery for any step-size. Combined with Theorem 1 which establishes full recovery for label noise SGD for two-layer linear networks, and the empirical results in Section 4, this strongly suggests that over-parameterization of the model itself plays a significant role in implicit regularization and generalization.

3 Why noise improves recovery

Theorem 1 gives a long-time separation between two algorithms that initially behave similarly. Both first approach the set of exact fits; after interpolation, deterministic gradient descent stops, whereas fresh label noise drives a much slower motion among interpolating solutions. The existence of such a post-interpolation drift is explained by the slow-time framework of Li et al. (2022). We show that in matrix factorization, this objective favors low-rank solutions and that the resulting dynamic actually reaches the target. We then transfer this deterministic picture back to label-noise SGD. The technical conditions are verified in the appendix.

3.1 Deterministic training selects a linearized interpolator far from the target

It is useful to first understand what large initialization does to deterministic training. Write the loss directly as a function of the represented matrix,

$$\ell(M) = \frac{1}{2n}\|\mathcal{A}(M) - y\|_2^2, \qquad L(U,V) = \ell(UV^{\mathsf{T}}).$$

For the analysis, it is convenient to normalize the row Grams. Write

$$\alpha = \alpha_{\text{init}}, \qquad P_0 = \alpha^{-2}U_0U_0^{\mathsf{T}}, \qquad Q_0 = \alpha^{-2}V_0V_0^{\mathsf{T}}. \tag{7}$$

The initialization in (2) then defines the positive-definite linear map

$$\mathcal{G}_0(N) = \frac{1}{2}(P_0N + NQ_0). \tag{8}$$

When α is large relative to the fixed normalized row Grams, the row Gram matrices UU^{T}/α^2 and VV^{T}/α^2 move very little during the fast fitting phase. Consequently the product approximately follows

$$M_{k+1} \simeq M_k - 2\alpha^2\eta\,\mathcal{G}_0\big(\nabla\ell(M_k)\big). \tag{9}$$

Since ℓ is quadratic and $\mathcal{G}_0$ is fixed, (9) is a linear iteration in matrix space. Its exact-fit endpoint is

$$M_{\text{lin}} = M_0 + \mathcal{G}_0\mathcal{A}^*(\mathcal{A}\mathcal{G}_0\mathcal{A}^*)^{-1}\big(y - \mathcal{A}(M_0)\big). \tag{10}$$

Thus, at large initialization, deterministic training behaves like an initialization-dependent linear estimator rather than like a low-rank method. This is analogous to the general kernel/lazy behavior induced by large initialization scale in homogeneous models Woodworth et al. (2020); large-initialization matrix factorization has also been studied in Chen et al. (2024). The particular endpoint (10) and the lower bound below are specific to our setting.

Proposition 2 (Deterministic training selects the linearized interpolator)**.** *Fix a sensing instance for which $\mathcal{A}$ has full row rank. Holding P_0, Q_0, M_0 fixed and increasing the scale α, for a sufficiently small fixed effective step $2\alpha^2\eta$, the endpoint $\widehat{M}_{\text{GD}}$ of deterministic gradient descent satisfies*

$$\|\widehat{M}_{\text{GD}} - M_{\text{lin}}\|_{\text{F}} = O(\alpha^{-2}) \qquad \text{as } \alpha \to \infty. \tag{11}$$

The same asymptotic statement holds for the endpoint of gradient flow. Moreover, under Gaussian sensing, if $n \le d_1 d_2/2$, then with high probability

$$\|M_{\mathrm{lin}} - M_\star\|_{\mathrm{F}} \ge \frac{1}{\sqrt{8}\,\kappa_{\mathrm{init}}} \|M_\star - M_0\|_{\mathrm{F}}. \tag{12}$$

Precise finite-α statements are given in Propositions 8 to 10.

The proposition explains the deterministic branch of Theorem 1. The observed measurements only constrain an n-dimensional subspace of the $d_1 d_2$-dimensional matrix space, and the linearized dynamics choose one particular interpolator according to the initialization geometry. When $n \ll d_1 d_2$, that interpolator need not be close to $M_\star$. In particular, large initialization removes the usual low-rank bias from the fast fitting phase, so any recovery advantage of label noise must come from the dynamics that remain after interpolation.

Why can label-noise SGD leave this solution? For the ith measurement define

$$r_i(U, V) = \langle A_i, UV^{\mathsf{T}} \rangle - y_i.$$

On the interpolation set

$$\Gamma_y = \{(U, V) : \mathcal{A}(UV^{\mathsf{T}}) = y\}, \tag{13}$$

every residual is zero. Hence deterministic gradient descent stops, and ordinary sample-index noise also disappears: the gradient of $\frac{1}{2} r_i^2$ is $r_i \nabla r_i = 0$ for every i. With a fresh label perturbation ξ, however,

$$\nabla \ell_{i,\xi}(U, V) = (r_i - \xi)\nabla r_i = -\xi \nabla r_i, \qquad (U, V) \in \Gamma_y. \tag{14}$$

Thus interpolation is absorbing for deterministic training, but not for label-noise SGD. The remaining question is what cumulative motion the zero-mean fluctuations in (14) produce.

3.2 Label noise induces a slow sharpness flow

We now identify the dynamics that remain after interpolation. Write $W = (U, V)$ and interpret $\nabla r_i(W)$ as the gradient with respect to both factors. At interpolation, (14) has mean zero, while its covariance is

$$\frac{\sigma^2}{n} \sum_{i=1}^{n} \nabla r_i(W) \nabla r_i(W)^{\mathsf{T}}.$$

For the squared loss this is exactly the Hessian. To write its trace explicitly, define

$$S_L = \frac{1}{n} \sum_{i=1}^{n} A_i A_i^{\mathsf{T}}, \qquad S_R = \frac{1}{n} \sum_{i=1}^{n} A_i^{\mathsf{T}} A_i, \tag{15}$$

and

$$J(U, V) = \frac{1}{2} \left(\mathrm{tr}(U^{\mathsf{T}} S_L U) + \mathrm{tr}(V^{\mathsf{T}} S_R V) \right). \tag{16}$$

Proposition 3 (Noise covariance and Hessian trace)**.** *For every $(U, V) \in \Gamma_y$,*

$$\mathbb{E}_{i,\xi}\left[\nabla \ell_{i,\xi}(U, V)\right] = 0, \qquad \mathrm{Cov}_{i,\xi}\left(\nabla \ell_{i,\xi}(U, V)\right) = \sigma^2 \nabla^2 L(U, V). \tag{17}$$

Moreover,

$$\mathrm{tr}\, \nabla^2 L(U, V) = 2J(U, V). \tag{18}$$

The trace identity is proved in Proposition 15; the covariance identity is derived in section E.

Equation (17) says that the surviving noise is highly anisotropic: directions of larger curvature receive larger fluctuations. This does not create a first-order drift, since the noise is centered. The effect appears one order higher. A stochastic step moves the parameters by $O(\eta)$ away from the interpolation manifold; the deterministic gradient then rapidly pulls the iterate back toward the manifold. The resulting second-order displacement along the manifold is $O(\eta^2)$ per step, and therefore accumulates over $O(\eta^{-2})$ iterations. At regular interpolating points, the tangent space of Γ_y is the kernel of $\nabla^2 L$. Hence the covariance in (17) is purely normal to the interpolation manifold: no Brownian component survives in tangent directions, and the limiting motion is deterministic.

Let W_{GF} denote the endpoint of gradient flow started from the common initialization. Applying the slow-time theorem of Li et al. (2022) gives the following limiting dynamics. The nontrivial point in applying the general theorem is that the relevant part of the interpolation set must remain regular and the endpoint map must be smooth; these properties are established in section E.

Proposition 4 (Slow-time limit of label-noise SGD). *Let W_k^η be the label-noise SGD iterates with factor step η. Under the regularity conditions established in section E, for every fixed time $s > 0$,*

$$W^\eta_{\lfloor s/\eta^2 \rfloor} \xrightarrow[\eta\to 0]{\mathbb{P}} Y(s), \tag{19}$$

where $Y(s) \in \Gamma_y$ solves

$$\frac{dY}{ds} = -\frac{\sigma^2}{4}\operatorname{grad}_{\Gamma_y} \operatorname{tr}\nabla^2 L(Y) = -\frac{\sigma^2}{2}\operatorname{grad}_{\Gamma_y} J(Y), \qquad Y(0) = W_{\mathrm{GF}}. \tag{20}$$

Here $\operatorname{grad}_{\Gamma_y}$ denotes the Euclidean gradient projected onto the tangent space of the interpolation manifold. A uniform finite-time version is given in Proposition 40.

The important point is the separation of time scales. On the fast scale, label-noise SGD follows the deterministic dynamics and reaches the neighborhood of Γ_y. On the slow scale $k \asymp \eta^{-2}$, it moves along Γ_y by decreasing J. Proposition 4 therefore reduces the noisy branch to a deterministic question: does this sharpness flow actually reach the low-rank target?

3.3 Why the sharpness flow reaches the low-rank solution

The slow-time theory identifies which flow to study, but does not establish whether that flow converges or what its limit is. Two additional facts are needed for matrix recovery. First, sharpness induces a data-dependent weighted nuclear norm, and we must show that it uniquely recovers $M_\star$. Second, the projected flow is nonconvex in the factors, so identifying its global minimizer does not by itself show that the dynamics reach it. These are the two main steps below.

Connections between flatness and low-rank structure in matrix factorization have appeared previously. Ding et al. (2024) show that Hessian-trace-flat solutions recover the ground truth in low-rank matrix recovery under standard statistical assumptions, while Gatmiry et al. (2023) study the end-to-end regularizer induced by trace-Hessian minimization in deep matrix factorization and give a closed-form characterization at depth two. Related work also shows that label-noise stochasticity can directly induce rank deficiency in two-layer linear networks Varre et al. (2024).

Proposition 5 (Sharpness and weighted nuclear norm). *Assume $S_L, S_R \succ 0$. For every matrix M,*

$$\min_{(U,V)\in\Gamma_y} J(U,V) = \min_{\mathcal{A}(M)=y} \left\| S_L^{1/2} M S_R^{1/2} \right\|_*. \tag{21}$$

In the Gaussian sensing regime of Theorem 1, if $n \geq Cr(d_1+d_2)$, then with high probability $M_\star$ is the unique minimizer of (21). See Propositions 17 and 24.

This proposition makes the low-rank bias explicit. For Gaussian sensing, $\mathbb{E}[S_L] = d_2 I_{d_1}$, $\mathbb{E}[S_R] = d_1 I_{d_2}$, and the empirical matrices concentrate around these values. Hence the induced penalty is close to $\sqrt{d_1 d_2}\,\|M\|_*$, so minimizing sharpness approximately asks for a low-nuclear-norm interpolator. The remaining issue is dynamical. Proposition 5 identifies the unique global solution, but the constrained problem remains nonconvex in the factors (U, V). Global convergence of sharpness minimization has been established in other two-layer settings Gatmiry et al. (2024); here we need to rule out incorrect critical limits in the matrix-sensing geometry.

Proposition 6 (The sharpness flow reaches the recovering solution). *On the high-probability Gaussian sensing event used in Theorem 1, let $W_{\mathrm{GF}} = (U_{\mathrm{cl}}, V_{\mathrm{cl}})$ be the gradient-flow endpoint. The projected gradient flow*

$$\frac{dY}{dt} = -\operatorname{grad}_{\Gamma_y} J(Y), \qquad Y(0) = W_{\mathrm{GF}}, \tag{22}$$

is defined for all $t \geq 0$ and, writing $Y(t) = (U_t, V_t)$,

$$U_t V_t^\mathsf{T} \longrightarrow M_\star. \tag{23}$$

Proposition 36 proves this conclusion from any factorization whose stacked factor matrix has full row rank.

The proof explains why the flow cannot get trapped at an incorrect interpolator. Since J decreases, the factors stay bounded, and the flow converges to a constrained critical point. At such a point, the

first-order conditions give a simple dichotomy: if there is no expanding direction, the point is a global minimizer; otherwise the dynamics have a strictly expanding direction. Because $M_\star$ is the unique global minimizer, any incorrect critical limit must fall into the second case. Full row rank ensures that the expanding direction is actually present in the factors, so their norm would grow exponentially, contradicting boundedness. Hence the limiting product must be $M_\star$; see section D.

The previous two subsections prove the noisy part of Theorem 1. Proposition 6 gives a time S at which the product along (20) is arbitrarily close to $M_\star$, while Proposition 4 shows that, for small enough η, label-noise SGD at $N \simeq S/\eta^2$ is close to the same point with high probability. Deterministic gradient descent, however, has already stopped near the bad linearized interpolator of Proposition 2.

4 Numerical Experiments

4.1 A depth-two illustration: deterministic training versus label-noise SGD

We first illustrate the qualitative separation in Theorem 1 on a small synthetic matrix-sensing problem. We take $d_1 = d_2 = 10$, rank $r = 1$, $n = 45$ Gaussian measurements, and width $m = d_1 + d_2 = 20$. The target $M_\star$ is a random rank-one matrix normalized so that $\|M_\star\|_{\mathrm{F}} = 1$, and every entry of every sensing matrix A_i is sampled independently from $\mathcal{N}(0, 1)$. We use the common initialization $P_0 = Q_0 = I$ and $M_0 = 0$. With $\alpha = 2$ and $m = d_1 + d_2$, we set

$$U_0 = \alpha \begin{bmatrix} I_{d_1} & 0 \end{bmatrix}, \qquad V_0 = \alpha \begin{bmatrix} 0 & I_{d_2} \end{bmatrix},$$

so that

$$U_0 U_0^\top = \alpha^2 I_{d_1}, \qquad V_0 V_0^\top = \alpha^2 I_{d_2}, \qquad U_0 V_0^\top = 0 = M_0.$$

We compare full-batch GD, ordinary single-sample SGD, and single-sample SGD with independent Rademacher label perturbations $\xi_k \in \{-\sigma, +\sigma\}$. All three use the same factor step size $\eta = 10^{-3}$, and for label-noise SGD we take $\sigma = 0.5$. Within each random instance, ordinary SGD and label-noise SGD use the same sequence of sampled measurement indices.

We report the deterministic training error and the generalization error:

$$e_{\mathrm{train}}(M) = \frac{\|\mathcal{A}(M) - y\|_2}{\|y\|_2}, \qquad e_{\mathrm{gen}}(M) = \frac{\|M - M_\star\|_{\mathrm{F}}}{\|M_\star\|_{\mathrm{F}}}.$$

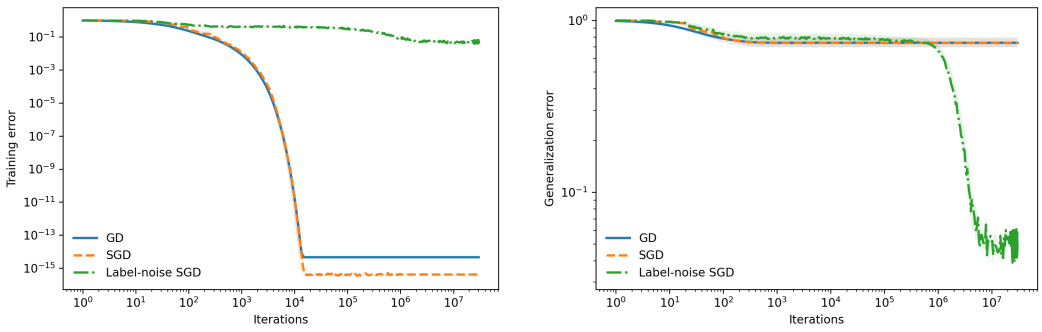


Figure 1: Training and generalization error for two layer network parameterization $M = UV^\top$

Overall, the experiment displays the two stages suggested by the proof. All three methods initially reduce the empirical residual and move toward an interpolator. Once interpolation is reached, the sampling noise of ordinary SGD vanishes and its predictor effectively freezes at the same solution as full-batch GD. Persistent label noise remains active near interpolation and induces a much slower evolution that substantially decreases the population error.

4.2 Effect of depth

We next examine empirically how network depth affects the recovery behavior of label-noise SGD. We keep the same sensing problem as above and $\sigma = 0.5$, and use five independent target/sensing instances. The network weights were initialized using Gaussian Xavier initialization, with independent entries sampled as $W_{ij} \sim \mathcal{N}(0,\, 2/(n_{\text{in}} + n_{\text{out}}))$, where n_{in} and n_{out} denote the input and output dimensions of the corresponding layer. For this comparison, all networks use Xavier initialization, with depth $p \in \{2, 3, 5, 10, 20, 50\}$ and one common factor step $\eta = 5 \times 10^{-4}$ for every depth.

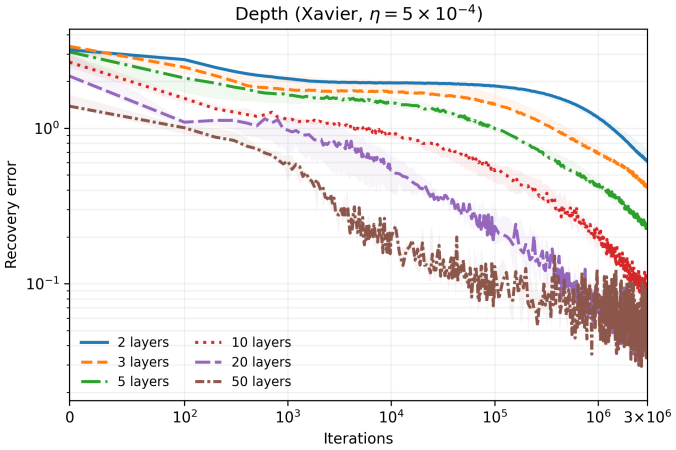


Figure 2: Depth ablation for label-noise SGD with Xavier initialization. Curves are medians over five independent target/sensing instances and shaded bands are interquartile ranges.

Figure 2 shows that empirically, recovery becomes substantially stronger with depth: at the same number of updates, deeper networks generally reach much smaller recovery error. Over the final 20% of the run, the median recovery error decreases from 0.655 at depth $p = 2$ to 0.051 at $p = 20$, before increasing slightly to 0.065 at $p = 50$. Thus, although final-window performance is not strictly monotone in depth, the trajectories exhibit a clear pattern: increasing depth seems to accelerates recovery, making the empirical effect of label-noise SGD more pronounced in deeper networks.

5 Discussion and future work

Our results provide a first step toward understanding how noise and over-parameterization interact beyond the highly structured setting of diagonal networks. Matrix factorization remains simple enough to analyze, while already capturing interactions between parameters that are absent in diagonal models. The fact that noise can still improve recovery in this setting suggests that the phenomenon may reflect a broader principle rather than an artifact of a specific parametrization. This opens the possibility of studying the same mechanism in deeper linear networks and ultimately, in more general over-parameterized models.

Our main result for two-layer linear networks and our lower bound in Proposition 11 for single-layer linear networks suggest an interesting relationship between depth and noise which is further substantiated by our empirical experiments. From our empirical results, increasing depth seems to accelerate recovery yet the behavior of the loss at larger depths is not purely monotone. This raises the possibility of a richer trade-off: depth may strengthen the beneficial effect of noise while also increasing the variability of the resulting dynamics. Quantifying this trade-off and understanding how it depends on the amount of noise, is a natural next step. More broadly, this could help identify which aspects of noise-induced regularization are specific to linear factorization and which persist as general principles of learning in over-parameterized models.

AI use statement. Generative AI tools were used to assist with drafting and refining portions of the manuscript, as well as with the development and presentation of technical mathematical arguments.

REFERENCES

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7413–7424, 2019.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an Ornstein–Uhlenbeck-like process. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 483–513, 2020.

Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

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 (4):2342–2359, 2011. doi: 10.1109/TIT.2011.2111771.

Hengchao Chen, Xin Chen, Mohamad Elmasri, and Qiang Sun. Gradient descent in matrix factorization: Understanding large initialization. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 619–647, 2024.

Guillaume Corlouer, Avi Semler, Alexander Strang, and Alexander Gietelink Oldenziel. Stochastic gradient descent in the saddle-to-saddle regime of deep linear networks. *arXiv preprint arXiv:2604.06366*, 2026.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2):iaae009, 2024. doi: 10.1093/imaiai/iaae009.

Khashayar Gatmiry, Zhiyuan Li, Tengyu Ma, Sashank J. Reddi, Stefanie Jegelka, and Ching-Yao Chuang. What is the inductive bias of flatness regularization? a study of deep matrix factorization models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28040–28052, 2023.

Khashayar Gatmiry, Zhiyuan Li, Sashank J. Reddi, and Stefanie Jegelka. Simplicity bias via global convergence of sharpness minimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15102–15129, 2024.

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6151–6159, 2017.

Jeff Z. HaoChen, Colin Wei, Jason D. Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2315–2357, 2021.

Morris W. Hirsch. *Differential Topology*, volume 33 of *Graduate Texts in Mathematics*. Springer, 1976.

Morris W. Hirsch, Charles C. Pugh, and Michael Shub. *Invariant Manifolds*, volume 583 of *Lecture Notes in Mathematics*. Springer, 1977.

Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon Shaolei Du, and Jason D. Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15200–15238, 2023.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2–47, 2018.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? a mathematical framework. In *International Conference on Learning Representations*, 2022.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: A provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29218–29230, 2021.

Loucas Pillaud-Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2127–2159, 2022.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. doi: 10.1137/070697835.

Aditya Varre, Margarita Sagitova, and Nicolas Flammarion. SGD vs GD: Rank deficiency in linear networks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 60133–60161, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Zihan Wang and Arthur Jacot. Implicit bias of sgd in l_2-regularized linear dnns: One-way jumps from high to low rank. In *International Conference on Learning Representations*, volume 2024, pp. 11141–11167, 2024.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673, 2020.

Tongcheng Zhang, Zhanpeng Zhou, Mingze Wang, Andi Han, Wei Huang, Taiji Suzuki, and Junchi Yan. On the learning dynamics of two-layer linear networks with label noise SGD. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(33):28400–28408, 2026. doi: 10.1609/aaai.v40i33.40069.

A THE DETERMINISTIC BRANCH: LARGE INITIALIZATION SELECTS AN AFFINE LINEARIZED INTERPOLATOR

It is useful to work first at the level of the represented matrix. Define

$$\ell(M) = \frac{1}{2n} \|\mathcal{A}(M) - y\|_2^2, \qquad \nabla\ell(M) = \frac{1}{n}\mathcal{A}^*(\mathcal{A}(M) - y). \tag{24}$$

Because $n \le d_1 d_2$, the vectorized Gaussian sensing matrix has full row rank almost surely. On that probability-one event define

$$\mu_{\mathcal{A}}^{-} = \lambda_{\min}\left(\frac{\mathcal{A}\mathcal{A}^*}{n}\right) > 0, \qquad \mu_{\mathcal{A}}^{+} = \lambda_{\max}\left(\frac{\mathcal{A}\mathcal{A}^*}{n}\right). \tag{25}$$

To connect the initialization in Theorem 1 with the scaled notation used below, set

$$\alpha := \alpha_{\text{init}}, \qquad P_0 := \alpha^{-2} U_0 U_0^{\mathsf{T}}, \qquad Q_0 := \alpha^{-2} V_0 V_0^{\mathsf{T}}. \tag{26}$$

By the definition of α_{init},

$$P_0 \succeq I_{d_1}, \qquad Q_0 \succeq I_{d_2}, \qquad \min\{\lambda_{\min}(P_0), \lambda_{\min}(Q_0)\} = 1. \tag{27}$$

The initialization defines the positive-definite linear operator

$$\mathcal{G}_0(N) = \frac{1}{2}(P_0 N + N Q_0). \tag{28}$$

We need only its lower spectral scale. Write

$$g_- := \lambda_{\min}(\mathcal{G}_0) = \frac{\lambda_{\min}(P_0) + \lambda_{\min}(Q_0)}{2} \ge 1. \tag{29}$$

The eigenvalues of $\mathcal{G}_0$ are the half-sums of eigenvalues of P_0 and Q_0, so the condition number in (4) gives $\|\mathcal{G}_0\|_{\text{op}} = \kappa_{\text{init}} g_-$. Hence, for every matrix N,

$$g_- \|N\|_{\text{F}}^2 \le \langle N, \mathcal{G}_0 N \rangle \le \kappa_{\text{init}} g_- \|N\|_{\text{F}}^2, \tag{30}$$

and therefore

$$g_- \mu_{\mathcal{A}}^{-} I_n \preceq \frac{\mathcal{A}\mathcal{G}_0\mathcal{A}^*}{n} \preceq \kappa_{\text{init}} g_- \mu_{\mathcal{A}}^{+} I_n. \tag{31}$$

The exactly linearized dynamics start from the same initial product M_0 and converge to the affine, initialization-weighted interpolator

$$M_{\text{lin}} = M_0 + \mathcal{G}_0\mathcal{A}^*(\mathcal{A}\mathcal{G}_0\mathcal{A}^*)^{-1}\big(y - \mathcal{A}(M_0)\big). \tag{32}$$

The rest of this section shows that the deterministic dynamics remain close to M_{lin} when α is large. The high-probability estimates needed to make "large" uniform over the sensing instance are deferred to the end of the argument.

A.1 One elementary geometric estimate for the factors

Write

$$W = \begin{pmatrix} U \\ V \end{pmatrix}, \qquad \bar{W} = \frac{W}{\alpha}, \qquad \bar{W}_0 = \frac{W_0}{\alpha}.$$

For the factor-size estimates below, set

$$p_+ := \max\{\lambda_{\max}(P_0), \lambda_{\max}(Q_0)\}. \tag{33}$$

Since $2\|\mathcal{G}_0\|_{\text{op}} = \lambda_{\max}(P_0) + \lambda_{\max}(Q_0)$,

$$p_+ \le 2\kappa_{\text{init}} g_-. \tag{34}$$

The initialization gives

$$\bar{W}_0 \bar{W}_0^{\mathsf{T}} = \begin{pmatrix} P_0 & \alpha^{-2} M_0 \\ \alpha^{-2} M_0^{\mathsf{T}} & Q_0 \end{pmatrix}.$$

Using (27) and (33),

$$\sigma_{\min}(\bar{W}_0)^2 \ge 1 - \frac{\|M_0\|_{\text{op}}}{\alpha^2}, \qquad \|\bar{W}_0\|_{\text{op}}^2 \le p_+ + \frac{\|M_0\|_{\text{op}}}{\alpha^2}. \tag{35}$$

Lemma 7 (What closeness to the initial factors implies)**.** *Suppose*

$$\frac{\|M_0\|_{\mathrm{op}}}{\alpha^2} \le \frac{1}{4}, \qquad \|\bar{W} - \bar{W}_0\|_{\mathrm{op}} \le q \le \frac{1}{2}.$$

Then W has full row rank. Moreover, if

$$P = \alpha^{-2} U U^{\mathsf{T}}, \qquad Q = \alpha^{-2} V V^{\mathsf{T}},$$

then

$$P \succeq (\sqrt{\lambda_{\min}(P_0)} - q)^2 I_{d_1}, \tag{36}$$

$$Q \succeq (\sqrt{\lambda_{\min}(Q_0)} - q)^2 I_{d_2}, \tag{37}$$

$$\|P - P_0\|_{\mathrm{op}} + \|Q - Q_0\|_{\mathrm{op}} \le 2q(2\sqrt{p_+} + q). \tag{38}$$

Proof. The two block rows of $\bar{W} - \bar{W}_0$ have operator norm at most q. Hence, for every $a \in \mathbb{R}^{d_1}$,

$$\begin{aligned} \|(U/\alpha)^{\mathsf{T}} a\|_2 &\ge \|(U_0/\alpha)^{\mathsf{T}} a\|_2 - \|(U/\alpha - U_0/\alpha)^{\mathsf{T}} a\|_2 \\ &\ge \sqrt{\lambda_{\min}(P_0)}\, \|a\|_2 - q\, \|a\|_2 . \end{aligned}$$

Squaring gives (36); the proof of (37) is identical.

For the difference of row Gram matrices,

$$P - P_0 = (U/\alpha - U_0/\alpha)(U/\alpha)^{\mathsf{T}} + (U_0/\alpha)(U/\alpha - U_0/\alpha)^{\mathsf{T}}.$$

By (33),

$$\|U_0/\alpha\|_{\mathrm{op}} \le \sqrt{p_+}, \qquad \|U/\alpha\|_{\mathrm{op}} \le \sqrt{p_+} + q.$$

Therefore

$$\|P - P_0\|_{\mathrm{op}} \le q\big(2\sqrt{p_+} + q\big).$$

The same estimate holds for $Q - Q_0$, proving (38).

Finally, (35) and the hypotheses give

$$\sigma_{\min}(\bar{W}) \ge \sigma_{\min}(\bar{W}_0) - \|\bar{W} - \bar{W}_0\|_{\mathrm{op}} \ge \frac{\sqrt{3}}{2} - \frac{1}{2} > 0.$$

Thus $\bar{W}$, and therefore $W = \alpha \bar{W}$, has full row rank. □

A.2 Deterministic gradient flow: the fast transient and its endpoint

The continuous-time flow is used only to identify the endpoint from which the slow noisy motion starts. The actual factor gradient flow is

$$\dot{U} = -\nabla \ell(U V^{\mathsf{T}}) V, \qquad \dot{V} = -\nabla \ell(U V^{\mathsf{T}})^{\mathsf{T}} U. \tag{39}$$

Multiplying time by $2\alpha^2$ does not change the trajectory or its endpoint; it only changes its speed. We therefore use the rescaled time in which

$$\bar{W}' = -\frac{1}{2\alpha^2} \begin{pmatrix} 0 & \nabla \ell(M) \\ \nabla \ell(M)^{\mathsf{T}} & 0 \end{pmatrix} \bar{W}, \tag{40}$$

$$M' = -\frac{1}{2} \left(P_t \nabla \ell(M) + \nabla \ell(M) Q_t \right), \qquad P_t = \alpha^{-2} U U^{\mathsf{T}}, \quad Q_t = \alpha^{-2} V V^{\mathsf{T}}. \tag{41}$$

Define the explicit radius

$$q_\alpha = \frac{2\left(\sqrt{2p_+} + \frac{1}{2}\right)}{\alpha^2 g_- \mu_{\mathcal{A}}^-} \sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\, \|\mathcal{A}(M_0) - y\|_2 . \tag{42}$$

Proposition 8 (The deterministic flow remains close to its initialization)**.** *If*

$$\frac{\|M_0\|_{\mathrm{op}}}{\alpha^2} \le \frac{1}{4}, \qquad q_\alpha \le \frac{1}{4}, \tag{43}$$

then the deterministic factor flow is global, converges in factor space to an exact fit $(U_{\alpha,\infty}, V_{\alpha,\infty})$*, and*

$$\sup_{t\ge 0} \left\|\bar{W}(t) - \bar{W}_0\right\|_{\mathrm{op}} \le q_\alpha. \tag{44}$$

Its limiting product satisfies

$$\left\|U_{\alpha,\infty}V_{\alpha,\infty}^{\mathsf{T}} - M_{\mathrm{lin}}\right\|_{\mathrm{F}} \le \frac{4\sqrt{\kappa_{\mathrm{init}}}}{g_-\mu_{\mathcal{A}}^-}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\ \|\mathcal{A}(M_0) - y\|_2\ q_\alpha\left(2\sqrt{p_+} + q_\alpha\right). \tag{45}$$

The limiting stack $\binom{U_{\alpha,\infty}}{V_{\alpha,\infty}}$ *has full row rank.*

Proof. The proof is a bootstrap: as long as $\bar{W}$ is near $\bar{W}_0$, the measurement residual decays exponentially; the integral of this residual bounds the total movement of $\bar{W}$; that movement is too small to leave the neighborhood.

Step 1: exponential loss decay inside the neighborhood. Assume temporarily that

$$\left\|\bar{W}(t) - \bar{W}_0\right\|_{\mathrm{op}} < \frac{1}{2}.$$

Using Lemma 7 with $q = 1/2$ gives

$$P_t \succeq \frac{\lambda_{\min}(P_0)}{4} I_{d_1}, \qquad Q_t \succeq \frac{\lambda_{\min}(Q_0)}{4} I_{d_2}.$$

Differentiate $\ell(M_t)$ and substitute (41):

$$\begin{aligned}
\frac{d}{dt}\ell(M_t) &= \langle \nabla\ell(M_t), M_t' \rangle \\
&= -\frac{1}{2}\left\langle \nabla\ell(M_t), P_t\nabla\ell(M_t)\right\rangle - \frac{1}{2}\left\langle \nabla\ell(M_t), \nabla\ell(M_t)Q_t\right\rangle \\
&= -\frac{1}{2}\left\|P_t^{1/2}\nabla\ell(M_t)\right\|_{\mathrm{F}}^2 - \frac{1}{2}\left\|\nabla\ell(M_t)Q_t^{1/2}\right\|_{\mathrm{F}}^2 \\
&\le -\frac{g_-}{4}\left\|\nabla\ell(M_t)\right\|_{\mathrm{F}}^2.
\end{aligned}$$

Now put $r_t = \mathcal{A}(M_t) - y$. From (24),

$$\nabla\ell(M_t) = \frac{1}{n}\mathcal{A}^* r_t,$$

so

$$\|\nabla\ell(M_t)\|_{\mathrm{F}}^2 = \frac{1}{n^2}\left\langle r_t, \mathcal{A}\mathcal{A}^* r_t\right\rangle.$$

Since the eigenvalues of $\mathcal{A}\mathcal{A}^*/n$ lie in $[\mu_{\mathcal{A}}^-, \mu_{\mathcal{A}}^+]$ and $\ell(M_t) = \|r_t\|_2^2/(2n)$,

$$2\mu_{\mathcal{A}}^-\ell(M_t) \le \|\nabla\ell(M_t)\|_{\mathrm{F}}^2 \le 2\mu_{\mathcal{A}}^+\ell(M_t). \tag{46}$$

Consequently

$$\frac{d}{dt}\ell(M_t) \le -\frac{g_-\mu_{\mathcal{A}}^-}{2}\ell(M_t).$$

At time zero, $r_0 = \mathcal{A}(M_0) - y$. Solving the scalar differential inequality gives

$$\ell(M_t) \le e^{-g_-\mu_{\mathcal{A}}^- t/2}\frac{\|\mathcal{A}(M_0) - y\|_2^2}{2n}.$$

The upper inequality in (46) therefore yields

$$\|\nabla \ell(M_t)\|_{\mathrm{F}} \leq \sqrt{\frac{\mu_{\mathcal{A}}^{+}}{n}}\, \|\mathcal{A}(M_0) - y\|_2\, e^{-g_- \mu_{\mathcal{A}}^{-} t/4}. \tag{47}$$

Integrating this expression from 0 to ∞ gives

$$\int_0^\infty \|\nabla \ell(M_t)\|_{\mathrm{F}}\, dt \leq \frac{4}{g_- \mu_{\mathcal{A}}^{-}} \sqrt{\frac{\mu_{\mathcal{A}}^{+}}{n}}\, \|\mathcal{A}(M_0) - y\|_2 . \tag{48}$$

Step 2: a first exit is impossible.

Let T be the supremum of the times for which $\left\|\bar{W}(t) - \bar{W}_0\right\|_{\mathrm{op}} < 1/2$. On $[0, T)$, (35) and (43) give

$$\left\|\bar{W}(t)\right\|_{\mathrm{op}} \leq \left\|\bar{W}_0\right\|_{\mathrm{op}} + \frac{1}{2} \leq \sqrt{2p_+} + \frac{1}{2}.$$

The block matrix in (40) has operator norm $\|\nabla \ell(M_t)\|_{\mathrm{op}} \leq \|\nabla \ell(M_t)\|_{\mathrm{F}}$. Hence, for $t < T$,

$$\begin{aligned}
\left\|\bar{W}(t) - \bar{W}_0\right\|_{\mathrm{op}} &\leq \int_0^t \left\|\bar{W}'(s)\right\|_{\mathrm{op}}\, ds \\
&\leq \frac{\sqrt{2p_+} + \frac{1}{2}}{2\alpha^2} \int_0^t \|\nabla \ell(M_s)\|_{\mathrm{F}}\, ds \\
&\leq \frac{\sqrt{2p_+} + \frac{1}{2}}{2\alpha^2} \frac{4}{g_- \mu_{\mathcal{A}}^{-}} \sqrt{\frac{\mu_{\mathcal{A}}^{+}}{n}}\, \|\mathcal{A}(M_0) - y\|_2 \\
&= q_\alpha \leq \frac{1}{4}.
\end{aligned}$$

This is strictly smaller than the radius $1/2$ used to define T. Continuity therefore rules out a finite first exit. Thus (44) holds for every $t \geq 0$.

Because $\bar{W}$ stays bounded, the polynomial vector field in (40) cannot blow up at a finite time, so the solution is global.

Step 3: the factors converge and the limit interpolates. The same calculation gives

$$\int_0^\infty \left\|\bar{W}'(t)\right\|_{\mathrm{op}}\, dt \leq \frac{\sqrt{2p_+} + \frac{1}{2}}{2\alpha^2} \int_0^\infty \|\nabla \ell(M_t)\|_{\mathrm{F}}\, dt < \infty.$$

Hence $\bar{W}(t)$ is Cauchy and converges. The unscaled factors $W(t) = \alpha \bar{W}(t)$ also converge. The loss estimate from Step 1 tends to zero, hence

$$\mathcal{A}(U_{\alpha,\infty} V_{\alpha,\infty}^{\mathsf{T}}) = y.$$

Finally, (44), (43), and Lemma 7 imply that the stack has full row rank for every t, and the same is true at the limit.

Step 4: compare the product with the exactly linearized flow. Add and subtract the initial row Gram matrices in (41):

$$M_t' = -\mathcal{G}_0 \nabla \ell(M_t) + E_t, \tag{49}$$

where

$$E_t = -\frac{1}{2}\Big((P_t - P_0)\nabla \ell(M_t) + \nabla \ell(M_t)(Q_t - Q_0)\Big).$$

By (44) and (38),

$$\|E_t\|_{\mathrm{F}} \leq q_\alpha\big(2\sqrt{p_+} + q_\alpha\big)\, \|\nabla \ell(M_t)\|_{\mathrm{F}}\, . \tag{50}$$

Now let $\overline{M}_t$ solve the linear equation

$$\overline{M}_t' = -\mathcal{G}_0 \nabla \ell(\overline{M}_t), \qquad \overline{M}_0 = M_0. \tag{51}$$

Applying $\mathcal{A}$ to (51) shows that

$$\frac{d}{dt}\big(\mathcal{A}(\overline{M}_t) - y\big) = -\frac{1}{n}\mathcal{A}\mathcal{G}_0\mathcal{A}^*\big(\mathcal{A}(\overline{M}_t) - y\big).$$

By (31), the coefficient matrix is positive definite, so $\mathcal{A}(\overline{M}_t) - y \to 0$ exponentially. Also, every $\overline{M}_t - M_0$ lies in the range of $\mathcal{G}_0\mathcal{A}^*$. There is exactly one interpolator in the affine space $M_0 + \mathrm{range}(\mathcal{G}_0\mathcal{A}^*)$: if $M = M_0 + \mathcal{G}_0\mathcal{A}^* z$ and $\mathcal{A}(M) = y$, then

$$\mathcal{A}\mathcal{G}_0\mathcal{A}^* z = y - \mathcal{A}(M_0), \qquad z = (\mathcal{A}\mathcal{G}_0\mathcal{A}^*)^{-1}\big(y - \mathcal{A}(M_0)\big),$$

which gives $M = M_{\mathrm{lin}}$ from (32). Therefore $\overline{M}_t \to M_{\mathrm{lin}}$.

To control the difference $D_t = M_t - \overline{M}_t$, use the norm

$$\|N\|^2_{\mathcal{G}_0^{-1}} = \left\langle N, \mathcal{G}_0^{-1} N \right\rangle.$$

The homogeneous part of the difference equation is

$$D' = -\frac{1}{n}\mathcal{G}_0\mathcal{A}^*\mathcal{A}(D).$$

Its squared $\mathcal{G}_0^{-1}$-norm satisfies

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\|D\|^2_{\mathcal{G}_0^{-1}} &= \left\langle \mathcal{G}_0^{-1} D, D' \right\rangle \\ &= -\frac{1}{n}\left\langle D, \mathcal{A}^*\mathcal{A}(D)\right\rangle \\ &= -\frac{1}{n}\|\mathcal{A}(D)\|_2^2 \le 0.\end{aligned}$$

Thus the homogeneous solution operator is nonexpansive in this norm. Since $D_0 = 0$, variation of constants gives

$$\|D_t\|_{\mathcal{G}_0^{-1}} \le \int_0^t \|E_s\|_{\mathcal{G}_0^{-1}}\, ds.$$

Using

$$\|N\|_{\mathrm{F}} \le \|\mathcal{G}_0\|_{\mathrm{op}}^{1/2}\|N\|_{\mathcal{G}_0^{-1}}, \qquad \|N\|_{\mathcal{G}_0^{-1}} \le \big\|\mathcal{G}_0^{-1}\big\|_{\mathrm{op}}^{1/2}\|N\|_{\mathrm{F}},$$

we get

$$\begin{aligned}\|D_t\|_{\mathrm{F}} &\le \sqrt{\kappa_{\mathrm{init}}}\int_0^t \|E_s\|_{\mathrm{F}}\, ds \\ &\le \sqrt{\kappa_{\mathrm{init}}}\, q_\alpha\big(2\sqrt{p_+} + q_\alpha\big)\int_0^\infty \|\nabla\ell(M_s)\|_{\mathrm{F}}\, ds \\ &\le \frac{4\sqrt{\kappa_{\mathrm{init}}}}{g_-\mu_{\mathcal{A}}^-}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\,\|\mathcal{A}(M_0) - y\|_2\, q_\alpha\big(2\sqrt{p_+} + q_\alpha\big),\end{aligned}$$

where the last line uses (48). Passing to the limit proves (45). □

A.3 Deterministic gradient descent: uniform control through convergence

Now return to the discrete deterministic algorithm from Theorem 1, whose factor step is η. For the analysis in this subsection, set

$$\gamma := 2\alpha^2\eta. \tag{52}$$

Thus γ is the effective matrix-level step, whereas η remains the actual factor step used by the algorithm. Let $\bar{W}_k = W_k/\alpha$ and $M_k = U_k V_k^{\mathsf{T}}$. A single update gives

$$\bar{W}_{k+1} = \left[I - \frac{\gamma}{2\alpha^2}\begin{pmatrix} 0 & \nabla\ell(M_k) \\ \nabla\ell(M_k)^{\mathsf{T}} & 0\end{pmatrix}\right]\bar{W}_k. \tag{53}$$

Define

$$P_k = \alpha^{-2}U_k U_k^{\mathsf{T}}, \qquad Q_k = \alpha^{-2}V_k V_k^{\mathsf{T}}.$$

Expanding the product $U_{k+1}V_{k+1}^{\mathsf{T}}$ without dropping the quadratic term gives

$$M_{k+1} = M_k - \gamma \mathcal{G}_0 \nabla \ell(M_k) + E_k, \tag{54}$$

where

$$\begin{aligned} E_k = & -\frac{\gamma}{2}\Big((P_k - P_0)\nabla\ell(M_k) + \nabla\ell(M_k)(Q_k - Q_0)\Big) \\ & + \frac{\gamma^2}{4\alpha^2}\nabla\ell(M_k)(V_k/\alpha)(U_k/\alpha)^{\mathsf{T}}\nabla\ell(M_k). \end{aligned} \tag{55}$$

Proposition 9 (The actual deterministic iterates stay close to the linearized iteration)**.** *Assume*

$$0 < \gamma \le (\kappa_{\text{init}} g_- - \mu_{\mathcal{A}}^+)^{-1}, \qquad \frac{\|M_0\|_{\text{op}}}{\alpha^2} \le \frac{1}{4}, \qquad q_\alpha \le \frac{1}{4}, \tag{56}$$

and

$$q_\alpha\big(2\sqrt{p_+} + q_\alpha\big) + \frac{\gamma\left(\sqrt{p_+} + \frac{1}{2}\right)^2}{4\alpha^2}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\,\|\mathcal{A}(M_0) - y\|_2 \le \frac{g_- - \mu_{\mathcal{A}}^-}{2\mu_{\mathcal{A}}^+}. \tag{57}$$

Then

$$\|\mathcal{A}(M_k) - y\|_2 \le \left(1 - \frac{\gamma g_- - \mu_{\mathcal{A}}^-}{2}\right)^k \|\mathcal{A}(M_0) - y\|_2, \tag{58}$$

$$\sup_{k\ge 0}\left\|\bar{W}_k - \bar{W}_0\right\|_{\text{op}} \le q_\alpha. \tag{59}$$

The factors converge to an exact-fit endpoint with full-row-rank stack. Denote its product by $\widehat{M}_{\text{GD}}(\alpha, \eta)$*. Then*

$$\begin{aligned} \left\|\widehat{M}_{\text{GD}}(\alpha,\eta) - M_{\text{lin}}\right\|_{\text{F}} \le & \frac{2\sqrt{\kappa_{\text{init}}}}{g_- - \mu_{\mathcal{A}}^-}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\,\|\mathcal{A}(M_0) - y\|_2 \\ & \times \Bigg[q_\alpha\big(2\sqrt{p_+} + q_\alpha\big) \\ & \quad + \frac{\gamma\left(\sqrt{p_+} + \frac{1}{2}\right)^2}{4\alpha^2}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\,\|\mathcal{A}(M_0) - y\|_2\Bigg]. \end{aligned} \tag{60}$$

Proof. Set $r_k = \mathcal{A}(M_k) - y$. We prove simultaneously, by induction on k, the two bounds in (58)–(59).

At $k = 0$, $r_0 = \mathcal{A}(M_0) - y$ and $\bar{W}_0 - \bar{W}_0 = 0$, so both statements are true. Assume they hold through index k.

Step 1: bound the error in the product iteration. For every $j \le k$, the two block rows satisfy

$$\|U_j/\alpha\|_{\text{op}}, \|V_j/\alpha\|_{\text{op}} \le \sqrt{p_+} + q_\alpha \le \sqrt{p_+} + \frac{1}{2}.$$

By Lemma 7,

$$\frac{1}{2}\big(\|P_j - P_0\|_{\text{op}} + \|Q_j - Q_0\|_{\text{op}}\big) \le q_\alpha\big(2\sqrt{p_+} + q_\alpha\big).$$

Also

$$\begin{aligned} \|\nabla\ell(M_j)\|_{\text{F}} &= \frac{1}{n}\|\mathcal{A}^* r_j\|_{\text{F}} \\ &\le \sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\,\|r_j\|_2 \\ &\le \sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\,\|\mathcal{A}(M_0) - y\|_2\left(1 - \frac{\gamma g_- - \mu_{\mathcal{A}}^-}{2}\right)^j. \end{aligned}$$

The first line of (55) is therefore at most

$$\gamma q_\alpha \big(2\sqrt{p_+} + q_\alpha\big) \, \|\nabla \ell(M_j)\|_{\mathrm{F}} \, .$$

For the quadratic line, use

$$\|U_j/\alpha\|_{\mathrm{op}} \, , \|V_j/\alpha\|_{\mathrm{op}} \leq \sqrt{p_+} + \frac{1}{2}$$

and

$$\|\nabla \ell(M_j)\|_{\mathrm{op}} \leq \|\nabla \ell(M_j)\|_{\mathrm{F}} \leq \sqrt{\frac{\mu_{\mathcal{A}}^+}{n}} \, \|\mathcal{A}(M_0) - y\|_2 \, .$$

Then

$$\frac{\gamma^2}{4\alpha^2} \left\|\nabla \ell(M_j)(V_j/\alpha)(U_j/\alpha)^{\mathsf{T}} \nabla \ell(M_j)\right\|_{\mathrm{F}} \leq \gamma \frac{\gamma \left(\sqrt{p_+} + \frac{1}{2}\right)^2}{4\alpha^2} \sqrt{\frac{\mu_{\mathcal{A}}^+}{n}} \, \|\mathcal{A}(M_0) - y\|_2 \, \|\nabla \ell(M_j)\|_{\mathrm{F}} \, .$$

Consequently

$$\|E_j\|_{\mathrm{F}} \leq \gamma \left[q_\alpha \big(2\sqrt{p_+} + q_\alpha\big) + \frac{\gamma \left(\sqrt{p_+} + \frac{1}{2}\right)^2}{4\alpha^2} \sqrt{\frac{\mu_{\mathcal{A}}^+}{n}} \, \|\mathcal{A}(M_0) - y\|_2 \right] \|\nabla \ell(M_j)\|_{\mathrm{F}} \, . \tag{61}$$

Step 2: the residual contracts. Applying $\mathcal{A}$ to (54),

$$r_{k+1} = \left(I_n - \frac{\gamma}{n} \mathcal{A}\mathcal{G}_0 \mathcal{A}^*\right) r_k + \mathcal{A}(E_k).$$

The eigenvalues of $\mathcal{A}\mathcal{G}_0\mathcal{A}^*/n$ lie in $[g_- \mu_{\mathcal{A}}^-, \kappa_{\mathrm{init}} g_- \mu_{\mathcal{A}}^+]$. Since $\gamma \leq (\kappa_{\mathrm{init}} g_- \mu_{\mathcal{A}}^+)^{-1}$,

$$\left\| I_n - \frac{\gamma}{n} \mathcal{A}\mathcal{G}_0 \mathcal{A}^* \right\|_{\mathrm{op}} \leq 1 - \gamma g_- \mu_{\mathcal{A}}^- .$$

Furthermore

$$\|\mathcal{A}(E_k)\|_2 \leq \sqrt{n \mu_{\mathcal{A}}^+} \, \|E_k\|_{\mathrm{F}} \, .$$

Combining this with (61) and $\|\nabla \ell(M_k)\|_{\mathrm{F}} \leq \sqrt{\mu_{\mathcal{A}}^+/n} \, \|r_k\|_2$ gives

$$\|\mathcal{A}(E_k)\|_2 \leq \gamma \mu_{\mathcal{A}}^+ \left[q_\alpha \big(2\sqrt{p_+} + q_\alpha\big) + \frac{\gamma \left(\sqrt{p_+} + \frac{1}{2}\right)^2}{4\alpha^2} \sqrt{\frac{\mu_{\mathcal{A}}^+}{n}} \, \|\mathcal{A}(M_0) - y\|_2 \right] \|r_k\|_2 \, .$$

By (57), the bracket multiplied by $\mu_{\mathcal{A}}^+$ is at most $g_- \mu_{\mathcal{A}}^-/2$. Hence

$$\|r_{k+1}\|_2 \leq \left(1 - \gamma g_- \mu_{\mathcal{A}}^- + \frac{\gamma g_- \mu_{\mathcal{A}}^-}{2}\right) \|r_k\|_2 = \left(1 - \frac{\gamma g_- \mu_{\mathcal{A}}^-}{2}\right) \|r_k\|_2 \, .$$

This proves the residual induction step. Summing the resulting geometric series gives a bound that will be used twice:

$$\begin{aligned} \sum_{j=0}^{\infty} \gamma \, \|\nabla \ell(M_j)\|_{\mathrm{F}} &\leq \gamma \sqrt{\frac{\mu_{\mathcal{A}}^+}{n}} \, \|\mathcal{A}(M_0) - y\|_2 \sum_{j=0}^{\infty} \left(1 - \frac{\gamma g_- \mu_{\mathcal{A}}^-}{2}\right)^j \\ &= \frac{2}{g_- \mu_{\mathcal{A}}^-} \sqrt{\frac{\mu_{\mathcal{A}}^+}{n}} \, \|\mathcal{A}(M_0) - y\|_2 \, . \end{aligned} \tag{62}$$

Step 3: the factors cannot drift far. From (53),

$$\left\|\bar{W}_{k+1} - \bar{W}_0\right\|_{\mathrm{op}} \leq \sum_{j=0}^{k} \left\|\bar{W}_{j+1} - \bar{W}_j\right\|_{\mathrm{op}}$$

$$\leq \frac{\sqrt{2p_+}+\frac{1}{2}}{2\alpha^2}\sum_{j=0}^{k}\gamma\left\|\nabla\ell(M_j)\right\|_{\mathrm{F}}$$
$$\leq \frac{\sqrt{2p_+}+\frac{1}{2}}{\alpha^2 g_- \mu_{\mathcal{A}}^-}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\left\|\mathcal{A}(M_0)-y\right\|_2$$
$$= \frac{1}{2}q_\alpha \leq q_\alpha.$$

This closes the induction.

Step 4: convergence and exact fitting. Equation (62) and the update (53) imply

$$\sum_{j=0}^{\infty}\left\|\bar{W}_{j+1}-\bar{W}_j\right\|_{\mathrm{op}} < \infty.$$

Thus $\bar{W}_k$ converges, and so do U_k and V_k. The residual bound tends to zero, so the limiting product is an exact fit. Moreover, (56), (59), and Lemma 7 imply that the endpoint stack has full row rank.

Step 5: comparison with the linearized discrete iteration. Let

$$\overline{M}_{k+1} = \overline{M}_k - \gamma\mathcal{G}_0\nabla\ell(\overline{M}_k), \qquad \overline{M}_0 = M_0.$$

Its residual is multiplied by $I_n - \gamma\mathcal{A}\mathcal{G}_0\mathcal{A}^*/n$, whose eigenvalues lie in $[0, 1-\gamma g_-\mu_{\mathcal{A}}^-]$. Hence the residual converges to zero. Every $\overline{M}_k - M_0$ belongs to the range of $\mathcal{G}_0\mathcal{A}^*$, so, exactly as in the flow argument, $\overline{M}_k \to M_{\mathrm{lin}}$.

In the $\mathcal{G}_0^{-1}$ norm, the linear update $I-\gamma\mathcal{G}_0\mathcal{A}^*\mathcal{A}/n$ is nonexpansive. To see this directly, conjugate by $\mathcal{G}_0^{1/2}$: it becomes

$$I - \frac{\gamma}{n}\mathcal{G}_0^{1/2}\mathcal{A}^*\mathcal{A}\mathcal{G}_0^{1/2},$$

a symmetric matrix whose eigenvalues lie in $[0,1]$ because $\gamma \leq (\kappa_{\mathrm{init}} g_- \mu_{\mathcal{A}}^+)^{-1}$. Iterating the difference equation from (54) therefore gives

$$\left\|M_k - \overline{M}_k\right\|_{\mathrm{F}} \leq \sqrt{\kappa_{\mathrm{init}}}\sum_{j=0}^{k-1}\left\|E_j\right\|_{\mathrm{F}}.$$

Use (61) and then (62). The resulting bound is uniform in k, and passing to the limit gives exactly (60). □

A.4 The Gaussian row space misses the initial displacement by a constant fraction

After vectorization, the sensing map is represented by an $n \times (d_1 d_2)$ standard Gaussian matrix. Its row space, denoted by S, is a uniformly random n-dimensional subspace of $\mathbb{R}^{d_1 d_2}$.

Proposition 10 (Lower bound for the linearized interpolator)**.** *For every fixed initialization (U_0, V_0), chosen independently of the sensing matrices, and every $1 \leq n < d_1 d_2$, with probability at least*

$$1 - 2e^{-c(d_1 d_2 - n)} - e^{-cd_1 d_2}, \tag{63}$$

one has

$$\left\|M_{\mathrm{lin}} - M_\star\right\|_{\mathrm{F}} \geq \sqrt{\frac{d_1 d_2 - n}{4 d_1 d_2}}\,\frac{\left\|M_\star - M_0\right\|_{\mathrm{F}}}{\kappa_{\mathrm{init}}}. \tag{64}$$

In particular, if $n \leq d_1 d_2/2$,

$$\left\|M_{\mathrm{lin}} - M_\star\right\|_{\mathrm{F}} \geq \frac{1}{\sqrt{8}\,\kappa_{\mathrm{init}}}\left\|M_\star - M_0\right\|_{\mathrm{F}}. \tag{65}$$

Proof. If $M_0 = M_\star$, both displayed lower bounds are zero and there is nothing to prove. Assume therefore that $M_0 \neq M_\star$.

By (32), $M_{\mathrm{lin}} - M_0$ belongs to the subspace $\mathcal{G}_0\mathsf{S}$. Hence

$$\|M_{\mathrm{lin}} - M_\star\|_{\mathrm{F}} \geq \operatorname{dist}(M_\star - M_0, \mathcal{G}_0\mathsf{S}) \tag{66}$$

$$= \inf_{Y \in \mathsf{S}} \left\|\mathcal{G}_0\big(\mathcal{G}_0^{-1}(M_\star - M_0) - Y\big)\right\|_{\mathrm{F}} \tag{67}$$

$$\geq \lambda_{\min}(\mathcal{G}_0) \operatorname{dist}(\mathcal{G}_0^{-1}(M_\star - M_0), \mathsf{S}). \tag{68}$$

The matrix $\mathcal{G}_0^{-1}(M_\star - M_0)$ is fixed before the Gaussian row space is drawn. Rotational invariance therefore implies

$$\frac{\left\|P_{\mathsf{S}^\perp}\mathcal{G}_0^{-1}(M_\star - M_0)\right\|_{\mathrm{F}}^2}{\left\|\mathcal{G}_0^{-1}(M_\star - M_0)\right\|_{\mathrm{F}}^2} \overset{d}{=} \frac{X_1}{X_1 + X_2},$$

where $X_1 \sim \chi^2_{d_1 d_2 - n}$ and $X_2 \sim \chi^2_n$ are independent. Standard chi-square concentration gives, outside an event of probability at most $2e^{-c(d_1 d_2 - n)} + e^{-c d_1 d_2}$,

$$X_1 \geq \frac{d_1 d_2 - n}{2}, \qquad X_1 + X_2 \leq 2 d_1 d_2.$$

Thus

$$\begin{aligned}\operatorname{dist}(\mathcal{G}_0^{-1}(M_\star - M_0), \mathsf{S}) &\geq \sqrt{\frac{d_1 d_2 - n}{4 d_1 d_2}} \left\|\mathcal{G}_0^{-1}(M_\star - M_0)\right\|_{\mathrm{F}} \\ &\geq \sqrt{\frac{d_1 d_2 - n}{4 d_1 d_2}} \frac{\|M_\star - M_0\|_{\mathrm{F}}}{\|\mathcal{G}_0\|_{\mathrm{op}}}.\end{aligned}$$

Combining this with (68) gives

$$\|M_{\mathrm{lin}} - M_\star\|_{\mathrm{F}} \geq \sqrt{\frac{d_1 d_2 - n}{4 d_1 d_2}} \frac{\lambda_{\min}(\mathcal{G}_0)}{\|\mathcal{G}_0\|_{\mathrm{op}}} \|M_\star - M_0\|_{\mathrm{F}}.$$

The ratio is $1/\kappa_{\mathrm{init}}$ by (4), proving (64). If $n \leq d_1 d_2/2$, then $(d_1 d_2 - n)/(4 d_1 d_2) \geq 1/8$, which gives (65). □

A.5 Proof of the one-layer neural network lower bound

Proposition 11. *Consider the setting of Theorem 1, and consider label-noise SGD applied directly to M,*

$$M_{k+1} = M_k - \eta \nabla \ell_{i_k, \xi_k}(M_k), \qquad \ell_{i,\xi}(M) = \frac{1}{2}\big(\langle A_i, M\rangle - y_i - \xi\big)^2,$$

where i_k is sampled uniformly from $\{1, \dots, n\}$ and $\xi_k \in \{-\sigma, +\sigma\}$. There exist numerical constants $c, C > 0$ such that, with probability at least $1 - Ce^{-cn}$ over the sensing matrices, for any step-size η,

$$\|M_k - M_\star\|_{\mathrm{F}} \geq \frac{1}{\sqrt{8}} \|M_0 - M_\star\|_{\mathrm{F}} \qquad \textit{for all } k \geq 0.$$

Proof. Let

$$\mathsf{S} = \operatorname{span}\{A_1, \dots, A_n\}.$$

Since

$$\nabla \ell_{i,\xi}(M) = \big(\langle A_i, M\rangle - y_i - \xi\big) A_i \in \mathsf{S},$$

every update lies in S. Hence, pathwise for every k,

$$P_{\mathsf{S}^\perp}(M_k - M_\star) = P_{\mathsf{S}^\perp}(M_0 - M_\star),$$

where $P_{\mathsf{S}^\perp}$ denotes the orthogonal projection onto $S^\perp$. Therefore

$$\|M_k - M_\star\|_{\mathrm{F}} \geq \|P_{\mathsf{S}^\perp}(M_0 - M_\star)\|_{\mathrm{F}}. \tag{69}$$

Set $D = d_1 d_2$ and $H = M_0 - M_\star$. If $H = 0$, the claim is immediate. Otherwise, S is a uniformly random n-dimensional subspace of $\mathbb{R}^D$ and

$$\frac{\|P_{\mathsf{S}^\perp} H\|_{\mathrm{F}}^2}{\|H\|_{\mathrm{F}}^2} \overset{d}{=} \frac{X_1}{X_1 + X_2}, \qquad X_1 \sim \chi^2_{D-n}, \quad X_2 \sim \chi^2_n,$$

with X_1 and X_2 independent. Standard chi-square concentration implies, with probability at least $1 - 2e^{-c(D-n)} - e^{-cn} \geqslant 1 - Ce^{-cn}$, we have

$$X_1 \geq \frac{D-n}{2}, \qquad X_1 + X_2 \leq 2D.$$

Since $n \leq D/2$,

$$\|P_{\mathsf{S}^\perp} H\|_{\mathrm{F}} \geq \sqrt{\frac{D-n}{4D}}\, \|H\|_{\mathrm{F}} \geq \frac{1}{\sqrt{8}}\, \|H\|_{\mathrm{F}}\,.$$

□

A.6 Conclusion of the deterministic branch

The preceding propositions are deterministic. We now record the single Gaussian estimate needed to choose the initialization scale uniformly over the sensing instance.

Lemma 12 (Gaussian conditioning for the deterministic branch)**.** *Assume $n \leq d_1 d_2/2$. There are numerical constants $c, C > 0$ such that, for every fixed $M_0, M_\star$, with probability at least $1 - Ce^{-cn}$,*

$$c\,\frac{d_1 d_2}{n} \leq \mu_{\mathcal{A}}^- \leq \mu_{\mathcal{A}}^+ \leq C\,\frac{d_1 d_2}{n}, \tag{70}$$

and

$$\frac{1}{\mu_{\mathcal{A}}^-}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\, \|\mathcal{A}(M_0) - y\|_2 \leq C\sqrt{\frac{n}{d_1 d_2}}\, \|M_\star - M_0\|_{\mathrm{F}}\,. \tag{71}$$

The event depends only on the sensing matrices and the fixed displacement $M_\star - M_0$, not on the initialization scale.

Proof. After vectorization, write the sensing map as an $n \times (d_1 d_2)$ standard Gaussian matrix G. Since $n \leq d_1 d_2/2$, the Gaussian singular-value bound in Proposition 41, applied with $t = (\sqrt{d_1 d_2} - \sqrt{n})/2$, gives

$$c\sqrt{d_1 d_2} \leq \sigma_{\min}(G) \leq \sigma_{\max}(G) \leq C\sqrt{d_1 d_2}$$

outside an event of probability at most $Ce^{-cd_1 d_2}$. Dividing the squared singular values by n proves (70).

Let $H = M_0 - M_\star$. If $H = 0$, then $\mathcal{A}(M_0) - y = 0$ and (71) is immediate. Otherwise, $G\,\mathrm{vec}(H)/\|H\|_{\mathrm{F}}$ is a standard Gaussian vector in $\mathbb{R}^n$, so standard chi-square concentration gives

$$\|\mathcal{A}(H)\|_2 \leq C\sqrt{n}\, \|H\|_{\mathrm{F}}$$

outside an event of probability at most Ce^{-cn}. Combining this with (70) and $y = \mathcal{A}(M_\star)$ proves (71). Since $d_1 d_2 \geq 2n$, the total failure probability is Ce^{-cn} after changing the numerical constants. □

Fix now a sensing instance for which $\mathcal{A}$ has full row rank and both Proposition 10 and Lemma 12 hold. Define the factor-step threshold

$$\eta_{\mathrm{GD}}(\alpha) := \frac{1}{2\alpha^2 \kappa_{\mathrm{init}} g_- \mu_{\mathcal{A}}^+}. \tag{72}$$

We show that, uniformly on the Gaussian conditioning event, all hypotheses of Propositions 8 and 9 hold once α is sufficiently large. Put $\Delta_0 := \|M_\star - M_0\|_{\mathrm{F}}$. By Lemma 12,

$$\frac{\mu_{\mathcal{A}}^-}{\mu_{\mathcal{A}}^+} \geq c, \qquad \frac{1}{\mu_{\mathcal{A}}^-}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\, \|\mathcal{A}(M_0) - y\|_2 \leq C\sqrt{\frac{n}{d_1 d_2}}\, \Delta_0. \tag{73}$$

Using $p_+ \le 2\kappa_{\mathrm{init}} g_-$ in the definition of q_α gives

$$q_\alpha \le \frac{C\sqrt{\kappa_{\mathrm{init}}}}{\alpha^2\sqrt{g_-}}\sqrt{\frac{n}{d_1 d_2}}\,\Delta_0 \le \frac{C\sqrt{\kappa_{\mathrm{init}}}}{\alpha^2}\sqrt{\frac{n}{d_1 d_2}}\,\Delta_0. \tag{74}$$

In particular, $q_\alpha \le 1/4$ for all sufficiently large α. Moreover, since $p_+ \ge 1$,

$$q_\alpha\big(2\sqrt{p_+} + q_\alpha\big) \le \frac{C\kappa_{\mathrm{init}}}{\alpha^2}\sqrt{\frac{n}{d_1 d_2}}\,\Delta_0, \tag{75}$$

where the factor $\sqrt{g_-}$ from $\sqrt{p_+}$ cancels the $g_-^{-1/2}$ in the first bound in (74).

Now let $0 < \eta \le \eta_{\mathrm{GD}}(\alpha)$ and set $\gamma = 2\alpha^2\eta$. Then $\gamma \le (\kappa_{\mathrm{init}} g_- \mu_{\mathcal{A}}^+)^{-1}$. The second term in the perturbation bracket in (57) satisfies

$$\begin{aligned}&\frac{\gamma\left(\sqrt{p_+} + \frac{1}{2}\right)^2}{4\alpha^2}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\;\|\mathcal{A}(M_0) - y\|_2 \\ &\quad\le \frac{C}{\alpha^2\mu_{\mathcal{A}}^+}\sqrt{\frac{\mu_{\mathcal{A}}^+}{n}}\;\|\mathcal{A}(M_0) - y\|_2 \le \frac{C}{\alpha^2}\sqrt{\frac{n}{d_1 d_2}}\,\Delta_0.\end{aligned} \tag{76}$$

Here we used $(\sqrt{p_+} + 1/2)^2 \le C\kappa_{\mathrm{init}} g_-$ and $\mu_{\mathcal{A}}^- \le \mu_{\mathcal{A}}^+$ in the last step. Combining (75) and (76), the entire bracket in (57) is at most

$$\frac{C\kappa_{\mathrm{init}}}{\alpha^2}\sqrt{\frac{n}{d_1 d_2}}\,\Delta_0, \tag{77}$$

uniformly for $0 < \eta \le \eta_{\mathrm{GD}}(\alpha)$.

Because $g_- \ge 1$ and $\mu_{\mathcal{A}}^-/\mu_{\mathcal{A}}^+ \ge c$, the bounds above show that, for all sufficiently large α,

$$\frac{\|M_0\|_{\mathrm{op}}}{\alpha^2} \le \frac{1}{4}, \qquad q_\alpha \le \frac{1}{4},$$

and (57) holds for every $0 < \eta \le \eta_{\mathrm{GD}}(\alpha)$. Thus Propositions 8 and 9 both apply, and deterministic GD converges to an exact fit. In addition, (60), (73), and (77) give

$$\left\|\widehat{M}_{\mathrm{GD}}(\alpha,\eta) - M_{\mathrm{lin}}\right\|_{\mathrm{F}} \le \frac{C\kappa_{\mathrm{init}}^{3/2}}{\alpha^2}\frac{n}{d_1 d_2}\,\Delta_0^2, \tag{78}$$

uniformly over the same step-size interval. If $\Delta_0 > 0$, enlarge the initialization threshold so that the right side is at most

$$\frac{1}{2\sqrt{8}\,\kappa_{\mathrm{init}}}\Delta_0. \tag{79}$$

If $\Delta_0 = 0$, this endpoint error is already zero.

Finally, if $M_0 = M_\star$, the lower bound in Theorem 1 is zero. Otherwise, the reverse triangle inequality, (65), and (79) give

$$\begin{aligned}\left\|\widehat{M}_{\mathrm{GD}}(\alpha,\eta) - M_\star\right\|_{\mathrm{F}} &\ge \|M_{\mathrm{lin}} - M_\star\|_{\mathrm{F}} - \left\|\widehat{M}_{\mathrm{GD}}(\alpha,\eta) - M_{\mathrm{lin}}\right\|_{\mathrm{F}} \\ &\ge \frac{1}{\sqrt{8}\,\kappa_{\mathrm{init}}}\Delta_0 - \frac{1}{2\sqrt{8}\,\kappa_{\mathrm{init}}}\Delta_0 \\ &= \frac{1}{4\sqrt{2}\,\kappa_{\mathrm{init}}}\Delta_0.\end{aligned}$$

This is part (a) of Theorem 1.

All bounds used to choose the initialization threshold depend only on κ_{init} and the fixed problem parameters, not on the realized sensing matrices. This is the meaning of "α_{init} sufficiently large" in Theorem 1. The cancellations in (75) and (78) are what remove any separate dependence on the absolute size of the normalized row Grams.

A.7 CONSTRUCTING LARGE INITIALIZATIONS

The theorem allows an arbitrary prescribed initial product M_0. The next lemma shows that such a product can always be represented with arbitrarily large intrinsic initialization scale and controlled spectral shape.

Lemma 13 (Existence of arbitrarily large initializations)**.** *Let $M_0 \in \mathbb{R}^{d_1 \times d_2}$ and $m \geq d_1 + d_2$. For every $a > 0$ there exist matrices $U_0 \in \mathbb{R}^{d_1 \times m}$ and $V_0 \in \mathbb{R}^{d_2 \times m}$ whose stack $W_0 = \binom{U_0}{V_0}$ has full row rank and such that*

$$U_0 V_0^\mathsf{T} = M_0, \qquad \alpha_{\text{init}} = a.$$

For the normalization (26),

$$P_0 = I_{d_1}, \qquad Q_0 = I_{d_2} + a^{-4} M_0^\mathsf{T} M_0, \tag{80}$$

and in particular

$$\kappa_{\text{init}} \leq 1 + \frac{\|M_0\|_{\text{op}}^2}{2a^4}. \tag{81}$$

Hence $\kappa_{\text{init}} \to 1$ as $a \to \infty$.

Proof. Take

$$U_0 = \begin{bmatrix} aI_{d_1} & 0 & 0 \end{bmatrix}, \qquad V_0 = \begin{bmatrix} a^{-1} M_0^\mathsf{T} & aI_{d_2} & 0 \end{bmatrix},$$

where the final zero blocks are omitted when $m = d_1 + d_2$. Then $U_0 V_0^\mathsf{T} = M_0$ and

$$U_0 U_0^\mathsf{T} = a^2 I_{d_1}, \qquad V_0 V_0^\mathsf{T} = a^2 I_{d_2} + a^{-2} M_0^\mathsf{T} M_0 \succeq a^2 I_{d_2}.$$

The identity blocks in disjoint column sets also show that $\binom{U_0}{V_0}$ has full row rank. Thus $\alpha_{\text{init}} = a$, and division by a^2 gives (80). Finally,

$$\kappa_{\text{init}} = \frac{\lambda_{\max}(P_0) + \lambda_{\max}(Q_0)}{\lambda_{\min}(P_0) + \lambda_{\min}(Q_0)} \leq \frac{2 + a^{-4} \|M_0\|_{\text{op}}^2}{2},$$

which is (81). □

For the asymptotic discussion in section 3.1, it is also useful to know that the normalized row Grams can be held fixed while the scale grows.

Lemma 14 (Realizability at fixed normalized row Grams)**.** *Let $P_0 \succ 0$, $Q_0 \succ 0$, and $M_0 \in \mathbb{R}^{d_1 \times d_2}$, with*

$$\min\{\lambda_{\min}(P_0), \lambda_{\min}(Q_0)\} = 1.$$

If $m \geq d_1 + d_2$, then for every $\alpha > 0$ such that

$$\alpha^4 > \left\| P_0^{-1/2} M_0 Q_0^{-1/2} \right\|_{\text{op}}^2, \tag{82}$$

there exist full-row-rank factors U_0, V_0 satisfying

$$U_0 V_0^\mathsf{T} = M_0, \qquad U_0 U_0^\mathsf{T} = \alpha^2 P_0, \qquad V_0 V_0^\mathsf{T} = \alpha^2 Q_0. \tag{83}$$

For these factors, $\alpha_{\text{init}} = \alpha$.

Proof. Consider

$$K_\alpha = \begin{pmatrix} \alpha^2 P_0 & M_0 \\ M_0^\mathsf{T} & \alpha^2 Q_0 \end{pmatrix}.$$

Its lower-right block is positive definite, and its Schur complement is

$$\alpha^2 P_0 - \alpha^{-2} M_0 Q_0^{-1} M_0^\mathsf{T}.$$

Condition (82) makes this Schur complement positive definite, hence $K_\alpha \succ 0$. Factor $K_\alpha = RR^\mathsf{T}$, pad R with zero columns if $m > d_1 + d_2$, and partition its rows as $R = \binom{U_0}{V_0}$. The identities in (83) follow from $RR^\mathsf{T} = K_\alpha$. Moreover,

$$\min\{\sigma_{\min}(U_0)^2, \sigma_{\min}(V_0)^2\} = \alpha^2 \min\{\lambda_{\min}(P_0), \lambda_{\min}(Q_0)\} = \alpha^2,$$

so $\alpha_{\text{init}} = \alpha$. □

The first lemma shows that the large-initialization regime in the theorem is non-vacuous for every fixed M_0. The second justifies the fixed-Gram, large-α family used only to state the asymptotic interpretation in Proposition 2. We now turn to the noisy branch.

B THE NOISY BRANCH: THE SLOW OBJECTIVE IS A WEIGHTED NUCLEAR NORM

B.1 EXACT TRACE OF THE HESSIAN AT INTERPOLATION

Define the empirical left and right sensing marginals

$$S_L = \frac{1}{n}\sum_{i=1}^{n} A_i A_i^\mathsf{T}, \qquad S_R = \frac{1}{n}\sum_{i=1}^{n} A_i^\mathsf{T} A_i. \tag{84}$$

At an exact fit, the quantity selected by label noise will be half the Hessian trace:

$$J(U,V) = \frac{1}{2}\Big(\operatorname{tr}(U^\mathsf{T} S_L U) + \operatorname{tr}(V^\mathsf{T} S_R V)\Big). \tag{85}$$

We now derive this formula directly from the loss. Define the manifold

$$\Gamma_y = \{(U,V) : \mathcal{A}(UV^\mathsf{T}) = y\}.$$

Proposition 15 (Trace-Hessian identity)**.** *If $(U,V) \in \Gamma_y$, then*

$$\operatorname{tr}\nabla^2 L(U,V) = \operatorname{tr}(U^\mathsf{T} S_L U) + \operatorname{tr}(V^\mathsf{T} S_R V) = 2J(U,V). \tag{86}$$

Proof. For each observation write

$$r_i(U,V) = \left\langle A_i, UV^\mathsf{T}\right\rangle - y_i.$$

For perturbations ΔU and ΔV,

$$\begin{aligned} Dr_i(U,V)[\Delta U, 0] &= \left\langle A_i, \Delta U V^\mathsf{T}\right\rangle = \left\langle A_i V, \Delta U\right\rangle, \\ Dr_i(U,V)[0, \Delta V] &= \left\langle A_i, U \Delta V^\mathsf{T}\right\rangle = \left\langle A_i^\mathsf{T} U, \Delta V\right\rangle. \end{aligned}$$

Therefore

$$\nabla_U r_i = A_i V, \qquad \nabla_V r_i = A_i^\mathsf{T} U.$$

Since $L = (2n)^{-1}\sum_i r_i^2$, differentiating twice gives

$$\nabla^2 L = \frac{1}{n}\sum_{i=1}^{n}\left(\nabla r_i \nabla r_i^\mathsf{T} + r_i \nabla^2 r_i\right).$$

At interpolation all $r_i = 0$, so the second term vanishes. Hence

$$\begin{aligned} \operatorname{tr}\nabla^2 L(U,V) &= \frac{1}{n}\sum_{i=1}^{n}\left(\|A_i V\|_\mathrm{F}^2 + \left\|A_i^\mathsf{T} U\right\|_\mathrm{F}^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\operatorname{tr}(V^\mathsf{T} A_i^\mathsf{T} A_i V) + \operatorname{tr}(U^\mathsf{T} A_i A_i^\mathsf{T} U)\right) \\ &= \operatorname{tr}(V^\mathsf{T} S_R V) + \operatorname{tr}(U^\mathsf{T} S_L U), \end{aligned}$$

which is (86). □

B.2 MINIMIZING THE SHARPNESS OVER FACTORIZATIONS

The next calculation is elementary but central: after the factors are minimized out, J becomes a weighted nuclear norm of the product.

Lemma 16 (Factorization formula for the nuclear norm)**.** *For every $M \in \mathbb{R}^{d_1 \times d_2}$ and every width $m \geq \operatorname{rank} M$,*

$$\min_{UV^\mathsf{T} = M} \frac{1}{2}\left(\|U\|_\mathrm{F}^2 + \|V\|_\mathrm{F}^2\right) = \|M\|_* . \tag{87}$$

Here $\|M\|_*$ is the sum of the singular values of M.

Proof. Let $M = E\Sigma F^\mathsf{T}$ be a compact singular-value decomposition. The factors

$$U = E\Sigma^{1/2}, \qquad V = F\Sigma^{1/2}$$

can be padded with zero columns to width m and satisfy $UV^\mathsf{T} = M$. Their cost is

$$\frac{1}{2}\left(\|U\|_\mathrm{F}^2 + \|V\|_\mathrm{F}^2\right) = \operatorname{tr}\Sigma = \|M\|_* .$$

Thus the left side of (87) is at most $\|M\|_*$.

For the reverse inequality, take any factorization $M = UV^\mathsf{T}$. Let G be a matrix with $\|G\|_{\mathrm{op}} \le 1$ that attains nuclear/operator duality for M, so $\langle G, M\rangle = \|M\|_*$. Then

$$\begin{aligned}\|M\|_* &= \left\langle G, UV^\mathsf{T}\right\rangle = \langle GV, U\rangle \\ &\le \|GV\|_\mathrm{F}\,\|U\|_\mathrm{F} \le \|G\|_{\mathrm{op}}\,\|V\|_\mathrm{F}\,\|U\|_\mathrm{F} \\ &\le \|U\|_\mathrm{F}\,\|V\|_\mathrm{F} \le \frac{1}{2}\left(\|U\|_\mathrm{F}^2 + \|V\|_\mathrm{F}^2\right).\end{aligned}$$

Since this holds for every factorization, equality follows. □

Proposition 17 (The sharpness induces a weighted nuclear norm)**.** *Assume $S_L, S_R \succ 0$. For every matrix M and every $m \ge \operatorname{rank} M$,*

$$\min_{UV^\mathsf{T}=M} J(U,V) = \left\|S_L^{1/2} M S_R^{1/2}\right\|_* . \tag{88}$$

Consequently, because $m \ge d_1 + d_2$ is at least the rank of every $d_1 \times d_2$ matrix,

$$\min_{(U,V)\in\Gamma_y} J(U,V) = \min_{\mathcal{A}(M)=y} \left\|S_L^{1/2} M S_R^{1/2}\right\|_* . \tag{89}$$

Proof. Make the invertible change of variables

$$\widetilde{U} = S_L^{1/2} U, \qquad \widetilde{V} = S_R^{1/2} V.$$

Then

$$\widetilde{U}\widetilde{V}^\mathsf{T} = S_L^{1/2} M S_R^{1/2}$$

and

$$J(U,V) = \frac{1}{2}\left(\left\|\widetilde{U}\right\|_\mathrm{F}^2 + \left\|\widetilde{V}\right\|_\mathrm{F}^2\right).$$

Because multiplication by invertible matrices does not change rank, the transformed product still has rank at most m. Applying Lemma 16 proves (88). Minimizing this identity over all exact-fit products gives (89). □

C Gaussian sensing makes the weighted nuclear norm recover the target

This section proves, on a single high-probability event, that $M_\star$ is the unique minimizer of (89). The only external probabilistic input is the Gaussian matrix-deviation estimate stated in Appendix F.2; everything after the two concentration events is deterministic.

Normalize the marginal matrices by their expectations:

$$\overline{S}_L = \frac{S_L}{d_2}, \qquad \overline{S}_R = \frac{S_R}{d_1}. \tag{90}$$

Define the whitened sensing map

$$\widetilde{\mathcal{A}}(X) = \mathcal{A}(\overline{S}_L^{-1/2} X \overline{S}_R^{-1/2}). \tag{91}$$

Under the same change of variables, the target becomes

$$\widetilde{M}_\star = \overline{S}_L^{1/2} M_\star \overline{S}_R^{1/2}.$$

Since

$$S_L^{1/2} M S_R^{1/2} = \sqrt{d_1 d_2}\,\overline{S}_L^{1/2} M \overline{S}_R^{1/2},$$

the irrelevant scalar $\sqrt{d_1 d_2}$ shows that uniqueness in (89) is equivalent to uniqueness of $\widetilde{M}_\star$ for ordinary nuclear-norm minimization under the constraint $\widetilde{\mathcal{A}}(X) = \widetilde{\mathcal{A}}(\widetilde{M}_\star)$.

C.1 THE TWO GAUSSIAN CONCENTRATION EVENTS

Lemma 18 (Marginal covariance concentration)**.** *Fix $0 < \rho < 1/2$. If*

$$nd_2 \geq C\rho^{-2}d_1, \qquad nd_1 \geq C\rho^{-2}d_2,$$

then, with probability at least

$$1 - 2e^{-c\rho^2 nd_2} - 2e^{-c\rho^2 nd_1},$$

we have

$$\left\|\overline{S}_L - I_{d_1}\right\|_{\mathrm{op}} \leq \rho, \qquad \left\|\overline{S}_R - I_{d_2}\right\|_{\mathrm{op}} \leq \rho. \tag{92}$$

Proof. We prove the left inequality; the right one is identical after interchanging d_1 and d_2. Form the $d_1 \times nd_2$ Gaussian matrix

$$G_L = [A_1\ A_2\ \cdots\ A_n].$$

Then

$$\overline{S}_L = \frac{1}{nd_2} G_L G_L^{\mathsf{T}}.$$

By the standard Gaussian singular-value bound, with probability at least $1 - 2e^{-t^2/2}$,

$$\sqrt{nd_2} - \sqrt{d_1} - t \leq \sigma_{\min}(G_L) \leq \sigma_{\max}(G_L) \leq \sqrt{nd_2} + \sqrt{d_1} + t.$$

Choose $t = c_0\rho\sqrt{nd_2}$ with c_0 sufficiently small. If the sample-size inequality is satisfied with a sufficiently large numerical constant, then

$$\sqrt{d_1} + t \leq \frac{\rho}{3}\sqrt{nd_2}.$$

Hence every singular value of $G_L/\sqrt{nd_2}$ lies in $[1 - \rho/3, 1 + \rho/3]$. Squaring, and using $\rho < 1/2$,

$$1 - \rho \leq (1 - \rho/3)^2 \leq \lambda_j(\overline{S}_L) \leq (1 + \rho/3)^2 \leq 1 + \rho.$$

This is the desired left-marginal bound. The same proof for $[A_1^{\mathsf{T}}\ \cdots\ A_n^{\mathsf{T}}]$ gives the right-marginal bound, and a union bound gives the stated probability. □

In this restricted isometry property (RIP) argument, s denotes the rank level and δ the distortion parameter. A linear map $\mathcal{C}$ has rank-s restricted isometry constant at most δ if

$$(1 - \delta)\|X\|_{\mathrm{F}}^2 \leq \frac{1}{n}\|\mathcal{C}(X)\|_2^2 \leq (1 + \delta)\|X\|_{\mathrm{F}}^2 \quad \text{for every } \operatorname{rank} X \leq s. \tag{93}$$

Lemma 19 (Gaussian rank RIP)**.** *For $0 < \delta < 1$, if*

$$n \geq C\delta^{-2}s(d_1 + d_2),$$

then the Gaussian sensing map $\mathcal{A}$ satisfies (93) with probability at least $1 - 2e^{-c\delta^2 n}$.

Proof. For a standard Gaussian matrix G, von Neumann's trace inequality and Cauchy–Schwarz give, for every X with $\operatorname{rank} X \leq s$ and $\|X\|_{\mathrm{F}} = 1$,

$$\langle G, X\rangle \leq \sum_{j=1}^{s} \sigma_j(G)\sigma_j(X) \leq \sqrt{s}\ \|G\|_{\mathrm{op}}.$$

Taking expectations and using $\mathbb{E}\,\|G\|_{\mathrm{op}} \leq \sqrt{d_1} + \sqrt{d_2}$,

$$\mathbb{E} \sup_{\substack{\operatorname{rank} X \leq s \\ \|X\|_{\mathrm{F}}=1}} \langle G, X\rangle \leq \sqrt{s}(\sqrt{d_1} + \sqrt{d_2}). \tag{94}$$

The matrix-deviation inequality for the vectorized Gaussian sensing matrix then gives, with probability at least $1 - 2e^{-u^2}$,

$$\sup_{\substack{\operatorname{rank} X \leq s \\ \|X\|_{\mathrm{F}}=1}} \left|\|\mathcal{A}(X)\|_2 - \sqrt{n}\right| \leq C\left(\sqrt{s}(\sqrt{d_1} + \sqrt{d_2}) + u\right).$$

Take $u = c_0\delta\sqrt{n}$. Under the stated sample-size condition, the right side is at most $(\delta/3)\sqrt{n}$. Hence

$$1 - \frac{\delta}{3} \leq \frac{\|\mathcal{A}(X)\|_2}{\sqrt{n}} \leq 1 + \frac{\delta}{3} \qquad (\operatorname{rank} X \leq s,\ \|X\|_{\mathrm{F}} = 1).$$

Squaring gives (93) for unit-Frobenius matrices of rank at most s, and homogeneity gives it for every rank-at-most-s matrix. □

C.2 WHITENING PRESERVES THE REQUIRED RESTRICTED ISOMETRY

Lemma 20 (RIP after data-dependent whitening)**.** *Suppose (92) holds and $\mathcal{A}$ has rank-s RIP constant δ. Then every rank-at-most-s matrix X satisfies*

$$\frac{1-\delta}{(1+\rho)^2}\,\|X\|_{\mathrm{F}}^2 \le \frac{1}{n}\left\|\widetilde{\mathcal{A}}(X)\right\|_2^2 \le \frac{1+\delta}{(1-\rho)^2}\,\|X\|_{\mathrm{F}}^2\,. \tag{95}$$

In particular, with $\rho=\delta=1/20$,

$$\frac{4}{5}\,\|X\|_{\mathrm{F}}^2 \le \frac{1}{n}\left\|\widetilde{\mathcal{A}}(X)\right\|_2^2 \le \frac{6}{5}\,\|X\|_{\mathrm{F}}^2\,, \tag{96}$$

so $\widetilde{\mathcal{A}}$ has rank-s RIP constant at most $1/5$.

Proof. Put

$$Y = \overline{S}_L^{-1/2} X \overline{S}_R^{-1/2}.$$

The transformation is invertible and preserves rank. On (92), the eigenvalues of both normalized marginals lie in $[1-\rho, 1+\rho]$, so

$$\frac{1}{1+\rho}\,\|X\|_{\mathrm{F}} \le \|Y\|_{\mathrm{F}} \le \frac{1}{1-\rho}\,\|X\|_{\mathrm{F}}\,.$$

Since $\widetilde{\mathcal{A}}(X) = \mathcal{A}(Y)$, applying the RIP of $\mathcal{A}$ to Y yields (95). Substituting $\rho=\delta=1/20$ gives bounds stronger than $4/5$ and $6/5$, which proves (96). □

C.3 RIP IMPLIES THE STRICT NULL-SPACE INEQUALITY NEEDED FOR NUCLEAR-NORM RECOVERY

Let $\widetilde{M}_\star = E\Sigma F^{\mathsf{T}}$ be a compact singular-value decomposition. Its rank-r tangent space is

$$\mathcal{T} = \{EB^{\mathsf{T}} + CF^{\mathsf{T}} : B \in \mathbb{R}^{d_2\times r},\ C \in \mathbb{R}^{d_1\times r}\}. \tag{97}$$

Every element of $\mathcal{T}$ has rank at most $2r$.

Lemma 21 (Restricted orthogonality)**.** *If a map $\mathcal{C}$ has rank-s RIP constant δ and X, Y are Frobenius-orthogonal with $\operatorname{rank} X + \operatorname{rank} Y \le s$, then*

$$\frac{1}{n}\,|\langle \mathcal{C}(X), \mathcal{C}(Y)\rangle| \le \delta\,\|X\|_{\mathrm{F}}\,\|Y\|_{\mathrm{F}}\,. \tag{98}$$

Proof. If either matrix is zero there is nothing to prove. Otherwise normalize them to unit Frobenius norm. Orthogonality gives $\left\|\widehat{X}+\widehat{Y}\right\|_{\mathrm{F}}^2 = \left\|\widehat{X}-\widehat{Y}\right\|_{\mathrm{F}}^2 = 2$, while the rank of each sum is at most s. The polarization identity gives

$$\frac{4}{n}\left\langle \mathcal{C}(\widehat{X}), \mathcal{C}(\widehat{Y})\right\rangle = \frac{1}{n}\left\|\mathcal{C}(\widehat{X}+\widehat{Y})\right\|_2^2 - \frac{1}{n}\left\|\mathcal{C}(\widehat{X}-\widehat{Y})\right\|_2^2.$$

Each squared norm lies in $[2(1-\delta), 2(1+\delta)]$, so the absolute value of their difference is at most 4δ. Multiply back by the two Frobenius norms to obtain (98). □

Lemma 22 (Rank-$6r$ RIP gives a strict nuclear null-space property)**.** *Suppose a map has rank-$6r$ RIP constant $\delta < 1/2$; if $6r$ exceeds the largest possible matrix rank, it is enough that the same RIP bound hold for all matrices. Then every nonzero Δ in its kernel satisfies*

$$\|P_{\mathcal{T}}\Delta\|_* < \|P_{\mathcal{T}^\perp}\Delta\|_*\,. \tag{99}$$

Proof. Write

$$\Delta_0 = P_{\mathcal{T}}\Delta.$$

Then $\operatorname{rank}\Delta_0 \le 2r$. Take a singular-value decomposition of the normal part $P_{\mathcal{T}^\perp}\Delta$ and partition its singular values, in decreasing order, into consecutive groups of at most $2r$ singular values. Let the

matrices formed by these groups be $\Delta_1, \Delta_2, \ldots$. They are mutually Frobenius-orthogonal, each has rank at most $2r$, and each is orthogonal to Δ_0.

Set

$$\Delta_{01} = \Delta_0 + \Delta_1.$$

Then $\operatorname{rank} \Delta_{01} \le 4r$. Since $\mathcal{C}(\Delta) = 0$,

$$\mathcal{C}(\Delta_{01}) = -\sum_{j \ge 2} \mathcal{C}(\Delta_j).$$

Take the inner product with $\mathcal{C}(\Delta_{01})/n$. The RIP lower bound applies to Δ_{01}; restricted orthogonality applies between Δ_{01} and each Δ_j, because their ranks sum to at most $6r$. Thus

$$(1-\delta)\,\|\Delta_{01}\|_{\mathrm{F}}^2 \le \frac{1}{n}\,\|\mathcal{C}(\Delta_{01})\|_2^2 \tag{100}$$

$$= -\sum_{j \ge 2} \frac{1}{n}\,\langle \mathcal{C}(\Delta_{01}), \mathcal{C}(\Delta_j)\rangle \tag{101}$$

$$\le \delta\,\|\Delta_{01}\|_{\mathrm{F}} \sum_{j \ge 2} \|\Delta_j\|_{\mathrm{F}}\,. \tag{102}$$

For the singular-value blocks, every singular value in block j is no larger than the average singular value in block $j-1$. Therefore

$$\|\Delta_j\|_{\mathrm{F}} \le \frac{1}{\sqrt{2r}}\,\|\Delta_{j-1}\|_*\,, \qquad j \ge 2.$$

Summing over $j \ge 2$ gives

$$\sum_{j \ge 2} \|\Delta_j\|_{\mathrm{F}} \le \frac{1}{\sqrt{2r}}\,\|P_{\mathcal{T}^\perp}\Delta\|_*\,. \tag{103}$$

If $\Delta_{01} \ne 0$, divide (102) by its Frobenius norm and use (103):

$$\|\Delta_{01}\|_{\mathrm{F}} \le \frac{\delta}{1-\delta}\frac{1}{\sqrt{2r}}\,\|P_{\mathcal{T}^\perp}\Delta\|_*\,.$$

Since $\operatorname{rank} \Delta_0 \le 2r$,

$$\begin{aligned}\|P_{\mathcal{T}}\Delta\|_* &= \|\Delta_0\|_* \\ &\le \sqrt{2r}\,\|\Delta_0\|_{\mathrm{F}} \\ &\le \sqrt{2r}\,\|\Delta_{01}\|_{\mathrm{F}} \\ &\le \frac{\delta}{1-\delta}\,\|P_{\mathcal{T}^\perp}\Delta\|_*\,.\end{aligned}$$

Because $\delta < 1/2$, the coefficient is strictly smaller than 1.

If $\Delta_{01} = 0$, orthogonality forces $\Delta_0 = \Delta_1 = 0$. Since $\Delta \ne 0$, at least one later normal block is nonzero, and then the left side of (99) is zero while the right side is positive. Thus (99) holds in all cases. □

Lemma 23 (Strict null-space property gives uniqueness)**.** *If every nonzero $\Delta \in \ker \widetilde{\mathcal{A}}$ satisfies (99), then $\widetilde{M}_\star$ is the unique minimizer of*

$$\min_{\widetilde{\mathcal{A}}(X) = \widetilde{\mathcal{A}}(\widetilde{M}_\star)} \|X\|_*\,.$$

Proof. Take a nonzero feasible perturbation $\Delta \in \ker \widetilde{\mathcal{A}}$. Let $G_\perp \in \mathcal{T}^\perp$ satisfy

$$\|G_\perp\|_{\mathrm{op}} \le 1, \qquad \langle G_\perp, P_{\mathcal{T}^\perp}\Delta\rangle = \|P_{\mathcal{T}^\perp}\Delta\|_*\,.$$

Such a matrix is obtained from the singular vectors of the normal component. The matrix $EF^{\mathsf{T}} + G_\perp$ is a subgradient of the nuclear norm at $\widetilde{M}_\star = E\Sigma F^{\mathsf{T}}$. Therefore

$$\left\|\widetilde{M}_\star + \Delta\right\|_* \ge \left\|\widetilde{M}_\star\right\|_* + \left\langle EF^{\mathsf{T}} + G_\perp, \Delta\right\rangle$$

$$
\begin{aligned}
&= \left\|\widetilde{M}_\star\right\|_* + \left\langle EF^\mathsf{T}, P_{\mathcal{T}}\Delta\right\rangle + \|P_{\mathcal{T}^\perp}\Delta\|_* \\
&\geq \left\|\widetilde{M}_\star\right\|_* - \|P_{\mathcal{T}}\Delta\|_* + \|P_{\mathcal{T}^\perp}\Delta\|_* \\
&> \left\|\widetilde{M}_\star\right\|_*,
\end{aligned}
$$

where the last strict inequality is (99). Thus every other feasible point has larger nuclear norm. □

Proposition 24 (Unique weighted recovery at optimal sample order)**.** *There are numerical constants $C, c > 0$ such that, if*

$$Cr(d_1 + d_2) \leq n \leq d_1 d_2,$$

then with probability at least $1 - Ce^{-cn}$, $S_L, S_R \succ 0$ and $M_\star$ is the unique minimizer of (89).

Proof. Choose

$$\rho = \delta = \frac{1}{20}, \qquad s = \min\{6r, \min(d_1, d_2)\}.$$

For a sufficiently large numerical constant in the sample lower bound, the hypotheses of Lemmas 18 and 19 hold. Work on the intersection of those two events. By Lemma 20, the whitened map $\widetilde{\mathcal{A}}$ has rank-s RIP constant at most $1/5 < 1/2$. If $6r$ is no larger than the maximal matrix rank, Lemma 22 applies directly. If $6r$ is larger, rank-s RIP is RIP on the entire matrix space, which is stronger than what the same proof needs. Hence in either case the strict null-space property (99) holds at $\widetilde{M}_\star$.

By Lemma 23, $\widetilde{M}_\star$ is the unique ordinary nuclear-norm minimizer in the whitened coordinates. Undoing the invertible whitening transformation shows that $M_\star$ is the unique minimizer of (89). The marginal event also gives $S_L, S_R \succ 0$.

The failure probability is at most the sum of the Wishart and RIP failure probabilities. Under $n \geq Cr(d_1 + d_2)$ these are bounded by Ce^{-cn} after changing numerical constants. No independence between the events is needed. □

D THE PROJECTED SHARPNESS FLOW REACHES THE RECOVERING SOLUTION

We now fix a sensing instance for which

(i) $\mathcal{A}$ has full row rank;

(ii) $S_L, S_R \succ 0$;

(iii) $M_\star$ is the unique minimizer of (89); and

(iv) the generic transversality property proved below holds.

All four properties hold on the high-probability event assembled in the final section.

The sharpness flow starts from a feasible endpoint of deterministic gradient flow whose stacked factor state has full row rank. For the common initialization in Theorem 1, existence of such an endpoint follows from Proposition 37 without any large-initialization assumption. The purpose of this section is to prove that the constrained negative-gradient flow of J from that point exists for all time and that its product converges to $M_\star$.

D.1 REGULAR POINTS AND THE EXACT PROJECTED VECTOR FIELD

For the stacked factor state $W = \binom{U}{V}$, write

$$g(W) = \mathcal{A}(UV^\mathsf{T}) - y.$$

We identify W with the factor pair (U, V) when convenient. A feasible state is called *regular* when $Dg(W)$ maps onto $\mathbb{R}^n$. Direct differentiation gives

$$Dg(W)[\Delta U, \Delta V] = \mathcal{A}(\Delta U V^\mathsf{T} + U\Delta V^\mathsf{T}). \tag{104}$$

Lemma 25 (A full-row-rank stack is automatically regular)**.** *If $W = \binom{U}{V}$ has full row rank and $\mathcal{A}$ has full row rank, then (U, V) is regular.*

Proof. A full-row-rank stack has every subset of its rows linearly independent. In particular, V has full row rank, so VV^T is invertible. Given any matrix $N \in \mathbb{R}^{d_1 \times d_2}$, choose

$$\Delta U = N(VV^\mathsf{T})^{-1}V, \qquad \Delta V = 0.$$

Then

$$\Delta U V^\mathsf{T} + U \Delta V^\mathsf{T} = N(VV^\mathsf{T})^{-1}VV^\mathsf{T} = N.$$

Thus the derivative of the product map $(U, V) \mapsto UV^\mathsf{T}$ is onto the full matrix space. Composing with the surjective map $\mathcal{A}$ proves that Dg is onto $\mathbb{R}^n$. □

Proposition 26 (Exact formula for the projected sharpness flow)**.** *At a regular feasible point, the Euclidean negative-gradient flow of J constrained to Γ_y is*

$$\dot{U} = -S_L U + ZV, \qquad \dot{V} = -S_R V + Z^\mathsf{T} U, \qquad Z = \mathcal{A}^*(\lambda), \tag{105}$$

where $\lambda \in \mathbb{R}^n$ is uniquely determined by

$$\mathcal{A}(UU^\mathsf{T} Z + ZVV^\mathsf{T}) = \mathcal{A}(S_L UV^\mathsf{T} + UV^\mathsf{T} S_R). \tag{106}$$

Along this flow,

$$\frac{d}{dt} J(U_t, V_t) = -\left\| \operatorname{grad}_{\Gamma_y} J(U_t, V_t) \right\|_{\mathrm{F}}^2, \tag{107}$$

and

$$\frac{d}{dt}(U^\mathsf{T} U - V^\mathsf{T} V) = -2\big(U^\mathsf{T} S_L U - V^\mathsf{T} S_R V\big). \tag{108}$$

Proof. The Euclidean gradient is

$$\nabla J(U, V) = (S_L U, S_R V).$$

From (104), if $Z = \mathcal{A}^*(\lambda)$, then

$$\begin{aligned} \langle \lambda, Dg[\Delta U, \Delta V] \rangle &= \left\langle Z, \Delta U V^\mathsf{T} + U \Delta V^\mathsf{T} \right\rangle \\ &= \langle ZV, \Delta U \rangle + \left\langle Z^\mathsf{T} U, \Delta V \right\rangle. \end{aligned}$$

Hence

$$Dg^*(\lambda) = (ZV, Z^\mathsf{T} U).$$

At a regular point, the orthogonal projection of $-\nabla J$ onto the tangent space $\ker Dg$ has the form

$$-(S_L U, S_R V) + Dg^*(\lambda),$$

with λ chosen so that the vector is tangent. This gives (105), and uniqueness follows because regularity makes $DgDg^*$ invertible.

Differentiate $M = UV^\mathsf{T}$ along (105):

$$\begin{aligned} \dot{M} &= (-S_L U + ZV)V^\mathsf{T} + U(-S_R V + Z^\mathsf{T} U)^\mathsf{T} \\ &= -S_L M - M S_R + ZVV^\mathsf{T} + UU^\mathsf{T} Z. \end{aligned}$$

Tangency is $\mathcal{A}(\dot{M}) = 0$, which is exactly (106).

Since the velocity is minus the orthogonal projection of ∇J onto the feasible tangent space,

$$\frac{d}{dt} J = \left\langle \operatorname{grad}_{\Gamma_y} J, -\operatorname{grad}_{\Gamma_y} J \right\rangle = -\left\| \operatorname{grad}_{\Gamma_y} J \right\|_{\mathrm{F}}^2.$$

Finally,

$$\begin{aligned} \frac{d}{dt}(U^\mathsf{T} U) &= -2U^\mathsf{T} S_L U + U^\mathsf{T} ZV + V^\mathsf{T} Z^\mathsf{T} U, \\ \frac{d}{dt}(V^\mathsf{T} V) &= -2V^\mathsf{T} S_R V + V^\mathsf{T} Z^\mathsf{T} U + U^\mathsf{T} ZV. \end{aligned}$$

Subtracting cancels the mixed terms and gives (108). □

D.2 THE FLOW CANNOT LOSE REGULARITY AT A FINITE TIME

The stacked equation corresponding to (105) is

$$\dot{W} = B_t W, \qquad B_t = \begin{pmatrix} -S_L & Z_t \\ Z_t^\mathsf{T} & -S_R \end{pmatrix}. \tag{109}$$

The matrix B_t is symmetric.

Proposition 27 (Global finite-time existence and regularity)**.** *Assume $S_L, S_R \succ 0$, $\mathcal{A}$ has full row rank, and the initial point is feasible with a full-row-rank stack W_0. Then the projected flow exists uniquely for every finite time, remains feasible and regular, and*

$$\det(W_t W_t^\mathsf{T}) = \det(W_0 W_0^\mathsf{T}) e^{-2t(\operatorname{tr} S_L + \operatorname{tr} S_R)}. \tag{110}$$

Moreover the factors are uniformly bounded for all $t \geq 0$.

Proof. By Lemma 25, the starting point is regular, so the projected vector field is smooth in a neighborhood and has a unique local solution. Let $[0, T_{\max})$ be the maximal interval on which this classical solution is defined.

Step 1: full row rank persists for every finite time. As long as $W_t W_t^\mathsf{T}$ is positive definite, (109) gives

$$\frac{d}{dt}(W_t W_t^\mathsf{T}) = B_t W_t W_t^\mathsf{T} + W_t W_t^\mathsf{T} B_t.$$

For a differentiable positive-definite matrix H_t, $\frac{d}{dt} \log\det H_t = \operatorname{tr}(H_t^{-1} \dot{H}_t)$. Applying this with $H_t = W_t W_t^\mathsf{T}$ and using cyclicity of trace,

$$\begin{aligned} \frac{d}{dt} \log\det(W_t W_t^\mathsf{T}) &= 2 \operatorname{tr}(B_t) \\ &= -2(\operatorname{tr} S_L + \operatorname{tr} S_R), \end{aligned}$$

because the off-diagonal blocks of B_t do not contribute to its trace. Integration gives (110). Its right side is strictly positive at every finite time, so the row Gram matrix cannot become singular at a first finite time. Therefore W_t retains full row rank on $[0, T_{\max})$.

Step 2: the factors are uniformly bounded. By (107), $J(U_t, V_t) \leq J(U_0, V_0)$. Positive definiteness of S_L, S_R gives

$$J(U_t, V_t) \geq \frac{1}{2} \min\{\lambda_{\min}(S_L), \lambda_{\min}(S_R)\} \left(\|U_t\|_{\mathrm{F}}^2 + \|V_t\|_{\mathrm{F}}^2 \right).$$

Therefore

$$\|W_t\|_{\mathrm{F}}^2 \leq \frac{2J(U_0, V_0)}{\min\{\lambda_{\min}(S_L), \lambda_{\min}(S_R)\}} \qquad (0 \leq t < T_{\max}). \tag{111}$$

Step 3: regularity is quantitatively uniform on each finite interval. Fix $T < T_{\max}$. There are $d_1 + d_2$ positive row singular values of W_t, and

$$\det(W_t W_t^\mathsf{T}) = \prod_{j=1}^{d_1+d_2} \sigma_j(W_t)^2.$$

All singular values other than the smallest are at most $\|W_t\|_{\mathrm{op}}$. Combining (110) with (111),

$$\inf_{0 \leq t \leq T} \sigma_{\min}(W_t) > 0. \tag{112}$$

We now show that this gives a uniform lower bound on $Dg Dg^*$. For $\lambda \in \mathbb{R}^n$, put $Z = \mathcal{A}^*(\lambda)$. The formula for Dg^* from the previous proof gives

$$\|Dg^* \lambda\|_{\mathrm{F}}^2 = \|Z V_t\|_{\mathrm{F}}^2 + \left\|Z^\mathsf{T} U_t\right\|_{\mathrm{F}}^2.$$

Since $W_t W_t^\mathsf{T} \succeq \sigma_{\min}(W_t)^2 I$, both diagonal principal blocks satisfy

$$U_t U_t^\mathsf{T} \succeq \sigma_{\min}(W_t)^2 I, \qquad V_t V_t^\mathsf{T} \succeq \sigma_{\min}(W_t)^2 I.$$

Hence

$$\begin{aligned} \|Dg^*\lambda\|_{\mathrm{F}}^2 &\geq \sigma_{\min}(W_t)^2 \, \|Z\|_{\mathrm{F}}^2 \\ &= \sigma_{\min}(W_t)^2 \, \|\mathcal{A}^*\lambda\|_{\mathrm{F}}^2 \\ &\geq \sigma_{\min}(W_t)^2 \lambda_{\min}(\mathcal{A}\mathcal{A}^*) \, \|\lambda\|_2^2 . \end{aligned}$$

The right side is uniformly positive on $[0, T]$ for every nonzero λ, by (112). Thus $DgDg^*$ remains uniformly invertible on every finite interval.

Step 4: a finite maximal time is impossible. At a regular point, the multiplier is

$$\lambda = (DgDg^*)^{-1} Dg(S_L U, S_R V). \tag{113}$$

Suppose $T_{\max} < \infty$. The determinant identity and the uniform factor bound give a positive lower bound for $\sigma_{\min}(W_t)$ on all of $[0, T_{\max})$, so $DgDg^*$ stays uniformly invertible there. Formula (113) then bounds the multiplier and hence the projected vector field on the bounded trajectory. The velocity is therefore bounded, so (U_t, V_t) is Cauchy as $t \uparrow T_{\max}$ and has a finite limit. The determinant lower bound implies that the limiting stack still has full row rank, hence the limiting point is regular by Lemma 25. Local existence from that limit extends the solution beyond $T_{\max}$, a contradiction. Therefore $T_{\max} = \infty$.

Finally, the vector field is tangent by construction, so $\frac{d}{dt} g(U_t, V_t) = 0$ and feasibility is preserved. Full row rank at every finite time and Lemma 25 give regularity at every finite time. □

D.3 The flow converges to one constrained critical point

Proposition 28 (Finite tail length and convergence)**.** *Under the hypotheses of Proposition 27, the trajectory has finite Euclidean tail length and converges to a single feasible limit* (U_∞, V_∞)*. The limit is a constrained critical point of* J*.*

Proof. The energy identity shows that $J(U_t, V_t)$ decreases and is bounded below by zero. Let

$$J_\infty = \lim_{t \to \infty} J(U_t, V_t).$$

At every finite time the feasible set is locally a smooth manifold. The flow velocity is minus the tangent projection of ∇J, so

$$\operatorname{dist}\big(0, \partial(J + \iota_{\Gamma_y})(U_t, V_t)\big) = \sqrt{\left\|\dot{U}_t\right\|_{\mathrm{F}}^2 + \left\|\dot{V}_t\right\|_{\mathrm{F}}^2}, \tag{114}$$

and

$$\frac{d}{dt} J(U_t, V_t) = -\left\|\dot{U}_t\right\|_{\mathrm{F}}^2 - \left\|\dot{V}_t\right\|_{\mathrm{F}}^2 . \tag{115}$$

Here ι_{Γ_y} is the indicator function of the feasible set.

The factor bound (111) makes the trajectory bounded, so its cluster set is compact. Every cluster point has objective value J_∞. The function $J + \iota_{\Gamma_y}$ is semialgebraic because J and the equations defining Γ_y are polynomial. The Kurdyka–Łojasiewicz theorem stated in Appendix F.2 therefore gives an increasing concave C^1 function φ such that, near the compact cluster set and whenever $J > J_\infty$,

$$\varphi'(J(U, V) - J_\infty) \operatorname{dist}\big(0, \partial(J + \iota_{\Gamma_y})(U, V)\big) \geq 1. \tag{116}$$

The trajectory eventually stays in this neighborhood; otherwise a sequence of late points outside it would have a convergent subsequence whose limit is itself a cluster point, a contradiction.

If J reaches J_∞ at finite time, (115) forces zero velocity thereafter, so convergence is immediate. Otherwise combine (116), (114), and (115):

$$-\frac{d}{dt} \varphi(J(U_t, V_t) - J_\infty) = \varphi'(J - J_\infty)\Big(\left\|\dot{U}_t\right\|_{\mathrm{F}}^2 + \left\|\dot{V}_t\right\|_{\mathrm{F}}^2\Big)$$

$$\geq \sqrt{\left\|\dot{U}_t\right\|_{\mathrm{F}}^2 + \left\|\dot{V}_t\right\|_{\mathrm{F}}^2}.$$

Integrating from a sufficiently large time T gives

$$\int_T^\infty \sqrt{\left\|\dot{U}_t\right\|_{\mathrm{F}}^2 + \left\|\dot{V}_t\right\|_{\mathrm{F}}^2}\, dt < \infty.$$

Thus the trajectory is Cauchy and converges to a single feasible limit.

Finally, integrating (115) over $[0, \infty)$ gives a finite integral of the squared speed. Hence there is a sequence $t_j \to \infty$ along which the speed tends to zero. By (114), there are subgradients of $J + \iota_{\Gamma_y}$ at (U_{t_j}, V_{t_j}) tending to zero. Closedness of the limiting subdifferential then gives

$$0 \in \partial(J + \iota_{\Gamma_y})(U_\infty, V_\infty).$$

At a regular limit this is precisely the usual Lagrange-multiplier first-order condition. □

Lemma 29 (Every limiting factorization is weighted-balanced)**.** *The limit in Proposition 28 satisfies*

$$U_\infty^\mathsf{T} S_L U_\infty = V_\infty^\mathsf{T} S_R V_\infty. \tag{117}$$

Proof. Integrate (108) from 0 to t and divide by t:

$$\frac{U_t^\mathsf{T} U_t - V_t^\mathsf{T} V_t - U_0^\mathsf{T} U_0 + V_0^\mathsf{T} V_0}{t} = -\frac{2}{t}\int_0^t \left(U_s^\mathsf{T} S_L U_s - V_s^\mathsf{T} S_R V_s\right) ds.$$

The numerator on the left converges and is therefore bounded, so the left side tends to zero. The integrand on the right converges to

$$U_\infty^\mathsf{T} S_L U_\infty - V_\infty^\mathsf{T} S_R V_\infty.$$

A convergent function and its long-time average have the same limit. Therefore the displayed limit must be zero, proving (117). □

D.4 Why an incorrect weighted-balanced exact fit is regular almost surely

For each $0 \leq k \leq \min(d_1, d_2)$, the set of rank-$k$ matrices is a smooth manifold. For a rank-k matrix with compact SVD $M = E\Sigma F^\mathsf{T}$, write T_M for its tangent space in this manifold. Explicitly,

$$T_M = \{EB^\mathsf{T} + CF^\mathsf{T} : B \in \mathbb{R}^{d_2 \times k},\ C \in \mathbb{R}^{d_1 \times k}\}. \tag{118}$$

Proposition 30 (Generic fixed-rank transversality)**.** *Except for a Lebesgue-null set of linear sensing maps $\mathcal{A}$, the following holds simultaneously for every k: if* $\operatorname{rank} M = k$, $M \neq M_\star$, *and $\mathcal{A}(M) = \mathcal{A}(M_\star)$, then*

$$\mathcal{A}(T_M) = \mathbb{R}^n. \tag{119}$$

In particular, the property holds almost surely for Gaussian sensing.

Proof. Fix k. Let $\mathscr{L}$ be the vector space of linear maps from matrix space to $\mathbb{R}^n$. On the product of $\mathscr{L}$ with the rank-k manifold excluding $M_\star$, define

$$F_k(\mathcal{A}, M) = \mathcal{A}(M - M_\star).$$

At any point of its domain and for any $v \in \mathbb{R}^n$, define a perturbation of the sensing map by

$$\Delta\mathcal{A}(X) = v\,\frac{\langle M - M_\star, X\rangle}{\|M - M_\star\|_{\mathrm{F}}^2}.$$

Then

$$D_{\mathcal{A}} F_k(\mathcal{A}, M)[\Delta\mathcal{A}] = v.$$

Thus the derivative in the parameter variable is onto everywhere. Proposition 43 therefore implies that, for almost every fixed $\mathcal{A}$, the derivative in the M variable is onto at every zero of F_k. But

$$D_M F_k(\mathcal{A}, M)[H] = \mathcal{A}(H),$$

so the conclusion is exactly (119). There are only finitely many possible ranks k, so we can impose the property simultaneously for all k. Finally, the Gaussian law has a density on $\mathscr{L}$, so it avoids the exceptional Lebesgue-null set with probability one. □

We need one elementary fact about balanced factorizations. In the next lemma, the word "balanced" refers to ordinary Euclidean balance, $\widetilde{U}^\mathsf{T}\widetilde{U} = \widetilde{V}^\mathsf{T}\widetilde{V}$.

Lemma 31 (Balanced factors have the canonical row geometry)**.** *If $\widetilde{M} = \widetilde{U}\widetilde{V}^\mathsf{T}$ and $\widetilde{U}^\mathsf{T}\widetilde{U} = \widetilde{V}^\mathsf{T}\widetilde{V}$, then*

$$\widetilde{U}\widetilde{U}^\mathsf{T} = (\widetilde{M}\widetilde{M}^\mathsf{T})^{1/2}, \qquad \widetilde{V}\widetilde{V}^\mathsf{T} = (\widetilde{M}^\mathsf{T}\widetilde{M})^{1/2}.$$

If $\widetilde{M} = E\Sigma F^\mathsf{T}$ has rank $k > 0$, then for some $R \in \mathbb{R}^{m\times k}$ with $R^\mathsf{T}R = I_k$,

$$\widetilde{U} = E\Sigma^{1/2}R^\mathsf{T}, \qquad \widetilde{V} = F\Sigma^{1/2}R^\mathsf{T}. \tag{120}$$

Proof. Balance gives

$$\widetilde{M}\widetilde{M}^\mathsf{T} = \widetilde{U}\widetilde{V}^\mathsf{T}\widetilde{V}\widetilde{U}^\mathsf{T} = \widetilde{U}(\widetilde{U}^\mathsf{T}\widetilde{U})\widetilde{U}^\mathsf{T} = (\widetilde{U}\widetilde{U}^\mathsf{T})^2.$$

The positive-semidefinite square root is unique, so $\widetilde{U}\widetilde{U}^\mathsf{T} = (\widetilde{M}\widetilde{M}^\mathsf{T})^{1/2}$. The second identity is identical with the roles interchanged.

If $\widetilde{M} = E\Sigma F^\mathsf{T}$, the first identity says that $\widetilde{U}$ has column space $\mathrm{span}(E)$ and nonzero singular values equal to the square roots of the diagonal entries of Σ. Thus $\widetilde{U} = E\Sigma^{1/2}R_U^\mathsf{T}$ for some R_U with orthonormal columns. Similarly $\widetilde{V} = F\Sigma^{1/2}R_V^\mathsf{T}$. Multiplying these factors gives

$$E\Sigma^{1/2}R_U^\mathsf{T}R_V\Sigma^{1/2}F^\mathsf{T} = E\Sigma F^\mathsf{T},$$

so $R_U^\mathsf{T}R_V = I_k$. Two matrices with orthonormal columns whose cross Gram is the identity must be equal, so $R_U = R_V =: R$, proving (120). □

Lemma 32 (Weighted-balanced factors generate every fixed-rank tangent direction)**.** *Suppose $S_L, S_R \succ 0$, $M = UV^\mathsf{T}$ has rank k, and*

$$U^\mathsf{T}S_LU = V^\mathsf{T}S_RV. \tag{121}$$

Then

$$\{\Delta UV^\mathsf{T} + U\Delta V^\mathsf{T} : \Delta U, \Delta V\} = T_M. \tag{122}$$

Proof. Whiten the factors:

$$\widetilde{U} = S_L^{1/2}U, \qquad \widetilde{V} = S_R^{1/2}V, \qquad \widetilde{M} = S_L^{1/2}MS_R^{1/2}.$$

Condition (121) becomes ordinary balance. Since the change of coordinates is invertible, it preserves rank and maps fixed-rank tangent spaces bijectively. It is enough to prove the result in the whitened variables.

If $k = 0$, balance forces both factors to be zero and both sides of (122) are $\{0\}$. Assume $k > 0$. Let $\widetilde{M} = E\Sigma F^\mathsf{T}$. By Lemma 31,

$$\widetilde{U} = E\Sigma^{1/2}R^\mathsf{T}, \qquad \widetilde{V} = F\Sigma^{1/2}R^\mathsf{T}$$

for some $R^\mathsf{T}R = I_k$.

An arbitrary tangent vector has the form $EB^\mathsf{T} + CF^\mathsf{T}$. Choose

$$\Delta\widetilde{U} = C\Sigma^{-1/2}R^\mathsf{T}, \qquad \Delta\widetilde{V} = B\Sigma^{-1/2}R^\mathsf{T}.$$

Then

$$\Delta\widetilde{U}\,\widetilde{V}^\mathsf{T} + \widetilde{U}\,\Delta\widetilde{V}^\mathsf{T} = CF^\mathsf{T} + EB^\mathsf{T}.$$

Thus every fixed-rank tangent vector is generated by factor perturbations. Conversely, because the row space of $\widetilde{V}^\mathsf{T}$ is contained in $\mathrm{span}(F)$ and the column space of $\widetilde{U}$ is contained in $\mathrm{span}(E)$, every factor perturbation has the form $CF^\mathsf{T} + EB^\mathsf{T}$ and hence lies in the tangent space. Undoing the whitening proves (122). □

Corollary 33 (An incorrect weighted-balanced exact fit is regular)**.** *Assume the transversality property of Proposition 30. If a feasible factor pair is weighted-balanced and has product $M \neq M_\star$, then it is regular.*

Proof. By Lemma 32, the image of the derivative of the product map is T_M. Applying $\mathcal{A}$ and using (119) gives all of $\mathbb{R}^n$, which is exactly regularity of Dg. □

D.5 AN INCORRECT CRITICAL LIMIT WOULD FORCE EXPONENTIAL GROWTH

At a regular constrained critical point, there is a multiplier λ and $Z = \mathcal{A}^*(\lambda)$ such that

$$S_L U = ZV, \qquad S_R V = Z^\mathsf{T} U. \tag{123}$$

Lemma 34 (A negative-semidefinite block coefficient is a global optimality certificate)**.** *At a regular critical point satisfying* (123)*, suppose*

$$\begin{pmatrix} -S_L & Z \\ Z^\mathsf{T} & -S_R \end{pmatrix} \preceq 0. \tag{124}$$

Then (U, V) *globally minimizes* J *over* Γ_y*, and its product globally minimizes* (89)*.*

Proof. Congruence by $\operatorname{diag}(S_L^{-1/2}, S_R^{-1/2})$ shows that (124) is equivalent to

$$\begin{pmatrix} -I & S_L^{-1/2} Z S_R^{-1/2} \\ S_R^{-1/2} Z^\mathsf{T} S_L^{-1/2} & -I \end{pmatrix} \preceq 0.$$

The negative of this matrix has quadratic form

$$\|x\|_2^2 + \|y\|_2^2 - 2\left\langle x, S_L^{-1/2} Z S_R^{-1/2} y \right\rangle.$$

This is nonnegative for every x, y if and only if

$$\left\| S_L^{-1/2} Z S_R^{-1/2} \right\|_{\mathrm{op}} \le 1. \tag{125}$$

Indeed, the inequality $2|\langle x, Cy\rangle| \le \|x\|^2 + \|y\|^2$ follows from $\|C\| \le 1$, and conversely taking unit singular vectors attaining $\|C\|_{\mathrm{op}}$ gives the reverse implication.

Take Frobenius inner products of the two KKT equations with U and V:

$$\operatorname{tr}(U^\mathsf{T} S_L U) = \left\langle Z, UV^\mathsf{T} \right\rangle, \qquad \operatorname{tr}(V^\mathsf{T} S_R V) = \left\langle Z, UV^\mathsf{T} \right\rangle.$$

Hence

$$J(U, V) = \left\langle Z, UV^\mathsf{T} \right\rangle. \tag{126}$$

For any feasible matrix N,

$$\langle Z, N\rangle = \langle \lambda, \mathcal{A}(N)\rangle = \langle \lambda, y\rangle = \left\langle Z, UV^\mathsf{T} \right\rangle = J(U, V).$$

By (125) and nuclear/operator duality,

$$\begin{aligned} \left\| S_L^{1/2} N S_R^{1/2} \right\|_* &\ge \left\langle S_L^{-1/2} Z S_R^{-1/2}, S_L^{1/2} N S_R^{1/2} \right\rangle \\ &= \langle Z, N\rangle = J(U, V). \end{aligned}$$

On the other hand, Proposition 17 gives, for the current product,

$$\left\| S_L^{1/2} U V^\mathsf{T} S_R^{1/2} \right\|_* \le J(U, V).$$

The preceding lower bound applied to $N = UV^\mathsf{T}$ gives the reverse inequality. Thus equality holds at the current product, while every feasible product has weighted nuclear norm at least this value. The current product is therefore globally optimal for (89), and the current factorization is globally optimal for J. □

Lemma 35 (A positive limiting eigenvalue creates persistent exponential expansion)**.** *Let* $B : [0, \infty) \to \mathbb{R}^{q \times q}$ *be continuous and symmetric, with* $B(t) \to B_\infty$*. If* $0 < \beta < \lambda_{\max}(B_\infty)$*, then for some* $T < \infty$ *the fundamental matrix* $\Psi(t, T)$ *solving*

$$\partial_t \Psi(t, T) = B(t)\Psi(t, T), \qquad \Psi(T, T) = I,$$

satisfies

$$\|\Psi(t, T)\|_{\mathrm{op}} \ge e^{\beta(t-T)} \qquad (t \ge T). \tag{127}$$

Proof. Choose θ with $\beta < \theta < \lambda_{\max}(B_\infty)$ and θ not an eigenvalue of B_∞. Let E_+ be the sum of eigenspaces with eigenvalues above θ, and let $E_- = E_+^\perp$. Put

$$a = \lambda_{\min}(B_\infty|_{E_+}), \qquad b = \lambda_{\max}(B_\infty|_{E_-}),$$

so $a > \theta > b$.

Because $B(t) \to B_\infty$, choose T so that

$$d := \sup_{t \geq T} \|B(t) - B_\infty\|_{\mathrm{op}} < \min\left\{\frac{a-b}{4}, \frac{a-\beta}{2}\right\}.$$

For a solution $x = u + v$ with $u \in E_+$ and $v \in E_-$, the one-sided derivatives of the norms satisfy

$$D^- \|u\|_2 \geq a \|u\|_2 - d(\|u\|_2 + \|v\|_2),$$
$$D^+ \|v\|_2 \leq b \|v\|_2 + d(\|u\|_2 + \|v\|_2).$$

At a time when $\|v\|_2 = \|u\|_2$, subtraction gives

$$D^+(\|v\|_2 - \|u\|_2) \leq (b - a + 4d) \|u\|_2 < 0.$$

Thus the region $\|v\|_2 \leq \|u\|_2$ is forward invariant.

Start at time T from any unit vector in E_+. Then $v(T) = 0$. Throughout the evolution,

$$\begin{aligned} D^- \|u\|_2 &\geq a \|u\|_2 - d(\|u\|_2 + \|v\|_2) \\ &\geq (a - 2d) \|u\|_2 \\ &\geq \beta \|u\|_2 . \end{aligned}$$

Hence $\|u\|(t) \geq e^{\beta(t-T)}$. Since $u \perp v$, also $\|x\|(t) \geq \|u\|(t)$. Taking the supremum over unit initial vectors gives (127). □

Proposition 36 (Global convergence of the sharpness flow)**.** *Under the four fixed-instance assumptions listed at the beginning of this section, if the sharpness flow starts from any feasible factor pair with full-row-rank stack, then*

$$U_t V_t^\mathsf{T} \longrightarrow M_\star. \tag{128}$$

Proof. By Proposition 27, the trajectory exists for every finite time, is feasible and regular, and remains bounded. By Proposition 28, the factors converge to one feasible constrained critical point (U_∞, V_∞). By Lemma 29, that limit is weighted-balanced.

Suppose for contradiction that

$$M_\infty := U_\infty V_\infty^\mathsf{T} \neq M_\star.$$

By Corollary 33, the limiting factor pair is regular. Therefore the limiting criticality condition is an ordinary KKT condition: there is a unique multiplier λ_∞ and $Z_\infty = \mathcal{A}^*(\lambda_\infty)$ satisfying (123) at the limit.

Near a regular point, $DgDg^*$ is invertible and the multiplier used by the flow is the smooth function (113). Since the factors converge, we have

$$\lambda_t \to \lambda_\infty, \qquad Z_t \to Z_\infty.$$

Because $M_\star$ is the *unique* global minimizer of the weighted nuclear-norm problem, the limiting product $M_\infty \neq M_\star$ is not globally optimal. By Lemma 34, the limiting block matrix

$$B_\infty = \begin{pmatrix} -S_L & Z_\infty \\ Z_\infty^\mathsf{T} & -S_R \end{pmatrix}$$

cannot be negative semidefinite. Since it is symmetric, it must have a strictly positive largest eigenvalue.

Choose $0 < \beta < \lambda_{\max}(B_\infty)$. The coefficient matrices B_t in (109) converge to B_∞, so Lemma 35 gives T such that

$$\|\Psi(t, T)\|_{\mathrm{op}} \geq e^{\beta(t-T)}, \qquad t \geq T,$$

where Ψ is the fundamental matrix of the stack equation. Since $W_t = \Psi(t,T)W_T$ and W_T has full row rank, its smallest row singular value is positive. For every matrix R and every full-row-rank W_T,

$$\|RW_T\|_{\mathrm{op}} \geq \sigma_{\min}(W_T)\,\|R\|_{\mathrm{op}}\,. \tag{129}$$

Indeed, choose a unit vector x with $\|Rx\| = \|R\|_{\mathrm{op}}$. Since W_T is onto from $\mathbb{R}^m$ to $\mathbb{R}^{d_1+d_2}$, there is z with $W_T z = x$ and $\|z\| \leq 1/\sigma_{\min}(W_T)$. Then

$$\|RW_T\|_{\mathrm{op}} \geq \frac{\|RW_T z\|_2}{\|z\|_2} \geq \sigma_{\min}(W_T)\,\|R\|_{\mathrm{op}}\,.$$

Applying this to $R = \Psi(t,T)$ gives

$$\|W_t\|_{\mathrm{op}} \geq \sigma_{\min}(W_T)e^{\beta(t-T)}.$$

Thus the factor stack grows exponentially. This contradicts the uniform factor bound (111). The assumption $M_\infty \neq M_\star$ is impossible, proving (128). □

E From label noise to the projected sharpness flow

The analysis of the deterministic branch is now complete. On a sensing instance satisfying the assumptions of section D, the endpoint of deterministic gradient flow is a regular feasible factor pair, and the projected negative-gradient flow of J from that endpoint converges in product to $M_\star$. It remains to justify that small-step label-noise SGD follows this projected flow on its slow time scale.

E.1 Exact centering and exact covariance

For a stacked factor state $W = \binom{U}{V}$, with $M(W) = UV^{\mathsf{T}}$, define the residual of the ith measurement by

$$r_i(W) = \langle A_i, UV^{\mathsf{T}} \rangle - y_i.$$

A direct differentiation gives

$$\nabla r_i(W) = \begin{pmatrix} A_i V \\ A_i^{\mathsf{T}} U \end{pmatrix}. \tag{130}$$

For a sampled index i and a fresh perturbation $\xi \in \{-\sigma, +\sigma\}$,

$$\ell_{i,\xi}(W) = \frac{1}{2}(r_i(W) - \xi)^2,$$

so

$$\nabla \ell_{i,\xi}(W) = (r_i(W) - \xi)\nabla r_i(W). \tag{131}$$

Averaging over the $2n$ equally likely states (i,ξ) gives

$$\frac{1}{2n}\sum_{i=1}^{n}\sum_{\xi\in\{-\sigma,+\sigma\}} \ell_{i,\xi}(W) = L(W) + \frac{\sigma^2}{2}.$$

The additive constant disappears after differentiation. Hence the average sampled gradient is exactly ∇L.

To put the recursion into the centered form used by the slow-time theorem, define

$$\omega_{i,\xi}(W) = \frac{\nabla \ell_{i,\xi}(W) - \nabla L(W)}{\sqrt{2n}}. \tag{132}$$

Then, identically for every ambient factor state,

$$\sum_{i=1}^{n}\sum_{\xi\in\{-\sigma,+\sigma\}} \omega_{i,\xi}(W) = 0, \qquad \nabla L(W) + \sqrt{2n}\,\omega_{i,\xi}(W) = \nabla \ell_{i,\xi}(W). \tag{133}$$

Using (131), the centered fluctuation can also be written as

$$\sqrt{2n}\,\omega_{i,\xi} = r_i \nabla r_i - \nabla L - \xi \nabla r_i.$$

The first two terms are the fluctuation caused by sampling i. They disappear on the interpolation set, but keeping them is essential for exact centering away from that set.

Now let $W \in \Gamma_y$. Then every $r_i(W) = 0$ and $\nabla L(W) = 0$, so

$$\sqrt{2n}\,\omega_{i,\xi}(W) = -\xi \nabla r_i(W).$$

Therefore

$$\sum_{i=1}^{n} \sum_{\xi \in \{-\sigma,+\sigma\}} \omega_{i,\xi}\omega_{i,\xi}^{\mathsf{T}} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{\xi \in \{-\sigma,+\sigma\}} \xi^2 \nabla r_i \nabla r_i^{\mathsf{T}}$$
$$= \frac{\sigma^2}{n} \sum_{i=1}^{n} \nabla r_i \nabla r_i^{\mathsf{T}}. \tag{134}$$

On the other hand, differentiating $L = (2n)^{-1} \sum_i r_i^2$ twice gives

$$\nabla^2 L = \frac{1}{n} \sum_{i=1}^{n} \big(\nabla r_i \nabla r_i^{\mathsf{T}} + r_i \nabla^2 r_i\big).$$

At interpolation the second term vanishes, and hence

$$\nabla^2 L(W) = \frac{1}{n} \sum_{i=1}^{n} \nabla r_i(W) \nabla r_i(W)^{\mathsf{T}}. \tag{135}$$

Combining (134) and (135), the covariance entering the slow-time theorem is exactly

$$\sum_{i,\xi} \omega_{i,\xi}(W)\omega_{i,\xi}(W)^{\mathsf{T}} = \sigma^2 \nabla^2 L(W), \qquad W \in \Gamma_y. \tag{136}$$

E.2 A REGULAR DETERMINISTIC ENDPOINT FROM ANY FULL-STACK INITIALIZATION

The large-initialization analysis of section A is needed to identify a bad deterministic interpolator, but it is not needed to start the slow noisy motion. For the latter it is enough that the stacked initialization has full row rank.

Proposition 37 (Full-stack initialization gives a regular interpolating endpoint)**.** *Assume that $\mathcal{A}$ has full row rank and that $W_0 = \binom{U_0}{V_0}$ has full row rank. Then deterministic factor gradient flow (39) is global and converges in factor space to an exact fit*

$$W_{\mathrm{det}} = \begin{pmatrix} U_{\mathrm{det}} \\ V_{\mathrm{det}} \end{pmatrix}.$$

Moreover, W_{det} has full row rank and hence is a regular point of Γ_y.

Proof. Put

$$C_0 = U_0^{\mathsf{T}} U_0 - V_0^{\mathsf{T}} V_0.$$

Let the eigenvalues of a symmetric matrix be ordered nonincreasingly. Since W_0 has full row rank, $C_0 = W_0^{\mathsf{T}} J W_0$, where $J = \mathrm{diag}(I_{d_1}, -I_{d_2})$, has inertia $(d_1, d_2, m - d_1 - d_2)$. Indeed, if R is any right inverse of W_0, then

$$R^{\mathsf{T}} C_0 R = J,$$

so C_0 has at least d_1 positive and at least d_2 negative eigenvalues; on the other hand $\mathrm{rank}(C_0) \le d_1 + d_2$. Therefore

$$\gamma_U := \lambda_{d_1}(C_0) > 0, \qquad \gamma_V := \lambda_{d_2}(-C_0) > 0. \tag{137}$$

Let $W_t = \binom{U_t}{V_t}$ be the maximal solution of the deterministic factor flow and set $G_t = \nabla \ell(U_t V_t^{\mathsf{T}})$. A direct calculation gives the usual balance invariant

$$\frac{d}{dt}\left(U_t^{\mathsf{T}} U_t - V_t^{\mathsf{T}} V_t\right) = 0, \qquad U_t^{\mathsf{T}} U_t - V_t^{\mathsf{T}} V_t = C_0. \tag{138}$$

Consequently

$$C_0 \preceq U_t^\mathsf{T} U_t, \qquad -C_0 \preceq V_t^\mathsf{T} V_t.$$

Eigenvalue monotonicity under the Loewner order and (137) yield, throughout the maximal interval,

$$\sigma_{\min}(U_t)^2 \geq \gamma_U, \qquad \sigma_{\min}(V_t)^2 \geq \gamma_V. \tag{139}$$

In particular both factors remain full row rank. Also, $C_0 = W_t^\mathsf{T} J W_t$ has rank $d_1 + d_2$, so $\operatorname{rank}(W_t) \geq d_1 + d_2$; hence the stack remains full row rank as well.

Now differentiate the loss along gradient flow. Using $\dot{U}_t = -G_t V_t$ and $\dot{V}_t = -G_t^\mathsf{T} U_t$ gives

$$\begin{aligned}\frac{d}{dt} L(U_t, V_t) &= -\left\|G_t V_t\right\|_{\mathrm{F}}^2 - \left\|G_t^\mathsf{T} U_t\right\|_{\mathrm{F}}^2 \\ &\leq -(\gamma_U + \gamma_V)\left\|G_t\right\|_{\mathrm{F}}^2 .\end{aligned}$$

By (46), $\|G_t\|_{\mathrm{F}}^2 \geq 2\mu_{\mathcal{A}}^- L(U_t, V_t)$, and therefore

$$L(U_t, V_t) \leq L(U_0, V_0) e^{-2(\gamma_U + \gamma_V)\mu_{\mathcal{A}}^- t}. \tag{140}$$

The upper bound in (46) shows more precisely that, if $[0, T_{\max})$ is the maximal interval of existence, then

$$\int_0^{T_{\max}} \|G_t\|_{\mathrm{F}}\, dt \leq \frac{\sqrt{2\mu_{\mathcal{A}}^+ L(U_0, V_0)}}{(\gamma_U + \gamma_V)\mu_{\mathcal{A}}^-} < \infty. \tag{141}$$

In stacked form the factor flow is

$$\dot{W}_t = -K_t W_t, \qquad K_t = \begin{pmatrix} 0 & G_t \\ G_t^\mathsf{T} & 0 \end{pmatrix}, \qquad \|K_t\|_{\mathrm{op}} = \|G_t\|_{\mathrm{op}} \leq \|G_t\|_{\mathrm{F}} .$$

Gronwall's inequality and (141) give a uniform bound for W_t on $[0, T_{\max})$. Since the vector field is polynomial, a finite maximal time could only occur through blow-up, so $T_{\max} = \infty$. Thus (141) holds with upper limit ∞. The same bounds imply finite total length:

$$\int_0^\infty \left\|\dot{W}_t\right\|_{\mathrm{F}} dt \leq \left(\sup_{t \geq 0} \|W_t\|_{\mathrm{F}}\right) \int_0^\infty \|G_t\|_{\mathrm{F}}\, dt < \infty.$$

Hence W_t converges to a limit $W_{\det}$. By (140), its product interpolates the data. Finally, continuity in (138) gives $C_0 = W_{\det}^\mathsf{T} J W_{\det}$, whose rank is $d_1 + d_2$. Therefore $W_{\det}$ has full row rank. Regularity follows from Lemma 25. □

E.3 Why the deterministic endpoint map is smooth on the required neighborhood

Write $W^{\det}(t; W_0)$ for deterministic factor gradient flow starting from W_0. Whenever this trajectory converges, write

$$\Phi(W_0) = \lim_{t \to \infty} W^{\det}(t; W_0).$$

The slow-time theorem needs Φ to be smooth on a neighborhood. We establish this in two elementary localization steps, with the final smoothness conclusion supplied by the normally attracting manifold theorem stated in Appendix F.2.

Proposition 38 (Uniform local attraction to a regular piece of the exact-fit set)**.** *Let $K \subset \Gamma_y$ be compact. Suppose that Dg is onto on an ambient neighborhood of K. Then there are open sets $K \subset \mathcal{U} \Subset \mathcal{N}$ such that every deterministic gradient-flow trajectory starting in $\mathcal{U}$ remains in $\mathcal{N}$, exists for all time, has exponentially decaying residual and finite total length, and converges to a point of $\Gamma_y \cap \mathcal{N}$. After shrinking $\mathcal{U}$ if necessary, the endpoint map Φ is C^2 on $\mathcal{U}$.*

Proof. Because Dg is onto near K and K is compact, choose a bounded open neighborhood $\mathcal{N}$ of K and constants $a, b > 0$ such that for every $W \in \mathcal{N}$ and every $z \in \mathbb{R}^n$,

$$\|Dg(W)^* z\|_{\mathrm{F}} \geq a \|z\|_2 , \qquad \|Dg(W)\|_{\mathrm{op}} \leq b. \tag{142}$$

Deterministic gradient flow is

$$\dot{W} = -\nabla L(W) = -\frac{1}{n} Dg(W)^* g(W). \tag{143}$$

As long as the trajectory stays in $\mathcal{N}$, the chain rule gives

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\|g(W_t)\|_2^2 &= \left\langle g(W_t), Dg(W_t)[\dot{W}_t]\right\rangle \\ &= -\frac{1}{n}\|Dg(W_t)^* g(W_t)\|_{\mathrm{F}}^2 \\ &\le -\frac{a^2}{n}\|g(W_t)\|_2^2 .\end{aligned}$$

Thus

$$\|g(W_t)\|_2 \le e^{-a^2 t/n}\|g(W_0)\|_2 . \tag{144}$$

Also, by (143) and (142),

$$\left\|\dot{W}_t\right\|_{\mathrm{F}} \le \frac{b}{n}\|g(W_t)\|_2 .$$

Integrating and using (144),

$$\int_0^T \left\|\dot{W}_t\right\|_{\mathrm{F}} dt \le \frac{b}{a^2}\|g(W_0)\|_2 \qquad (T>0). \tag{145}$$

Choose $\mathcal{U}$ sufficiently close to K that the right side of (145) is smaller than the distance needed to leave $\mathcal{N}$. A first-exit argument then shows that no trajectory from $\mathcal{U}$ can leave $\mathcal{N}$. Since the vector field is smooth and the trajectory remains in a bounded set, the solution exists for all $t \ge 0$.

The bound (145) with $T \to \infty$ gives finite total length. Hence W_t is Cauchy and converges to a limit. Equation (144) forces the limiting residual to be zero, so the limit belongs to Γ_y.

For completeness, the tail length after time T obeys the uniform estimate

$$\int_T^\infty \left\|\dot{W}_t\right\|_{\mathrm{F}} dt \le \frac{b}{a^2} e^{-a^2 T/n}\|g(W_0)\|_2 . \tag{146}$$

Finite-time solutions depend continuously on their starting point, while (146) makes the remaining tails uniformly small on a smaller closed starting neighborhood. This proves continuity of the endpoint map.

It remains only to justify its differentiability. At a regular feasible point, (135) can be written

$$\nabla^2 L = \frac{1}{n} Dg^* Dg.$$

Its kernel is $\ker Dg = T_W \Gamma_y$, and (142) gives a uniform positive spectral gap in the normal directions. Thus the linearization of the negative gradient flow is zero tangent to the equilibrium manifold and uniformly contracting transverse to it. The normally attracting manifold result stated in Appendix F.2 therefore gives a C^2 endpoint projection after shrinking $\mathcal{U}$. This is exactly the claimed smoothness of Φ. □

Proposition 39 (One smooth localization for the deterministic transient and a finite sharpness segment)**.** *Suppose the deterministic trajectory from the initialization converges to a regular endpoint W_{det}. Let $\mathcal{K} \subset \Gamma_y$ be any compact set consisting of regular points. Then there is an open set $\mathcal{U}$ containing the complete reference deterministic trajectory and $\mathcal{K}$ such that deterministic gradient flow is global for every starting point in $\mathcal{U}$, converges to Γ_y, and has a C^2 endpoint map Φ on $\mathcal{U}$.*

Proof. By Proposition 38, choose a neighborhood $\mathcal{V}_\infty$ of the regular deterministic endpoint W_{det} on which the endpoint map is C^2. Since the reference deterministic trajectory converges to W_{det}, there is a finite time T_0 after which the whole reference trajectory stays in $\mathcal{V}_\infty$.

Consider the compact initial segment

$$\{W^{\mathrm{det}}(t; W_0) : 0 \le t \le T_0\}.$$

For a point on this segment, the remaining reference evolution up to time T_0 reaches the interior of $\mathcal{V}_\infty$. Smooth finite-time dependence on initial conditions therefore gives a neighborhood of that point from which the corresponding finite-time flow also reaches $\mathcal{V}_\infty$. On that neighborhood the endpoint equals

$$\text{(endpoint map on } \mathcal{V}_\infty) \circ \text{(a finite-time deterministic flow map)},$$

so it is C^2. Compactness of this initial segment reduces these neighborhoods to a finite subcover. On overlaps the formulas agree because the deterministic ODE has unique trajectories. Together with $\mathcal{V}_\infty$, they therefore patch to a single C^2 endpoint map around the entire reference trajectory.

For each point of $\mathcal{K}$, Proposition 38 gives another local C^2 endpoint neighborhood. Compactness of $\mathcal{K}$ again gives a finite subcover. Add these neighborhoods to the previous finite union. The endpoint maps agree wherever two neighborhoods overlap, again by uniqueness of the deterministic trajectory. Their union is the required $\mathcal{U}$. □

E.4 The stopped slow-time limit, with the stopping removed

For the noisy recursion (6), write its factor iterate as W_k^η. The slow-time theorem removes the rapidly contracting deterministic motion by considering the compensated process

$$\widehat{Y}^\eta(s) = W^\eta_{\lfloor s/\eta^2 \rfloor} - W^{\det}\big(\eta\lfloor s/\eta^2\rfloor; W_0\big) + \Phi(W_0). \tag{147}$$

The second term is the deterministic transient at fast time $\eta\lfloor s/\eta^2\rfloor$. For every fixed $s > 0$ this fast time is asymptotic to s/η and therefore goes to infinity as $\eta \downarrow 0$.

Proposition 40 (Label-noise SGD converges to the sharpness flow on slow time)**.** *Fix a sensing instance and a finite $S > 0$. Suppose an open set $\mathcal{U}$ contains the reference deterministic trajectory and the limiting path below, and suppose the deterministic endpoint map is C^2 on $\mathcal{U}$. Choose a compact $K \Subset \mathcal{U}$ whose interior contains both paths. Then, after stopping at the first slow time at which* (147) *leaves* $\operatorname{int} K$*, the compensated process converges in probability, uniformly for $0 \le s \le S$, to the unique solution of*

$$\frac{dY}{ds} = -\frac{\sigma^2}{4}\operatorname{grad}_{\Gamma_y} \operatorname{tr}\nabla^2 L(Y) = -\frac{\sigma^2}{2}\operatorname{grad}_{\Gamma_y} J(Y), \qquad Y(0) = \Phi(W_0). \tag{148}$$

The probability of hitting the stopping boundary before S tends to zero. Consequently the actual *terminal iterate satisfies*

$$W^\eta_{\lfloor S/\eta^2\rfloor} \longrightarrow Y(S) \quad \textit{in probability as } \eta \downarrow 0. \tag{149}$$

Proof. We apply the stopped weak-convergence theorem and its label-noise specialization stated in Appendix F.2. We check each model-specific hypothesis before using its conclusion.

Smoothness and centering. Each sample loss $\ell_{i,\xi}$ is a polynomial in the entries of U, V. Hence every derivative needed by the external theorem is bounded on the fixed compact set K. The exact centered representation is (133); in particular, the stochastic recursion contains no omitted deterministic or random term.

Normal attraction to the interpolation manifold. On the regular part of Γ_y, (135) gives

$$\nabla^2 L = \frac{1}{n} Dg^* Dg.$$

Because Dg is onto there, this Hessian has rank n, its kernel is $\ker Dg = T_W\Gamma_y$, and it is positive definite on normal directions. Equivalently, deterministic negative-gradient flow has zero linearization along the minimum manifold and strictly negative linearization in every normal direction. The required C^2 endpoint map is available by hypothesis and Proposition 39.

Noise covariance and identification of the limit. By (136), on Γ_y the exact centered covariance equals $\sigma^2\nabla^2 L$. The label-noise specialization of the external theorem therefore identifies every stopped subsequential limit with

$$\frac{dY}{ds} = -\frac{\sigma^2}{4}\operatorname{grad}_{\Gamma_y} \operatorname{tr}\nabla^2 L(Y).$$

By Proposition 15, $\frac{1}{2}\operatorname{tr}\nabla^2 L = J$ on Γ_y, which gives the second equality in (148). On the localized regular set the projected vector field is smooth, so this ODE has a unique solution on the finite interval under consideration. Consequently all convergent subsequences have the same deterministic limit. The entire stopped family therefore converges to Y; convergence in distribution to a deterministic path is convergence in probability. In particular,

$$\sup_{0\le s\le S}\left\|\widehat{Y}^{\eta}(s\wedge\tau_\eta)-Y(s)\right\|_{\mathrm{F}}\longrightarrow 0\quad\text{in probability,}\tag{150}$$

where τ_η is the first slow exit time from $\operatorname{int} K$.

The stopping boundary is asymptotically irrelevant. Since $Y([0,S])$ is compact and lies in $\operatorname{int} K$,

$$d=\operatorname{dist}\big(Y([0,S]),(\operatorname{int}K)^c\big)>0.$$

If $\tau_\eta\le S$, the stopped stochastic path has reached the complement of $\operatorname{int} K$, whereas Y is still at least d away from that complement. Hence

$$\{\tau_\eta\le S\}\subseteq\left\{\sup_{0\le s\le S}\left\|\widehat{Y}^{\eta}(s\wedge\tau_\eta)-Y(s)\right\|_{\mathrm{F}}\ge d\right\}.$$

The probability of the event on the right tends to zero by (150). Thus

$$\mathbb{P}(\tau_\eta\le S)\longrightarrow 0.\tag{151}$$

On $\{\tau_\eta>S\}$ the stopped and unstopped paths are identical up to time S, so (150) and (151) imply

$$\sup_{0\le s\le S}\left\|\widehat{Y}^{\eta}(s)-Y(s)\right\|_{\mathrm{F}}\longrightarrow 0\quad\text{in probability.}\tag{152}$$

Return from the compensated process to the actual iterate. From (147), at $s=S$ we have the exact identity

$$W^{\eta}_{\lfloor S/\eta^2\rfloor}-\widehat{Y}^{\eta}(S)=W^{\det}\big(\eta\lfloor S/\eta^2\rfloor;W_0\big)-\Phi(W_0).$$

Moreover,

$$\eta\left\lfloor\frac{S}{\eta^2}\right\rfloor=\frac{S}{\eta}+O(\eta)\longrightarrow\infty.$$

The deterministic trajectory converges to $\Phi(W_0)$, so the right side tends to zero deterministically. Together with (152), this proves (149). □

We now derive a valid choice of the iteration count N_η in Theorem 1. Put

$$t=\frac{\sigma^2}{2}s.\tag{153}$$

Then (148) becomes

$$\frac{dY}{dt}=-\operatorname{grad}_{\Gamma_y}J(Y),$$

which is exactly the sharpness flow analyzed in section D. To observe that flow at sharpness time $t=T$, one must take

$$s=\frac{2T}{\sigma^2}.$$

Since slow time s corresponds to $\lfloor s/\eta^2\rfloor$ stochastic iterations, the required number is

$$N_\eta=\left\lfloor\frac{2T}{\sigma^2\eta^2}\right\rfloor.\tag{154}$$

E.5 Completion of the noisy branch

Assume only the full-row-rank condition on the common stacked initialization W_0 in (2). By Proposition 37, the deterministic factor flow from this initialization converges to a feasible endpoint W_{det} whose stacked factor state has full row rank. By Lemma 25, this endpoint is regular. No lower bound on α_{init} is used here.

Let $W^{\mathrm{sh}}(t)$ denote the projected sharpness flow starting from W_{det}, and write

$$M^{\mathrm{sh}}(t) = U^{\mathrm{sh}}(t)V^{\mathrm{sh}}(t)^{\mathsf{T}}.$$

By Proposition 36,

$$M^{\mathrm{sh}}(t) \longrightarrow M_\star.$$

Given $\varepsilon \in (0,1)$, choose a finite T such that

$$\left\|M^{\mathrm{sh}}(T) - M_\star\right\|_{\mathrm{F}} \le \frac{\varepsilon}{2}\left\|M_\star\right\|_{\mathrm{F}}. \tag{155}$$

This is the only role of T in the proof: it is a deterministic time at which the limiting sharpness flow is already accurate enough.

The finite sharpness path $\{W^{\mathrm{sh}}(t) : 0 \le t \le T\}$ is compact, and every point on it is regular by Proposition 27. Applying Proposition 39 to this compact path therefore supplies one open neighborhood containing both the entire deterministic reference trajectory and the sharpness path, with a C^2 deterministic endpoint map on that neighborhood. Choose a slightly smaller compact set K whose interior still contains both paths. The hypotheses of Proposition 40 are now satisfied for slow horizon

$$S = \frac{2T}{\sigma^2}.$$

We finally translate convergence of the factors into convergence of their product. Matrix multiplication is continuous; more explicitly, for any two factor pairs (U,V) and $(\widetilde{U},\widetilde{V})$,

$$\begin{aligned}\left\|UV^{\mathsf{T}} - \widetilde{U}\widetilde{V}^{\mathsf{T}}\right\|_{\mathrm{F}} &\le \left\|(U-\widetilde{U})V^{\mathsf{T}}\right\|_{\mathrm{F}} + \left\|\widetilde{U}(V-\widetilde{V})^{\mathsf{T}}\right\|_{\mathrm{F}} \\ &\le \left\|U-\widetilde{U}\right\|_{\mathrm{F}}\|V\|_{\mathrm{op}} + \left\|\widetilde{U}\right\|_{\mathrm{op}}\left\|V-\widetilde{V}\right\|_{\mathrm{F}}.\end{aligned} \tag{156}$$

By Proposition 40 and the time change (153),

$$W^{\eta}_{N_\eta} \longrightarrow W^{\mathrm{sh}}(T) \quad \text{in probability}, \qquad N_\eta = \left\lfloor \frac{2T}{\sigma^2\eta^2} \right\rfloor.$$

The continuous mapping theorem and (156) therefore give

$$U_{N_\eta}V_{N_\eta}^{\mathsf{T}} \longrightarrow M^{\mathrm{sh}}(T) \quad \text{in probability}.$$

Consequently, for the prescribed $\delta \in (0,1)$, every sufficiently small $\eta > 0$ satisfies

$$\mathbb{P}\Big(\left\|U_{N_\eta}V_{N_\eta}^{\mathsf{T}} - M^{\mathrm{sh}}(T)\right\|_{\mathrm{F}} \le \frac{\varepsilon}{2}\left\|M_\star\right\|_{\mathrm{F}}\Big) \ge 1-\delta. \tag{157}$$

On the event in (157), (155) gives

$$\begin{aligned}\left\|U_{N_\eta}V_{N_\eta}^{\mathsf{T}} - M_\star\right\|_{\mathrm{F}} &\le \left\|U_{N_\eta}V_{N_\eta}^{\mathsf{T}} - M^{\mathrm{sh}}(T)\right\|_{\mathrm{F}} + \left\|M^{\mathrm{sh}}(T) - M_\star\right\|_{\mathrm{F}} \\ &\le \frac{\varepsilon}{2}\left\|M_\star\right\|_{\mathrm{F}} + \frac{\varepsilon}{2}\left\|M_\star\right\|_{\mathrm{F}} \\ &= \varepsilon\left\|M_\star\right\|_{\mathrm{F}}.\end{aligned}$$

This proves part (b) on every fixed sensing instance satisfying the deterministic assumptions of the sharpness-flow analysis.

F Assembly of the proof of Theorem 1

F.1 High-probability event and proof of Theorem 1

Fix the common initialization U_0, V_0 from the statement of Theorem 1, chosen independently of the sensing matrices, and use the normalization (26). By the final estimates in section A.6, there is a deterministic large-initialization threshold depending only on κ_{init} and the fixed problem parameters, not on the realized sensing matrices. This threshold will be imposed only for part (a); part (b) uses only the full-row-rank condition on the stacked initialization in (2).

We now collect the random events for this fixed initialization. First, the vectorized Gaussian sensing matrix is an $n \times (d_1 d_2)$ Gaussian matrix. Since $n \le d_1 d_2/2 < d_1 d_2$, it has row rank n almost surely. Thus $\mathcal{A}\mathcal{A}^*$ is invertible almost surely, as required throughout the deterministic analysis.

Second, Proposition 24 gives, under $n \ge Cr(d_1 + d_2)$, an event of probability at least

$$1 - Ce^{-cn} \tag{158}$$

on which S_L, S_R are positive definite and $M_\star$ is the unique minimizer of the weighted nuclear-norm program (89).

Third, the transversality property of Proposition 30 fails only on a Lebesgue-null set of sensing maps and therefore has probability one under the Gaussian law.

Fourth, Lemma 12 gives an event of probability at least

$$1 - Ce^{-cn} \tag{159}$$

on which the deterministic conditioning bounds (70)–(71) hold. If α_{init} is above the deterministic threshold just described, the uniform estimates in section A.6 imply that the hypotheses of Propositions 8 and 9 hold throughout this event for every $0 < \eta \le \eta_{\text{GD}}(\alpha)$.

Fifth, for the present fixed initialization, Proposition 10 gives an event of probability at least

$$1 - 2e^{-c(d_1 d_2 - n)} - e^{-cd_1 d_2} \tag{160}$$

on which, because $n \le d_1 d_2/2$,

$$\|M_{\text{lin}} - M_\star\|_{\text{F}} \ge \frac{1}{\sqrt{8}\,\kappa_{\text{init}}} \|M_\star - M_0\|_{\text{F}}\,. \tag{161}$$

Intersect the three high-probability events (158), (159), and (160) with the two probability-one events. A union bound gives probability at least

$$1 - Ce^{-cn} - 2e^{-c(d_1 d_2 - n)} - e^{-cd_1 d_2}.$$

Since $n \le d_1 d_2/2$, the last two exponential terms are also bounded by constant multiples of e^{-cn}. After adjusting the universal constants, the intersection therefore has probability at least $1 - Ce^{-cn}$, as stated in Theorem 1.

On every sensing instance in this common event, if α_{init} is above the deterministic threshold, the conclusion of section A.6 applies. Consequently, for every $0 < \eta \le \eta_{\text{GD}}(\alpha)$, deterministic full-batch gradient descent with factor step η converges to an exact interpolator and

$$\frac{\left\|\widehat{M}_{\text{GD}}(\alpha,\eta) - M_\star\right\|_{\text{F}}}{\|M_\star\|_{\text{F}}} \ge \frac{1}{4\sqrt{2}\,\kappa_{\text{init}}} \frac{\|M_\star - M_0\|_{\text{F}}}{\|M_\star\|_{\text{F}}}.$$

Now fix any $\sigma > 0$ and any $\varepsilon, \delta \in (0,1)$. By Proposition 37, the deterministic factor flow from the common initialization converges to a full-row-rank interpolating stack, with no large-initialization assumption. All four fixed-instance assumptions listed at the beginning of section D are therefore available. The argument of the preceding subsection produces a finite T such that, for all sufficiently small $\eta > 0$,

$$\mathbb{P}\left(\frac{\left\|U_{N_\eta} V_{N_\eta}^{\mathsf{T}} - M_\star\right\|_{\text{F}}}{\|M_\star\|_{\text{F}}} \le \varepsilon\right) \ge 1 - \delta, \qquad N_\eta = \left\lfloor \frac{2T}{\sigma^2 \eta^2} \right\rfloor.$$

The noisy conclusion holds on an interval of all sufficiently small positive steps. If the hypothesis of part (a) on α_{init} is also imposed, the deterministic conclusion holds whenever $\eta \le \eta_{\text{GD}}(\alpha)$, and the intersection again contains all sufficiently small positive η. The required smallness may depend on the fixed initialization (and, for the noisy branch, on $\varepsilon, \delta, \sigma$), consistent with the statement of the theorem. The probability in part (b) is conditional on the already fixed sensing instance, so it concerns only the random indices and label perturbations used by the algorithm.

Thus part (b) holds for every full-stack-rank common initialization, while part (a) holds under its additional large-initialization hypothesis, on the same sensing event. This completes the proof of Theorem 1.

F.2 External results used in the proof

This appendix records the general results that are invoked without proof. We state only the versions needed in the main argument. Each subsection first gives the mathematical statement, then explains any terminology that is not standard and finally indicates where the result is used.

Gaussian singular values and matrix deviation

We use two standard facts about Gaussian matrices; see (Vershynin, 2018, Chapters 4 and 9).

Proposition 41 (Extreme singular values of a Gaussian matrix)**.** *Let $G \in \mathbb{R}^{a\times b}$ have independent $N(0,1)$ entries, with $a \le b$. Then, for every $t \ge 0$, with probability at least $1 - 2e^{-t^2/2}$,*

$$\sqrt{b} - \sqrt{a} - t \le \sigma_{\min}(G) \le \sigma_{\max}(G) \le \sqrt{b} + \sqrt{a} + t.$$

For the second statement, let $T \subset \mathbb{R}^D$ be a symmetric set, meaning $T = -T$. Its *Gaussian width* is

$$w(T) = \mathbb{E} \sup_{x\in T} \langle g, x\rangle, \qquad g \sim N(0, I_D).$$

The width is a measure of the effective size of T: a set with many possible directions has larger Gaussian width.

Proposition 42 (Gaussian matrix deviation on a set)**.** *There are numerical constants $c, C > 0$ such that the following holds. Let $G \in \mathbb{R}^{n\times D}$ have independent $N(0,1)$ entries and let $T \subset \{x \in \mathbb{R}^D : \|x\|_2 = 1\}$ be symmetric. Then, for every $u \ge 0$, with probability at least $1 - 2e^{-cu^2}$,*

$$\sup_{x\in T} \big|\|Gx\|_2 - \sqrt{n}\big| \le C\big(w(T) + u\big).$$

Use in this proof. Proposition 41 is used in Lemma 18 to control the two empirical marginal covariance matrices. Proposition 42 is used in Lemma 19 with T equal to the unit-Frobenius-norm matrices of rank at most s. The bound on the width of that particular set is proved directly in (94).

Parametric transversality

We use the following standard form of the parametric transversality theorem; see Hirsch (1976).

For a smooth map $H : X \to \mathbb{R}^n$ defined on a smooth manifold X, the value $0 \in \mathbb{R}^n$ is called a *regular value* if, at every x satisfying $H(x) = 0$, the derivative $DH(x)$ maps the tangent space T_xX onto all of $\mathbb{R}^n$. In this special case, saying that H is *transverse to* 0 means exactly the same thing. "Onto" means that every vector in the target space is the derivative of some admissible perturbation in the domain.

Proposition 43 (Parametric transversality, form used here)**.** *Let P be a finite-dimensional Euclidean parameter space, let X be a smooth manifold, and let*

$$F : P \times X \to \mathbb{R}^n$$

be smooth. Suppose that for every $(p, x) \in P \times X$, the derivative with respect to the parameter,

$$D_pF(p, x) : P \to \mathbb{R}^n,$$

is onto. Then, for Lebesgue-almost every fixed parameter $p \in P$, the map $F_p(x) = F(p, x)$ has 0 as a regular value. Equivalently, whenever $F(p, x) = 0$, the derivative $D_xF(p, x) : T_xX \to \mathbb{R}^n$ is onto.

Use in this proof. In Proposition 30, the parameter p is the sensing map $\mathcal{A}$, the state x is a fixed-rank matrix M, and $F(\mathcal{A}, M) = \mathcal{A}(M - M_\star)$. The proof there explicitly shows that $D_{\mathcal{A}}F$ is onto, so Proposition 43 gives the required generic regularity of every incorrect fixed-rank exact fit.

The Kurdyka–Łojasiewicz inequality

The convergence argument uses the Kurdyka–Łojasiewicz (KL) inequality for semialgebraic functions; see Bolte et al. (2007). The cited theorem is phrased using Clarke subgradients; it implies the form below for the limiting subdifferential used in this proof. The compact version follows by uniformizing the pointwise inequality over a finite cover of the compact set.

A function f is *lower semicontinuous* if $f(x) \le \liminf_j f(x_j)$ whenever $x_j \to x$; informally, its value cannot jump downward at a limit. A set is *semialgebraic* if it can be described by finitely many polynomial equalities and inequalities, together with finitely many unions and intersections. A function is semialgebraic when its graph is semialgebraic. For a closed constraint set C, its indicator function is

$$\iota_C(x) = \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C. \end{cases}$$

Thus minimizing $J + \iota_C$ is the same as minimizing J subject to $x \in C$. The *limiting subdifferential* $\partial f(x)$ is the standard generalized gradient used for nonsmooth constrained optimization. At a smooth point of a constraint manifold C,

$$\partial(J + \iota_C)(x) = \nabla J(x) + N_C(x),$$

where $N_C(x)$ is the normal space to C. We write $\operatorname{dist}(0, \partial f(x)) = \inf\{\|v\|_2 : v \in \partial f(x)\}$.

Proposition 44 (Uniform KL inequality on a compact level set)**.** *Let f be a lower-semicontinuous semialgebraic function and let Ω be a compact set on which f is finite and constant, with value $f_\star$. Then there are a neighborhood $\mathcal{V}$ of Ω, a number $\eta > 0$, and an increasing concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that $\varphi(0) = 0$, φ is C^1 on $(0, \eta)$, $\varphi'(s) > 0$, and*

$$\varphi'\big(f(x) - f_\star\big)\operatorname{dist}\big(0, \partial f(x)\big) \ge 1$$

whenever $x \in \mathcal{V}$ and $f_\star < f(x) < f_\star + \eta$.

Use in this proof. We apply Proposition 44 in Proposition 28 to $f = J + \iota_{\Gamma_y}$ and to the compact cluster set of the sharpness-flow trajectory. Since J and the equations defining Γ_y are polynomial, this f is semialgebraic. The resulting inequality is exactly (116).

Smooth endpoint map near a normally attracting manifold

The next input is a local consequence of the normally attracting invariant manifold theorem; see Hirsch et al. (1977). We state it directly for gradient flow, which is the only case needed here.

Let $\Gamma \subset \mathbb{R}^D$ be a smooth manifold. Its tangent space at x is $T_x\Gamma$, and its *normal space* is the orthogonal complement $N_x\Gamma = (T_x\Gamma)^\perp$. A manifold of equilibria is *normally attracting* when there is no first-order motion along the manifold, while perturbations normal to it are contracted at a uniform exponential rate.

Proposition 45 (Smooth endpoint map near a normally attracting equilibrium manifold)**.** *Let $L : \mathbb{R}^D \to \mathbb{R}$ be C^3, let Γ be a C^2 manifold of critical points of L, and let $K \subset \Gamma$ be compact. Suppose that, on a neighborhood of K inside Γ,*

$$\ker \nabla^2 L(x) = T_x\Gamma$$

and that there is a constant $\lambda > 0$ such that

$$\big\langle v, \nabla^2 L(x) v \big\rangle \ge \lambda \|v\|_2^2 \qquad \text{for every } v \in N_x\Gamma.$$

Assume also that, on some neighborhood $\mathcal{U}_0$ of K, the negative gradient flow is defined for all $t \ge 0$, stays in the region where these conditions hold, and converges to a point of Γ. Then, after shrinking $\mathcal{U}_0$ if necessary, the endpoint map

$$\Phi(x) = \lim_{t \to \infty} x(t), \qquad \dot{x}(t) = -\nabla L(x(t)),$$

is C^2.

Use in this proof. In Proposition 38, the identity $\nabla^2 L = n^{-1} Dg^* Dg$ on Γ_y shows that the Hessian vanishes exactly in tangent directions and is uniformly positive in normal directions on the compact regular set under consideration. Proposition 45 then supplies the C^2 endpoint map required by the slow-time SGD theorem.

SLOW-TIME LIMIT OF LABEL-NOISE SGD

Finally, we use the stopped weak-convergence theorem of Li, Wang, and Arora and its label-noise specialization; see (Li et al., 2022, Theorems B.8–B.9 and Corollary 5.2). The following formulation extracts exactly the consequence needed in this proof.

Consider a centered SGD recursion

$$x_{k+1}^\eta = x_k^\eta - \eta\Big(\nabla L(x_k^\eta) + \sqrt{N_{\mathrm{sgd}}}\,\omega_{\xi_k}(x_k^\eta)\Big), \qquad \xi_k \sim \mathrm{Unif}\{1,\dots,N_{\mathrm{sgd}}\}, \tag{162}$$

where N_{sgd} is the number of possible stochastic-gradient components and

$$\sum_{j=1}^{N_{\mathrm{sgd}}} \omega_j(x) = 0, \qquad \Sigma(x) = \sum_{j=1}^{N_{\mathrm{sgd}}} \omega_j(x)\omega_j(x)^{\mathsf{T}}.$$

Here Σ is the covariance of the centered stochastic-gradient fluctuation. Let $\phi(x,t)$ denote deterministic gradient flow at time t, and let $\Phi(x) = \lim_{t\to\infty} \phi(x,t)$ be its endpoint whenever the limit exists.

Proposition 46 (Slow-time limit, specialized form used here)**.** *Assume that L is C^3 and that, on an open set $\mathcal{U}$, the set Γ of minima is a C^2 manifold, the Hessian has constant rank equal to the codimension of Γ, deterministic gradient flow from $\mathcal{U}$ converges to Γ, and the endpoint map Φ is C^2. Assume that each ω_j in (162) is C^2 on $\mathcal{U}$.*

Fix a compact $K \Subset \mathcal{U}$, an initial point $x_0 \in K$, and a finite slow-time horizon $S > 0$. Define the compensated slow-time path

$$\widehat{Y}^\eta(s) = x^\eta_{\lfloor s/\eta^2 \rfloor} - \phi\big(x_0, \eta\lfloor s/\eta^2 \rfloor\big) + \Phi(x_0), \qquad 0 \le s \le S.$$

Stop this path at its first exit from the interior of K. As $\eta \downarrow 0$, every weak limit of the stopped paths stays on Γ and is given by the limiting diffusion of Li, Wang, and Arora.

If, in addition,

$$\Sigma(x) = c\,\nabla^2 L(x) \qquad \text{for every } x \in \Gamma \tag{163}$$

for a constant $c > 0$, then the limiting diffusion has no Brownian component along Γ and reduces to the deterministic ODE

$$\frac{dY}{ds} = -\frac{c}{4}\,\mathrm{grad}_\Gamma \,\mathrm{tr}\,\nabla^2 L(Y), \qquad Y(0) = \Phi(x_0). \tag{164}$$

Here grad_Γ means the Euclidean gradient projected onto the tangent space of Γ. If this ODE has a unique solution on a fixed finite interval, the stopped SGD paths converge in distribution to that solution; because the limit is deterministic, this is equivalently convergence in probability in the uniform path norm.

Terminology. "Weak convergence" here means convergence in distribution of the whole path on a fixed finite time interval, rather than pointwise convergence of a single iterate. The stopping at K is a localization device: it ensures all derivatives are controlled on one compact set. In the main proof, the deterministic limiting path stays a positive distance from the boundary of K, which makes the stopping asymptotically irrelevant.

Use in this proof. In section E, we take $N_{\mathrm{sgd}} = 2n$ and the functions $\omega_{i,\xi}$ from (132). Exact centering is verified in (133), while (136) gives $c = \sigma^2$ in (163). Therefore (164) becomes

$$\frac{dY}{ds} = -\frac{\sigma^2}{4}\,\mathrm{grad}_{\Gamma_y} \,\mathrm{tr}\,\nabla^2 L(Y),$$

which is the slow sharpness flow used in Proposition 40.