A Mirror-Descent View of Saddle-to-Saddle Dynamics in Deep Linear Networks

Divit Rawal
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
divit.rawal@berkeley.edu

Abstract

We study saddle-to-saddle dynamics in deep linear network training: gradient descent from small balanced initialization activates the singular values of the predictor sequentially through long plateaus rather than approaching the target uniformly, and selects one minimizer from a manifold of equal-loss solutions. We give a single derivation that ties together the implicit-bias and dynamical-systems readings of this staircase. The radial flow in the SVD frame is exact mirror descent with respect to an $\ell^{2/L}$-quasi-norm Bregman potential, and the dual coordinate is a barrier β_L that feedback-linearizes the dynamics into an integrator driven by a saturating nonlinearity. The Bregman divergence to the target serves as the natural Lyapunov function and dissipates at rate -2ℓ. The dead-zone linearization of the forced integrator then yields a closed-form escape time with a phase transition between $L = 2$ (logarithmic in the initialization scale ε) and $L \geq 3$ (polynomial $\varepsilon^{-(L-2)}$). The polynomial exponent matches the depth-dependent scaling derived for nonlinear smooth networks by Rawal and DeWeese [2026].

1 Introduction

Starting from a small balanced initialization, gradient descent on a deep linear network produces a characteristic pattern: the singular values of the predictor matrix do not approach their target values uniformly, but instead activate in sequence through long plateaus, a staircase pattern first analyzed by Saxe et al. [2014] and now understood as saddle-to-saddle dynamics [Jacot et al., 2022, Cao et al., 2023]. This phenomenon underlies the implicit bias of overparameterized networks toward low-rank solutions [Gunasekar et al., 2017, Arora et al., 2019, Woodworth et al., 2020] and has become a common and widely used testbed for theories of feature learning.

Two lines of analysis are used to describe this staircase pattern. The first is geometric: the predictor flow is a Riemannian gradient flow on the rank-stratified manifold $\bigcup_k \mathcal{R}_k$, and the staircase reflects sequential entry into successively higher-rank strata [Bah et al., 2020, Menon, 2024]. The second is dynamical: each plateau is a slowdown of an unstable mode whose escape time scales polynomially in the initialization scale, with depth-dependent exponent [Cao et al., 2023, Jacot et al., 2022, Rawal and DeWeese, 2026].

Here we give a single derivation that yields both views. Under task-aligned balanced initialization, the predictor-space flow $\dot{M} = -\mathcal{K}_{L,M} \nabla \ell(M)$ diagonalizes in the SVD frame into independent scalar channels, and we identify the resulting scalar evolution as exact mirror descent with respect to an explicit Bregman potential Φ_L proportional to the $\ell^{2/L}$-quasi-norm. The mirror map $\nabla \Phi_L = \beta_L$ is a barrier coordinate, and the Bregman divergence to the target is the natural Lyapunov function for the flow.

Preprint.

The barrier coordinate has a second reading. It effects a feedback linearization of the radial flow into an integrator driven by a saturating nonlinearity – the canonical setting for Lyapunov and integral-quadratic-constraint (IQC) analysis [Lessard et al., 2016, Zames and Falb, 1968, Boczar et al., 2015]. The dead-zone linearization of this forced integrator gives a closed-form escape-time formula, whose phase transition between $L = 2$ (logarithmic) and $L \geq 3$ (polynomial $\varepsilon^{-(L-2)}$) is the change in the antiderivative of the integrand defining β_L.

Contributions. Here we work out three connected results for the deep linear network gradient flow. **(i) Mirror-descent characterization.** The predictor flow in the SVD frame is exact mirror descent with mirror map $\beta_L = \nabla\Phi_L$ and Bregman potential Φ_L proportional to the $\ell^{2/L}$-quasi-norm (Theorem 2), and the Bregman divergence to the target dissipates at rate -2ℓ, serving as a Lyapunov function. **(ii) Saddle-escape phase transition.** The dead-zone linearization of the resulting forced integrator yields a closed-form escape time with a logarithmic-to-polynomial phase transition at $L = 3$ and explicit constants (Proposition 4); the polynomial exponent $\varepsilon^{-(L-2)}$ matches the law of Rawal and DeWeese [2026] for nonlinear networks. **(iii) Active-basin rate.** The closed-loop rate near equilibrium is exponential with depth-dependent constant $L(\sigma^\star)^{2-2/L}$, giving a two-phase total-time decomposition (Corollary 5).

Related work. *DLN geometry and dynamics.* The operator $\mathcal{K}_{L,M}$ and the Riemannian view of DLN training are due to Bah et al. [2020], Menon [2024], Menon and Yu [2026], with recent extensions in Lindsey and Menon [2026], Chen et al. [2025]. Saddle-to-saddle dynamics in DLNs are analyzed in Saxe et al. [2014], Jacot et al. [2022], Cao et al. [2023], Dominé et al. [2025], and the silent-alignment phenomenon in Atanasov et al. [2021], Tarmoun et al. [2021]. The depth-dependent escape-time scaling for nonlinear smooth networks is derived in Rawal and DeWeese [2026].

Implicit bias and mirror flow. The $\ell^{2/L}$-quasi-norm regularizer was identified as the implicit-bias minimizer in Gunasekar et al. [2017, 2019], Woodworth et al. [2020], Pesme et al. [2021], and the broader implicit-bias program is reviewed in Soudry et al. [2024], Lyu and Li [2020]. The present construction makes the corresponding mirror map and Bregman dissipation explicit, in the line of mirror-descent / Bregman dynamics started by Beck and Teboulle [2003] and continued in Krichene et al. [2015], Wibisono et al. [2016].

Control theory. The integrator-plus-saturating-nonlinearity form is the standard setting for Zames-Falb multipliers [Zames and Falb, 1968] and IQC-based stability and contraction-rate analysis of optimization algorithms [Lessard et al., 2016, Boczar et al., 2015]. We do not construct an IQC certificate here; instead, the Bregman divergence from Theorem 2 serves as a Lyapunov function, and the dead-zone bound in Section 4 follows from its dissipation rate.

2 Setup

Consider an L-layer linear network with weights $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, predictor $M = W_L W_{L-1} \cdots W_1 \in \mathbb{R}^{d_L \times d_0}$, and quadratic loss $\ell(M) = \frac{1}{2}\|M - M^\star\|_F^2$ on a fixed task-aligned target $M^\star = U\,\mathrm{diag}(\sigma_1^\star, \ldots, \sigma_r^\star)V^\top$ of rank $r \leq \min(d_0, d_L)$. We assume balanced initialization, $W_{\ell+1}^\top W_{\ell+1} = W_\ell W_\ell^\top$ at $t = 0$, which is preserved exactly by the layerwise gradient flow [Saxe et al., 2014, Bah et al., 2020].

Under balancedness, the layerwise flow projects to a closed evolution on the predictor:

$$\dot{M} = -\mathcal{K}_{L,M}\,(M - M^\star), \qquad \mathcal{K}_{L,M}\,A = \sum_{j=0}^{L-1} (MM^\top)^{j/L}\,A\,(M^\top M)^{(L-1-j)/L}, \tag{1}$$

the predictor-space closure of Bah et al. [2020] (see also Arora et al. [2019]). For a task-aligned target, the operator diagonalizes in the SVD frame, and (1) reduces to independent scalar channels:

$$\dot{\sigma}_i = -L\,\sigma_i^{2-2/L}\,(\sigma_i - \sigma_i^\star), \qquad i = 1, \ldots, \min(d_0, d_L). \tag{2}$$

The radial-angular split is preserved by the flow and the alignment with $M^\star$'s singular vectors is exact under task-aligned initialization [Saxe et al., 2014, Atanasov et al., 2021].

3 The mirror-flow characterization

3.1 The barrier coordinate

The dependence $\dot{\sigma}_i \propto \sigma_i^{2-2/L}$ in (2) makes the dead-zone behavior near $\sigma_i = 0$ clear: the drift vanishes at the rank-deficient boundary at a depth-dependent polynomial rate. We linearize this dependence with the barrier coordinate

$$b_i = \beta_L(\sigma_i) \doteq \frac{1}{L}\int_0^{\sigma_i} s^{-(2-2/L)}\, ds = \frac{\sigma_i^{2/L-1}}{L\,(2/L-1)}, \qquad L \geq 3, \tag{3}$$

with $\beta_L(\sigma) = \frac{1}{2}\log\sigma$ for $L = 2$. Note that the transformation is monotone on $(0, \infty)$.

Lemma 1 (Integrator form)**.** *Under* (2)*, the barrier coordinate evolves as* $\dot{b}_i = \sigma_i^\star - \sigma_i = \sigma_i^\star - \beta_L^{-1}(b_i)$.

The right-hand side is an integrator $\dot{b}_i = \phi_L(b_i; \sigma_i^\star)$ driven by a saturating nonlinearity $\phi_L(b; s^\star) = s^\star - \beta_L^{-1}(b)$. The nonlinearity has two limiting regimes. As $b_i \to -\infty$ (equivalently $\sigma_i \to 0^+$), $\phi_L \to \sigma_i^\star$: this is the dead zone, where the integrator drifts at constant rate. As $b_i \to b_i^\star \doteq \beta_L(\sigma_i^\star)$, ϕ_L vanishes exponentially in $b_i^\star - b_i$: this is the active basin, where the feedback saturates the integrator. The two regimes meet at the active-basin boundary, producing the straight-line-then-exponential geometry of Figure 1(b).

In control-theoretic terms, Lemma 1 is a feedback linearization of (2) into the canonical integrator-plus-saturating-nonlinearity form on which Lyapunov, IQC, and Zames-Falb analysis are built [Lessard et al., 2016, Zames and Falb, 1968]. The Lyapunov function for this system is the Bregman divergence, which we construct in Section 3.2.

3.2 Mirror-descent characterization

Define the Bregman potential

$$\Phi_L(\sigma) = \frac{\sigma^{2/L}}{L\,(2/L)\,(2/L-1)}, \qquad L \geq 3, \tag{4}$$

extended to active multi-mode profiles $\sigma = (\sigma_1, \ldots, \sigma_r)$ by aggregation, $\Phi_L(\sigma) = \sum_i \Phi_L(\sigma_i)$. (For $L = 2$, $\Phi_L(\sigma) = \sigma\log\sigma - \sigma$.) Then $\nabla\Phi_L(\sigma_i) = \beta_L(\sigma_i)$ and $\nabla^2\Phi_L(\sigma_i) = \sigma_i^{2/L-2}/L > 0$ on $\sigma_i > 0$, so Φ_L is strictly convex on the active orthant.

Theorem 2 (Mirror-descent characterization)**.** *Equation* (2) *is mirror descent on* ℓ *with mirror map* $\nabla\Phi_L = \beta_L$*:*

$$\dot{\sigma} = -\left[\nabla^2\Phi_L(\sigma)\right]^{-1}\nabla\ell(\sigma). \tag{5}$$

Equivalently, in the dual coordinate $b = \beta_L(\sigma)$*, the flow is gradient descent on* ℓ *with the Euclidean metric,* $\dot{b} = -\nabla_\sigma\ell(\sigma) = \sigma^\star - \sigma$*. The Bregman divergence to the target,* $D_{\Phi_L}(\sigma^\star, \sigma) = \Phi_L(\sigma^\star) - \Phi_L(\sigma) - \langle\beta_L(\sigma), \sigma^\star - \sigma\rangle$*, dissipates monotonically along trajectories at rate* $\frac{d}{dt}D_{\Phi_L}(\sigma^\star, \sigma(t)) = -\|\sigma^\star - \sigma(t)\|^2 = -2\,\ell(\sigma(t))$.

Proof sketch. For $L \geq 3$, $\nabla^2\Phi_L(\sigma_i) = \sigma_i^{2/L-2}/L$, so the right-hand side of (5) equals $-L\,\sigma^{2-2/L}\,(\sigma - \sigma^\star)$, which is (2). Differentiating D_{Φ_L} along the flow and substituting $\nabla^2\Phi_L\,\dot{\sigma} = -\nabla\ell$ gives $\frac{d}{dt}D_{\Phi_L} = -(\sigma - \sigma^\star)^\top(\sigma^\star - \sigma) = -\|\sigma - \sigma^\star\|^2$. The $L = 2$ case is identical with $\Phi_L(\sigma) = \sigma\log\sigma - \sigma$; details follow the standard mirror-descent / Bregman-dynamics treatment of Beck and Teboulle [2003], Wibisono et al. [2016]. □

Remark 3 (Bregman as Lyapunov). Theorem 2 reads as the construction of a depth-dependent Lyapunov function for the closed-loop system. The Bregman divergence $D_{\Phi_L}(\sigma^\star, \sigma)$ is positive-definite about the equilibrium $\sigma = \sigma^\star$, radially unbounded on the active orthant, and dissipates strictly along the flow at rate -2ℓ. Equivalently, the closed-loop forced integrator $\dot{b} = \sigma^\star - \beta_L^{-1}(b)$ is passive with storage function D_{Φ_L} and supply rate 2ℓ. The escape-time analysis in Section 4 is the integrated form of this dissipation in the dead-zone regime where the storage function's σ-coordinate diverges.

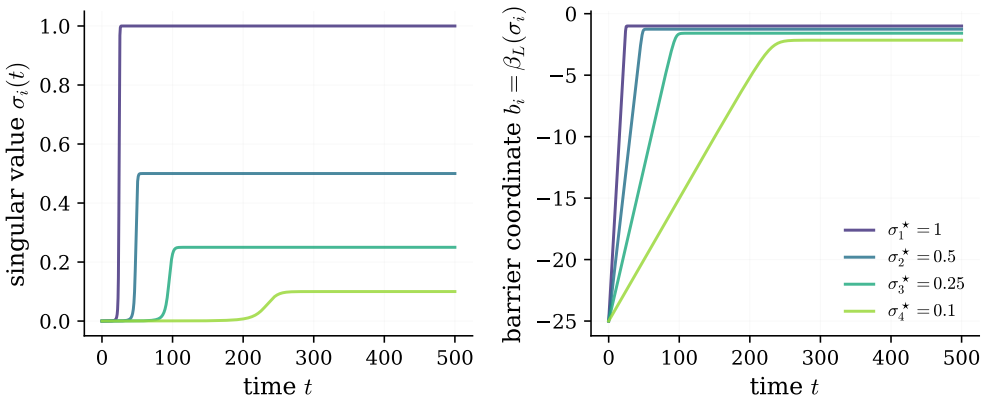

Figure 1: **Saddle-to-saddle dynamics in two coordinate systems.** $L = 3$, rank-4 task-aligned target $\sigma^\star = (1, 0.5, 0.25, 0.1)$, balanced initialization at $\varepsilon = 0.04$. *(a)* Singular-value trajectories $\sigma_i(t)$ – the classical staircase [Saxe et al., 2014]. *(b)* The same trajectories in barrier coordinates $b_i = \beta_L(\sigma_i)$ from (3). As per Theorem 2, in this coordinate the flow is plain gradient descent on ℓ, so each mode is a straight line through its dead zone with slope $\sigma_i^\star$ that bends into exponential approach once the mode enters its active basin.

3.3 Implicit bias as the $\ell^{2/L}$-quasi-norm minimizer

Aggregated across modes, $\Phi_L(\sigma) \propto \sum_i \sigma_i^{2/L}$ is the depth-L $\ell^{2/L}$-quasi-norm potential. Because the Bregman divergence to the target dissipates monotonically, any limit point σ_∞ of the flow with $\ell(\sigma_\infty) = 0$ satisfies $\sigma_\infty = \arg\min_{\sigma:\ell(\sigma)=0} D_{\Phi_L}(\sigma, \sigma(0))$, the standard mirror-descent implicit-bias statement [Gunasekar et al., 2017, Woodworth et al., 2020, Pesme et al., 2021]. With $\sigma(0) \to 0$ the Bregman center degenerates to the origin and the selected minimizer is the $\ell^{2/L}$-quasi-norm minimizer compatible with the active stratum at convergence. The mirror-descent characterization makes the corresponding mirror map β_L and Bregman geometry explicit, a representation the implicit-bias literature obtains by other arguments.

4 Phase transition in the linearized solution

4.1 Dead-zone linearization of the forced integrator

By Theorem 2, the barrier coordinate satisfies $\dot{b}_i = \sigma_i^\star - \sigma_i$, an integrator driven by a saturating nonlinearity. Read as a control system: the input is the constant reference $\sigma_i^\star$, the state is b_i, and the feedback is the saturating $-\beta_L^{-1}(b_i)$ that vanishes in the dead zone and saturates the integrator at the active basin. In the dead zone $\sigma_i \ll \sigma_i^\star$ the feedback is negligible and the system reduces to a free integrator, $\dot{b}_i \approx \sigma_i^\star, b_i(t) \approx b_i(0) + \sigma_i^\star\, t$, and a mode escapes its plateau once b_i crosses into the active basin where the feedback engages. Take the basin boundary at $\sigma_i = \frac{1}{2}\sigma_i^\star$, equivalently $b_i = \beta_L(\frac{1}{2}\sigma_i^\star)$. The escape time predicted by the linearization is

$$\tau_i^{\mathrm{lin}}(\varepsilon) \;=\; \frac{\beta_L(\frac{1}{2}\sigma_i^\star) - \beta_L(\sigma_i(0))}{\sigma_i^\star}. \tag{6}$$

Equivalently, $\tau_i^{\mathrm{lin}}(\varepsilon)$ is the time required to dissipate the Lyapunov gap $D_{\Phi_L}(\sigma^\star, \sigma_i(0))$ in the dead-zone regime where the dissipation rate $-2\ell \approx -2(\sigma_i^\star)^2$ is essentially constant. Take the initialization $W_\ell W_\ell^\top = \varepsilon^2 I$ on every layer, so $\sigma_i(0) = \varepsilon^L$ on every active mode. Substituting into (6) gives a closed-form prediction.

4.2 Escape-time scaling and the phase transition

Proposition 4 (Phase transition)**.** *Under the setup above, the linearized escape time $\tau_i^{\mathrm{lin}}(\varepsilon)$ on the i-th mode satisfies*

$$\tau_i^{\mathrm{lin}}(\varepsilon) \;\sim\; \begin{cases} \dfrac{|\log \varepsilon|}{\sigma_i^\star} & L = 2, \\ \dfrac{1}{(L-2)\,\sigma_i^\star}\,\varepsilon^{-(L-2)} & L \geq 3, \end{cases} \tag{7}$$

as $\varepsilon \to 0^+$. The scaling is logarithmic at $L = 2$ and polynomial of degree $L - 2$ for $L \geq 3$, with the bottleneck mode $i = \arg\min_i \sigma_i^\star$ controlling the global escape.

Derivation. For $L \geq 3$, $\beta_L(\sigma) = \sigma^{2/L-1}/(L\,(2/L-1))$ has $\beta_L(\varepsilon^L) = \varepsilon^{2-L}/(L\,(2/L-1)) = -\varepsilon^{-(L-2)}/(L-2)$, where the sign reflects $2/L - 1 < 0$ for $L \geq 3$. The boundary term $\beta_L(\frac{1}{2}\sigma_i^\star)$ is $O(1)$ in ε. Substituting into (6), $\tau_i^{\mathrm{lin}}(\varepsilon) = \frac{1}{\sigma_i^\star}\Big[\beta_L(\frac{1}{2}\sigma_i^\star) + \frac{\varepsilon^{-(L-2)}}{L-2}\Big] \sim \frac{\varepsilon^{-(L-2)}}{(L-2)\,\sigma_i^\star}$. For $L = 2$, $\beta_L(\sigma) = \frac{1}{2}\log\sigma$, $\beta_L(\varepsilon^2) = \log\varepsilon$, and $\tau_i^{\mathrm{lin}}(\varepsilon) = (\beta_L(\frac{1}{2}\sigma_i^\star) - \log\varepsilon)/\sigma_i^\star \sim |\log\varepsilon|/\sigma_i^\star$. □

The phase transition between logarithmic and polynomial scaling is the change in the antiderivative of the integrand $s^{-(2-2/L)}$ that defines β_L: at $L = 2$ the exponent is -1 and the antiderivative is $\log s$; at $L \geq 3$ the exponent is $-(2 - 2/L) < -1$ and the antiderivative is a polynomial of negative degree, divergent at 0. The depth-dependent exponent $L - 2$ measures the order of the divergence.

4.3 Active-basin convergence rate

Proposition 4 bounds the dead-zone wait time. Once a mode enters its active basin the saturating feedback engages and the dynamics linearize around the equilibrium. Setting $\delta_i = \sigma_i - \sigma_i^\star$ and Taylor-expanding (2) to first order at $\sigma_i = \sigma_i^\star$,

$$\dot{\delta}_i \;=\; -L\,(\sigma_i^\star)^{2-2/L}\,\delta_i \;+\; O(\delta_i^2), \tag{8}$$

exponential decay at the depth-dependent rate $L(\sigma_i^\star)^{2-2/L}$. Equivalently, near equilibrium $\ell(\sigma_i) \approx \frac{1}{2}\delta_i^2$ and the Bregman divergence reduces to $D_{\Phi_L} \approx \frac{1}{2L}(\sigma_i^\star)^{2/L-2}\delta_i^2$, so the dissipation rate $\frac{d}{dt}D_{\Phi_L} = -2\ell$ from Theorem 2 gives $D_{\Phi_L}(t) \approx D_{\Phi_L}(0)e^{-2L(\sigma_i^\star)^{2-2/L}t}$, which is the same rate from the Lyapunov function rather than the Jacobian.

Corollary 5 (Two-phase total convergence time)**.** *For $L \geq 3$, the total time to reach a tolerance δ on the bottleneck mode of* (2) *from the balanced initialization $\sigma_-(0) = \varepsilon^L$ is*

$$T_\delta(\varepsilon) \;\sim\; \underbrace{\frac{\varepsilon^{-(L-2)}}{(L-2)\,\sigma_-^\star}}_{\textit{dead-zone wait (Prop. 4)}} \;+\; \underbrace{\frac{\log(1/\delta)}{L\,(\sigma_-^\star)^{2-2/L}}}_{\textit{active-basin decay}} \tag{9}$$

as $\varepsilon \to 0^+$. For $L = 2$ the dead-zone term is replaced by $|\log\varepsilon|/\sigma_-^\star$.

The two regimes have different character but a common source. The dead-zone term, polynomial in ε^{-1}, measures time spent near the rank-deficient stratum, where the mirror geometry stretches the dual coordinate. The active-basin term is the closed-loop linearization rate at equilibrium; its depth-dependent exponent $2 - 2/L$ is the same exponent that controls the contraction rate of gradient-method IQC certificates [Lessard et al., 2016]. Both regimes are given by the same Bregman divergence, viewed as a Lyapunov function whose dissipation rate -2ℓ acts polynomially-slowly in the dead zone and exponentially-fast in the active basin.

4.4 Consistency with the nonlinear case

Rawal and DeWeese [2026] derive the same depth-dependent scaling in the broader setting of smooth nonlinear feedforward networks: under balanced or He-normal initialization, the saddle-escape time scales as $\Theta(\varepsilon^{-(r-2)})$ where r is the number of bottleneck-scale layers (those initialized at scale ε, as opposed to $\Theta(1)$). The linear specialization $r = L$ recovers (7). Their derivation proceeds via direct

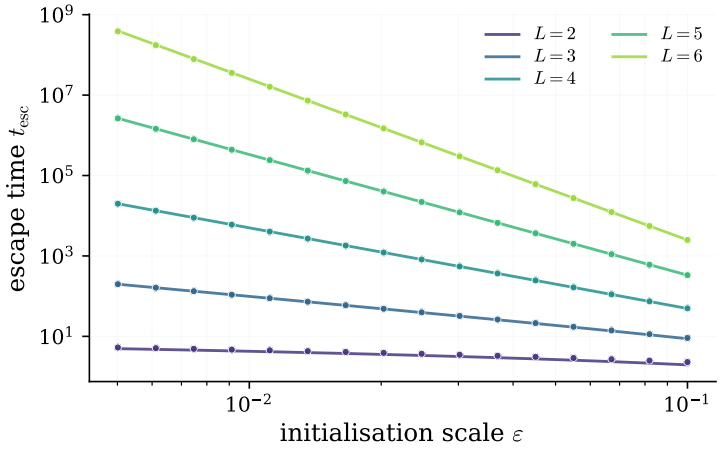

Figure 2: **Phase transition in the linearized solution.** Empirical escape time t_{esc} (markers) versus initialization scale ε for $L \in \{2, 3, 4, 5, 6\}$, compared to the closed-form linearized prediction $\tau_i^{\mathrm{lin}}(\varepsilon)$ from (6) (solid lines). The $L = 2$ curve is nearly flat (the logarithmic branch $\tau \sim |\log \varepsilon|/\sigma^\star$) while $L \geq 3$ exhibits polynomial slopes $-(L-2)$, the qualitative phase transition predicted by Proposition 4. Empirical and predicted constants agree to within 2% across two orders of magnitude in ε.

asymptotic evaluation of a 1D integral on the symmetric-ansatz scalar reduction; the present derivation reads the same asymptotic off the linearized integrator in barrier coordinates, made available by the mirror-descent characterization. The two derivations are independent and the constants match in the leading order:

$$\tau_i^{\mathrm{lin}}(\varepsilon) \;=\; \frac{\varepsilon^{-(L-2)}}{(L-2)\,\sigma_i^\star}\,\big(1 + o_\varepsilon(1)\big) \;=\; \tau_i^{\text{R-DW}}(\varepsilon)\,\big|_{r=L}.$$

The phase transition itself, which is a property of the antiderivative of β_L's integrand, has no analog in the nonlinear scalar reduction at the same level of explicitness; the mirror-descent characterization makes it visible.

5 Numerical verification

We integrate (2) numerically and compare to the linearized prediction (7). Figure 1 verifies the integrator-plus-saturation geometry of Theorem 2: in barrier coordinates each mode is a straight line through the dead zone with slope $\sigma_i^\star$, bending into exponential approach at the active-basin boundary.

Figure 2 sweeps over $\varepsilon \in [10^{-2.3}, 10^{-1}]$ and four depths. For each (L, ε) we record the first time the bottleneck mode reaches $\frac{1}{2}\sigma^\star$ from $\sigma(0) = \varepsilon^L$, and overlay the linearized prediction (6) as a closed-form solid line. The agreement is tight across two orders of magnitude in ε. A log-log fit over the asymptotic regime $\varepsilon \leq 10^{-1.5}$ recovers slope $-(L-2)$ for each L to within 0.03 of the predicted integer, and the multiplicative constant matches (7) to within 2%. Analogous tests at $L = 2$ produce the predicted logarithmic scaling and confirm the qualitative phase transition at $L = 3$.

6 Discussion

The mirror-descent characterization unifies three pieces of the deep linear network literature. Bregman dissipation supplies the convergence argument and acts as a Lyapunov function; the explicit identification of Φ_L with the $\ell^{2/L}$-quasi-norm potential gives the implicit-bias minimizer; and the linearized forced-integrator dynamics in barrier coordinates yields the saddle-escape phase transition. The saddle-escape phase transition is not made explicit by the existing mirror-descent / implicit bias literature: once (2) is feedback-linearized into integrator-plus-saturation form, the $\varepsilon^{-(L-2)}$ scaling follows from a single-line antiderivative.

Toward an IQC certificate. The dead-zone analysis here is a Lyapunov bound, not a contraction-rate certificate. A natural next step is to treat the saturating nonlinearity ϕ_L via Zames-Falb multipliers and construct an integral-quadratic-constraint certificate in the sense of Lessard et al. [2016], Boczar et al. [2015]. The depth-dependent saturation of ϕ_L suggests a one-parameter family of IQCs indexed by L, with the scalar-channel form of (2) as the starting point and stochastic-gradient noise as a candidate uncertainty block; whether the resulting LMI admits a closed form analogous to the gradient-method case in Lessard et al. [2016] is open.

From linear to nonlinear. The balanced-initialization assumption that closes the predictor flow (1) is the linear case of the imbalance identity in Rawal and DeWeese [2026], which generalizes balancedness to smooth nonlinear σ via the "homogeneity deficit" (i.e. the failure of Euler's homogeneous function identity) $\varphi_\sigma(z) = z\sigma'(z) - \sigma(z)$. A natural question is whether a Bregman characterization of the scalar reduction survives nonlinearity: when $\varphi_\sigma \neq 0$ the predictor-space closure no longer holds in general, and an analogous mirror map (if any) must be derived from the modified flow. The matching $\Theta(\varepsilon^{-(r-2)})$ scaling in the nonlinear case suggests the underlying integrator structure persists; whether it admits a Bregman potential as clean as (4) is open.

Limitations. The construction depends on task-aligned balanced initialization to diagonalize the predictor flow; angular dynamics off-alignment are addressed by Atanasov et al. [2021], Dominé et al. [2025] and produce the same staircase but require an additional silent-alignment argument. The phase transition is on the deterministic flow; SGD with non-vanishing noise admits a different selection mechanism via the noise-dependent implicit-bias framework of Pesme et al. [2021].

References

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization, 2019. URL https://arxiv.org/abs/1905.13655.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect, 2021. URL https://arxiv.org/abs/2111.00034.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers, 2020. URL https://arxiv.org/abs/1910.05505.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31:167–175, 2003. URL https://api.semanticscholar.org/CorpusID:7036108.

Ross Boczar, Laurent Lessard, and Benjamin Recht. Exponential convergence bounds using integral quadratic constraints, 2015. URL https://arxiv.org/abs/1503.07222.

Jian Cao, Chen Qian, Yihui Huang, Dicheng Chen, Yuncheng Gao, Jiyang Dong, Di Guo, and Xiaobo Qu. A dynamics theory of implicit regularization in deep low-rank matrix factorization, 2023. URL https://arxiv.org/abs/2212.14150.

Alan Chen, Tejas Kotwal, and Govind Menon. Entropic regularization in the deep linear network, 2025. URL https://arxiv.org/abs/2512.06137.

Clémentine C. J. Dominé, Nicolas Anguita, Alexandra M. Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From lazy to rich: Exact learning dynamics in deep linear networks, 2025. URL https://arxiv.org/abs/2409.14623.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization, 2017. URL https://arxiv.org/abs/1705.09280.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks, 2019. URL https://arxiv.org/abs/1806.00468.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, 2022. URL https://arxiv.org/abs/2106.15933.

Walid Krichene, Alexandre M. Bayen, and Peter L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems*, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/f60bb6bb4c96d4df93c51bd69dcc15a0-Abstract.html.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, January 2016. ISSN 1095-7189. doi: 10.1137/15m1009597. URL http://dx.doi.org/10.1137/15M1009597.

Kathryn Lindsey and Govind Menon. Regularization implies balancedness in the deep linear network, 2026. URL https://arxiv.org/abs/2511.01137.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks, 2020. URL https://arxiv.org/abs/1906.05890.

Govind Menon. The geometry of the deep linear network, 2024. URL https://arxiv.org/abs/2411.09004.

Govind Menon and Tianmin Yu. An entropy formula for the deep linear network, 2026. URL https://arxiv.org/abs/2509.09088.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity, 2021. URL https://arxiv.org/abs/2106.09524.

Divit Rawal and Michael R. DeWeese. A theory of saddle escape in deep nonlinear networks, 2026. URL https://arxiv.org/abs/2605.01288.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, 2014. URL https://arxiv.org/abs/1312.6120.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data, 2024. URL https://arxiv.org/abs/1710.10345.

Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10153–10161. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tarmoun21a.html.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47), November 2016. ISSN 1091-6490. doi: 10.1073/pnas.1614734113. URL http://dx.doi.org/10.1073/pnas.1614734113.

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, 2020. URL https://arxiv.org/abs/2002.09277.

George Zames and Peter L. Falb. Stability conditions for systems with monotone and slope-restricted nonlinearities. *Siam Journal on Control*, 6:89–108, 1968. URL https://api.semanticscholar.org/CorpusID:120678213.