Two-Step Optimal Learning-Rate Geometry under μP

Generic Uniqueness, Exceptional Nonuniqueness, and Perturbative Stability

Siyuan Feng

Mathematics Institute
University of Warwick

Abstract

Width-stable learning-rate loss curves do not by themselves determine a transferable scalar hyperparameter: a single learning rate is identified only when the limiting loss has a unique global minimizer. We study this optimizer-identification problem in the first exactly solvable nonlinear multisample setting, a deep linear μP network with two trained matrices and two full-batch gradient updates. For every fixed sample size, we derive the exact limiting learning-rate landscape and show that its global minimizer is generically unique over positive-definite data. This scalar conclusion is not universal: we construct a full-rank multisample family with exactly three distinct positive nondegenerate global minimizers. We then establish perturbative and finite-width stability, separating a generic regime with single-learning-rate transfer from an exceptional regime naturally described by optimizer sets and persistent local-minimum branches.

1 Introduction

The maximal-update parametrization (μP) preserves feature learning in the infinite-width limit and has motivated zero-shot hyperparameter transfer across model widths [17, 18]. Learning-rate transfer is a central example: one tunes a learning rate on a narrower model and reuses it at larger widths, avoiding a new search at every scale. A mathematical theory of this phenomenon must therefore explain not only why the learning-rate loss stabilizes, but also why it selects a reproducible learning rate.

This distinction is essential. Compact convergence $\mathcal{L}_n(\eta) \to \mathcal{L}_\infty(\eta)$ controls objective values and the possible accumulation points of minimizers, but it does not by itself identify a scalar limit. If $\arg\min \mathcal{L}_\infty$ is a singleton, compact argmin consistency yields one canonical learning rate. If the limiting loss has several equal-depth minimizers, the same limiting objective specifies only an optimizer set: finite-width fluctuations, data perturbations, or the search procedure may select different branches. Thus uniqueness is not merely a technical convenience in an existing theorem. It is the identifiability condition that turns a width-stable objective into a well-defined scalar hyperparameter prescription. Nonuniqueness need not imply worse attainable loss; it means that "the optimal learning rate" is not intrinsically single-valued without an additional selection rule.

Hayou [8] proves compact learning-rate transfer for deep linear μP networks under uniqueness of the limiting global minimizer. The one-step case does not expose the geometry behind this assumption: its limiting loss is quadratic and hence has a uniquely determined optimizer whenever the problem is nondegenerate. After two updates, the first learned weights feed back into the second gradient, higher-order interactions survive at infinite width, and the learning-rate loss becomes a data-dependent polynomial with a genuinely nonlinear global-minimum geometry. The case $L = 2, t = 2$ is therefore the first setting in which the infinite-width limit may fail to identify one transferable learning rate, while remaining explicit enough for an exact analysis.

We ask three questions. First, is a unique limiting optimal learning rate typical for positive-definite multisample data? Second, can exact nonuniqueness persist for genuine full-rank datasets rather than only in scalar or rank-deficient examples? Third, how do these two regimes behave under perturbations of the data and at finite width?

Main contributions. Our answers are as follows.

1. We derive an exact finite-width state compression and a closed arbitrary-m formula for the two-step limiting residual. The entire limiting landscape lies in $\mathrm{span}\{y, Ky, K^2y\}$ and admits a finite spectral-measure representation.
2. We introduce a critical-value discriminant and prove that uniqueness holds on a nonempty Zariski-open, Euclidean-open dense, full-Lebesgue-measure subset of the positive-definite parameter space for every fixed $m \geq 2$.
3. We construct a positive-definite full-rank multisample family with exactly three distinct positive nondegenerate global minimizers. Hence generic uniqueness is not a universal single-learning-rate principle.
4. We prove perturbative argmin-set stability and finite-width consequences. Generic data support ordinary compact transfer to one scalar learning rate, whereas exceptional data exhibit set-valued global behavior and persistent local-minimum branches.

2 Related work and positioning

μP and hyperparameter transfer. Tensor Programs IV identifies parametrizations that retain feature learning at infinite width and introduces μP as the maximal-update choice [4, 10, 13, 17]. Subsequent work develops zero-shot hyperparameter transfer and studies width and depth scaling, optimizer scaling, and sharpness-related consistency [2, 5, 7, 12, 14, 15, 18–20]. These works establish or motivate stabilization across scale. Our question is downstream of that stabilization: when the limiting learning-rate objective exists, does it identify one hyperparameter or only a set of equally optimal choices?

Learning-rate transfer and argmin geometry. Hayou [8] formulates learning-rate transfer as convergence of an optimal learning rate and proves a compact result conditional on uniqueness of the limiting minimizer. We study the geometry responsible for that condition rather than assuming it. This separates three logically distinct phenomena: convergence of loss curves, convergence of optimal values, and identification of a unique optimizer. Our results show that the last property is generic in the first nonlinear two-step model, but that structured exceptions require a set-valued or branchwise formulation.

Exact deep-linear dynamics. Deep linear networks provide tractable models of nonconvex training dynamics [1, 3, 6, 9, 11, 16]. Existing exact analyses primarily describe training trajectories, mode learning, or convergence in time. Here the optimization variable is different: after exactly two gradient updates, we regard the base learning rate itself as the variable and analyze the equal-depth global minima of the resulting degree-ten loss. The spectral reduction connects this hyperparameter geometry to the eigenspaces of the data Gram matrix, while the discriminant and perturbation arguments connect it to optimizer identifiability and stability.

3 Problem setup and why uniqueness matters

For fixed inputs $x_i \in \mathbb{R}^d$ and labels $y_i \in \mathbb{R}$, consider

$$f_n(x) = V^\top W_2 W_1 W_0 x, \tag{1}$$

where $W_0 \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^n$ are frozen and $W_1, W_2 \in \mathbb{R}^{n\times n}$ are trained. All initialization variables are independent, with

$$(W_0)_{ab} \sim N(0, d^{-1}), \qquad (W_\ell)_{ab} \sim N(0, n^{-1})\ (\ell = 1, 2), \qquad V_a \sim N(0, n^{-2}). \tag{2}$$

The full-batch loss and simultaneous gradient updates are

$$\mathcal{L}_n(W_1, W_2) = \frac{1}{2m}\sum_{i=1}^{m}\big(f_n(x_i) - y_i\big)^2, \tag{3}$$

$$W_\ell^{(s+1)} = W_\ell^{(s)} - \eta\nabla_{W_\ell}\mathcal{L}_n, \qquad \ell = 1, 2, \quad s = 0, 1. \tag{4}$$

Let $f_n^{(s)} \in \mathbb{R}^m$ collect the outputs after s updates, $r_n^{(s)} = f_n^{(s)} - y$, and

$$K_{ij} = \frac{\langle x_i, x_j \rangle}{d}, \qquad K \succeq 0. \tag{5}$$

The two-step learning-rate loss is $\mathcal{L}_n^{(2)}(\eta) = \|r_n^{(2)}(\eta)\|^2/(2m)$.

Scalar versus set-valued transfer. Whenever a deterministic limiting loss $\mathcal{L}_\infty^{(2)}$ exists, its relevant object is

$$\mathcal{A}_\infty = \arg\min_{\eta \geq 0} \mathcal{L}_\infty^{(2)}(\eta).$$

If $\mathcal{A}_\infty = \{\eta_\infty\}$, the limiting landscape identifies one scalar learning rate and compact argmin convergence can yield $\eta_n \to \eta_\infty$. If $\mathcal{A}_\infty$ contains several points, convergence of the loss curves determines only the set of admissible limiting optimizers; convergence of a particular finite-width selection then requires an additional rule or a branchwise description. Throughout, *nonunique* means at least two distinct equal-depth global minimizers. A finite-width *branch* is a local minimum in a prescribed neighborhood of an isolated limiting local minimum; it need not be globally optimal at finite width.

3.1 The first failure of scalar identifiability

The uniqueness issue already appears in the smallest two-step model.

Proposition 3.1 (Minimal counterexample)**.** *Let $m = 1$, $L = 2$, and $t = 2$, with $x \neq 0$ and $y \neq 0$. Put*

$$\kappa = \frac{\|x\|^2}{d}, \qquad a = \kappa\eta, \qquad \theta = \frac{y^2}{\kappa} > 0.$$

The limiting residual factorizes as

$$\frac{r_\infty^{(2)}(\eta)}{y} = (2a-1)Q_\theta(a), \qquad Q_\theta(a) = -(2a-1) + 4\theta a^3(a-1). \tag{6}$$

Consequently, the limiting loss has at least three distinct positive global minimizers, with one in each of $(0, \frac{1}{2})$, $\{\frac{1}{2}\}$, and $(1, \infty)$ in the rescaled variable a.

Proof sketch. The factorization (6) and the sign changes of Q_θ give two further positive zeros; all three zeros have zero loss. See Appendix A.1. □

Thus nonuniqueness is not a merely logical possibility. The remaining questions are whether it survives for positive-definite multisample data and whether a single optimizer is nevertheless generic.

4 Exact two-step learning-rate landscape

The two-step network contains two trained $n \times n$ matrices, but its dependence on the learning rate admits an exact low-dimensional compression. This reduction is the bridge between the neural-network dynamics and the global-minimum geometry studied later.

4.1 Exact finite-width compression

Let

$$\begin{gathered} U = [W_0 x_1, \ldots, W_0 x_m], \qquad \widetilde{v} = nV, \\ A_s = W_1^{(s)} U, \qquad q_s = (W_2^{(s)})^\top \widetilde{v}, \qquad K_n = \frac{1}{n} U^\top U, \qquad \nu_n = \frac{1}{n}\|\widetilde{v}\|^2. \end{gathered} \tag{7}$$

Proposition 4.1 (Exact compressed recursion)**.** *For every width, initialization, and learning rate, and for* $s = 0, 1$*,*

$$f_n^{(s)} = \frac{1}{n} A_s^\top q_s, \qquad r_n^{(s)} = f_n^{(s)} - y, \tag{8}$$

$$A_{s+1} = A_s - \frac{\eta}{m} q_s (r_n^{(s)})^\top K_n, \tag{9}$$

$$q_{s+1} = q_s - \frac{\eta}{m} \nu_n A_s r_n^{(s)}. \tag{10}$$

Proof sketch. Direct differentiation and multiplication of the two matrix updates by U and $\widetilde{\nu}$ yield the closed recursion. See Appendix A.2. □

The proposition removes the ambient $n \times n$ matrices from the subsequent analysis: once the initial empirical moments are known, the two-step output is determined by a finite collection of forward, backward, and Gram-type quantities.

4.2 Closed limiting residual

Write

$$z = \frac{\eta}{m}. \tag{11}$$

The normalized initialization moments in Proposition 4.1 converge almost surely to deterministic Gaussian moments; these limits are also direct instances of the Tensor Programs IV Master Theorem [17, Theorem 7.4].

Theorem 4.2 (Exact limiting two-step dynamics)**.** *For fixed* m, d *and deterministic data, almost surely,*

$$f_\infty^{(1)}(z) = 2zKy, \qquad r_\infty^{(1)}(z) = (2zK - I)y, \tag{12}$$

$$f_\infty^{(2)}(z) = A(z)Ky + B(z)K^2 y, \qquad r_\infty^{(2)}(z) = -y + A(z)Ky + B(z)K^2 y. \tag{13}$$

where

$$\begin{aligned} A(z) &= 4z + 4(y^\top Ky)z^3 - 6\|Ky\|^2 z^4, \\ B(z) &= -4z^2 - 6(y^\top Ky)z^4 + 8\|Ky\|^2 z^5. \end{aligned} \tag{14}$$

The deterministic limiting loss is

$$\mathcal{L}_\infty^{(2)}(\eta; K, y) = \frac{1}{2m} \left\| -y + A(\eta/m)Ky + B(\eta/m)K^2 y \right\|^2. \tag{15}$$

Proof sketch. At initialization, the row laws converge to independent Gaussian variables with covariances K and 1. The first compressed update determines the three moments $\mathbb{E}[a_1 q_1]$, $\mathbb{E}[a_1 a_1^\top]$, and $\mathbb{E}[q_1^2]$; substituting these into the second update and collecting the Ky and $K^2 y$ components gives (13). See Appendices A.3 and B. □

The structural point is more important than the expanded polynomial: independently of width, the limiting residual belongs to $\mathrm{span}\{y, Ky, K^2 y\}$. Thus the two-step hyperparameter landscape is governed by a low-order Krylov geometry of the data rather than by the ambient parameter dimension.

4.3 Polynomial geometry

Proposition 4.3 (Exact degree and boundary behavior)**.** *If* $Ky \neq 0$*, then* $r_\infty^{(2)}$ *has exact degree five and* $\mathcal{L}_\infty^{(2)}$ *has exact degree ten, with*

$$[z^5] r_\infty^{(2)}(z) = 8\|Ky\|^2 K^2 y \neq 0, \tag{16}$$

$$[\eta^{10}] \mathcal{L}_\infty^{(2)}(\eta) = \frac{32\|Ky\|^4 \|K^2 y\|^2}{m^{11}} > 0, \tag{17}$$

$$\left(\mathcal{L}_\infty^{(2)}\right)'(0) = -\frac{4}{m^2} y^\top Ky < 0. \tag{18}$$

If $Ky = 0$*, then* $f_\infty^{(2)} \equiv 0$ *and* $\mathcal{L}_\infty^{(2)} \equiv \|y\|^2/(2m)$.

Corollary 4.4 (Coercivity and minimizer cardinality)**.** *If $Ky \neq 0$, the limiting loss is coercive on $[0, \infty)$ and its global-minimizer set is a nonempty finite subset of $(0, \infty)$ containing at most five points. If $Ky = 0$, every $\eta \geq 0$ is a global minimizer.*

The positive leading coefficient controls the large-learning-rate boundary, while the negative derivative at zero excludes the origin. The cardinality bound follows because each distinct global minimizer is a root of even multiplicity of a degree-ten polynomial. Full proofs are given in Appendix A.4.

4.4 Spectral reduction and interpretation

Let $\lambda_1, \ldots, \lambda_q \geq 0$ be the distinct eigenvalues of K, with orthogonal spectral projectors $P_1, \ldots, P_q$, and put

$$w_a = \|P_a y\|^2.$$

Proposition 4.5 (Spectral representation)**.** *Define*

$$p_\lambda(z) = -1 + \lambda A(z) + \lambda^2 B(z). \tag{19}$$

Then

$$r_\infty^{(2)}(z) = \sum_{a=1}^{q} p_{\lambda_a}(z) P_a y, \qquad \mathcal{L}_\infty^{(2)}(mz) = \frac{1}{2m} \sum_{a=1}^{q} w_a p_{\lambda_a}(z)^2. \tag{20}$$

Hence the landscape depends on (K, y) only through the eigenvalues λ_a and the corresponding weights w_a.

Proof sketch. Apply the spectral projectors to (13) and use orthogonality. See Appendix A.5. □

This representation separates the contribution of each eigenspace of K: the term associated with λ_a appears precisely when $P_a y \neq 0$, with weight $w_a = \|P_a y\|^2$. In Theorem 6.1, we specialize the general spectral decomposition $y = \sum_a P_a y$ to $y = y_0 + y_\lambda$, where $y_0 \in \ker K$ and y_λ lies in a single positive eigenspace of K; equivalently, all other spectral components of y vanish. Hence only the weights corresponding to 0 and λ remain. Since $p_0(z) = -1$, the zero-eigenspace contributes only a constant, while the remaining term is a positive multiple of $p_\lambda(z)^2$. This reduction is the basis of the three-minimizer construction in Section 6.

5 Generic uniqueness

5.1 Critical-value discriminant

Nonuniqueness can be detected through a collision of critical values. By Corollary 4.4, every global minimizer in the nondegenerate case lies in $(0, \infty)$ and is therefore a critical point. Hence two distinct global minimizers force two critical points to have the same function value. The resultant below eliminates the location variable z and records the critical values as roots in u, while its discriminant detects whether such roots collide.

Write

$$P_{K,y}(z) = \mathcal{L}_\infty^{(2)}(mz; K, y). \tag{21}$$

For $Ky \neq 0$, this is a degree-ten polynomial. Define the critical-value resultant

$$C_{K,y}(u) = \mathrm{Res}_z\big(P'_{K,y}(z), u - P_{K,y}(z)\big) \tag{22}$$

and its discriminant

$$\Delta_m(K, y) = \mathrm{disc}_u\, C_{K,y}(u). \tag{23}$$

The roots of $C_{K,y}$ are the critical values of $P_{K,y}$, counted with multiplicity; thus $\Delta_m(K, y) \neq 0$ means that these critical values are pairwise distinct. After clearing fixed powers of m and numerical denominators, Δ_m is a polynomial in the independent entries of K and the entries of y.

Theorem 5.1 (Critical-value criterion)**.** *Assume $K \succeq 0$ and $Ky \neq 0$. If $\Delta_m(K, y) \neq 0$, then $\mathcal{L}_\infty^{(2)}$ has a unique global minimizer on $[0, \infty)$. Conversely, two or more distinct global minimizers imply $\Delta_m(K, y) = 0$.*

Proof sketch. Up to a nonzero factor, $C_{K,y}(u)$ is the product of $u - P_{K,y}(\zeta)$ over the critical points ζ of $P_{K,y}$. Two distinct global minimizers are interior by Corollary 4.4 and have the same critical value, producing a repeated root of $C_{K,y}$ and hence $\Delta_m = 0$. Conversely, $\Delta_m \neq 0$ makes all critical values distinct, so at most one of them can be the global minimum. See Appendix A.6. □

The condition $\Delta_m(K, y) \neq 0$ is sufficient, but not necessary, for uniqueness: Δ_m may also vanish because of a degenerate non-global critical point or a collision between non-global critical values. A complete classification of the discriminant-zero locus is unnecessary for the genericity argument below; it is enough to identify a generic subset on which uniqueness holds.

5.2 Generic uniqueness theorem

For fixed $m \geq 2$, define the positive-definite parameter space[1]

$$\mathcal{P}_m = \{(K, y) : K \in \mathbb{S}^m_{++},\ y \in \mathbb{R}^m \setminus \{0\}\}. \tag{24}$$

Theorem 5.2 (Generic uniqueness for every fixed $m \geq 2$)**.** *For every fixed $m \geq 2$, the set*

$$\mathcal{U}_m = \{(K, y) \in \mathcal{P}_m : \Delta_m(K, y) \neq 0\}$$

is nonempty and Zariski open. It is also Euclidean open dense and has full Lebesgue measure in $\mathcal{P}_m$. Every point of $\mathcal{U}_m$ has a unique limiting global minimizer.

Proof sketch. For $m = 2$, the exact witness $K = \mathrm{diag}(1, 2)$ and $y = (1, 1)^\top$ has $\Delta_2(K, y) \neq 0$. A block-diagonal embedding with zero label coordinates preserves the witness landscape up to a positive scalar for every $m > 2$, so Δ_m is not the zero polynomial. Its nonvanishing set is therefore Zariski open; the zero set of a nonzero real polynomial has empty Euclidean interior and Lebesgue measure zero, which gives density and full measure. See Appendix A.7. □

The three genericity statements have complementary meanings. Euclidean openness says that a point certified by $\Delta_m \neq 0$ remains in the unique regime under all sufficiently small perturbations, while density says that every positive-definite dataset can be approximated arbitrarily closely by such points. Full Lebesgue measure means that any absolutely continuous random choice of (K, y) lies in the certified unique regime with probability one. The exceptional nonunique cases are therefore genuine but confined to a proper algebraic subset.

For literal datasets, the full positive-definite statement is realizable whenever $d \geq m$. No rank-constrained genericity claim is made when $d < m$.

5.3 Arbitrarily small restoration of uniqueness

Equip $\mathcal{P}_m$ with the metric

$$\|(K, y) - (\widetilde{K}, \widetilde{y})\| = \|K - \widetilde{K}\|_F + \|y - \widetilde{y}\|_2. \tag{25}$$

Corollary 5.3 (Arbitrarily small perturbations restore uniqueness)**.** *Fix $m \geq 2$. For every $(K, y) \in \mathcal{P}_m$ and every $\varepsilon > 0$, there exists $(\widetilde{K}, \widetilde{y}) \in \mathcal{P}_m$ such that*

$$\|K - \widetilde{K}\|_F + \|y - \widetilde{y}\|_2 < \varepsilon$$

and the limiting two-step loss for $(\widetilde{K}, \widetilde{y})$ has a unique global minimizer on $[0, \infty)$.

Proof. By Theorem 5.2, $\mathcal{U}_m$ is dense in $\mathcal{P}_m$. Therefore every open ε-ball about (K, y) contains a point of $\mathcal{U}_m$, and Theorem 5.1 gives uniqueness at that point. □

Thus generic perturbations restore uniqueness and the nonunique locus has no interior. The statement is existential: it does not assert that every perturbation, or all sufficiently small perturbations, restore uniqueness.

[1] Here $\mathbb{S}^m_{++} = \{K \in \mathbb{R}^{m \times m} : K^\top = K,\ x^\top K x > 0 \text{ for every } x \neq 0\}$ denotes the set of real symmetric positive-definite $m \times m$ matrices.

6 Structured full-rank nonuniqueness

Theorem 6.1 (Exactly three minima on one positive eigenspace)**.** *Let $K \succeq 0$ and suppose*

$$y = y_0 + y_\lambda, \qquad y_0 \in \ker K, \qquad K y_\lambda = \lambda y_\lambda, \qquad \lambda > 0, \qquad y_\lambda \neq 0. \tag{26}$$

Then $\mathcal{L}_\infty^{(2)}$ has exactly three distinct positive global minimizers. With

$$a = \frac{\lambda \eta}{m}, \qquad \theta = \frac{\|y_\lambda\|^2}{\lambda} > 0,$$

their rescaled locations are the three positive roots of

$$p_\theta(a) = (2a-1)Q_\theta(a), \qquad Q_\theta(a) = -(2a-1) + 4\theta a^3(a-1). \tag{27}$$

There is exactly one root in each of $(0, \frac{1}{2})$, $\{\frac{1}{2}\}$, and $(1, \infty)$. All three minima are isolated and nondegenerate.

Proof sketch. Proposition 4.5 reduces the loss to a constant plus a positive multiple of $p_\theta(a)^2$; the root analysis gives exactly three simple positive zeros. See Appendix A.8. □

Corollary 6.2 (Full-rank multisample nonuniqueness)**.** *For every $m \geq 2$, let*[2]

$$K = I_m + \mathbf{1}\mathbf{1}^\top, \qquad y = c\mathbf{1}, \qquad c \neq 0. \tag{28}$$

Then $K \succ 0$, all labels are nonzero, the inputs realizing K can be chosen pairwise distinct, and the limiting loss has exactly three distinct positive nondegenerate global minimizers.

Proof. $K\mathbf{1} = (m+1)\mathbf{1}$, while the other eigenvalues equal 1. Thus Theorem 6.1 applies with $\lambda = m+1$ and $y_0 = 0$. Moreover, $K_{ii} + K_{jj} - 2K_{ij} = 2 > 0$ for $i \neq j$, which is the squared normalized distance between any two realizing inputs; hence they are distinct. □

This nonunique family is positive-dimensional but not ambient-open. Theorem 5.2 places every positive-definite nonunique point inside the proper algebraic set $\{\Delta_m = 0\}$.

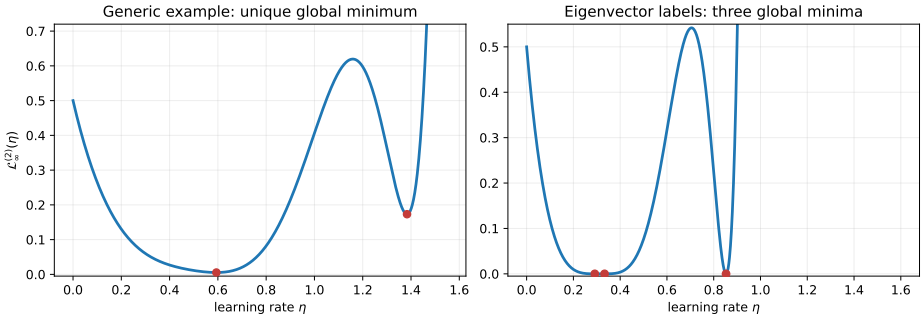

Figure 1: **Unique and exceptional landscapes.** Left: $K = \mathrm{diag}(1,2)$ and $y = (1,1)$ give the generic witness in Theorem 5.2, with a unique global minimizer. Right: $K = \left[\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right]$ and $y = (1,1)$ give the full-rank example in Corollary 6.2, with exactly three zero-loss global minimizers.

7 Perturbative argmin-set stability

For $\vartheta = (K, y) \in \mathcal{P}_m$, write

$$F(\vartheta, \eta) = \mathcal{L}_\infty^{(2)}(\eta; K, y), \qquad \mathcal{A}(\vartheta) = \arg\min_{\eta \geq 0} F(\vartheta, \eta). \tag{29}$$

[2]Here $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^m$ denotes the all-ones vector.

Theorem 7.1 (Global argmin-set stability under data perturbations)**.** *Fix $m \geq 2$. If $\vartheta_j = (K_j, y_j) \to \vartheta = (K, y)$ in $\mathcal{P}_m$, then*

$$\sup_{\eta \in \mathcal{A}(\vartheta_j)} \operatorname{dist}\big(\eta, \mathcal{A}(\vartheta)\big) \longrightarrow 0. \tag{30}$$

Equivalently, every choice $\eta_j \in \mathcal{A}(\vartheta_j)$ satisfies $\operatorname{dist}(\eta_j, \mathcal{A}(\vartheta)) \to 0$.

Proof sketch. Coefficient continuity gives compact-uniform convergence of the perturbed losses on every bounded interval. Continuity and positivity of the leading coefficient provide one common coercive tail bound, so all nearby global minimizers lie in a fixed compact interval. Every accumulation point of such minimizers then minimizes the limiting loss, and a contradiction argument yields (30). See Appendix A.9. □

The direction of (30) is important. If the limiting problem has three global minimizers and each perturbed problem has a unique one, the theorem says that the perturbed optimizer must approach one of the three original branches. It need not approach all three, and the theorem does not guarantee that every original branch is realized by some perturbation path.

Corollary 7.2 (Unique perturbed optimizers)**.** *If each nearby problem has a unique minimizer $\mathcal{A}(\vartheta_j) = \{\eta_j\}$, then*

$$\operatorname{dist}(\eta_j, \mathcal{A}(\vartheta)) \to 0.$$

If in addition $\mathcal{A}(\vartheta) = \{\eta_\star\}$, then $\eta_j \to \eta_\star$.

Proof. The first conclusion is Theorem 7.1; distance to the singleton $\{\eta_\star\}$ is $|\eta_j - \eta_\star|$. □

Combining Corollary 5.3 with Theorem 7.1, every $\vartheta \in \mathcal{P}_m$ admits a sequence of arbitrarily small generic perturbations with unique optimizers whose distance to $\mathcal{A}(\vartheta)$ tends to zero. This is only the upper, or one-sided, argmin inclusion. It does not assert that every point of $\mathcal{A}(\vartheta)$ is realized by some perturbation path, nor that a selected optimizer must oscillate among the limiting branches.

Figure 2 gives an exact numerical illustration based on (15). For the structured example of Corollary 6.2, the unperturbed three-way tie is lifted by each displayed nearby perturbation, and the unique global optimizer remains close to the original argmin set. In these particular perturbations the middle branch is selected; the figure is not a claim that arbitrarily small perturbations realize all three limiting branches.

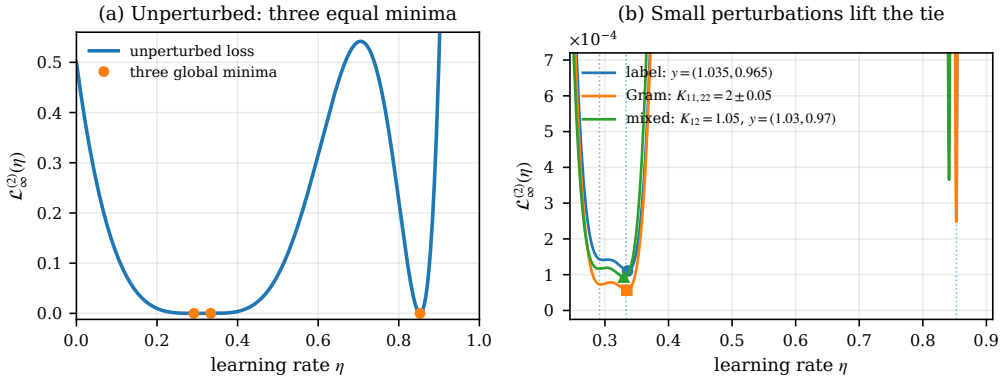

Figure 2: **Perturbative resolution of the structured nonunique landscape.** (a) The unperturbed full-rank example $K = \left[\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right]$, $y = (1, 1)$ has three zero-loss global minimizers. (b) Exact evaluations of (15) for the three nearby perturbations specified in the legend. Dotted lines mark the original argmin set and filled markers mark the unique global minimizers of the perturbed losses. All displayed optimizers lie near the middle branch, illustrating the one-sided conclusion of Theorem 7.1 without asserting reverse branch realization.

8 Consequences for learning-rate transfer

The finite-width argument follows a four-step chain. The exact recursion gives a width-independent polynomial degree bound; Tensor Programs IV at finitely many learning rates, followed by Vandermonde inversion, gives coefficient convergence on one probability-one event. Fixed degree upgrades this to compact-uniform convergence of every fixed derivative. Uniqueness then yields scalar compact transfer, while nondegeneracy yields persistence of separate local-minimum branches.

8.1 Finite-width coefficient convergence

Proposition 8.1 (Model-specific compact-uniform convergence)**.** *For the $L = 2, t = 2$ model, $\mathcal{L}_n^{(2)}(\eta)$ is a polynomial in η of degree at most 16. On one probability-one event, its coefficients converge to those of $\mathcal{L}_\infty^{(2)}$, padded by zero coefficients in degrees $11, \dots, 16$. Consequently, for every compact interval $I \subset [0, \infty)$ and every fixed derivative order r,*

$$\sup_{\eta \in I} \left| \frac{d^r}{d\eta^r} \mathcal{L}_n^{(2)}(\eta) - \frac{d^r}{d\eta^r} \mathcal{L}_\infty^{(2)}(\eta) \right| \longrightarrow 0 \qquad \textit{almost surely.} \tag{31}$$

Proof sketch. The compressed recursion bounds the finite-width loss degree by 16, uniformly in n. Tensor Programs IV gives simultaneous almost-sure convergence at 17 distinct learning rates, and inversion of the resulting deterministic Vandermonde system recovers all coefficients on the same event. Since the degree is fixed, coefficient convergence implies compact-uniform convergence of every fixed derivative. See Appendices A.10 and D. □

8.2 Unique regime

Theorem 8.2 (Compact learning-rate transfer)**.** *Assume $K \succeq 0$, $Ky \neq 0$, and that $\mathcal{L}_\infty^{(2)}$ has a unique global minimizer η_∞. Let $I \subset [0, \infty)$ be a deterministic compact interval containing η_∞. Then every selection*

$$\eta_n \in \arg\min_{\eta \in I} \mathcal{L}_n^{(2)}(\eta)$$

converges almost surely to η_∞.

Proof sketch. Uniform convergence on I and uniqueness of the limiting minimizer give compact argmin consistency. See Appendix A.11. □

In particular, Theorem 5.2 verifies the uniqueness hypothesis for every $(K, y) \in \mathcal{U}_m$. This is deliberately a compact-interval statement; it does not claim that unrestricted finite-width global minimizers remain bounded.

8.3 Nonunique regime and branchwise transfer

Lemma 8.3 (Persistence of one isolated nondegenerate minimum)**.** *Let $\eta_* > 0$ be an isolated local minimum of $\mathcal{L}_\infty^{(2)}$ satisfying*

$$(\mathcal{L}_\infty^{(2)})'(\eta_*) = 0, \qquad (\mathcal{L}_\infty^{(2)})''(\eta_*) > 0.$$

Then there exists a deterministic compact interval $I \subset (0, \infty)$ with $\eta_* \in \operatorname{int} I$ such that, almost surely for all sufficiently large n, $\mathcal{L}_n^{(2)}$ has a unique critical point $\eta_n \in \operatorname{int} I$. This point is a strict local minimizer, and $\eta_n \to \eta_*$.

Proof sketch. Choose I so that the limiting second derivative is strictly positive throughout I and the limiting first derivative has opposite signs at its endpoints. Compact-uniform C^2 convergence preserves strict convexity and the endpoint signs, giving a unique finite-width critical point and hence a strict local minimum. Shrinking the neighborhood, or equivalently taking accumulation points, gives $\eta_n \to \eta_*$. See Appendix A.12. □

Theorem 8.4 (Branchwise persistence of isolated nondegenerate global minimizers)**.** *Assume that, for some finite $k \geq 1$,*

$$\arg\min_{\eta \geq 0} \mathcal{L}_\infty^{(2)}(\eta) = \{\eta_{\infty,1}, \ldots, \eta_{\infty,k}\}, \qquad 0 < \eta_{\infty,1} < \cdots < \eta_{\infty,k},$$

and that every global minimizer is nondegenerate:

$$(\mathcal{L}_\infty^{(2)})''(\eta_{\infty,j}) > 0, \qquad j = 1, \ldots, k.$$

Then there exist pairwise disjoint deterministic compact intervals $I_1, \ldots, I_k$ such that, almost surely for all sufficiently large n, each I_j contains a unique finite-width strict local minimizer $\eta_{n,j} \in \operatorname{int} I_j$, and

$$\eta_{n,j} \longrightarrow \eta_{\infty,j}, \qquad j = 1, \ldots, k.$$

These branches are not asserted to be finite-width global minimizers. In particular, their finite-width loss values need not be equal.

Proof sketch. Choose pairwise disjoint neighborhoods of the finitely many limiting minimizers and apply Lemma 8.3 in each one. See Appendix A.13. □

Applying Theorem 8.4 to the structured family of Theorem 6.1 yields the three persistent finite-width branches illustrated in Figure 3.

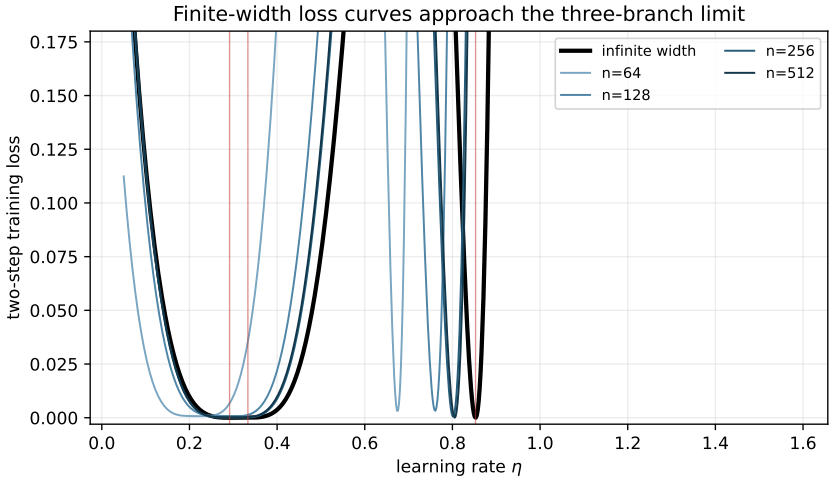

Figure 3: **Finite-width branch persistence.** For the structured example in Figure 1, the finite-width losses approach the limiting three-minimum landscape on the plotted compact interval; vertical lines mark the limiting branches. The branches need not be finite-width global minima (Theorem 8.4).

Thus the unique regime supports convergence to one deterministic learning rate on every fixed compact interval containing the limiting optimizer. At exceptional nonunique data, the proved finite-width object is a collection of local-minimum branches, while Theorem 7.1 describes how global optimizers behave under perturbations of the limiting data.

9 Discussion and limitations

We have analyzed the optimizer-identification problem for the first nonlinear but exactly solvable learning-rate landscape under μP. In the $L = 2$, $t = 2$ deep-linear model, the limiting loss is an explicit coercive degree-ten polynomial. Its global minimizer is generically unique on the positive-definite parameter space, while a structured full-rank family has exactly three nondegenerate global minimizers. The perturbation and finite-width results then separate ordinary scalar transfer from the set-valued or branchwise behavior required at exceptional data.

Limitations. The analysis is deliberately restricted to two trained matrices, two full-batch gradient updates, squared loss, simultaneous gradient descent, frozen input and output layers, and the Gaussian μP initialization in (2). These assumptions make the complete global geometry accessible, but the present discriminant and spectral arguments do not directly extend to general depth L or training time t. Although polynomial dependence on the learning rate persists for every fixed L and t, the degree and coefficient geometry grow rapidly. The exact limiting formula itself allows $K \succeq 0$, but the generic-uniqueness and perturbation results are formulated on $K \succ 0$. Thus they do not yet provide a rank-constrained genericity theory for singular Gram matrices or for the common regime $d < m$. The stability conclusions are also qualitative: we prove only the one-sided argmin inclusion in Theorem 7.1, not reverse or Hausdorff convergence, and we do not show that every limiting minimizer can be selected by a vanishing perturbation. The branchwise result is stated for isolated nondegenerate limiting minima. Degenerate limiting minima may have a more delicate finite-width local structure, since compact-uniform C^2 convergence alone does not determine a unique nearby branch; such further exceptional local configurations are outside the scope of the present branchwise analysis. This does not affect the general perturbative restoration of uniqueness and stability results of Sections 5–7, or the compact transfer result of Theorem 8.2 once the limiting minimizer is unique. Finally, the finite-width transfer theorem is restricted to a prescribed compact interval and gives no quantitative rate in width.

Future directions. Natural next steps are to develop genericity on fixed-rank positive-semidefinite strata, to determine whether the generic-unique/exceptional-nonunique dichotomy persists for larger L and t, and to obtain quantitative coefficient and argmin rates. A finer perturbation theory could characterize which limiting branches are realizable and establish two-sided set convergence under suitable nondegeneracy conditions. Extending the analysis beyond deep-linear networks, squared loss, and gradient descent will likely require different proof machinery, since the exact polynomial reduction used here is then no longer available in its present form.

References

[1] S. Arora, N. Cohen, N. Golowich, and W. Hu. A convergence analysis of gradient descent for deep linear neural networks. *International Conference on Learning Representations*, 2019.

[2] B. Bordelon, L. Noci, M. B. Li, B. Hanin, and C. Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. *International Conference on Learning Representations*, 2024.

[3] B. Bordelon and C. Pehlevan. Deep linear network training dynamics from random initialization: Data, width, depth, and hyperparameter transfer. *arXiv preprint arXiv:2502.02531*, 2025.

[4] L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] L. Chizat and P. Netrapalli. The feature speed formula: A flexible approach to scale hyper-parameters of deep neural networks. *Advances in Neural Information Processing Systems*, 37:62362–62383, 2024.

[6] L. Chizat, M. Colombo, X. Fernández-Real, and A. Figalli. Infinite-width limit of deep linear neural networks. *Communications on Pure and Applied Mathematics*, 77(10):3958–4007, 2024.

[7] K. Everett, L. Xiao, M. Wortsman, A. A. Alemi, R. Novak, P. J. Liu, I. Gur, J. Sohl-Dickstein, L. P. Kaelbling, J. Lee, and J. Pennington. Scaling exponents across parameterizations and optimizers. *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, 2024.

[8] S. Hayou. A proof of learning rate transfer under μP. *arXiv preprint arXiv:2511.01734*, 2025.

[9] W. Huang, W. Chen, Z. Xu, Z. Wang, and T. Suzuki. Exact and rich feature learning dynamics of two-layer linear networks. *Conference on Parsimony and Learning*, PMLR 280:1087–1111, 2025.

[10] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[11] Z. Ji and M. Telgarsky. Gradient descent aligns the layers of deep linear networks. *International Conference on Learning Representations*, 2019.

[12] A. Kosson, J. Welborn, Y. Liu, M. Jaggi, and X. Chen. Weight decay may matter more than μP for learning rate transfer in practice. *arXiv preprint arXiv:2510.19093*, 2025.

[13] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, 32:8572–8583, 2019.

[14] L. Lingle. An empirical study of μP learning rate transfer. *arXiv preprint arXiv:2404.05728*, 2024.

[15] L. Noci, A. Meterez, T. Hofmann, and A. Orvieto. Super consistency of neural network landscapes and learning rate transfer. *Advances in Neural Information Processing Systems*, 37, 2024.

[16] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[17] G. Yang and E. J. Hu. Tensor Programs IV: Feature learning in infinite-width neural networks. *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:11727–11737, 2021.

[18] G. Yang, E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao. Tensor Programs V: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

[19] G. Yang, D. Yu, C. Zhu, and S. Hayou. Tensor Programs VI: Feature learning in infinite-depth neural networks. *arXiv preprint arXiv:2310.02244*, 2023.

[20] Y. Zhang, P. E. Latham, L. C. Vankadara, and A. Saxe. Optimal learning rate scaling depends on data in deep scalar linear networks. *arXiv preprint arXiv:2607.07884*, 2026.

A Proofs of the main results

A.1 Proof of Proposition 3.1

In the scalar wide limit, let p and q denote the independent initial row variables with $\mathbb{E}[p^2] = \kappa$ and $\mathbb{E}[q^2] = 1$. The initial output vanishes, so the first update gives

$$p_1 = p + \eta\kappa y q, \qquad q_1 = q + \eta y p, \qquad r_\infty^{(1)} = y(2a - 1).$$

Consequently,

$$\mathbb{E}[p_1 q_1] = 2ay, \qquad \mathbb{E}[p_1^2] = \kappa(1 + \theta a^2), \qquad \mathbb{E}[q_1^2] = 1 + \theta a^2.$$

Expanding the second simultaneous update and subtracting y gives

$$\begin{aligned}\frac{r_\infty^{(2)}(\eta)}{y} &= 2a - 1 - 2a(2a-1)(1+\theta a^2) + 2\theta a^3(2a-1)^2 \\ &= (2a-1)\big[-(2a-1) + 4\theta a^3(a-1)\big].\end{aligned}$$

The factor $2a - 1$ gives the root $a = 1/2$. Since $Q_\theta(0) = 1$ and $Q_\theta(1/2) = -\theta/4 < 0$, there is a root in $(0, 1/2)$. Since $Q_\theta(1) = -1$ and $Q_\theta(a) \to +\infty$ as $a \to \infty$, there is another root in $(1, \infty)$. At all three roots the residual vanishes, so the nonnegative limiting loss is globally minimized.

A.2 Proof of Proposition 4.1

At step s,

$$\nabla_{W_1}\mathcal{L}_n = \frac{1}{m}\sum_i r_{n,i}^{(s)}\big((W_2^{(s)})^\top V\big)(W_0 x_i)^\top, \qquad \nabla_{W_2}\mathcal{L}_n = \frac{1}{m}\sum_i r_{n,i}^{(s)} V(A_s[:,i])^\top.$$

Since $(W_2^{(s)})^\top V = q_s/n$, multiplying the first matrix update by U gives (9). Transposing the second update and multiplying by $\widetilde{v}$ gives (10), because $V^\top\widetilde{v} = n\|V\|^2 = \nu_n$. Finally, $V^\top W_2^{(s)} A_s = n^{-1} q_s^\top A_s$, proving (8).

A.3 Proof of Theorem 4.2

At initialization, the row law of A_0 converges to a centered Gaussian vector $p \in \mathbb{R}^m$ with covariance K, and the coordinate law of q_0 converges to an independent standard Gaussian q. In particular,

$$K_n \to K, \quad \nu_n \to 1, \quad \frac{1}{n}A_0^\top A_0 \to K, \quad \frac{1}{n}\|q_0\|^2 \to 1, \quad \frac{1}{n}A_0^\top q_0 \to 0 \tag{32}$$

almost surely. Write the limiting row variables after the first update as

$$a_1 = p + zqKy, \qquad q_1 = q + zp^\top y.$$

Using independence and $\mathbb{E}[pp^\top] = K$ gives

$$\mathbb{E}[a_1 q_1] = 2zKy, \qquad C_1 := \mathbb{E}[a_1 a_1^\top] = K + z^2(Ky)(Ky)^\top, \qquad S_1 := \mathbb{E}[q_1^2] = 1 + z^2 y^\top K y. \tag{33}$$

With $r = 2zKy - y$, the second update is $a_2 = a_1 - zq_1 Kr$ and $q_2 = q_1 - za_1^\top r$. Therefore

$$\mathbb{E}[a_2 q_2] = 2zKy - zC_1 r - zS_1 Kr + z^2(r^\top 2zKy)Kr. \tag{34}$$

Substitution of (33) and $r = 2zKy - y$, followed by the expansion in Appendix B, yields $f_\infty^{(2)} = A(z)Ky + B(z)K^2 y$. Subtracting y and squaring give (13) and (15).

A.4 Proofs of Proposition 4.3 and Corollary 4.4

For $K \succeq 0$,

$$Ky \neq 0 \quad \Longleftrightarrow \quad y^\top Ky > 0, \qquad \ker K^2 = \ker K.$$

Thus $Ky \neq 0$ implies both $\|Ky\|^2 > 0$ and $K^2 y \neq 0$. Equation (13) gives (16); squaring its norm and using $z = \eta/m$ gives (17). The constant and linear residual coefficients are $-y$ and $4Ky$, so differentiating $\|r_\infty^{(2)}\|^2/(2m)$ gives (18). If $Ky = 0$, then $K^2 y = 0$, $y^\top Ky = 0$, and $\|Ky\|^2 = 0$, so the limiting output vanishes identically.

Under $Ky \neq 0$, the positive leading coefficient gives coercivity and the negative derivative at zero excludes the boundary point 0. Hence a positive global minimizer exists. If the minimum value is c, every distinct minimizer is a root of even multiplicity at least two of the degree-ten polynomial $\mathcal{L}_\infty^{(2)} - c$, so there are at most five. The case $Ky = 0$ is immediate from the constant-loss formula.

A.5 Proof of Proposition 4.5

On the range of P_a, the operators K and K^2 act by λ_a and λ_a^2. Applying P_a to (13) gives

$$P_a r_\infty^{(2)}(z) = p_{\lambda_a}(z) P_a y.$$

Summing over the orthogonal eigenspaces gives the residual formula, and Pythagoras gives the loss formula in (20).

A.6 Proof of Theorem 5.1

Because $P_{K,y}$ has exact degree ten, the resultant is, up to a nonzero factor,

$$C_{K,y}(u) = \prod_{P'_{K,y}(\zeta)=0} \big(u - P_{K,y}(\zeta)\big),$$

where complex critical points are counted with multiplicity. Hence $\Delta_m(K, y) \neq 0$ implies that the critical points are simple and their critical values are pairwise distinct. By Corollary 4.4, every global minimizer is positive and interior, hence critical. Two distinct global minimizers would therefore have the same critical value, contradicting $\Delta_m(K, y) \neq 0$. Conversely, two distinct global minimizers have the same value and hence produce a repeated root of $C_{K,y}$, so $\Delta_m(K, y) = 0$.

A.7 Proof of Theorem 5.2

For $m = 2$, choose $K = \mathrm{diag}(1, 2)$ and $y = (1, 1)^\top$. Exact expansion gives

$$\begin{aligned} P_{K,y}(z) = \frac{1}{2}\big(&13600z^{10} - 23040z^9 + 14184z^8 - 6464z^7 + 4104z^6 \\ &- 1880z^5 + 556z^4 - 180z^3 + 60z^2 - 12z + 1\big). \end{aligned} \tag{35}$$

Exact resultant arithmetic modulo $p = 1{,}000{,}003$ gives

$$\mathrm{disc}_u\, C_{K,y}(u) \equiv 96809 \pmod{p}, \tag{36}$$

which is nonzero. Appendix C records the complete coefficient certificate, so Δ_2 is not the zero polynomial.

For $m > 2$, take K block diagonal with the leading block $\mathrm{diag}(1, 2)$ and arbitrary positive diagonal entries in the remaining coordinates, and take $y = (1, 1, 0, \ldots, 0)^\top$. The quantities Ky, $K^2 y$, $y^\top Ky$, and $\|Ky\|^2$ are unchanged, and $P_{K,y}$ is the $m = 2$ witness polynomial multiplied by the positive constant $2/m$. Scaling all critical values by a nonzero constant preserves their distinctness, so Δ_m is nonzero at this embedded point. Hence Δ_m is a nonzero polynomial for every fixed $m \geq 2$.

The complement of the zero set of a nonzero real polynomial is Zariski open, Euclidean open dense, and full measure. Intersecting with the open set $\mathcal{P}_m$ preserves these properties. Theorem 5.1 gives uniqueness throughout $\mathcal{U}_m$.

A.8 Proof of Theorem 6.1

On $\ker K$ the residual is always $-y_0$. On the λ-eigenspace, Proposition 4.5, together with

$$y^\top K y = \lambda \|y_\lambda\|^2, \qquad \|Ky\|^2 = \lambda^2 \|y_\lambda\|^2,$$

gives

$$r_\infty^{(2)} = -y_0 + p_\theta(a) y_\lambda. \tag{37}$$

Hence

$$\mathcal{L}_\infty^{(2)}(\eta) = \frac{1}{2m}\left(\|y_0\|^2 + \|y_\lambda\|^2 p_\theta(a)^2\right), \tag{38}$$

so its global minimizers are exactly the positive zeros of p_θ.

The linear factor gives $a = 1/2$. Also $Q_\theta(0) = 1$ and $Q_\theta(1/2) = -\theta/4 < 0$. Since

$$Q_\theta'(a) = -2 + 4\theta a^2(4a - 3) < 0 \qquad (0 < a < 1/2),$$

there is exactly one simple root in $(0, 1/2)$. On $(1/2, 1)$,

$$Q_\theta''(a) = 24\theta a(2a - 1) > 0,$$

and both endpoint values are negative; strict convexity places the graph below its negative endpoint chord, so there is no root. On $(1, \infty)$ the derivative is strictly increasing. If $Q_\theta'(1) \geq 0$, the function is increasing there; if $Q_\theta'(1) < 0$, it decreases once to its unique minimum and then increases. In either case, $Q_\theta(1) = -1$ and $Q_\theta(a) \to +\infty$ give exactly one crossing, with positive derivative. Finally $Q_\theta(1/2) \neq 0$, so the middle root is simple. At every simple zero $\eta_\star$ of the residual factor,

$$\left(\mathcal{L}_\infty^{(2)}\right)''(\eta_\star) = \frac{\|y_\lambda\|^2}{m}\left(\frac{\lambda}{m} p_\theta'(a_\star)\right)^2 > 0,$$

which proves nondegeneracy.

A.9 Proof of Theorem 7.1

Lemma A.1 (Coefficient and compact-uniform continuity)**.** *For fixed m, write*

$$F(\vartheta, \eta) = \sum_{k=0}^{10} c_k(\vartheta)\eta^k.$$

Every c_k is a polynomial in the entries of K and y divided only by a fixed power of m. Hence, if $\vartheta_j \to \vartheta$ in the norm (25), then for every $R < \infty$,

$$\sup_{0 \leq \eta \leq R} |F(\vartheta_j, \eta) - F(\vartheta, \eta)| \longrightarrow 0. \tag{39}$$

Proof. Equations (14) and (15) use only matrix multiplication, inner products, addition, and multiplication, followed by fixed powers of m^{-1}. Expansion therefore gives the claimed coefficient form. Coefficient continuity yields

$$\sup_{0 \leq \eta \leq R} |F(\vartheta_j, \eta) - F(\vartheta, \eta)| \leq \sum_{k=0}^{10} |c_k(\vartheta_j) - c_k(\vartheta)| R^k \to 0.$$

□

Proof of Theorem 7.1. We verify the required global compactness rather than assuming it from pointwise continuity.

Local uniform coercivity. By (17), the leading coefficient is

$$c_{10}(K, y) = \frac{32\|Ky\|^4 \|K^2 y\|^2}{m^{11}} > 0 \qquad \text{on } \mathcal{P}_m. \tag{40}$$

Choose a closed parameter ball N centered at ϑ, of sufficiently small radius that $N \subset \mathcal{P}_m$. It is compact. By coefficient continuity there are finite constants

$$a_0 = \min_{\vartheta' \in N} c_{10}(\vartheta') > 0, \qquad M = \max_{\vartheta' \in N} \sum_{k=0}^{9} |c_k(\vartheta')| < \infty, \qquad B = \max_{\vartheta' \in N} F(\vartheta', 0) < \infty.$$

For every $\vartheta' \in N$ and $\eta \geq 1$,

$$F(\vartheta', \eta) \geq a_0 \eta^{10} - M\eta^9.$$

Choose one $R > 1$ such that

$$R \geq \frac{2M}{a_0}, \qquad \frac{a_0}{2} R^{10} > B.$$

Then for all $\eta \geq R$ and all $\vartheta' \in N$,

$$F(\vartheta', \eta) \geq \frac{a_0}{2} \eta^{10} > B \geq F(\vartheta', 0).$$

Thus every global minimizer for every $\vartheta' \in N$ lies in the common compact interval $[0, R]$. In particular, all $\mathcal{A}(\vartheta')$ are nonempty compact sets.

Limit points of minimizers. For all sufficiently large j, $\vartheta_j \in N$. Take any $\eta_j \in \mathcal{A}(\vartheta_j)$. The common interval gives a subsequence $\eta_{j_k} \to \bar{\eta} \in [0, R]$. For every fixed $\eta \geq 0$,

$$F(\vartheta_{j_k}, \eta_{j_k}) \leq F(\vartheta_{j_k}, \eta).$$

Lemma A.1 gives uniform convergence on $[0, R]$, so the left-hand side converges to $F(\vartheta, \bar{\eta})$. Applying the same lemma on $[0, \max\{R, \eta\}]$ makes the right-hand side converge to $F(\vartheta, \eta)$. Hence

$$F(\vartheta, \bar{\eta}) \leq F(\vartheta, \eta) \qquad \text{for every } \eta \geq 0,$$

and therefore $\bar{\eta} \in \mathcal{A}(\vartheta)$.

Set-distance conclusion. If (30) failed, then for some $\varepsilon > 0$ and a subsequence one could choose $\eta_j \in \mathcal{A}(\vartheta_j)$ with $\operatorname{dist}(\eta_j, \mathcal{A}(\vartheta)) \geq \varepsilon$. The common interval yields a further convergent subsequence. Its limit belongs to $\mathcal{A}(\vartheta)$ by the previous paragraph, contradicting continuity of the distance to the closed set $\mathcal{A}(\vartheta)$. This proves the one-sided set convergence. □

A.10 Proof of Proposition 8.1

Starting from A_0, q_0 of degree zero in η, the exact recursion (8)–(10) gives

$$\deg(A_1), \deg(q_1) \leq 1, \qquad \deg r_n^{(1)} \leq 2, \qquad \deg(A_2), \deg(q_2) \leq 4.$$

Thus $\deg r_n^{(2)} \leq 8$ and $\deg \mathcal{L}_n^{(2)} \leq 16$, uniformly in n. Expanding the recursion coefficient by coefficient shows that every loss coefficient is a finite sum of products of K_n, ν_n and normalized empirical inner products of polynomial coordinate functions of A_0 and q_0. Only finitely many empirical moments of fixed degree occur.

The Gaussian initialization, together with the finite program (8)–(10), satisfies the hypotheses of Tensor Programs IV [17, Theorem 7.4]; Appendix D records the model-specific verification. Hence all required empirical moments converge almost surely. Equivalently, pointwise convergence at 17 fixed distinct learning rates and Vandermonde inversion identifies all coefficients on one probability-one event. The limit is the degree-ten polynomial (15), so the six higher coefficients vanish. For fixed degree, coefficient convergence implies (31) by termwise differentiation and a finite geometric bound on I.

A.11 Proof of Theorem 8.2

Work on the probability-one event of Proposition 8.1. If a subsequence failed to converge to η_∞, compactness of I would give a further subsequence converging to some $\bar{\eta} \in I$, $\bar{\eta} \neq \eta_\infty$. For every $\eta \in I$,

$$\mathcal{L}_n^{(2)}(\eta_n) \leq \mathcal{L}_n^{(2)}(\eta).$$

Uniform convergence on I and continuity of the limiting polynomial imply $\mathcal{L}_\infty^{(2)}(\bar{\eta}) \leq \mathcal{L}_\infty^{(2)}(\eta)$ for every $\eta \in I$. Since $\eta_\infty \in I$ is the unique global minimizer, this forces $\bar{\eta} = \eta_\infty$, a contradiction.

A.12 Proof of Lemma 8.3

By continuity of the limiting second derivative, choose a compact interval $I = [a, b] \subset (0, \infty)$ with $a < \eta_* < b$ and a constant $c > 0$ such that $(\mathcal{L}_\infty^{(2)})''(\eta) \geq c$ throughout I. Since $(\mathcal{L}_\infty^{(2)})'(\eta_*) = 0$, strict positivity of the second derivative gives $(\mathcal{L}_\infty^{(2)})'(a) < 0 < (\mathcal{L}_\infty^{(2)})'(b)$.

Work on the probability-one event of Proposition 8.1. Its compact-uniform convergence for the first two derivatives implies that, for all sufficiently large n, $(\mathcal{L}_n^{(2)})'' \geq c/2$ throughout I and the endpoint derivative signs persist. The intermediate value theorem therefore gives a critical point $\eta_n \in (a, b)$. The derivative $(\mathcal{L}_n^{(2)})'$ is strictly increasing on I, so this critical point is unique, and the positive second derivative makes it a strict local minimizer.

It remains to prove convergence. Any subsequence of (η_n) has, by compactness, a further subsequence converging to some $\bar{\eta} \in I$. Uniform convergence of the first derivatives and $(\mathcal{L}_n^{(2)})'(\eta_n) = 0$ give $(\mathcal{L}_\infty^{(2)})'(\bar{\eta}) = 0$. But the limiting first derivative is strictly increasing on I and vanishes at η_*, so $\bar{\eta} = \eta_*$. Every accumulation point is therefore η_*, which proves $\eta_n \to \eta_*$.

A.13 Proof of Theorem 8.4

The construction in the proof of Lemma 8.3 can be carried out inside any prescribed neighborhood of the relevant limiting minimizer. Choose pairwise disjoint neighborhoods of $\eta_{\infty,1}, \ldots, \eta_{\infty,k}$ and apply the lemma in each one. Because k is finite, the probability-one event and the sufficiently-large-n threshold may be chosen simultaneously for all $j = 1, \ldots, k$. This gives pairwise disjoint intervals I_j, unique strict local minimizers $\eta_{n,j}$ within them, and $\eta_{n,j} \to \eta_{\infty,j}$ for every j.

B Expanded algebra for the second update

Starting from (34) and $r = 2zKy - y$,

$$\begin{aligned}
f_\infty^{(2)} &= 2zKy - z\big(K + z^2(Ky)(Ky)^\top\big)r - z\big(1 + z^2 y^\top Ky\big)Kr + z^2(r^\top 2zKy)Kr \\
&= 2zKy - 2z(2zK^2y - Ky) \\
&\quad - z^3\Big\{(Ky)\big(2z\|Ky\|^2 - y^\top Ky\big) + (y^\top Ky)(2zK^2y - Ky)\Big\} \\
&\quad + z^2\big(4z^2\|Ky\|^2 - 2zy^\top Ky\big)(2zK^2y - Ky) \\
&= 4zKy - 4z^2K^2y + 4z^3(y^\top Ky)Ky - 6z^4\|Ky\|^2Ky \\
&\quad - 6z^4(y^\top Ky)K^2y + 8z^5\|Ky\|^2K^2y.
\end{aligned}$$

The symbolic verification script independently compares this coefficient expansion with (13), the scalar factorization in Proposition 3.1, the leading coefficient, the derivative at zero, and the spectral formula.

C Exact nonvanishing certificate

For (35), let $C(u) = \mathrm{Res}_z(P'(z), u - P(z))$. Modulo $p = 1{,}000{,}003$, the coefficients of C from highest to lowest degree are

$$(577638, 381302, 745024, 781663, 753402, 293667, 765438, 295834, 210088, 871189).$$

The Euclidean algorithm over $\mathbb{F}_p[u]$ gives $\gcd(C, C') = 1$, equivalently

$$\mathrm{disc}(C) \equiv 96809 \not\equiv 0 \pmod{p}.$$

The calculation starts from rational coefficients; clearing their denominators only multiplies the discriminant by a nonzero factor. The displayed resultant coefficients and the nonzero discriminant residue provide a self-contained exact certificate that Δ_2 is not identically zero; no numerical approximation is used.

D Finite-width coefficient convergence

The program needed for Proposition 8.1 is finite because m, L, t are fixed. The coordinate slices of

$$(W_0 x_1, \dots, W_0 x_m, \widetilde{v})$$

are centered Gaussian with limiting covariance $\mathrm{diag}(K, 1)$. They are independent of $W_1^{(0)}$ and $W_2^{(0)}$, whose entries are independent $N(0, 1/n)$ variables. Initial matrix multiplications are Tensor Program MatMul operations; the normalized inner products in (7) and (8) are Moment operations; and the two updates (9)–(10) are polynomial coordinate operations. The nonlinearities have fixed polynomial degree and satisfy the polynomial-control hypothesis used in Tensor Programs IV.

The Master Theorem therefore gives almost-sure limits for every normalized empirical moment appearing in the two-step recursion at each fixed learning rate [17, Theorem 7.4]. Because the finite-width loss has degree at most 16, choose 17 distinct deterministic learning rates $\eta_0, \dots, \eta_{16}$. Intersecting their probability-one events still has probability one. If b_n is the vector of the 17 loss coefficients and v_n the vector of the 17 evaluated losses, then

$$v_n = V b_n, \qquad V_{jk} = \eta_j^k.$$

The Vandermonde matrix V is deterministic and invertible, so $b_n = V^{-1} v_n$ converges almost surely. Theorem 4.2 identifies the limit with (15). This supplies the common event and derivative-level compact uniformity used in Section 8, without citing an unpublished general-step result.