A STOCHASTIC–GEOMETRIC THEORY OF SCALING LAWS IN GROKKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Delayed generalization (*i.e.* grokking) refers to the phenomenon in which a neural network fits its training data early in training but only begins to generalize after a prolonged delay, often through an abrupt transition. Despite extensive empirical study, its underlying mechanism remains poorly understood. In this work, we first theoretically characterize a shell–core topological configuration of the reachable solution space induced by Adam's optimization dynamics with weight-shrinkage regularization, supported by empirical evidence. This optimization-induced topological configuration gives rise to grokking. In model's parameter space, random initialization solutions concentrate on a thin outer spherical shell, enclosing another spherical shell of memorization solutions, which in turn contains a core corresponding to the generalization solutions. Leveraging stopping-time theory, we then analyze the geometry of this topological configuration and the solution transition time at which optimization trajectories escape the memorization manifold and first reach the boundary of the generalization manifold. Our theoretical analysis derives grokking scaling laws for the learning rate, batch size, and ℓ_2 regularization coefficient, which are further validated through experiments and shown to recover results from prior literature.

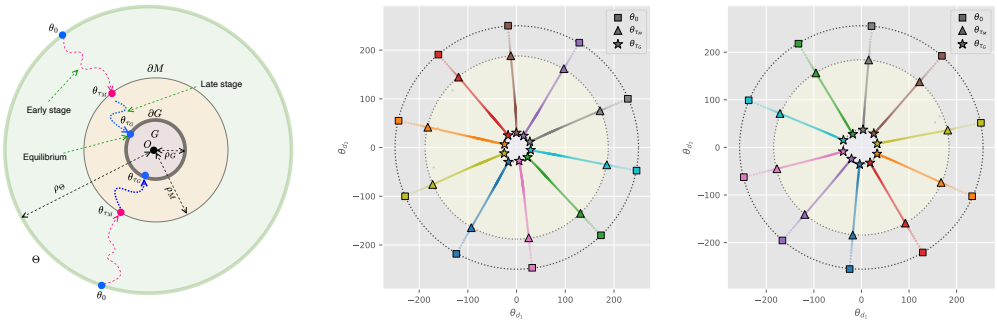
1 INTRODUCTION

Neural networks trained on noise-free, highly structured learning tasks have been observed to exhibit an *epiphany* phenomenon known as delayed generalization, or grokking (Power et al., 2022). In these settings, models rapidly achieve near-zero training loss, often early in training, yet fail to generalize for an extended period before abruptly transitioning to strong test performance. This behavior has been reported across a range of tasks with exact underlying structure (Hwang & Park, 2026), such as modular arithmetic (Power et al., 2022; Liu et al., 2022; Zhong et al., 2023), algorithmic reasoning (Nanda et al., 2023), and group-theoretic learning (Stander et al., 2024; Notsawo et al., 2026). Moreover, prior work has shown that the onset of grokking is highly sensitive to hyperparameter choices (Zhong et al., 2023), including the fraction of training data, learning rate, batch size, and regularization coefficient (Power et al., 2022; Liu et al., 2022). Despite extensive empirical investigation, however, its underlying mechanism remains poorly understood.

Empirical findings consistently suggest that training dynamics in the grokking regime exhibit three-stage behavior (Power et al., 2022; Nanda et al., 2023; Kumar et al., 2024). In the initial stage, the model rapidly interpolates the training set and reaches the memorization regime. In the second stage, the trajectory remains near the memorization solutions for an extended period before escaping and eventually reaching the generalization regime. The length of this extended period scales with the task, learning rate, batch size, and regularization coefficient. In the third stage, drift and diffusion equilibrate in the optimization-induced dynamics, and the trajectory settles within the generalization solution region centered at its minimizer.

Under the regime with ℓ_2 weight-shrinkage regularization, these observations imply the existence of two distinct classes of solutions: *memorization solutions*, which only interpolate the training data and does not generalize, and *generalization solutions*, which additionally achieve low loss with respect to the underlying data distribution. Their topological configuration is thus induced by optimization dynamics, and characterized into three stages: the *early stage*, the *late stage*, and the *equilibrium*, as shown in Figure 1a. The *early-stage* dynamics are driven by deterministic drift from

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.



(a) Weight-Shrinkage Grokking (b) Trajectories w/ S_5 (c) Trajectories w/ $\mathbb{Z}_{127}$

Figure 1: **Optimization-Induced Shell–Core Topology and Empirical Evidence.** Left (a): grokking dynamics under weight-shrinkage regularization induce a shell–core topology in parameter space, where initialization points θ_0 concentrate on a hyperspherical shell Θ with radius $\rho_\Theta = O(k)$ and thickness $O(k/\tilde{p})$, where k denotes the number of parameter matrices and $\tilde{p}$ denotes effective parameter dimension (see Section 3), enclosing the memorization shell $M \setminus G$ and the generalization core G; the radii ρ_G, ρ_M, and ρ_Θ are determined by the task and optimization configuration. Middle–right (b–c): empirical evidence on S_5 and $\mathbb{Z}_{127}$ using isometric manifold learning (*e.g.*, MDS) to visualize ten optimization parameter trajectories, where θ_0, θ_{τ_M}, and θ_{τ_G} denote initialization, first memorization, and first generalization states. The projections show that Adam-reachable M and G are approximately symmetric embedded in ambient space $\mathbb{R}^p$.

large gradients, which dominates the optimization trajectory and rapidly drives it toward the memorization manifold. At this stage, the gradient fluctuations are anisotropic. The *late-stage* dynamics are driven more slowly by decaying but nonzero gradients, together with isotropic gradient fluctuations, which guide the optimization trajectory across the trivial memorization solutions. Once the trajectory reaches the generalization set, the drift and diffusion in the optimization dynamics equilibrate and confine the trajectory within the generalization manifold, referred to as the *equilibrium*. Figure 1b–1c provide empirical evidence consistent with this optimization-induced topological configuration by visualizing optimization trajectories via isometric manifold learning. This optimizer-induced topological configuration, as illustrated in Figure 1a, gives rise to grokking and forms a shell–core configuration: initialization solutions concentrate on a thin hyperspherical shell, which encloses a shell of memorization solutions containing a compact core of generalization solutions. Detailed empirical results of Adam–induced shell–core radii are provided in Appendix A.2.

We theoretically analyze the geometry of this topological configuration and derive scaling laws for the manifold radii and the solution transition time from memorization to generalization through a *stopping-time analysis* (Dynkin, 1965) of stochastic differential equations (SDEs) (Øksendal, 2003). The remainder of the paper is organized as follows. Section 3 presents initialization concentration, and Section 4.1 establishes the joint-state SDE for Adam (Kingma & Ba, 2015) designed to analyze grokking dynamics. Building on this formulation, Section 4.2 derives the scaling law for the memorization-manifold radius. Under the late-stage reduction observations of grokking, Section 4.3 derives a closed-form preconditioned radius SDE for the evolution of manifold radii. Sections 4.4–4.5 derive the scaling laws for the generalization-manifold radius and the solution transition time from memorization to generalization. Finally, Section 5 validates the theory through experiments on modular arithmetic and group-theoretic learning tasks, and by comparison with the grokking literature. The contributions are summarized below:

1. **Optimization-Induced Shell–Core Topology.** We introduce an optimizer-induced shell–core prior for grokking, where initialization concentrates on a thin hyperspherical shell enclosing memorization solutions, which in turn contain a compact core of generalization solutions.
2. **Stochastic–Geometric Characterization.** Leveraging the topological prior, we formulate grokking as a stochastic transition problem and characterize the geometry by the first stopping time at which stochastic gradient flows escape the memorization manifold and reach the generalization manifold.

3. **Empirical, Literature, and Symbolic-Algebra Validation.** We validate our theory through experiments on group-theoretic learning and modular arithmetic tasks, and by comparison with the literature, recovering known scaling laws for the learning rate, batch size, and regularization coefficient. We additionally machine-check all closed-form and asymptotic results appearing in the proofs by a symbolic algebra system (SAS) with Julia.

2 NOTATIONS

Data and Network. Let $\{(x_i, y_i)\}_{i=1}^n$ denote a finite training set of size n, where each pair $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ is drawn i.i.d. from the empirical distribution $\widehat{P}_{\mathcal{X}\mathcal{Y}}$, with underlying true distribution $P_{\mathcal{X}\mathcal{Y}}$. Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ be a model parameterized by $\theta \in \mathbb{R}^p$. Throughout, we use X_t to denote the time-indexed value of a variable X at time t; for example, θ_t denotes the parameter at time t.

Instance, Batch, Empirical, and True Loss. Let $\ell_f(x, y; \theta)$ denote the instance loss evaluated at a point (x, y). Let $\mathscr{L}_f(\xi; \theta)$ denote the *batch loss* evaluated on a mini-batch $\xi = \{(x_{\xi_i}, y_{\xi_i})\}_{i=1}^b$ of size b. Let $\widehat{\mathscr{L}_f}(\theta) := \mathbb{E}_{\xi \sim \widehat{P}_{\mathcal{X}\mathcal{Y}}}[\mathscr{L}_f(\xi; \theta)]$ denote the *empirical loss* evaluated over the training set. Let $\mathscr{L}_f(\theta)$ denote the *true loss* with respect to the underlying data distribution $P_{\mathcal{X}\mathcal{Y}}$.

Initialization, Memorization, and Generalization Manifolds. We study the solution manifolds reachable under the dynamics induced by Adam (Kingma & Ba, 2015). The ambient parameter space is $\mathbb{R}^p$ endowed with the Euclidean metric. Let Θ denote the set of *initialization solutions* under some initialization distribution P_Θ. Let $M := \{\theta \in \mathbb{R}^p \mid \widehat{\mathscr{L}_f}(\theta) < \epsilon \ \wedge \ p_{\text{Adam}}(\theta \mid \theta_0 \in \Theta) > 0\}$ denote the set of *memorization solutions* under Adam's dynamics with a loss tolerance ϵ. Similarly, let $G := \{\theta \in \mathbb{R}^p \mid \mathscr{L}_f(\theta) < \epsilon \ \wedge \ p_{\text{Adam}}(\theta \mid \theta_0 \in \Theta) > 0\}$ denote the set of *generalization solutions*. Throughout, ∂S and S° denote the boundary and interior of a set S, respectively; ∂M is the outer boundary of $M \setminus G$, and ∂G is its inner boundary. Let $\rho_M^2 := \mathbb{E}_{\theta_0 \in \Theta, \theta \in \partial M}\left[\|\theta\|_2^2 \mid \theta_0\right]$ and $\rho_G^2 := \mathbb{E}_{\theta_0 \in \Theta, \theta \in \partial G}\left[\|\theta\|_2^2 \mid \theta_0\right]$ denote the mean squared radii of the memorization and generalization manifolds, respectively.

3 INITIALIZATION CONCENTRATION

We first demonstrate the topological concentration of network-parameter initialization under normal and uniform distributions. Consider that $\theta_0 = (\theta_0^{(1)}, \cdots, \theta_0^{(k)}) \in \mathbb{R}^p$ consists of k sub-vectors such that $\theta_0^{(j)} \in \mathbb{R}^{p_j}$ where $\sum_{j=1}^k p_j = p$. Define effective dimension $\tilde{p}$ such that $\frac{1}{\tilde{p}} := \frac{1}{k}\sum_{j=1}^k \frac{1}{p_j}$. For coordinate-wise initialization $[\theta_0^{(j)}]_i \sim \mathcal{N}(0, \sigma_j^2)$, the initialization radius square

$$\rho_\Theta^2(\theta_0) = \|\theta_0\|_2^2 = \sum_{j=1}^k \left(\sum_{i=1}^{p_j} \left| [\theta_0^{(j)}]_i \right|^2 \right) \sim \sum_{j=1}^k \sigma_j^2 \chi_{p_j}^2, \tag{1}$$

follows a sum of scaled chi-square distributions, where each $\sigma_j^2 \chi_{p_j}^2$ has p_j degrees of freedom. Hence, the radius of Θ admits the closed-form expression

$$\tilde{\sigma}^2 := \frac{1}{p}\sum_{j=1}^k p_j \sigma_j^2, \quad \mathbb{E}_{\theta_0}\left[\rho_\Theta^2(\theta_0)\right] \ = \ \tilde{\sigma}^2 p = O(k), \quad \text{Var}_{\theta_0}\left[\rho_\Theta^2(\theta_0)\right] = 2\sum_{j=1}^k p_j \sigma_j^4 = O(\frac{k}{\tilde{p}}),$$

under a coordinate scaling $\sigma_j = O(1/\sqrt{p_j})$ (LeCun et al., 1998; Glorot & Bengio, 2010). This suggests that initialization solutions concentrate on a thin annulus of radius $O(k)$ with thickness $O(\sqrt{k/\tilde{p}})$ in parameter space. Similar concentration also holds for coordinate-wise uniform initialization distribution $[\theta_0^{(j)}]_i \sim \mathcal{U}(-\varepsilon_j, \varepsilon_j)$,

$$\tilde{\varepsilon}^2 := \frac{1}{p}\sum_{j=1}^k p_j \varepsilon_j^2, \quad \mathbb{E}_{\theta_0}\left[\rho_\Theta^2(\theta_0)\right] \ = \ \frac{p\tilde{\varepsilon}^2}{3} = O(k), \quad \text{Var}_{\theta_0}\left[\rho_\Theta^2(\theta_0)\right] = \frac{4}{45}\sum_{j=1}^k p_j \varepsilon_j^4 = O(\frac{k}{\tilde{p}}),$$

under a coordinate scaling $\varepsilon_j = O(1/\sqrt{p_j})$. Proofs for the normal and uniform concentration results are provided in Appendices A.3 and A.4, respectively.

4 MANIFOLD RADIUS AND SOLUTION TRANSITION ANALYSIS

We characterize Adam's dynamics through its continuous-time limit, formulated as an SDE on the joint state of the parameters and first- and second-moment estimates. Based on this joint-state SDE, we derive the scaling law for the *memorization radius* ρ_M, defined as the parameter norm at the first-hitting time of the empirical-loss boundary ∂M, via perturbation method (Bender & Orszag, 1999). Further leveraging two reduction properties of the grokking dynamics, we reduce the intractable joint-state SDE for Adam (Section 4.1) to a tractable radial SDE. This reduction enables an analytical characterization of the *generalization radius* ρ_G, defined as the parameter norm at the first-hitting time of the generalization boundary ∂G, and the *solution transition time* $\mathbb{E}[\tau_{M\to G}]$, defined as the mean time required to transition from memorization boundary ∂M to generalization boundary ∂G. A high-level sketch for theoretical analysis framework is illustrated as in Figure 6 of the appendix.

4.1 ADAM'S JOINT-STATE CONTINUOUS-TIME SDE LIMIT

We consider Adam dynamics in the small learning rate and large batch size regime $\eta \to 0$, $b \to \infty$, with exponential moving average (EMA) coefficients $(\beta_1, \beta_2) \in (0,1)^2$ and an ℓ_2 regularizer $R(\theta) = \frac{1}{2}\|\theta\|_2^2$ with coefficient $\lambda > 0$. Let $g_t := \nabla[\mathscr{L}_f^*(\xi_t;\theta_t) + \lambda R(\theta_t)]$ denote the regularized mini-sample gradient flow, with per-sample mean $\bar{g}_t := \mathbb{E}_s[\nabla \ell_f(s;\theta_t)]$ and covariance $\Sigma_t := \mathrm{Cov}_s[\nabla \ell_f(s;\theta_t)]$. For sufficiently large batch size b, the mini-batch gradient admits the distributional limit $g_t \sim \mathcal{N}(\bar{g}_t, \Sigma_t/b)$. We refer to this setting as *coupled weight decay*, since the ℓ_2 regularizer contributes to the gradient used to update Adam's first- and second-moment estimates. In contrast, the *decoupled weight decay* scheme applies weight decay as a separate parameter-shrinkage step and is commonly referred to as *AdamW* (Loshchilov & Hutter, 2019).

Adam's Parameter Update Rules. Let $\theta \in \mathbb{R}^p$ be model's parameters. Let $m, v \in \mathbb{R}^p$ be the first- and second-moment estimates, respectively. Let $\hat{m}, \hat{v}$ be the m, v with bias corrections. Adam's discrete update rules at iteration k are given as

$$m_{k+1} = \beta_1 m_k + (1-\beta_1) g_k, \qquad v_{k+1} = \beta_2 v_k + (1-\beta_2)(g_k \odot g_k),$$
$$\hat{m}_{k+1} = m_{k+1} / (1-\beta_1^{k+1}), \qquad \hat{v}_{k+1} = v_{k+1} / (1-\beta_2^{k+1}),$$
$$\theta_{k+1} = \theta_k - \eta\, \hat{m}_{k+1} \oslash \left(\sqrt{\hat{v}_{k+1}} + \varepsilon\, \mathbf{1}_p\right),$$

with initial values $\theta_0 \in \Theta, m_0 = 0, v_0 = 0$, where $\odot$ is element-wise product and $\oslash$ is element-wise division. $\hat{m}$ and $\hat{v}$ are Adam's bias corrections, which compensate the warm-up underestimation of (m_k, v_k) caused by the zero initialization $m_0 = v_0 = 0$. We drop Adam's numerical-stability constant ε throughout the analysis as it does not affect the analysis. Defining the Adam joint state as $S_t := (\theta_t, m_t, v_t) \in \mathbb{R}^{3p}$ and using the continuous-time interpolation $t = \eta k$, we state the resulting joint-state SDE for Adam in Lemma 1. The proof sketch is illustrated in Figure 7. The proof is provided in Appendix A.5. The technical correctness is verified through an induced radius SDE with Itô's lemma in Appendix A.6. Lemma 1 provides a stochastic–theoretical framework for studying grokking dynamics induced by Adam optimizer.

Lemma 1 (Adam's Joint-State Continuous-Time SDE Limit)**.** *Let $S_t := (\theta_t^\top, m_t^\top, v_t^\top)^\top \in \mathbb{R}^{3p}$ denote the joint state of Adam, where θ_t, m_t, and v_t are the parameters, first-moment estimates, and second-moment estimates, respectively. Taking the continuous-time interpolation $t = \eta k$, the discrete Adam updates admit the following Itô SDE limit on the joint state*

$$\mathrm{d}S_t \;=\; \mu(S_t)\,\mathrm{d}t \;+\; \sqrt{\frac{\eta}{b}}\,\sigma(S_t)\,\mathrm{d}W_t, \tag{2}$$

where $\mathrm{d}W_t \in \mathbb{R}^p$ denotes the infinitesimal increment of a Wiener process adapted to the filtration generated by the mini-batch sampling process $\{\xi_t\}$, with drift factor $\mu(S_t) \in \mathbb{R}^{3p}$ and diffusion factor $\sigma(S_t) \in \mathbb{R}^{3p\times p}$

$$\mu(S_t) \;=\; \begin{pmatrix} -B(t)\, m_t \oslash \sqrt{v_t} \\ -\alpha_1 (m_t - \bar{g}_t) \\ -\alpha_2 (v_t - \overline{g \odot g}_t) \end{pmatrix}, \quad \sigma(S_t) \;=\; \begin{pmatrix} 0 \\ \alpha_1 \Sigma_t^{1/2} \\ \alpha_2 D_t \end{pmatrix}, \quad B(t) := \frac{\sqrt{1-e^{-\alpha_2 t}}}{1-e^{-\alpha_1 t}}, \tag{3}$$

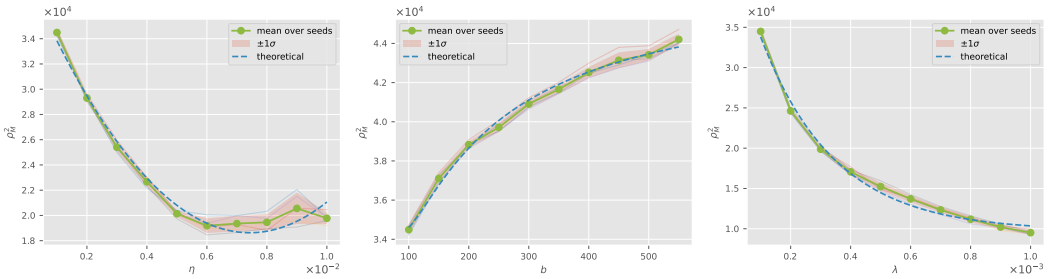


Figure 2: **Scaling Laws of Manifold Radius ρ_M^2 on S_5.** We show the scaling law of ρ_M^2 with respect to the learning rate η, batch size b, and ℓ_2 regularization coefficient λ on the S_5 task. For each hyperparameter configuration, we train for ten runs. The results show that larger η/b modifies the dynamics with a stronger diffusion variance, whereas λ does not affect the diffusion variance. We also overlay the theoretical fits. **As predicted by theory, the scaling law of ρ_M^2 with respect to the learning rate η exhibits a U-shaped curve.** An additional experimental results for $\mathbb{Z}_{127}$ are provided in Appendix A.15.

where $\alpha_i := (1-\beta_i)/\eta$ and $B(t)$ is the bias-correction factor in continuous time and

$$\overline{g \odot g}_t \;:=\; \bar{g}_t \odot \bar{g}_t \;+\; \frac{1}{b}\,\mathrm{diag}(\Sigma_t) \in \mathbb{R}^p, \quad D_t \;:=\; 2\,\mathrm{diag}\,(\bar{g}_t)\;\Sigma_t^{1/2} \in \mathbb{R}^{p\times p}. \tag{4}$$

In particular, to decompose this SDE into a tractable form, we define $\pi(\theta_t) := \mathrm{diag}(\overline{g \odot g}_t)^{-1/2} \in \mathbb{R}^{p\times p}$ as Adam's preconditioner. We write $\theta_t(\theta_0), \theta_t(S_0)$ to denote the Adam's θ-evolution starting from initial states $\theta_0 \in \Theta, S_0 \in \Theta_S$, respectively, where Θ_S is the initialization distribution of S.

Remark 1. While the continuous-time SDE limit of ordinary stochastic gradient descent (SGD) is now well established (Li et al., 2017; Mandt et al., 2017; Luo et al., 2025), a rigorous continuous-time treatment of Adam that explicitly accounts for stochasticity remains, to the best of our knowledge, comparatively less developed. For example, the joint (m, v) deterministic limit of Adam appears in (Da Silva & Gazeau, 2020), and approximations for adaptive methods such as RMSprop and Adam were established in (Malladi et al., 2022; Compagnoni et al., 2025). Our joint-state SDE limit only partially overlaps with this line of work and yields an SDE limit that accounts for stochasticity without relying on a first-order approximation.

4.2 Memorization Radius

Starting from the initial state $S_0 = (\theta_0^\top, 0_p^\top, 0_p^\top)^\top$, Adam's joint-state SDE (2) generates a stochastic trajectory $\{S_t\}$ adapted to the filtration generated by mini-batch sampling. The first time at which θ_t reaches the memorization boundary ∂M defines the *stopping time* (Karatzas & Shreve, 1991) and mean squared memorization radius as

$$\tau_M(S_0) := \inf\{t \geq 0 : \theta_t(S_0) \in \partial M\}, \qquad \text{and} \qquad \rho_M^2 := \mathbb{E}_{S_0\in\Theta_S}\left[\|\theta_{\tau_M(S_0)}\|_2^2\right], \tag{5}$$

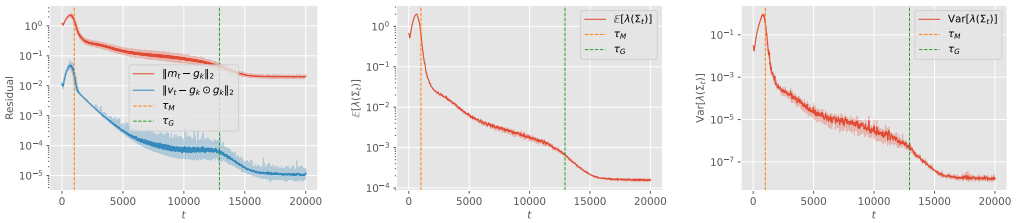
respectively. This defines a *Dirichlet exit problem* on the state S_t, with absorbing boundary ∂M and the *infinitesimal generator* (*i.e.*, a differential operator) (Dynkin, 1965) induced by Adam's SDE in Lemma 1 as:

$$\mathcal{L}_S[\bullet] \;:=\; \mu(S)^\top \nabla_S[\bullet] \;+\; \frac{1}{2}\,\frac{\eta}{b}\,\mathrm{tr}\big(\Sigma_S(S)\,\nabla_S^2\,[\bullet]\,\big), \tag{6}$$

which contains an operator modulated by a coefficient η/b with $\Sigma_S(S) := \sigma(S)\sigma(S)^\top$. The resulting Dirichlet partial differential equation (PDE) admits no closed-form solution, we apply a low-order regular perturbation expansion (Bender & Orszag, 1999) in $\varepsilon = \eta/b$ to obtain Theorem 1. The proof is provided in Appendix A.12. Experimental verification is provided for S_5 in Figure 2 and for $\mathbb{Z}_{127}$ in Appendix A.15.

Theorem 1 (Scaling Law of Memorization Radius (Perturbation Solution))**.** *The memorization radius approximately admits the scaling law with respect to η/b as*

$$\rho_M^2 \;=\; (\rho_M^{(0)})^2 \;+\; \frac{\eta}{b}\,c_M \;+\; \left(\frac{\eta}{b}\right)^2 c_M^{(2)} \;+\; O\!\left(\left(\frac{\eta}{b}\right)^3\right), \tag{7}$$



(a) Reduction on (m_t, v_t) (b) Gradient noise magnitude (c) Gradient noise isotropy

Figure 3: **Observations in Grokking Dynamics.** The markers τ_M and τ_G denote the first-hitting times of ∂M and ∂G, respectively, and the y-axis is shown on a logarithmic scale. Left (a) shows the exponential reduction of moment estimates; (m_t, v_t) rapidly converges to $\big(\bar{g}_t(1-e^{-\alpha_1 t}), \overline{g \odot g}_t(1-e^{-\alpha_2 t})\big)$, with exponential decay $\exp(-O(t))$. Middle–right (b–c) shows the exponential reduction of the gradient noise; the mean and variance of the singular values of the gradient covariance Σ_t decay as $\exp(-O(t))$, indicating small and isotropic gradient fluctuations in late-stage dynamics.

where $(\rho_M^{(0)})^2$, c_M, $c_M^{(2)}$ *are task-determined constants independent of* η *and* b, *as defined in Appendix A.12. The quadratic regularizer* $\frac{\lambda}{2}\|\theta_t\|_2^2$ *induces deterministic contraction. The leading constant* $(\rho_M^{(0)})^2$ *depends on the deterministic gradient flow* $\nabla \mathscr{L}_f(\theta_t)$, *when the task gradient* $\nabla \mathscr{L}_f^*(\theta_t)$ *is small,* λ *enters* $(\rho_M^{(0)})^2$ *through* $\mathrm{d}\|\theta_t\|_2^2 = 2\theta_t^\top \mathrm{d}\theta_t \approx -2\lambda\|\theta_t\|_2^2\,\mathrm{d}t \Rightarrow \frac{1}{(\rho_M^{(0)})^2}\,\mathrm{d}\Big(\rho_M^{(0)}\Big)^2 = -2\lambda\,\mathrm{d}t$. *Thus,* $(\rho_M^{(0)})^2 \propto \exp(-O(\lambda))$, *and hence* $\rho_M^2 \propto \exp(-O(\lambda))$.

Remark 2. In the radius-shrinkage regime, diffusion first shifts the hitting point inward, giving $c_M < 0$, while boundary curvature produces a stabilizing second-order correction, giving $c_M^{(2)} > 0$. Our theory predicts that ρ_M^2 is U-shaped in η/b with two regimes: a sub-linear decrease $\rho_M^2 \approx (\rho_M^{(0)})^2 + (\eta/b)c_M$ for $\eta/b \ll |c_M|/(2c_M^{(2)})$, and a sub-quadratic increase $\rho_M^2 \approx (\rho_M^{(0)})^2 + (\eta/b)^2 c_M^{(2)}$ for $\eta/b \gtrsim |c_M|/(2c_M^{(2)})$, with minimum at $(\eta/b)^\star = |c_M|/(2c_M^{(2)})$. This U-shaped effect is observed in Figure 2, where the scaling law with respect to η exhibits a U-shaped curve under large b.

4.3 Late-Stage Radius Dynamics

To characterize the manifold geometry, we derive the radius SDE induced by Adam's SDE (2) via Itô's lemma under the late-stage reduction observations. Empirically, as shown in Figure 3, once the trajectory $\{S_t\}$ reaches the memorization manifold M (*i.e.*, memorization regime), the late-stage grokking dynamics exhibit two reduction properties. First, refer to equation (3), the first- and second-moment estimates m_t and v_t rapidly converge to $\bar{g}_t\big(1 - e^{-\alpha_1 t}\big)$ and $\overline{g \odot g}_t$, respectively, at an exponential rate $\exp(-O(t))$. We refer to this property as *slow-manifold reduction*, formalized in Observation 1. Second, refer to equation (4), the gradient fluctuations, characterized by Σ_t, become small and isotropic at an exponential rate $\exp(-O(t))$. We refer to this property as *small-isotropic gradient covariance*, formalized in Observation 2.

Observation 1 (Slow-Manifold Reduction on (m_t, v_t))**.** If (m_t, v_t) varies slowly as $t \to \infty$, the first- and second-moment estimates m_t, v_t rapidly converge to its mean-field limit

$$m_t \;\to\; \bar{g}_t\left(1 - e^{-\alpha_1 t}\right) \to \bar{g}_t \quad \text{and} \quad v_t \;\to\; \overline{g \odot g}_t(1 - e^{-\alpha_2 t}) \to \overline{g \odot g}_t, \tag{8}$$

respectively, at an exponential rate $\exp(-O(t))$. The proof for this mean-field limit is provided in Lemma 7 of Appendix A.7.

Observation 2 (Small-Isotropic Gradient Covariance)**.** In Adam's late-stage dynamics, the gradient covariance Σ_t rapidly becomes small and approximately isotropic as $t \to \infty$. With this isotropy and by equation (4), the Adam's preconditioner is therefore approximated by $\pi(\theta) \approx s(\theta)^{-1} I_p$, as $t \to \infty$, where the state-dependent scalar $s(\theta) > 0$ is defined by

$$\frac{1}{s(\theta)} := \frac{1}{p}\mathrm{tr}\big(\pi(\theta)\big), \qquad \pi(\theta) := \mathrm{diag}\big(\overline{g \odot g}(\theta)\big)^{-1/2}, \tag{9}$$

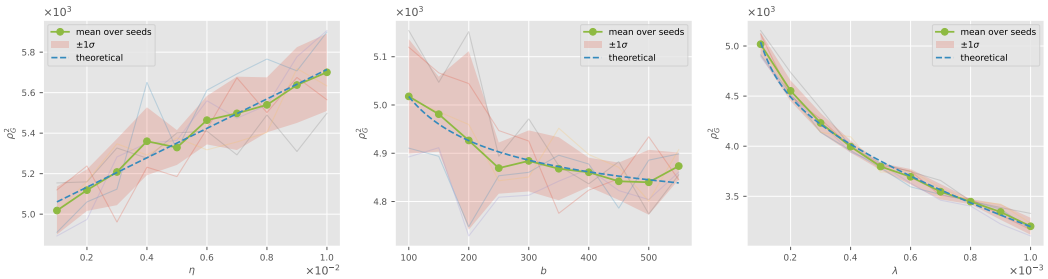


Figure 4: **Scaling Laws of Manifold Radius ρ_G^2 on S_5.** We show the scaling laws of ρ_G^2 with respect to the learning rate η, batch size b, and ℓ_2 regularization coefficient λ on the S_5 task. For each hyperparameter configuration, we train for ten runs. The results show that larger η induces stronger diffusion variance, whereas λ does not affect the diffusion variance. We also overlay the theoretical fits. An additional experimental results for $\mathbb{Z}_{127}$ are provided in Appendix A.16.

where p is the parameter dimension. In particular, let $G(\theta) := \frac{1}{\sqrt{b}} B(t)\pi(\theta)\Sigma(\theta)^{1/2}$ be the preconditioned diffusion factor, which admits $G(\theta)G(\theta)^\top \approx I_p$. See Appendix A.10.

Under Observations 1–2, Lemma 2 gives the approximate closed-form dynamics of $r_t^2 = \|\theta_t\|_2^2$. The proof is provided in Appendix A.11, where we also experimentally verify that the lemma accurately characterizes the late-stage dynamics of manifold radii in Figure 10.

Lemma 2 (Reduced Late-Stage Radius SDE). *Let $r_t^2 := \|\theta_t\|_2^2$. In Adam's late-stage regime underlying grokking, assume that the diffusion covariance becomes isotropic $G(\theta_t)G(\theta_t)^\top \approx I_p$ as $t \to \infty$, then the reduced late-stage squared-radius SDE admits*

$$\mathrm{d}r_t^2 \approx \left[-\frac{2\lambda}{s(\theta_t)} r_t^2 - \frac{2}{s(\theta_t)} \theta_t^\top \left(\bar{g}_t - \lambda \theta_t \right) + \eta\, p + \mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t) \right] \mathrm{d}t + 2\sqrt{\eta}\, r_t\, \mathrm{d}W_t^{(r)}, \quad (10)$$

where slow-manifold residual $\mathcal{R}_{\mathrm{SM}}(t)$ and preconditioner residual $\mathcal{R}_\pi(t)$ are

$$\mathcal{R}_{\mathrm{SM}}(t) := -2B(t)\theta_t^\top \left[\mathbb{E}\left[m_t \oslash \sqrt{v_t} \right] - \bar{g}_t \left(1 - e^{-\alpha_1 t} \right) \oslash \sqrt{\overline{g \odot g}_t} \right], \quad (11)$$

$$\mathcal{R}_\pi(t) := -2\theta_t^\top \left[\pi(\theta_t) - s(\theta_t)^{-1} I_p \right] \bar{g}_t, \quad (12)$$

respectively, and a one-dimensional Wiener process $W_t^{(r)} := \int_0^t e_r(\theta_u)^\top G(\theta_u) \mathrm{d}W_u$ with $e_r(\theta) := \frac{\theta}{\|\theta\|_2}$ by Lévy's characterization: $\langle W_t^{(r)} \rangle_t = t$. Empirically, in the memorization regime, the residual sum $\mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t)$ becomes negligible. Hence, we set $\mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t) = 0$ in the subsequent analysis, see Appendix A.11.

4.4 Generalization Radius

Let S_{τ_M} denote the state at which the optimization trajectory first reaches ∂M. We treat S_{τ_M} as the initial state for the subsequent *first-hitting problem* (Karatzas & Shreve, 1991) on ∂G. Define the *stopping time* for first hitting ∂G and the corresponding mean squared generalization radius as

$$\tau_G(S_0) := \inf\{ t \geq \tau_M : \theta_t(S_0) \in \partial G \}, \qquad \rho_G^2 := \mathbb{E}_{S_0 \in \Theta_S} \left[\|\theta_{\tau_G(S_0)}\|_2^2 \right]. \quad (13)$$

Once the trajectory enters ∂G, the drift and diffusion in equation (2) equilibrate, and Adam's SDE approaches its equilibrium. The center of this equilibrium starting from S_0 is the regularized local minimizer

$$\theta^\star(S_0) := \arg\min_{\theta(S_0)} \mathscr{L}_f^*(\theta(S_0)) + \frac{\lambda}{2} \|\theta(S_0)\|_2^2, \quad (14)$$

where $\mathscr{L}_f^*(\theta(S_0))$ is task loss. This yields an equilibrium-distribution problem for the reduced late-stage dynamics around $\theta^\star(S_0)$, initialized from the post-memorization state $S_{\tau_M}(S_0)$. Linearizing the reduced dynamics around $\theta^\star(S_0)$ solves Theorem 2. The proof is provided in Appendix A.13. Experimental verification is provided for S_5 in Figure 4 and for $\mathbb{Z}_{127}$ in Appendix A.16.

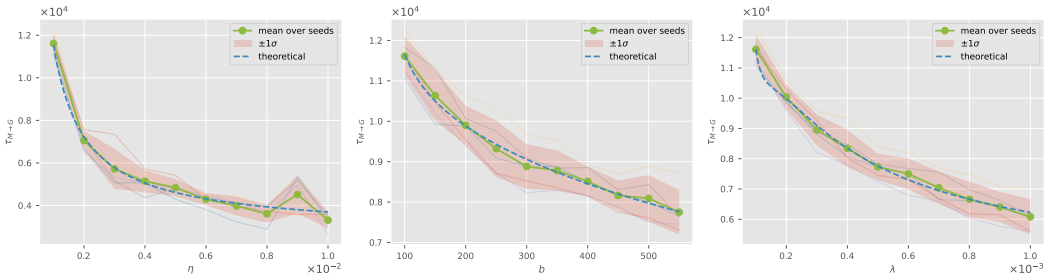


Figure 5: **Scaling laws of solution transition time on S_5.** We show that the solution transition time $\tau_{M\to G}$ from the memorization manifold M to the generalization manifold G scales with the learning rate η, batch size b, and ℓ_2 regularization coefficient λ. For each hyperparameter configuration, we train for ten runs. We also overlay the theoretical fits. An additional experimental results for $\mathbb{Z}_{127}$ are provided in Figure 13.

Theorem 2 (Scaling Law of Generalization Radius)**.** *The mean squared generalization manifold radius admits the asymptotic expansion*

$$\rho_G^2 \;\approx\; \big(\rho_G^{(0)}\big)^2 \;+\; \frac{\eta}{\lambda}\, c_G \;+\; O\Big(\frac{\eta^2}{\lambda^2}\Big), \tag{15}$$

with $(\rho_G^{(0)})^2 := \mathbb{E}_{S_0\in\Theta_S}\left[\|\theta^\star(S_0)\|_2^2\right]$ *the equilibrium radius modulated by* λ *and* $c_G > 0$ *a task-determined constant independent of* η *and* λ*, as defined in Appendix A.13. A weak dependence on* λ *enters* $(\rho_G^{(0)})^2$ *through* $(\rho_G^{(0)})^2 \propto \exp(-O(\lambda))$*; and a weak batch-size dependence enters through* $c_G \propto O(\frac{1}{\sqrt{b}})$.

4.5 Solution Transition Time

Starting from $S_0 \in \Theta_S$, the transition time on the trivial annulus $M \setminus G$ is given as

$$\tau_{M\to G}(S_0) \;:=\; \tau_G(S_0) - \tau_M(S_0), \tag{16}$$

where $\tau_M(S_0)$ and $\tau_G(S_0)$ are the stopping times that the trajectory first hits ∂M and ∂G, respectively, defined in Sections 4.2–4.4. Taking $\theta_{\tau_M}(S_0) \in \partial M$ as the initial condition (Theorem 1), we formulate the transition from ∂M to ∂G as a *mean first-passage problem* (Karatzas & Shreve, 1991) for the radial process $r_t := \|\theta_t\|_2$ on the annulus $M \setminus G$. Its equilibrium distribution characterizes the equilibrium around the generalization region, while the mean first-passage time from the memorization radius to the generalization radius gives the solution transition time. Solving this first-passage problem yields Theorem 3. The proof is provided in Appendix A.14. Experimental verification is provided for S_5 in Figure 5 and for $\mathbb{Z}_{127}$ in Appendix A.17.

Theorem 3 (Scaling Law of Solution Transition Time)**.** *The mean number of optimizer iterations between the first hit at* ∂M *and the first hit at* ∂G *admits the asymptotic expansion*

$$\tau_{M\to G} := \mathbb{E}_{S_0\in\Theta_S}\left[\tau_{M\to G}(S_0)\right] \;\approx\; \frac{\bar{s}}{\eta\,\lambda}\,\log\frac{\rho_M^{(0)}}{\rho_G^{(0)}} \;+\; \frac{c_\tau}{b\,\lambda} \;+\; \frac{c_\tau^{(2)}}{\lambda^2} \;+\; O\Big(\frac{\eta}{\lambda^3}\Big), \tag{17}$$

where $\bar{s}$ *is determined by the harmonic mean of the effective deterministic gradient flow, while* c_τ *and* $c_\tau^{(2)}$ *collect the trajectory corrections, as defined in Appendix A.14. All three are task-determined constants independent of* η *and* λ*; a weak batch-size* b*-dependence enters through the scalar preconditioning scale* $\bar{s} \propto O(\frac{1}{\sqrt{b}})$*, and hence* $c_\tau \propto O(\frac{1}{\sqrt{b}})$ *and* $c_\tau^{(2)} \propto O(\frac{1}{b})$.

5 Validation

We validate the three scaling laws in Section 5.1 by directly measuring ρ_M^2, ρ_G^2, and $\tau_{M\to G}$ on the group-theoretic learning task S_5 and modular arithmetic tasks over $\mathbb{Z}_{127}$, and further compare them with published results from the grokking literature in Section 5.2.

5.1 Experimental Validation

Experimental Setting. We use a two-layer MLP: each input token is embedded into 256 dimensions, the two embeddings are concatenated, passed through a width-512 ReLU hidden layer, and projected to $|\mathcal{Y}|$ output logits, where $|\mathcal{Y}| = 120$ for S_5 and 127 for $\mathbb{Z}_{127}$. The embedding layers are initialized from a standard normal distribution, while the linear layers use *Kaiming* initialization. Training is performed in 32-bit floating-point precision. We find that 16-bit precision can introduce numerical instability in grokking, suggesting that grokking dynamics are sensitive to numerical precision errors. We use base hyperparameters $\eta = 10^{-3}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, ℓ_2 regularization coefficient $\lambda = 10^{-4}$, and batch size 100. These base values are varied as needed to study scaling laws.

Learning Tasks. We adopt two structured learning tasks: a group-theoretic learning task on S_n and a modular arithmetic learning task on $\mathbb{Z}_p$. For S_n, each input is an ordered pair of permutations $(f, g) \in S_n^2$, and the model is trained to predict their group product $f \circ g$ under permutation composition. For $\mathbb{Z}_p$, each input is a pair $(a, b) \in \mathbb{Z}_p^2$, and the model is trained to predict the modular sum $a + b \bmod p$. In both tasks, the dataset consists of all possible input–output pairs, with a randomly sampled subset used for training and the remaining pairs used for evaluation. More details are provided in Appendix A.1.

Results. Figures 2 and 4 validate the predicted scaling laws of the manifold radii ρ_M^2 and ρ_G^2 with respect to the learning rate η, batch size b, and regularization coefficient λ, as stated in Theorems 1 and 2. Figure 5 validates the predicted scaling law of the solution transition time $\tau_{M\to G}$ with respect to the same hyperparameters, as stated in Theorem 3. Additional results on $\mathbb{Z}_{127}$ are provided in Appendices A.15, A.16, and A.17. Overall, the experiments support the theoretical predictions; in particular, ρ_M^2 exhibits the predicted U-shaped dependence on η/b.

Discussions and Limitations. The scaling laws originate from how η, b, and λ modulate Adam's SDE: η and b module the stochastic diffusion by scaling the gradient covariance with a factor $\sqrt{\eta/b}$, whereas λ modulates the deterministic gradient drift flow by regularization. Our results inherit the conditions required by Adam's SDE limit, including an adapted Wiener-process approximation. These conditions require a sufficiently small learning rate and a sufficiently large batch size; thus, our results do not directly apply to the large-η/b regime.

5.2 Literature Validation

Theorem 1 characterizes the scaling of the memorization radius ρ_M with respect to η/b, which, to the best of our knowledge, has not been derived in prior work. Theorem 2 derives the scaling law of the generalization radius ρ_G, showing that it decomposes into the landscape-dependent term $\left(\rho_G^{(0)}\right)^2 = \mathbb{E}_{\theta_0 \in \Theta}\left[\|\theta^\star(\theta_0)\|_2^2\right]$ and an $O(\eta/(\sqrt{b}\lambda))$ fluctuation correction; this is consistent with the weight-norm "Goldilocks zone" reported in prior studies (Liu et al., 2022; Varma et al., 2023). Theorem 3 recovers the inverse dependence of the grokking delay on the regularization coefficient, including the scaling $\tau \propto 1/(\eta\lambda)$ observed or implied in prior empirical and mechanistic studies (Liu et al., 2022). Together, these results show that the proposed stopping-time analysis recovers known scaling behavior while refining it into separate contributions from learning rate, batch size, and regularization coefficient.

6 Conclusions

Optimization dynamics induce a topological configuration of solution manifolds underlying grokking. In this work, we characterize this configuration through a shell–core prior of the reachable solution space, in which training first reaches memorization solutions and then transitions, after a prolonged delay, toward generalization solutions. We derive scaling laws for the manifold radii and the solution transition time with respect to the learning rate, batch size, and regularization coefficient through stopping-time and equilibrium-distribution analyses of Adam's joint-state SDE. These results connect the geometry of solution manifolds with the stochastic dynamics of Adam and provide a theoretical framework for understanding grokking through optimization dynamics.

AI USE STATEMENT

In this work, we used generative AI tools for: polishing writing drafted by the human authors; checking for typographical and grammatical errors; refining wording choices; partially assisting with the development of experimental code; assisting human mathematicians in verifying and normalizing proofs (*e.g.*, standardizing notations); and retrieving related literature. We have not used generative AI tools for drafting the manuscript, fully deriving proofs without the supervision of the human mathematician (the first author), or developing the theoretical model, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. For example, we carefully reviewed LLM-polished sentences; verified LLM-normalized proofs through numerical experiments, symbolic algebra, and independent examination by human mathematicians; and manually verified LLM-aided literature retrieval by consulting the ACM Digital Library or the literature's original publication venues, such as *OpenReview*, and having human experts read the retrieved works. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Carl M. Bender and Steven A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Springer New York, New York, NY, 1 edition, 1999. ISBN 978-0-387-98931-0. doi: 10.1007/978-1-4757-3069-2. Originally published by McGraw Hill, 1978.

Enea Monzio Compagnoni, Tianlin Liu, Rustem Islamov, Frank Norbert Proske, Antonio Orvieto, and Aurelien Lucchi. Adaptive methods through the lens of SDEs: Theoretical insights on the role of noise. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ww3CLRhF1v.

André Belotto Da Silva and Maxime Gazeau. A general system of differential equations to model first-order adaptive algorithms. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

E. B. Dynkin. *Markov Processes: Volume II*. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, Berlin, Heidelberg, 1st edition, 1965. ISBN 978-3-662-23320-7. doi: 10.1007/978-3-662-25360-1.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

Hyeonbin Hwang and Yeachan Park. Intrinsic task symmetry drives generalization in algorithmic tasks. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.01968.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer New York, New York, NY, 2nd edition, 1991. ISBN 978-0-387-97655-6. doi: 10.1007/978-1-4612-0949-2.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vt5mnLVIVo.

Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, pp. 9–50, Berlin, Heidelberg, 1998. Springer-Verlag. ISBN 3540653112.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2101–2110. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/li17f.html.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: an effective theory of representation learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Róisín Luo, James McDermott, Christian Gagné, Qiang Sun, and Colm O'Riordan. Optimization-induced dynamics of Lipschitz continuity in neural networks, 2025. URL https://arxiv.org/abs/2506.18588.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017. URL http://jmlr.org/papers/v18/17-214.html.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Pascal Jr Tikeng Notsawo, Guillaume Dumas, and Guillaume Rabusseau. Grokking finite-dimensional algebra. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.19533.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer Berlin, Heidelberg, Berlin, Heidelberg, 6 edition, 2003. ISBN 978-3-540-04758-2. doi: 10.1007/978-3-642-14394-6.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency, 2023. URL https://arxiv.org/abs/2309.02390.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S5wmbQc1We.

A APPENDIX

This section provides proofs, experimental verification of the technical derivations, and additional experimental results. A high-level sketch of theoretical analysis framework is illustrated as in Figure 6.

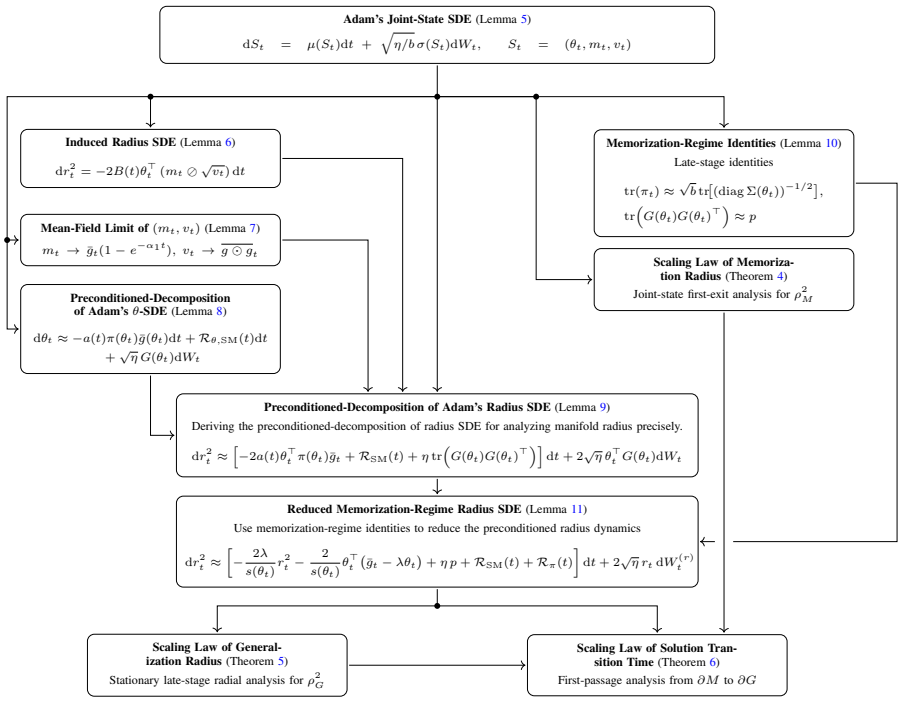


Figure 6: **High-Level Sketch of Theoretical Analysis Framework.**

Table of Appendix Contents:

A.1 Experimental Settings: Learning Tasks 13
A.2 Additional Results: Adam–Induced Shell–Core Radius & Stopping-Time Concentration 14
A.3 Proof: Concentration of Normal Initialization 14
A.4 Proof: Concentration of Uniform Initialization 17
A.5 Proof: Adam's Closed-Form Continuous-Time SDE Limit 19
A.6 Proof: Induced Radius SDE 27
A.7 Proof: Mean-Field Limit of First- and Second-Moment Estimates 28
A.8 Proof: Preconditioned-Decomposition of Adam's θ-SDE 29
A.9 Proof: Preconditioned-Decomposition of Adam's Radius SDE 33
A.10 Proof: Memorization-Regime Preconditioner and Effective Diffusion Identities 35
A.11 Proof: Reduced Memorization-Regime Radius SDE 38
A.12 Proof: Scaling Law of Memorization Radius 40
A.13 Proof: Scaling Law of Generalization Radius 46
A.14 Proof: Scaling Law of Solution Transition Time 50
A.15 Additional Results: Scaling Laws of Manifold Radius ρ_M^2 on $\mathbb{Z}_{127}$ 56
A.16 Additional Results: Scaling Laws of Manifold Radius ρ_G^2 on $\mathbb{Z}_{127}$ 57

A.17 Additional Results: Scaling Laws of Solution Transition Time on $\mathbb{Z}_{127}$ 58

A.1 Experimental Settings: Learning Tasks

Group-theoretic learning task on S_n. Let S_n denote the symmetric group on n elements. Each group element $f \in S_n$ is a permutation sending its index set $[n] := \{1, 2, \ldots, n\}$ to a fixed enumeration of all $n!$ permutations. Let $f[i]$ denote the indexed permutation $i \mapsto j$ from an index $i \in [n]$ to a target $j \in [n]$. For example, given two permutations $f, g \in S_5$:

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 5 & 1 & 4 & 3 \end{pmatrix} \quad g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & 1 & 5 & 2 & 4 \end{pmatrix},$$

the permutations are given by $f[1] : 1 \mapsto 2$, $f[2] : 2 \mapsto 5$, etc. The product of f and g is defined by permutation composition $(f \circ g)[i] = g\big[f[i]\big]$ given by

$$f \circ g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & 4 & 3 & 2 & 5 \end{pmatrix}.$$

The learning task is to predict the group product $f \circ g$ given the pair (f, g). The dataset consists of all ordered pairs $(f, g) \in S_n^2$, with a randomly sampled subset used for training and the remaining pairs used for evaluation.

Modular arithmetic learning task on $\mathbb{Z}_p$. Let $\mathbb{Z}_p$ denote the cyclic group of integers modulo p. Each input consists of a pair $(a, b) \in \mathbb{Z}_p^2$, and the task is to predict their modular sum $a+b \bmod p$. As in the group-theoretic task, training is performed on a randomly selected subset of all p^2 input–output pairs, while evaluation is conducted on the full set.

A.2 Additional Results: Adam–Induced Shell–Core Radius & Stopping-Time Concentration

Table 1: **Adam–Induced Shell–Core Radius & Stopping-Time Concentration on Learning Task S_5.** The experiment shows the concentration of Adam–induced shell–core radii and stopping times on the learning task S_5, across 10 seeds, with learning rate $\eta = 10^{-3}, (\beta_1, \beta_2) = (0.9, 0.999)$, ℓ_2 regularization coefficient 10^{-4}, and batch size 100.

Seed (S_0)	$\rho_0(S_0)$	$\rho_M(S_0)$	$\tau_M(S_0)$	$\rho_G(S_0)$	$\tau_G(S_0)$	$\tau_{M\to G}(S_0)$
1	249.27	186.07	1006	71.39	14000	12994
2	247.09	184.59	1009	69.22	14850	13841
3	247.87	185.19	1007	68.84	12150	11143
4	248.13	185.16	1009	69.67	13950	12941
5	248.85	185.64	1008	70.72	12500	11492
6	248.44	186.61	939	69.34	12100	11161
7	247.98	185.05	1009	69.50	15450	14441
8	248.03	185.19	1009	68.33	13850	12841
9	247.95	186.23	937	67.80	13050	12113
10	248.88	185.59	1004	70.15	12750	11746
$\mathbb{E}[\cdot]$	248.25	185.53	993.7	69.50	13465.0	12471.3
$\mathrm{Std}[\cdot]$	0.60	0.59	27.9	1.02	1083.3	1071.5

Table 2: **Adam–Induced Shell–Core Radius & Stopping-Time Concentration on Learning Task $\mathbb{Z}_{127}$.** The experiment shows the concentration of Adam–induced shell–core radii and stopping times on the learning task S_5, across 10 seeds, with learning rate $\eta = 10^{-3}, (\beta_1, \beta_2) = (0.9, 0.999)$, ℓ_2 regularization coefficient 10^{-4}, and batch size 100.

Seed (S_0)	$\rho_0(S_0)$	$\rho_M(S_0)$	$\tau_M(S_0)$	$\rho_G(S_0)$	$\tau_G(S_0)$	$\tau_{M\to G}(S_0)$
1	256.25	186.57	1041	71.47	8000	6959
2	254.24	183.95	1115	72.40	8600	7485
3	254.76	184.10	1121	71.35	8300	7179
4	255.47	184.79	1118	71.32	8300	7182
5	256.10	185.09	1119	70.61	8400	7281
6	255.76	184.74	1121	71.06	8450	7329
7	255.09	184.36	1116	72.32	8400	7284
8	255.45	184.88	1114	72.21	8500	7386
9	255.46	185.97	1047	72.66	7650	6603
10	256.06	186.53	1042	73.44	7800	6758
$\mathbb{E}[\cdot]$	255.46	185.10	1095.4	71.88	8240.0	7144.6
$\mathrm{Std}[\cdot]$	0.60	0.90	34.2	0.82	299.8	269.4

A.3 Proof: Concentration of Normal Initialization

Lemma 3 (Gaussian Initialization Concentration)**.** *Let*

$$\theta_0 = \left(\theta_0^{(1)}, \dots, \theta_0^{(k)}\right) \in \mathbb{R}^p, \qquad \theta_0^{(j)} \in \mathbb{R}^{p_j}, \qquad \sum_{j=1}^{k} p_j = p, \tag{18}$$

where the subvectors are independent and $[\theta_0^{(j)}]_i \sim \mathcal{N}(0, \sigma_j^2)$. Define the effective coordinate variance $\tilde{\sigma}^2$ and the effective dimension parameter $\tilde{p}$ by

$$\tilde{\sigma}^2 := \frac{1}{p}\sum_{j=1}^{k} p_j \sigma_j^2, \qquad \frac{1}{\tilde{p}} := \frac{1}{k}\sum_{j=1}^{k} \frac{1}{p_j}. \tag{19}$$

Then the squared initialization radius satisfies

$$\mathbb{E}\big[\|\theta_0\|_2^2\big] = \tilde{\sigma}^2 p, \qquad \mathrm{Var}\big[\|\theta_0\|_2^2\big] = 2\sum_{j=1}^{k} p_j \sigma_j^4. \tag{20}$$

In particular, for initialization schemes used for numerical stability, each block-wise coordinate scale is often chosen so that $\sigma_j = O(p_j^{-1/2})$. Under this scaling,

$$\tilde{\sigma}^2 = O\left(\frac{k}{p}\right), \qquad \mathbb{E}\big[\|\theta_0\|_2^2\big] = O(k), \qquad \mathrm{Var}\big[\|\theta_0\|_2^2\big] = O\left(\frac{k}{\tilde{p}}\right). \tag{21}$$

Proof. We consider a Gaussian initialization with block-wise coordinate scales. Write

$$\theta_0 = \left(\theta_0^{(1)}, \ldots, \theta_0^{(k)}\right), \qquad \theta_0^{(j)} = \sigma_j Z^{(j)}, \qquad Z^{(j)} \sim \mathcal{N}(0, I_{p_j}), \tag{22}$$

where the subvectors $Z^{(j)}$ are independent. Then

$$\|\theta_0\|_2^2 = \sum_{j=1}^{k} \|\theta_0^{(j)}\|_2^2 = \sum_{j=1}^{k} \sigma_j^2 \|Z^{(j)}\|_2^2, \tag{23}$$

and it suffices to study the distribution of each $\|Z^{(j)}\|_2^2$. Let

$$R_j := \|Z^{(j)}\|_2^2. \tag{24}$$

The random variable R_j follows a chi-square distribution with p_j degrees of freedom, whose density is

$$f_{R_j}(r) = \frac{1}{2^{\frac{p_j}{2}}\Gamma\left(\frac{p_j}{2}\right)} r^{\frac{p_j}{2}-1} e^{-r/2}, \qquad r \geq 0. \tag{25}$$

Computing $\mathbb{E}\big[\|\theta_0\|_2^2\big]$. The expected squared Euclidean norm of the initialization is therefore

$$\mathbb{E}\big[\|\theta_0\|_2^2\big] = \sum_{j=1}^{k} \sigma_j^2 \, \mathbb{E}[R_j] \tag{26}$$

$$= \sum_{j=1}^{k} \sigma_j^2 \int_0^\infty r f_{R_j}(r) \, \mathrm{d}r \tag{27}$$

$$= \sum_{j=1}^{k} \frac{\sigma_j^2}{2^{\frac{p_j}{2}}\Gamma\left(\frac{p_j}{2}\right)} \int_0^\infty r^{\frac{p_j}{2}} e^{-r/2} \, \mathrm{d}r. \tag{28}$$

Applying the change of variables $u = r/2$ yields

$$\int_0^\infty r^{\frac{p_j}{2}} e^{-r/2} \, \mathrm{d}r = 2^{\frac{p_j}{2}+1}\Gamma\left(\frac{p_j}{2}+1\right). \tag{29}$$

Substituting back, we obtain

$$\mathbb{E}\big[\|\theta_0\|_2^2\big] = \sum_{j=1}^{k} 2\sigma_j^2 \frac{\Gamma\left(\frac{p_j}{2}+1\right)}{\Gamma\left(\frac{p_j}{2}\right)} \tag{30}$$

$$= \sum_{j=1}^{k} p_j \sigma_j^2 \tag{31}$$

$$= \tilde{\sigma}^2 p. \tag{32}$$

Computing $\mathrm{Var}\big[\|\theta_0\|_2^2\big]$. We next compute the fluctuations of the squared initialization radius. Since the subvectors are independent and $\|\theta_0\|_2^2 = \sum_{j=1}^{k} \sigma_j^2 R_j$, it follows that

$$\mathrm{Var}\big[\|\theta_0\|_2^2\big] = \sum_{j=1}^{k} \sigma_j^4 \mathrm{Var}[R_j]. \tag{33}$$

To compute $\mathrm{Var}[R_j]$, we first compute the second moment:

$$\mathbb{E}[R_j^2] = \int_0^\infty r^2 f_{R_j}(r)\,\mathrm{d}r \tag{34}$$

$$= \frac{1}{2^{\frac{p_j}{2}}\Gamma\left(\frac{p_j}{2}\right)} \int_0^\infty r^{\frac{p_j}{2}+1} e^{-r/2}\,\mathrm{d}r. \tag{35}$$

Applying again the change of variables $u = r/2$ gives

$$\int_0^\infty r^{\frac{p_j}{2}+1} e^{-r/2}\,\mathrm{d}r = 2^{\frac{p_j}{2}+2}\Gamma\left(\frac{p_j}{2}+2\right). \tag{36}$$

Thus,

$$\mathbb{E}[R_j^2] = 4\frac{\Gamma\left(\frac{p_j}{2}+2\right)}{\Gamma\left(\frac{p_j}{2}\right)} \tag{37}$$

$$= 4\left(\frac{p_j}{2}+1\right)\left(\frac{p_j}{2}\right) \tag{38}$$

$$= p_j(p_j+2). \tag{39}$$

Therefore,

$$\mathrm{Var}[R_j] = \mathbb{E}[R_j^2] - \left(\mathbb{E}[R_j]\right)^2 \tag{40}$$

$$= p_j(p_j+2) - p_j^2 \tag{41}$$

$$= 2p_j. \tag{42}$$

Consequently,

$$\mathrm{Var}\left[\|\theta_0\|_2^2\right] = \sum_{j=1}^k \sigma_j^4 \mathrm{Var}[R_j] \tag{43}$$

$$= 2\sum_{j=1}^k p_j \sigma_j^4. \tag{44}$$

Coordinate Scaling. Finally, in common initialization schemes used for numerical stability, each block-wise per-coordinate variance is scaled inversely with its block dimension. Equivalently, one may write

$$\sigma_j = O\left(\frac{1}{\sqrt{p_j}}\right), \qquad \sigma_j^2 = O\left(\frac{1}{p_j}\right). \tag{45}$$

Substituting this scaling into the definition of the effective coordinate variance gives

$$\tilde{\sigma}^2 = \frac{1}{p}\sum_{j=1}^k p_j \sigma_j^2 = O\left(\frac{k}{p}\right). \tag{46}$$

Therefore,

$$\mathbb{E}\left[\|\theta_0\|_2^2\right] = \sum_{j=1}^k p_j \sigma_j^2 = \tilde{\sigma}^2 p = O(k), \tag{47}$$

$$\mathrm{Var}\left[\|\theta_0\|_2^2\right] = 2\sum_{j=1}^k p_j \sigma_j^4 = O\left(\sum_{j=1}^k \frac{1}{p_j}\right). \tag{48}$$

Using the definition

$$\frac{1}{\tilde{p}} = \frac{1}{k}\sum_{j=1}^k \frac{1}{p_j}, \tag{49}$$

we obtain

$$\mathrm{Var}\left[\|\theta_0\|_2^2\right] = O\left(\frac{k}{\tilde{p}}\right). \tag{50}$$

□

A.4 Proof: Concentration of Uniform Initialization

Lemma 4 (Uniform Initialization Concentration)**.** *Let*

$$\theta_0 = \left(\theta_0^{(1)}, \dots, \theta_0^{(k)}\right) \in \mathbb{R}^p, \qquad \theta_0^{(j)} \in \mathbb{R}^{p_j}, \qquad \sum_{j=1}^{k} p_j = p, \tag{51}$$

where the subvectors are independent and the coordinates of each subvector satisfy $[\theta_0^{(j)}]_i \sim \mathcal{U}(-\varepsilon_j, \varepsilon_j)$*. Define the effective coordinate variance* $\tilde{\varepsilon}^2$ *and the effective dimension parameter* $\tilde{p}$ *by*

$$\tilde{\varepsilon}^2 := \frac{1}{p}\sum_{j=1}^{k} p_j \varepsilon_j^2, \qquad \frac{1}{\tilde{p}} := \frac{1}{k}\sum_{j=1}^{k} \frac{1}{p_j}. \tag{52}$$

Then the squared initialization radius satisfies

$$\mathbb{E}[\|\theta_0\|_2^2] = \frac{\tilde{\varepsilon}^2 p}{3}, \qquad \mathrm{Var}(\|\theta_0\|_2^2) = \frac{4}{45}\sum_{j=1}^{k} p_j \varepsilon_j^4. \tag{53}$$

Moreover, by the central limit theorem,

$$\|\theta_0\|_2^2 \overset{d}{\approx} \mathcal{N}\left(\frac{\tilde{\varepsilon}^2 p}{3}, \frac{4}{45}\sum_{j=1}^{k} p_j \varepsilon_j^4\right), \qquad p \to \infty. \tag{54}$$

In particular, for initialization schemes with block-wise coordinate half-widths $\varepsilon_j = O(p_j^{-1/2})$*, the squared initialization radius satisfies*

$$\tilde{\varepsilon}^2 = O\left(\frac{k}{p}\right), \qquad \mathbb{E}[\|\theta_0\|_2^2] = O(k), \qquad \mathrm{Var}(\|\theta_0\|_2^2) = O\left(\frac{k}{\tilde{p}}\right). \tag{55}$$

For example, a common numerically stable block-wise choice is $\varepsilon_j = p_j^{-1/2}$*, in which case*

$$\|\theta_0\|_2^2 \overset{d}{\approx} \mathcal{N}\left(\frac{k}{3}, \frac{4k}{45\tilde{p}}\right). \tag{56}$$

Thus uniform initialization concentrates on a thin hyperspherical shell with squared radius $k/3$ *and squared-radius variance* $4k/(45\tilde{p})$*.*

Proof. For each coordinate $X_i^{(j)} := [\theta_0^{(j)}]_i \sim \mathcal{U}(-\varepsilon_j, \varepsilon_j)$, the density of $X_i^{(j)}$ is $f_{X_i^{(j)}}(x) = \frac{1}{2\varepsilon_j}$ for $x \in [-\varepsilon_j, \varepsilon_j]$. Then the second moment is computed as

$$\mathbb{E}[(X_i^{(j)})^2] = \int_{-\varepsilon_j}^{\varepsilon_j} x^2 f_{X_i^{(j)}}(x)\,\mathrm{d}x = \frac{1}{2\varepsilon_j}\int_{-\varepsilon_j}^{\varepsilon_j} x^2\,\mathrm{d}x = \frac{\varepsilon_j^2}{3}. \tag{57}$$

The fourth moment is computed as

$$\mathbb{E}[(X_i^{(j)})^4] = \int_{-\varepsilon_j}^{\varepsilon_j} x^4 f_{X_i^{(j)}}(x)\,\mathrm{d}x = \frac{1}{2\varepsilon_j}\int_{-\varepsilon_j}^{\varepsilon_j} x^4\,\mathrm{d}x = \frac{\varepsilon_j^4}{5}. \tag{58}$$

Defining $S_i^{(j)} := (X_i^{(j)})^2$, we have $\mathbb{E}[S_i^{(j)}] = \mathbb{E}[(X_i^{(j)})^2]$ and $\mathbb{E}[(S_i^{(j)})^2] = \mathbb{E}[(X_i^{(j)})^4]$, so that

$$\mathrm{Var}((X_i^{(j)})^2) = \mathrm{Var}(S_i^{(j)}) = \mathbb{E}[(S_i^{(j)})^2] - \mathbb{E}[S_i^{(j)}]^2 \tag{59}$$

$$= \mathbb{E}[(X_i^{(j)})^4] - \left(\mathbb{E}[(X_i^{(j)})^2]\right)^2 \tag{60}$$

$$= \frac{\varepsilon_j^4}{5} - \left(\frac{\varepsilon_j^2}{3}\right)^2 \tag{61}$$

$$= \frac{4\varepsilon_j^4}{45}. \tag{62}$$

Now set

$$R^2 = \sum_{j=1}^{k} \sum_{i=1}^{p_j} (X_i^{(j)})^2 = \sum_{j=1}^{k} \sum_{i=1}^{p_j} S_i^{(j)}. \tag{63}$$

Since the random variables $S_i^{(j)}$ are independent, we obtain

$$\mathbb{E}[R^2] = \sum_{j=1}^{k} p_j \, \mathbb{E}[S_i^{(j)}] = \frac{1}{3} \sum_{j=1}^{k} p_j \varepsilon_j^2 = \frac{\tilde{\varepsilon}^2 p}{3}, \tag{64}$$

$$\mathrm{Var}(R^2) = \sum_{j=1}^{k} p_j \, \mathrm{Var}(S_i^{(j)}) = \frac{4}{45} \sum_{j=1}^{k} p_j \varepsilon_j^4. \tag{65}$$

Moreover, by the central limit theorem,

$$R^2 \overset{d}{\approx} \mathcal{N}\left(\sum_{j=1}^{k} p_j \, \mathbb{E}[S_i^{(j)}], \sum_{j=1}^{k} p_j \, \mathrm{Var}(S_i^{(j)}) \right) = \mathcal{N}\left(\frac{\tilde{\varepsilon}^2 p}{3}, \frac{4}{45} \sum_{j=1}^{k} p_j \varepsilon_j^4 \right), \qquad p \to \infty. \tag{66}$$

For numerically stable uniform initialization, the coordinate half-width of each block is often chosen to scale as

$$\varepsilon_j = O\left(\frac{1}{\sqrt{p_j}} \right). \tag{67}$$

Substituting this scaling into the preceding identities gives

$$\tilde{\varepsilon}^2 = \frac{1}{p} \sum_{j=1}^{k} p_j \varepsilon_j^2 = O\left(\frac{k}{p} \right), \tag{68}$$

and therefore

$$\mathbb{E}[\|\theta_0\|_2^2] = \frac{\tilde{\varepsilon}^2 p}{3} = O(k), \qquad \mathrm{Var}(\|\theta_0\|_2^2) = \frac{4}{45} \sum_{j=1}^{k} p_j \varepsilon_j^4 = O\left(\sum_{j=1}^{k} \frac{1}{p_j} \right). \tag{69}$$

Using the definition

$$\frac{1}{\tilde{p}} = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{p_j}, \tag{70}$$

we obtain

$$\mathrm{Var}(\|\theta_0\|_2^2) = O\left(\frac{k}{\tilde{p}} \right). \tag{71}$$

For example, taking the common block-wise choice

$$\varepsilon_j = \frac{1}{\sqrt{p_j}}, \tag{72}$$

we obtain

$$R^2 \overset{d}{\approx} \mathcal{N}\left(\frac{k}{3}, \frac{4k}{45\tilde{p}} \right). \tag{73}$$

□

A.5 Proof: Adam's Closed-Form Continuous-Time SDE Limit

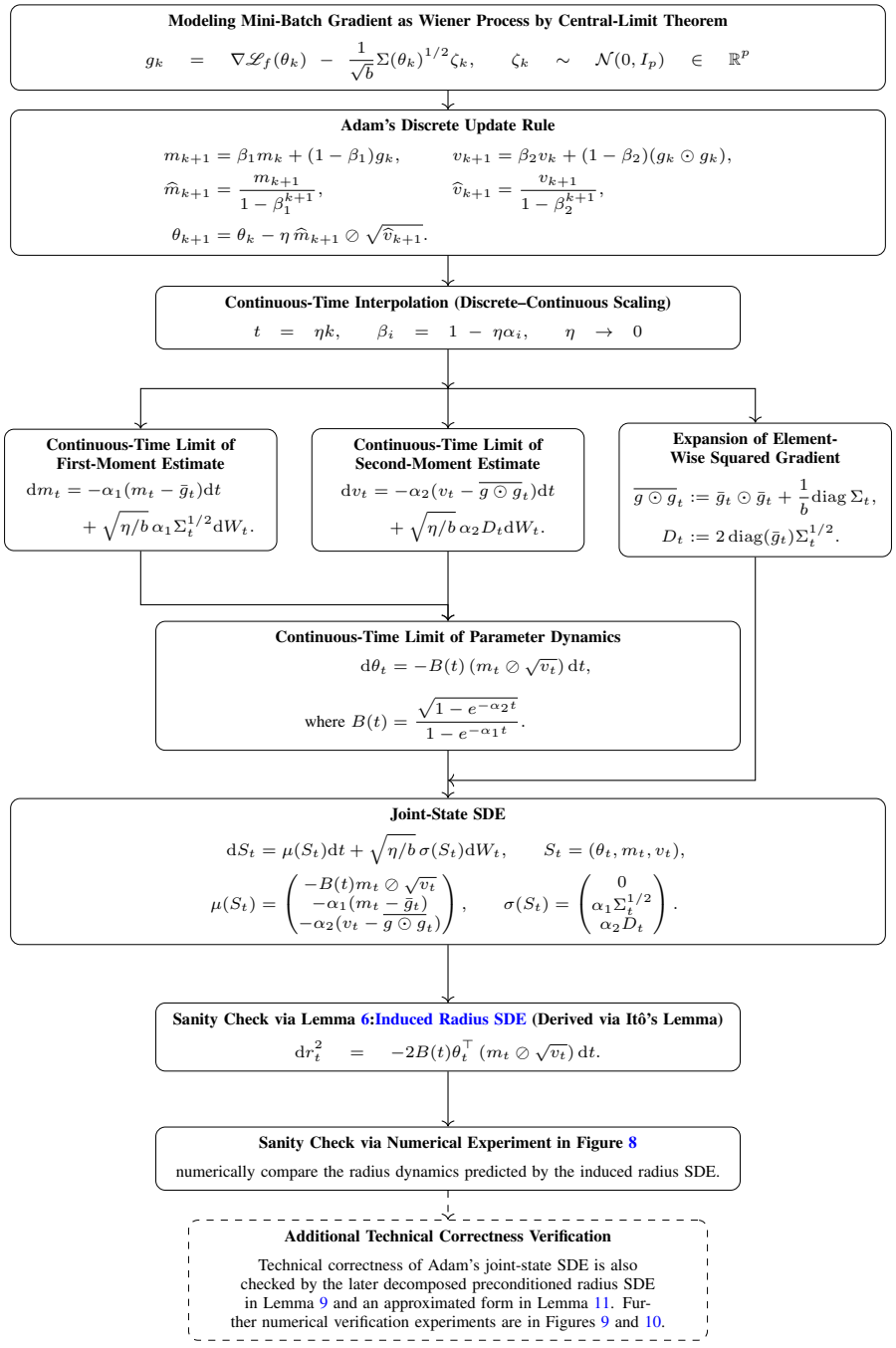


Figure 7: **Proof Sketch for Adam's Closed-Form Continuous-Time SDE Limit.** This diagram illustrates the proof sketch and the corresponding correctness checks for Adam's continuous-time SDE limit. The mini-batch gradient is modeled as a Wiener process by **Central-Limit Theorem**. Combining this stochastic gradient model with Adam's discrete update rules and the continuous-time interpolation yields the continuous-time limits of the first- and second-moment estimates, as well as the continuous-time parameter dynamics. Together with the expansion of the element-wise squared gradient, these components give the joint-state SDE for $S_t := (m_t, v_t, \theta_t)$.

Lemma 5 (Adam's Joint-State Continuous-Time SDE Limit)**.** *For a network f parameterized with $\theta \in \mathbb{R}^p$, let $\ell_f(s;\theta)$ be the loss for sample s, let $\mathscr{L}_f(\xi;\theta)$ be the mini-batch loss for batch ξ, and*

let $\mathscr{L}_f(\theta)$ be the population loss over true dataset distribution. Let θ_k, θ_t denote the parameter at iteration k and time t, respectively.

Discrete-Time Update Rules. *Consider Adam with small learning rate η, large batch size b, and coefficients (β_1, β_2). Let*

$$S_k := (\theta_k^\top, m_k^\top, v_k^\top)^\top \in \mathbb{R}^{3p}, \quad m_0 = 0_p, \quad v_0 = 0_p, \tag{74}$$

denote the discrete-time representation of the parameter, first-moment, and second-moment estimates at iteration k. For a batch ξ_k at step k, the discrete Adam update rules with the bias-correction step are:

$$g_k = \nabla \mathscr{L}_f(\xi_k; \theta_k) \qquad (\textit{mini-batch gradient}) \tag{75}$$

$$m_{k+1} = \beta_1 m_k + (1-\beta_1) g_k \qquad (\textit{first-moment estimate}) \tag{76}$$

$$v_{k+1} = \beta_2 v_k + (1-\beta_2)(g_k \odot g_k) \qquad (\textit{second-moment estimate}) \tag{77}$$

$$\hat{m}_{k+1} = \frac{m_{k+1}}{1-\beta_1^{k+1}} \qquad (\textit{bias-corrected first-moment estimate}) \tag{78}$$

$$\hat{v}_{k+1} = \frac{v_{k+1}}{1-\beta_2^{k+1}} \qquad (\textit{bias-corrected second-moment estimate}) \tag{79}$$

$$\theta_{k+1} = \theta_k - \eta\, h(\hat{m}_{k+1}, \hat{v}_{k+1}) \qquad (\textit{bias-corrected parameter update}), \tag{80}$$

where $h(\hat{m}_{k+1}, \hat{v}_{k+1})$ is the per-coordinate Adam update direction, defined as

$$h(\hat{m}, \hat{v}) \;=\; \hat{m} \oslash (\sqrt{\hat{v}} + \epsilon \mathbf{1}_p) \;=\; \begin{bmatrix} \dfrac{\hat{m}^{(1)}}{\sqrt{\hat{v}^{(1)}} + \epsilon} \\ \vdots \\ \dfrac{\hat{m}^{(p)}}{\sqrt{\hat{v}^{(p)}} + \epsilon} \end{bmatrix}, \tag{81}$$

$\oslash$ denotes element-wise division, $\odot$ denotes element-wise product, and ε is a small positive constant for numerical stability.

Distributional Limit of Mini-Batch Gradient. *For the parameter θ_k, the mini-batch gradient $g_k := \nabla \mathscr{L}_f(\xi_k; \theta_k)$ is an empirical average of b i.i.d. per-sample gradient $g_k(s) := \nabla \ell_f(s; \theta_k)$. Hence, by the central limit theorem, as batch size $b \to \infty$*

$$g_k = \frac{1}{b}\sum_{i=1}^{b} g_k(s_i) \xrightarrow{d} \mathcal{N}\left(\bar{g}_k, \frac{1}{b}\Sigma_k\right), \quad \bar{g}_k := \mathbb{E}_s[g_k(s)], \quad \Sigma_k := \mathrm{Cov}_s[g_k(s)], \tag{82}$$

where $\bar{g}_k$ and Σ_k are the per-sample gradient expectation and covariance at step k, respectively.

Continuous-Time SDE Limit. *Under the continuous-time interpolation $t = \eta k$, with the conditions of sufficiently small learning rate $\eta \to 0$ and large batch size $b \to \infty$, the discrete Adam updates admit the Itô SDE limit*

$$\mathrm{d}S_t = \mu(S_t)\mathrm{d}t + \sqrt{\frac{\eta}{b}}\,\sigma(S_t)\mathrm{d}W_t, \tag{83}$$

where $W_t \in \mathbb{R}^p$ is a Wiener process adapted to the filtration generated by the mini-batch sampling process $\{\xi_t\}$. The drift and diffusion factors are

$$\mu(S_t) = \begin{pmatrix} -B(t)\, m_t \oslash \sqrt{v_t} \\ -\alpha_1\big(m_t - \bar{g}_t\big) \\ -\alpha_2\big(v_t - \overline{g \odot g}_t\big) \end{pmatrix}, \qquad \sigma(S_t) = \begin{pmatrix} 0 \\ \alpha_1 \Sigma_t^{1/2} \\ \alpha_2 D_t \end{pmatrix}, \tag{84}$$

with

$$B(t) := \frac{\sqrt{1-e^{-\alpha_2 t}}}{1-e^{-\alpha_1 t}}, \qquad \alpha_i := \frac{1-\beta_i}{\eta}, \qquad i = 1, 2, \tag{85}$$

and

$$\overline{g \odot g}_t := \bar{g}_t \odot \bar{g}_t + \frac{1}{b}\,\mathrm{diag}(\Sigma_t), \qquad D_t := 2\,\mathrm{diag}(\bar{g}_t)\Sigma_t^{1/2}. \tag{86}$$

Proof. We derive the continuous-time limit of the Adam optimizer starting from its discrete update rules. For a network f parameterized with $\theta \in \mathbb{R}^p$, let $\ell_f(s;\theta)$ be the loss for sample s, let $\mathscr{L}_f(\xi;\theta)$ be the mini-batch loss for batch ξ, and let $\mathscr{L}_f(\theta)$ be the population loss over true dataset distribution. Let θ_k, θ_t denote the parameter at iteration k and time t, respectively.

Mini-Batch Gradient as Wiener Process. Let $g_k(s) := \nabla \ell_f(s;\theta_k)$ be the per-sample gradient for sample s with the mean and covariance

$$\bar{g}_k = \mathbb{E}_s[g_k(s)], \qquad \Sigma_k = \mathrm{Cov}_s\big[g_k(s)\big], \tag{87}$$

respectively. For a batch $\xi_k := \{s_i\}_{i=1}^b$ of size b, the gradient at time k is given by:

$$g_k = \frac{1}{b}\sum_{i=1}^{b} g_k(s_i), \tag{88}$$

by central limit theorem, so that g_k distributionally converge to

$$g_k \xrightarrow{d} \mathcal{N}\left(\bar{g}_k, \frac{1}{b}\Sigma_k\right), \tag{89}$$

as batch size is sufficiently large $b \to \infty$. We therefore can model mini-batch gradient g_k as a Wiener process

$$g_k = \bar{g}_k - \frac{1}{\sqrt{b}}\Sigma_k^{1/2}\zeta_k, \qquad (\Sigma_k^{1/2})(\Sigma_k^{1/2})^\top = \Sigma_k, \tag{90}$$

where $\zeta_k \sim \mathcal{N}(0, I_p)$ is Wiener process adapted to the filtration generated from mini-batch sampling sequence $\{\xi_k\}$. Because gradient descent uses negative gradient, we take a negative sign on ζ_k for simplifying the later analysis algebraically.

Discrete Adam with Bias Correction. Let $\eta > 0$ be sufficiently small learning rate, $\beta_1, \beta_2 \in (0,1)$ be exponential decay rates, and $\epsilon > 0$ be numerical stability constant. Let $\theta_t \in \mathbb{R}^p$, m_t, v_t be parameter, first-moment, and second-moment estimates, with the initialization $m_0 = 0$ and $v_0 = 0$. Let $k = 0, 1, 2, \ldots$ be discrete iteration time points. The discrete Adam update rules with the bias-correction step written explicitly are:

$$\bar{g}_k := \nabla[\mathscr{L}_f(\theta_k)] = \nabla[\mathscr{L}_f^*(\theta_k) + \frac{\lambda}{2}\|\theta_k\|_2^2] \qquad \text{(coupled } \ell_2\text{-regularizer)} \tag{91}$$

$$g_k = \bar{g}_k - \frac{1}{\sqrt{b}}\Sigma_k^{1/2}\zeta_k \qquad \text{(mini-batch gradient)} \tag{92}$$

$$m_{k+1} = \beta_1 m_k + (1-\beta_1) g_k \qquad \text{(first-moment estimate)} \tag{93}$$

$$v_{k+1} = \beta_2 v_k + (1-\beta_2)(g_k \odot g_k) \qquad \text{(second-moment estimate)} \tag{94}$$

$$\hat{m}_{k+1} = \frac{m_{k+1}}{1-\beta_1^{k+1}} \qquad \text{(bias-corrected first-moment estimate)} \tag{95}$$

$$\hat{v}_{k+1} = \frac{v_{k+1}}{1-\beta_2^{k+1}} \qquad \text{(bias-corrected second-moment estimate)} \tag{96}$$

$$\theta_{k+1} = \theta_k - \eta\, h(\hat{m}_{k+1}, \hat{v}_{k+1}) \qquad \text{(bias-corrected parameter update)}, \tag{97}$$

where λ is the regularization coefficient, and $\mathscr{L}_f(\theta_k)$ is the population coupled loss and $\mathscr{L}_f^*(\theta_k)$ is the population task loss, then the per-coordinate Adam update direction is

$$h(\hat{m}, \hat{v}) \;=\; \hat{m} \oslash (\sqrt{\hat{v}} + \epsilon \mathbf{1}_p) \;=\; \begin{bmatrix} \dfrac{\hat{m}^{(1)}}{\sqrt{\hat{v}^{(1)}} + \epsilon} \\ \vdots \\ \dfrac{\hat{m}^{(p)}}{\sqrt{\hat{v}^{(p)}} + \epsilon} \end{bmatrix}, \tag{98}$$

$\oslash$ denotes element-wise division, and ε is a small positive constant for numerical stability.

Remark 3. Throughout this paper, we use the convention that the ℓ_2 regularizer is coupled through the loss function. Thus, the gradients induced by the regularizer enter the computation of the optimizer's first- and second-moment estimates. When the ℓ_2 penalty is instead decoupled and applied as a standalone update term,

$$\theta_{k+1} = \theta_k - \eta\, h(\hat{m}_{k+1}, \hat{v}_{k+1}) - \eta\, \lambda\, \theta_k, \tag{99}$$

where the resulting optimizer is referred to as *AdamW*.

Continuous–Time Interpolation. Taking continuous-time interpolation in discrete iteration k by:

$$t = k\eta, \tag{100}$$

therefore:

$$\mathrm{d}t = \lim_{\Delta t \to 0} \Delta t = \lim_{\Delta t \to 0} \eta = \eta. \tag{101}$$

Wiener Increment in Gradient. Let $g_k(s) := \nabla \ell_f(s; \theta_k)$ be the per-sample gradient for sample s with the mean and covariance

$$\bar{g}_k = \mathbb{E}_s[g_k(s)], \qquad \Sigma_k = \mathrm{Cov}_s\big[g_k(s)\big], \tag{102}$$

respectively. For a batch $\xi_k := \{s_i\}_{i=1}^b$ of size b, the gradient at time k is given by:

$$g_k = \frac{1}{b} \sum_{i=1}^{b} g_k(s_i), \tag{103}$$

by central limit theorem, so that g_k distributionally converge to

$$g_k \xrightarrow{d} \mathcal{N}\left(\bar{g}_k, \frac{1}{b}\Sigma_k\right), \tag{104}$$

as batch size is sufficiently large $b \to \infty$. We therefore can model mini-batch gradient g_k as a Wiener process

$$g_k = \bar{g}_k - \frac{1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k, \qquad (\Sigma_k^{1/2})(\Sigma_k^{1/2})^\top = \Sigma_k, \tag{105}$$

where $\zeta_k \sim \mathcal{N}(0, I_p)$ is Wiener process adapted to the filtration generated from mini-batch sampling sequence $\{\xi_k\}$. Because gradient descent uses negative gradient, we take a negative sign on ζ_k for simplifying the later analysis algebraically.

For the adapted noise vector $\zeta_k \sim \mathcal{N}(0, I_p) \in \mathbb{R}^p$ at iteration k (time $t = k\eta$) corresponds to the Wiener increment over $[t, t+\eta]$:

$$\zeta_k = \frac{1}{\sqrt{\eta}} \zeta_k \sqrt{\eta} = \frac{1}{\sqrt{\eta}} \zeta_k \sqrt{\Delta t} = \frac{1}{\sqrt{\eta}} \Delta W_k, \tag{106}$$

where $\Delta t = \eta$ and the increment follows the property of Brownian motion:

$$\Delta W_k = \zeta_k \sqrt{\Delta t} \sim \mathcal{N}(0, \Delta t\ I_p). \tag{107}$$

Remark 4 (Sanity-Check Reference: Continuous SDE for SGD). Therefore, for SGD, the continuous-time SDE limit is, with $t = k\eta$ and $\Delta t = \eta$:

$$\begin{aligned}
\mathrm{d}\theta_t &= \lim_{\Delta t \to 0} \big(\theta_{t+\Delta t} - \theta_t\big) && (108)\\
&= \lim_{\Delta t \to 0} \left[-\eta \Big(\bar{g}_k - \frac{1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k \Big) \right] && (109)\\
&= \lim_{\Delta t \to 0} \left[-\eta \bar{g}_k + \sqrt{\frac{\eta}{b}} \Sigma_k^{1/2} \sqrt{\eta} \zeta_k \right] && (110)\\
&= \lim_{\Delta t \to 0} \left[-\eta \bar{g}_k + \sqrt{\frac{\eta}{b}} \Sigma_k^{1/2} \Delta W_k \right] && (111)\\
&= -\bar{g}_t \mathrm{d}t + \sqrt{\frac{\eta}{b}} \Sigma_t^{1/2} \mathrm{d}W_t. && (112)
\end{aligned}$$

Discrete Increment of First-Moment Estimate. From equation (93):

$$m_{k+1} = \beta_1 m_k + (1-\beta_1) g_k \tag{113}$$
$$m_{k+1} - m_k = \beta_1 m_k + (1-\beta_1) g_k - m_k \tag{114}$$
$$= (\beta_1 - 1) m_k + (1-\beta_1) g_k \tag{115}$$
$$= -(1-\beta_1) m_k + (1-\beta_1) g_k \tag{116}$$
$$= (1-\beta_1)(g_k - m_k) \tag{117}$$
$$= -(1-\beta_1)(m_k - g_k) \tag{118}$$

Substituting equation (105):

$$g_k = \bar{g}_k - \frac{1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k \tag{119}$$

into equation (118) yields:

$$m_{k+1} - m_k = -(1-\beta_1) \left[m_k - \bar{g}_k + \frac{1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k \right] \tag{120}$$
$$= -(1-\beta_1)[m_k - \bar{g}_k] - (1-\beta_1) \frac{1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k \tag{121}$$
$$= -(1-\beta_1)[m_k - \bar{g}_k] - \frac{1-\beta_1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k. \tag{122}$$

Continuous-Time Limit of First-Moment Estimate. Use $\Delta t = \eta$ and $\Delta W_k = \sqrt{\Delta} \zeta_k$,

$$\mathrm{d}m = \lim_{\Delta t \to 0} \Delta m \tag{123}$$
$$= \lim_{\Delta t \to 0} (m_{k+1} - m_k) \tag{124}$$
$$= \lim_{\Delta t \to 0} \left[-(1-\beta_1)[m_k - \bar{g}_k] - \frac{1-\beta_1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k \right] \tag{125}$$
$$= \lim_{\Delta t \to 0} \left[-(1-\beta_1)[m_k - \bar{g}_k] \frac{1}{\eta} \Delta t - \frac{1-\beta_1}{\sqrt{b}} \Sigma_k^{1/2} \frac{1}{\sqrt{\eta}} \sqrt{\Delta t} \zeta_k \right] \tag{126}$$
$$= -\frac{1-\beta_1}{\eta} [m_t - \bar{g}_k] \mathrm{d}t - \frac{1-\beta_1}{\sqrt{b\eta}} \Sigma_k^{1/2} \mathrm{d}W_t. \tag{127}$$

Expansion of Element-Wise Squared Gradient. To derive the continuous-time limit of the second-moment estimate $\mathrm{d}v$. We need to expand $g_k \odot g_k$ of equation (96) into a tractable form. Consider:

$$g_k = \bar{g}_k - \frac{1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k, \tag{128}$$

we expand the element-wise quadratic product $g_k \odot g_k$ by:

$$g_k \odot g_k = \left[\bar{g}_k - \frac{1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k \right] \odot \left[\bar{g}_k - \frac{1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k \right] \tag{129}$$
$$= \bar{g}_k \odot \bar{g}_k - 2 \left(\bar{g}_k \odot \frac{1}{\sqrt{b}} \Sigma_k^{1/2} \zeta_k \right) + \frac{1}{b} \left(\Sigma_k^{1/2} \zeta_k \right) \odot \left(\Sigma_k^{1/2} \zeta_k \right) \tag{130}$$
$$= [\bar{g}_k]^{\odot 2} - \frac{2}{\sqrt{b}} \bar{g}_k \odot \left[\Sigma_k^{1/2} \zeta_k \right] + \frac{1}{b} (\Sigma_k^{1/2} \zeta_k)^{\odot 2}. \tag{131}$$

Using the fundamental identity, for two vectors u and v:

$$u \odot v = \mathrm{diag}(u) v, \tag{132}$$

then the Hadamard cross product can be written as:

$$\bar{g}_k \odot \left[\Sigma_k^{1/2} \zeta_k \right] = \mathrm{diag}(\bar{g}_k) \Sigma_k^{1/2} \zeta_k. \tag{133}$$

We write $g_k \odot g_k$ as:

$$g_k \odot g_k = \bar{g}_k^{\odot 2} - \frac{2}{\sqrt{b}} \mathrm{diag}(\bar{g}_k) \left[\Sigma_k^{1/2} \zeta_k \right] + \frac{1}{b} (\Sigma_k^{1/2} \zeta_k)^{\odot 2}. \tag{134}$$

Statistics of Element-Wise Squared Gradient. Fix θ_k over all batches, the element-wise squared gradient $g_k \odot g_k$ admits a mean

$$\begin{aligned}\mathbb{E}_{\xi_k}[g_k \odot g_k \mid \theta_k] &= \mathbb{E}\left[\bar{g}_k^{\odot 2} - \frac{2}{\sqrt{b}}\operatorname{diag}(\bar{g}_k)\,\Sigma_k^{1/2}\zeta_k + \frac{1}{b}\left(\Sigma_k^{1/2}\zeta_k\right)^{\odot 2}\right] \\ &= \bar{g}_k^{\odot 2} - \frac{2}{\sqrt{b}}\operatorname{diag}(\bar{g}_k)\,\mathbb{E}\left[\Sigma_k^{1/2}\zeta_k\right] + \frac{1}{b}\mathbb{E}\left[\left(\Sigma_k^{1/2}\zeta_k\right)^{\odot 2}\right] \\ &= \bar{g}_k^{\odot 2} + \frac{1}{b}\operatorname{diag}(\Sigma_k)\,, \end{aligned} \tag{135}$$

and a covariance

$$\begin{aligned}\operatorname{Cov}_{\xi_k}(g_k \odot g_k \mid \theta_k) &= \operatorname{Cov}\left(-\frac{2}{\sqrt{b}}\operatorname{diag}(\bar{g}_k)\,\Sigma_k^{1/2}\zeta_k + \frac{1}{b}\left(\Sigma_k^{1/2}\zeta_k\right)^{\odot 2}\right) \\ &= \frac{4}{b}\operatorname{diag}(\bar{g}_k)\,\Sigma_k\operatorname{diag}(\bar{g}_k) + \frac{1}{b^2}\operatorname{Cov}\left(\left(\Sigma_k^{1/2}\zeta_k\right)^{\odot 2}\right) \\ &= \frac{4}{b}\operatorname{diag}(\bar{g}_k)\,\Sigma_k\operatorname{diag}(\bar{g}_k) + \frac{1}{b^2}\left[2\left(\Sigma_k \odot \Sigma_k\right)\right] \\ &= \frac{4}{b}\operatorname{diag}(\bar{g}_k)\,\Sigma_k\operatorname{diag}(\bar{g}_k) + \frac{2}{b^2}\left(\Sigma_k \odot \Sigma_k\right), \end{aligned} \tag{136}$$

where the last identity follows componentwise from

$$\begin{aligned}\left[\operatorname{Cov}\left(\left(\Sigma_k^{1/2}\zeta_k\right)^{\odot 2}\right)\right]_{ij} &= \mathbb{E}\left[\left(\Sigma_k^{1/2}\zeta_k\right)_i^2\left(\Sigma_k^{1/2}\zeta_k\right)_j^2\right] - \mathbb{E}\left[\left(\Sigma_k^{1/2}\zeta_k\right)_i^2\right]\mathbb{E}\left[\left(\Sigma_k^{1/2}\zeta_k\right)_j^2\right] \\ &= \left[(\Sigma_k)_{ii}(\Sigma_k)_{jj} + 2(\Sigma_k)_{ij}^2\right] - (\Sigma_k)_{ii}(\Sigma_k)_{jj} \\ &= 2(\Sigma_k)_{ij}^2 = 2(\Sigma_k \odot \Sigma_k)_{ij}. \end{aligned} \tag{137}$$

Continuous-Time Limit of Second-Moment Estimate. From equation (96):

$$v_{k+1} - v_k = \beta_2 v_k + (1-\beta_2)(g_k \odot g_k) - v_k \tag{138}$$
$$= -(1-\beta_2)v_k + (1-\beta_2)(g_k \odot g_k) \tag{139}$$
$$= -(1-\beta_2)[v_k - g_k \odot g_k], \tag{140}$$

substituting equation (134) yields:

$$v_{k+1} - v_k = -(1-\beta_2)\left[v_k - \bar{g}_k^{\odot 2} + \frac{2}{\sqrt{b}}\operatorname{diag}(\bar{g}_k)\Sigma_k^{1/2}\zeta_k - \frac{1}{b}(\Sigma_k^{1/2}\zeta_k)^{\odot 2}\right] \tag{141}$$

Expanding yields:

$$v_{k+1} - v_k = -(1-\beta_2)\left[v_k - \bar{g}_k^{\odot 2}\right] - \frac{2(1-\beta_2)}{\sqrt{b}}\operatorname{diag}(\bar{g}_k)\Sigma_k^{1/2}\zeta_k + \frac{1-\beta_2}{b}(\Sigma_k^{1/2}\zeta_k)^{\odot 2}. \tag{142}$$

Consider the diffusion terms:

$$\frac{2(1-\beta_2)}{\sqrt{b}}\operatorname{diag}(\bar{g}_k)\Sigma_k^{1/2}\zeta_k = \frac{2(1-\beta_2)}{\sqrt{b}}\operatorname{diag}(\bar{g}_k)\Sigma_k^{1/2}\frac{\Delta W_k}{\sqrt{\eta}} \tag{143}$$
$$= \frac{2(1-\beta_2)}{\sqrt{b\eta}}\operatorname{diag}(\bar{g}_k)\Sigma_k^{1/2}\Delta W_k, \tag{144}$$

and:

$$\left[\Sigma_k^{1/2}\zeta_k\right]_i^2 = \left[\Sigma_k^{1/2}\frac{\Delta W_k}{\sqrt{\eta}}\right]_i^2 \tag{145}$$

$$= \left(\sum_{j=1}^{p}(\Sigma_k)_{ij}^{1/2}\frac{\Delta W_k^{(j)}}{\sqrt{\eta}}\right)^2 \tag{146}$$

$$= \frac{1}{\eta}\sum_{j=1}^{p}\sum_{h=1}^{p}(\Sigma_k)_{ij}^{1/2}(\Sigma_k)_{ih}^{1/2}\Delta W_k^{(j)}\Delta W_k^{(h)} \tag{147}$$

$$= \frac{1}{\eta}\sum_{j=1}^{p}\sum_{h=1}^{p}(\Sigma_k)_{ij}^{1/2}(\Sigma_k)_{ih}^{1/2}\delta_{jh}\Delta t \tag{148}$$

$$= \frac{1}{\eta}\sum_{j=1}^{p}[(\Sigma_k)_{ij}^{1/2}]^2\,\Delta t \tag{149}$$

$$= \frac{\Delta t}{\eta}\,(\Sigma_k)_{ii} \;=\; \frac{\Delta t}{\eta}\,\mathrm{diag}\big(\Sigma_k\big)_i\,, \tag{150}$$

so $(\Sigma^{1/2}\zeta_k)^{\odot 2} = (\Delta t/\eta)\,\mathrm{diag}(\Sigma_k)$.

Take limits $\Delta t \to \mathrm{d}t, \Delta W_k \to \mathrm{d}W_t$, then the continuous-time limit of $\Delta v = v_{k+1} - v_k$ is:

$$\mathrm{d}v_t = -\frac{1-\beta_2}{\eta}\Big[v_t - \bar{g}_k^{\odot 2}\Big]\mathrm{d}t - \frac{2(1-\beta_2)}{\sqrt{b\eta}}\,\mathrm{diag}(\bar{g}_t)\,\Sigma_t^{1/2}\mathrm{d}W_t + \frac{1-\beta_2}{\eta b}\,\mathrm{diag}\big(\Sigma_t\big)\mathrm{d}t. \tag{151}$$

Continuous-Time Limit of Parameter Dynamics. From equation (97) (with ε dropped), the discrete update uses the bias-corrected moments

$$\hat{m}_k = m_k/(1-\beta_1^k), \qquad \hat{v}_k = v_k/(1-\beta_2^k), \tag{152}$$

then

$$\theta_{k+1} \;=\; \theta_k - \eta\,\hat{m}_k \oslash \sqrt{\hat{v}_k} \;=\; \theta_k - \eta\,\frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k}\,m_k \oslash \sqrt{v_k}. \tag{153}$$

Under the joint scaling $\eta \to 0$ with $(1-\beta_i)/\eta = \alpha_i$ fixed and $t = \eta k$,

$$1-\beta_i^k \;=\; 1-(1-\alpha_i\eta)^{t/\eta} \;\xrightarrow[\eta\to 0]{}\; 1-e^{-\alpha_i t},$$

$$\frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k} \;\xrightarrow[\eta\to 0]{}\; \frac{\sqrt{1-e^{-\alpha_2 t}}}{1-e^{-\alpha_1 t}} =: B(t), \tag{154}$$

with $B(t)$ defined in equation (85). The continuous-time θ-dynamics is therefore

$$\mathrm{d}\theta_t \;=\; -\,B(t)\,m_t \oslash \sqrt{v_t}\,\mathrm{d}t, \tag{155}$$

which is time-inhomogeneous through $B(t)$.

Joint-State SDE. To simplify discussion, by equations (134), (135), (136),

$$g_t \odot g_t = \bar{g}_t \odot \bar{g}_t - \frac{2}{\sqrt{b}}\mathrm{diag}(\bar{g}_t)\left[\Sigma_t^{1/2}\zeta_t\right] + \frac{1}{b}(\Sigma_t^{1/2}\zeta_t)^{\odot 2}, \tag{156}$$

$$\mathbb{E}_{\xi_t}[g_t \odot g_t \mid \theta_t] = \bar{g}_t \odot \bar{g}_t + \frac{1}{b}\,\mathrm{diag}(\Sigma_t)\,, \tag{157}$$

$$\mathrm{Cov}_{\xi_t}(g_t \odot g_t \mid \theta_t) = \frac{4}{b}\,\mathrm{diag}(\bar{g}_t)\,\Sigma_t\,\mathrm{diag}(\bar{g}_t) + \frac{2}{b^2}\,(\Sigma_t \odot \Sigma_t)\,, \tag{158}$$

we set

$$\overline{g \odot g}_t \;:=\; \mathbb{E}[g_t \odot g_t \mid \theta_k] = \bar{g}_t \odot \bar{g}_t \;+\; \tfrac{1}{b}\,\mathrm{diag}\,(\Sigma_t), \tag{159}$$

$$D_t \;:=\; 2\,\mathrm{diag}\big(\bar{g}_t\big)\,\Sigma_t^{1/2}, \quad \frac{1}{b} D_t D_t^\top = \mathrm{Cov}[g_k \odot g_k \mid \theta_k] - O(\frac{1}{b^2}), \tag{160}$$

where D_t is the diffusion factor of the leading cross-term in $g_k \odot g_k$; the sub-leading term has Wick's variance at a scale by $O(1/b^2)$ and is dropped if $b \to \infty$.

Set

$$\alpha_i := (1-\beta_i)/\eta, \tag{161}$$

and combining equations (155), (127), (151),

$$\mathrm{d}\theta_t = -\,B(t)\, m_t \oslash \sqrt{v_t}\,\mathrm{d}t, \tag{162}$$

$$\begin{aligned}
\mathrm{d}m_t &= -\frac{1-\beta_1}{\eta}\,[m_t - \bar{g}_t]\,\mathrm{d}t - \frac{1-\beta_1}{\sqrt{b\eta}}\Sigma_t^{1/2}\mathrm{d}W_t \\
&= -\alpha_1(m_t - \bar{g}_t)\mathrm{d}t - \sqrt{\frac{\eta}{b}}\,\alpha_1\Sigma_t^{1/2}\mathrm{d}W_t,
\end{aligned} \tag{163}$$

$$\begin{aligned}
\mathrm{d}v_t &= -\frac{1-\beta_2}{\eta}\Big[v_t - \bar{g}_t \odot \bar{g}_t\Big]\mathrm{d}t \\
&\quad - \frac{2(1-\beta_2)}{\sqrt{b\eta}}\,\mathrm{diag}(\bar{g}_t)\,\Sigma_t^{1/2}\mathrm{d}W_t + \frac{1-\beta_2}{\eta b}\,\mathrm{diag}\big(\Sigma_t\big)\mathrm{d}t \\
&= -\alpha_2\Big[v_t - \bar{g}_t \odot \bar{g}_t - \frac{1}{b}\,\mathrm{diag}(\Sigma_t)\Big]\mathrm{d}t - \sqrt{\frac{\eta}{b}}\,\alpha_2 D_t \mathrm{d}W_t \\
&= -\alpha_2\,(v_t - \overline{g \odot g}_t)\,\mathrm{d}t - \sqrt{\frac{\eta}{b}}\,\alpha_2 D_t \mathrm{d}W_t,
\end{aligned} \tag{164}$$

so that:

$$\mathrm{d}\begin{pmatrix}\theta_t\\ m_t\\ v_t\end{pmatrix} = \begin{pmatrix}-B(t)\, m_t \oslash \sqrt{v_t}\\ -\alpha_1\,(m_t - \bar{g}_t)\\ -\alpha_2\,(v_t - \overline{g \odot g}_t)\end{pmatrix}\mathrm{d}t + \sqrt{\frac{\eta}{b}}\begin{pmatrix}0_{p\times p}\\ \alpha_1\,\Sigma_t^{1/2}\\ \alpha_2\,D_t\end{pmatrix}\mathrm{d}W_t. \tag{165}$$

Write the joint-state as

$$S_t := \begin{pmatrix}\theta_t\\ m_t\\ v_t\end{pmatrix} \in \mathbb{R}^{3p}, \tag{166}$$

then Adam's optimization is characterized by

$$\mathrm{d}S_t \;=\; \mu(S_t)\,\mathrm{d}t \;+\; \sqrt{\frac{\eta}{b}}\,\sigma(S_t)\,\mathrm{d}W_t, \tag{167}$$

with W_t a p-dimensional Wiener process and

$$\mu(S_t) \;=\; \begin{pmatrix}-B(t)\, m_t \oslash \sqrt{v_t}\\ -\alpha_1\,(m_t - \bar{g}_t)\\ -\alpha_2\,(v_t - \overline{g \odot g}_t)\end{pmatrix} \in \mathbb{R}^{3p}, \qquad \sigma(S_t) \;=\; \begin{pmatrix}0_{p\times p}\\ \alpha_1\,\Sigma_t^{1/2}\\ \alpha_2\,D_t\end{pmatrix} \in \mathbb{R}^{3p\times p}. \tag{168}$$

The technical correctness verification is provided in later sections and experiments, such as Appendix A.6 through an numerical experiment.

□

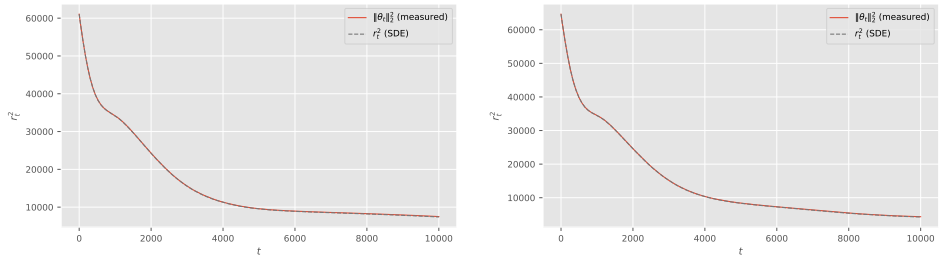
Remark 5 (SGD's SDE as Special Case). The joint-state SDE in equation (167) contains the SGD's SDE as a limiting case. Specifically, take the no-momentum limit $\beta_1 = 0$, so that $m_t = g_t$, and ignore the second-moment state v_t equivalently, take $\beta_2 = 0$ and replace $\sqrt{v_t}$ by $\mathbf{1}_p$. Then Adam's θ-update reduces to the SGD update

$$\theta_{k+1} = \theta_k - \eta g_k, \tag{169}$$

and its continuous-time interpolation gives the standard SGD SDE

$$\mathrm{d}\theta_t = -\bar{g}_t\mathrm{d}t + \sqrt{\eta/b}\,\Sigma_t^{1/2}\mathrm{d}W_t. \tag{170}$$

A.6 Proof: Induced Radius SDE



(a) Problem S_5 (b) Problem $\mathbb{Z}_{127}$

Figure 8: **Sanity Check with Radius SDE.** We use the induced radius SDE (Lemma 6), derived via Itô's lemma, as a sanity check for Adam's continuous-time SDE limit.

We use the induced radius SDE as a sanity check on the joint-state SDE of Lemma 1. Applying Itô's lemma (Øksendal, 2003) to the quadratic form $r_t^2 = S_t^\top E\, S_t$ with the θ-block projector $E \in \mathbb{R}^{3p\times 3p}$ defined in equation (176) below must reproduce the Itô decomposition stated in Lemma 6; empirical verification on Adam runs is reported in Figure 8.

Lemma 6 (Induced Radius SDE)**.** *Let $S_t = (\theta_t^\top, m_t^\top, v_t^\top)^\top \in \mathbb{R}^{3p}$ evolve according to Adam's joint-state SDE*

$$\mathrm{d}S_t = \mu(S_t)\mathrm{d}t + \sqrt{\frac{\eta}{b}}\,\sigma(S_t)\mathrm{d}W_t, \tag{171}$$

where

$$\mu(S_t) = \begin{pmatrix} -B(t)\, m_t \oslash \sqrt{v_t} \\ -\alpha_1(m_t - \bar{g}_t) \\ -\alpha_2(v_t - \overline{g \odot g}_t) \end{pmatrix}, \qquad \sigma(S_t) = \begin{pmatrix} 0 \\ \alpha_1 \Sigma_t^{1/2} \\ \alpha_2 D_t \end{pmatrix}, \tag{172}$$

$W_t \in \mathbb{R}^p$ is a Wiener process, and $B(t)$ is the Adam bias-correction factor. Define

$$r_t^2 := \|\theta_t\|_2^2, \tag{173}$$

then r_t^2 admits the dynamics

$$\mathrm{d}r_t^2 = -2B(t)\theta_t^\top\,(m_t \oslash \sqrt{v_t})\,\mathrm{d}t. \tag{174}$$

This is not an ordinary differential equation (ODE) as the m_t and v_t are stochastic through its dependence on the joint state S_t.

Proof. Recall from Lemma 5 the joint-state SDE on $S_t = (\theta_t^\top, m_t^\top, v_t^\top)^\top \in \mathbb{R}^{3p}$,

$$\mathrm{d}S_t \;=\; \mu(S_t)\,\mathrm{d}t \;+\; \sqrt{\eta/b}\,\sigma(S_t)\,\mathrm{d}W_t, \tag{175}$$

with $\mu(S_t)$ and $\sigma(S_t)$ as in equations (168).

Define a projector $E \in \mathbb{R}^{3p\times 3p}$

$$E \;:=\; \begin{pmatrix} I_p & 0_p & 0_p \\ 0_p & 0_p & 0_p \\ 0_p & 0_p & 0_p \end{pmatrix}, \tag{176}$$

where each block is $p \times p$. Then E is symmetric and idempotent, $E^\top = E$ and $E^2 = E$, and for $S = (\theta^\top, m^\top, v^\top)^\top$ the squared parameter norm is the quadratic form

$$r^2 \;=\; \|\theta\|_2^2 \;=\; S^\top E\, S. \tag{177}$$

Apply Itô's lemma to $f(S) := S^\top E\, S$ under equation (175). The gradient and Hessian of f are

$$\nabla_S f(S) \;=\; 2\,E\,S, \qquad \nabla_S^2 f(S) \;=\; 2\,E, \tag{178}$$

so

$$\mathrm{d}r_t^2 \;=\; 2\,S_t^\top E\,\mu(S_t)\,\mathrm{d}t \;+\; 2\sqrt{\eta/b}\,S_t^\top E\,\sigma(S_t)\,\mathrm{d}W_t \;+\; \tfrac{\eta}{b}\,\mathrm{tr}\big(\sigma(S_t)^\top E\,\sigma(S_t)\big)\,\mathrm{d}t. \tag{179}$$

Each contraction with E keeps only the θ-block:

$$S_t^\top E\,\mu(S_t) \;=\; \theta_t^\top\,\mu_\theta(S_t) \;=\; -\,B(t)\,\theta_t^\top\big(m_t \oslash \sqrt{v_t}\big), \tag{180}$$

$$S_t^\top E\,\sigma(S_t) \;=\; \theta_t^\top\,\sigma_\theta(S_t) \;=\; 0, \tag{181}$$

$$\mathrm{tr}\big(\sigma(S_t)^\top E\,\sigma(S_t)\big) \;=\; \mathrm{tr}\big(\sigma_\theta(S_t)^\top\sigma_\theta(S_t)\big) \;=\; 0. \tag{182}$$

Equations (181)–(182) use $\sigma_\theta \equiv 0$. Substituting equations (180)–(182) into equation (179) yields

$$\mathrm{d}r_t^2 \;=\; -\,2\,B(t)\,\theta_t^\top\big(m_t \oslash \sqrt{v_t}\big)\,\mathrm{d}t, \tag{183}$$

which is equation (174). □

A.7 Proof: Mean-Field Limit of First- and Second-Moment Estimates

Lemma 7 (Mean-Field Limit of First- and Second-Moment Estimates)**.** *Let m_t and v_t be Adam's first- and second-moment states, let g_t be the mini-batch gradient at time t, let $\bar{g}_t := \mathbb{E}[g_t \mid \theta_t]$ be the mean of g_t, and let $\overline{g \odot g}_t := \mathbb{E}[g_t \odot g_t \mid \theta_t]$ be the mean of $g_t \odot g_t$. Assume that $\bar{g}_t$ and $\overline{g \odot g}_t$ vary slowly in the long-term dynamics as $t \to \infty$. Then, under initialization $m_0 = 0, v_0 = 0$, the first- and second-moment estimates admit the mean-field limit*

$$m_t \to \bar{g}_t\left(1 - e^{-\alpha_1 t}\right), \qquad v_t \to \overline{g \odot g}_t\left(1 - e^{-\alpha_2 t}\right), \tag{184}$$

so that

$$\begin{aligned} m_t \oslash \sqrt{v_t} &\to \bar{g}_t\big(1 - e^{-\alpha_1 t}\big) \oslash \sqrt{\overline{g \odot g}_t\big(1 - e^{-\alpha_2 t}\big)} \\ &\approx \bar{g}_t\big(1 - e^{-\alpha_1 t}\big) \oslash \sqrt{\overline{g \odot g}_t}. \end{aligned} \tag{185}$$

Proof. We start from the stochastic moment dynamics of the joint-state Adam SDE,

$$\mathrm{d}m_t = -\alpha_1(m_t - \bar{g}_t)\mathrm{d}t - \sqrt{\frac{\eta}{b}}\,\alpha_1\Sigma_t^{1/2}\mathrm{d}W_t, \tag{186}$$

$$\mathrm{d}v_t = -\alpha_2\left(v_t - \overline{g \odot g}_t\right)\mathrm{d}t - \sqrt{\frac{\eta}{b}}\,\alpha_2 D_t\mathrm{d}W_t. \tag{187}$$

The mean-field approximation replaces the stochastic moment dynamics by their conditional mean dynamics. Since the stochastic terms are martingale increments,

$$\mathbb{E}[\mathrm{d}W_t \mid \theta_t, m_t, v_t] = 0. \tag{188}$$

Taking conditional expectation in (186) and (187) gives

$$\mathbb{E}[\mathrm{d}m_t \mid \theta_t, m_t, v_t] = -\alpha_1(m_t - \bar{g}_t)\mathrm{d}t, \tag{189}$$

$$\mathbb{E}[\mathrm{d}v_t \mid \theta_t, m_t, v_t] = -\alpha_2\left(v_t - \overline{g \odot g}_t\right)\mathrm{d}t. \tag{190}$$

Thus, under the mean-field dynamics,

$$\mathrm{d}m_t = -\alpha_1(m_t - \bar{g}_t)\mathrm{d}t, \tag{191}$$

$$\mathrm{d}v_t = -\alpha_2\left(v_t - \overline{g \odot g}_t\right)\mathrm{d}t. \tag{192}$$

Because $\bar{g}_t$ varies slowly on the m_t relaxation time scale, we freeze $\bar{g}_t$ when solving equation (191). Hence

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = -\alpha_1 m_t + \alpha_1\bar{g}_t. \tag{193}$$

Multiplying by $e^{\alpha_1 t}$ gives

$$\begin{aligned} e^{\alpha_1 t}\frac{\mathrm{d}m_t}{\mathrm{d}t} + \alpha_1 e^{\alpha_1 t}m_t &= \alpha_1 e^{\alpha_1 t}\bar{g}_t \\ \frac{\mathrm{d}}{\mathrm{d}t}\left(e^{\alpha_1 t}m_t\right) &= \alpha_1 e^{\alpha_1 t}\bar{g}_t. \end{aligned} \tag{194}$$

Integrating from 0 to t,

$$e^{\alpha_1 t} m_t - m_0 = \int_0^t \alpha_1 e^{\alpha_1 u} \bar{g}_t \, \mathrm{d}u$$
$$= \bar{g}_t \left(e^{\alpha_1 t} - 1 \right). \tag{195}$$

Using $m_0 = 0$, we obtain

$$m_t = \bar{g}_t \left(1 - e^{-\alpha_1 t} \right). \tag{196}$$

Similarly, because $\overline{g \odot g}_t$ varies slowly on the v_t relaxation time scale, we freeze $\overline{g \odot g}_t$ when solving (192). Hence

$$\frac{\mathrm{d}v_t}{\mathrm{d}t} = -\alpha_2 v_t + \alpha_2 \overline{g \odot g}_t. \tag{197}$$

Multiplying by $e^{\alpha_2 t}$ gives

$$e^{\alpha_2 t} \frac{\mathrm{d}v_t}{\mathrm{d}t} + \alpha_2 e^{\alpha_2 t} v_t = \alpha_2 e^{\alpha_2 t} \overline{g \odot g}_t$$
$$\frac{\mathrm{d}}{\mathrm{d}t} \left(e^{\alpha_2 t} v_t \right) = \alpha_2 e^{\alpha_2 t} \overline{g \odot g}_t. \tag{198}$$

Integrating from 0 to t,

$$e^{\alpha_2 t} v_t - v_0 = \int_0^t \alpha_2 e^{\alpha_2 u} \overline{g \odot g}_t \, \mathrm{d}u$$
$$= \overline{g \odot g}_t \left(e^{\alpha_2 t} - 1 \right). \tag{199}$$

Therefore,

$$v_t = \overline{g \odot g}_t + \left(v_0 - \overline{g \odot g}_t \right) e^{-\alpha_2 t}. \tag{200}$$

Using $v_0 = 0$, this becomes

$$v_t = \overline{g \odot g}_t \left(1 - e^{-\alpha_2 t} \right). \tag{201}$$

Combining equation (196) and (201), we get

$$m_t \oslash \sqrt{v_t} = \bar{g}_t \big(1 - e^{-\alpha_1 t} \big) \oslash \sqrt{\overline{g \odot g}_t (1 - e^{-\alpha_2 t})}. \tag{202}$$

For large t on the v_t relaxation time scale, $\sqrt{1 - e^{-\alpha_2 t}}$ decays faster than $1 - e^{-\alpha_1 t}$, so that

$$m_t \oslash \sqrt{v_t} \approx \bar{g}_t \big(1 - e^{-\alpha_1 t} \big) \oslash \sqrt{\overline{g \odot g}_t}. \tag{203}$$

□

A.8 Proof: Preconditioned-Decomposition of Adam's θ-SDE

We decompose Adam's exact θ-dynamics into a preconditioned drift and diffusion with a residual. This decomposition helps us to formulate the late-stage evolution of training dynamics with Adam in grokking. The technical correctness is verified via numerical experiments in Appendix A.9.

Lemma 8 (Preconditioned-Decomposition of Adam's θ-SDE)**.** *Let $S_t = (\theta_t^\top, m_t^\top, v_t^\top)^\top$ evolve under Adam's joint-state SDE. Define*

$$a(t) := \sqrt{1 - e^{-\alpha_2 t}}, \qquad B(t) := \frac{\sqrt{1 - e^{-\alpha_2 t}}}{1 - e^{-\alpha_1 t}}, \tag{204}$$

$$\pi(\theta_t) := \mathrm{diag}(\overline{g \odot g}_t)^{-1/2}, \qquad G(\theta_t) = \frac{1}{\sqrt{b}} B(t) \pi(\theta_t) \Sigma_t^{1/2}. \tag{205}$$

where $\pi(\theta_t)$ is referred to as the preconditioner of Adam's SDE and $G(\theta_t)$ is referred to as the preconditioned diffusion factor. By dropping higher correction orders, Adam's parameter dynamics admit the preconditioned decomposition

$$\mathrm{d}\theta_t \approx -a(t) \pi(\theta_t) \bar{g}_t \mathrm{d}t + \mathcal{R}_{\theta,\mathrm{SM}}(t) \mathrm{d}t + \sqrt{\eta} \, G(\theta_t) \mathrm{d}W_t, \tag{206}$$

where

$$\mathcal{R}_{\theta,\mathrm{SM}}(t) := -B(t)\left[\mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] - \bar{g}_t\left(1-e^{-\alpha_1 t}\right) \oslash \sqrt{\overline{g \odot g}_t}\right], \tag{207}$$

is referred to as the slow-manifold residual, which represents the residual term to the preconditioned term. This result immediately recovers standard SGD's SDE with $B(t) = 1, m_t = g_t, v_t = \mathbf{1}_p, \pi(\theta_t) = I_p$.

Proof. We decompose the θ-component of Adam's joint-state SDE

$$\mathrm{d}\theta_t = -B(t)\left(m_t \oslash \sqrt{v_t}\right)\mathrm{d}t, \tag{208}$$

where

$$B(t) = \frac{\sqrt{1-e^{-\alpha_2 t}}}{1-e^{-\alpha_1 t}}, \qquad a(t) = \sqrt{1-e^{-\alpha_2 t}}. \tag{209}$$

The SDE preconditioner is defined as

$$\pi(\theta_t) := \mathrm{diag}(\overline{g \odot g}_t)^{-1/2}. \tag{210}$$

Mean-Field Decomposition. By Lemma 7 (Mean-Field Limit of First- and Second-Moment Estimates), consider the evolution

$$m_t \oslash \sqrt{v_t} \to \bar{g}_t\left(1-e^{-\alpha_1 t}\right) \oslash \sqrt{\overline{g \odot g}_t}, \tag{211}$$

we add and subtract the mean-field limit

$$\bar{g}_t\left(1-e^{-\alpha_1 t}\right) \oslash \sqrt{\overline{g \odot g}_t} \tag{212}$$

from equation (208), then

$$\begin{aligned}
\mathrm{d}\theta_t &= -B(t)\left(m_t \oslash \sqrt{v_t}\right)\mathrm{d}t \\
&= \underbrace{-B(t)\left[\bar{g}_t\left(1-e^{-\alpha_1 t}\right) \oslash \sqrt{\overline{g \odot g}_t}\right]}_{\text{mean-field component}}\mathrm{d}t \\
&\quad \underbrace{-B(t)\left[m_t \oslash \sqrt{v_t} - \bar{g}_t\left(1-e^{-\alpha_1 t}\right) \oslash \sqrt{\overline{g \odot g}_t}\right]}_{\text{mean-field residual}}\mathrm{d}t.
\end{aligned} \tag{213}$$

Since

$$B(t)\left(1-e^{-\alpha_1 t}\right) = a(t), \tag{214}$$

and

$$\bar{g}_t \oslash \sqrt{\overline{g \odot g}_t} = \pi(\theta_t)\bar{g}_t, \tag{215}$$

the first term (mean-field component) in equation (213) becomes

$$-B(t)\left[\bar{g}_t\left(1-e^{-\alpha_1 t}\right) \oslash \sqrt{\overline{g \odot g}_t}\right]\mathrm{d}t = -a(t)\pi(\theta_t)\bar{g}_t\mathrm{d}t. \tag{216}$$

Martingale Decomposition of Mean-Field Residual. Let

$$\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \theta_t]. \tag{217}$$

Define

$$\mathcal{R}_{\theta,\mathrm{SM}}(t) := -B(t)\left[\mathbb{E}_t[m_t \oslash \sqrt{v_t}] - \bar{g}_t\left(1-e^{-\alpha_1 t}\right) \oslash \sqrt{\overline{g \odot g}_t}\right] \tag{218}$$

as the small-manifold residual. Then the second term in equation (213) admits a martingale decomposition

$$\begin{aligned}
&-B(t)\left[m_t \oslash \sqrt{v_t} - \bar{g}_t\left(1-e^{-\alpha_1 t}\right) \oslash \sqrt{\overline{g \odot g}_t}\right]\mathrm{d}t \\
&\quad = \mathcal{R}_{\theta,\mathrm{SM}}(t)\mathrm{d}t - B(t)\left[m_t \oslash \sqrt{v_t} - \mathbb{E}_t[m_t \oslash \sqrt{v_t}]\right]\mathrm{d}t.
\end{aligned} \tag{219}$$

Effective Diffusion. To analyze the martingale residual

$$m_t \oslash \sqrt{v_t} - \mathbb{E}_t[m_t \oslash \sqrt{v_t}]\,, \tag{220}$$

define

$$\bar{m}_t := \left(1 - e^{-\alpha_1 t}\right)\bar{g}_t, \qquad \bar{v}_t := \overline{g \odot g}_t, \tag{221}$$

and

$$\delta m_t := m_t - \bar{m}_t, \qquad \delta v_t := v_t - \bar{v}_t. \tag{222}$$

A componentwise Taylor expansion of $m \oslash \sqrt{v}$ around $(\bar{m}_t, \bar{v}_t)$ gives

$$m_t \oslash \sqrt{v_t} = \bar{m}_t \oslash \sqrt{\bar{v}_t} + \pi(\theta_t)\delta m_t - \frac{1}{2}\left(\bar{m}_t \oslash \bar{v}_t^{3/2}\right) \odot \delta v_t + \mathcal{Q}_\theta(t), \tag{223}$$

where

$$\mathcal{Q}_\theta(t) = O\left(\|\delta m_t\|\,\|\delta v_t\| + \|\delta v_t\|_2^2\right). \tag{224}$$

Taking the conditional mean of equation (223) and subtracting it from equation (223), we obtain

$$\begin{aligned} m_t \oslash \sqrt{v_t} - \mathbb{E}_t[m_t \oslash \sqrt{v_t}] = \pi(\theta_t)\left(\delta m_t - \mathbb{E}_t[\delta m_t]\right) \\ - \frac{1}{2}\left(\bar{m}_t \oslash \bar{v}_t^{3/2}\right) \odot \left(\delta v_t - \mathbb{E}_t[\delta v_t]\right) \\ + \mathcal{Q}_\theta(t) - \mathbb{E}_t[\mathcal{Q}_\theta(t)]. \end{aligned} \tag{225}$$

Variation Residual of First-Moment Estimate. We now analyze the variation residual of first-moment estimate

$$\delta m_t - \mathbb{E}_t[\delta m_t]. \tag{226}$$

From Adam's joint-state SDE, the first-moment component satisfies

$$\mathrm{d}m_t = -\alpha_1 m_t \mathrm{d}t + \alpha_1 g_t \mathrm{d}t. \tag{227}$$

Rearranging equation (227) gives

$$m_t \mathrm{d}t = g_t \mathrm{d}t - \frac{1}{\alpha_1}\mathrm{d}m_t. \tag{228}$$

Taking conditional expectations in equation (228) gives

$$\mathbb{E}_t[m_t]\mathrm{d}t = \bar{g}_t \mathrm{d}t - \frac{1}{\alpha_1}\mathrm{d}\mathbb{E}_t[m_t]. \tag{229}$$

Subtracting equation (229) from equation (228) yields

$$\left(m_t - \mathbb{E}_t[m_t]\right)\mathrm{d}t = \left(g_t - \bar{g}_t\right)\mathrm{d}t - \frac{1}{\alpha_1}\mathrm{d}\left(m_t - \mathbb{E}_t[m_t]\right). \tag{230}$$

Since

$$\delta m_t - \mathbb{E}_t[\delta m_t] = m_t - \mathbb{E}_t[m_t], \tag{231}$$

we get

$$\left(\delta m_t - \mathbb{E}_t[\delta m_t]\right)\mathrm{d}t = \left(g_t - \bar{g}_t\right)\mathrm{d}t - \frac{1}{\alpha_1}\mathrm{d}\left(\delta m_t - \mathbb{E}_t[\delta m_t]\right). \tag{232}$$

In continuous-time diffusion scaling,

$$\left(g_t - \bar{g}_t\right)\mathrm{d}t = \sqrt{\eta/b}\,\Sigma_t^{1/2}\mathrm{d}W_t. \tag{233}$$

Substituting equation (233) into equation (232) gives

$$\left(\delta m_t - \mathbb{E}_t[\delta m_t]\right)\mathrm{d}t = \sqrt{\eta/b}\,\Sigma_t^{1/2}\mathrm{d}W_t - \frac{1}{\alpha_1}\mathrm{d}\left(\delta m_t - \mathbb{E}_t[\delta m_t]\right). \tag{234}$$

Define

$$\mathrm{d}\mathcal{R}_m(t) := -\frac{1}{\alpha_1}\mathrm{d}\left(\delta m_t - \mathbb{E}_t[\delta m_t]\right). \tag{235}$$

Then

$$\left(\delta m_t - \mathbb{E}_t[\delta m_t]\right)\mathrm{d}t = \sqrt{\eta/b}\,\Sigma_t^{1/2}\mathrm{d}W_t + \mathrm{d}\mathcal{R}_m(t). \tag{236}$$

Substituting equation (236) into the first term of equation (225) gives

$$\begin{aligned} &- B(t)\pi(\theta_t)\left(\delta m_t - \mathbb{E}_t[\delta m_t]\right)\mathrm{d}t \\ &\quad = -B(t)\sqrt{\eta/b}\,\pi(\theta_t)\Sigma_t^{1/2}\mathrm{d}W_t - B(t)\pi(\theta_t)\mathrm{d}\mathcal{R}_m(t). \end{aligned} \tag{237}$$

Absorbing the sign into W_t, and using the definition

$$G(\theta_t) := \frac{1}{\sqrt{b}}B(t)\pi(\theta_t)\Sigma_t^{1/2}, \tag{238}$$

we obtain

$$-B(t)\pi(\theta_t)\left(\delta m_t - \mathbb{E}_t[\delta m_t]\right)\mathrm{d}t = \sqrt{\eta}\,G(\theta_t)\mathrm{d}W_t - B(t)\pi(\theta_t)\mathrm{d}\mathcal{R}_m(t). \tag{239}$$

Produce Claims. Combining equations (225) and (239), we get

$$\begin{aligned} &- B(t)\left[m_t \oslash \sqrt{v_t} - \mathbb{E}_t[m_t \oslash \sqrt{v_t}]\right]\mathrm{d}t \\ &\quad = \sqrt{\eta}\,G(\theta_t)\mathrm{d}W_t + \mathrm{d}\mathcal{W}_\theta(t), \end{aligned} \tag{240}$$

where

$$\begin{aligned} \mathrm{d}\mathcal{W}_\theta(t) := &-B(t)\pi(\theta_t)\mathrm{d}\mathcal{R}_m(t) \\ &+ \frac{B(t)}{2}\left[\left(\bar{m}_t \oslash \bar{v}_t^{3/2}\right)\odot\left(\delta v_t - \mathbb{E}_t[\delta v_t]\right)\right]\mathrm{d}t \\ &- B(t)\left[\mathcal{Q}_\theta(t) - \mathbb{E}_t[\mathcal{Q}_\theta(t)]\right]\mathrm{d}t. \end{aligned} \tag{241}$$

Combining equations (216), (219), and (240), we obtain

$$\mathrm{d}\theta_t = -a(t)\pi(\theta_t)\bar{g}_t\mathrm{d}t + \mathcal{R}_{\theta,\mathrm{SM}}(t)\mathrm{d}t + \sqrt{\eta}\,G(\theta_t)\mathrm{d}W_t + \mathrm{d}\mathcal{W}_\theta(t). \tag{242}$$

Dropping higher-order terms and keeping the leading diffusion order gives

$$\mathrm{d}\theta_t \approx -a(t)\pi(\theta_t)\bar{g}_t\mathrm{d}t + \mathcal{R}_{\theta,\mathrm{SM}}(t)\mathrm{d}t + \sqrt{\eta}\,G(\theta_t)\mathrm{d}W_t. \tag{243}$$

This is the claimed preconditioned decomposition. □

Remark 6. In particular, in late-stage training with sufficiently large batch size b, the reduction of m_t, v_t

$$m_t \to \bar{g}_t\big(1 - e^{-\alpha_1 t}\big), \qquad v_t \to \overline{g \odot g}_t, \tag{244}$$

is referred to as the slow-manifold reduction. When the residual $\mathcal{R}_{\theta,\mathrm{SM}}(t)$ is negligible, then the reduced preconditioned θ-SDE can further be simplified into

$$\mathrm{d}\theta_t \approx -a(t)\pi(\theta_t)\bar{g}_t\mathrm{d}t + \sqrt{\eta}\,G(\theta_t)\mathrm{d}W_t. \tag{245}$$

Remark 7 (Quick Sanity Check). As a sanity check, if $\pi(\theta_t) = I_p$, $B(t) \to 1$, and the residual is ignored at leading order, then

$$G(\theta_t) \to \frac{1}{\sqrt{b}}\Sigma_t^{1/2}, \tag{246}$$

and the preconditioned θ-SDE reduces to

$$\mathrm{d}\theta_t = -\bar{g}_t\mathrm{d}t + \sqrt{\eta/b}\,\Sigma_t^{1/2}\mathrm{d}W_t, \tag{247}$$

which is the standard continuous-time SDE approximation of mini-batch SGD.

A.9 Proof: Preconditioned-Decomposition of Adam's Radius SDE

We decompose Adam's exact radius SDE into preconditioned and residual terms in Lemma 9. This decomposition is used to derive the reduced radius SDE under the slow-manifold reduction and small-isotropic gradient covariance observations. Its technical correctness is verified in Figure 9, in which the predicted radius dynamics exactly match the theoretical values computed from Lemma 9.

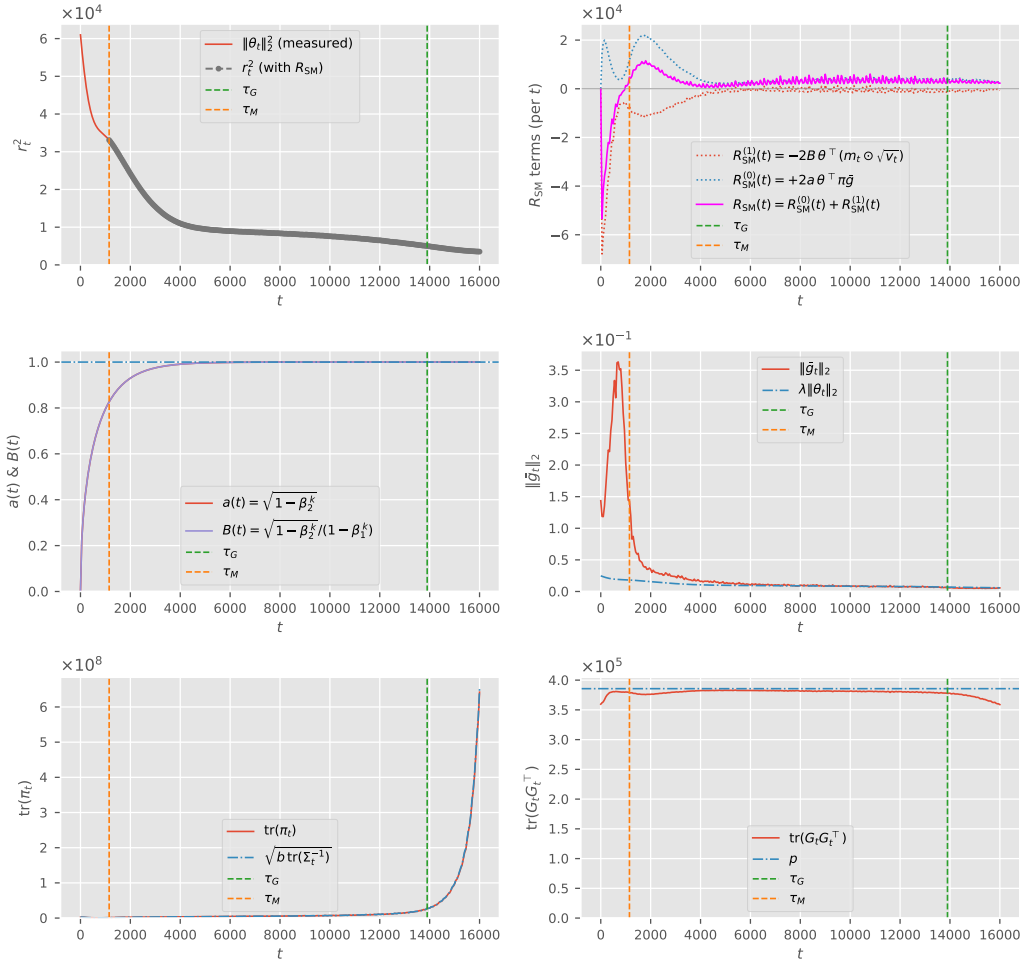


Figure 9: **Preconditioned Radius SDE.** This experiment shows the dynamics of the preconditioned radius SDE in Lemma 9, and two memorization-regime identities in Lemma 10 where $\mathrm{tr}(\pi(\theta_t)) \approx \sqrt{b}\,\mathrm{tr}\big[(\mathrm{diag}\,\Sigma_t)^{-1/2}\big]$ and $\mathrm{tr}(G(\theta_t)G(\theta_t)^\top) \approx p$. The predicted radius dynamics match the theoretical values computed from Lemma 9 and Lemma 10, validating the correctness of the decomposition.

Lemma 9 (Preconditioned-Decomposition of Adam's Radius SDE)**.** *Let* $S_t = (\theta_t^\top, m_t^\top, v_t^\top)^\top$ *evolve under Adam's joint-state SDE, and let* $r_t^2 := \|\theta_t\|_2^2$*. Define*

$$a(t) := \sqrt{1-e^{-\alpha_2 t}}, \qquad B(t) := \frac{\sqrt{1-e^{-\alpha_2 t}}}{1-e^{-\alpha_1 t}}$$
$$\pi(\theta) := \mathrm{diag}(\overline{g \odot g}(\theta))^{-1/2}, \qquad G(\theta_t) = \frac{1}{\sqrt{b}} B(t)\pi(\theta_t)\Sigma_t^{1/2}. \tag{248}$$

Then the residual-corrected preconditioned squared-radius dynamics are

$$\mathrm{d}r_t^2 \approx \big[-2a(t)\theta_t^\top \pi(\theta_t)\bar{g}_t + \mathcal{R}_{\mathrm{SM}}(t) + \eta\,\mathrm{tr}\big(G(\theta_t)G(\theta_t)^\top\big)\big]\,\mathrm{d}t + 2\sqrt{\eta}\,\theta_t^\top G(\theta_t)\mathrm{d}W_t, \tag{249}$$

where the slow-manifold residual is

$$\mathcal{R}_{\mathrm{SM}}(t) := -2B(t)\theta_t^\top \left[\mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] - \bar{g}_t\big(1 - e^{-\alpha_1 t}\big) \oslash \sqrt{\overline{g \odot g}_t}\right]. \tag{250}$$

Proof. Let

$$E_\theta := \begin{pmatrix} I_p & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{3p\times 3p} \tag{251}$$

be a projector. Since $S_t = (\theta_t^\top, m_t^\top, v_t^\top)^\top$, the squared radius can be written as the joint-state quadratic form

$$r_t^2 = \|\theta_t\|_2^2 = S_t^\top E_\theta S_t. \tag{252}$$

Therefore,

$$\nabla_S r_t^2 = 2E_\theta S_t = \begin{pmatrix} 2\theta_t \\ 0 \\ 0 \end{pmatrix}, \qquad \nabla_S^2 r_t^2 = 2E_\theta. \tag{253}$$

Remark 8. Since θ_t is the θ-component of the joint state $S_t = (\theta_t^\top, m_t^\top, v_t^\top)^\top$, and the components of S_t are coupled through Adam's dynamics, we should regard $\|\theta_t\|_2^2$ as a function of the full joint state rather than of an isolated variable θ_t. Thus, when applying Itô's lemma to the joint-state SDE, we introduce the projection map $E_\theta : S_t \mapsto \theta_t$ and write the squared radius as the bilinear form $S_t^\top E_\theta S_t$.

Radius Dynamics. Applying Itô's lemma to $r_t^2 = S_t^\top E_\theta S_t$ under Adam's joint-state SDE gives

$$\mathrm{d}r_t^2 = (\nabla_S r_t^2)^\top \mathrm{d}S_t + \frac{1}{2}\mathrm{tr}\left[\left(\frac{\eta}{b}\sigma(S_t)\sigma(S_t)^\top\right)\nabla_S^2 r_t^2\right]\mathrm{d}t. \tag{254}$$

The joint-state diffusion has zero θ-block, because the Brownian noise enters only through the (m, v) components. Since $\nabla_S^2 r_t^2 = 2E_\theta$ only selects the θ-block, the second-order term vanishes:

$$\mathrm{tr}\left[\left(\frac{\eta}{b}\sigma(S_t)\sigma(S_t)^\top\right)\nabla_S^2 r_t^2\right] = 0. \tag{255}$$

Thus,

$$\begin{aligned}\mathrm{d}r_t^2 &= (\nabla_S r_t^2)^\top \mathrm{d}S_t \\ &= 2\theta_t^\top \mathrm{d}\theta_t.\end{aligned} \tag{256}$$

From the θ-component of Adam's joint-state SDE,

$$\mathrm{d}\theta_t = -B(t)\left(m_t \oslash \sqrt{v_t}\right)\mathrm{d}t, \tag{257}$$

we obtain the exact induced squared-radius identity

$$\mathrm{d}r_t^2 = -2B(t)\theta_t^\top\left(m_t \oslash \sqrt{v_t}\right)\mathrm{d}t. \tag{258}$$

Mean-Field Decomposition. By Lemma 7 (Mean-Field Limit of First- and Second-Moment Estimates), consider the mean-field limit

$$m_t \oslash \sqrt{v_t} \to \bar{g}_t\big(1 - e^{-\alpha_1 t}\big) \oslash \sqrt{\overline{g \odot g}_t}, \tag{259}$$

we add and subtract

$$\bar{g}_t\big(1 - e^{-\alpha_1 t}\big) \oslash \sqrt{\overline{g \odot g}_t} \tag{260}$$

inside the Adam drift:

$$\begin{aligned}-2B(t)\theta_t^\top\left(m_t \oslash \sqrt{v_t}\right) = &-2B(t)\theta_t^\top\left[\bar{g}_t\big(1 - e^{-\alpha_1 t}\big) \oslash \sqrt{\overline{g \odot g}_t}\right] \\ &- 2B(t)\theta_t^\top\left[m_t \oslash \sqrt{v_t} - \bar{g}_t\big(1 - e^{-\alpha_1 t}\big) \oslash \sqrt{\overline{g \odot g}_t}\right].\end{aligned} \tag{261}$$

Using the continuous-time bias-correction factor

$$B(t) = \frac{\sqrt{1 - e^{-\alpha_2 t}}}{1 - e^{-\alpha_1 t}}, \tag{262}$$

and the definition $a(t) = \sqrt{1 - e^{-\alpha_2 t}}$, we have

$$B(t)\big(1 - e^{-\alpha_1 t}\big) = a(t). \tag{263}$$

Moreover, since

$$\pi(\theta_t) = \operatorname{diag}(\overline{g \odot g}_t)^{-1/2}, \tag{264}$$

we have the element-wise identity

$$\bar{g}_t \oslash \sqrt{\overline{g \odot g}_t} = \pi(\theta_t)\bar{g}_t. \tag{265}$$

Therefore, the first term in equation (261) becomes

$$-2B(t)\theta_t^\top \left[\bar{g}_t\big(1 - e^{-\alpha_1 t}\big) \oslash \sqrt{\overline{g \odot g}_t}\right] = -2a(t)\theta_t^\top \pi(\theta_t)\bar{g}_t. \tag{266}$$

Martingale Residual of Mean-Field Decomposition. The second term in equation (261) is exactly the slow-manifold residual:

$$\mathcal{R}_{\mathrm{SM}}(t) := -2B(t)\theta_t^\top \left[\mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] - \bar{g}_t\big(1 - e^{-\alpha_1 t}\big) \oslash \sqrt{\overline{g \odot g}_t}\right]. \tag{267}$$

Combining equations (261)–(267), the exact induced squared-radius has a drift

$$\left[-2a(t)\theta_t^\top \pi(\theta_t)\bar{g}_t + \mathcal{R}_{\mathrm{SM}}(t)\right] \mathrm{d}t. \tag{268}$$

Effective Diffusion. Finally, in the reduced preconditioned diffusion approximation, the mini-batch gradient fluctuations induce the θ-diffusion term $\sqrt{\eta}\, G(\theta_t)\mathrm{d}W_t$. Applying Itô's lemma to this diffusion contribution gives the additional correction

$$\eta \operatorname{tr}\big(G(\theta_t)G(\theta_t)^\top\big)\, \mathrm{d}t + 2\sqrt{\eta}\, \theta_t^\top G(\theta_t)\mathrm{d}W_t. \tag{269}$$

Produce Claims. Adding equation (269) to equation (268) yields

$$\mathrm{d}r_t^2 \approx \left[-2a(t)\theta_t^\top \pi(\theta_t)\bar{g}_t + \mathcal{R}_{\mathrm{SM}}(t) + \eta \operatorname{tr}\big(G(\theta_t)G(\theta_t)^\top\big)\right] \mathrm{d}t + 2\sqrt{\eta}\, \theta_t^\top G(\theta_t)\mathrm{d}W_t. \tag{270}$$

This proves equation (249). □

A.10 Proof: Memorization-Regime Preconditioner and Effective Diffusion Identities

The validation for Lemma 10 is provided in Figure 9. The experiment shows that, in the memorization regime, the approximated identities

$$\left.\begin{array}{l} \pi(\theta_t) := \operatorname{diag}(\overline{g \odot g}_t)^{-1/2} \\ \overline{g \odot g}_t \approx (1/b)\operatorname{diag}\Sigma_t \end{array}\right\} \Rightarrow \operatorname{tr}(\pi(\theta_t)) \approx \sqrt{b}\operatorname{tr}\big[(\operatorname{diag}\Sigma_t)^{-1/2}\big] \tag{271}$$

and

$$G(\theta_t) = \frac{1}{\sqrt{b}} B(t)\pi(\theta_t)\Sigma_t^{1/2} \Rightarrow \operatorname{tr}\big(G(\theta_t)G(\theta_t)^\top\big) \approx p \tag{272}$$

hold in the memorization regime.

Lemma 10 (Memorization-Regime Preconditioner and Effective Diffusion Identities)**.** *Let*

$$\pi(\theta_t) := \operatorname{diag}(\overline{g \odot g}_t)^{-1/2}, \tag{273}$$

and define

$$G(\theta_t) := \frac{1}{\sqrt{b}} B(t)\pi(\theta_t)\Sigma_t^{1/2}, \qquad B(t) := \frac{\sqrt{1-e^{-\alpha_2 t}}}{1-e^{-\alpha_1 t}}. \tag{274}$$

Assume that, in the memorization regime,

$$\overline{g \odot g}_t = \bar{g}_t \odot \bar{g}_t + \frac{1}{b}\mathrm{diag}(\Sigma_t) \approx \frac{1}{b}\mathrm{diag}(\Sigma_t), \tag{275}$$

and that the diagonal Adam preconditioner admits the trace-matched scalar approximation

$$\pi(\theta_t) \approx s(\theta_t)^{-1} I_p, \qquad \frac{1}{s(\theta)} := \frac{1}{p}\mathrm{tr}\big(\pi(\theta)\big). \tag{276}$$

Equivalently, under the corresponding scalar covariance approximation,

$$\frac{1}{b}\Sigma_t \approx s(\theta_t)^2 I_p. \tag{277}$$

Then the following two late-stage identities hold:

$$\mathrm{tr}(\pi(\theta_t)) \approx \sqrt{b}\,\mathrm{tr}\big[(\mathrm{diag}(\Sigma_t))^{-1/2}\big], \tag{278}$$

$$\mathrm{tr}\big(G(\theta_t)G(\theta_t)^\top\big) \approx p. \tag{279}$$

Moreover,

$$G(\theta_t)G(\theta_t)^\top \approx I_p \qquad (t \to \infty). \tag{280}$$

Proof. From equation (275),

$$\overline{g \odot g}_t \approx \frac{1}{b}\mathrm{diag}(\Sigma_t). \tag{281}$$

Therefore

$$\begin{aligned}\pi(\theta_t) &= \mathrm{diag}(\overline{g \odot g}_t)^{-1/2} \\ &\approx \mathrm{diag}\left(\frac{1}{b}\mathrm{diag}(\Sigma_t)\right)^{-1/2} \\ &= \sqrt{b}\,\mathrm{diag}(\mathrm{diag}(\Sigma_t))^{-1/2}.\end{aligned} \tag{282}$$

Taking traces gives

$$\mathrm{tr}(\pi(\theta_t)) \approx \sqrt{b}\,\mathrm{tr}\big[(\mathrm{diag}(\Sigma_t))^{-1/2}\big]. \tag{283}$$

This proves equation (278).

Next, by definition,

$$G(\theta_t) = \frac{1}{\sqrt{b}} B(t)\pi(\theta_t)\Sigma_t^{1/2}. \tag{284}$$

Hence

$$\begin{aligned}G(\theta_t)G(\theta_t)^\top &= \frac{B(t)^2}{b}\pi(\theta_t)\Sigma_t^{1/2}\left(\Sigma_t^{1/2}\right)^\top \pi(\theta_t)^\top \\ &= \frac{B(t)^2}{b}\pi(\theta_t)\Sigma_t\pi(\theta_t)^\top.\end{aligned} \tag{285}$$

Since $\pi(\theta_t)$ is diagonal, it is symmetric. Thus

$$\pi(\theta_t)^\top = \pi(\theta_t), \tag{286}$$

and equation (285) becomes

$$G(\theta_t)G(\theta_t)^\top = \frac{B(t)^2}{b}\pi(\theta_t)\Sigma_t\pi(\theta_t). \tag{287}$$

Using the trace-matched scalar approximation

$$\pi(\theta_t) \approx s(\theta_t)^{-1} I_p, \tag{288}$$

we obtain

$$\begin{aligned} G(\theta_t)G(\theta_t)^\top &\approx \frac{B(t)^2}{b} \left(s(\theta_t)^{-1} I_p\right) \Sigma_t \left(s(\theta_t)^{-1} I_p\right) \\ &= \frac{B(t)^2}{b\, s(\theta_t)^2} \Sigma_t. \end{aligned} \tag{289}$$

By

$$\frac{1}{b}\Sigma_t \approx s(\theta_t)^2 I_p, \tag{290}$$

or equivalently,

$$\Sigma_t \approx b\, s(\theta_t)^2 I_p. \tag{291}$$

Substituting equation (291) into equation (289) gives

$$\begin{aligned} G(\theta_t)G(\theta_t)^\top &\approx \frac{B(t)^2}{b\, s(\theta_t)^2} b\, s(\theta_t)^2 I_p \\ &= B(t)^2 I_p. \end{aligned} \tag{292}$$

Finally,

$$B(t)^2 = \frac{1 - e^{-\alpha_2 t}}{\left(1 - e^{-\alpha_1 t}\right)^2} \to 1 \qquad (t \to \infty). \tag{293}$$

Therefore,

$$G(\theta_t)G(\theta_t)^\top \approx I_p \qquad (t \to \infty). \tag{294}$$

Taking traces yields

$$\operatorname{tr}\left(G(\theta_t)G(\theta_t)^\top\right) \approx \operatorname{tr}(I_p) = p. \tag{295}$$

This proves equations (279) and (280). □

A.11 Proof: Reduced Memorization-Regime Radius SDE

We derive the exact closed-form reduced late-stage radius SDE with residual terms in Lemma 11. Its technical correctness is verified in Figure 10, where the predicted radius dynamics match the theoretical values computed from Lemma 11.

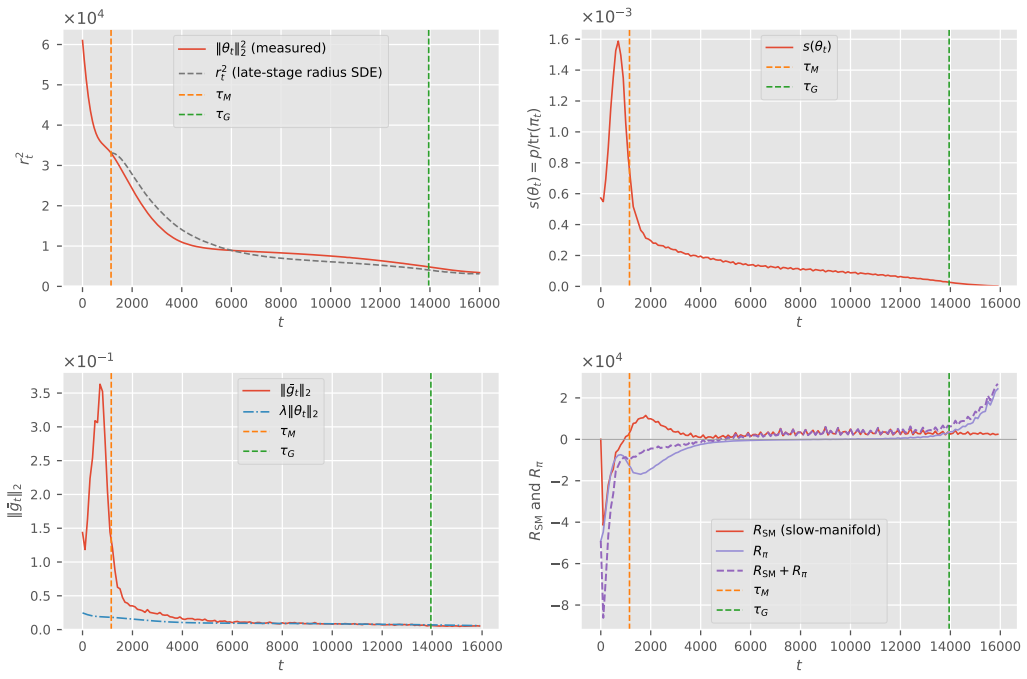


Figure 10: **Late-Stage Radius SDE.** This experiment shows the dynamics of the late-stage radius SDE in Lemma 11. The predicted radius dynamics closely match the theoretical values computed from Lemma 11 without counting residual terms, validating the correctness of the decomposition. In particular, the late-stage residual sum $\mathcal{R}_{\mathrm{SM}} + \mathcal{R}_{\pi}$ is negligible, as hypothesized.

Lemma 11 (Reduced Late-Stage Radius SDE with Slow-Manifold, Preconditioned Residual (Restated))**.** *Let $r_t := \|\theta_t\|_2$, where r_t^2 follows the residual-corrected preconditioned radius SDE in equation (249) of Lemma 9 (Preconditioned-Decomposition of Adam's Radius SDE). In the late-stage regime, assume*

$$B(t) := \frac{\sqrt{1 - e^{-\alpha_2 t}}}{1 - e^{-\alpha_1 t}} \to 1, \qquad G(\theta_t)G(\theta_t)^\top \approx I_p, \tag{296}$$

so that the residuals from $a(t)$ and the diffusion covariance are neglected. Then the reduced late-stage squared-radius SDE is

$$\mathrm{d}r_t^2 \approx \left[-\frac{2\lambda}{s(\theta_t)} r_t^2 - \frac{2}{s(\theta_t)} \theta_t^\top \left(\bar{g}_t - \lambda \theta_t \right) + \eta\, p + \mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_{\pi}(t) \right] \mathrm{d}t + 2\sqrt{\eta}\, r_t\, \mathrm{d}W_t^{(r)}, \tag{297}$$

where

$$\mathcal{R}_{\mathrm{SM}}(t) := -2B(t)\theta_t^\top \left[\mathbb{E}\left[m_t \oslash \sqrt{v_t} \right] - \bar{g}_t \left(1 - e^{-\alpha_1 t} \right) \oslash \sqrt{\overline{g \odot g}_t} \right], \tag{298}$$

$$\mathcal{R}_{\pi}(t) := -2\theta_t^\top \left[\pi(\theta_t) - s(\theta_t)^{-1} I_p \right] \bar{g}_t, \tag{299}$$

$$W_t^{(r)} := \int_0^t e_r(\theta_u)^\top G(\theta_u) \mathrm{d}W_u, \qquad e_r(\theta) := \frac{\theta}{\|\theta\|_2}, \tag{300}$$

and under $G(\theta_t)G(\theta_t)^\top \approx I_p$, $W_t^{(r)}$ is a one-dimensional Brownian motion.

Remark 9. Moreover,

$$\begin{aligned}\mathcal{R}_{\mathrm{SM}}(t)+\mathcal{R}_{\pi}(t) &= -2B(t)\theta_t^{\top}\mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] + 2a(t)\theta_t^{\top}\pi(\theta_t)\bar{g}_t \\ &\quad - 2\theta_t^{\top}\pi(\theta_t)\bar{g}_t + \frac{2}{s(\theta_t)}\theta_t^{\top}\bar{g}_t \\ &= -2B(t)\theta_t^{\top}\mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] + 2\big(a(t)-1\big)\theta_t^{\top}\pi(\theta_t)\bar{g}_t + \frac{2}{s(\theta_t)}\theta_t^{\top}\bar{g}_t. \end{aligned} \tag{301}$$

In particular, in the late-stage limit $a(t) \to 1$,

$$\mathcal{R}_{\mathrm{SM}}(t)+\mathcal{R}_{\pi}(t) \approx -2B(t)\theta_t^{\top}\mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] + \frac{2}{s(\theta_t)}\theta_t^{\top}\bar{g}_t. \tag{302}$$

Proof. Starting from the residual-corrected preconditioned radius SDE in equation (249) of Lemma 9 (Preconditioned-Decomposition of Adam's Radius SDE),

$$\begin{aligned}\mathrm{d}r_t^2 \approx \big[&-2a(t)\theta_t^{\top}\pi(\theta_t)\bar{g}_t + \mathcal{R}_{\mathrm{SM}}(t) + \eta\,\mathrm{tr}\big(G(\theta_t)G(\theta_t)^{\top}\big)\big]\,\mathrm{d}t \\ &+ 2\sqrt{\eta}\,\theta_t^{\top}G(\theta_t)\mathrm{d}W_t. \end{aligned} \tag{303}$$

Under $a(t) \to 1$,

$$-2a(t)\theta_t^{\top}\pi(\theta_t)\bar{g}_t \approx -2\theta_t^{\top}\pi(\theta_t)\bar{g}_t. \tag{304}$$

Decompose the preconditioner as

$$\pi(\theta_t) = s(\theta_t)^{-1}I_p + \left[\pi(\theta_t) - s(\theta_t)^{-1}I_p\right]. \tag{305}$$

Substituting this decomposition gives

$$-2\theta_t^{\top}\pi(\theta_t)\bar{g}_t = -\frac{2}{s(\theta_t)}\theta_t^{\top}\bar{g}_t + \mathcal{R}_{\pi}(t), \tag{306}$$

where

$$\mathcal{R}_{\pi}(t) = -2\theta_t^{\top}\left[\pi(\theta_t) - s(\theta_t)^{-1}I_p\right]\bar{g}_t. \tag{307}$$

Next, decompose the mean gradient as

$$\bar{g}_t = \lambda\theta_t + \big(\bar{g}_t - \lambda\theta_t\big). \tag{308}$$

Then

$$-\frac{2}{s(\theta_t)}\theta_t^{\top}\bar{g}_t = -\frac{2\lambda}{s(\theta_t)}r_t^2 - \frac{2}{s(\theta_t)}\theta_t^{\top}\big(\bar{g}_t - \lambda\theta_t\big). \tag{309}$$

Combining equations (306) and (309), we obtain

$$-2a(t)\theta_t^{\top}\pi(\theta_t)\bar{g}_t \approx -\frac{2\lambda}{s(\theta_t)}r_t^2 - \frac{2}{s(\theta_t)}\theta_t^{\top}\big(\bar{g}_t - \lambda\theta_t\big) + \mathcal{R}_{\pi}(t). \tag{310}$$

Under $G(\theta_t)G(\theta_t)^{\top} \approx I_p$,

$$\eta\,\mathrm{tr}\big(G(\theta_t)G(\theta_t)^{\top}\big) \approx \eta\,p. \tag{311}$$

Also,

$$\begin{aligned}2\sqrt{\eta}\,\theta_t^{\top}G(\theta_t)\mathrm{d}W_t &= 2\sqrt{\eta}\,r_t e_r(\theta_t)^{\top}G(\theta_t)\mathrm{d}W_t \\ &= 2\sqrt{\eta}\,r_t\,\mathrm{d}W_t^{(r)}. \end{aligned} \tag{312}$$

The quadratic variation is

$$\begin{aligned}\langle W^{(r)}\rangle_t &= \int_0^t e_r(\theta_u)^{\top}G(\theta_u)G(\theta_u)^{\top}e_r(\theta_u)\mathrm{d}u \\ &\approx \int_0^t e_r(\theta_u)^{\top}e_r(\theta_u)\mathrm{d}u = t. \end{aligned} \tag{313}$$

Thus, by Lévy's characterization (Karatzas & Shreve, 1991), $W_t^{(r)}$ is a one-dimensional Brownian motion.

Substituting equations (310), (311), and (312) into equation (303) yields equation (297)

$$\mathrm{d}r_t^2 \approx \left[-\frac{2\lambda}{s(\theta_t)} r_t^2 - \frac{2}{s(\theta_t)} \theta_t^\top \left(\bar{g}_t - \lambda \theta_t\right) + \eta\, p + \mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t)\right] \mathrm{d}t + 2\sqrt{\eta}\, r_t\, \mathrm{d}W_t^{(r)}. \quad (314)$$

□

Remark 10. Empirical study in Figure 10 shows that $\mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t)$ is negligible during the memorization regime, so that the radius SDE in grokking can be approximated through

$$\mathrm{d}r_t^2 \approx \left[-\frac{2\lambda}{s(\theta_t)} r_t^2 - \frac{2}{s(\theta_t)} \theta_t^\top \left(\bar{g}_t - \lambda \theta_t\right) + \eta\, p\right] \mathrm{d}t + 2\sqrt{\eta}\, r_t\, \mathrm{d}W_t^{(r)}. \quad (315)$$

We now analyze the cancellation of $\mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t)$. Expanding $\mathcal{R}_{\mathrm{SM}}(t)$ gives

$$\begin{aligned}\mathcal{R}_{\mathrm{SM}}(t) = &-2B(t)\theta_t^\top \mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] \\ &+ 2B(t)\theta_t^\top \left[\bar{g}_t\left(1 - e^{-\alpha_1 t}\right) \oslash \sqrt{\overline{g \odot g}_t}\right]. \end{aligned} \quad (316)$$

Using

$$B(t)\left(1 - e^{-\alpha_1 t}\right) = a(t), \qquad \bar{g}_t \oslash \sqrt{\overline{g \odot g}_t} = \pi(\theta_t)\bar{g}_t, \quad (317)$$

we obtain

$$\mathcal{R}_{\mathrm{SM}}(t) = -2B(t)\theta_t^\top \mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] + 2a(t)\theta_t^\top \pi(\theta_t)\bar{g}_t. \quad (318)$$

Adding the explicit expression for $\mathcal{R}_\pi(t)$,

$$\begin{aligned}\mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t) &= -2B(t)\theta_t^\top \mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] + 2a(t)\theta_t^\top \pi(\theta_t)\bar{g}_t \\ &\quad - 2\theta_t^\top \left[\pi(\theta_t) - s(\theta_t)^{-1} I_p\right] \bar{g}_t \\ &= -2B(t)\theta_t^\top \mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] + 2\big(a(t) - 1\big)\theta_t^\top \pi(\theta_t)\bar{g}_t + \frac{2}{s(\theta_t)} \theta_t^\top \bar{g}_t. \end{aligned} \quad (319)$$

This proves equation (301). In the late-stage limit $a(t), B(t) \to 1$, the middle term vanishes, and therefore

$$\mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t) \approx -2B(t)\theta_t^\top \mathbb{E}\left[m_t \oslash \sqrt{v_t}\right] + \frac{2}{s(\theta_t)} \theta_t^\top \bar{g}_t. \quad (320)$$

Hence the residual sum is small when gradient and its first-moment estimate are near zeros.

A.12 Proof: Scaling Law of Memorization Radius

Theorem 4 (Scaling Law of Memorization Radius (Restated))**.** *Consider the joint Adam SDE in equation (167) of Lemma 5 (Adam's Joint-State Continuous-Time SDE Limit) on the joint state* $S_t = (\theta_t^\top, m_t^\top, v_t^\top)^\top \in \mathbb{R}^{3p}$,

$$\mathrm{d}S_t = \mu(S_t)\mathrm{d}t + \sqrt{\eta/b}\,\sigma(S_t)\mathrm{d}W_t, \quad (321)$$

and assume that the bias-correction factor satisfies $B(t) \to 1$ *in the early hitting regime. Let*

$$\tau_M(S_0) := \inf\{t \geq 0 : \theta_t(S_0) \in \partial M\} \quad (322)$$

be the first-hitting time of the memorization boundary. Define the exit-value function

$$u(S) := \mathbb{E}\left[\|\theta_{\tau_M(S_0)}\|_2^2 \,\middle|\, S_0 = S\right], \quad (323)$$

and the mean squared memorization radius

$$\rho_M^2 := \mathbb{E}_{S_0 \sim \Theta_S}[u(S_0)], \quad (324)$$

where Θ_S is the initialization distribution of the joint state, with $S_0 = (\theta_0^\top, 0_p^\top, 0_p^\top)^\top$.

Let $E_\theta \in \mathbb{R}^{3p\times 3p}$ be the θ-block projector

$$E_\theta := \begin{pmatrix} I_p & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{325}$$

so that for any joint state $S = (\theta, m, v)$,

$$\|\theta\|_2^2 = S^\top E_\theta S. \tag{326}$$

Let $\Phi_t(S_0)$ denote the deterministic Adam flow generated by the drift field $\mu(S)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(S_0) = \mu(\Phi_t(S_0)), \qquad \Phi_0(S_0) = S_0. \tag{327}$$

Define the deterministic first-hitting time

$$\tau_M^{(0)}(S_0) := \inf\{t \geq 0 : \theta(\Phi_t(S_0)) \in \partial M\}, \tag{328}$$

and the deterministic exit-radius-squared map

$$R^2(S_0) := \Phi_{\tau_M^{(0)}(S_0)}(S_0)^\top E_\theta \Phi_{\tau_M^{(0)}(S_0)}(S_0). \tag{329}$$

Let

$$\Sigma_S(S) := \sigma(S)\sigma(S)^\top \tag{330}$$

be the joint-state diffusion covariance. Then, for sufficiently small η/b,

$$\rho_M^2 = \rho_M^{(0)\,2} + (\eta/b)c_M + (\eta/b)^2 c_M^{(2)} + O\big((\eta/b)^3\big), \tag{331}$$

where

$$\rho_M^{(0)\,2} := \mathbb{E}_{S_0\sim\Theta_S}\left[R^2(S_0)\right], \tag{332}$$

$$c_M := \mathbb{E}_{S_0\sim\Theta_S}\int_0^{\tau_M^{(0)}(S_0)} \frac{1}{2}\operatorname{tr}\big(\Sigma_S(\Phi_t(S_0))\nabla_S^2 R^2(\Phi_t(S_0))\big)\,\mathrm{d}t, \tag{333}$$

$$c_M^{(2)} := \mathbb{E}_{S_0\sim\Theta_S}\int_0^{\tau_M^{(0)}(S_0)} \frac{1}{2}\operatorname{tr}\big(\Sigma_S(\Phi_t(S_0))\nabla_S^2 u_1(\Phi_t(S_0))\big)\,\mathrm{d}t. \tag{334}$$

Here u_1 is the first-order perturbation corrector

$$u_1(S_0) := \int_0^{\tau_M^{(0)}(S_0)} \frac{1}{2}\operatorname{tr}\big(\Sigma_S(\Phi_t(S_0))\nabla_S^2 R^2(\Phi_t(S_0))\big)\,\mathrm{d}t. \tag{335}$$

The constants $\rho_M^{(0)\,2}$, c_M, and $c_M^{(2)}$ are task-determined and independent of η and b; ℓ_2 regularization coefficient λ enters $(\rho_M^{(0)})^2$ through deterministic gradient flow with scaling law $(\rho_M^{(0)})^2 \propto \exp(-O(\lambda))$.

Proof. Let

$$\tau_M(S_0) := \inf\{t \geq 0 : \theta_t(S_0) \in \partial M\} \tag{336}$$

be the first-hitting time of the memorization boundary for the trajectory $\{\theta_t\}$ starting from $S_0 = (\theta_0^\top, 0_p^\top, 0_p^\top)^\top$.

Bilinear Radius Projector. Since the parameter θ_t is the θ-component of the joint state S_t, define the θ-block projector

$$E_\theta := \begin{pmatrix} I_p & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{3p\times 3p}. \tag{337}$$

Thus, for any joint state S, the squared parameter radius is the joint-state quadratic form

$$r^2(S) := \|\theta\|_2^2 = S^\top E_\theta S. \tag{338}$$

In particular,

$$\nabla_S r^2(S) = 2E_\theta S, \qquad \nabla_S^2 r^2(S) = 2E_\theta. \tag{339}$$

Define Dirichlet Boundary Problem. We introduce the exit-value function

$$u(S) := \mathbb{E}\big[r^2(S_{\tau_M}) \,\big|\, S_0 = S\big] = \mathbb{E}\big[S_{\tau_M}^\top E_\theta S_{\tau_M} \,\big|\, S_0 = S\big]\,, \tag{340}$$

which is the expected squared parameter norm at the first-hitting time of ∂M, starting from S.

We now derive the PDE satisfied by u. The joint Adam SDE has the form

$$\mathrm{d}S_t = \mu(S_t)\,\mathrm{d}t + \sqrt{\eta/b}\,\sigma(S_t)\,\mathrm{d}W_t. \tag{341}$$

Therefore its infinitesimal generator is

$$\mathcal{L}_S f(S) = \mu(S)^\top \nabla_S f(S) + \frac{1}{2}\boxed{\frac{\eta}{b}}\operatorname{tr}\big(\Sigma_S(S)\nabla_S^2 f(S)\big)\,, \qquad \Sigma_S(S) := \sigma(S)\sigma(S)^\top. \tag{342}$$

Learning rate η and batch size b modulate infinitesimal generator

Remark 11. The structure of the infinitesimal generator $\mathcal{L}_S f(S)$ shows that this operator contains a term modulated by the learning rate η and the batch size b through the coefficient η/b. This motivates us to treat $\mathcal{L}_S f(S)$ as an operator expansion with respect to the coefficient η/b, so that we can apply a regular perturbation expansion to study the corresponding expansion of the memorization radius with respect to η/b.

Dirichlet Boundary PDE. Since $u(S)$ is the expected boundary value of the process stopped at ∂M, it solves the Dirichlet boundary value problem (Øksendal, 2003)

$$\mathcal{L}_S u(S) = 0, \qquad \theta \in M^\circ, \tag{343}$$

$$u(S) = S^\top E_\theta S, \qquad \theta \in \partial M. \tag{344}$$

Here equation (343) states that $u(S_t)$ is harmonic with respect to the stopped Adam diffusion before hitting the boundary, while equation (344) assigns the squared radius as the boundary payoff.

Let $\varepsilon := \eta/b$. We split the generator into its deterministic and stochastic parts:

$$\mathcal{L}_S = \mathcal{L}_S^{(0)} + \varepsilon\mathcal{L}_S^{(1)}, \qquad \mathcal{L}_S^{(0)} f := \mu^\top \nabla_S f, \qquad \mathcal{L}_S^{(1)} f := \frac{1}{2}\operatorname{tr}\big(\Sigma_S \nabla_S^2 f\big)\,. \tag{345}$$

Thus the Dirichlet PDE becomes

$$\left(\mathcal{L}_S^{(0)} + \varepsilon\mathcal{L}_S^{(1)}\right) u(S) = 0, \qquad \theta \in M^\circ, \tag{346}$$

with boundary condition

$$u(S) = S^\top E_\theta S, \qquad \theta \in \partial M. \tag{347}$$

Solving PDE via Regular Perturbation Expansion. We solve this PDE by a regular perturbation expansion in the small parameter ε. Write

$$u(S) = u_0(S) + \varepsilon u_1(S) + \varepsilon^2 u_2(S) + O(\varepsilon^3). \tag{348}$$

Substituting equation (348) into equation (346) gives

$$\begin{aligned} 0 &= \left(\mathcal{L}_S^{(0)} + \varepsilon\mathcal{L}_S^{(1)}\right)\left(u_0 + \varepsilon u_1 + \varepsilon^2 u_2 + O(\varepsilon^3)\right) \\ &= \mathcal{L}_S^{(0)} u_0 + \varepsilon\left(\mathcal{L}_S^{(0)} u_1 + \mathcal{L}_S^{(1)} u_0\right) + \varepsilon^2\left(\mathcal{L}_S^{(0)} u_2 + \mathcal{L}_S^{(1)} u_1\right) + O(\varepsilon^3). \end{aligned} \tag{349}$$

Since this identity must hold for all sufficiently small ε, each coefficient of ε must vanish. Therefore,

$$\varepsilon^0: \quad \mathcal{L}_S^{(0)} u_0 = 0, \tag{350}$$

$$\varepsilon^1: \quad \mathcal{L}_S^{(0)} u_1 = -\mathcal{L}_S^{(1)} u_0, \tag{351}$$

$$\varepsilon^2: \quad \mathcal{L}_S^{(0)} u_2 = -\mathcal{L}_S^{(1)} u_1. \tag{352}$$

The boundary condition is expanded in the same way:

$$u_0 + \varepsilon u_1 + \varepsilon^2 u_2 + O(\varepsilon^3) = S^\top E_\theta S, \qquad \theta \in \partial M. \tag{353}$$

Matching powers of ε on the boundary gives

$$u_0\big|_{\theta\in\partial M} = S^\top E_\theta S, \tag{354}$$

$$u_1\big|_{\theta\in\partial M} = 0, \tag{355}$$

$$u_2\big|_{\theta\in\partial M} = 0. \tag{356}$$

Zero-Order Expansion. We now solve these equations by characteristics. The characteristic curves of $\mathcal{L}_S^{(0)} = \mu^\top \nabla_S$ are the deterministic Adam trajectories

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(S_0) = \mu(\Phi_t(S_0)), \qquad \Phi_0(S_0) = S_0, \tag{357}$$

so that

$$\begin{aligned}
\dot\theta^{(0)} &= -m^{(0)} \oslash \sqrt{v^{(0)}},\\
\dot m^{(0)} &= -\alpha_1\big(m^{(0)} - \bar g(\theta^{(0)})\big),\\
\dot v^{(0)} &= -\alpha_2\big(v^{(0)} - \overline{g\odot g}(\theta^{(0)})\big).
\end{aligned} \tag{358}$$

For any smooth function w, along the deterministic flow we have

$$\frac{\mathrm{d}}{\mathrm{d}t} w(\Phi_t(S_0)) = \mathcal{L}_S^{(0)} w(\Phi_t(S_0)). \tag{359}$$

Let $\tau_M^{(0)}(S_0)$ be the deterministic hitting time

$$\tau_M^{(0)}(S_0) := \inf\{t \ge 0 : (\Phi_t(S_0))_\theta \in \partial M\}. \tag{360}$$

At order ε^0, equation (350) gives

$$\frac{\mathrm{d}}{\mathrm{d}t} u_0(\Phi_t(S_0)) = 0. \tag{361}$$

Thus u_0 is constant along the deterministic trajectory. Evaluating it at the deterministic hitting time and using the boundary condition in equation (354), we obtain

$$\begin{aligned}
u_0(S_0) &= u_0(\Phi_{\tau_M^{(0)}(S_0)}(S_0))\\
&= \Phi_{\tau_M^{(0)}(S_0)}(S_0)^\top E_\theta \Phi_{\tau_M^{(0)}(S_0)}(S_0)\\
&= \Big\| \big(\Phi_{\tau_M^{(0)}(S_0)}(S_0)\big)_\theta \Big\|_2^2 .
\end{aligned} \tag{362}$$

Define the deterministic exit-radius-squared map

$$R^2(S_0) := \Phi_{\tau_M^{(0)}(S_0)}(S_0)^\top E_\theta \Phi_{\tau_M^{(0)}(S_0)}(S_0). \tag{363}$$

Equivalently,

$$R^2(S_0) = \Big\| \big(\Phi_{\tau_M^{(0)}(S_0)}(S_0)\big)_\theta \Big\|_2^2 . \tag{364}$$

Then $u_0(S_0) = R^2(S_0)$.

First-Order Expansion. At order ε^1, equation (351) and the transport identity give

$$\frac{\mathrm{d}}{\mathrm{d}t} u_1(\Phi_t(S_0)) = -\mathcal{L}_S^{(1)} u_0(\Phi_t(S_0)). \tag{365}$$

Integrating from 0 to $\tau_M^{(0)}(S_0)$ gives

$$u_1(\Phi_{\tau_M^{(0)}(S_0)}(S_0)) - u_1(S_0) = -\int_0^{\tau_M^{(0)}(S_0)} \mathcal{L}_S^{(1)} u_0(\Phi_t(S_0))\,\mathrm{d}t. \tag{366}$$

Since $u_1 = 0$ on ∂M, the first term on the left vanishes. Therefore,

$$u_1(S_0) = \int_0^{\tau_M^{(0)}(S_0)} \mathcal{L}_S^{(1)} u_0(\Phi_t(S_0))\,\mathrm{d}t. \tag{367}$$

Using the definition of $\mathcal{L}_S^{(1)}$ and $u_0 = R^2$ along the deterministic flow,

$$u_1(S_0) = \int_0^{\tau_M^{(0)}(S_0)} \frac{1}{2}\,\mathrm{tr}\big(\Sigma_S(\Phi_t(S_0))\nabla_S^2 R^2(\Phi_t(S_0))\big)\,\mathrm{d}t. \tag{368}$$

Second-Order Expansion. At order ε^2, the same argument gives

$$\frac{\mathrm{d}}{\mathrm{d}t} u_2(\Phi_t(S_0)) = -\mathcal{L}_S^{(1)} u_1(\Phi_t(S_0)). \tag{369}$$

Since $u_2 = 0$ on ∂M, integrating along the deterministic characteristic yields

$$u_2(S_0) = \int_0^{\tau_M^{(0)}(S_0)} \mathcal{L}_S^{(1)} u_1(\Phi_t(S_0))\,\mathrm{d}t. \tag{370}$$

Thus,

$$u_2(S_0) = \int_0^{\tau_M^{(0)}(S_0)} \frac{1}{2}\,\mathrm{tr}\big(\Sigma_S(\Phi_t(S_0))\nabla_S^2 u_1(\Phi_t(S_0))\big)\,\mathrm{d}t. \tag{371}$$

Produce Claims. Combining the three orders, we have

$$u(S_0) = R^2(S_0) + \varepsilon u_1(S_0) + \varepsilon^2 u_2(S_0) + O(\varepsilon^3). \tag{372}$$

Finally, the mean squared memorization radius is obtained by averaging this exit-value function over the initialization distribution:

$$\begin{aligned}\rho_M^2 &:= \mathbb{E}\big[\|\theta_{\tau_M}\|_2^2\big] = \mathbb{E}\big[S_{\tau_M}^\top E_\theta S_{\tau_M}\big] = \mathbb{E}_{S_0\sim\Theta_S}[u(S_0)] \\ &= \mathbb{E}_{S_0\sim\Theta_S}\big[R^2(S_0)\big] + \varepsilon\mathbb{E}_{S_0\sim\Theta_S}[u_1(S_0)] + \varepsilon^2\mathbb{E}_{S_0\sim\Theta_S}[u_2(S_0)] + O(\varepsilon^3).\end{aligned} \tag{373}$$

Define

$$\begin{aligned}\rho_M^{(0)\,2} &:= \mathbb{E}_{S_0\sim\Theta_S}\big[R^2(S_0)\big]\,, \\ c_M &:= \mathbb{E}_{S_0\sim\Theta_S}[u_1(S_0)]\,, \\ c_M^{(2)} &:= \mathbb{E}_{S_0\sim\Theta_S}[u_2(S_0)]\,.\end{aligned} \tag{374}$$

Substituting these definitions into equation (373) gives

$$\rho_M^2 = \rho_M^{(0)\,2} + \varepsilon c_M + \varepsilon^2 c_M^{(2)} + O(\varepsilon^3). \tag{375}$$

Since $\varepsilon = \eta/b$, we obtain

$$\rho_M^2 = \rho_M^{(0)\,2} + (\eta/b)c_M + (\eta/b)^2 c_M^{(2)} + O\big((\eta/b)^3\big), \tag{376}$$

which proves the theorem. □

Dependence of $\rho_M^{(0)}$ on λ under weak task gradient. The leading constant $(\rho_M^{(0)})^2$ is determined by the deterministic gradient flow. In the regime where the task gradient is small,

$$\nabla\mathscr{L}_f^*(\theta_t) \approx 0, \tag{377}$$

the deterministic flow is dominated by the ℓ_2 drift:

$$\mathrm{d}\theta_t \approx -\lambda\theta_t\,\mathrm{d}t. \tag{378}$$

Hence

$$\begin{aligned}\mathrm{d}\|\theta_t\|_2^2 &= 2\theta_t^\top \mathrm{d}\theta_t \\ &\approx -2\lambda\|\theta_t\|_2^2\,\mathrm{d}t.\end{aligned} \tag{379}$$

Equivalently,

$$\frac{1}{\|\theta_t\|_2^2}\mathrm{d}\|\theta_t\|_2^2 \approx -2\lambda\,\mathrm{d}t. \tag{380}$$

Integrating along the deterministic memorization path gives

$$\|\theta_{\tau_M^{(0)}}\|_2^2 \approx \|\theta_0\|_2^2 \exp\Big(-2\lambda\tau_M^{(0)}\Big)\,. \tag{381}$$

Therefore,

$$(\rho_M^{(0)})^2 \propto \exp\Big(-O(\lambda\tau_M^{(0)})\Big)\,. \tag{382}$$

When $\tau_M^{(0)}$ is weakly dependent on λ in the considered regime,

$$(\rho_M^{(0)})^2 \propto \exp(-O(\lambda))\,. \tag{383}$$

Since

$$\rho_M^2 = (\rho_M^{(0)})^2 + O(\eta), \tag{384}$$

the same leading dependence gives

$$\rho_M^2 \propto \exp(-O(\lambda))\,. \tag{385}$$

A.13 Proof: Scaling Law of Generalization Radius

Theorem 5 (Scaling Law of Generalization Radius, restated)**.** *We do not assume a global minizer, for each initialization $S_0 \in \Theta_S$, let*

$$\theta^\star(S_0) := \arg\min_{\theta(S_0)} \left\{ \mathscr{L}_f^*(\theta(S_0)) + \frac{\lambda}{2}\|\theta(S_0)\|_2^2 \right\} \in G \tag{386}$$

denote the regularized local minimizer selected by the trajectory starting from initial state S_0, and $\mathscr{L}_f^*(\theta(S_0))$ is the task loss. The trajectory $\{S_t\}$ will be confined in the basin centered at $\theta^\star(S_0)$. Define

$$\left(\rho_G^{(0)}\right)^2 := \mathbb{E}_{S_0 \in \Theta_S}\left[\|\theta^\star(S_0)\|_2^2\right]. \tag{387}$$

Let

$$\tau_G(S_0) := \inf\{t \geq 0 : \theta_t(S_0) \in \partial G\}, \tag{388}$$

and define the generalization radius by

$$\rho_G^2 := \mathbb{E}_{S_0 \in \Theta_S}\left[\mathbb{E}\left[\|\theta_{\tau_G(S_0)}\|_2^2 \mid S_0\right]\right]. \tag{389}$$

Let

$$s^\star(S_0) := s(\theta^\star(S_0)), \qquad D^\star(S_0) := G(\theta^\star(S_0))G(\theta^\star(S_0))^\top, \tag{390}$$

where the effective scalar preconditioner at the selected minimizer is defined by

$$\frac{1}{s^\star(S_0)} = \frac{1}{p}\mathrm{tr}\big(\pi(\theta^\star(S_0))\big), \qquad \pi(\theta^\star(S_0)) = \mathrm{diag}(\overline{g \odot g}(\theta^\star(S_0)))^{-1/2}. \tag{391}$$

Then, generalization radius approximately admits,

$$\rho_G^2 \approx \left(\rho_G^{(0)}\right)^2 + \frac{\eta}{\lambda}\, c_G + O\left(\frac{\eta^2}{\lambda^2}\right), \tag{392}$$

where

$$\|\theta^\star(S_0)\|_2^2 \propto \exp(-O(\lambda)) \tag{393}$$

and

$$c_G := \mathbb{E}_{S_0 \in \Theta_S}\left[\tfrac{1}{2} s^\star(S_0)\,\mathrm{tr}(D^\star(S_0))\right] > 0. \tag{394}$$

Proof. For a trajectory $\{S_t\}$ starting from $S_0 \in \Theta_S$, write

$$\theta^\star := \theta^\star(S_0), \qquad \theta_t = \theta^\star + \delta_t. \tag{395}$$

By definition of the local minimizer,

$$\bar{g}^*(\theta^\star) + \lambda\theta^\star = 0, \tag{396}$$

where $\bar{g}^*(\bullet)$ denotes the task loss mean. By Lemma 11 (Reduced Late-Stage Radius SDE with Slow-Manifold, Preconditioned Residual (Restated)), together with Lemma 10 (Memorization-Regime Preconditioner and Effective Diffusion Identities),

$$G(\theta_t)G(\theta_t)^\top \approx I_p, \tag{397}$$

and the residual

$$\mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t) \approx 0, \tag{398}$$

the reduced late-stage squared-radius SDE gives

$$\mathrm{d}r_t^2 \approx \left[-\frac{2}{s(\theta_t)}\theta_t^\top(\bar{g}_t^* + \lambda\theta_t) + \eta p\right]\mathrm{d}t + 2\sqrt{\eta}\, r_t\, \mathrm{d}W_t^{(r)}, \tag{399}$$

where

$$r_t^2 = \|\theta_t\|_2^2. \tag{400}$$

Linearization Near Local Minizer. The Taylor expansion of $\bar{g}^*$ around $\theta^\star$ gives

$$\begin{aligned}\bar{g}_t^* &= \bar{g}^*(\theta^\star + \delta_t) \\ &= \bar{g}^*(\theta^\star) + \nabla\bar{g}^*(\theta^\star)\delta_t + O\big(\|\delta_t\|_2^2\big)\,. \end{aligned} \tag{401}$$

Therefore,

$$\begin{aligned}\bar{g}_t^* + \lambda\theta_t &= \bar{g}^*(\theta^\star) + \nabla\bar{g}^*(\theta^\star)\delta_t + \lambda\theta^\star + \lambda\delta_t + O\big(\|\delta_t\|_2^2\big) \\ &= \left(\nabla\bar{g}^*(\theta^\star) + \lambda I_p\right)\delta_t + O\big(\|\delta_t\|_2^2\big)\,, \end{aligned} \tag{402}$$

where equation (396) was used. Under the dominant-regularization approximation,

$$\nabla\bar{g}^*(\theta^\star) + \lambda I_p \approx \lambda I_p, \tag{403}$$

$$\bar{g}_t^* + \lambda\theta_t \approx \lambda\delta_t + O\big(\|\delta_t\|_2^2\big)\,. \tag{404}$$

Using $\theta_t = \theta^\star + \delta_t$,

$$\begin{aligned}\theta_t^\top\left(\bar{g}_t^* + \lambda\theta_t\right) &\approx (\theta^\star + \delta_t)^\top\left[\lambda\delta_t + O\big(\|\delta_t\|_2^2\big)\right] \\ &= \lambda\theta^{\star\top}\delta_t + \lambda\|\delta_t\|_2^2 + O\big(\|\delta_t\|_2^2\big)\,. \end{aligned} \tag{405}$$

Local Stationary Approximation. Under the local stationary approximation,

$$s(\theta_t) = s^\star(S_0) + O(\|\delta_t\|_2)\,, \qquad s^\star(S_0) := s(\theta^\star(S_0)). \tag{406}$$

Taking conditional expectation in equation (399) gives

$$0 \approx -\frac{2}{s^\star(S_0)}\mathbb{E}\left[\theta_t^\top\left(\bar{g}_t^* + \lambda\theta_t\right) \mid S_0\right] + \eta p + O\left(\frac{\eta^2}{\lambda^2}\right), \tag{407}$$

since

$$\mathbb{E}\left[r_t\,\mathrm{d}W_t^{(r)} \mid S_0\right] = 0. \tag{408}$$

The local stationary fluctuation is centered, so

$$\mathbb{E}[\delta_t \mid S_0] = 0, \qquad \mathbb{E}[\theta^{\star\top}\delta_t \mid S_0] = 0. \tag{409}$$

Combining equations (405), (407), and (409),

$$0 \approx -\frac{2\lambda}{s^\star(S_0)}\mathbb{E}\left[\|\delta_t\|_2^2 \mid S_0\right] + \eta p + O\left(\frac{\eta^2}{\lambda^2}\right). \tag{410}$$

Hence

$$\mathbb{E}\left[\|\delta_t\|_2^2 \mid S_0\right] \approx \frac{\eta}{\lambda}\frac{s^\star(S_0)}{2}p + O\left(\frac{\eta^2}{\lambda^2}\right). \tag{411}$$

Also,

$$\begin{aligned}\|\theta_t\|_2^2 &= \|\theta^\star + \delta_t\|_2^2 \\ &= \|\theta^\star\|_2^2 + 2\theta^{\star\top}\delta_t + \|\delta_t\|_2^2. \end{aligned} \tag{412}$$

Taking conditional expectation gives

$$\begin{aligned}\mathbb{E}\left[\|\theta_t\|_2^2 \mid S_0\right] &= \|\theta^\star(S_0)\|_2^2 + \mathbb{E}\left[\|\delta_t\|_2^2 \mid S_0\right] \\ &\approx \|\theta^\star(S_0)\|_2^2 + \frac{\eta}{\lambda}\frac{s^\star(S_0)}{2}p + O\left(\frac{\eta^2}{\lambda^2}\right). \end{aligned} \tag{413}$$

Solving Generalization Radius. Under the local stationary approximation in the generalization basin,

$$\mathbb{E}\left[\|\theta_{\tau_G(S_0)}\|_2^2 \mid S_0\right] \approx \mathbb{E}\left[\|\theta_t\|_2^2 \mid S_0\right]. \tag{414}$$

Therefore,

$$\mathbb{E}\left[\|\theta_{\tau_G(S_0)}\|_2^2 \mid S_0\right] \approx \|\theta^\star(S_0)\|_2^2 + \frac{\eta}{\lambda}\frac{s^\star(S_0)}{2}p + O\left(\frac{\eta^2}{\lambda^2}\right). \tag{415}$$

Averaging over $S_0 \in \Theta_S$,

$$\begin{aligned}\rho_G^2 &:= \mathbb{E}_{S_0\in\Theta_S}\left[\mathbb{E}\left[\|\theta_{\tau_G}\|_2^2 \mid S_0\right]\right] \\ &\approx \mathbb{E}_{S_0\in\Theta_S}\left[\|\theta^\star(S_0)\|_2^2\right] + \frac{\eta}{\lambda}\mathbb{E}_{S_0\in\Theta_S}\left[\frac{s^\star(S_0)}{2}p\right] + O\left(\frac{\eta^2}{\lambda^2}\right).\end{aligned} \tag{416}$$

By definition,

$$\left(\rho_G^{(0)}\right)^2 := \mathbb{E}_{S_0\in\Theta_S}\left[\|\theta^\star(S_0)\|_2^2\right], \tag{417}$$

$$c_G := \mathbb{E}_{S_0\in\Theta_S}\left[\frac{s^\star(S_0)}{2}p\right]. \tag{418}$$

Thus,

$$\rho_G^2 \approx \left(\rho_G^{(0)}\right)^2 + \frac{\eta}{\lambda}c_G + O\left(\frac{\eta^2}{\lambda^2}\right), \tag{419}$$

which proves the claimed scaling law.

Regularizer λ-Dependence of $\rho_G^{(0)}$. By definition,

$$\left(\rho_G^{(0)}\right)^2 := \mathbb{E}_{S_0\in\Theta_S}\left[\|\theta^\star(S_0)\|_2^2\right]. \tag{420}$$

The dependence on λ enters $(\rho_G^{(0)})^2$ through the deterministic regularized trajectory selecting the local minimizer $\theta^\star(S_0)$. Along the deterministic component of the regularized flow,

$$\mathrm{d}\theta_t \approx -\left(\nabla\mathscr{L}_f^*(\theta_t) + \lambda\theta_t\right)\mathrm{d}t. \tag{421}$$

When the task-gradient contribution is small in the generalization basin, the radial dynamics satisfy

$$\mathrm{d}\|\theta_t\|_2^2 = 2\theta_t^\top \mathrm{d}\theta_t \approx -2\lambda\|\theta_t\|_2^2\,\mathrm{d}t. \tag{422}$$

Therefore,

$$\frac{1}{\|\theta_t\|_2^2}\mathrm{d}\|\theta_t\|_2^2 \approx -2\lambda\,\mathrm{d}t, \tag{423}$$

and integration along the deterministic approach to the generalization basin gives

$$\|\theta^\star(S_0)\|_2^2 \propto \exp(-O(\lambda)). \tag{424}$$

Averaging over $S_0 \in \Theta_S$, we obtain

$$\left(\rho_G^{(0)}\right)^2 = \mathbb{E}_{S_0\in\Theta_S}\left[\|\theta^\star(S_0)\|_2^2\right] \propto \exp(-O(\lambda)). \tag{425}$$

Batch-Size b-Dependence of c_G. The weak batch-size dependence of c_G enters through the effective scalar preconditioner $s^\star(S_0)$. From the definition of c_G,

$$c_G := \mathbb{E}_{S_0\in\Theta_S}\left[\frac{s^\star(S_0)}{2}p\right]. \tag{426}$$

By the late-stage preconditioning identity,

$$\frac{1}{s^\star(S_0)} = \frac{1}{p}\mathrm{tr}\left(\pi(\theta^\star(S_0))\right), \tag{427}$$

$$\pi(\theta^\star(S_0)) = \mathrm{diag}\left(\overline{g\odot g}(\theta^\star(S_0))\right)^{-1/2}. \tag{428}$$

In the memorization regime,

$$\overline{g\odot g}(\theta^\star(S_0)) \approx \frac{1}{b}\mathrm{diag}(\Sigma(\theta^\star(S_0))). \tag{429}$$

Therefore,

$$\pi(\theta^\star(S_0)) \approx \sqrt{b}\,\mathrm{diag}\left(\mathrm{diag}(\Sigma(\theta^\star(S_0)))\right)^{-1/2}, \tag{430}$$

$$\frac{1}{s^\star(S_0)} = O(\sqrt{b}), \tag{431}$$

$$s^\star(S_0) = O\left(\frac{1}{\sqrt{b}}\right). \tag{432}$$

Using the late-stage identity $G(\theta)G(\theta)^\top \approx I_p$, the diffusion trace contribution is $\mathrm{tr}(G(\theta)G(\theta)^\top) \approx p$. Hence

$$c_G = \mathbb{E}_{S_0 \in \Theta_S}\left[\frac{s^\star(S_0)}{2} p\right] = O\left(\frac{1}{\sqrt{b}}\right). \tag{433}$$

□

A.14 Proof: Scaling Law of Solution Transition Time

Theorem 6 (Scaling Law of Solution Transition Time, restated)**.** *Assume the reduced late-stage radius SDE in Lemma 11 (Reduced Late-Stage Radius SDE with Slow-Manifold, Preconditioned Residual (Restated)), with conditions in Lemma 10 (Memorization-Regime Preconditioner and Effective Diffusion Identities)*

$$G(\theta_t)G(\theta_t)^\top \approx I_p, \tag{434}$$

and the residual approximation

$$\mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t) \approx 0. \tag{435}$$

Let $\bar{g}(\theta)$ be the population mean gradient and let $s(\theta) > 0$ denote the scalar preconditioning factor satisfying

$$\frac{1}{s(\theta)} = \frac{1}{p}\mathrm{tr}\big(\pi(\theta)\big). \tag{436}$$

Let $\theta_t^{(0)}(S_0)$ be the deterministic late-stage trajectory from ∂M to ∂G starting with initial state S_0, and let $\tau^{(0)}_{M\to G}(S_0)$ be its deterministic transition time. For each deterministic late-stage trajectory $\theta_t^{(0)}(S_0)$ from ∂M to ∂G, define its pathwise effective radial contraction rate by

$$\kappa(S_0) := \frac{1}{\tau^{(0)}_{M\to G}(S_0)} \int_0^{\tau^{(0)}_{M\to G}(S_0)} \left[\frac{\lambda}{s(\theta_t^{(0)}(S_0))} + \frac{\theta_t^{(0)\top}(S_0)\big(\bar{g}(\theta_t^{(0)}(S_0)) - \lambda\theta_t^{(0)}(S_0)\big)}{s(\theta_t^{(0)}(S_0))\|\theta_t^{(0)}(S_0)\|_2^2} \right] \mathrm{d}t. \tag{437}$$

The scalar $\bar{s}$ is defined through the ensemble-averaged effective contraction rate

$$\frac{\lambda}{\bar{s}} := \mathbb{E}_{S_0\in\Theta_S}\left[\kappa(S_0)\right]. \tag{438}$$

Let ρ_M and ρ_G be the memorization and generalization radii. Assume the asymptotic expansions

$$\rho_M^2 = (\rho_M^{(0)})^2 + \frac{\eta}{b}c_M + O\big((\eta/b)^2\big), \tag{439}$$

$$\rho_G^2 \approx \big(\rho_G^{(0)}\big)^2 + \frac{\eta}{\lambda}c_G + O\left(\frac{\eta^2}{\lambda^2}\right), \tag{440}$$

where c_M is the memorization-radius correction from Theorem 4, and

$$\big(\rho_G^{(0)}\big)^2 := \mathbb{E}_{S_0\in\Theta_S}\left[\|\theta^\star(S_0)\|_2^2\right], \tag{441}$$

$$\theta^\star(S_0) := \arg\min_{\theta(S_0)} \left\{ \mathscr{L}_f^*(\theta(S_0)) + \frac{\lambda}{2}\|\theta(S_0)\|_2^2 \right\}, \tag{442}$$

$$c_G := \mathbb{E}_{S_0\in\Theta_S}\left[\frac{s^\star(S_0)}{2}\mathrm{tr}(D^\star(S_0))\right], \tag{443}$$

$$s^\star(S_0) := s(\theta^\star(S_0)), \qquad D^\star(S_0) := G(\theta^\star(S_0))G(\theta^\star(S_0))^\top. \tag{444}$$

Under the late-stage identity $G(\theta^\star(S_0))G(\theta^\star(S_0))^\top \approx I_p$, equation (443) reduces to

$$c_G = \mathbb{E}_{S_0\in\Theta_S}\left[\frac{s^\star(S_0)}{2}p\right]. \tag{445}$$

Then, in iteration time, the expected transition time from ∂M to ∂G satisfies

$$\mathbb{E}_{S_0\in\Theta_S}\big[\tau_{M\to G}\big] \approx \frac{\bar{s}}{\eta\,\lambda}\log\frac{\rho_M^{(0)}}{\rho_G^{(0)}} + \frac{c_\tau}{b\,\lambda} + \frac{c_\tau^{(2)}}{\lambda^2} + O\Big(\frac{\eta}{\lambda^3}\Big), \tag{446}$$

where

$$c_\tau := \bar{s}\,\frac{c_M}{2(\rho_M^{(0)})^2}, \tag{447}$$

$$c_\tau^{(2)} := -\bar{s}\,\frac{c_G}{2(\rho_G^{(0)})^2} + \frac{\bar{s}^2(p-2)}{4}\left[\frac{1}{(\rho_G^{(0)})^2} - \frac{1}{(\rho_M^{(0)})^2}\right]. \tag{448}$$

The constants $\bar{s}$, c_τ, and $c_\tau^{(2)}$ are independent of η and λ in the considered scaling regime; a weak batch-size b-dependence enters through the scalar preconditioning scale $\bar{s} = O(\frac{1}{\sqrt{b}})$, and hence $c_\tau = O(\frac{1}{\sqrt{b}})$ and $c_\tau^{(2)} = O(\frac{1}{b})$.

Proof. Let

$$r_t := \|\theta_t\|_2. \tag{449}$$

By Lemma 11 (Reduced Late-Stage Radius SDE with Slow-Manifold, Preconditioned Residual (Restated)), under the late-stage observations

$$a(t) \to 1, \qquad G(\theta_t)G(\theta_t)^\top \approx I_p, \qquad \mathcal{R}_{\mathrm{SM}}(t) + \mathcal{R}_\pi(t) \approx 0, \tag{450}$$

the squared-radius process satisfies

$$\mathrm{d}r_t^2 \approx \left[-\frac{2\lambda}{s(\theta_t)} r_t^2 - \frac{2}{s(\theta_t)} \theta_t^\top \left(\bar{g}_t - \lambda\theta_t\right) + \eta\, p\right] \mathrm{d}t + 2\sqrt{\eta}\, r_t\, \mathrm{d}W_t^{(r)}. \tag{451}$$

Effective Deterministic Radius Flow. The leading deterministic radial flow is obtained from equation (451) by keeping the $O(1)$ drift terms and dropping the $O(\eta)$ Itô correction. Thus,

$$\mathrm{d}r_t^2 \approx \left[-\frac{2\lambda}{s(\theta_t)} r_t^2 - \frac{2}{s(\theta_t)} \theta_t^\top \left(\bar{g}_t - \lambda\theta_t\right)\right] \mathrm{d}t. \tag{452}$$

Since

$$\mathrm{d}r_t^2 = 2r_t\, \mathrm{d}r_t + O(\eta)\mathrm{d}t, \tag{453}$$

the leading deterministic radial drift is

$$\begin{aligned} \dot{r}_t &\approx -\frac{\lambda}{s(\theta_t)} r_t - \frac{1}{s(\theta_t) r_t} \theta_t^\top \left(\bar{g}_t - \lambda\theta_t\right) \\ &= -\left[\frac{\lambda}{s(\theta_t)} + \frac{\theta_t^\top \left(\bar{g}_t - \lambda\theta_t\right)}{s(\theta_t)\|\theta_t\|_2^2}\right] r_t. \end{aligned} \tag{454}$$

For each deterministic late-stage trajectory $\theta_t^{(0)}(S_0)$ from ∂M to ∂G, define its pathwise effective radial contraction rate by

$$\kappa(S_0) := \frac{1}{\tau_{M\to G}^{(0)}(S_0)} \int_0^{\tau_{M\to G}^{(0)}(S_0)} \left[\frac{\lambda}{s(\theta_t^{(0)}(S_0))} + \frac{\theta_t^{(0)\top}(S_0)\left(\bar{g}(\theta_t^{(0)}(S_0)) - \lambda\theta_t^{(0)}(S_0)\right)}{s(\theta_t^{(0)}(S_0))\|\theta_t^{(0)}(S_0)\|_2^2}\right] \mathrm{d}t. \tag{455}$$

The scalar $\bar{s}$ is defined through the ensemble-averaged effective contraction rate

$$\frac{\lambda}{\bar{s}} := \mathbb{E}_{S_0 \in \Theta_S}\left[\kappa(S_0)\right]. \tag{456}$$

Thus, in the averaged effective radial model, the leading deterministic late-stage radius flow is

$$\dot{r}_t^{(0)} = -\frac{\lambda}{\bar{s}} r_t^{(0)}. \tag{457}$$

Therefore, along the averaged late-stage transition dynamics,

$$-\frac{\lambda}{s(\theta_t)} r_t - \frac{1}{s(\theta_t) r_t} \theta_t^\top \left(\bar{g}_t - \lambda\theta_t\right) \approx -\frac{\lambda}{\bar{s}} r_t. \tag{458}$$

Effective Radius SDE. We now pass from the squared-radius SDE to the radius SDE. Applying Itô's lemma to $r_t = (r_t^2)^{1/2}$ gives

$$\mathrm{d}r_t = \frac{1}{2r_t} \mathrm{d}r_t^2 - \frac{1}{8r_t^3} \mathrm{d}\langle r^2\rangle_t. \tag{459}$$

From the martingale term in equation (451),

$$\mathrm{d}\langle r^2\rangle_t = 4\eta r_t^2\,\mathrm{d}t. \tag{460}$$

Substituting equations (451) and (460) into equation (459), and using equation (458), yields

$$\begin{aligned}\mathrm{d}r_t &= \left[-\frac{\lambda}{\bar{s}}r_t + \eta\frac{p}{2r_t} - \eta\frac{1}{2r_t}\right]\mathrm{d}t + \sqrt{\eta}\,\mathrm{d}W_t^{(r)} \\ &= \left[-\frac{\lambda}{\bar{s}}r_t + \eta\frac{p-1}{2r_t}\right]\mathrm{d}t + \sqrt{\eta}\,\mathrm{d}W_t^{(r)}.\end{aligned} \tag{461}$$

Let

$$\kappa := \frac{\lambda}{\bar{s}}, \qquad \varepsilon := \eta. \tag{462}$$

Then equation (461) becomes

$$\mathrm{d}r_t = \left[-\kappa r_t + \frac{\varepsilon(p-1)}{2r_t}\right]\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}W_t^{(r)}. \tag{463}$$

Mean First-Passage Time PDE. Let $T(r)$ be the expected continuous time for the process in equation (463), initialized at radius r, to hit the absorbing boundary r_G, defined as

$$T(r) := \mathbb{E}[\tau_G \mid r_0 = r], \qquad r > r_G. \tag{464}$$

The transition starts at the memorization boundary and ends at the generalization boundary:

$$r_0 := \rho_M, \qquad r_G := \rho_G. \tag{465}$$

Its infinitesimal generator is

$$\mathcal{L}_r[\bullet] = \left(-\kappa r + \frac{\varepsilon(p-1)}{2r}\right)\frac{\mathrm{d}}{\mathrm{d}r}[\bullet] + \frac{\varepsilon}{2}\frac{\mathrm{d}^2}{\mathrm{d}r^2}[\bullet]. \tag{466}$$

Then $T(r)$ solves a Dirichlet problem

$$\mathcal{L}_r T(r) = -1, \qquad T(r_G) = 0. \tag{467}$$

Solving Absorbing Boundary Dirichlet PDE. Equivalently,

$$\frac{\varepsilon}{2}T''(r) + \left(-\kappa r + \frac{\varepsilon(p-1)}{2r}\right)T'(r) = -1. \tag{468}$$

Writing

$$q(r) := T'(r), \tag{469}$$

we obtain

$$q'(r) + \left(-\frac{2\kappa}{\varepsilon}r + \frac{p-1}{r}\right)q(r) = -\frac{2}{\varepsilon}. \tag{470}$$

The integrating factor is

$$I(r) = r^{p-1}\exp\!\left(-\frac{\kappa}{\varepsilon}r^2\right). \tag{471}$$

Using

$$I(r)q(r) \to 0 \qquad \text{as } r \to \infty, \tag{472}$$

we obtain

$$q(r) = \frac{2}{\varepsilon}\frac{1}{I(r)}\int_r^\infty I(y)\mathrm{d}y. \tag{473}$$

Since $T(r_G) = 0$, for an initial radius $r_0 > r_G$,

$$T(r_0) = \frac{2}{\varepsilon} \int_{r_G}^{r_0} \frac{1}{I(x)} \int_x^\infty I(y) \mathrm{d}y \, \mathrm{d}x. \tag{474}$$

Define

$$\beta := \frac{\kappa}{\varepsilon} = \frac{\lambda}{\bar{s}\,\eta}. \tag{475}$$

Then

$$I(r) = r^{p-1} e^{-\beta r^2}, \tag{476}$$

and

$$\int_x^\infty y^{p-1} e^{-\beta y^2} \mathrm{d}y = \frac{1}{2} \beta^{-p/2} \Gamma(p/2, \beta x^2), \tag{477}$$

where

$$\Gamma(a, z) := \int_z^\infty t^{a-1} e^{-t} \mathrm{d}t. \tag{478}$$

Thus,

$$T(r_0) = \frac{1}{\varepsilon} \beta^{-p/2} \int_{r_G}^{r_0} \frac{e^{\beta x^2}}{x^{p-1}} \Gamma(p/2, \beta x^2) \mathrm{d}x. \tag{479}$$

Small-Learning-Rate Regime. The learning rate $\eta \to 0$ is small, so

$$\beta x^2 = \frac{\lambda}{\bar{s}\,\eta} x^2 \gg 1. \tag{480}$$

Therefore,

$$\Gamma(p/2, \beta x^2) = (\beta x^2)^{p/2-1} e^{-\beta x^2} \left[1 + \frac{p-2}{2\beta x^2} + O\big((\beta x^2)^{-2}\big) \right]. \tag{481}$$

Substituting equation (481) into equation (479) gives

$$\begin{aligned} T(r_0) &= \frac{1}{\varepsilon \beta} \int_{r_G}^{r_0} \left[\frac{1}{x} + \frac{p-2}{2\beta x^3} + O\big((\beta x^2)^{-2} x^{-1}\big) \right] \mathrm{d}x \\ &= \frac{1}{\kappa} \log \frac{r_0}{r_G} + \frac{\varepsilon(p-2)}{4\kappa^2} \left[\frac{1}{r_G^2} - \frac{1}{r_0^2} \right] + O\left(\frac{\varepsilon^2}{\kappa^3} \right). \end{aligned} \tag{482}$$

Substituting $\kappa = \lambda/\bar{s}$ and $\varepsilon = \eta$ yields

$$T(r_0) = \frac{\bar{s}}{\lambda} \log \frac{r_0}{r_G} + \eta \frac{\bar{s}^2(p-2)}{4\lambda^2} \left[\frac{1}{r_G^2} - \frac{1}{r_0^2} \right] + O\left(\frac{\eta^2}{\lambda^3} \right). \tag{483}$$

Asymptotic Expansions. The transition starts at the memorization radius and ends at the generalization radius. By Theorem 4,

$$\rho_M^2 = (\rho_M^{(0)})^2 + \frac{\eta}{b} c_M + O\big((\eta/b)^2\big). \tag{484}$$

Taking the square root gives

$$r_0 = \rho_M = \rho_M^{(0)} + \frac{\eta}{b} \frac{c_M}{2\rho_M^{(0)}} + O\big((\eta/b)^2\big). \tag{485}$$

Similarly, by Theorem 5,

$$\rho_G^2 = (\rho_G^{(0)})^2 + \frac{\eta}{\lambda} c_G + O\left(\frac{\eta^2}{\lambda^2} \right), \tag{486}$$

where

$$(\rho_G^{(0)})^2 = \mathbb{E}_{S_0 \in \Theta_S}\left[\|\theta^\star(S_0)\|_2^2\right], \tag{487}$$

and

$$\theta^\star(S_0) = \arg\min_{\theta(S_0)} \left\{ \mathscr{L}_f^*(\theta(S_0)) + \frac{\lambda}{2}\|\theta(S_0)\|_2^2 \right\}. \tag{488}$$

Therefore,

$$r_G = \rho_G = \rho_G^{(0)} + \frac{\eta}{\lambda}\frac{c_G}{2\rho_G^{(0)}} + O\left(\frac{\eta^2}{\lambda^2}\right). \tag{489}$$

Using

$$\log(x + \Delta x) = \log x + \frac{\Delta x}{x} + O(\Delta x^2), \tag{490}$$

we have

$$\log r_0 = \log \rho_M^{(0)} + \frac{\eta}{b}\frac{c_M}{2(\rho_M^{(0)})^2} + O\big((\eta/b)^2\big), \tag{491}$$

$$\log r_G = \log \rho_G^{(0)} + \frac{\eta}{\lambda}\frac{c_G}{2(\rho_G^{(0)})^2} + O\left(\frac{\eta^2}{\lambda^2}\right). \tag{492}$$

Hence

$$\log\frac{r_0}{r_G} = \log\frac{\rho_M^{(0)}}{\rho_G^{(0)}} + \frac{\eta}{b}\frac{c_M}{2(\rho_M^{(0)})^2} - \frac{\eta}{\lambda}\frac{c_G}{2(\rho_G^{(0)})^2} + O\left(\frac{\eta^2}{\lambda^2}\right). \tag{493}$$

In the explicit $O(\eta)$ Itô correction term of equation (483), it is sufficient to use the leading asymptotics:

$$\frac{1}{r_G^2} - \frac{1}{r_0^2} = \frac{1}{(\rho_G^{(0)})^2} - \frac{1}{(\rho_M^{(0)})^2} + O(\eta/b) + O(\eta/\lambda). \tag{494}$$

Substituting equations (493) and (494) into equation (483) gives

$$\begin{aligned} T_{M\to G} = {} & \frac{\bar{s}}{\lambda}\log\frac{\rho_M^{(0)}}{\rho_G^{(0)}} \\ & + \frac{\eta}{b\lambda}\bar{s}\frac{c_M}{2(\rho_M^{(0)})^2} \\ & + \frac{\eta}{\lambda^2}\left\{-\bar{s}\frac{c_G}{2(\rho_G^{(0)})^2} + \frac{\bar{s}^2(p-2)}{4}\left[\frac{1}{(\rho_G^{(0)})^2} - \frac{1}{(\rho_M^{(0)})^2}\right]\right\} \\ & + O\left(\frac{\eta^2}{\lambda^3}\right). \end{aligned} \tag{495}$$

Define

$$\begin{aligned} c_\tau &:= \bar{s}\,\frac{c_M}{2(\rho_M^{(0)})^2}, \\ c_\tau^{(2)} &:= -\bar{s}\,\frac{c_G}{2(\rho_G^{(0)})^2} + \frac{\bar{s}^2(p-2)}{4}\left[\frac{1}{(\rho_G^{(0)})^2} - \frac{1}{(\rho_M^{(0)})^2}\right]. \end{aligned} \tag{496}$$

Then

$$T_{M\to G} \approx \frac{\bar{s}}{\lambda}\log\frac{\rho_M^{(0)}}{\rho_G^{(0)}} + \frac{\eta}{b\lambda}c_\tau + \frac{\eta}{\lambda^2}c_\tau^{(2)} + O\left(\frac{\eta^2}{\lambda^3}\right). \tag{497}$$

Computing Number of Iterations. We compute the number of iterations, since the continuous-time interpolation is $t = \eta k$:

$$\mathbb{E}_{S_0 \in \Theta_S}[\tau_{M \to G}] = \frac{T_{M \to G}}{\eta}. \tag{498}$$

Dividing equation (497) by η gives

$$\mathbb{E}_{S_0 \in \Theta_S}[\tau_{M \to G}] \approx \frac{\bar{s}}{\eta\,\lambda} \log \frac{\rho_M^{(0)}}{\rho_G^{(0)}} + \frac{c_\tau}{b\,\lambda} + \frac{c_\tau^{(2)}}{\lambda^2} + O\Big(\frac{\eta}{\lambda^3}\Big), \tag{499}$$

which proves equation (446).

ℓ_2 Regularization Coefficient Dependence. In the late-stage radial SDE, the diffusion scale is $\varepsilon = \eta$ and is independent of batch size; the leading first-passage time is obtained by setting $\varepsilon = 0$ and is determined only by the deterministic radial drift $-\lambda r/\bar{s}$. Batch size enters the leading term through the scalar preconditioning scale $\bar{s} = O(1/\sqrt{b})$, and the corrections through the memorization asymptotic ρ_M — whose $O(\eta/b)$ correction becomes $O(1/b)$ after converting to iteration time — and through $c_\tau = O(1/\sqrt{b})$ and $c_\tau^{(2)} = O(1/b)$. □

A.15 Additional Results: Scaling Laws of Manifold Radius ρ_M^2 on $\mathbb{Z}_{127}$

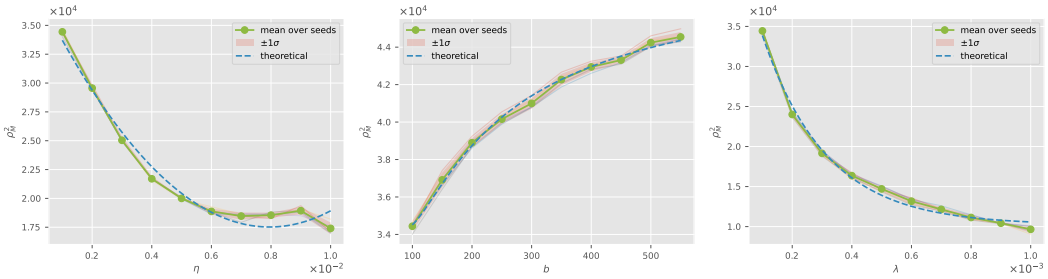


Figure 11: **Scaling Law of Manifold Radius ρ_M^2 on $\mathbb{Z}_{127}$.** We show the scaling law of ρ_M^2 with respect to the learning rate η, batch size b, and ℓ_2 regularization coefficient λ on the $\mathbb{Z}_{127}$ task. For each hyperparameter configuration, we train for ten runs. The results show that larger η/b induces stronger diffusion variance, whereas λ does not affect the diffusion variance. We also overlay the theoretical fits.

A.16 Additional Results: Scaling Laws of Manifold Radius ρ_G^2 on $\mathbb{Z}_{127}$

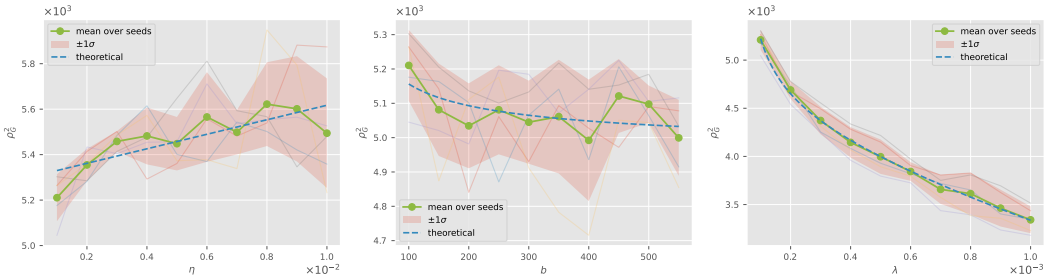


Figure 12: **Scaling Law of Manifold Radius ρ_G^2 on $\mathbb{Z}_{127}$.** We show the scaling law of ρ_G^2 with respect to the learning rate η, batch size b, and ℓ_2 regularization coefficient λ on the $\mathbb{Z}_{127}$ task. For each hyperparameter configuration, we train for ten runs. The results show that larger η induces stronger diffusion variance, whereas λ does not affect the diffusion variance. We also overlay the theoretical fits.

A.17 Additional Results: Scaling Laws of Solution Transition Time on $\mathbb{Z}_{127}$

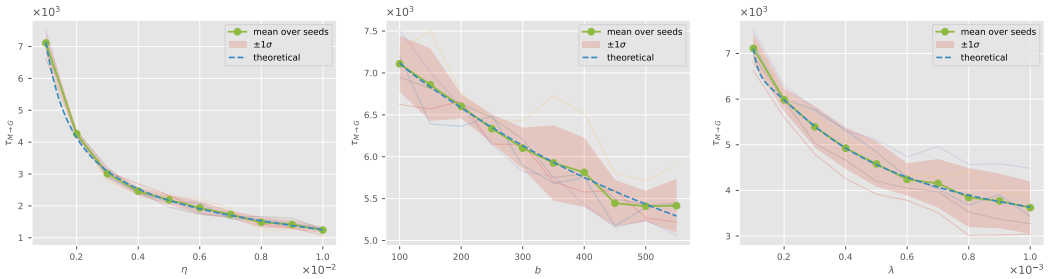


Figure 13: **Scaling laws of solution transition time on $\mathbb{Z}_{127}$.** We show that the solution transition time $\tau_{M\to G}$ from the memorization manifold M to the generalization manifold G scales with the learning rate η, batch size b, and ℓ_2 regularization coefficient λ. For each hyperparameter configuration, we train for ten runs. We also overlay the theoretical fits.