The Multiscale Single-Index Model: An MSE Analysis of Online SGD Dynamics

Gordon Dai
New York University
td2568@nyu.edu

Joan Bruna
New York University
jb4496@nyu.edu

Abstract

We consider the Multiscale Single-Index Model (MSIM), first introduced in [27], as a stylized model for hierarchical feature learning with scale separation. Each layer extracts a shared single-index feature at one physical scale and passes it to the next, giving a tractable setting in which to study how deep architectures learn local multiscale representations. Under non-degeneracy and delocalization assumptions on the link function and planted features respectively, for the two hidden-layer case (K=2), we analyze joint online spherical SGD under MSE loss, which is a canonical regression objective, in discrete time and show that it achieves weak recovery from random initialization in $O(d\log^2 d)$ samples, followed by strong recovery to accuracy ε in $O(d\log(1/\varepsilon))$ additional samples. The main technical challenge arises from the non-Gaussian learned features observed by the second layer. We address this challenge through a quantitative Gaussian comparison using delocalization, combined with sharp trajectory-level control of the coupled SGD dynamics. This work is a companion to [5], which studies recovery under a correlation objective for general K.

1 Introduction

Feature learning, informally defined as the ability to extract informative low-dimensional representations out of high-dimensional data, is the defining characteristic of Neural Networks trained with gradient descent —and the basis for the modern AI paradigm of pretraining then fine-tuning. Over the recent years, a feature learning theory has emerged, focusing on a stylized model for high-dimensional learning based on multi-index models [2, 16, 12, 13, 14, 3, 11]; see [6] for a recent survey, demonstrating the ability of shallow neural networks to perform efficient feature learning, and even identifying sharp phase transitions [25].

A natural shortcoming for such theory is that shallow NNs are limited idealizations of modern NNs, where depth is a critical component. Towards addressing this gap, several works [1, 26, 28, 15, 31] have studied feature learning in deeper architectures, raising novel questions such as (i) elucidating the hierarchy between features at different layers, or (ii) how to adapt gradient methods to facilitate learning across all layers.

In this work, motivated by applications in data domains such as images or text, we consider a stylized model for deep feature learning based on the principle of *scale separation*, where each layer of the network can be associated with a different physical scale. In the context of computer vision, the model becomes a Convolutional Neural Network (CNN) with disjoint receptive fields, as originally introduced in [27]; the first hidden-layer extracts local features at each patch, and subsequent layers capture local information at the next scale, progressively coarse-graining the domain [7, 22].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We instantiate each feature learning stage using a single-index model. Specifically, we consider a setting with Gaussian inputs $X \in (\mathbb{R}^d)^{\otimes K}$, capturing a multiscale structure with K scales and 'patches' of size d, and a target obtained by a K-layer network, whereby each layer $k \in [K]$ coarse-grains the feature map $Z^{(k-1)} \in (\mathbb{R}^d)^{\otimes K-k+1}$ as $Z^{(k)} = \phi(Z^{(k-1)}\theta_k^*) \in (\mathbb{R}^d)^{\otimes(K-k)}$; see Figure 1. The model $X \mapsto f_{\theta^*}(X) := Z^{(K)} \in \mathbb{R}$ is thus parametrized by hidden directions or features $\theta_1^*, \dots, \theta_K^* \in \mathbb{S}^{d-1}$, and a non-linear activation function $\phi : \mathbb{R} \to \mathbb{R}$.

By considering the Wiener chaos expansion of the target $f(X)$, this model can be viewed as a non-linear analogue of Tensor PCA [24], where the planted directions $\theta_1^*, \dots, \theta_K^*$ may be estimated from a spike in the associated chaos decomposition. This naturally leads to the study of two different algorithmic frameworks: (i) a spectral method, motivated by its efficiency in related problems, eg single- and multi-index models [23, 13, 20], spiked tensor PCA, and tensor-decomposition methods for non-overlapping CNNs [27]; and (ii) stochastic gradient-descent, inspired by the practical connection with deep learning. The latter setting is not addressed by these tensor-decomposition analyses and therefore is the main focus of our work.

Our analysis makes two key structural assumptions: we consider the setting where ϕ is *non-degenerate*, in the sense that $\mathbb{E}_{G\sim N(0,1)}[\phi'(G)] := \kappa \neq 0$ [1], and further assume that the planted directions are *delocalized*, unaligned with the basis on which the non-linear activations act. Exploiting these two assumptions, we analyze online SGD in the case (K=2), the immediate multiscale extension of the single-index setting, we establish a sample complexity of $n = O(d(\log(d)^2 + \log \varepsilon^{-1}))$ to achieve error ε, nearly matching the spectral estimator up to logarithmic factor. It is important to emphasize that the machinery of Ben Arous et al. [2] does *not* directly apply in this setting, due to the non-Gaussian features seen by the second hidden layer activations.

Related Works. Understanding the benefits of depth in Neural Networks is a longstanding question in the field, with initial progress centered on approximation questions [30, 17]. The efforts then moved towards incorporating optimization aspects, with early notable works such as [29, 21], and later developed in [32, 26, 1]. One aspect shared by all these works is the fact that the input high-dimensional distribution is rotation-invariant, ie with no particular physical structure. Early works that also exploited tensor representations of data to capture scale separation are [10], which focused on expressivity.

The closest works to ours are [15] and [28], which we discuss briefly. Dandi et al. [15] study Single/Multi-Index Gaussian Hierarchical Targets $f^\star(x) = g^\star(a^{\star\top} P_k(W^\star x)/\sqrt{d^{\epsilon_1}})$ with $W^\star \in \mathbb{R}^{d^{\epsilon_1}\times d}$, where the hierarchy is one of *dimensions* $d \to d^{\epsilon_1} \to 1$. They train a three-layer MLP *layer-wise* and propagate Hermite decompositions across layers via the asymptotic Gaussianity of the inner statistic $h^\star$, which holds because $\mathrm{IE}(P_k) = 2$. Our setup differs in three essentials: (1) the hierarchy is one of *physical scales* (within- vs. across-patch), with both layers acting in the same dimension d on single-index teachers $\beta^\star, \theta^\star \in \mathbb{S}^{d-1}$; (2) both layers are trained *jointly* by online SGD on the same timescale, so the second-layer student sees a moving, non-Gaussian feature map throughout; and (3) at $\mathrm{IE}(\phi) = 1$ the inner statistic is not asymptotically Gaussian, so we replace the CLT by a quantitative Stein/Lindeberg comparison driven by the delocalization. Ren et al. [28] prove that L-layer convolutional networks trained layer-wise learn L-level Random Hierarchy Models with $O(m^{(1+o(1))L})$ samples, via a problem-independent "shallow-to-deep chaining" reduction: layer-wise training succeeds whenever the target correlates with lower-level latents, the signal to lower layers is clean, and the lower-level features are weakly identifiable. Their setting is complementary to ours: The data are discrete tokens drawn from a probabilistic context-free grammar rather than continuous Gaussian patches, and the targets are categorical class labels rather than scalar regressions, so the relevant geometry is a tree of synonym classes rather than the product sphere. Our companion work [5] shows that under correlation loss between teacher model and student model $\mathbb{E}\big[f_\theta(X) f_{\theta^\star}(X)\big]$ when initialized in a favorable basin of constant-in-d probability, online SGD achieves achieves weak recovery in $\widetilde{O}(d^{K-1})$ samples, and continues to $1 - o_d(1)$ overlap after a lower-order continuation phase, for any fixed depth K. Yet it remains unclear how the online SGD behaves under standard loss such as Mean Square Error (MSE) loss, which is the main focus of this work.

[1] This corresponds to information exponent 1 in the language of [2].

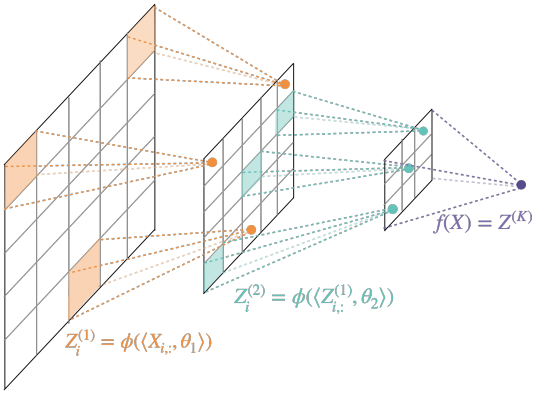


Figure 1: Schema of the Multiscale SIM model. Each patch within the same layer k shares the same local feature, parametrized by a SIM of direction θ_k.

Notation: All expectations $\mathbb{E}_X$ are with respect to X, with the directions held fixed, unless explicitly stated otherwise. For unit vectors u, v, define the sign-invariant distance $\mathrm{dist}_{\pm}(u, v) := \min\{\|u - v\|_2, \|u + v\|_2\}$.

2 The Multiscale Single-Index Model

Let $K \geq 1$ be fixed. We model the input as a K-th order Gaussian tensor of dimension d, ie $X \in (\mathbb{R}^d)^{\otimes K}$ have independent $N(0,1)$ entries. Given unit vectors $\theta_1, \ldots, \theta_K \in S^{d-1}$, define $Z^{(0)} = X$, and recursively

$$Y^{(r)}_{i_{r+1},\ldots,i_K} = \sum_{j=1}^{d} \theta_r(j) Z^{(r-1)}_{j,i_{r+1},\ldots,i_K}, \qquad 1 \leq r \leq K, \tag{1}$$
$$Z^{(r)} = \phi(Y^{(r)}) \quad \text{entrywise.}$$

At the final layer $Z^{(K)} \in \mathbb{R}$, and we write $f_\theta(X) := Z^{(K)}$. As illustrated in Figure 1, this is a stylized model of a NN with physical structure: each layer computes *local* and *translation-invariant* features at a fixed spatial *scale*. In the case where $K = 2$, for instance, the first layer extracts a single shared 'local' feature θ_1^* capturing internal correlations within each patch, while the second layer extracts a coarse feature θ_2^* capturing correlations across patches.

Our algorithmic approaches will be based on two main assumptions:

Assumption 2.1 (Non-degeneracy and regularity, informal)**.** *Let $G \sim N(0,1)$. We assume*

$$\mathbb{E}\phi(G) = 0, \qquad \mathbb{E}\phi(G)^2 = 1, \qquad \kappa := \mathbb{E}\phi'(G) \neq 0. \tag{A1}$$

Further, ϕ satisfies certain standard regularity assumptions, which differ slightly on each setting (discussed in each section).

In the language of Ben Arous et al. [2], non-degeneracy corresponds to the case where the information exponent of ϕ is 1. In the shallow setting $K = 1$, this assumption greatly simplifies the geometry of the model, ensuring a descent direction from random initializations. In our setting $K > 1$, the non-degeneracy is instrumental for gradient methods to escape the mediocrity, as will be shown in later sections.

Assumption 2.2 (Incoherence, informal)**.** *The planted directions are* delocalized*, in the sense that $\|\theta_k\|_\infty \ll 1$.*

Incoherence corresponds to the notion that no patch exerts an dominant influence on the output, and enables a mean-field approximation, or equivalently a *Gaussian universality* in our SGD analysis, similarly as in [18].

3 Main Result

We now turn to verify that, indeed, $K = 2$ is a tractable setting for SGD dynamics to achieve strong recovery under the natural MSE loss rather than linear correlation in our companion paper [5]. For ease of presentation, we name $\theta_1^* = \theta^*$ and $\theta_2^* = \beta^*$ in the remainder of this section.

3.1 Setup

Given the teacher direction $(\theta^*, \beta^*) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ we want to study the ability of *online stochastic gradient descent* (online SGD) to minimize the population loss

$$\mathcal{L}(\beta_t, \theta_t) := \|f_{\beta_t,\theta_t} - f_{\beta^*,\theta^*}\|^2 = \mathbb{E}_{x\sim\gamma_D}[f_{\beta_t,\theta_t}(x) - f_{\beta^*,\theta^*}(x)]^2, \tag{2}$$

where the label is generated by $z_i := \phi(X_i \cdot \beta^*), f_{\beta^*,\theta^*}(X) := \psi(z \cdot \theta^*), i \in [d]$. To recover the teacher, we run online SGD on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$:

$$\beta_{t+1} = \frac{\beta_t - \delta\nabla_\beta^{\mathcal{S}}\ell(\beta_t, \theta_t; x_t)}{\left\|\beta_t - \delta\nabla_\beta^{\mathcal{S}}\ell(\beta_t, \theta_t; x_t)\right\|}, \qquad \theta_{t+1} = \frac{\theta_t - \delta\nabla_\theta^{\mathcal{S}}\ell(\beta_t, \theta_t; x_t)}{\left\|\theta_t - \delta\nabla_\theta^{\mathcal{S}}\ell(\beta_t, \theta_t; x_t)\right\|}, \tag{3}$$

for $t = 0, 1, 2, \ldots$, where $(x_t \in \mathbb{R}^{d\times d})$ are fresh samples independent of the past, δ is the step-size, and the spherical gradient is defined as:

$$\nabla_\beta^{\mathcal{S}}\ell(\beta, \theta; x) := (I - \beta\beta^\top)\nabla_\beta\ell(\beta, \theta; x), \qquad \nabla_\theta^{\mathcal{S}}\ell(\beta, \theta; x) := (I - \theta\theta^\top)\nabla_\theta\ell(\beta, \theta; x).$$

Following [2, 8], throughout the learning process, we track the evolution of summary statistics, defined as overlaps $\rho_t := \beta_t \cdot \beta^*$ and $\chi_t := \theta_t \cdot \theta^*$ for the first and second layer's parameter respectively.

Theorem 3.1 (Informal)**.** *With our assumptions (A1)-(A4) (see Section B), for any $\varepsilon > 0$, under online SGD training (3), we have:*

- Weak Recovery*: starting from initialization, with probability at least $1 - K\varepsilon$, choosing step size $\delta \asymp 1/(d \log d)$, SGD achieves weak recovery: $\min\{\rho_t, \chi_t\} = \Theta(1)$, using $O(d \log^2 d)$ samples.*
- Strong Recovery*: conditioned on the weak recovery event, with probability at least $1 - O\left(\log(1/\varepsilon)/d\right)$, choosing step size $\delta \asymp 1/d$, SGD achieves strong recovery: $\min\{\rho_t, \chi_t\} \geq 1 - \varepsilon$, using $O\left(d \log(1/\varepsilon)\right)$ additional samples.*

We build our analysis on top of the analytical framework introduced in Ben Arous et al. [2] by modeling the learning process as drifts with fluctuations. The main technical point is to leverage delocalization (which, as remarked in B.2, will be true w.h.p. under (A4)) to control the gaussian approximation $\mathcal{L}_G$ of the true loss $\mathcal{L}$ to derive the continuous-population level on the gradient flow dynamics: Based on Stein's Coupling (see Theorem E.1), for each layer, in Lemma C.2, we bound the distance between the spherical gradient of the true loss $\nabla_{\mathbb{S},\beta}\mathcal{L}$ and its Gaussian proxy $\nabla_{\mathbb{S},\beta}\mathcal{L}_G$ by some constant factor of the incoherence $\eta_t := \|\theta_t\|_\infty + \|\theta^*\|_\infty$ that holds true uniformly in time:

$$|\langle\nabla_{\mathbb{S},\beta}(L - L_G)(\theta_t, \beta_t), \beta^*\rangle| \leq C_\beta^{(1)}\eta_t, \qquad |\langle\nabla_{\mathbb{S},\theta}(L - L_G)(\theta_t, \beta_t), \theta^*\rangle| \leq C_\theta^{(1)}\eta_t. \tag{4}$$

Continuous-time Analysis In continuous time, we consider the population-level spherical gradient flow defined as

$$\dot{\beta}_t = -\nabla_{\mathbb{S},\beta}L(\theta_t, \beta_t), \quad \dot{\theta}_t = -\nabla_{\mathbb{S},\theta}L(\theta_t, \beta_t). \tag{5}$$

Based on (4), in Lemma C.3 we derive a system of coupled ODE inequalities that captures the evolution of overlaps as

$$\dot{\rho}_t \geq a_\rho\chi_t(1 - \rho_t^2) - b_\rho\eta_t, \qquad \dot{\chi}_t \geq a_\chi\rho_t(1 - \chi_t^2) - b_\chi\eta_t, \tag{6}$$

where $a_\rho, a_\chi, b_\rho, b_\chi > 0$ are independent of d and (θ_t, β_t). From here, we can see that for the RHS to be non-vacuous and provide enough signal for the growth of (ρ_t, θ_t), terms $b_\rho\eta_t$ and $b_\chi\eta_t$ must not exceed $a_\rho\chi_t(1 - \rho_t^2)$ and $a_\chi\rho_t(1 - \chi_t^2)$, respectively. This requires us to provide a quantitative control of η_t, derived in Lemma C.4:

$$\dot{\eta}_t \leq C_1\rho_t\|\theta^*\|_\infty + C_2\chi_t\rho_t\eta_t + O(\eta_t^2)\,. \tag{7}$$

We make the following remarks on our setting:

- *On the Delocalization hypothesis*: our proof crucially relies on the Gaussian comparison in order to obtain an overlap growth differential inequality. Delocalization is not necessary for efficient recovery, e.g. if $\|\theta^*\|_0 = O_d(1)$ the target is a multi-index model, for which gradient-based guarantees exist [12, 4]. A natural follow-up question is to understand whether our SGD can obtain recovery guarantees uniform on θ^*.
- *On non-degeneracy* The key mechanism for learning is described in (6). Whenever the information exponent of ϕ is $s > 1$, one would instead obtain

$$\dot{\rho}_t \geq a_\rho \chi_t \rho_t^{s-1}(1-\rho_t^2) - b_\rho \eta_t, \qquad \dot{\chi}_t \geq a_\chi \rho_t \chi_t^{s-1}(1-\chi_t^2) - b_\chi \eta_t, \tag{8}$$

  in which the perturbative η_t term dominates the positive drift at initialization. This difficulty is not merely technical; our numerical simulations in Section A show that online SGD does not manage to learn in this setting.
- *On Gaussian inputs* For simplicity we state our result in the Gaussian setting, but using delocalization in the first layer feature, our arguments naturally extend to the class of well-behaved product measures for which Stein's CLT coupling applies.

3.2 Main Elements of the Proof

Below, we present the high-level roadmap of proofs in getting Theorem 3.1.

3.2.1 Weak Recovery

The proof has two dynamical regimes, Phase I and Phase II, separated by

$$\tau_A := \inf\{t : Z_t^2 \geq A\eta_t\}, \quad Z_t := \rho_t + \chi_t.$$

In Phase I, Z_t is still below the incoherence scale. In this phase, the crude $O(\eta_t)$ comparison error is too large under spherical initialization; in Lemma C.5 we bound the deviation of true population drifts from its Gaussian proxy by a refined third-moment quantity

$$\omega_t := \sum_{i=1}^{d} \big(|\theta_{t,i}| + |\theta_i^*|\big)^3, \tag{9}$$

which itself is controlled in Lemma C.6 and Lemma C.7. Under the small-overlap regime of weak recovery, the dynamics is well approximated by the coupled drift inequalities

$$\dot{\rho}_t \geq a\chi_t - b\omega_t, \qquad \dot{\chi}_t \geq a\rho_t - b\omega_t. \tag{10}$$

In Lemma C.8, we showed that (i) starting from initialization, (10) pulls Z_t from $d^{-1/2}$ to the signal-dominant scale $Z_t^2 \sim \eta_t$ in $O(\log d)$ time; (ii) after τ_A, the incoherence error is of a lower order compared with the positive Z-drift. The sum Z_t then grows from order $d^{-1/4}$ to a fixed small constant, while the imbalance $D_t = \rho_t - \chi_t$ contracts. The entry estimate $D_{\tau_A} = o(Z_{\tau_A})$ ensures that both overlaps become positive constants simultaneously under another $O(\log d)$ time, obtaining weak recovery.

Extension to Discrete-Time SGD The discrete-time weak-recovery proof mirrors the same two-regime mechanism, with additional martingale and discretization controls: We start from the one-step expansion in Lemma D.2,

$$\rho_{t+1} = \rho_t + \delta F_t^{(\rho)} + \delta M_t^{(\rho)} + R_t^{(\rho)} \qquad \chi_{t+1} = \chi_t + \delta F_t^{(\chi)} + \delta M_t^{(\chi)} + R_t^{(\chi)}, \tag{11}$$

where we take the step size $\delta \asymp 1/d\log d$. During Phase I, plugging the ω-controlled population drift in Lemma C.6 into (11), we obtained the aggregate one-step update expression

$$Z_{t+1} \geq (1+\beta\delta)Z_t - C\delta\omega_t + \varepsilon_t^{(Z)}, \tag{12}$$

where $\varepsilon_t^{(Z)}$ as the one-step increment of martingale and discretization errors, with their time-accumulation is controlled by Abel summation from the event $E_T^{(Z)}$. Besides, for the third-moment term w_t in (12), under Lemma D.12 as the discretized version of third-moment control made in Lemma C.7 we showed that for Phase I, $\omega_t \lesssim (1+t\delta)/\sqrt{d}$, making the weighted forcing remains

of order $d^{-1/2}$. Therefore Z_t grows from its initialization scale $d^{-1/2}$ to the signal-dominant scale $Z_{\tau_A} \gtrsim d^{-1/4}$ in $O(\delta^{-1}\log d)$ iterations, while $\eta_t \lesssim (\log d)/\sqrt{d}$. Same as continuous case, we arrived at an entry-balance estimate $D_{\tau_A} = o(Z_{\tau_A})$ using the small-overlap D-drift derived in D.15.

Entering Phase II, as shown in Lemma D.16, once $Z_t^2 \gtrsim \eta_t$, the incoherence error is negligible compared with the positive Z-drift, so Z_t grows from order $d^{-1/4}$ to a fixed constant c_Z within another $O(\delta^{-1}\log d)$ iterations. At the same time, the D-recursion is contractive up to lower-order forcing, which keeps $|D_t| \le Z_t/2$. Hence, as shown in Theorem D.17, at the hitting time $t_{\mathrm{esc}} := \inf\{t : Z_t \ge c_Z\}$, both overlaps are bounded below with high probability:

$$\min\{\rho_{t_{\mathrm{esc}}}, \chi_{t_{\mathrm{esc}}}\} = \frac{Z_{t_{\mathrm{esc}}} - |D_{t_{\mathrm{esc}}}|}{2} \ge \frac{c_Z}{4}.$$

3.2.2 Strong Recovery

Continuous-time Strong Recovery Results After weak recovery, we show in Lemma C.9 that under additional $O(1)$ time, the population dynamics drives (θ_t, β_t) into a fixed local cap of arbitrary level around $y^* := (\theta^*, \beta^*)$, the global minimizer of the true population loss. While the small-overlap estimate from weak recovery no longer holds, we employ the uniform bounds in Lemma C.3 and Lemma C.4 to show $\eta_t = o_d(1)$, which guarantees both overlaps enter the claimed local cap.

Once inside this local cap, the proof switches from overlap drift to local geometry. We analyze the Hessian of the loss function by decomposing it as:

$$H_L(y) = \underbrace{\big(H_L(y) - H_L(y^*)\big)}_{(\mathrm{A})} + \underbrace{\big(H_L(y^*) - H_{L_G}(y^*)\big)}_{(\mathrm{B})} + \underbrace{H_{L_G}(y^*)}_{(\mathrm{C})}. \tag{13}$$

From (13), we derive Lemma C.12 to lower bound term (B) by some constant factor of $-\|\theta^*\|_\infty$, and Lemma C.13 to lower bound term (A) by some constant factor of $-\operatorname{dist}_{\mathcal{M}}(y, y^*) := -\sqrt{\arccos(\rho)^2 + \arccos(\chi)^2}$, taken arbitrarily small. Finally, as (C) is lower bounded by a Hermite-coefficient quantity that dominates the other constants, in Lemma C.14, we obtain the positivity of the Hessian for the true loss $H_L(y)$ throughout the local cap.

Hence, under this implied local strong convexity, in Lemma C.15, by exploiting the the log-drift inequality

$$\big\langle \nabla_{\mathcal{M}} L(y_t), \operatorname{Log}_{y_t}(y^*) \big\rangle \le -2\mu_{\mathrm{loc}} Q_t, \quad Q_t := \frac{1}{2}\arccos(\chi_t)^2 + \frac{1}{2}\arccos(\rho_t)^2, \tag{14}$$

we make a standard Riemannian gradient-flow argument to obtain the exponential contraction of $\operatorname{dist}_{\mathcal{M}}(y, y^*)$, in the sense that

$$\operatorname{dist}_{\mathcal{M}}(y_t, y^*)^2 \le e^{-2\mu_{\mathrm{loc}}(t - t_{\mathrm{esc}})} \operatorname{dist}_{\mathcal{M}}(y_{t_{\mathrm{esc}}}, y^*)^2. \tag{15}$$

Converting the product distance back to overlap gives the strong recovery of

$$\min\{\rho_t, \chi_t\} \ge 1 - \varepsilon \tag{16}$$

after an additional $O(\log(1/\varepsilon))$ time from weak recovery. We formalize this result in Theorem C.16.

Extension to Discrete-Time SGD Same as the extension to discrete-time SGD in Section 3.2.1, here we must control stochastic fluctuations of SGD inside the strongly-convex basin to ensure the local contraction. Finishing the discrete-time weak recovery gave by Theorem D.17, we switch the SGD step size to $\delta_{\mathrm{loc}} = \frac{c_{\mathrm{loc}}}{d}$, that is optimal to strong recovery as remarked in RemarkD.18.

Mirroring Lemma C.9, Proposition D.21 shows that after an additional $O(d)$ iterations beyond weak recovery, the trajectory enters the local strongly-convex cap in discrete-time. Inside this cap, we track the *local distance*

$$Q_t := \frac{1}{2}\arccos(\chi_t)^2 + \frac{1}{2}\arccos(\rho_t)^2. \tag{17}$$

From its one-step recursion in Lemma D.22:

$$Q_{t+1} \le Q_t + \delta_{\mathrm{loc}}\langle \bar{g}_t, \mathcal{L}_t\rangle + \delta_{\mathrm{loc}}\Xi_t + C_{\mathrm{ret}}\delta_{\mathrm{loc}}^2 G_t, \tag{18}$$

we bound the population drift $\langle \bar{g}_t, \mathcal{L}_t \rangle$ by the log-drift inequality (14) in continuous time, and we bound both the martingale fluctuation term $\Xi_t := \langle g_t - \bar{g}_t, \mathcal{L}_t \rangle$ and the quadratic gradient term $G_t := \|g_t\|^2$ in Lemma D.23 by the sharp local self-quenching estimate:

$$\mathbb{E}[\Xi_t^2 \mid \mathcal{F}_t] \lesssim Q_t^2, \text{ and } \mathbb{E}[G_t \mid \mathcal{F}_t] \lesssim dQ_t, \tag{19}$$

which show that both the gradient energy and the martingale variance vanish proportionally to the current distance to the teacher. Thus the SGD noise becomes weaker as the iterates approach the teacher: Proposition D.24 proves that one local SGD epoch of length $O(\delta_{\mathrm{loc}}^{-1})$ contracts the distance scale by a factor of $1/2$ with probability $1 - O(\delta_{\mathrm{loc}})$, while the trajectory remains inside the local cap. Iterating dyadic epochs in Proposition D.25 yields $Q_t \le \varepsilon$ after $O\big(d \log \frac{1}{\varepsilon}\big)$ additional samples. We conclude our discrete-time result in Theorem D.26, where we obtained (16), namely the strong recovery under discrete-time SGD, with a total sample complexity of $O(d \log^2 d) + O\big(d \log \frac{1}{\varepsilon}\big)$.

4 Conclusions and Future Directions

In this work, we studied the Multiscale Single-Index Model, in which a d^K-dimensional input with multiscale structure is iteratively coarse-grained via local, translation-invariant single-index models, and where the number of scales K corresponds directly to the depth of the associated neural network. Provided the non-linearities are non-degenerate and the planted directions are delocalized, we proved weak and strong recovery of online SGD in the three-layer setting $K = 2$ using $O(d(\log(d)^2)$ samples.

Our work leaves several open questions ahead of us: the first, most pressing one is to extend the SGD recoverability to general depth $K > 2$. While [5] provides d-asymptotic strong recovery for general K under the correlation loss, it remains unclear how a general K model behaves under canonical regression objectives such as MSE loss. In our current SGD analysis, the restriction to $K = 2$ is tied to the weak-recovery phase near random initialization. Indeed, the MSE objective decomposes as

$$\mathcal{L}_{\mathrm{MSE}}(\Theta) = \mathbb{E}\big[f_\Theta(X)^2\big] - 2\mathbb{E}\big[f_\Theta(X) f_{\Theta^\star}(X)\big] + \mathbb{E}\big[f_{\Theta^\star}(X)^2\big]. \tag{20}$$

When $m_j = \langle \theta_j, \theta_j^\star \rangle \asymp d^{-1/2}$, the teacher-dependent correlation drift in mode i has size $\prod_{\ell \neq i} m_\ell \asymp d^{-(K-1)/2}$. In contrast, our current first-order bounds allow teacher-independent finite-d self-energy corrections at scale $d^{-1/2}$. For $K = 2$, these terms occur at the same scale, and our refined third-moment and incoherence estimates allow us to prove that the resulting two-overlap dynamics still has a positive cooperative drift to escape mediocrity. For $K > 2$, however, the planted drift is of strictly smaller order, and a first-order Gaussian or norm comparison no longer determines the sign of the overlap evolution. Moreover, unlike the perturbations in the correlation objective, the self-energy corrections are not naturally accompanied by planted-overlap factors. Figure 3, Section A shows the empirical evidence of hardness of MSE loss for $K > 2$.

The next natural question is to consider the regime of fixed d (the so-called 'receptive field') and diverging depth K; this will also require substantial technical innovation, since our CLT/universality arguments do not directly apply.

Finally, an important limitation of our work is that it imposes a form of scale separation: each layer lives on its dedicated scale. While this is a valid assumption for certain physical data (e.g images), most popular deep architectures do not have scale separation; for instance in ResNets [19]/ NeuralODEs [9] the features remain at the same scale throughout several layers. Building toy models that explain feature learning in these settings remains an outstanding challenge.

5 Acknowledgments

GD was partially supported by the Dean's Undergradaute Research Fund from NYU's College of Arts and Science. This work was completed during the JB sabbatical visit at the Flatiron Institute, which also provided the computing resources used in the project. The authors would also like to thank Hengzhi Li for helping them running experiments, and Dr. Denny Wu, Yunwei Ren, Hanzhe Wu for reviewing our draft and providing helpful feedback.

References

[1] Z. Allen-Zhu and Y. Li. What Can ResNet Learn Efficiently, Going Beyond Kernels? 2019.

[2] G. B. Arous, R. Gheissari, and A. Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference, 2021.

[3] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[4] A. Bietti, J. Bruna, and L. Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

[5] J. Bruna. The multiscale single-index model: A stylized model for hierarchical feature learning, 2026.

[6] J. Bruna and D. Hsu. Survey on algorithms for multi-index models, 2025.

[7] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013.

[8] J. Bruna, L. Pillaud-Vivien, and A. Zweig. On single index models beyond gaussian data, 2023.

[9] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations, 2019.

[10] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on learning theory*, pages 698–728. PMLR, 2016.

[11] H. Cui, L. Pesce, Y. Dandi, F. Krzakala, Y. M. Lu, L. Zdeborová, and B. Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. *arXiv preprint arXiv:2402.04980*, 2024.

[12] A. Damian, J. Lee, and M. Soltanolkotabi. Neural Networks can Learn Representations with Gradient Descent. 2022.

[13] A. Damian, L. Pillaud-Vivien, J. D. Lee, and J. Bruna. Computational-statistical gaps in gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.

[14] Y. Dandi, F. Krzakala, B. Loureiro, L. Pesce, and L. Stephan. Learning two-layer neural networks, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

[15] Y. Dandi, L. Pesce, L. Zdeborová, and F. Krzakala. The computational advantage of depth: Learning high-dimensional hierarchical functions with gradient descent, 2025.

[16] R. Dudeja and D. Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pages 1887–1930. PMLR, 2018.

[17] R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.

[18] R. Gheissari and A. Jagannath. Universality of high-dimensional scaling limits of stochastic gradient descent. *arXiv preprint arXiv:2512.13634*, 2025.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.

[20] N. Joshi, T. Misiakiewicz, and N. Srebro. On the complexity of learning sparse functions with statistical and gradient queries. *arXiv preprint arXiv:2407.05622*, 2024.

[21] E. Malach, G. Yehudai, S. Shalev-Schwartz, and O. Shamir. The connection between approximation, depth separation and learnability in neural networks. In *Conference on Learning Theory*, pages 3265–3295. PMLR, 2021.

[22] S. Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203, 2016.

[23] M. Mondelli and A. Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Conference On Learning Theory*, pages 1445–1450. PMLR, 2018.

[24] A. Montanari and E. Richard. A statistical model for tensor pca. *Advances in neural information processing systems*, 27, 2014.

[25] A. Montanari and Z. Wang. Phase transitions for feature learning in neural networks, 2026.

[26] E. Nichani, A. Damian, and J. D. Lee. Provable guarantees for nonlinear feature learning in three-layer neural networks. *Advances in Neural Information Processing Systems*, 36, 2023.

[27] S. Oymak and M. Soltanolkotabi. Learning a deep convolutional neural network via tensor decomposition. *Information and Inference: A Journal of the IMA*, 10(3):1031–1071, 2021.

[28] Y. Ren, Y. Dandi, F. Krzakala, and J. D. Lee. Provable learning of random hierarchy models and hierarchical shallow-to-deep chaining, 2026.

[29] I. Safran and J. Lee. Optimization-based separations for neural networks. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3–64. PMLR, 02–05 Jul 2022.

[30] M. Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

[31] Z. Wang, E. Nichani, and J. D. Lee. Learning hierarchical polynomials with three-layer neural networks. *arXiv preprint arXiv:2311.13774*, 2023.

[32] Z. Wang, E. Nichani, and J. D. Lee. Learning hierarchical polynomials with three-layer neural networks, 2023.

Appendix

Table of Contents

A Numerical Experiments **10**

B Formalized Assumptions **11**

C Proof of Continuous-Time Population Dynamics **13**
C.1 Initialization Event 13
C.2 General Results on Dynamics 14
C.3 Weak Recovery 20
C.4 Strong Recovery 32

D Proof of Discrete-Time Online SGD Dynamics **46**
D.1 Weak Recovery Setups 46
D.2 Weak Recovery Dynamics 60
D.3 Strong Recovery Setups 75
D.4 Strong Recovery Dynamics 89

E Auxiliary Lemmas **97**
E.1 Stein Bounds 97
E.2 Gradient Bounds 101
E.3 Summation by Parts 104

A Numerical Experiments

We illustrate the performance of online SGD on several representative experiments. In order to improve the computational efficiency, we implement an online minibatch version, with batchsize$= 64$. We set up a learning rate of order $\eta = C/(d \log d)$ with a parameter C optimized using a small hyper-parameter search. We ran all the experiments on an Apple Macbook pro running on the GPU. We initialize the parameters (θ^*, β^*) drawn i.i.d. from the uniform measure, and we initialize (θ_0, β_0) uniformly at random from the non-negative-overlap hemispheres $\{\theta; \theta \cdot \theta^* \geq 0\}, \{\beta; \beta \cdot \beta^* \geq 0\}$.

Sample complexity In Figure 2 we plot the recovery performance of online SGD for varying input $d \in \{10, 20, 50, 100, 200\}$, for the choice $\phi = \mathrm{GeLU}$. We illustrate the performance by tracking $\min(\chi_t, \rho_t)$, running 10 independent seeds and plotting error bars. As predicted by our theory, the algorithm successfully recovers the planted parameters, at a rate that is roughly $n \simeq d \log d$.

Comparing different K, d pair choices To empirically see the potential of online SGD under MSE loss extends beyond $K > 2$ and compare its performance with correlation loss as studied in our companion work [5], we ran a set of experiments with varying K. Specifically, based on our model set-up, we consider a fixed dimension of model input, $d^K = 16777216$, and tested pair choices $(K, d) \in \{(2, 4096), (3, 256), (4, 64), (6, 16)\}$ for both correlation and MSE loss. All experiments use normalized $\phi = \mathrm{GeLU}$ activations, a learning rate of $\eta = 0.000391$, minibatches of size 16, and 2^{27} training examples across 10 random seeds. As the figure shows, only for $K = 2$ do we see reliable recovery for both losses. Their performance diverges substantially as depth increases, with correlation loss exhibits meaningful recovery among all remaining trials.

Testing non-degeneracy assumption Finally, we investigate the effect of removing the non-degeneracy assumption, by using instead $\phi = \mathrm{He}_2$, which has information exponent 2. Figure 4 shows that, in this setting, online SGD does not make significant progress, even at timescales $n \gg d^2$.

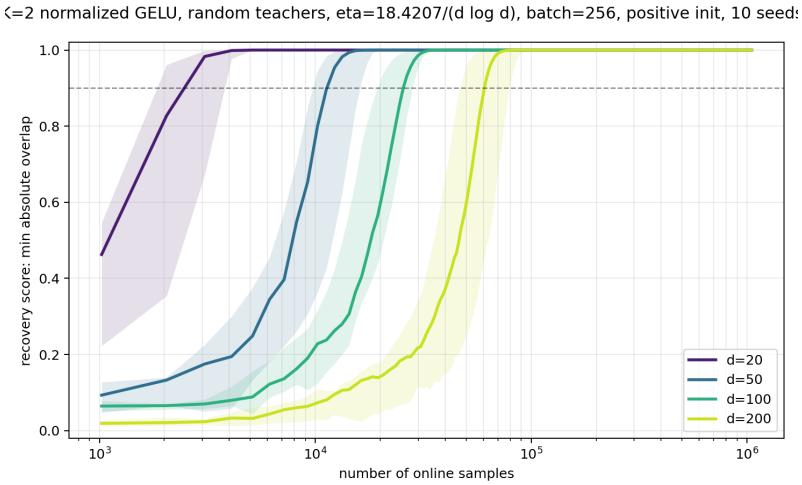


Figure 2: Tracking the overlaps with varying d for the choice $\phi = \mathrm{Gelu}$

B Formalized Assumptions

Let $G \sim N(0,1)$, and write the Hermite expansions $\phi = \sum_{k\geq 0} a_k h_k$ and $\psi = \sum_{k\geq 0} \alpha_k h_k$ in the normalized Hermite basis. We assume:

(A1) **Normalization.** $\mathbb{E}\phi(G) = 0$ and $\mathbb{E}\phi(G)^2 = 1$.

(A2) **Non-degeneracy.** The first Hermite coefficients are nonzero:

$$a_1 = \mathbb{E}[G\phi(G)] = \mathbb{E}[\phi'(G)] \neq 0, \qquad \alpha_1 = \mathbb{E}[G\psi(G)] = \mathbb{E}[\psi'(G)] \neq 0.$$

(A3) **Regularity of activation function.** The inner activation $\phi \in C^3(\mathbb{R})$ and outer activation $\psi \in C_b^4(\mathbb{R})$ have finite regularity constants

$$L_\phi := \|\phi'\|_\infty + \|\phi''\|_\infty + \|\phi'''\|_\infty, \qquad L_\psi := \max_{0\leq j\leq 4} \|\psi^{(j)}\|_\infty.$$

Here $C_b^4(\mathbb{R})$ requires ψ and its derivatives through order four to be bounded; ϕ itself need not be bounded.

(A4) **Initialization.** The initial second layer student parameter and the teacher parameter are uniformly sampled from the unit sphere: $\theta_0, \theta^* \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and set the initial overlap to be of the order $d^{-1/2}$ The initialization distribution of the overlap satisfies

$$\mathcal{E}_{\mathrm{ov}} := \left\{ \rho_0, \chi_0 \in [c_{\mathrm{init}} d^{-1/2}, C_{\mathrm{init}} d^{-1/2}] \right\}.$$

Below, we made two remarks, for (A3) and for (A4) respectively:

Remark B.1. Assumption (A3) implies all moment conditions used below. Indeed, since $\|\phi'\|_\infty < \infty$,

$$|\phi(x)| \leq |\phi(0)| + \|\phi'\|_\infty |x|.$$

Hence $\phi(G)$ has finite Gaussian moments of every order. Moreover, for $1 \leq j \leq 3$, the derivatives $\phi^{(j)}$ are bounded. Therefore, for every finite $p \geq 1$,

$$\max_{0\leq j\leq 3} \mathbb{E}\left[(1+|G|)^p |\phi^{(j)}(G)|^p\right] < \infty, \qquad G \sim N(0,1),$$

where $\phi^{(0)} = \phi$. In particular,

$$\mathbb{E}|\phi(G)|^8 < \infty, \qquad \mathbb{E}|\phi'(G)|^8 < \infty,$$

and

$$\mathbb{E}\big[(1+|\phi(G)|^2)|\phi'(G)|^2\big] < \infty, \qquad \mathbb{E}\big[|\phi(G)|\,|\phi'(G)|^2\big] < \infty.$$

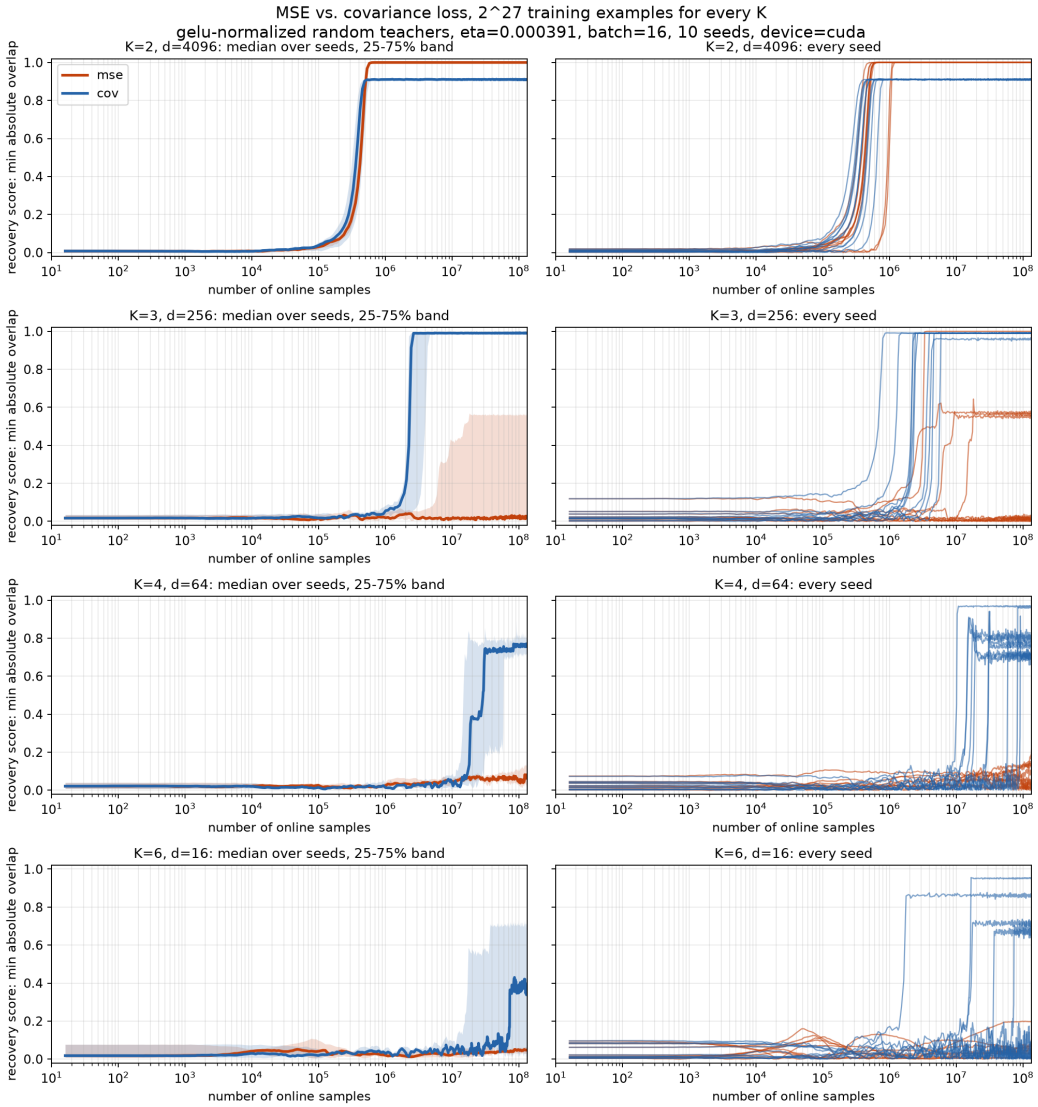


Figure 3: Comparison between MSE and Correlation Loss, for various pairs of K, d with $d^K = 16777216$, using $\phi = \text{GeLU}$.

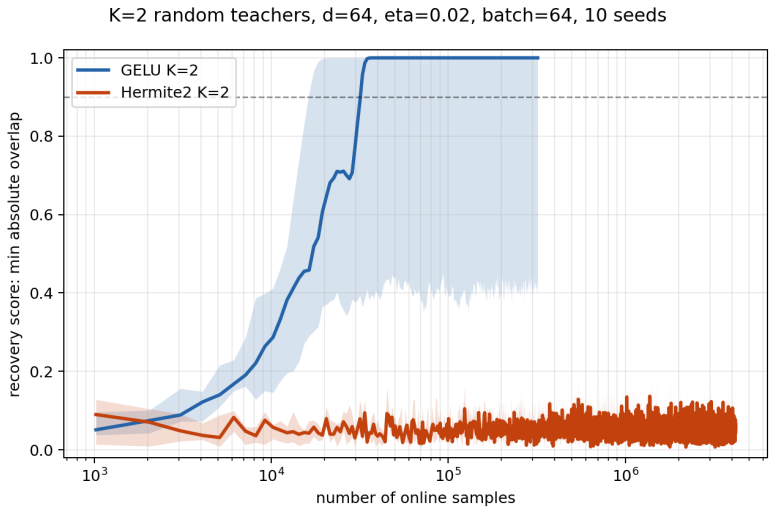


Figure 4: Comparison between $\phi = \mathrm{GeLU}$ and $\phi = \mathrm{He}_2$.

Finally, if Y is sub-Gaussian, then $\phi(Y)$ is sub-Gaussian, with

$$\|\phi(Y)\|_{\psi_2} \le C\big(|\phi(0)| + L_\phi \|Y\|_{\psi_2}\big),$$

for a universal constant C. This follows from the same Lipschitz bound.

Remark B.2. By standard concentration of the uniform measure (e.g. [4]), the event $\mathcal{E}_{\mathrm{ov}}$ in (A4) will happen with probability at least $p_0 > 0$ independent of d for the uniform initialization.

Fix $K > 0$. We work on the event

$$\mathcal{E}_0(K) := \mathcal{E}_{\mathrm{ov}} \cap \mathcal{E}_{\mathrm{deloc}}(K),$$

where

$$\mathcal{E}_{\mathrm{deloc}}(K) := \left\{ \|\theta_0\|_\infty + \|\theta^*\|_\infty \le C_K \sqrt{\frac{\log d}{d}}, \quad \|\theta_0\|_3^3 + \|\theta^*\|_3^3 \le C_K d^{-1/2} \right\}.$$

By Lemma C.1, given $\theta_0, \theta^* \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, $\mathcal{E}_{\mathrm{deloc}}(K)$ holds with probability at least $1 - d^{-K}$.

C Proof of Continuous-Time Population Dynamics

C.1 Initialization Event

Lemma C.1 (Delocalization of random spherical vectors)**.** *Let $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Then, for every fixed $K > 0$, there exists a constant $C_K < \infty$, independent of d, such that*

$$\mathbb{P}\left(\|u\|_\infty \le C_K \sqrt{\frac{\log d}{d}} \quad \textit{and} \quad \|u\|_3^3 \le C_K d^{-1/2} \right) \ge 1 - d^{-K}. \tag{21}$$

Consequently, if θ_0, θ^* are independent uniform random vectors on $\mathbb{S}^{d-1}$, then, after increasing C_K, with probability at least $1 - d^{-K}$,

$$\|\theta_0\|_\infty + \|\theta^*\|_\infty \le C_K \sqrt{\frac{\log d}{d}}, \tag{22}$$

and

$$\|\theta_0\|_3^3 + \|\theta^*\|_3^3 \le C_K d^{-1/2}. \tag{23}$$

Proof. Write

$$u = \frac{g}{\|g\|_2}, \qquad g = (g_1, \dots, g_d), \qquad g_i \overset{\text{iid}}{\sim} N(0,1).$$

Standard Gaussian concentration gives, for every fixed $K > 0$, after choosing C_K large enough,

$$\max_{1 \le i \le d} |g_i| \le C_K \sqrt{\log d}, \qquad \|g\|_2^2 \ge \frac{d}{2}$$

with probability at least $1 - d^{-K}$. Hence

$$\|u\|_\infty = \frac{\max_i |g_i|}{\|g\|_2} \le C_K \sqrt{\frac{\log d}{d}}.$$

For the third moment,

$$\|u\|_3^3 = \sum_{i=1}^d |u_i|^3 = \frac{\sum_{i=1}^d |g_i|^3}{\|g\|_2^3}.$$

By a standard concentration bound for sums of iid variables with finite moments, for every fixed $K > 0$ there exists $C_K < \infty$ such that

$$\sum_{i=1}^d |g_i|^3 \le C_K d$$

with probability at least $1 - d^{-K}$. Combining this with $\|g\|_2^2 \ge d/2$, we get

$$\|u\|_3^3 \le \frac{C_K d}{(d/2)^{3/2}} \le C_K d^{-1/2}.$$

This proves (21). By a union bound applied to θ_0 and θ^*, we get (22) and (23), respectively. □

C.2 General Results on Dynamics

Lemma C.2 (Uniform first-order gradient comparison)**.** *Assume (A1) and (A3), then there exist constants $C_\beta^{(1)}, C_\theta^{(1)} > 0$, independent of d, such that uniformly for all (θ, β) with $\rho, \chi \ge 0$,*

$$|\langle \nabla_{\mathbb{S},\beta}(L - L_G)(\theta, \beta), \beta^* \rangle| \le C_\beta^{(1)} \eta. \tag{24}$$

$$|\langle \nabla_{\mathbb{S},\theta}(L - L_G)(\theta, \beta), \theta^* \rangle| \le C_\theta^{(1)} \eta. \tag{25}$$

Proof. Throughout the proof, constants may change from line to line but depend only on the model and moment constants, not on $d, \theta, \theta^*, \beta, \beta^*$.

Set

$$U_i := x^{(i)} \cdot \beta, \qquad U_i^* := x^{(i)} \cdot \beta^*,$$

so that $(U_i, U_i^*)_{i=1}^d$ are iid centered Gaussian pairs with correlation ρ. Define

$$X_i := \phi(U_i), \qquad X_i^* := \phi(U_i^*),$$

and

$$S := \sum_{i=1}^d \theta_i X_i, \qquad S^* := \sum_{i=1}^d \theta_i^* X_i^*.$$

Then

$$\mathbb{E}[X_i] = \mathbb{E}[X_i^*] = 0, \qquad \mathbb{E}[X_i^2] = \mathbb{E}[(X_i^*)^2] = 1, \qquad \mathbb{E}[X_i X_i^*] = \kappa(\rho).$$

Let $(\Gamma_i, \Gamma_i^*)_{i=1}^d$ be iid Gaussian pairs with the same covariance matrix

$$\mathbb{E}[\Gamma_i^2] = \mathbb{E}[(\Gamma_i^*)^2] = 1, \qquad \mathbb{E}[\Gamma_i \Gamma_i^*] = \kappa(\rho),$$

and write

$$S_G := \sum_{i=1}^d \theta_i \Gamma_i, \qquad S_G^* := \sum_{i=1}^d \theta_i^* \Gamma_i^*.$$

Step 1: the θ-gradient comparison. Let

$$w := P_{\theta^\perp}\theta^* = \theta^* - \chi\theta.$$

Then

$$\langle \nabla_{\mathbb{S},\theta} L(\theta,\beta), \theta^* \rangle = 2\mathbb{E}\left[(\psi(S) - \psi(S^*))\psi'(S) \sum_{i=1}^{d} w_i X_i \right].$$

The same formula holds for L_G, with (X, X^*) replaced by (Γ, Γ^*). Hence it is enough to compare the expectations of

$$h(x, x^*) := 2\big(\psi(x\cdot\theta) - \psi(x^*\cdot\theta^*)\big)\psi'(x\cdot\theta)(x\cdot w).$$

We apply Corollary E.2. Fix a block j, condition on all other blocks, and write

$$s := x_j, \qquad t := x_j^*,$$

$$\ell_1 = \theta_j s + S_1, \qquad \ell_2 = w_j s + S_2, \qquad \ell_3 = \theta_j^* t + S_3,$$

where S_1, S_2, S_3 are the leave-out sums. Then

$$h(s,t) = 2(\psi(\ell_1) - \psi(\ell_3))\psi'(\ell_1)\ell_2.$$

Since $\psi \in C_b^4$, every third partial derivative $\partial_s^a \partial_t^b h$, $a+b=3$, is a finite sum of terms of the two following types:

$$C\,|\theta_j|^a |\theta_j^*|^b |\ell_2|, \qquad a+b=3,$$

and

$$C\,|w_j|\,|\theta_j|^{a'} |\theta_j^*|^{b'}, \qquad a'+b'=2.$$

At every hybrid vector in the Lindeberg interpolation,

$$\mathbb{E}|\ell_2| \le C\|w\|_2 \le C.$$

Therefore

$$\mathbb{E}\left[\left|\partial_s^a \partial_t^b h(W)\right|\right] \le C\left(|\theta_j|^a |\theta_j^*|^b + |w_j| \sum_{a'+b'=2} |\theta_j|^{a'} |\theta_j^*|^{b'} \right).$$

Summing over j and over $a+b=3$, we get

$$|\mathbb{E}h(X,X^*) - \mathbb{E}h(\Gamma,\Gamma^*)| \le C\left[\sum_j \sum_{a+b=3} |\theta_j|^a |\theta_j^*|^b + \sum_j |w_j| \sum_{a'+b'=2} |\theta_j|^{a'} |\theta_j^*|^{b'} \right]. \tag{26}$$

The first sum is bounded by

$$C(\|\theta\|_\infty + \|\theta^*\|_\infty).$$

Indeed,

$$\sum_j |\theta_j|^3 \le \|\theta\|_\infty, \qquad \sum_j |\theta_j|^2 |\theta_j^*| \le \|\theta^*\|_\infty,$$

and similarly for the remaining degree-three monomials. For the second sum, use

$$\|w\|_\infty = \|\theta^* - \chi\theta\|_\infty \le \|\theta^*\|_\infty + \|\theta\|_\infty = \eta,$$

and $\|w\|_2, \|\theta\|_2, \|\theta^*\|_2 \le C$, to obtain

$$\sum_j |w_j| \sum_{a'+b'=2} |\theta_j|^{a'} |\theta_j^*|^{b'} \le C\eta.$$

Thus

$$|\langle \nabla_{\mathbb{S},\theta}(L - L_G)(\theta,\beta), \theta^* \rangle| \le C\eta.$$

This proves the θ-gradient comparison.

Step 2: exact formula for the β-gradient. Let

$$b := P_{\beta^\perp}\beta^* = \beta^* - \rho\beta.$$

Then

$$\langle \nabla_{\mathbb{S},\beta} L(\theta,\beta), \beta^* \rangle = \langle \nabla_\beta L(\theta,\beta), b \rangle.$$

Differentiating S in the direction b gives

$$D_b S = \sum_{i=1}^d \theta_i \phi'(U_i) H_i, \qquad H_i := x^{(i)} \cdot b.$$

Hence

$$\langle \nabla_{\mathbb{S},\beta} L(\theta,\beta), \beta^* \rangle = 2\mathbb{E}\left[(\psi(S) - \psi(S^*))\psi'(S) \sum_{i=1}^d \theta_i \phi'(U_i) H_i \right].$$

For each i, $H_i = U_i^* - \rho U_i$, so H_i is centered Gaussian with variance $1-\rho^2$, independent of U_i. Moreover $U_i^* = \rho U_i + H_i$. Applying Gaussian integration by parts in the H_i-variable gives

$$\mathbb{E}\left[(\psi(S) - \psi(S^*))\psi'(S)\theta_i \phi'(U_i) H_i \right] = -(1-\rho^2)\theta_i \theta_i^* \mathbb{E}\left[\psi'(S)\psi'(S^*)\phi'(U_i)\phi'(U_i^*) \right].$$

Therefore

$$-\langle \nabla_{\mathbb{S},\beta} L(\theta,\beta), \beta^* \rangle = 2(1-\rho^2) \sum_{i=1}^d \theta_i \theta_i^* \mathbb{E}\left[\psi'(S)\psi'(S^*)\phi'(U_i)\phi'(U_i^*) \right]. \tag{27}$$

Step 3: leave-one-out reduction for the true β-gradient. Set

$$F(r,s) := \psi'(r)\psi'(s), \qquad q_i := \phi'(U_i)\phi'(U_i^*), \qquad c_i := \theta_i \theta_i^*.$$

Also define

$$S_{-i} := S - \theta_i X_i, \qquad S_{-i}^* := S^* - \theta_i^* X_i^*.$$

Since S_{-i}, S_{-i}^* are independent of (U_i, U_i^*),

$$\mathbb{E}[F(S_{-i}, S_{-i}^*) q_i] = \mathbb{E}[F(S_{-i}, S_{-i}^*)]\mathbb{E}[q_i].$$

By the Hermite identity,

$$\mathbb{E}[q_i] = \mathbb{E}[\phi'(U_i)\phi'(U_i^*)] = \kappa'(\rho).$$

Moreover, since F is Lipschitz and the stated moment assumptions imply

$$\sup_{\rho \in [0,1]} \mathbb{E}\left[(|X_i| + |X_i^*|) |q_i| \right] < \infty,$$

we have

$$\left| \mathbb{E}[F(S,S^*)q_i] - \mathbb{E}[F(S_{-i}, S_{-i}^*)q_i] \right| \le C(|\theta_i| + |\theta_i^*|).$$

Multiplying by $|c_i|$ and summing,

$$\sum_{i=1}^d |c_i|(|\theta_i| + |\theta_i^*|) \le \sum_i |\theta_i|^2 |\theta_i^*| + \sum_i |\theta_i| |\theta_i^*|^2 \le \|\theta^*\|_\infty + \|\theta\|_\infty = \eta.$$

Thus

$$\sum_{i=1}^d c_i \mathbb{E}[F(S,S^*)q_i] = \kappa'(\rho) \sum_{i=1}^d c_i \mathbb{E}[F(S_{-i}, S_{-i}^*)] + O(\eta).$$

Similarly,

$$\left| \mathbb{E}[F(S,S^*)] - \mathbb{E}[F(S_{-i}, S_{-i}^*)] \right| \le C(|\theta_i| + |\theta_i^*|),$$

and the same summation estimate gives

$$\sum_{i=1}^d c_i \mathbb{E}[F(S_{-i}, S_{-i}^*)] = \left(\sum_{i=1}^d c_i \right) \mathbb{E}[F(S,S^*)] + O(\eta).$$

Since $\sum_i c_i = \theta \cdot \theta^* = \chi$, we conclude

$$\sum_{i=1}^d c_i \mathbb{E}[F(S,S^*)q_i] = \kappa'(\rho)\chi\, \mathbb{E}[F(S,S^*)] + O(\eta). \tag{28}$$

Step 4: Stein comparison for the two-dimensional statistic. Apply Corollary E.2 to

$$h(x, x^*) := F(x \cdot \theta, x^* \cdot \theta^*).$$

For a fixed block j, the third partials have the form

$$\partial_s^a \partial_t^b h = \theta_j^a (\theta_j^*)^b \partial_1^a \partial_2^b F(\ell_1, \ell_2), \qquad a + b = 3.$$

Since $F \in C_b^3(\mathbb{R}^2)$,

$$\mathbb{E}\left[|\partial_s^a \partial_t^b h(W)|\right] \le C|\theta_j|^a |\theta_j^*|^b.$$

Therefore

$$|\mathbb{E}[F(S, S^*)] - \mathbb{E}[F(S_G, S_G^*)]| \le C \sum_j \sum_{a+b=3} |\theta_j|^a |\theta_j^*|^b \le C\eta.$$

Combining this with (28), and using that $\kappa'(\rho)$ and χ are bounded on the entry corridor, gives

$$\sum_{i=1}^d c_i \mathbb{E}[F(S, S^*) q_i] = \kappa'(\rho)\chi\, \mathbb{E}[F(S_G, S_G^*)] + O(\eta). \tag{29}$$

Step 5: identify the Gaussian β-gradient. In the Gaussian equivalent model,

$$L_G(\theta, \beta) = \Phi(\chi, \rho), \qquad \Phi(\chi, \rho) = 2 \sum_{k \ge 1} \alpha_k^2 \left[1 - (\kappa(\rho)\chi)^k\right].$$

Equivalently, by Price's identity applied to the Gaussian pair (S_G, S_G^*),

$$-\Phi_\rho(\chi, \rho) = 2\kappa'(\rho)\chi \mathbb{E}[\psi'(S_G)\psi'(S_G^*)].$$

Since

$$\nabla_{\mathbb{S},\beta} L_G = \Phi_\rho(\chi, \rho) P_{\beta^\perp} \beta^*,$$

we have

$$-\left\langle \nabla_{\mathbb{S},\beta} L_G(\theta, \beta), \beta^* \right\rangle = 2(1 - \rho^2)\kappa'(\rho)\chi \mathbb{E}[\psi'(S_G)\psi'(S_G^*)]. \tag{30}$$

Comparing (27), (29), and (30), we obtain

$$|\langle \nabla_{\mathbb{S},\beta}(L - L_G)(\theta, \beta), \beta^* \rangle| \le C(1 - \rho^2)\eta \le C\eta.$$

This proves the β-gradient comparison.

Combining the two block estimates gives the lemma. □

Lemma C.3 (Uniform positive-overlap drift)**.** *Assume Lemma C.2. On the positive overlap region $\rho, \chi \ge 0$, the population gradient flow satisfies*

$$\dot\rho_t \ge a_\rho \chi_t (1 - \rho_t^2) - b_\rho \eta_t, \qquad \dot\chi_t \ge a_\chi \rho_t (1 - \chi_t^2) - b_\chi \eta_t, \tag{31}$$

where $a_\rho, a_\chi, b_\rho, b_\chi > 0$ are independent of d.

Proof. With the population flow defined in (5), we have

$$\dot\rho_t = -\left\langle \nabla_{\mathbb{S},\beta} L(\theta_t, \beta_t), \beta^* \right\rangle, \qquad \dot\chi_t = -\left\langle \nabla_{\mathbb{S},\theta} L(\theta_t, \beta_t), \theta^* \right\rangle.$$

We first compute the Gaussian drifts exactly. Recall that

$$L_G(\theta, \beta) = \Phi(\chi, \rho), \qquad \Phi(\chi, \rho) = 2 \sum_{k \ge 1} \alpha_k^2 \left[1 - (\kappa(\rho)\chi)^k\right],$$

where $\kappa(\rho) = \sum_{\ell \ge 1} a_\ell^2 \rho^\ell$.

For the first layer,

$$\nabla_{\mathbb{S},\beta} L_G = \Phi_\rho(\chi, \rho) P_{\beta^\perp} \beta^* = \Phi_\rho(\chi, \rho)(\beta^* - \rho\beta).$$

Hence

$$-\left\langle \nabla_{\mathbb{S},\beta} L_G, \beta^* \right\rangle = -\Phi_\rho(\chi, \rho)(1 - \rho^2).$$

Moreover,

$$-\Phi_\rho(\chi,\rho) = 2\chi\kappa'(\rho)\sum_{k\geq 1} k\alpha_k^2(\kappa(\rho)\chi)^{k-1}.$$

On the positive-overlap region $\rho, \chi \geq 0$, all terms are nonnegative. Keeping only the $k = 1$ term gives

$$-\Phi_\rho(\chi,\rho) \geq 2\alpha_1^2\chi\kappa'(\rho).$$

Since

$$\kappa'(\rho) = \sum_{\ell\geq 1} \ell a_\ell^2 \rho^{\ell-1} \geq a_1^2 \qquad (\rho \geq 0),$$

we obtain

$$-\langle \nabla_{\mathbb{S},\beta} L_G, \beta^*\rangle \geq 2\alpha_1^2 a_1^2 \chi(1-\rho^2).$$

Similarly,

$$\nabla_{\mathbb{S},\theta} L_G = \Phi_\chi(\chi,\rho) P_{\theta^\perp}\theta^* = \Phi_\chi(\chi,\rho)(\theta^* - \chi\theta),$$

so

$$-\langle \nabla_{\mathbb{S},\theta} L_G, \theta^*\rangle = -\Phi_\chi(\chi,\rho)(1-\chi^2).$$

Now

$$-\Phi_\chi(\chi,\rho) = 2\kappa(\rho)\sum_{k\geq 1} k\alpha_k^2(\kappa(\rho)\chi)^{k-1}.$$

Again all terms are nonnegative on $\rho, \chi \geq 0$, so

$$-\Phi_\chi(\chi,\rho) \geq 2\alpha_1^2\kappa(\rho).$$

Since

$$\kappa(\rho) = \sum_{\ell\geq 1} a_\ell^2 \rho^\ell \geq a_1^2\rho \qquad (\rho \geq 0),$$

we get

$$-\langle \nabla_{\mathbb{S},\theta} L_G, \theta^*\rangle \geq 2\alpha_1^2 a_1^2 \rho(1-\chi^2).$$

Now transfer the Gaussian drift to the true population loss. By (24) from Lemma C.2,

$$\begin{aligned}\dot\rho_t &= -\langle \nabla_{\mathbb{S},\beta} L_G(\theta_t,\beta_t), \beta^*\rangle - \langle \nabla_{\mathbb{S},\beta}(L - L_G)(\theta_t,\beta_t), \beta^*\rangle \\ &\geq 2\alpha_1^2 a_1^2 \chi_t(1-\rho_t^2) - C_\beta^{(1)}\eta_t.\end{aligned}$$

Likewise, by (25) from Lemma C.2,

$$\begin{aligned}\dot\chi_t &= -\langle \nabla_{\mathbb{S},\theta} L_G(\theta_t,\beta_t), \theta^*\rangle - \langle \nabla_{\mathbb{S},\theta}(L - L_G)(\theta_t,\beta_t), \theta^*\rangle \\ &\geq 2\alpha_1^2 a_1^2 \rho_t(1-\chi_t^2) - C_\theta^{(1)}\eta_t.\end{aligned}$$

Thus the claimed inequalities hold with

$$a_\rho = a_\chi := 2\alpha_1^2 a_1^2, \qquad b_\rho := C_\beta^{(1)}, \qquad b_\chi := C_\theta^{(1)}.$$

□

Lemma C.4 (ℓ_∞ control)**.** *Under our assumptions,*

$$\dot\eta_t \leq C_1\rho_t\|\theta^*\|_\infty + C_2\chi_t\rho_t\eta_t + O(\eta_t^2)\,. \tag{32}$$

Proof. By definition, we have

$$\mathcal{L} = \mathbb{E}_{(z,z^*)\sim\phi_\#\gamma_{\beta\cdot\beta^*}}[|\psi(z\cdot\theta) - \psi(z^*\cdot\theta^*)|^2]$$

so

$$\nabla_\theta\mathcal{L} = \mathbb{E}_{(z,z^*)\sim\phi_\#\gamma_{\beta\cdot\beta^*}}[(\psi(z\cdot\theta) - \psi(z^*\cdot\theta^*))\psi'(z\cdot\theta)z]\,.$$

Call ν the law of $\phi(w)$, where $w \sim N(0,1)$. The random vector z is then drawn from $\nu^{\otimes d}$. Let us now analyze the two terms in the RHS.

The first term reads $\mathbb{E}[F(z\cdot\theta)z]$ for F of the form $F=\psi\psi'$. Assume that F' is Lipschitz. A second-order Taylor expansion of F yields

$$F(z\cdot\theta) = F(z_1\theta_1 + z_{-1}\cdot\theta_{-1}) \tag{33}$$

$$= F(z_{-1}\cdot\theta_{-1}) + F'(z_{-1}\cdot\theta_{-1})\theta_1 z_1 + \frac{1}{2}\theta_1^2 z_1^2 F''(M)\,, \tag{34}$$

and therefore, for all $i\in[d]$,

$$\mathbb{E}[F(z\cdot\theta)z_i] = \theta_i\mathbb{E}[F'(z_{-i}\cdot\theta_{-i})] + O(\theta_i^2)\,. \tag{35}$$

Now, for each θ and $i\in[d]$, consider $\mu_i = \mathrm{Law}(\sum_{j\neq i}\theta_j z_j)\in\mathcal{P}(\mathbb{R})$, where $z_j\sim\nu$ iid. We claim that

$$\sup_{i,i'} W_1(\mu_i,\mu_{i'}) = O(\|\theta\|_\infty)\,. \tag{36}$$

Indeed, fix i,i' and consider the coupling $(Y_i,Y_{i'})\sim\Gamma_{\mu_i,\mu_{i'}}$ given by $Y_i = W+\theta_i Z$, $Y_{i'} = W+\theta_{i'}Z$, where $Z\sim\nu$ and $W=\sum_{j\neq i,i'}\theta_j z_j$, with $z_j\sim\nu$ independently of Z. We then have

$$W_1(\mu_i,\mu_{i'}) \le \mathbb{E}|Y_i - Y_{i'}| = |\theta_i-\theta_{i'}|\mathbb{E}|Z| \le 2\sqrt{\mathbb{E}[|\phi(G)|^2]}\|\theta\|_\infty \le 2\|\theta\|_\infty\,.$$

Now, since F' is Lipschitz, we have

$$\sup_{i,j}|\mathbb{E}[F'(z_{-i}\cdot\theta_{-i})] - \mathbb{E}[F'(z_{-j}\cdot\theta_{-j})]| \le \mathrm{Lip}(F')\sup_{i,i'} W_1(\mu_i,\mu_{i'}) = O(\|\theta\|_\infty)\,, \tag{37}$$

and therefore, denoting by $\upsilon(\theta) = \frac{1}{d}\sum_i \mathbb{E}[F'(z_{-i}\cdot\theta_{-i})]$, we have

$$\mathbb{E}[F(z\cdot\theta)z] = \theta\upsilon(\theta) + \varepsilon\,,\ \text{with } |\varepsilon_i|\le C|\theta_i|\|\theta\|_\infty\,. \tag{38}$$

It follows that

$$P_{\theta^\perp}\mathbb{E}[F(z\cdot\theta)z] = P_{\theta^\perp}\varepsilon = \varepsilon - (\varepsilon\cdot\theta)\theta \tag{39}$$

satisfies

$$\|P_{\theta^\perp}\mathbb{E}[F(z\cdot\theta)z]\|_\infty \le \tilde{C}\|\theta\|_\infty^2\,. \tag{40}$$

Now we repeat the argument for a term of the form $\mathbb{E}[G(z\cdot\theta,\tilde{z}\cdot\theta^*)z]$ for G of the form $G(t_1,t_2) = \psi'(t_1)\psi(t_2)$. Doing a second-order Taylor expansion leads to

$$G(z\cdot\theta,\tilde{z}\cdot\theta^*) = G(z_1\theta_1 + z_{-1}\cdot\theta_{-1}, \tilde{z}_1\theta_1^* + \tilde{z}_{-1}\cdot\theta_{-1}^*) \tag{41}$$

$$= \psi'(z_{-1}\cdot\theta_{-1})\psi(\tilde{z}_{-1}\cdot\theta_{-1}^*) + \psi''(z_{-1}\cdot\theta_{-1})\psi(\tilde{z}_{-1}\cdot\theta_{-1}^*)\theta_1 z_1 + \psi'(z_{-1}\cdot\theta_{-1})\psi'(\tilde{z}_{-1}\cdot\theta_{-1}^*)\tilde{z}_1\theta_1^* \tag{42}$$

$$+ \frac{1}{2}[z_1\theta_1\ \ \tilde{z}_1\theta_1^*]^\top\nabla^2 G(M)[z_1\theta_1\ \ \tilde{z}_1\theta_1^*]\,, \tag{43}$$

and thus, for all $i\in[d]$, denoting $\alpha = \mathbb{E}[z_i\tilde{z}_i]$, and assuming that ∇G is Lipschitz, we have

$$\mathbb{E}[G(z\cdot\theta,\tilde{z}\cdot\theta^*)z_i] = \theta_i\mathbb{E}[\psi''(z_{-i}\cdot\theta_{-i})\psi(\tilde{z}_{-i}\cdot\theta_{-i}^*)] + \alpha\theta_i^*\mathbb{E}[\psi'(z_{-i}\cdot\theta_{-i})\psi'(\tilde{z}_{-i}\cdot\theta_{-i}^*)] + O(\theta_i^2 + (\theta_i^*)^2)\,. \tag{44}$$

Following our previous argument, we now consider for each θ,θ^* and each $i\in[d]$ the bivariate distributions $\bar{\mu}_i = \mathrm{Law}(\sum_{j\neq i}\theta_j z_j, \sum_{j\neq i}\theta_j^*\tilde{z}_j)$, where $(z_i,\tilde{z}_i)$ are drawn i.i.d. from $\bar{\nu}$. We now have

$$\sup_{i,i'} W_1(\bar{\mu}_i,\bar{\mu}_{i'}) = O(\|\theta\|_\infty + \|\theta^*\|_\infty)\,, \tag{45}$$

leading to

$$\mathbb{E}[G(z\cdot\theta,\tilde{z}\cdot\theta^*)z] = \theta\upsilon_1(\theta,\theta^*) + \alpha\theta^*\upsilon_2(\theta,\theta^*) + \varepsilon\,, \tag{46}$$

with $|\varepsilon_i|\le(C_1|\theta_i| + C_2|\alpha||\theta_i^*|)(\|\theta\|_\infty + \|\theta^*\|_\infty)$ and $\upsilon_2 = O(1)$. We then arrive at

$$P_{\theta^\perp}\mathbb{E}[G(z\cdot\theta,\tilde{z}\cdot\theta^*)z] = \alpha\theta^*\upsilon_2(\theta,\theta^*) - \chi\alpha\upsilon_2\theta + \varepsilon - (\varepsilon\cdot\theta)\theta\,. \tag{47}$$

Again, since $|\varepsilon\cdot\theta| = O(\|\theta\|_\infty + \|\theta^*\|_\infty)$, we have

$$\|\nabla_\theta^S\mathcal{L}\|_\infty \le C_1\rho\|\theta^*\|_\infty + C_2\rho\chi\|\theta\|_\infty + O(\|\theta\|_\infty^2 + \|\theta^*\|_\infty^2)\,, \tag{48}$$

proving the lemma. □

C.3 Weak Recovery

At the weak recovery, the total overlap Z_t is still below the incoherence scale. In this regime the crude $O(\eta_t)$ comparison error from Lemma C.3 is too large under spherical initialization. We instead track the third moment quantity $\omega_t := \sum_{i=1}^d \big(|\theta_{t,i}| + |\theta_i^*|\big)^3$ and show it is controlled, at population level, under the order $\frac{1+t}{\sqrt{d}}$.

Lemma C.5 (First-order comparison in the small-overlap regime)**.** *Assume the hypotheses of Lemma C.2 and the regularity needed for Corollary E.2. Define*

$$\omega(\theta, \theta^*) := \sum_{i=1}^d \big(|\theta_i| + |\theta_i^*|\big)^3.$$

Let

$$\dot\rho := -\left\langle \nabla_{\mathbb{S},\beta} L(\theta,\beta), \beta^* \right\rangle, \qquad \dot\chi := -\left\langle \nabla_{\mathbb{S},\theta} L(\theta,\beta), \theta^* \right\rangle,$$

and define $\dot\rho_G, \dot\chi_G$ analogously with L_G. Then, uniformly on the small-overlap region $0 \le \rho, \chi$ and $Z := \rho + \chi \le c_{\mathrm{sm}}$,

$$|\dot\rho - \dot\rho_G| + |\dot\chi - \dot\chi_G| \le C\omega(\theta, \theta^*), \tag{49}$$

where C is independent of d.

Proof. Throughout, constants may change from line to line and depend only on the model and moment constants. Set

$$U_i := x^{(i)} \cdot \beta, \qquad U_i^* := x^{(i)} \cdot \beta^*,$$

so that $(U_i, U_i^*)_{i=1}^d$ are iid centered Gaussian pairs with correlation ρ. Let

$$X_i := \phi(U_i), \qquad X_i^* := \phi(U_i^*),$$

and

$$S := \sum_{i=1}^d \theta_i X_i, \qquad S^* := \sum_{i=1}^d \theta_i^* X_i^*.$$

Let $(\Gamma_i, \Gamma_i^*)_{i=1}^d$ be iid centered Gaussian pairs with the same block covariance as (X_i, X_i^*):

$$\mathbb{E}[\Gamma_i^2] = \mathbb{E}[(\Gamma_i^*)^2] = 1, \qquad \mathbb{E}[\Gamma_i \Gamma_i^*] = \kappa(\rho),$$

and write

$$S_G := \sum_{i=1}^d \theta_i \Gamma_i, \qquad S_G^* := \sum_{i=1}^d \theta_i^* \Gamma_i^*.$$

We prove the two block comparisons separately.

The θ-block. Let

$$w := P_{\theta^\perp}\theta^* = \theta^* - \chi\theta.$$

Since $w \in T_\theta \mathbb{S}^{d-1}$,

$$\langle \nabla_{\mathbb{S},\theta}(L - L_G)(\theta,\beta), \theta^* \rangle = \langle \nabla_\theta (L - L_G)(\theta,\beta), w \rangle .$$

Thus the θ-comparison is the difference between the true and Gaussian expectations of a function of the form

$$h(x, x^*) = 2\big(\psi(x\cdot\theta) - \psi(x^*\cdot\theta^*)\big)\psi'(x\cdot\theta)\,(x\cdot w).$$

Apply Corollary E.2 to h. For a fixed block j, every third partial derivative $\partial_s^a \partial_t^b h$, $a + b = 3$, is a finite sum of terms of the following two types:

$$C|\theta_j|^a |\theta_j^*|^b\, |\ell_2|, \qquad a + b = 3,$$

and

$$C|w_j|\, |\theta_j|^{a'} |\theta_j^*|^{b'}, \qquad a' + b' = 2,$$

where ℓ_2 denotes the linear form $w \cdot W$ evaluated at the corresponding Lindeberg hybrid vector. Since $\|w\|_2 \le 1$, the moment assumptions imply

$$\mathbb{E}|\ell_2| \le C.$$

Therefore

$$|\langle \nabla_{\mathbb{S},\theta}(L - L_G)(\theta,\beta), \theta^*\rangle| \le C\sum_{j=1}^{d}\sum_{a+b=3}|\theta_j|^a|\theta_j^*|^b + C\sum_{j=1}^{d}|w_j|\sum_{a'+b'=2}|\theta_j|^{a'}|\theta_j^*|^{b'}.$$

Now

$$|w_j| = |\theta_j^* - \chi\theta_j| \le |\theta_j^*| + |\theta_j|,$$

because $|\chi| \le 1$. Hence, for every j,

$$\sum_{a+b=3}|\theta_j|^a|\theta_j^*|^b \le C\big(|\theta_j| + |\theta_j^*|\big)^3,$$

and

$$|w_j|\sum_{a'+b'=2}|\theta_j|^{a'}|\theta_j^*|^{b'} \le C\big(|\theta_j| + |\theta_j^*|\big)^3.$$

Summing over j, we obtain

$$|\langle \nabla_{\mathbb{S},\theta}(L - L_G)(\theta,\beta), \theta^*\rangle| \le C\omega(\theta,\theta^*).$$

Equivalently,

$$|\dot\chi - \dot\chi_G| \le C\omega(\theta,\theta^*).$$

The β-block. Let

$$b := P_{\beta^\perp}\beta^* = \beta^* - \rho\beta.$$

Then

$$\langle \nabla_{\mathbb{S},\beta}L(\theta,\beta), \beta^*\rangle = \langle \nabla_\beta L(\theta,\beta), b\rangle\,.$$

As in the proof of Lemma C.2, Gaussian integration by parts in the b-direction gives the exact formula

$$-\langle \nabla_{\mathbb{S},\beta}L(\theta,\beta), \beta^*\rangle = 2(1-\rho^2)\sum_{i=1}^{d}\theta_i\theta_i^*\mathbb{E}[\psi'(S)\psi'(S^*)\phi'(U_i)\phi'(U_i^*)]\,.$$

Set

$$F(s,t) := \psi'(s)\psi'(t), \qquad q_i := \phi'(U_i)\phi'(U_i^*), \qquad c_i := \theta_i\theta_i^*.$$

We compare

$$\sum_{i=1}^{d} c_i\mathbb{E}[F(S,S^*)q_i]$$

with its Gaussian analogue.

Let

$$S_{-i} := S - \theta_i X_i, \qquad S_{-i}^* := S^* - \theta_i^* X_i^*.$$

Since (S_{-i}, S_{-i}^*) is independent of (U_i, U_i^*),

$$\mathbb{E}[F(S_{-i},S_{-i}^*)q_i] = \mathbb{E}[F(S_{-i},S_{-i}^*)]\mathbb{E}[q_i].$$

By the Hermite identity,

$$\mathbb{E}[q_i] = \mathbb{E}[\phi'(U_i)\phi'(U_i^*)] = \kappa'(\rho).$$

Moreover, by the Lipschitz bound on F and the moment assumptions,

$$\big|\mathbb{E}[F(S,S^*)q_i] - \mathbb{E}[F(S_{-i},S_{-i}^*)q_i]\big| \le C(|\theta_i| + |\theta_i^*|).$$

Multiplying by $|c_i| = |\theta_i\theta_i^*|$ and summing yields

$$\sum_i |c_i|(|\theta_i| + |\theta_i^*|) = \sum_i |\theta_i\theta_i^*|(|\theta_i| + |\theta_i^*|) \le C\omega(\theta,\theta^*).$$

Hence

$$\sum_{i=1}^{d} c_i\mathbb{E}[F(S,S^*)q_i] = \kappa'(\rho)\sum_{i=1}^{d} c_i\mathbb{E}[F(S_{-i},S_{-i}^*)] + O(\omega).$$

The same leave-one-out argument gives

$$\sum_{i=1}^{d} c_i \mathbb{E}[F(S_{-i}, S^*_{-i})] = \left(\sum_{i=1}^{d} c_i\right) \mathbb{E}[F(S, S^*)] + O(\omega).$$

Since

$$\sum_{i=1}^{d} c_i = \theta \cdot \theta^* = \chi,$$

we obtain

$$\sum_{i=1}^{d} c_i \mathbb{E}[F(S, S^*) q_i] = \kappa'(\rho)\chi\, \mathbb{E}[F(S, S^*)] + O(\omega).$$

It remains to replace (S, S^*) by (S_G, S^*_G). Applying Corollary E.2 to

$$h(x, x^*) = F(x \cdot \theta, x^* \cdot \theta^*)$$

gives

$$|\mathbb{E}[F(S, S^*)] - \mathbb{E}[F(S_G, S^*_G)]| \le C \sum_{j=1}^{d} \sum_{a+b=3} |\theta_j|^a |\theta^*_j|^b \le C\omega(\theta, \theta^*).$$

Since $\kappa'(\rho)$ and χ are bounded, we conclude

$$\sum_{i=1}^{d} c_i \mathbb{E}[F(S, S^*) q_i] = \kappa'(\rho)\chi\, \mathbb{E}[F(S_G, S^*_G)] + O(\omega).$$

In the Gaussian equivalent model, Price's identity gives

$$-\langle \nabla_{\mathbb{S},\beta} L_G(\theta, \beta), \beta^* \rangle = 2(1 - \rho^2)\kappa'(\rho)\chi \mathbb{E}[F(S_G, S^*_G)].$$

Comparing the last two displays and using $0 \le 1 - \rho^2 \le 1$, we get

$$|\langle \nabla_{\mathbb{S},\beta}(L - L_G)(\theta, \beta), \beta^* \rangle| \le C\omega(\theta, \theta^*).$$

Equivalently,

$$|\dot\rho - \dot\rho_G| \le C\omega(\theta, \theta^*).$$

Combining the θ- and β-block estimates proves

$$|\dot\rho - \dot\rho_G| + |\dot\chi - \dot\chi_G| \le C\omega(\theta, \theta^*).$$

□

Lemma C.6 (Small-overlap drift estimates)**.** *Under the assumptions of Lemma C.2, with the same definition of $\omega(\theta, \theta^*)$ in C.5. There exist constants $c_{\mathrm{sm}}, a, b, C > 0$, independent of d, such that whenever $\rho, \chi \ge 0, \quad Z := \rho + \chi \le c_{\mathrm{sm}}$, the population overlap drifts satisfy*

$$\dot\rho \ge a\chi - b\omega(\theta, \theta^*), \qquad \dot\chi \ge a\rho - b\omega(\theta, \theta^*). \tag{50}$$

Moreover, with $\widetilde{C} := 2\alpha_1^2 a_1^2 > 0$, one has the refined two-sided expansion

$$\dot\rho = \widetilde{C}\,\chi(1 - \rho^2) + R_\rho, \qquad \dot\chi = \widetilde{C}\,\rho(1 - \chi^2) + R_\chi, \tag{51}$$

where $|R_\rho| + |R_\chi| \le C\big(Z^2 + \omega(\theta, \theta^*)\big)$. Consequently, for $D := \rho - \chi$,

$$\dot D = -\widetilde{C}(1 + \rho\chi)D + R_D, \qquad |R_D| \le C\big(Z^2 + \omega(\theta, \theta^*)\big). \tag{52}$$

Proof. Let L_G denote the Gaussian-equivalent population loss. As in the proof of Lemma C.3, the Gaussian overlap drifts are

$$\dot\rho_G = 2\chi\,\kappa'(\rho)A(q)(1 - \rho^2), \qquad \dot\chi_G = 2\kappa(\rho)A(q)(1 - \chi^2),$$

where

$$q := \kappa(\rho)\chi, \qquad A(q) := \sum_{k \ge 1} k\alpha_k^2 q^{k-1}, \qquad \kappa(\rho) := \sum_{\ell \ge 1} a_\ell^2 \rho^\ell.$$

Since $a_1 \neq 0$ and $\alpha_1 \neq 0$,

$$\kappa'(0) = a_1^2 > 0, \qquad A(0) = \alpha_1^2 > 0.$$

First we prove the cooperative lower bounds. On the positive-overlap region $\rho, \chi \geq 0$, all terms in the Gaussian drift formulas are nonnegative. Keeping only the leading terms gives

$$\dot{\rho}_G \geq 2\alpha_1^2 a_1^2 \, \chi(1-\rho^2), \qquad \dot{\chi}_G \geq 2\alpha_1^2 a_1^2 \, \rho(1-\chi^2).$$

Choose $c_{\mathrm{sm}} > 0$ sufficiently small so that $Z \leq c_{\mathrm{sm}}$ implies $\rho, \chi \leq c_{\mathrm{sm}}$ and hence

$$1-\rho^2 \geq \frac{1}{2}, \qquad 1-\chi^2 \geq \frac{1}{2}.$$

Therefore

$$\dot{\rho}_G \geq c_0 \chi, \qquad \dot{\chi}_G \geq c_0 \rho$$

for $c_0 := \alpha_1^2 a_1^2 > 0$.

It remains to transfer these estimates from L_G to the true population loss. By Lemma C.5,

$$|\dot{\rho} - \dot{\rho}_G| + |\dot{\chi} - \dot{\chi}_G| \leq C\omega(\theta, \theta^*).$$

Hence, after setting $a := c_0/2$ and increasing b,

$$\dot{\rho} \geq a\chi - b\omega(\theta, \theta^*), \qquad \dot{\chi} \geq a\rho - b\omega(\theta, \theta^*),$$

which proves (50).

We now prove the two-sided expansion. Since $Z \leq c_{\mathrm{sm}}$, we have

$$|\rho| + |\chi| \leq c_{\mathrm{sm}}.$$

Taylor expanding κ, κ', and A near the origin gives

$$\kappa(\rho) = a_1^2 \rho + O(\rho^2), \qquad \kappa'(\rho) = a_1^2 + O(|\rho|),$$

and since

$$q = \kappa(\rho)\chi = O(Z^2),$$

also

$$A(q) = \alpha_1^2 + O(Z^2).$$

Substituting these expansions into the exact Gaussian drift formulas yields

$$\dot{\rho}_G = 2\alpha_1^2 a_1^2 \, \chi(1-\rho^2) + O(Z^2),$$

and

$$\dot{\chi}_G = 2\alpha_1^2 a_1^2 \, \rho(1-\chi^2) + O(Z^2).$$

Indeed, in the first formula the error from $\kappa'(\rho) - a_1^2$ is $O(|\rho||\chi|) = O(Z^2)$, while the error from $A(q) - \alpha_1^2$ is $O(|\chi|Z^2) = O(Z^3)$. In the second formula, the error from $\kappa(\rho) - a_1^2\rho$ is $O(\rho^2) = O(Z^2)$, and the error from $A(q) - \alpha_1^2$ is $O(|\rho|Z^2) = O(Z^3)$.

Thus, with $\widetilde{C} := 2\alpha_1^2 a_1^2$, we have

$$\dot{\rho}_G = \widetilde{C} \, \chi(1-\rho^2) + O(Z^2), \qquad \dot{\chi}_G = \widetilde{C} \, \rho(1-\chi^2) + O(Z^2).$$

Using again the refined comparison estimate

$$|\dot{\rho} - \dot{\rho}_G| + |\dot{\chi} - \dot{\chi}_G| \leq C\omega(\theta, \theta^*),$$

we obtain

$$\dot{\rho} = \widetilde{C} \, \chi(1-\rho^2) + R_\rho, \qquad \dot{\chi} = \widetilde{C} \, \rho(1-\chi^2) + R_\chi,$$

with $|R_\rho| + |R_\chi| \leq C\big(Z^2 + \omega(\theta, \theta^*)\big)$. This proves (51).

Finally, subtract the two expansions. Since

$$\chi(1-\rho^2) - \rho(1-\chi^2) = \chi - \rho - \rho^2\chi + \rho\chi^2 = -(\rho - \chi) - \rho\chi(\rho - \chi) = -(1+\rho\chi)D,$$

we get

$$\dot{D} = -\widetilde{C}(1+\rho\chi)D + R_D,$$

where

$$|R_D| \leq |R_\rho| + |R_\chi| \leq C\big(Z^2 + \omega(\theta, \theta^*)\big).$$

This proves (52). □

Lemma C.7 (Third-moment control at population level)**.** *Let* $\Omega_t := \|\theta_t\|_3^3 = \sum_{i=1}^d |\theta_{t,i}|^3, \quad \omega_t := \sum_{i=1}^d \big(|\theta_{t,i}| + |\theta_i^*|\big)^3$. *Work on the initialization event of Lemma C.1. Then, on the positive-overlap region* $\rho_t, \chi_t \geq 0$*, with* $Z_t := \rho_t + \chi_t$*, the population flow satisfies*

$$\dot{\Omega}_t \leq C\left(Z_t\|\theta^*\|_\infty + Z_t^2\eta_t + \eta_t^2\right). \tag{53}$$

Consequently, suppose that on an interval $0 \leq t \leq T$*, with* $T \leq C_T \log d$*, one has*

$$Z_t^2 \leq A\eta_t, \qquad \eta_t \leq C_\eta \frac{\sqrt{\log d} + 1 + t}{\sqrt{d}}.$$

Then, for all sufficiently large d*,*

$$\Omega_t \leq C_\Omega \frac{1+t}{\sqrt{d}}, \quad \omega_t \leq C_\omega \frac{1+t}{\sqrt{d}}, \quad 0 \leq t \leq T. \tag{54}$$

The constants C_Ω, C_ω *may depend on* K, A, C_η, C_T*, but not on* d*.*

Proof. The population θ-flow is

$$\dot{\theta}_t = -\nabla_{\mathbb{S},\theta} L(\theta_t, \beta_t).$$

Let

$$v_t = (|\theta_{t,i}|\theta_{t,i})_{i=1}^d.$$

Since $x \mapsto |x|^3$ is C^1, we have

$$\frac{d}{dt}\Omega_t = 3\sum_i |\theta_{t,i}|\theta_{t,i}\dot{\theta}_{t,i} = -3\langle v_t, \nabla_{\mathbb{S},\theta} L(\theta_t, \beta_t)\rangle.$$

Moreover,

$$\|v_t\|_1 = \sum_i |\theta_{t,i}|^2 = 1.$$

Hence

$$\dot{\Omega}_t \leq 3\|\nabla_{\mathbb{S},\theta} L(\theta_t, \beta_t)\|_\infty.$$

By Lemma C.4, more precisely by (48), we have

$$\|\nabla_{\mathbb{S},\theta} L(\theta_t, \beta_t)\|_\infty \leq C_1|\rho_t|\|\theta^*\|_\infty + C_2|\rho_t\chi_t|\|\theta_t\|_\infty + C_3\eta_t^2.$$

On the positive-overlap region, $\rho_t, \chi_t \geq 0$, and hence

$$|\rho_t| \leq Z_t, \qquad |\rho_t\chi_t| \leq Z_t^2, \qquad \|\theta_t\|_\infty \leq \eta_t.$$

Therefore

$$\dot{\Omega}_t \leq C\left(Z_t\|\theta^*\|_\infty + Z_t^2\eta_t + \eta_t^2\right),$$

which proves (53).

Now with the conditions

$$Z_t^2 \leq A\eta_t, \qquad \eta_t \leq C_\eta \frac{\sqrt{\log d} + 1 + t}{\sqrt{d}}, \qquad 0 \leq t \leq T \leq C_T \log d.$$

Then

$$Z_t\|\theta^*\|_\infty \leq \sqrt{A\eta_t}\, C_K \sqrt{\frac{\log d}{d}}.$$

Using the envelope for η_t,

$$Z_t\|\theta^*\|_\infty \leq C\sqrt{\frac{\sqrt{\log d} + 1 + t}{\sqrt{d}}}\sqrt{\frac{\log d}{d}}.$$

Since $t \leq C_T \log d$, the right-hand side is bounded by

$$C\frac{\log d}{d^{3/4}} \leq \frac{C}{\sqrt{d}}$$

for all sufficiently large d. Also,

$$Z_t^2 \eta_t \le A\eta_t^2.$$

Therefore, throughout the Phase-I interval,

$$\dot{\Omega}_t \le \frac{C}{\sqrt{d}} + C\eta_t^2.$$

Integrating from 0 to $t \le T$, we get

$$\Omega_t \le \Omega_0 + \frac{Ct}{\sqrt{d}} + C\int_0^t \eta_s^2\, ds.$$

By Lemma C.1,

$$\Omega_0 = \|\theta_0\|_3^3 \le C_K d^{-1/2}.$$

Moreover,

$$\int_0^t \eta_s^2\, ds \le \frac{C}{d}\int_0^t (\sqrt{\log d} + 1 + s)^2\, ds.$$

For $t \le C_T \log d$,

$$\int_0^t (\sqrt{\log d} + 1 + s)^2\, ds \le C(\log d)^3.$$

Thus

$$\int_0^t \eta_s^2\, ds \le \frac{C(\log d)^3}{d} = o(d^{-1/2}).$$

Hence

$$\Omega_t \le C_\Omega \frac{1+t}{\sqrt{d}}, \qquad 0 \le t \le T.$$

Finally, by the elementary inequality

$$(a+b)^3 \le 4a^3 + 4b^3 \qquad (a, b \ge 0),$$

we have

$$\omega_t = \sum_i \big(|\theta_{t,i}| + |\theta_i^*|\big)^3 \le 4\|\theta_t\|_3^3 + 4\|\theta^*\|_3^3.$$

Using the bound on Ω_t and the initialization event

$$\|\theta^*\|_3^3 \le C_K d^{-1/2},$$

we obtain

$$\omega_t \le C_\omega \frac{1+t}{\sqrt{d}}.$$

This proves (54). □

Lemma C.8 (Escape from mediocrity at population level)**.** *Assume Lemmas C.4, C.6, and C.7. Work on the random-spherical delocalization event of Lemma C.1, namely*

$$\|\theta_0\|_\infty + \|\theta^*\|_\infty \le C_K\sqrt{\frac{\log d}{d}}, \qquad \|\theta_0\|_3^3 + \|\theta^*\|_3^3 \le C_K d^{-1/2},$$

and the favorable initialization event defined in (A4). Then there exist constants $c_{\mathrm{esc}}, C_{\mathrm{esc}} > 0$, independent of d, such that for all sufficiently large d, there is a time

$$t_{\mathrm{esc}} \le C_{\mathrm{esc}} \log d \tag{55}$$

with

$$\min\{\rho_{t_{\mathrm{esc}}}, \chi_{t_{\mathrm{esc}}}\} \ge c_{\mathrm{esc}}. \tag{56}$$

Moreover, throughout the escape window,

$$\eta_t \le C\frac{\log d}{\sqrt{d}} = o_d(1). \tag{57}$$

Proof. Fix a large constant $C_T > 0$, to be chosen below, and set

$$T := C_T \log d.$$

Let $A > 0$ be a fixed large constant and define the stopping times

$$\tau_A := \inf\{t : Z_t^2 \geq A\eta_t\},$$
$$\tau_B := \inf\{t : \rho_t \geq 1/2 \text{ or } \chi_t \geq 1/2\},$$
$$\tau_+ := \inf\{t : \rho_t < 0 \text{ or } \chi_t < 0\}.$$

We first work on Phase I, defined as when $t \in [0, \sigma]$ where $\sigma := \tau_A \wedge \tau_B \wedge \tau_+ \wedge T$.

Step 1: Incoherence Control in Phase I We begin with the Phase-I control of η_t. By Lemma C.4, on the positive-overlap region,

$$\dot{\eta}_t \leq CZ_t\|\theta^*\|_\infty + CZ_t^2\eta_t + C\eta_t^2.$$

For $t < \tau_A$, we have $Z_t^2 < A\eta_t$. Therefore

$$Z_t\|\theta^*\|_\infty \leq \sqrt{A\eta_t}\,\|\theta^*\|_\infty.$$

Define

$$H_I(t) := K_I \frac{\sqrt{\log d} + 1 + t}{\sqrt{d}},$$

where K_I is a large constant. We claim that

$$\eta_t \leq H_I(t) \qquad (0 \leq t \leq \sigma).$$

Indeed, at initialization,

$$\eta_0 = \|\theta_0\|_\infty + \|\theta^*\|_\infty \leq C_K\sqrt{\frac{\log d}{d}} \leq \frac{1}{2}H_I(0)$$

after choosing K_I large. Suppose the estimate holds up to a first violation time. Since $t \leq T = C_T \log d$, the bootstrap bound gives

$$\eta_t \leq H_I(t) \leq C\frac{\log d}{\sqrt{d}}.$$

Thus

$$Z_t\|\theta^*\|_\infty \leq \sqrt{AH_I(t)}\,C_K\sqrt{\frac{\log d}{d}} \leq C\frac{\log d}{d^{3/4}} \leq \frac{C}{\sqrt{d}}$$

for all sufficiently large d. Also,

$$Z_t^2\eta_t \leq A\eta_t^2.$$

Consequently, up to the first violation time,

$$\dot{\eta}_t \leq \frac{C}{\sqrt{d}} + C\eta_t^2.$$

Integrating,

$$\eta_t \leq \eta_0 + \frac{Ct}{\sqrt{d}} + C\int_0^t \eta_s^2\,ds.$$

Under the bootstrap $\eta_s \leq H_I(s)$,

$$\int_0^t \eta_s^2\,ds \leq \frac{C}{d}\int_0^{C_T \log d} (\sqrt{\log d} + 1 + s)^2\,ds \leq \frac{C(\log d)^3}{d} = o\left(\frac{\sqrt{\log d} + 1 + t}{\sqrt{d}}\right).$$

Choosing K_I large gives

$$\eta_t \leq \frac{1}{2}H_I(t)$$

at the putative first violation, a contradiction. Hence for $0 \leq t \leq \sigma$,

$$\eta_t \leq H_I(t).$$

So by Lemma C.7, we have for $0 \leq t \leq \sigma$,

$$\Omega_t := \|\theta_t\|_3^3 \leq C_\Omega\frac{1+t}{\sqrt{d}},$$

and

$$\omega_t \leq C_\omega\frac{1+t}{\sqrt{d}}.$$

Step 2: Phase I Positivity, Exponential Pull-Up, and Balancing. We next prove positivity and exponential pull-up in Phase I. On $t < \sigma$, we have

$$0 \le \rho_t, \chi_t \le 1/2.$$

By Lemma C.6, since $0 \le \rho_t, \chi_t \le 1/2$ on the stopped Phase-I interval and d is large, after decreasing $a > 0$ and increasing $b > 0$,

$$\dot{\rho}_t \ge a\chi_t - b\omega_t, \qquad \dot{\chi}_t \ge a\rho_t - b\omega_t.$$

Let

$$u_t := (\rho_t, \chi_t)^\top, \qquad J := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathbf{1} := (1,1)^\top.$$

Then, componentwise,

$$\dot{u}_t \ge aJu_t - b\omega_t \mathbf{1}.$$

Since e^{aJr} has nonnegative entries and

$$e^{aJr}\mathbf{1} = e^{ar}\mathbf{1},$$

variation of constants gives, for $0 \le t \le \sigma$,

$$u_t \ge e^{aJt}u_0 - b\int_0^t e^{aJ(t-s)}\omega_s \mathbf{1}\, ds.$$

Using $\rho_0, \chi_0 \ge z_{\text{init}} d^{-1/2}$, from (A4), we get componentwise

$$e^{aJt}u_0 \ge e^{at}\frac{z_{\text{init}}}{\sqrt{d}}\mathbf{1}.$$

Also,

$$b\int_0^t e^{aJ(t-s)}\omega_s \mathbf{1}\, ds = be^{at}\left(\int_0^t e^{-as}\omega_s\, ds\right)\mathbf{1}.$$

Since

$$\omega_s \le C_\omega \frac{1+s}{\sqrt{d}},$$

we have

$$\int_0^\infty e^{-as}\omega_s\, ds \le \frac{C}{\sqrt{d}}.$$

Let

$$M_\omega := bC_\omega \int_0^\infty e^{-as}(1+s)\, ds.$$

By choosing $z_{\text{init}} > 2M_\omega$, we obtain

$$\rho_t, \chi_t \ge e^{at}\frac{z_{\text{init}} - M_\omega}{\sqrt{d}} \ge \frac{c_-}{\sqrt{d}}e^{at} \qquad (0 \le t \le \sigma),$$

for some $c_- > 0$. In particular,

$$Z_t \ge \frac{2c_-}{\sqrt{d}}e^{at} \qquad (0 \le t \le \sigma),$$

and τ_+ cannot be the first stopping time.

Choose

$$t_0 := C_0 \log d$$

with $C_0 > 0$ sufficiently large, and then choose $C_T > C_0$. Here we show that

$$\sigma \le t_0. \tag{58}$$

Suppose instead that $\sigma > t_0$. Then $t_0 < \tau_A$, so

$$Z_{t_0}^2 < A\eta_{t_0}.$$

But the lower bound above gives

$$Z_{t_0}^2 \geq \frac{c}{d} e^{2at_0} = c\, d^{2aC_0 - 1},$$

whereas the Phase-I η-bound gives

$$A\eta_{t_0} \leq C \frac{\sqrt{\log d} + 1 + t_0}{\sqrt{d}} \leq C \frac{\log d}{\sqrt{d}}.$$

For C_0 sufficiently large, this is impossible for large d. Hence (58) holds.

Since $t_0 < T$ and τ_+ cannot occur first, the first stopping time is either τ_A or τ_B. We claim it must be τ_A. If $\sigma = \tau_B \leq \tau_A$, then by continuity and the definition of τ_B,

$$Z_{\tau_B} \geq \frac{1}{2}.$$

On the other hand, since $\tau_B \leq \tau_A$,

$$Z_{\tau_B}^2 \leq A\eta_{\tau_B} \leq C \frac{\log d}{\sqrt{d}},$$

and hence

$$Z_{\tau_B} \leq C\sqrt{\log d}\, d^{-1/4} = o(1),$$

a contradiction. Therefore

$$\sigma = \tau_A, \qquad \tau_A \leq C_0 \log d, \qquad \tau_A < \tau_B.$$

Since $Z_0 \leq C_{\text{init}} d^{-1/2}$ by assumption and given the lower bound $\eta_0 \geq \|\theta^*\|_\infty \geq d^{-1/2}$, we have $Z_0^2 < A\eta_0$ for all large d. Thus, by continuity,

$$Z_{\tau_A}^2 = A\eta_{\tau_A}.$$

Moreover,

$$\eta_{\tau_A} \geq \|\theta^*\|_\infty \geq d^{-1/2},$$

and from the Phase-I bound,

$$\eta_{\tau_A} \leq C \frac{\log d}{\sqrt{d}}.$$

Therefore

$$Z_{\tau_A} \geq c_A d^{-1/4}, \qquad Z_{\tau_A} \leq C_A \sqrt{\log d}\, d^{-1/4} = o(1),$$

and

$$\eta_{\tau_A} \leq C \frac{\log d}{\sqrt{d}}.$$

We also need the two overlaps to be balanced at τ_A. On $[0, \tau_A]$, by the D-drift consequence (52) of Lemma C.6, we have

$$\dot{D}_t = -\widetilde{C}(1 + \rho_t \chi_t) D_t + O(Z_t^2 + \omega_t). \tag{59}$$

Integrating (59) on $[0, \tau_A]$, and using $Z_t^2 \leq A\eta_t$, gives

$$|D_{\tau_A}| \leq |D_0| + C \int_0^{\tau_A} (Z_s^2 + \omega_s)\, ds.$$

Using

$$|D_0| \leq C_{\text{init}} d^{-1/2}, \qquad \eta_s \leq C \frac{\sqrt{\log d} + 1 + s}{\sqrt{d}}, \qquad \omega_s \leq C \frac{1+s}{\sqrt{d}},$$

and $\tau_A \leq C_0 \log d$, we obtain

$$|D_{\tau_A}| \leq \frac{C}{\sqrt{d}} + C \int_0^{C_0 \log d} \left(\frac{\sqrt{\log d} + 1 + s}{\sqrt{d}} + \frac{1+s}{\sqrt{d}} \right) ds \leq C \frac{(\log d)^2}{\sqrt{d}}.$$

Since $Z_{\tau_A} \geq c_A d^{-1/4}$, we have

$$|D_{\tau_A}| = o(Z_{\tau_A}).$$

In particular, for all sufficiently large d,

$$|D_{\tau_A}| \leq \frac{1}{8} Z_{\tau_A}.$$

Step 3: Phase II Incoherence Control and Drift Pull-Up Exiting Phase I, we next show that Z_t will be pulled from the $d^{-1/4}$-scale to a fixed constant. We define Phase II as when $t \in [\tau_A, t^*]$, where $t^* := \inf\{t \geq \tau_A : Z_t \geq c_Z\}$. Choose a small constant $c_Z > 0$, to be fixed below, and define

$$H := K_2 \frac{\log d}{\sqrt{d}}, \qquad T_2 := \tau_A + C_2 \log d,$$

where K_2, C_2 are large constants. Define the bootstrap stopping time

$$\sigma_2 := \inf \left\{ t \geq \tau_A : \eta_t > H \text{ or } Z_t > 2c_Z \text{ or } |D_t| > \frac{1}{2} Z_t \text{ or } t = T_2 \right\}.$$

At $t = \tau_A$, none of the first three clauses holds for large d. Hence $\sigma_2 > \tau_A$. For $t < \sigma_2$, we have

$$|D_t| \leq \frac{1}{2} Z_t,$$

and therefore

$$\frac{1}{4} Z_t \leq \rho_t, \chi_t \leq \frac{3}{4} Z_t.$$

Also $Z_t \leq 2c_Z$.

We first obtain a crude bound on ω_t on the Phase-II bootstrap interval. From the third-moment differential inequality,

$$\dot{\Omega}_t \leq C \left(Z_t \|\theta^*\|_\infty + Z_t^2 \eta_t + \eta_t^2 \right).$$

For $t < \sigma_2$, we have $Z_t \leq 2c_Z$, $\eta_t \leq H$, and $t - \tau_A \leq C_2 \log d$. Thus

$$\Omega_t \leq \Omega_{\tau_A} + C\|\theta^*\|_\infty \int_{\tau_A}^t Z_s \, ds + C \int_{\tau_A}^t Z_s^2 \eta_s \, ds + C \int_{\tau_A}^t \eta_s^2 \, ds.$$

Using

$$\Omega_{\tau_A} \leq C \frac{1 + \tau_A}{\sqrt{d}} \leq C \frac{\log d}{\sqrt{d}},$$

$$\|\theta^*\|_\infty \leq C_K \sqrt{\frac{\log d}{d}},$$

and the crude bounds $Z_s \leq 2c_Z$, $\eta_s \leq H$, we get

$$\Omega_t \leq C \frac{\log d}{\sqrt{d}} + C c_Z \frac{(\log d)^{3/2}}{\sqrt{d}} + C c_Z^2 \frac{(\log d)^2}{\sqrt{d}} + C \frac{(\log d)^3}{d}.$$

After increasing the constant, this yields

$$\Omega_t \leq C \frac{(\log d)^2}{\sqrt{d}} \qquad (t < \sigma_2).$$

Since $\omega_t \leq 4\Omega_t + 4\|\theta^*\|_3^3$ and $\|\theta^*\|_3^3 \leq C_K d^{-1/2}$ by Lemma C.1, we obtain

$$\omega_t \leq W_d := C \frac{(\log d)^2}{\sqrt{d}} \qquad (t < \sigma_2).$$

We next prove that Z_t grows exponentially on the Phase-II bootstrap interval. Let

$$\tau_- := \inf\{t \geq \tau_A : Z_t < Z_{\tau_A}/2\}.$$

On $[\tau_A, \sigma_2 \wedge \tau_-)$, we have $Z_t \geq Z_{\tau_A}/2$, the balance bounds imply $\rho_t, \chi_t \geq Z_t/4$, and, by choosing c_Z small, $\rho_t, \chi_t \leq 1/2$. Hence the refined drift gives

$$\dot{Z}_t = \dot{\rho}_t + \dot{\chi}_t \geq a_Z Z_t - b_Z \omega_t \geq a_Z Z_t - b_Z W_d.$$

Since

$$\frac{W_d}{Z_{\tau_A}} \leq C \frac{(\log d)^2/\sqrt{d}}{d^{-1/4}} = C \frac{(\log d)^2}{d^{1/4}} = o(1),$$

we have, for all sufficiently large d,

$$b_Z W_d \le \frac{1}{2} a_Z Z_t \qquad \text{on } [\tau_A, \sigma_2 \wedge \tau_-).$$

Thus

$$\dot{Z}_t \ge \mu Z_t \qquad (\tau_A \le t < \sigma_2 \wedge \tau_-)$$

for some $\mu > 0$. This implies $\tau_- > \sigma_2$, and hence

$$Z_t \ge Z_{\tau_A} e^{\mu(t-\tau_A)} \qquad (\tau_A \le t < \sigma_2).$$

In particular, for $s \le t < \sigma_2$,

$$Z_s \le e^{-\mu(t-s)} Z_t.$$

We now close the η-bootstrap before t^*. For $\tau_A \le t < \sigma_2 \wedge t^*$, Lemma C.4 gives

$$\eta_t \le \eta_{\tau_A} + C\|\theta^*\|_\infty \int_{\tau_A}^t Z_s \, ds + C \int_{\tau_A}^t Z_s^2 \eta_s \, ds + C \int_{\tau_A}^t \eta_s^2 \, ds.$$

Since $t < t^*$, we have $Z_t < c_Z$. Using the exponential comparison $Z_s \le e^{-\mu(t-s)} Z_t$, we obtain

$$\int_{\tau_A}^t Z_s \, ds \le C c_Z, \qquad \int_{\tau_A}^t Z_s^2 \, ds \le C c_Z^2.$$

Therefore

$$\eta_t \le C \frac{\log d}{\sqrt{d}} + C c_Z \sqrt{\frac{\log d}{d}} + C c_Z^2 H + C H^2 \log d.$$

Choose $c_Z > 0$ small enough that

$$C c_Z^2 H \le \frac{1}{8} H.$$

Then choose K_2 large enough so that

$$C \frac{\log d}{\sqrt{d}} \le \frac{1}{4} H.$$

Moreover,

$$C c_Z \sqrt{\frac{\log d}{d}} = o(H), \qquad H^2 \log d = K_2^2 \frac{(\log d)^3}{d} = o(H).$$

Thus, for all sufficiently large d,

$$\eta_t \le \frac{1}{2} H \qquad (\tau_A \le t < \sigma_2 \wedge t^*).$$

By a first-violation argument, the clause $\eta_t > H$ cannot trigger before t^*.

Next we close the D/Z-balance bootstrap. On the balanced small-overlap band, the same subtraction estimate as above gives

$$\dot{D}_t = -\widetilde{C}(1 + \rho_t \chi_t) D_t + O(Z_t^2 + \omega_t).$$

Since c_Z is small, $1 + \rho_t \chi_t \ge 1/2$. Hence, for $\tau_A \le t < \sigma_2 \wedge t^*$,

$$|D_t| \le e^{-c(t-\tau_A)} |D_{\tau_A}| + C \int_{\tau_A}^t e^{-c(t-s)} (Z_s^2 + \omega_s) \, ds.$$

The entry term is bounded by

$$e^{-c(t-\tau_A)} |D_{\tau_A}| \le |D_{\tau_A}| \le \frac{1}{8} Z_{\tau_A} \le \frac{1}{8} Z_t.$$

The ω-term is bounded by

$$C W_d = o(Z_{\tau_A}) \le o(Z_t).$$

For the Z_s^2-term, using

$$Z_s \le e^{-\mu(t-s)} Z_t,$$

we get

$$\int_{\tau_A}^{t} e^{-c(t-s)} Z_s^2 \, ds \le C Z_t^2 \le C c_Z Z_t.$$

Choosing c_Z smaller if necessary, and then d large, gives

$$|D_t| \le \frac{3}{8} Z_t < \frac{1}{2} Z_t \qquad (\tau_A \le t < \sigma_2 \wedge t^*).$$

Thus the clause $|D_t| > Z_t/2$ cannot trigger before t^*.

We now show that $t^* \le T_2$. Suppose, for contradiction, that $t^* > T_2$. Then $Z_t < c_Z$ for all $t \le T_2$. The clauses $\eta_t > H$ and $|D_t| > Z_t/2$ cannot trigger before t^*, as just shown, and the clause $Z_t > 2c_Z$ cannot trigger while $Z_t < c_Z$. Hence

$$T_2 < \sigma_2.$$

The exponential lower bound then gives

$$Z_{T_2} \ge Z_{\tau_A} e^{\mu C_2 \log d} \ge c_A d^{-1/4} d^{\mu C_2}.$$

Choosing $C_2 > 1/(4\mu)$ large enough gives

$$Z_{T_2} \ge c_Z$$

for all sufficiently large d, contradicting $t^* > T_2$. Therefore

$$t^* \le T_2.$$

By continuity,

$$Z_{t^*} = c_Z.$$

The bootstrap balance gives

$$|D_{t^*}| \le \frac{1}{2} Z_{t^*}.$$

Therefore

$$\min\{\rho_{t^*}, \chi_{t^*}\} = \frac{Z_{t^*} - |D_{t^*}|}{2} \ge \frac{c_Z}{4}.$$

Set

$$t_{\mathrm{esc}} := t^*, \qquad c_{\mathrm{esc}} := \frac{c_Z}{4}.$$

Since

$$\tau_A \le C_0 \log d$$

and

$$t^* \le \tau_A + C_2 \log d,$$

we have

$$t_{\mathrm{esc}} \le (C_0 + C_2) \log d.$$

Thus the theorem holds with

$$C_{\mathrm{esc}} := C_0 + C_2.$$

Finally, the escape-window η_t-bound follows by combining the Phase-I estimate

$$\eta_t \le K_I \frac{\sqrt{\log d} + 1 + t}{\sqrt{d}} \le C \frac{\log d}{\sqrt{d}} \qquad (0 \le t \le \tau_A)$$

with the Phase-II bootstrap bound

$$\eta_t \le H = K_2 \frac{\log d}{\sqrt{d}} \qquad (\tau_A \le t \le t^*).$$

Hence, throughout weak recovery window $0 \le t \le t_{\mathrm{esc}}$,

$$\eta_t \le C \frac{\log d}{\sqrt{d}} = o_d(1).$$

This completes the proof. □

C.4 Strong Recovery

Building on Lemma C.8, we show that after weak recovery the population dynamics enters the fixed local strong-convexity cap around the teacher in an additional $O(1)$ time. Once inside this cap, local strong convexity gives exponential contraction to the teacher, achiving strong recovery.

Lemma C.9 (Entry into the local cap after weak recovery)**.** *Let $r_{\mathrm{loc}} > 0$ be the radius from Lemma C.14. After possibly shrinking r_{loc}, assume*

$$r_{\mathrm{loc}} < \frac{\pi}{2}.$$

Set

$$s_{\mathrm{loc}} := \cos(r_{\mathrm{loc}}/4).$$

Assume C.8 with t_{esc}, c_{esc}, and C_{esc}. Under the same assumptions as Lemmas C.3 and C.4, for all sufficiently large d, there exists a time

$$t_{\mathrm{loc}} \le t_{\mathrm{esc}} + C_{\mathrm{in}}$$

where $C_{\mathrm{in}} < \infty$ is independent of d, such that

$$\rho_{t_{\mathrm{loc}}} \ge s_{\mathrm{loc}}, \qquad \chi_{t_{\mathrm{loc}}} \ge s_{\mathrm{loc}}.$$

Consequently,

$$\mathrm{dist}_{\mathcal{M}}(y_{t_{\mathrm{loc}}}, y^*)^2 \le \frac{r_{\mathrm{loc}}^2}{8}.$$

In particular, since $t_{\mathrm{esc}} \le C_{\mathrm{esc}} \log d$,

$$t_{\mathrm{loc}} \le C_{\mathrm{loc}} \log d$$

for some constant $C_{\mathrm{loc}} < \infty$ independent of d.

Proof. The proof is deterministic on the favorable event supplied by Lemma C.8. By that lemma, there are constants $c_{\mathrm{esc}} \in (0,1)$ and $C_{\mathrm{esc}} < \infty$, independent of d, and a time t_{esc} such that

$$t_{\mathrm{esc}} \le C_{\mathrm{esc}} \log d, \qquad \rho_{t_{\mathrm{esc}}} \ge c_{\mathrm{esc}}, \qquad \chi_{t_{\mathrm{esc}}} \ge c_{\mathrm{esc}}, \qquad \eta_{t_{\mathrm{esc}}} = o_d(1).$$

Set

$$c_- := \frac{c_{\mathrm{esc}}}{2}.$$

Since $r_{\mathrm{loc}} > 0$, we have

$$s_{\mathrm{loc}} = \cos(r_{\mathrm{loc}}/4) < 1.$$

Hence

$$m_{\mathrm{loc}} := 1 - s_{\mathrm{loc}}^2 > 0.$$

By Lemma C.3, on the positive-overlap region $\rho_t, \chi_t \ge 0$, the flow satisfies

$$\dot\rho_t \ge a_\rho \chi_t (1 - \rho_t^2) - b_\rho \eta_t,$$

$$\dot\chi_t \ge a_\chi \rho_t (1 - \chi_t^2) - b_\chi \eta_t,$$

where $a_\rho, a_\chi, b_\rho, b_\chi > 0$ are independent of d.

Choose $\varepsilon_* > 0$, independent of d, so small that

$$b_\rho \varepsilon_* \le \frac{1}{2} a_\rho c_- m_{\mathrm{loc}}, \qquad b_\chi \varepsilon_* \le \frac{1}{2} a_\chi c_- m_{\mathrm{loc}},$$

and also

$$b_\rho \varepsilon_* \le \frac{1}{2} a_\rho c_- (1 - c_-^2), \qquad b_\chi \varepsilon_* \le \frac{1}{2} a_\chi c_- (1 - c_-^2).$$

Define

$$v_\rho := \frac{1}{2} a_\rho c_- m_{\mathrm{loc}}, \qquad v_\chi := \frac{1}{2} a_\chi c_- m_{\mathrm{loc}}.$$

Let

$$T_\rho := \frac{(s_{\mathrm{loc}} - c_{\mathrm{esc}})_+}{v_\rho}, \qquad T_\chi := \frac{(s_{\mathrm{loc}} - c_{\mathrm{esc}})_+}{v_\chi},$$

and set

$$C_{\mathrm{in}} := 1 + \max\{T_\rho, T_\chi\}.$$

This constant depends only on the fixed population constants and on r_{loc}, and therefore is independent of d.

We first show that η_t remains smaller than ε_* on the fixed window

$$[t_{\mathrm{esc}}, t_{\mathrm{esc}} + C_{\mathrm{in}}].$$

By Lemma C.4,

$$\dot{\eta}_t \le C_1|\rho_t|\|\theta^*\|_\infty + C_2|\chi_t\rho_t|\eta_t + C_3\eta_t^2.$$

Since $|\rho_t|, |\chi_t| \le 1$, on the present fixed window this gives

$$\dot{\eta}_t \le A_d + B\eta_t + C\eta_t^2,$$

where

$$A_d := C_1\|\theta^*\|_\infty \le C\sqrt{\frac{\log d}{d}} = o_d(1),$$

and $B, C < \infty$ are independent of d.

Define

$$\tau_\eta := \inf\{t \ge t_{\mathrm{esc}} : \eta_t \ge \varepsilon_*\} \wedge (t_{\mathrm{esc}} + C_{\mathrm{in}}).$$

Since $\eta_{t_{\mathrm{esc}}} = o_d(1)$, for all sufficiently large d,

$$\eta_{t_{\mathrm{esc}}} < \varepsilon_*.$$

Thus $\tau_\eta > t_{\mathrm{esc}}$. On $[t_{\mathrm{esc}}, \tau_\eta]$, we have $\eta_t \le \varepsilon_*$, and therefore for almost every $t \in [t_{\mathrm{esc}}, \tau_\eta]$,

$$\dot{\eta}_t \le A_d + (B + C\varepsilon_*)\eta_t.$$

Set

$$B_* := B + C\varepsilon_*.$$

By Gronwall's inequality, for every $t \in [t_{\mathrm{esc}}, \tau_\eta]$,

$$\eta_t \le e^{B_*(t-t_{\mathrm{esc}})}\eta_{t_{\mathrm{esc}}} + A_d\int_{t_{\mathrm{esc}}}^{t} e^{B_*(t-u)}\,du.$$

Hence

$$\eta_t \le e^{B_*C_{\mathrm{in}}}\eta_{t_{\mathrm{esc}}} + A_dC_{\mathrm{in}}e^{B_*C_{\mathrm{in}}}.$$

The right-hand side is $o_d(1)$. Therefore, for all sufficiently large d,

$$\eta_t < \varepsilon_* \qquad \text{for all } t \in [t_{\mathrm{esc}}, \tau_\eta].$$

By the definition of τ_η, this forces

$$\tau_\eta = t_{\mathrm{esc}} + C_{\mathrm{in}}.$$

Consequently,

$$\eta_t \le \varepsilon_* \qquad \text{for all } t \in [t_{\mathrm{esc}}, t_{\mathrm{esc}} + C_{\mathrm{in}}].$$

Next we prove that the rectangle

$$\mathcal{R}_- := \{(\rho, \chi) : \rho \ge c_-,\ \chi \ge c_-\}$$

is forward invariant on the same window. At time t_{esc},

$$\rho_{t_{\mathrm{esc}}} \ge c_{\mathrm{esc}} > c_-, \qquad \chi_{t_{\mathrm{esc}}} \ge c_{\mathrm{esc}} > c_-.$$

Suppose, for contradiction, that the trajectory exits $\mathcal{R}_-$ before $t_{\mathrm{esc}} + C_{\mathrm{in}}$. Let

$$\tau_- := \inf\left\{t \in [t_{\mathrm{esc}}, t_{\mathrm{esc}} + C_{\mathrm{in}}] : \min\{\rho_t, \chi_t\} \le c_-\right\}.$$

Then $\tau_- < \infty$, and by continuity,

$$\rho_{\tau_-} \ge c_-, \qquad \chi_{\tau_-} \ge c_-,$$

with at least one equality.

If $\rho_{\tau_-} = c_-$, then $\rho_{\tau_-}, \chi_{\tau_-} \geq 0$, so Lemma C.3 applies. Using also $\eta_{\tau_-} \leq \varepsilon_*$, we get

$$\dot{\rho}_{\tau_-} \geq a_\rho c_-(1 - c_-^2) - b_\rho \varepsilon_* \geq \frac{1}{2} a_\rho c_-(1 - c_-^2) > 0.$$

Thus ρ_t cannot cross downward through c_- at τ_-. Similarly, if $\chi_{\tau_-} = c_-$, then

$$\dot{\chi}_{\tau_-} \geq a_\chi c_-(1 - c_-^2) - b_\chi \varepsilon_* \geq \frac{1}{2} a_\chi c_-(1 - c_-^2) > 0,$$

so χ_t cannot cross downward through c_-. This contradicts the definition of τ_-. Therefore,

$$\rho_t, \chi_t \geq c_- \qquad \text{for all } t \in [t_{\text{esc}}, t_{\text{esc}} + C_{\text{in}}].$$

We now show that ρ_t reaches s_{loc} within time T_ρ and then cannot fall below s_{loc} during the window. Define

$$\tau_\rho := \inf\{t \geq t_{\text{esc}} : \rho_t \geq s_{\text{loc}}\}.$$

If $\rho_{t_{\text{esc}}} \geq s_{\text{loc}}$, then $\tau_\rho = t_{\text{esc}}$. Otherwise, before ρ_t reaches s_{loc}, we have $\rho_t < s_{\text{loc}}$. On this interval, the rectangle invariance gives $\chi_t \geq c_-$, and the previous step gives $\eta_t \leq \varepsilon_*$. Hence Lemma C.3 implies

$$\dot{\rho}_t \geq a_\rho c_-(1 - \rho_t^2) - b_\rho \varepsilon_*.$$

Since $\rho_t < s_{\text{loc}}$,

$$1 - \rho_t^2 \geq 1 - s_{\text{loc}}^2 = m_{\text{loc}}.$$

Therefore

$$\dot{\rho}_t \geq a_\rho c_- m_{\text{loc}} - b_\rho \varepsilon_* \geq \frac{1}{2} a_\rho c_- m_{\text{loc}} = v_\rho.$$

Thus ρ_t reaches s_{loc} no later than

$$t_{\text{esc}} + T_\rho.$$

Moreover, after ρ_t reaches s_{loc}, it cannot cross downward through s_{loc} inside the same window. Indeed, at a hypothetical first downward crossing through s_{loc}, we would have

$$\rho_t = s_{\text{loc}}, \qquad \chi_t \geq c_-, \qquad \eta_t \leq \varepsilon_*.$$

Therefore

$$\dot{\rho}_t \geq a_\rho c_-(1 - s_{\text{loc}}^2) - b_\rho \varepsilon_* \geq v_\rho > 0,$$

which is incompatible with a first downward crossing.

The same argument applies to χ_t. Define

$$\tau_\chi := \inf\{t \geq t_{\text{esc}} : \chi_t \geq s_{\text{loc}}\}.$$

If $\chi_{t_{\text{esc}}} \geq s_{\text{loc}}$, then $\tau_\chi = t_{\text{esc}}$. Otherwise, before χ_t reaches s_{loc}, we have $\chi_t < s_{\text{loc}}$, while $\rho_t \geq c_-$ and $\eta_t \leq \varepsilon_*$. Hence

$$\dot{\chi}_t \geq a_\chi c_-(1 - \chi_t^2) - b_\chi \varepsilon_* \geq a_\chi c_- m_{\text{loc}} - b_\chi \varepsilon_* \geq v_\chi.$$

Thus χ_t reaches s_{loc} no later than

$$t_{\text{esc}} + T_\chi.$$

Once it reaches s_{loc}, it cannot cross downward through s_{loc}, because at such a crossing the same estimate gives

$$\dot{\chi}_t \geq v_\chi > 0.$$

Now set

$$t_{\text{loc}} := \max\{\tau_\rho, \tau_\chi\}.$$

By the previous estimates,

$$t_{\text{loc}} \leq t_{\text{esc}} + \max\{T_\rho, T_\chi\} \leq t_{\text{esc}} + C_{\text{in}},$$

and by the no-downcrossing arguments,

$$\rho_{t_{\text{loc}}} \geq s_{\text{loc}}, \qquad \chi_{t_{\text{loc}}} \geq s_{\text{loc}}.$$

It remains to convert the overlap bounds into a product-manifold distance bound. Since

$$\rho_{t_{\mathrm{loc}}} = \langle \beta_{t_{\mathrm{loc}}}, \beta^* \rangle \geq s_{\mathrm{loc}} = \cos(r_{\mathrm{loc}}/4),$$

and since $\arccos$ is decreasing on $[-1, 1]$,

$$\mathrm{dist}_{\mathbb{S}^{d-1}}(\beta_{t_{\mathrm{loc}}}, \beta^*) = \arccos(\rho_{t_{\mathrm{loc}}}) \leq \frac{r_{\mathrm{loc}}}{4}.$$

Similarly,

$$\mathrm{dist}_{\mathbb{S}^{d-1}}(\theta_{t_{\mathrm{loc}}}, \theta^*) = \arccos(\chi_{t_{\mathrm{loc}}}) \leq \frac{r_{\mathrm{loc}}}{4}.$$

Using the product metric on $\mathcal{M} = \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, we get

$$\mathrm{dist}_{\mathcal{M}}(y_{t_{\mathrm{loc}}}, y^*)^2 = \mathrm{dist}_{\mathbb{S}^{d-1}}(\theta_{t_{\mathrm{loc}}}, \theta^*)^2 + \mathrm{dist}_{\mathbb{S}^{d-1}}(\beta_{t_{\mathrm{loc}}}, \beta^*)^2.$$

Therefore,

$$\mathrm{dist}_{\mathcal{M}}(y_{t_{\mathrm{loc}}}, y^*)^2 \leq \left(\frac{r_{\mathrm{loc}}}{4}\right)^2 + \left(\frac{r_{\mathrm{loc}}}{4}\right)^2 = \frac{r_{\mathrm{loc}}^2}{8}.$$

Finally, since Lemma C.8 gives $t_{\mathrm{esc}} \leq C_{\mathrm{esc}} \log d$, we have

$$t_{\mathrm{loc}} \leq C_{\mathrm{esc}} \log d + C_{\mathrm{in}}.$$

For $d \geq e$,

$$C_{\mathrm{in}} \leq C_{\mathrm{in}} \log d,$$

and hence

$$t_{\mathrm{loc}} \leq (C_{\mathrm{esc}} + C_{\mathrm{in}}) \log d.$$

Thus the last claim holds with

$$C_{\mathrm{loc}} := C_{\mathrm{esc}} + C_{\mathrm{in}}.$$

The proof is complete. □

Curvature near the teacher. Here, we prove the local curvature estimate used for the contraction phase. The comparison with the Gaussian model is needed only at the teacher; the extension to a small cap uses Hessian continuity of the true loss.

Lemma C.10 (Gaussian local strong convexity)**.** *Let $\mathcal{M} := \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}, y = (\theta, \beta), y^* = (\theta^*, \beta^*)$. Let L_G be the Gaussian population loss*

$$L_G(\theta, \beta) = \mathbb{E}\left[(\psi(g \cdot \theta) - \psi(g^* \cdot \theta^*))^2\right],$$

where $(g, g^*) \in \mathbb{R}^d \times \mathbb{R}^d$ is centered Gaussian with $\mathbb{E}[g_i g_j] = \delta_{ij}, \mathbb{E}[g_i^* g_j^*] = \delta_{ij}, \mathbb{E}[g_i g_j^*] = \kappa(\rho)\delta_{ij}$, with

$$\rho := \beta \cdot \beta^*, \qquad \kappa(\rho) := \sum_{\ell \geq 1} a_\ell^2 \rho^\ell.$$

Assume $\|\phi\|_{L^2(\gamma)} = 1$, so $\kappa(1) = 1$, and assume the information-exponent-one conditions $\alpha_1 \neq 0, a_1 \neq 0$, where

$$\psi = \sum_{k \geq 0} \alpha_k h_k, \qquad \phi = \sum_{\ell \geq 1} a_\ell h_\ell.$$

Assume also enough regularity so that the function below is C^2 in a neighborhood of $(1, 1)$.

Then there exist constants $r_G > 0$ and $\mu_G > 0$, independent of d, such that for every

$$(\theta, \beta) \in \mathcal{U}_{r_G} := \left\{(\theta, \beta) : \mathrm{dist}_{\mathbb{S}}(\theta, \theta^*)^2 + \mathrm{dist}_{\mathbb{S}}(\beta, \beta^*)^2 \leq r_G^2\right\},$$

and every product tangent vector $\xi = (u, v) \in T_\theta \mathbb{S}^{d-1} \times T_\beta \mathbb{S}^{d-1}$, one has

$$\mathrm{Hess}_{\mathcal{M}} L_G(\theta, \beta)[\xi, \xi] \geq \mu_G \left(\|u\|^2 + \|v\|^2\right). \tag{60}$$

Proof. In the Gaussian model, the scalar pair $(g \cdot \theta,\ g^* \cdot \theta^*)$ is centered Gaussian with unit variances and correlation $m = \kappa(\rho)\chi$. Therefore, using the Hermite expansion of ψ,

$$\mathbb{E}[\psi(g \cdot \theta)\psi(g^* \cdot \theta^*)] = \sum_{k \geq 0} \alpha_k^2 (\kappa(\rho)\chi)^k.$$

Since both marginals have variance one,

$$\mathbb{E}[\psi(g \cdot \theta)^2] = \mathbb{E}[\psi(g^* \cdot \theta^*)^2] = \sum_{k \geq 0} \alpha_k^2.$$

Hence

$$L_G(\theta, \beta) = 2 \sum_{k \geq 0} \alpha_k^2 \left[1 - (\kappa(\rho)\chi)^k\right], \tag{61}$$

with $k = 0$ term is zero. We define

$$\Phi(\chi, \rho) := 2 \sum_{k \geq 1} \alpha_k^2 \left[1 - (\kappa(\rho)\chi)^k\right] = L_G(\theta, \beta).$$

Now we first compute the Hessian at the planted point $(\theta, \beta) = (\theta^*, \beta^*)$: we have $\chi = 1, \rho = 1, \kappa(1) = 1$. Also $\Phi_\chi(1, 1) = -2 \sum_{k \geq 1} k\alpha_k^2$, and $\Phi_\rho(1, 1) = -2\kappa'(1) \sum_{k \geq 1} k\alpha_k^2$. Define

$$\lambda_\theta := -\Phi_\chi(1, 1) = 2 \sum_{k \geq 1} k\alpha_k^2,$$

and

$$\lambda_\beta := -\Phi_\rho(1, 1) = 2\kappa'(1) \sum_{k \geq 1} k\alpha_k^2.$$

The information-exponent-one condition $\alpha_1 \neq 0$ implies $\lambda_\theta > 0$. Moreover,

$$\kappa'(1) = \sum_{\ell \geq 1} \ell a_\ell^2 > 0,$$

because $a_1 \neq 0$. Hence $\lambda_\beta > 0$.

Now let $u \in T_{\theta^*}\mathbb{S}^{d-1}, v \in T_{\beta^*}\mathbb{S}^{d-1}$ and let (θ_t, β_t) be the product geodesic with

$$(\theta_0, \beta_0) = (\theta^*, \beta^*), \qquad (\dot{\theta}_0, \dot{\beta}_0) = (u, v).$$

Then

$$\chi(t) := \theta_t \cdot \theta^* = \cos(t\|u\|),$$

and

$$\rho(t) := \beta_t \cdot \beta^* = \cos(t\|v\|).$$

Thus

$$\chi'(0) = 0, \qquad \chi''(0) = -\|u\|^2,$$

and

$$\rho'(0) = 0, \qquad \rho''(0) = -\|v\|^2.$$

Since the curve is a product geodesic, the Riemannian Hessian is computed by the second derivative along this curve:

$$\mathrm{Hess}_{\mathcal{M}} L_G(y^*)[(u, v), (u, v)] = \left.\frac{d^2}{dt^2}\right|_{t=0} \Phi(\chi(t), \rho(t)).$$

By the chain rule and the fact that $\chi'(0) = \rho'(0) = 0$,

$$\left.\frac{d^2}{dt^2}\right|_{t=0} \Phi(\chi(t), \rho(t)) = \Phi_\chi(1, 1)\chi''(0) + \Phi_\rho(1, 1)\rho''(0).$$

Therefore

$$\mathrm{Hess}_{\mathcal{M}} L_G(y^*)[(u, v), (u, v)] = \lambda_\theta \|u\|^2 + \lambda_\beta \|v\|^2.$$

Consequently, with

$$\lambda_0 := \min\{\lambda_\theta, \lambda_\beta\} > 0,$$

we have

$$\mathrm{Hess}_{\mathcal{M}} L_G(y^*)[(u,v),(u,v)] \geq \lambda_0(\|u\|^2 + \|v\|^2).$$

It remains to extend this lower bound to a small product cap around y^*. We do this explicitly.

Let

$$(\theta,\beta) \in \mathcal{M}, \qquad u \in T_\theta \mathbb{S}^{d-1}, \qquad v \in T_\beta \mathbb{S}^{d-1}.$$

Let (θ_t, β_t) be the product geodesic starting from (θ,β) with initial velocity (u,v). Then

$$\chi'(0) = u \cdot \theta^*, \qquad \chi''(0) = -\chi\|u\|^2,$$

and

$$\rho'(0) = v \cdot \beta^*, \qquad \rho''(0) = -\rho\|v\|^2.$$

Therefore

$$\mathrm{Hess}_{\mathcal{M}} L_G(\theta,\beta)[(u,v),(u,v)] \tag{62}$$

$$= \Phi_{\chi\chi}(\chi,\rho)(u\cdot\theta^*)^2 + 2\Phi_{\chi\rho}(\chi,\rho)(u\cdot\theta^*)(v\cdot\beta^*) + \Phi_{\rho\rho}(\chi,\rho)(v\cdot\beta^*)^2 - \chi\Phi_\chi(\chi,\rho)\|u\|^2 - \rho\Phi_\rho(\chi,\rho)\|v\|^2. \tag{63}$$

We now estimate the terms in this expression. Choose a small neighborhood

$$\mathcal{N} := \{(\chi,\rho) : |\chi - 1| \leq r_0,\ |\rho - 1| \leq r_0\}$$

on which all second derivatives of Φ are bounded:

$$|\Phi_{\chi\chi}| + |\Phi_{\chi\rho}| + |\Phi_{\rho\rho}| \leq M.$$

By continuity of Φ_χ, Φ_ρ, after shrinking r_0 if necessary, we also have

$$-\chi\Phi_\chi(\chi,\rho) \geq \frac{\lambda_\theta}{2}, \qquad -\rho\Phi_\rho(\chi,\rho) \geq \frac{\lambda_\beta}{2}$$

throughout $\mathcal{N}$.

Now suppose

$$\mathrm{dist}_{\mathbb{S}}(\theta,\theta^*) \leq r, \qquad \mathrm{dist}_{\mathbb{S}}(\beta,\beta^*) \leq r,$$

with $r > 0$ small enough so that $(\chi,\rho) \in \mathcal{N}$. Since $u \perp \theta$,

$$|u\cdot\theta^*| = |u \cdot P_{\theta^\perp}\theta^*| \leq \|u\|\|P_{\theta^\perp}\theta^*\| = \|u\|\sqrt{1-\chi^2} \leq r\|u\|.$$

Similarly,

$$|v\cdot\beta^*| \leq r\|v\|.$$

Hence the second-derivative terms are bounded in absolute value by

$$Mr^2\|u\|^2 + 2Mr^2\|u\|\|v\| + Mr^2\|v\|^2 \leq 2Mr^2(\|u\|^2 + \|v\|^2).$$

Therefore,

$$\mathrm{Hess}_{\mathcal{M}} L_G(\theta,\beta)[(u,v),(u,v)] \geq \frac{\lambda_\theta}{2}\|u\|^2 + \frac{\lambda_\beta}{2}\|v\|^2 - 2Mr^2(\|u\|^2 + \|v\|^2).$$

Choose $r_G > 0$ small enough that $2Mr_G^2 \leq \frac{1}{4}\min\{\lambda_\theta, \lambda_\beta\}$. Then, for every $r \leq r_G$,

$$\mathrm{Hess}_{\mathcal{M}} L_G(\theta,\beta)[(u,v),(u,v)] \geq \frac{1}{4}\min\{\lambda_\theta,\lambda_\beta\}(\|u\|^2 + \|v\|^2).$$

Thus the lemma holds with

$$\mu_G := \frac{1}{4}\min\{\lambda_\theta,\lambda_\beta\} > 0.$$

□

Lemma C.11 (Teacher Hessian block formula)**.** *Let $y^* := (\theta^*, \beta^*)$. Define*

$$G_i := x^{(i)} \cdot \beta^*, \qquad S := \sum_{i=1}^{d} \theta_i^* \phi(G_i).$$

For $u \in T_{\theta^*}\mathbb{S}^{d-1}$, set

$$A_u := \sum_{i=1}^{d} u_i \phi(G_i).$$

Then, for $u_1, u_2 \in T_{\theta^*}\mathbb{S}^{d-1}$ and $v_1, v_2 \in T_{\beta^*}\mathbb{S}^{d-1}$, the Hessian blocks of L at the teacher satisfy

$$H_{L,\theta\theta}(y^*)[u_1, u_2] = 2\mathbb{E}\left[\psi'(S)^2 A_{u_1} A_{u_2}\right], \tag{64}$$

$$H_{L,\theta\beta}(y^*)[u, v] = 0, \tag{65}$$

$$H_{L,\beta\beta}(y^*)[v_1, v_2] = 2\langle v_1, v_2\rangle \sum_{i=1}^{d} (\theta_i^*)^2 \mathbb{E}\left[\psi'(S)^2 \phi'(G_i)^2\right]. \tag{66}$$

Consequently, for every $(u, v) \in T_{\theta^*}\mathbb{S}^{d-1} \times T_{\beta^*}\mathbb{S}^{d-1}$,

$$\mathrm{Hess}_{\mathcal{M}} L(y^*)[(u, v), (u, v)] = 2\mathbb{E}\left[\psi'(S)^2 A_u^2\right] + 2\|v\|^2 \sum_{i=1}^{d} (\theta_i^*)^2 \mathbb{E}\left[\psi'(S)^2 \phi'(G_i)^2\right]. \tag{67}$$

Proof. We first compute the full quadratic form at the teacher and then polarize.

Let $y_t = (\theta_t, \beta_t)$ be any product geodesic on $\mathcal{M} = \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ with

$$y_0 = y^*, \qquad \dot{y}_0 = (u, v).$$

Write

$$f_t(x) := f_{\beta_t, \theta_t}(x), \qquad f_*(x) := f_{\beta^*, \theta^*}(x).$$

Since $f_0 = f_*$ pointwise,

$$L(y_t) = \mathbb{E}[(f_t - f_*)^2],$$

and therefore

$$\frac{d^2}{dt^2}\Big|_{t=0} L(y_t) = 2\mathbb{E}[\dot{f}_0^2] + 2\mathbb{E}[(f_0 - f_*)\ddot{f}_0] = 2\mathbb{E}[\dot{f}_0^2].$$

Because y_t is a geodesic, this second derivative is exactly the Riemannian Hessian:

$$\mathrm{Hess}_{\mathcal{M}} L(y^*)[(u, v), (u, v)] = 2\mathbb{E}[\dot{f}_0^2].$$

Now compute $\dot{f}_0$. At y^*,

$$f_t(x) = \psi\left(\sum_{i=1}^{d} \theta_{t,i} \phi(x^{(i)} \cdot \beta_t)\right).$$

Thus

$$\dot{f}_0 = \psi'(S)\left[\sum_{i=1}^{d} u_i \phi(G_i) + \sum_{i=1}^{d} \theta_i^* \phi'(G_i)(x^{(i)} \cdot v)\right].$$

Define, for the proof,

$$B_v := \sum_{i=1}^{d} \theta_i^* \phi'(G_i) H_i(v), \qquad H_i(v) := x^{(i)} \cdot v.$$

Then

$$\dot{f}_0 = \psi'(S)(A_u + B_v),$$

and hence

$$\mathrm{Hess}_{\mathcal{M}} L(y^*)[(u, v), (u, v)] = 2\mathbb{E}\left[\psi'(S)^2 (A_u + B_v)^2\right]. \tag{68}$$

Since the right-hand side of (68) is a quadratic form in (u, v), its associated symmetric bilinear form is obtained by polarization:

$$\mathrm{Hess}_{\mathcal{M}} L(y^*)[(u_1, v_1), (u_2, v_2)] = 2\mathbb{E}\big[\psi'(S)^2(A_{u_1} + B_{v_1})(A_{u_2} + B_{v_2})\big]. \tag{69}$$

Now, we derive the three Hessian blocks: First, take $v_1 = v_2 = 0$ in (69). This gives

$$H_{L,\theta\theta}(y^*)[u_1, u_2] = 2\mathbb{E}\big[\psi'(S)^2 A_{u_1} A_{u_2}\big],$$

which proves (64), the $\theta\theta$-block.

Second, take $(u_1, v_1) = (u, 0)$ and $(u_2, v_2) = (0, v)$. Then

$$H_{L,\theta\beta}(y^*)[u, v] = 2\mathbb{E}\big[\psi'(S)^2 A_u B_v\big].$$

We show this expectation is zero. Since $v \in T_{\beta^*}\mathbb{S}^{d-1}$, we have $v \perp \beta^*$. For each i, the pair

$$G_i = x^{(i)} \cdot \beta^*, \qquad H_i(v) = x^{(i)} \cdot v$$

is jointly Gaussian with covariance

$$\mathbb{E}[G_i H_i(v)] = \langle \beta^*, v \rangle = 0.$$

Hence G_i and $H_i(v)$ are independent. Moreover, conditional on $(G_1, \ldots, G_d)$, the variables $H_i(v)$ are centered Gaussian and independent across i. Therefore

$$\mathbb{E}[B_v \mid G_1, \ldots, G_d] = \sum_{i=1}^{d} \theta_i^* \phi'(G_i) \mathbb{E}[H_i(v) \mid G_1, \ldots, G_d] = 0.$$

Since $\psi'(S)^2 A_u$ is measurable with respect to $\sigma(G_1, \ldots, G_d)$, we obtain

$$\mathbb{E}[\psi'(S)^2 A_u B_v] = \mathbb{E}\big[\psi'(S)^2 A_u \mathbb{E}[B_v \mid G_1, \ldots, G_d]\big] = 0.$$

Thus we have (65), the $\theta\beta$-block:

$$H_{L,\theta\beta}(y^*)[u, v] = 0.$$

Finally, take $u_1 = u_2 = 0$ in (69). Then

$$H_{L,\beta\beta}(y^*)[v_1, v_2] = 2\mathbb{E}\big[\psi'(S)^2 B_{v_1} B_{v_2}\big].$$

Conditioning again on $G := (G_1, \ldots, G_d)$,

$$B_{v_1} B_{v_2} = \sum_{i,j=1}^{d} \theta_i^* \theta_j^* \phi'(G_i) \phi'(G_j) H_i(v_1) H_j(v_2).$$

For $i \neq j$, conditional independence and centering give

$$\mathbb{E}[H_i(v_1) H_j(v_2) \mid G] = 0.$$

For $i = j$, since $v_1, v_2 \perp \beta^*$,

$$\mathbb{E}[H_i(v_1) H_i(v_2) \mid G] = \langle v_1, v_2 \rangle.$$

Hence

$$\mathbb{E}[B_{v_1} B_{v_2} \mid G] = \langle v_1, v_2 \rangle \sum_{i=1}^{d} (\theta_i^*)^2 \phi'(G_i)^2.$$

Because $\psi'(S)^2$ is G-measurable,

$$\mathbb{E}[\psi'(S)^2 B_{v_1} B_{v_2}] = \langle v_1, v_2 \rangle \sum_{i=1}^{d} (\theta_i^*)^2 \mathbb{E}\big[\psi'(S)^2 \phi'(G_i)^2\big].$$

Therefore we have the final $\beta\beta$-block in (66):

$$H_{L,\beta\beta}(y^*)[v_1, v_2] = 2\langle v_1, v_2 \rangle \sum_{i=1}^{d} (\theta_i^*)^2 \mathbb{E}\big[\psi'(S)^2 \phi'(G_i)^2\big].$$

Combining the $\theta\theta$-block, the vanishing mixed block, and the $\beta\beta$-block with $u_1 = u_2 = u$, $v_1 = v_2 = v$, gives the final quadratic formula. This completes the proof. □

Stein inputs for the teacher Hessian. We next state the two blockwise teacher-level comparisons used below. Their proofs are deferred to Subsection E.1.

Lemma C.12 (Teacher Hessian comparison)**.** *Let*

$$A_\psi := \mathbb{E}[\psi'(G)^2], \qquad B_\phi := \mathbb{E}[\phi'(G)^2] = \kappa'(1), \qquad \lambda_\theta := 2A_\psi, \qquad \lambda_\beta := 2A_\psi B_\phi.$$

Under the regularity assumptions of Lemma E.4 and Lemma E.6, there exists $C_* > 0$, independent of d, such that

$$\sup_{\|\xi\|=1} |\mathrm{Hess}_{\mathcal{M}} L(y^*)[\xi,\xi] - \mathrm{Hess}_{\mathcal{M}} L_G(y^*)[\xi,\xi]| \le C_* \|\theta^*\|_\infty. \tag{70}$$

Moreover,

$$\mathrm{Hess}_{\mathcal{M}} L_G(y^*)[(u,v),(u,v)] = \lambda_\theta \|u\|^2 + \lambda_\beta \|v\|^2. \tag{71}$$

Proof. Let

$$\xi = (u,v) \in T_{y^*}\mathcal{M}, \qquad \|\xi\|^2 = \|u\|^2 + \|v\|^2.$$

First, the Gaussian teacher Hessian formula follows from the explicit Gaussian loss:

$$L_G(\theta,\beta) = 2\sum_{k\ge 1} \alpha_k^2 \left(1 - (\kappa(\rho)\chi)^k\right).$$

At y^*, tangent geodesics satisfy

$$\chi''(0) = -\|u\|^2, \qquad \rho''(0) = -\|v\|^2, \qquad \chi'(0) = \rho'(0) = 0.$$

Therefore

$$\mathrm{Hess}_{\mathcal{M}} L_G(y^*)[(u,v),(u,v)] = \lambda_\theta \|u\|^2 + \lambda_\beta \|v\|^2,$$

with

$$\lambda_\theta = 2\sum_{k\ge 1} k\alpha_k^2 = 2A_\psi, \qquad \lambda_\beta = 2\kappa'(1)\sum_{k\ge 1} k\alpha_k^2 = 2A_\psi B_\phi.$$

For the true loss, decompose the Hessian into blocks. By Lemma C.11,

$$H_{L,\theta\beta}(y^*) = 0.$$

The same geodesic computation above gives

$$H_{L_G,\theta\beta}(y^*) = 0.$$

For the $\theta\theta$-block, Lemma E.4 at $(\theta,\beta) = (\theta^*,\beta^*)$ gives

$$|H_{L,\theta\theta}(y^*)[u,u] - H_{L_G,\theta\theta}(y^*)[u,u]| \le C_\theta \|\theta^*\|_\infty \|u\|^2.$$

For the $\beta\beta$-block, Lemma E.6 gives

$$|H_{L,\beta\beta}(y^*)[v,v] - H_{L_G,\beta\beta}(y^*)[v,v]| \le C_\beta \|\theta^*\|_\infty \|v\|^2.$$

Combining the three block estimates,

$$|H_L(y^*)[\xi,\xi] - H_{L_G}(y^*)[\xi,\xi]| \le (C_\theta + C_\beta)\|\theta^*\|_\infty (\|u\|^2 + \|v\|^2).$$

Taking $\|\xi\| = 1$ and setting $C_* := C_\theta + C_\beta$ proves the claim. □

Lemma C.13 (Uniform Hessian continuity near the teacher)**.** *Assume $\psi \in C_b^3(\mathbb{R})$, and assume that $\phi \in C^3(\mathbb{R})$ satisfies*

$$\max_{0\le j\le 3} \mathbb{E}\left[(1+|G|)^{12} |\phi^{(j)}(G)|^{12}\right] < \infty, \qquad G \sim N(0,1).$$

Assume also

$$\mathbb{E}[\phi(G)] = 0, \qquad \mathbb{E}[\phi(G)^2] = 1.$$

Then there exist constants $r_0 > 0$ and $C_3 > 0$, independent of d, such that for every $y \in \mathcal{U}_{r_0}$,

$$\|P_{y\to y^*} \mathrm{Hess}_{\mathcal{M}} L(y) P_{y^*\to y} - \mathrm{Hess}_{\mathcal{M}} L(y^*)\|_{\mathrm{op}} \le C_3 \,\mathrm{dist}_{\mathcal{M}}(y, y^*). \tag{72}$$

Here $P_{y\to y^*}$ denotes parallel transport along the minimizing product geodesic from y to y^*.

Proof. It is enough to prove that the covariant derivative of the Hessian is uniformly bounded on a fixed product cap. Indeed, let $\gamma : [0, \ell] \to \mathcal{M}$ be the minimizing product geodesic from y^* to y, parametrized by arclength, so that

$$\ell = \operatorname{dist}_{\mathcal{M}}(y, y^*).$$

Let $\xi_0 \in T_{y^*}\mathcal{M}$ be a unit vector, and let ξ_t be its parallel transport along γ. Then

$$\frac{d}{dt} \operatorname{Hess}_{\mathcal{M}} L(\gamma(t))[\xi_t, \xi_t] = (\nabla \operatorname{Hess}_{\mathcal{M}} L)(\gamma(t))[\dot{\gamma}(t), \xi_t, \xi_t].$$

Thus, if

$$\sup_{\gamma(t) \in \mathcal{U}_{r_0}} \sup_{\|\zeta\| = \|\xi\| = 1} |(\nabla \operatorname{Hess}_{\mathcal{M}} L)(\gamma(t))[\zeta, \xi, \xi]| \le C_3,$$

then integration along γ gives the desired transported-Hessian Lipschitz bound.

We prove this uniform bound. Write

$$s_y(x) := \sum_{i=1}^{d} \theta_i \phi(x^{(i)} \cdot \beta), \qquad s_*(x) := \sum_{i=1}^{d} \theta_i^* \phi(x^{(i)} \cdot \beta^*),$$

so that

$$L(y) = \mathbb{E}\left[\big(\psi(s_y(x)) - \psi(s_*(x))\big)^2 \right].$$

Let $y_t = (\theta_t, \beta_t)$ be any unit-speed product geodesic contained in $\mathcal{U}_{r_0}$, and set

$$s_t := s_{y_t}(x).$$

We first record the basic moment estimate

$$\sup_t \left(\|\dot{s}_t\|_{L^6} + \|\ddot{s}_t\|_{L^3} + \|s_t^{(3)}\|_{L^1} \right) \le C, \tag{73}$$

where C is independent of d and of the chosen unit-speed product geodesic.

To prove (73), observe that

$$s_t = \sum_{i=1}^{d} \theta_{t,i} \phi(U_{t,i}), \qquad U_{t,i} := x^{(i)} \cdot \beta_t.$$

Along a unit-speed spherical geodesic,

$$\|\dot{\theta}_t\| + \|\ddot{\theta}_t\| + \|\theta_t^{(3)}\| \le C, \qquad \|\dot{\beta}_t\| + \|\ddot{\beta}_t\| + \|\beta_t^{(3)}\| \le C.$$

Moreover,

$$\dot{U}_{t,i} = x^{(i)} \cdot \dot{\beta}_t, \qquad \ddot{U}_{t,i} = x^{(i)} \cdot \ddot{\beta}_t, \qquad U_{t,i}^{(3)} = x^{(i)} \cdot \beta_t^{(3)}.$$

The variables coming from distinct blocks i are independent, and each block has moments bounded uniformly in t. Differentiating

$$\theta_{t,i} \phi(U_{t,i})$$

up to order three gives a finite sum of terms whose coefficient vectors have uniformly bounded ℓ_2-norm and whose block variables have uniformly bounded moments by the assumptions on ϕ.

The only point that needs care is centering. For each i,

$$\mathbb{E}[\phi(U_{t,i})] = \mathbb{E}[\phi(G)] = 0$$

for every t, because $U_{t,i} \sim N(0,1)$. Differentiating this identity in t shows that the differentiated block variables appearing in $\dot{s}_t, \ddot{s}_t, s_t^{(3)}$ are centered. Therefore Rosenthal's inequality applies to each differentiated sum and gives the bounds in (73), uniformly in d.

Now define

$$R_t := \psi(s_t) - \psi(s_*).$$

Since ψ and its derivatives up to order three are bounded,

$$\dot{R}_t = \psi'(s_t)\dot{s}_t,$$

$$\ddot{R}_t = \psi''(s_t)\dot{s}_t^2 + \psi'(s_t)\ddot{s}_t,$$

and

$$R_t^{(3)} = \psi^{(3)}(s_t)\dot{s}_t^3 + 3\psi''(s_t)\dot{s}_t\ddot{s}_t + \psi'(s_t)s_t^{(3)}.$$

Since $L(y_t) = \mathbb{E}[R_t^2]$,

$$\frac{d^3}{dt^3}L(y_t) = 2\mathbb{E}\left[3\dot{R}_t\ddot{R}_t + R_t R_t^{(3)}\right].$$

Using the boundedness of $\psi, \psi', \psi'', \psi^{(3)}$, together with (73) and Hölder's inequality, we obtain

$$\sup_t \left|\frac{d^3}{dt^3}L(y_t)\right| \le C.$$

The same argument, applied to geodesic variations and then polarized, gives

$$\sup_{y\in\mathcal{U}_{r_0}} \sup_{\|\zeta\|=\|\xi\|=1} |(\nabla \operatorname{Hess}_{\mathcal{M}} L)(y)[\zeta,\xi,\xi]| \le C.$$

Equivalently, the transported Hessian is C-Lipschitz along every minimizing product geodesic in $\mathcal{U}_{r_0}$. Integrating this bound along the geodesic from y^* to y gives

$$\|P_{y\to y^*} \operatorname{Hess}_{\mathcal{M}} L(y) P_{y^*\to y} - \operatorname{Hess}_{\mathcal{M}} L(y^*)\|_{\mathrm{op}} \le C \operatorname{dist}_{\mathcal{M}}(y, y^*).$$

This proves the lemma, with $C_3 = C$. □

Lemma C.14 (True local product strong convexity)**.** *Assume the teacher Hessian comparison of Lemma C.12 and the uniform Hessian continuity of Lemma C.13. On the random-spherical delocalization event of Lemma C.1, so that*

$$\|\theta^*\|_\infty \le C_K\sqrt{\frac{\log d}{d}},$$

there exist constants $r_{\mathrm{loc}} > 0$ and $\mu_{\mathrm{loc}} > 0$, independent of d, such that for all sufficiently large d and every $y \in \mathcal{U}_{r_{\mathrm{loc}}}$,

$$\operatorname{Hess}_{\mathcal{M}} L(y) \succeq \mu_{\mathrm{loc}} I. \tag{74}$$

Proof. Let

$$\lambda_0 := \min\{\lambda_\theta, \lambda_\beta\} = 2A_\psi \min\{1, B_\phi\}.$$

By the information-exponent-one assumptions, $A_\psi > 0$ and $B_\phi > 0$, so $\lambda_0 > 0$.

By Lemma C.12, for every $\xi \in T_{y^*}\mathcal{M}$,

$$\operatorname{Hess}_{\mathcal{M}} L(y^*)[\xi,\xi] \ge \operatorname{Hess}_{\mathcal{M}} L_G(y^*)[\xi,\xi] - C_*\|\theta^*\|_\infty\|\xi\|^2.$$

Since by Lemma C.10

$$\operatorname{Hess}_{\mathcal{M}} L_G(y^*)[\xi,\xi] \ge \lambda_0\|\xi\|^2,$$

we obtain

$$\operatorname{Hess}_{\mathcal{M}} L(y^*)[\xi,\xi] \ge (\lambda_0 - C_*\|\theta^*\|_\infty)\,\|\xi\|^2.$$

By Lemma C.1,

$$\|\theta^*\|_\infty \le C_K\sqrt{\frac{\log d}{d}}.$$

Hence, for all sufficiently large d,

$$C_*\|\theta^*\|_\infty \le C_* C_K\sqrt{\frac{\log d}{d}} \le \frac{\lambda_0}{4}.$$

Hence

$$\operatorname{Hess}_{\mathcal{M}} L(y^*) \succeq \frac{3\lambda_0}{4} I.$$

Now choose

$$r_{\mathrm{loc}} \le r_0 \qquad \text{and} \qquad C_3 r_{\mathrm{loc}} \le \frac{\lambda_0}{4}.$$

Let $y \in \mathcal{U}_{r_{\mathrm{loc}}}$, and let $\xi \in T_y\mathcal{M}$ be a unit tangent vector. Parallel transport ξ to $T_{y^*}\mathcal{M}$, writing

$$\widetilde{\xi} := P_{y \to y^*}\xi.$$

Then $\|\widetilde{\xi}\| = 1$. By Lemma C.13,

$$\mathrm{Hess}_{\mathcal{M}} L(y)[\xi, \xi] \geq \mathrm{Hess}_{\mathcal{M}} L(y^*)[\widetilde{\xi}, \widetilde{\xi}] - C_3 \,\mathrm{dist}_{\mathcal{M}}(y, y^*).$$

Since $y \in \mathcal{U}_{r_{\mathrm{loc}}}$,

$$\mathrm{dist}_{\mathcal{M}}(y, y^*) \leq r_{\mathrm{loc}}.$$

Therefore

$$\mathrm{Hess}_{\mathcal{M}} L(y)[\xi, \xi] \geq \frac{3\lambda_0}{4} - C_3 r_{\mathrm{loc}} \geq \frac{\lambda_0}{2}.$$

Thus the lemma holds with

$$\mu_{\mathrm{loc}} := \frac{\lambda_0}{2}.$$

□

Strong Recovery Dynamics.

Lemma C.15 (Local contraction under product strong convexity)**.** *Let $\mathcal{M} = \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ equipped with the product Riemannian metric, and let $y^* = (\theta^*, \beta^*) \in \mathcal{M}$. Fix $r \in (0, \pi/2)$ and suppose $L : \mathcal{M} \to \mathbb{R}$ is C^2 and satisfies*

$$\mathrm{Hess}_{\mathcal{M}} L(y) \succeq \mu I \qquad \text{for all } y \in \mathcal{U}_r := \{y \in \mathcal{M} : \mathrm{dist}_{\mathcal{M}}(y, y^*) \leq r\},$$

for some $\mu > 0$ (Lemma C.14, with $r = r_{\mathrm{loc}}$ and $\mu = \mu_{\mathrm{loc}}$), and that y^* is a critical point of L, i.e., $\nabla_{\mathcal{M}} L(y^*) = 0$.

Let y_t be the population gradient flow

$$\dot{y}_t = -\nabla_{\mathcal{M}} L(y_t), \qquad y_{t_0} \in \mathcal{M}, \qquad \mathrm{dist}_{\mathcal{M}}(y_{t_0}, y^*) \leq r/2.$$

Then $y_t \in \mathcal{U}_r$ for all $t \geq t_0$, and

$$\mathrm{dist}_{\mathcal{M}}(y_t, y^*)^2 \leq e^{-2\mu(t-t_0)} \,\mathrm{dist}_{\mathcal{M}}(y_{t_0}, y^*)^2. \tag{75}$$

Proof. Write

$$D_t := \mathrm{dist}_{\mathcal{M}}(y_t, y^*), \qquad E_t := \frac{1}{2} D_t^2.$$

Since $r < \pi/2$, every point $y = (\theta, \beta) \in \mathcal{U}_r$ satisfies

$$\mathrm{dist}_{\mathbb{S}}(\theta, \theta^*) < \pi/2, \qquad \mathrm{dist}_{\mathbb{S}}(\beta, \beta^*) < \pi/2.$$

Thus neither component is antipodal to its target. Hence the minimizing geodesic from y to y^* is unique and is obtained by taking the componentwise minimizing geodesics on the two spheres. In particular, $\mathrm{Log}_y(y^*)$ is well-defined, and

$$\|\,\mathrm{Log}_y(y^*)\| = \mathrm{dist}_{\mathcal{M}}(y, y^*).$$

Moreover, if $\gamma : [0, 1] \to \mathcal{M}$ is this minimizing product geodesic from y to y^*, then

$$\mathrm{dist}_{\mathcal{M}}(\gamma(s), y^*) = (1 - s)\,\mathrm{dist}_{\mathcal{M}}(y, y^*),$$

so $\gamma(s) \in \mathcal{U}_r$ whenever $y \in \mathcal{U}_r$.

Define the exit time

$$\tau := \inf\{t \geq t_0 : \; D_t > r\}.$$

For $t \in [t_0, \tau)$, we have $y_t \in \mathcal{U}_r$. We will prove

$$\frac{d}{dt} E_t \leq -2\mu E_t \qquad \text{for all } t \in [t_0, \tau).$$

Once this is shown, Gronwall gives

$$E_t \leq e^{-2\mu(t-t_0)} E_{t_0} \qquad \text{for all } t \in [t_0, \tau).$$

Since $D_{t_0} \le r/2$, this implies $D_t \le r/2$ on $[t_0, \tau)$. If $\tau < \infty$, then by continuity $D_\tau \le r/2 < r$, contradicting the definition of τ. Hence $\tau = \infty$, and the trajectory stays in $\mathcal{U}_r$ for all $t \ge t_0$.

It remains to prove the differential inequality. Fix $t < \tau$, and let $\gamma : [0,1] \to \mathcal{M}$ be the minimizing geodesic from y_t to y^*, parametrized with constant speed:

$$\gamma(0) = y_t, \qquad \gamma(1) = y^*, \qquad \|\gamma'(s)\| \equiv D_t.$$

Then

$$\gamma'(0) = \mathrm{Log}_{y_t}(y^*).$$

Set

$$\varphi(s) := L(\gamma(s)).$$

Since γ is a geodesic,

$$\varphi''(s) = \mathrm{Hess}_{\mathcal{M}} L(\gamma(s))[\gamma'(s), \gamma'(s)].$$

The geodesic remains in $\mathcal{U}_r$, so the Hessian lower bound gives

$$\varphi''(s) \ge \mu \|\gamma'(s)\|^2 = \mu D_t^2.$$

Integrating from 0 to 1,

$$\varphi'(1) - \varphi'(0) \ge \mu D_t^2.$$

Because y^* is critical,

$$\varphi'(1) = \langle \nabla_{\mathcal{M}} L(y^*), \gamma'(1) \rangle = 0.$$

Therefore

$$\varphi'(0) \le -\mu D_t^2.$$

But

$$\varphi'(0) = \left\langle \nabla_{\mathcal{M}} L(y_t), \mathrm{Log}_{y_t}(y^*) \right\rangle.$$

Thus

$$\left\langle \nabla_{\mathcal{M}} L(y_t), \mathrm{Log}_{y_t}(y^*) \right\rangle \le -\mu D_t^2.$$

Finally, the first-variation formula for the squared distance gives

$$\frac{d}{dt} E_t = -\left\langle \dot{y}_t, \mathrm{Log}_{y_t}(y^*) \right\rangle.$$

Since $\dot{y}_t = -\nabla_{\mathcal{M}} L(y_t)$,

$$\frac{d}{dt} E_t = \left\langle \nabla_{\mathcal{M}} L(y_t), \mathrm{Log}_{y_t}(y^*) \right\rangle \le -\mu D_t^2 = -2\mu E_t.$$

This proves the desired differential inequality, and hence the contraction estimate. □

Theorem C.16 (Continuous-time strong recovery)**.** *Assume the weak-recovery conclusion of Lemma C.8, together with the assumptions of Lemmas C.3 and C.4. Assume also the regularity and incoherence assumptions under which Lemmas C.12, C.13, C.14, and C.15 hold.*

Let $y_t := (\theta_t, \beta_t) \in \mathcal{M} := \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, $y^* := (\theta^*, \beta^*)$, and write $R_t := \mathrm{dist}_{\mathcal{M}}(y_t, y^*)$. Then there exist constants $r_{\mathrm{loc}} > 0$, $\mu_{\mathrm{loc}} > 0$, $C_{\mathrm{loc}} < \infty$, independent of d, such that for all sufficiently large d, on the favorable initialization event of (A4), there exists a time $t_{\mathrm{loc}} \le C_{\mathrm{loc}} \log d$ for which

$$R_{t_{\mathrm{loc}}}^2 \le \frac{r_{\mathrm{loc}}^2}{8}. \tag{76}$$

Moreover, for every $t \ge t_{\mathrm{loc}}$,

$$R_t^2 \le e^{-2\mu_{\mathrm{loc}}(t - t_{\mathrm{loc}})} R_{t_{\mathrm{loc}}}^2.$$

Consequently,

$$(1 - \chi_t) + (1 - \rho_t) \le \frac{r_{\mathrm{loc}}^2}{16} e^{-2\mu_{\mathrm{loc}}(t - t_{\mathrm{loc}})}.$$

In particular, there exists $\varepsilon_0 > 0$, independent of d, such that for every $\varepsilon \in (0, \varepsilon_0)$,

$$\min\{\chi_t, \rho_t\} \ge 1 - \varepsilon \tag{77}$$

whenever

$$t \ge C \log d + C \log(1/\varepsilon),$$

where $C < \infty$ is independent of d and ε.

Proof. By Lemma C.14, there exist constants $r_{\mathrm{loc}} > 0$ and $\mu_{\mathrm{loc}} > 0$, independent of d, such that for all sufficiently large d,

$$\mathrm{Hess}_{\mathcal{M}} L(y) \succeq \mu_{\mathrm{loc}} I \qquad \text{for every } y \in \mathcal{U}_{r_{\mathrm{loc}}} := \{y \in \mathcal{M} : \mathrm{dist}_{\mathcal{M}}(y, y^*) \le r_{\mathrm{loc}}\}.$$

After shrinking r_{loc}, if necessary, we may assume

$$r_{\mathrm{loc}} < \frac{\pi}{2}.$$

The same Hessian lower bound remains valid on the smaller neighborhood.

Apply Lemma C.9 with this value of r_{loc}. On the favorable event of Lemma C.8, it gives a time t_{loc} satisfying

$$t_{\mathrm{loc}} \le t_{\mathrm{esc}} + C_{\mathrm{in}},$$

where $C_{\mathrm{in}} < \infty$ is independent of d, and

$$\rho_{t_{\mathrm{loc}}} \ge \cos(r_{\mathrm{loc}}/4), \qquad \chi_{t_{\mathrm{loc}}} \ge \cos(r_{\mathrm{loc}}/4).$$

Equivalently

$$R^2_{t_{\mathrm{loc}}} = \mathrm{dist}_{\mathcal{M}}(y_{t_{\mathrm{loc}}}, y^*)^2 \le \frac{r^2_{\mathrm{loc}}}{8},$$

which proves (76). Since Lemma C.8 gives

$$t_{\mathrm{esc}} \le C_{\mathrm{esc}} \log d,$$

we have, for $d \ge e$,

$$t_{\mathrm{loc}} \le C_{\mathrm{esc}} \log d + C_{\mathrm{in}} \le (C_{\mathrm{esc}} + C_{\mathrm{in}}) \log d.$$

Thus the claimed bound holds with

$$C_{\mathrm{loc}} := C_{\mathrm{esc}} + C_{\mathrm{in}}.$$

Next, observe that y^* is a critical point of L. Indeed, at the planted pair the residual vanishes pointwise, so

$$L(y^*) = 0.$$

Since $L \ge 0$, y^* is a global minimizer of L, and therefore

$$\nabla_{\mathcal{M}} L(y^*) = 0.$$

The entry estimate gives

$$R_{t_{\mathrm{loc}}} \le \frac{r_{\mathrm{loc}}}{2\sqrt{2}} < \frac{r_{\mathrm{loc}}}{2}.$$

Hence we may apply Lemma C.15 with

$$r = r_{\mathrm{loc}}, \qquad \mu = \mu_{\mathrm{loc}}, \qquad t_0 = t_{\mathrm{loc}}.$$

It follows that the trajectory remains in $\mathcal{U}_{r_{\mathrm{loc}}}$ for all $t \ge t_{\mathrm{loc}}$, and that

$$R^2_t \le e^{-2\mu_{\mathrm{loc}}(t - t_{\mathrm{loc}})} R^2_{t_{\mathrm{loc}}} \qquad \text{for every } t \ge t_{\mathrm{loc}}.$$

We now convert the product-manifold distance decay into overlap decay. Since $\chi_t = \langle \theta_t, \theta^* \rangle = \cos\left(\mathrm{dist}_{\mathbb{S}^{d-1}}(\theta_t, \theta^*)\right)$, and $\rho_t = \langle \beta_t, \beta^* \rangle = \cos\left(\mathrm{dist}_{\mathbb{S}^{d-1}}(\beta_t, \beta^*)\right)$, the elementary bound $1 - \cos x \le \frac{x^2}{2} \quad (x \ge 0)$ gives

$$1 - \chi_t \le \frac{1}{2} \mathrm{dist}_{\mathbb{S}^{d-1}}(\theta_t, \theta^*)^2,$$

and

$$1 - \rho_t \le \frac{1}{2} \mathrm{dist}_{\mathbb{S}^{d-1}}(\beta_t, \beta^*)^2.$$

Summing these two estimates and using the product metric, we have

$$(1 - \chi_t) + (1 - \rho_t) \le \frac{1}{2} \mathrm{dist}_{\mathcal{M}}(y_t, y^*)^2 = \frac{1}{2} R^2_t.$$

Therefore,

$$(1-\chi_t)+(1-\rho_t) \le \frac{1}{2} e^{-2\mu_{\mathrm{loc}}(t-t_{\mathrm{loc}})} R_{t_{\mathrm{loc}}}^2.$$

Using 76 that we proved, we obtain

$$(1-\chi_t)+(1-\rho_t) \le \frac{r_{\mathrm{loc}}^2}{16} e^{-2\mu_{\mathrm{loc}}(t-t_{\mathrm{loc}})}.$$

Finally, set

$$\varepsilon_0 := \frac{r_{\mathrm{loc}}^2}{16}.$$

Let $\varepsilon \in (0, \varepsilon_0)$. If

$$\frac{r_{\mathrm{loc}}^2}{16} e^{-2\mu_{\mathrm{loc}}(t-t_{\mathrm{loc}})} \le \varepsilon,$$

then

$$(1-\chi_t)+(1-\rho_t) \le \varepsilon.$$

Since both summands are nonnegative, this implies

$$1-\chi_t \le \varepsilon, \qquad 1-\rho_t \le \varepsilon,$$

and hence we arrive at (77):

$$\min\{\chi_t, \rho_t\} \ge 1-\varepsilon.$$

The displayed exponential condition is equivalent to

$$t - t_{\mathrm{loc}} \ge \frac{1}{2\mu_{\mathrm{loc}}} \log \frac{r_{\mathrm{loc}}^2}{16\varepsilon}.$$

Using $t_{\mathrm{loc}} \le C_{\mathrm{loc}} \log d$, we conclude that

$$\min\{\chi_t, \rho_t\} \ge 1-\varepsilon$$

whenever

$$t \ge C_{\mathrm{loc}} \log d + \frac{1}{2\mu_{\mathrm{loc}}} \log \frac{r_{\mathrm{loc}}^2}{16\varepsilon}.$$

Since r_{loc} and μ_{loc} are fixed constants independent of d and ε, this is of the form

$$t \ge C \log d + C \log(1/\varepsilon),$$

for a constant $C < \infty$ independent of d and ε. This proves the theorem. □

D Proof of Discrete-Time Online SGD Dynamics

We now move to analyze the learning dynamics under SGD discretization.

D.1 Weak Recovery Setups

Assumption D.1 (SGD regime and initialization for weak recovery)**.** *Fix $0 < \varepsilon \le \varepsilon_*$. Consider online* blockwise *spherical SGD on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ follows* (3)*. The horizon and step-size follows*

$$T\delta \le C_T \log d, \qquad \delta \le \frac{c_0 \varepsilon}{d \log d}, \qquad 0 < \varepsilon \le \varepsilon_*.$$

for an absolute constant $c_0 > 0$, so we have the weak recovery horizon $T := \lceil C_T \delta^{-1} \log d \rceil \in \mathbb{N}$. Initialization condition (A4) also follows.

We start the discretizations analysis by deriving the one-step evolution of the overlaps under SGD:

Lemma D.2 (One-step evolution of the overlaps under SGD)**.** *Under Assumption D.1, on the event*

$$E_t^{(T)} := \{\delta\|g_t^{(\beta)}\| \le 1/2,\ \delta\|g_t^{(\theta)}\| \le 1/2\}, \tag{78}$$

there exist $\mathcal{F}_t$-measurable predictable drifts $F_t^{(\rho)}, F_t^{(\chi)}$, martingale increments $M_t^{(\rho)}, M_t^{(\chi)}$, and remainders $R_t^{(\rho)}, R_t^{(\chi)}$ such that for all t,

$$\rho_{t+1} = \rho_t + \delta F_t^{(\rho)} + \delta M_t^{(\rho)} + R_t^{(\rho)} \tag{79}$$

$$\chi_{t+1} = \chi_t + \delta F_t^{(\chi)} + \delta M_t^{(\chi)} + R_t^{(\chi)} \tag{80}$$

where $\mathbb{E}[M_t^{(\cdot)} \mid \mathcal{F}_t] = 0$ for each component, $R_t^{(\rho)} = O(\delta^3\|g_t^{(\beta)}\|^3)$, and $R_t^{(\chi)} = O(\delta^3\|g_t^{(\theta)}\|^3)$.

Proof. Let $g_t^{(\beta)} := \nabla_\beta^S l(\beta_t, \theta_t; x_t)$ and $g_t^{(\theta)} := \nabla_\theta^S l(\beta_t, \theta_t; x_t)$ be the spherical gradients taken with respect β and θ respectively, which implies $g_t^{(\beta)} \perp \beta_t$ and $g_t^{(\theta)} \perp \theta_t$.From (3) and under $\mathcal{E}_t$, we have the normalized update as

$$\beta_{t+1} = \frac{\beta_t - \delta g_t^{(\beta)}}{\|\beta_t - \delta g_t^{(\beta)}\|} = \beta_t - \delta g_t^{(\beta)} - \frac{\delta^2}{2}\|g_t^{(\beta)}\|^2\beta_t + O(\delta^3\|g_t^{(\beta)}\|^3) \tag{81}$$

$$\theta_{t+1} = \frac{\theta_t - \delta g_t^{(\theta)}}{\|\theta_t - \delta g_t^{(\theta)}\|} = \theta_t - \delta g_t^{(\theta)} - \frac{\delta^2}{2}\|g_t^{(\theta)}\|^2\theta_t + O(\delta^3\|g_t^{(\theta)}\|^3) \tag{82}$$

Hence taking overlap with the teacher direction β^* and θ^*, respectively, we have

$$\langle\beta_{t+1}, \beta^*\rangle = \langle\beta_t, \beta^*\rangle - \delta\langle g_t^{(\beta)}, \beta^*\rangle - \frac{\delta^2}{2}\|g_t^{(\beta)}\|^2\langle\beta_t, \beta^*\rangle + O(\delta^3\|g_t^{(\beta)}\|^3) \tag{83}$$

$$\langle\theta_{t+1}, \theta^*\rangle = \langle\theta_t, \theta^*\rangle - \delta\langle g_t^{(\theta)}, \theta^*\rangle - \frac{\delta^2}{2}\|g_t^{(\theta)}\|^2\langle\theta_t, \theta^*\rangle + O(\delta^3\|g_t^{(\theta)}\|^3), \tag{84}$$

where $\rho_t := \langle\beta_t, \beta^*\rangle$ and $\chi_t := \langle\theta_t, \theta^*\rangle$ Now, we decompose the middle two terms involving spherical gradient as the population drift plus the martingale increment, namely,

$$\langle g_t^{(\beta)}, \beta^*\rangle + \frac{\delta}{2}\|g_t^{(\beta)}\|^2\rho_t = \underbrace{\mathbb{E}\left[\langle g_t^{(\beta)}, \beta^*\rangle + \frac{\delta}{2}\|g_t^{(\beta)}\|^2\rho_t|\mathcal{F}_t\right]}_{-F_t^{(\rho)}} + \underbrace{\left(\langle g_t^{(\beta)}, \beta^*\rangle + \frac{\delta}{2}\|g_t^{(\beta)}\|^2\rho_t - \mathbb{E}[\langle g_t^{(\beta)}, \beta^*\rangle + \frac{\delta}{2}\|g_t^{(\beta)}\|^2\rho_t|\mathcal{F}_t]\right)}_{-M_t^{(\rho)}} \tag{85}$$

$$\langle g_t^{(\theta)}, \theta^*\rangle + \frac{\delta}{2}\|g_t^{(\theta)}\|^2\chi_t = \underbrace{\mathbb{E}\left[\langle g_t^{(\theta)}, \theta^*\rangle + \frac{\delta}{2}\|g_t^{(\theta)}\|^2\chi_t|\mathcal{F}_t\right]}_{-F_t^{(\chi)}} + \underbrace{\left(\langle g_t^{(\theta)}, \theta^*\rangle + \frac{\delta}{2}\|g_t^{(\theta)}\|^2\chi_t - \mathbb{E}[\langle g_t^{(\theta)}, \theta^*\rangle + \frac{\delta}{2}\|g_t^{(\theta)}\|^2\chi_t|\mathcal{F}_t]\right)}_{-M_t^{(\chi)}}. \tag{86}$$

Therefore, we can write (83) and (84) respectively as

$$\rho_{t+1} = \rho_t + \delta F_t^{(\rho)} + \delta M_t^{(\rho)} + O(\delta^3\|g_t^{(\beta)}\|^3) \tag{87}$$

$$\chi_{t+1} = \chi_t + \delta F_t^{(\chi)} + \delta M_t^{(\chi)} + O(\delta^3\|g_t^{(\theta)}\|^3) \tag{88}$$

Defining $R_t^{(\rho)} := O(\delta^3\|g_t^{(\beta)}\|^3)$ and $R_t^{(\chi)} := O(\delta^3\|g_t^{(\theta)}\|^3)$ gives the claimed expression. □

Next we provide a class of concentration inequalities to better control the iteration of (ρ_t, χ_t) that has been used in studying the aggregate SGD dynamics, on top of (79) and (80). In particular, Lemma D.5 controls the aggregate martingale term, and Lemma D.7 controls the aggregate discretization remainder term. Besides, As the one-step expansion in Lemma D.2 relies on the event $\mathcal{E}_t$, we provide Lemma D.3 to show that $\mathcal{E}_t$ is true with high probability. All of these inequalities are obtained via using the technical bounds specified in Lemma E.7 in E.2.

D.1.1 Taylor Expansion Error Control

Lemma D.3 (Uniform Taylor event on the Weak-Recovery Horizon)**.** *Assume the moment bounds of Lemma E.7 and following Assumption D.1 suppose*

$$T\delta \le C_T \log d, \qquad \delta \le \frac{c_0 \varepsilon}{d \log d}, \qquad 0 < \varepsilon \le \varepsilon_*.$$

Then there is a constant K_{Tay}, depending only on model constants, C_T, c_0, ε_*, such that for all sufficiently large d, the Taylor Event $E_T^{(T)}$ satisfies:

$$\mathbb{P}((E_T^{(T)})^c) \le 32KT\delta^4 d^2 \le K_{\mathrm{Tay}} \frac{\varepsilon^3}{d \log^2 d} \le K_{\mathrm{Tay}} \varepsilon. \tag{89}$$

Moreover, on $E_T^{(T)}$, the one-step expansions (79) *and* (80) *in Lemma D.2 are valid simultaneously for every $t < T$, with remainders*

$$|R_t^{(\rho)}| \le C\delta^3 \|g_t^{(\beta)}\|^3, \qquad |R_t^{(\chi)}| \le C\delta^3 \|g_t^{(\theta)}\|^3.$$

Proof. By Lemma E.7, conditionally on $\mathcal{F}_t$, we have

$$\mathbb{E}\left[\|g_t^{(\beta)}\|^4 \mid \mathcal{F}_t\right] \le Kd^2, \qquad \mathbb{E}\left[\|g_t^{(\theta)}\|^4 \mid \mathcal{F}_t\right] \le Kd^2.$$

Hence, by Markov's inequality, $\mathbb{P}\left(\delta\|g_t^{(\beta)}\| > \frac{1}{2} \mid \mathcal{F}_t\right) \le 16\delta^4 \mathbb{E}\left[\|g_t^{(\beta)}\|^4 \mid \mathcal{F}_t\right] \le 16K\delta^4 d^2$. The same bound holds for the θ-block. Therefore

$$\mathbb{P}({E_t^{(T)}}^c) \le 32K\delta^4 d^2.$$

Union bounding over $t = 0, \dots, T-1$,

$$\mathbb{P}({E_T^{(T)}}^c) \le 32KT\delta^4 d^2.$$

Since $T\delta \le C_T \log d$,

$$T\delta^4 d^2 = (T\delta)\delta^3 d^2 \le C_T \delta^3 d^2 \log d.$$

Using $\delta \le \frac{c_0 \varepsilon}{d \log d}$, we get

$$\delta^3 d^2 \log d \le \frac{c_0^3 \varepsilon^3}{d \log^2 d}.$$

Thus

$$\mathbb{P}({E_T^{(T)}}^c) \le K_{\mathrm{Tay}} \frac{\varepsilon^3}{d \log^2 d} \le K_{\mathrm{Tay}} \varepsilon$$

for all sufficiently large d. This proves the lemma. □

D.1.2 Martingale and Discretization Error Control for Z_t

The Martingale Error for Z_t depends on this intermediate Corollary, derived from Lemma E.7.

Corollary D.4 (Negligible remainder for drift)**.** *With the martingale increments $M_t^{(\rho)}$ and $M_t^{(\chi)}$ we defined in Lemma D.2 and adopting the technical gradient bounds in Lemma E.7, we have*

$$\mathbb{E}\Big[(M_t^{(\rho)})^2 \mid \mathcal{F}_t\Big] \le K_M^{\rho}, \ \mathbb{E}\Big[(M_t^{(\chi)})^2 \mid \mathcal{F}_t\Big] \le K_M^{\chi}, \tag{90}$$

where K_M^{ρ}, K_M^{χ} absolute for large d under step size $\delta \le c_0 \varepsilon / d \log d$.

Proof. Let $A_t^{(\rho)} = \langle g_t^{(\beta)}, \beta^* \rangle + \frac{\delta}{2} \|g_t^{(\beta)}\|^2 \rho_t$, then the martingale term writes as $M_t^{(\rho)} = -(A_t^{(\rho)} - \mathbb{E}[A_t^{(\rho)} \mid \mathcal{F}_t])$, so the conditional second moment of $M_t^{(\rho)}$ can be written as

$$\mathbb{E}[(M_t^{(\rho)})^2 \mid \mathcal{F}_t] = \mathbb{E}\big[\big(A_t^{(\rho)} - \mathbb{E}[A_t^{(\rho)} \mid \mathcal{F}_t]\big)^2 \mid \mathcal{F}_t\big] \tag{91}$$

Since we have:

$$A_t^{(\rho)} - \mathbb{E}[A_t^{(\rho)} \mid \mathcal{F}_t] = \big(\langle g_t^{(\beta)}, \beta^* \rangle + \frac{\delta}{2}\|g_t^{(\beta)}\|^2\rho_t\big) - \mathbb{E}\big[\langle g_t^{(\beta)}, \beta^* \rangle + \frac{\delta}{2}\|g_t^{(\beta)}\|^2\rho_t \mid \mathcal{F}_t\big]$$
$$= \big(\langle g_t^{(\beta)}, \beta^* \rangle - \mathbb{E}[\langle g_t^{(\beta)}, \beta^* \rangle \mid \mathcal{F}_t]\big) + \frac{\delta}{2}\rho_t\big(\|g_t^{(\beta)}\|^2 - \mathbb{E}[\|g_t^{(\beta)}\|^2 \mid \mathcal{F}_t]\big) \text{ by regrouping}$$
$$= \langle \Delta g_t^{(\beta)}, \beta^* \rangle + \frac{\delta}{2}\rho_t \Delta q_t^{(\beta)},$$

where we defined $\Delta g_t^{(\beta)} := g_t^{(\beta)} - \mathbb{E}[g_t^{(\beta)} \mid \mathcal{F}_t]$, and $\Delta q_t^{(\beta)} := \|g_t^{(\beta)}\|^2 - \mathbb{E}[\|g_t^{(\beta)}\|^2 \mid \mathcal{F}_t]$, (91) can be written as:

$$\mathbb{E}[(M_t^{(\rho)})^2 \mid \mathcal{F}_t] = \mathbb{E}[\big(\langle \Delta g_t^{(\beta)}, \beta^* \rangle + \frac{\delta}{2}\rho_t \Delta q_t^{(\beta)}\big)^2 \mid \mathcal{F}_t] \tag{92}$$
$$= 2\mathbb{E}[\langle \Delta g_t^{(\beta)}, \beta^* \rangle^2 \mid \mathcal{F}_t] + \frac{\delta^2}{2}\rho_t^2 \mathbb{E}[(\Delta q_t^{(\beta)})^2], \text{ given } (a+b)^2 \le 2a^2 + 2b^2 \tag{93}$$

By Lemma E.7, we can bound the first term of (93) by K. For the second term of (93), note that $(\Delta q)^2 = \big(\|g_t^{(\beta)}\|^2 - \mathbb{E}[\|g_t^{(\beta)}\|^2 \mid \mathcal{F}_t]\big)^2 \le \|g_t^{(\beta)}\|^4$, we can use the forth moment bound in Lemma E.7 to obtain the Kd^2 bound, so

$$\mathbb{E}[(M_t^{(\rho)})^2 \mid \mathcal{F}_t] \le 2K + \frac{(\delta\rho_t)^2}{2} K d^2 \tag{94}$$
$$\le 2K + \frac{\delta^2}{2} K d^2. \tag{95}$$

Given our step size $\delta \le \frac{c_0 \varepsilon}{d \log d}$ by Assumption D.1, $\frac{\delta^2}{2} K d^2 \le \frac{c_0^2 \varepsilon^2}{2 \log d} K$, enabling us to write the bound

$$\mathbb{E}[(M_t^{(\rho)})^2 \mid \mathcal{F}_t] \le K_M^\rho, \tag{96}$$

for an absolute K_M^ρ that depends only on model constants and c_0, ε_*. The exact argument for χ gives

$$\mathbb{E}[(M_t^{(\chi)})^2 \mid \mathcal{F}_t] \le K_M^\chi, \tag{97}$$

for an absolute K_M^χ. □

Lemma D.5 (Martingale inequality for Z_t)**.** *Let*

$$\mathcal{M}_t^{(Z)} := \sum_{k=0}^{t-1} \big(M_k^{(\rho)} + M_k^{(\chi)}\big), \qquad t \ge 0.$$

Under Assumption D.1 and the bounds of Corollary D.4, for all $T \in \mathbb{N}$ and all $\lambda > 0$,

$$\mathbb{P}\Big(\sup_{0 \le t \le T} \big|\delta \mathcal{M}_t^{(Z)}\big| \ \ge\ \lambda\Big) \ \le\ \frac{KT\delta^2}{\lambda^2}, \tag{98}$$

for some constant $K > 0$ depending only on the model.

Proof. Let $D_k := M_k^{(\rho)} + M_k^{(\chi)}$, so that $\mathcal{M}_t^{(Z)} = \sum_{k=0}^{t-1} D_k$ is a martingale with respect to $(\mathcal{F}_t)$. By Doob's L^2 maximal inequality,

$$\mathbb{E}\Big[\sup_{0 \le t \le T} \big|\mathcal{M}_t^{(Z)}\big|^2\Big] \le 4\,\mathbb{E}\big[\big|\mathcal{M}_T^{(Z)}\big|^2\big]. \tag{99}$$

Moreover, using $(a+b)^2 \le 2a^2 + 2b^2$ and Corollary D.4,

$$\mathbb{E}[D_k^2 \mid \mathcal{F}_k] \le 2\mathbb{E}[(M_k^{(\rho)})^2 \mid \mathcal{F}_k] + 2\mathbb{E}[(M_k^{(\chi)})^2 \mid \mathcal{F}_k] \le 4K.$$

Since martingale increments are orthogonal in L^2,

$$\mathbb{E}\big[\big|\mathcal{M}_T^{(Z)}\big|^2\big] = \sum_{k=0}^{T-1} \mathbb{E}[D_k^2] \le \sum_{k=0}^{T-1} 4K = 4KT.$$

Combining with (99) and applying Markov's inequality yields

$$\mathbb{P}\Big(\sup_{0\le t\le T}\big|\delta\mathcal{M}_t^{(Z)}\big|\ge\lambda\Big)\le\frac{\delta^2}{\lambda^2}\,\mathbb{E}\Big[\sup_{0\le t\le T}\big|\mathcal{M}_t^{(Z)}\big|^2\Big]$$
$$\le\frac{\delta^2}{\lambda^2}\cdot 4\,\mathbb{E}\big[\big|\mathcal{M}_T^{(Z)}\big|^2\big]\le\frac{16KT\delta^2}{\lambda^2}.$$

Absorbing the numerical factor into the constant completes the proof. □

Remark D.6. With $\delta\le\frac{c_0\varepsilon}{d\log d}$ and $T:=\lceil C_T\delta^{-1}\log d\rceil$ and picking $\lambda=C_0/\sqrt{d}$, we have

$$\mathbb{P}\Big(\sup_{0\le t\le T}\big|\delta\mathcal{M}_t^{(Z)}\big|\ \ge\ \frac{C_0}{\sqrt{d}}\Big)\ \le\ \frac{KT\delta^2}{\lambda^2}=\frac{KT\delta^2 d}{C_0^2}\le\frac{KC_T\delta d\log d}{C_0^2}\le\frac{KC_Tc_0}{C_0^2}\varepsilon.$$

Lemma D.7 (Discretization terms for Z_t)**.** *Under Assumption D.1 and Lemma E.7, there exists $K>0$ such that for all $T\in\mathbb{N}$ and all $\lambda>0$,*

$$\mathbb{P}\left(E_T^{(T)}\cap\left\{\sup_{t\le T}|\mathcal{R}_t^{(Z)}|\ge\lambda\right\}\right)\le\frac{KT\delta^3d^{3/2}}{\lambda}.$$

Consequently, under the assumptions of Lemma D.3,

$$\mathbb{P}\left(\sup_{t\le T}|\mathcal{R}_t^{(Z)}|\ge\lambda\right)\le\mathbb{P}\left((E_T^{(T)})^c\right)+\frac{KT\delta^3d^{3/2}}{\lambda}.\tag{100}$$

Proof. Let

$$r_k^{(Z)}:=R_k^{(\rho)}+R_k^{(\chi)},\qquad\mathcal{R}_t^{(Z)}:=\sum_{k=0}^{t-1}r_k^{(Z)}.$$

On the Taylor event $E_T^{(T)}$, Lemma D.3 gives, simultaneously for all $k<T$,

$$|R_k^{(\rho)}|\le C\delta^3\|g_k^{(\beta)}\|^3,\qquad|R_k^{(\chi)}|\le C\delta^3\|g_k^{(\theta)}\|^3.$$

Therefore, on $E_T^{(T)}$,

$$|r_k^{(Z)}|\le C\delta^3\left(\|g_k^{(\beta)}\|^3+\|g_k^{(\theta)}\|^3\right).$$

Hence

$$E_T^{(T)}\cap\left\{\sup_{t\le T}|\mathcal{R}_t^{(Z)}|\ge\lambda\right\}\subseteq\left\{C\delta^3\sum_{k<T}\left(\|g_k^{(\beta)}\|^3+\|g_k^{(\theta)}\|^3\right)\ge\lambda\right\}.$$

By Markov's inequality,

$$\mathbb{P}\left(E_T^{(T)}\cap\left\{\sup_{t\le T}|\mathcal{R}_t^{(Z)}|\ge\lambda\right\}\right)\le\frac{C\delta^3\sum_{k<T}\mathbb{E}\left[\|g_k^{(\beta)}\|^3+\|g_k^{(\theta)}\|^3\right]}{\lambda}.$$

By Hölder's inequality and Lemma E.7,

$$\mathbb{E}\left[\|g_k^{(\beta)}\|^3\mid\mathcal{F}_k\right]\le\left(\mathbb{E}\left[\|g_k^{(\beta)}\|^4\mid\mathcal{F}_k\right]\right)^{3/4}\le(Kd^2)^{3/4}\le Kd^{3/2}.$$

The same bound holds for $g_k^{(\theta)}$. Therefore

$$\sum_{k<T}\mathbb{E}\left[\|g_k^{(\beta)}\|^3+\|g_k^{(\theta)}\|^3\right]\le KTd^{3/2}.$$

Thus

$$\mathbb{P}\left(E_T^{(T)}\cap\left\{\sup_{t\le T}|\mathcal{R}_t^{(Z)}|\ge\lambda\right\}\right)\le\frac{KT\delta^3d^{3/2}}{\lambda}.$$

Finally,

$$\left\{\sup_{t\le T}|\mathcal{R}_t^{(Z)}|\ge\lambda\right\}\subseteq(E_T^{(T)})^c\cup\left(E_T^{(T)}\cap\left\{\sup_{t\le T}|\mathcal{R}_t^{(Z)}|\ge\lambda\right\}\right),$$

so

$$\mathbb{P}\left(\sup_{t\le T}|\mathcal{R}_t^{(Z)}|\ge\lambda\right)\le\mathbb{P}\left((E_T^{(T)})^c\right)+\frac{KT\delta^3d^{3/2}}{\lambda}.$$

The second displayed bound follows from Lemma D.3. □

Remark D.8. Taking $T\delta\le c_2\log d$, $\delta\le c_0\varepsilon/(d\log d)$, and $\lambda=C_0/\sqrt{d}$, Lemma D.7 gives

$$\frac{KT\delta^3d^{3/2}}{\lambda}=\frac{KT\delta^3d^2}{C_0}\le\frac{Kc_2}{C_0}\delta^2d^2\log d\le\frac{Kc_2c_0^2}{C_0}\frac{\varepsilon^2}{\log d}.$$

Together with Lemma D.3,

$$\mathbb{P}\left((E_T^{(T)})^c\right)\le K_{\mathrm{Tay}}\frac{\varepsilon^3}{d\log^2 d}.$$

Hence the discretization contribution is absorbed into the existing $O(\varepsilon)$ good-event failure probability.

Good Events for Z_t Now, we formalize the definition and the high-likelihood of good events for Z_t iterate: From the two-layer aggregate martingale and discretization remainder for each time step defined as $M_t^{(Z)}:=M_t^{(\rho)}+M_t^{(\chi)}, r_t^{(Z)}:=R_t^{(\rho)}+R_t^{(\chi)}$, we write their time accumulation as

$$\mathcal{M}_t^{(Z)}:=\sum_{k<t}M_k^{(Z)},\qquad\mathcal{R}_t^{(Z)}:=\sum_{k<t}r_k^{(Z)},\qquad N_t^{(Z)}:=\delta\mathcal{M}_t^{(Z)}+\mathcal{R}_t^{(Z)}.$$

For a threshold $\lambda_Z=C_0/\sqrt{d}$, define

$$E_T^{(Z,M)}:=\left\{\sup_{t\le T}|\delta\mathcal{M}_t^{(Z)}|\le\lambda_Z\right\},\qquad E_T^{(Z,R)}:=\left\{\sup_{t\le T}|\mathcal{R}_t^{(Z)}|\le\lambda_Z\right\},$$

and

$$E_T^{(Z)}:=E_T^{(Z,M)}\cap E_T^{(Z,R)}.\tag{101}$$

Then on $E_T^{(Z)}$,

$$\sup_{t\le T}|N_t^{(Z)}|\le 2\lambda_Z.\tag{102}$$

By Lemma D.5 and Lemma D.7, together with Lemma D.3, we have

$$\mathbb{P}\big((E_T^{(Z)})^c\big)\le\mathbb{P}\big((E_T^{(Z,M)})^c\big)+\mathbb{P}\big((E_T^{(Z,R)})^c\big)\le K\varepsilon,\tag{103}$$

for some constant K.

Good Events for D_t Besides, we keep track of the good events for $D_t:=\rho_t-\chi_t$: for $a\in\{\rho,\chi\}$, we have the individual-layer time accumulation as:

$$\mathcal{M}_t^{(a)}:=\sum_{k<t}M_k^{(a)},\qquad\mathcal{R}_t^{(a)}:=\sum_{k<t}R_k^{(a)},\qquad N_t^{(a)}:=\delta\mathcal{M}_t^{(a)}+\mathcal{R}_t^{(a)}.$$

We then track their difference $N_t^{(D)}:=N_t^{(\rho)}-N_t^{(\chi)}$: With parameter $\lambda_\rho=\lambda_\chi:=\frac{C_{\mathrm{ind}}}{\sqrt{d}}$, $\lambda_D:=\lambda_\rho+\lambda_\chi$, we define the good events here as

$$E_T^{(\rho,\chi)}:=\left\{\sup_{t\le T}|N_t^{(\rho)}|\le 2\lambda_\rho\right\}\cap\left\{\sup_{t\le T}|N_t^{(\chi)}|\le 2\lambda_\chi\right\},\tag{104}$$

So on $E_T^{(\rho,\chi)}$,

$$\sup_{t\le T}|N_t^{(D)}|\le\sup_{t\le T}|N_t^{(\rho)}|+\sup_{t\le T}|N_t^{(\chi)}|\le 2(\lambda_\rho+\lambda_\chi)=2\lambda_D.$$

By the individual versions of Lemmas D.5 and D.7, choosing C_{ind} large enough gives

$$\mathbb{P}\big((E_T^{(\rho,\chi)})^c\big)\le K\varepsilon.\tag{105}$$

D.1.3 Martingale and Discretization Error Control for η_t

Deriving the discrete-time evolution η_t requires a different stragegy compares to that of Z_t. To start with, we derive the one-step evolution η_t:

Lemma D.9 (Evolution of η_t)**.** *Work under Assumption D.1. Let $g_t^{(\theta)} := \nabla_\theta^{\mathcal{S}} \ell(\beta_t, \theta_t; x_t)$ so that $g_t^{(\theta)} \perp \theta_t$, and define the mean-field spherical gradient $\bar{g}_t^{(\theta)} := \mathbb{E}[g_t^{(\theta)} \mid \mathcal{F}_t]$ and $\Delta g_t^{(\theta)} := g_t^{(\theta)} - \bar{g}_t^{(\theta)}$. Then for all t, we have*

$$\|\theta_t\|_\infty \le \|\theta_0\|_\infty + \delta \sum_{k<t} \|\bar{g}_k\|_\infty + \max_i \left|\delta \sum_{k<t} \xi_{k,i}\right| + \frac{\delta^2}{2} \sum_{k<t} \|g_k\|_2^2 \|\theta_k\|_\infty + \frac{\delta^3}{2} \sum_{k<t} \|g_k\|_2^3. \tag{106}$$

Proof. Since $g_t^{(\theta)} \perp \theta_t$ and $\|\theta_t\|_2 = 1$, we have $a_t := 1/\|\theta_t - \delta g_t^{(\theta)}\|_2 = 1/\sqrt{1 + \delta^2 \|g_t^{(\theta)}\|_2^2}$ and

$$\theta_{t+1} = a_t(\theta_t - \delta g_t^{(\theta)}).$$

Define the "telescoping correlation"

$$r_t = \theta_{t+1} - (\theta_t - \delta g_t^{(\theta)}) = (a_t - 1)(\theta_t - \delta g_t^{(\theta)}), \tag{107}$$

so as $\theta_{t+1} = \theta_t - \delta g_t^{(\theta)} + r_t$, under time-cumulation we have

$$\theta_t = \theta_0 - \delta \sum_{k<t} g_k^{(\theta)} + \sum_{k<t} r_k.$$

Now we decompose the gradient $g_k^{(\theta)}$ as the mean-field and martingale noise $g_k^{(\theta)} = \bar{g}_k^{(\theta)} + \xi_k^{(\theta)}$, where $\bar{g}_k := \mathbb{E}[g_k^{(\theta)} \mid \mathcal{F}_k]$ and $\xi_k := g_k - \bar{g}_k$. Thus, taking the l_∞-norm

$$\|\theta_t\|_\infty \le \|\theta_0\|_\infty + \delta \sum_{k<t} \|\bar{g}_k^{(\theta)}\|_\infty + \max_i \left|\delta \sum_{k<t} \xi_{k,i}^{(\theta)}\right| + \sum_{k<t} \|r_k^{(\theta)}\|_\infty \tag{108}$$

To bound $\|r_k^{(\theta)}\|_\infty$, taking l_∞-norm of (107):

$$\begin{aligned}
\|r_t^{(\theta)}\|_\infty &\le |a_t - 1| \cdot \|\theta_t - \delta g_t^{(\theta)}\|_\infty && (109)\\
&\le |1 - a_t| \cdot (\|\theta_t\|_\infty + \delta \|g_t^{(\theta)}\|_\infty) && (110)\\
&\le \frac{\delta^2}{2} \|g_t^{(\theta)}\|_2^2 \cdot (\|\theta_t\|_\infty + \delta \|g_t^{(\theta)}\|_\infty), \text{ since for } x \ge 0, \sqrt{1+x} \ge 1 + x/2. && (111)\\
&= \frac{\delta^2}{2} \|g_t^{(\theta)}\|_2^2 \|\theta_t\|_\infty + \frac{\delta^3}{2} \|g_t^{(\theta)}\|_2^3, \text{ given } \|g_t^{(\theta)}\|_\infty \le \|g_t^{(\theta)}\|. && (112)
\end{aligned}$$

So we have the bound

$$\|\theta_t\|_\infty \le \|\theta_0\|_\infty + \delta \sum_{k<t} \|\bar{g}_k^{(\theta)}\|_\infty + \max_i \left|\delta \sum_{k<t} \xi_{k,i}^{(\theta)}\right| + \frac{\delta^2}{2} \sum_{k<t} \|g_k^{(\theta)}\|_2^2 \|\theta_k\|_\infty + \frac{\delta^3}{2} \sum_{k<t} \|g_k^{(\theta)}\|_2^3. \tag{113}$$

□

Now, we analyze the aggregation through time: let $\Theta_t := \max_{s \le t} \|\theta_s\|_\infty$ to be the maximum over the trajectory, we then have for any $t \le T$,

$$\Theta_t \le \|\theta_0\|_\infty + \underbrace{\delta \sum_{k<t} \|\bar{g}_k^{(\theta)}\|_\infty}_{1} + \underbrace{\max_i \sup_{s \le t} \left|\delta \sum_{k<s} \xi_{k,i}^{(\theta)}\right|}_{2} + \underbrace{\Big(\frac{\delta^2}{2} \sum_{k<t} \|g_k^{(\theta)}\|_2^2\Big)\Theta_t}_{3} + \underbrace{\frac{\delta^3}{2} \sum_{k<t} \|g_k^{(\theta)}\|_2^3}_{4}. \tag{114}$$

To bound term 3 in (114), we have by Markov Inequality,

$$\mathbb{P}\Big(\frac{\delta^2}{2}\sum_{k<T}\|g_k^{(\theta)}\|_2^2 \geq \lambda\Big) \leq \frac{\mathbb{E}\big[\frac{\delta^2}{2}\sum_{k<T}\|g_k^{(\theta)}\|_2^2\big]}{\lambda} \leq \frac{\delta^2 TKd}{\lambda}$$

Taking $\lambda = 1/4$, we have

$$\mathbb{P}\Big(\frac{\delta^2}{2}\sum_{k<T}\|g_k^{(\theta)}\|_2^2 \geq \frac{1}{4}\Big) = O(\varepsilon), \tag{115}$$

giving that w.p. $1 - O(\varepsilon)$,

$$\Theta_t \leq \|\theta_0\|_\infty + \underbrace{\delta\sum_{k<t}\|\bar{g}_k^{(\theta)}\|_\infty}_{1} + \underbrace{\max_i \sup_{s\leq t}\big|\delta\sum_{k<s}\xi_{k,i}^{(\theta)}\big|}_{2} + \frac{1}{4}\Theta_t + \underbrace{\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3}_{4}.$$

$$(1-\frac{1}{4})\Theta_t \leq \|\theta_0\|_\infty + \underbrace{\delta\sum_{k<t}\|\bar{g}_k^{(\theta)}\|_\infty}_{1} + \underbrace{\max_i \sup_{s\leq t}\big|\delta\sum_{k<s}\xi_{k,i}^{(\theta)}\big|}_{2} + \underbrace{\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3}_{4}.$$

$$\Theta_t \leq \frac{4}{3}\Big(\|\theta_0\|_\infty + \underbrace{\delta\sum_{k<t}\|\bar{g}_k^{(\theta)}\|_\infty}_{1} + \underbrace{\max_i \sup_{s\leq t}\big|\delta\sum_{k<s}\xi_{k,i}^{(\theta)}\big|}_{2} + \underbrace{\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3}_{4}\Big). \tag{116}$$

We start by bounding term 2 using a Freedman Inequality. Yet to show it, we need this corollary below:

Corollary D.10 (Coordinate variance and 8th-moment bounds for $\xi^{(\theta)}$)**.** *Work under Assumption D.1 and Lemma E.7. Let $\xi_t^{(\theta)} := g_t^{(\theta)} - \mathbb{E}[g_t^{(\theta)} \mid \mathcal{F}_t]$ and write $\xi_{t,i}^{(\theta)}$ for its i-th coordinate. Then there exist constants $\sigma^2 > 0$ and $K_8 > 0$ (depending only on model constants) such that for all t and all $i \leq d$,*

$$\mathbb{E}\Big[(\xi_{t,i}^{(\theta)})^2 \mid \mathcal{F}_t\Big] \leq \sigma^2, \qquad \mathbb{E}\Big[|\xi_{t,i}^{(\theta)}|^8 \mid \mathcal{F}_t\Big] \leq K_8. \tag{117}$$

Proof. **Conditional variance.** Lemma E.7 gives the directional variance bound

$$\sup_{\|u\|=1}\mathbb{E}\big[((\Delta g_t^{(\theta)})\cdot u)^2 \mid \mathcal{F}_t\big] \leq K.$$

Taking $u = e_i$ yields $\mathbb{E}[(\xi_{t,i}^{(\theta)})^2 \mid \mathcal{F}_t] \leq K$. Set $\sigma^2 := K$.

8-th moment. Recall that in Lemma E.7, from the per-sample gradient $\nabla_\theta \ell(\beta_t, \theta_t; x_t) = A_t z_t$ with $A_t := 2(f_t - f_t^*)\psi'(z_t \cdot \theta_t)$ defined in (215), using the boundedness of ψ, ψ' we showed $|A_t| \leq C_A$ in (219). From there, the spherical gradient is

$$g_t^{(\theta)} = (I - \theta_t\theta_t^\top)\nabla_\theta \ell = A_t\big(z_t - (z_t\cdot\theta_t)\theta_t\big),$$

so for each coordinate

$$|g_{t,i}^{(\theta)}| \leq |A_t|\big(|z_{t,i}| + |z_t\cdot\theta_t|\big) \leq C_A\big(|z_{t,i}| + |z_t\cdot\theta_t|\big). \tag{118}$$

Condition on $\mathcal{F}_t$. Then θ_t is deterministic with $\|\theta_t\|_2 = 1$ and $z_t = (z_{t,1}, \ldots, z_{t,d})$ has i.i.d. coordinates $z_{t,j} = \phi(\gamma_{t,j})$ with $\gamma_{t,j} \sim \mathcal{N}(0,1)$. In particular,

$$\mu_2 := \mathbb{E}[z_{t,1}^2] = \mathbb{E}[\phi(\gamma)^2] < \infty, \qquad \mu_8 := \mathbb{E}[|z_{t,1}|^8] = \mathbb{E}[|\phi(\gamma)|^8] < \infty.$$

Let $X_j := \theta_{t,j} z_{t,j}$. Then $(X_j)_{j\leq d}$ are independent and mean-zero, and

$$\sum_{j=1}^d \mathbb{E}[X_j^2 \mid \mathcal{F}_t] = \sum_{j=1}^d \theta_{t,j}^2\,\mathbb{E}[z_{t,1}^2] = \mu_2\sum_{j=1}^d \theta_{t,j}^2 = \mu_2.$$

By Rosenthal's inequality for $p = 8$, there exists a constant $C_8 > 0$ such that

$$\mathbb{E}\Big[\Big|\sum_{j=1}^{d} X_j\Big|^8 \;\Big|\; \mathcal{F}_t\Big] \le C_8\Big(\Big(\sum_{j=1}^{d} \mathbb{E}[X_j^2 \mid \mathcal{F}_t]\Big)^4 + \sum_{j=1}^{d} \mathbb{E}[|X_j|^8 \mid \mathcal{F}_t]\Big).$$

Moreover,

$$\sum_{j=1}^{d} \mathbb{E}[|X_j|^8 \mid \mathcal{F}_t] = \sum_{j=1}^{d} |\theta_{t,j}|^8\, \mathbb{E}[|z_{t,1}|^8] = \mu_8 \sum_{j=1}^{d} |\theta_{t,j}|^8 \le \mu_8 \sum_{j=1}^{d} \theta_{t,j}^2 = \mu_8,$$

since $|\theta_{t,j}| \le 1$ implies $|\theta_{t,j}|^8 \le |\theta_{t,j}|^2$. Therefore,

$$\mathbb{E}\big[|z_t \cdot \theta_t|^8 \mid \mathcal{F}_t\big] = \mathbb{E}\Big[\Big|\sum_{j=1}^{d} X_j\Big|^8 \;\Big|\; \mathcal{F}_t\Big] \le C_8(\mu_2^4 + \mu_8) =: C,$$

which is a constant independent of t and d. Hence, plugging into (118), we have $\mathbb{E}[|g_{t,i}^{(\theta)}|^8 \mid \mathcal{F}_t] \le K_8'$ uniformly.

Finally, by Jensen and $(a-b)^8 \le 2^7(a^8 + b^8)$,

$$\mathbb{E}[|\xi_{t,i}^{(\theta)}|^8 \mid \mathcal{F}_t] \le 2^8 \mathbb{E}[|g_{t,i}^{(\theta)}|^8 \mid \mathcal{F}_t] \le 2^8 K_8'.$$

Set $K_8 := 2^8 K_8'$ gives the claimed result. □

Now we move to prove the Freedman bound:

Lemma D.11 (Coordinate martingale control for the η-bootstrap)**.** *Assume Corollary D.10. Let $T\delta \le C_T \log d$ and $\delta \le c_0\varepsilon/(d \log d)$. Then there is a constant C_M such that*

$$\mathbb{P}\left(\max_{i \le d} \sup_{t \le T} \left|\delta \sum_{k<t} \xi_{k,i}^{(\theta)}\right| > C_M \sqrt{\frac{\log(40d/\varepsilon)}{d}}\right) \le \varepsilon/5. \tag{119}$$

Proof. Let

$$L := \log \frac{40d}{\varepsilon}.$$

The case $T = 0$ is trivial, so assume $T \ge 1$. Let σ^2, K_8 be the constants from Corollary D.10. Define the truncation level

$$B := \left(\frac{40dTK_8}{\varepsilon}\right)^{1/8}.$$

For $k < T$ and $i \le d$, write

$$\xi_{k,i} := \xi_{k,i}^{(\theta)}.$$

Define the centered truncated martingale difference

$$Y_{k,i} := \xi_{k,i}\mathbf{1}_{\{|\xi_{k,i}| \le B\}} - \mathbb{E}\big[\xi_{k,i}\mathbf{1}_{\{|\xi_{k,i}| \le B\}} \mid \mathcal{F}_k\big].$$

Then

$$\mathbb{E}[Y_{k,i} \mid \mathcal{F}_k] = 0, \qquad |Y_{k,i}| \le 2B,$$

and

$$\mathbb{E}[Y_{k,i}^2 \mid \mathcal{F}_k] \le \mathbb{E}[\xi_{k,i}^2 \mid \mathcal{F}_k] \le \sigma^2.$$

Hence, for each fixed i, the process

$$M_t^{(i)} := \sum_{k<t} Y_{k,i}$$

is a martingale with increments bounded by $2B$ and predictable quadratic variation at most $\sigma^2 t \le \sigma^2 T$.

By the maximal form of Freedman's inequality, for every $x > 0$,

$$\mathbb{P}\left(\sup_{t \le T} M_t^{(i)} \ge x\right) \le \exp\left(-\frac{x^2}{2(\sigma^2 T + 2Bx/3)}\right),$$

and the same estimate holds for $-M_t^{(i)}$. Set

$$x_T := \sqrt{2\sigma^2 TL} + \frac{4}{3}BL.$$

Then

$$\frac{x_T^2}{2(\sigma^2 T + 2Bx_T/3)} \geq L.$$

Indeed, writing $a = \sigma^2 T$, $u = \sqrt{2aL}$, and $v = \frac{4}{3}BL$, we have $x_T = u + v$, while

$$2L\left(a + \frac{2Bx_T}{3}\right) = 2aL + v(u+v).$$

Since

$$x_T^2 = (u+v)^2 = 2aL + 2uv + v^2,$$

we get

$$x_T^2 - 2L\left(a + \frac{2Bx_T}{3}\right) = uv \geq 0.$$

Therefore

$$\mathbb{P}\left(\sup_{t\leq T} |M_t^{(i)}| \geq x_T\right) \leq 2e^{-L}.$$

Taking a union bound over $i \leq d$, and using $L = \log(40d/\varepsilon)$, gives

$$\mathbb{P}\left(\max_{i\leq d}\sup_{t\leq T} |M_t^{(i)}| \geq x_T\right) \leq 2de^{-L} = \frac{\varepsilon}{20}.$$

It remains to compare the original martingale to the truncated one. Let

$$\mathcal{J} := \left\{\max_{k<T}\max_{i\leq d} |\xi_{k,i}| \leq B\right\}.$$

By Corollary D.10 and Markov's inequality,

$$\mathbb{P}(\mathcal{J}^c) \leq \sum_{k<T}\sum_{i\leq d}\mathbb{P}(|\xi_{k,i}| > B) \leq dT\frac{K_8}{B^8} = \frac{\varepsilon}{40}.$$

Now define

$$b_{k,i} := \mathbb{E}\left[\xi_{k,i}\mathbf{1}_{\{|\xi_{k,i}|>B\}} \mid \mathcal{F}_k\right].$$

Since $\mathbb{E}[\xi_{k,i} \mid \mathcal{F}_k] = 0$,

$$\mathbb{E}\left[\xi_{k,i}\mathbf{1}_{\{|\xi_{k,i}|\leq B\}} \mid \mathcal{F}_k\right] = -b_{k,i}.$$

Hence

$$Y_{k,i} = \xi_{k,i}\mathbf{1}_{\{|\xi_{k,i}|\leq B\}} + b_{k,i}.$$

On $\mathcal{J}$, the indicator $\mathbf{1}_{\{|\xi_{k,i}|\leq B\}}$ is identically one for all $k < T$ and $i \leq d$. Therefore, on $\mathcal{J}$,

$$\xi_{k,i} = Y_{k,i} - b_{k,i}.$$

Moreover, by the conditional eighth-moment bound,

$$|b_{k,i}| \leq \mathbb{E}\left[|\xi_{k,i}|\mathbf{1}_{\{|\xi_{k,i}|>B\}} \mid \mathcal{F}_k\right] \leq \frac{\mathbb{E}[|\xi_{k,i}|^8 \mid \mathcal{F}_k]}{B^7} \leq \frac{K_8}{B^7}.$$

Thus, on $\mathcal{J}$,

$$\sup_{t\leq T}\left|\sum_{k<t}\xi_{k,i}\right| \leq \sup_{t\leq T}|M_t^{(i)}| + \frac{TK_8}{B^7}.$$

Combining the previous estimates, with probability at least $1 - \varepsilon/20 - \varepsilon/40$, we have

$$\max_{i\leq d}\sup_{t\leq T}\left|\sum_{k<t}\xi_{k,i}\right| \leq x_T + \frac{TK_8}{B^7}.$$

Multiplying by δ, it remains to show that

$$\delta x_T + \delta \frac{TK_8}{B^7} \le C_M \sqrt{\frac{L}{d}}.$$

First,

$$\delta\sqrt{2\sigma^2 TL} = \sqrt{2\sigma^2 L\,\delta(\delta T)} \le \sqrt{2\sigma^2 C_T c_0 \varepsilon}\sqrt{\frac{L}{d}} \le C\sqrt{\frac{L}{d}},$$

because $T\delta \le C_T \log d$, $\delta \le c_0\varepsilon/(d\log d)$, and $\varepsilon \le 1$.

Next, using $T \le C_T(\log d)/\delta$,

$$\delta BL = \delta L \left(\frac{40dTK_8}{\varepsilon}\right)^{1/8} \le CL\,\delta^{7/8}\left(\frac{d\log d}{\varepsilon}\right)^{1/8}.$$

Using again $\delta \le c_0\varepsilon/(d\log d)$, this gives

$$\delta BL \le C c_0^{7/8} \varepsilon^{3/4} L\, d^{-3/4} (\log d)^{-3/4}.$$

Since $L = \log(40d/\varepsilon)$ and $0 < \varepsilon \le 1$,

$$\varepsilon^{3/4} L^{1/2} \le C(\log d)^{1/2}$$

for $d \ge 3$. Hence

$$\delta BL \le C c_0^{7/8} L^{1/2} d^{-3/4} (\log d)^{-1/4} \le C\sqrt{\frac{L}{d}}.$$

Finally,

$$\delta \frac{TK_8}{B^7} = \delta T K_8 \left(\frac{\varepsilon}{40dTK_8}\right)^{7/8} \le C\delta T^{1/8} \varepsilon^{7/8} d^{-7/8}.$$

Using $T \le C_T(\log d)/\delta$ and $\delta \le c_0\varepsilon/(d\log d)$, we get

$$\delta \frac{TK_8}{B^7} \le C c_0^{7/8} \varepsilon^{7/4} d^{-7/4} (\log d)^{-3/4} \le C\sqrt{\frac{L}{d}}.$$

Therefore, after increasing C_M to absorb constants depending only on σ^2, K_8, C_T, c_0, we have

$$\delta x_T + \delta \frac{TK_8}{B^7} \le C_M \sqrt{\frac{L}{d}} = C_M \sqrt{\frac{\log(40d/\varepsilon)}{d}}.$$

Since

$$\frac{\varepsilon}{20} + \frac{\varepsilon}{40} < \frac{\varepsilon}{5},$$

the desired bound follows. □

We plug this proved Freedman bound into (116) to obtain

$$\Theta_t \le \frac{4}{3}\Big(\|\theta_0\|_\infty + \underbrace{\delta \sum_{k<t} \|\bar{g}_k^{(\theta)}\|_\infty}_{1} + \Lambda_\eta + \underbrace{\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3}_{4}\Big). \tag{120}$$

Since $\eta_t = \|\theta_t\|_\infty + \|\theta^*\|_\infty \le \Theta_t + \|\theta^*\|_\infty$, we re-write it as:

$$\eta_t \le \frac{4}{3}\Big(\eta_0 + \underbrace{\delta \sum_{k<t} \|\bar{g}_k^{(\theta)}\|_\infty}_{1} + \Lambda_\eta + \underbrace{\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3}_{4}\Big). \tag{121}$$

Next we bound 4: notice that since $\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3$ is non decreasing in t, we have

$$\sup_{t\le T} \frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3 = \frac{\delta^3}{2}\sum_{k<T}\|g_k^{(\theta)}\|_2^3.$$

On the other hand,

$$\mathbb{E}\big[\|g_k^{(\theta)}\|_2^3\big] \le \mathbb{E}\big[\|g_k^{(\theta)}\|_2^4\big]^{3/4} \text{ by Holder inequality} \tag{122}$$

$$\le (Kd^2)^{3/4} \text{ by the technical bounds in Lemma E.7} \tag{123}$$

$$= K^{3/4}d^{3/2}. \tag{124}$$

Therefore, for a given λ_{cub},

$$\mathbb{P}\Big(\sup_{t\le T}\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3 \ge \lambda_{\text{cub}}\Big) = \mathbb{P}\Big(\frac{\delta^3}{2}\sum_{k<T}\|g_k^{(\theta)}\|_2^3 \ge \lambda_{\text{cub}}\Big) \text{ by monotonicity} \tag{125}$$

$$\le \frac{\mathbb{E}\big[\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3\big]}{\lambda_{\text{cub}}}, \text{ by Markov Inequality} \tag{126}$$

$$\le \frac{K^{3/4}d^{3/2}T\frac{\delta^3}{2}}{\lambda_{\text{cub}}}, \text{ following (124)} \tag{127}$$

Taking $\lambda_{\text{cub}} = C_{\text{cub}}/\sqrt{d}$, and plugging in $T \le c_2(\log d)/\delta$ and $\delta \le c_0\varepsilon/(d\log d)$, we have

$$\mathbb{P}\left(\sup_{t\le T}\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3 \ge \frac{C_{\text{cub}}}{\sqrt{d}}\right) \le \frac{c_2, c_0^2, K^{3/4}}{2C_{\text{cub}}}\cdot\frac{\varepsilon^2}{\log d}. \tag{128}$$

Choose C_{cub} large enough (absorbing constants) and take d sufficiently large so that $\frac{\varepsilon^2}{\log d} \le \varepsilon/10$, we obtain

$$\mathbb{P}\left(\sup_{t\le T}\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3 \ge \frac{C_{\text{cub}}}{\sqrt{d}}\right) \le \frac{\varepsilon}{10} \tag{129}$$

Plugging it in the η bound in (116) it becomes

$$\eta_t \le C_\eta\Big(\eta_0 + \underbrace{\delta\sum_{k<t}\|\bar{g}_k^{(\theta)}\|_\infty}_{1} + \Lambda_\eta + \frac{C_{\text{cub}}}{\sqrt{d}}\Big), \tag{130}$$

with high probability. The bound of 1 is contingent to dynamics's phase that we will derive it later. We later bound them in Step 1 of the Phase I proof (Lemma D.13) and Step 1(a) of the Phase II proof (Lemma D.16).

Good event for η iteration Define

$$E_T^{(\eta)} := E_T^{(\eta,M)} \cap E_T^{(\eta,2)} \cap E_T^{(\eta,3)}, \tag{131}$$

where

$$E_T^{(\eta,M)} := \left\{\max_{i\le d}\sup_{t\le T}\left|\delta\sum_{k<t}\xi_{k,i}^{(\theta)}\right| \le \Lambda_\eta\right\}, \tag{132}$$

$$E_T^{(\eta,2)} := \left\{\frac{\delta^2}{2}\sum_{k<T}\|g_k^{(\theta)}\|_2^2 \le \frac{1}{4}\right\}, \tag{133}$$

$$E_T^{(\eta,3)} := \left\{\sup_{t\le T}\frac{\delta^3}{2}\sum_{k<t}\|g_k^{(\theta)}\|_2^3 \le \frac{C_{\text{cub}}}{\sqrt{d}}\right\}. \tag{134}$$

Therefore, with Bound (119) from Lemma E.7 to ensure (132), Bound (115) to ensure (133), and Bound (129) to ensure (134), we obtain a union bound,

$$\mathbb{P}\big((E_T^{(\eta)})^c\big) \le \mathbb{P}\big((E_T^{(\eta,M)})^c\big) + \mathbb{P}\big((E_T^{(\eta,2)})^c\big) + \mathbb{P}\big((E_T^{(\eta,3)})^c\big) \le K\varepsilon, \tag{135}$$

after increasing K.

D.1.4 Martingale and Discretization Error Control for Ω_t and ω_t

As a discrete analogy of Lemma C.7, we next provide the control of third moment quantities $\Omega_t := \|\theta_t\|_3^3 = \sum_{i=1}^d |\theta_{t,i}|^3$, and $\omega_t := \sum_{i=1}^d \big(|\theta_{t,i}| + |\theta_i^*|\big)^3$, upon their martingale and discretizatio noises.

Lemma D.12 (Third-moment control for SGD through Phase I)**.** *Let Ω_t and ω_t follow the definition made in Lemma C.7. Fix $T \in \mathbb{N}$ with $T\delta \le C_T \log d$. Under Assumption D.1, Lemma E.7, and Corollary D.10, there is an event $E_T^{(\Omega)}$ satisfying*

$$\mathbb{P}\big((E_T^{(\Omega)})^c\big) \le K_\Omega \varepsilon$$

such that the following holds. Work on $\mathcal{E}_0 \cap E_T^{(\Omega)}$. Let

$$\tau_A := \inf\{t : Z_t^2 \ge A\eta_t\}, \qquad \tau_B := \inf\{t : \rho_t \ge 1/2 \text{ or } \chi_t \ge 1/2\}, \qquad \tau_+ := \inf\{t : \rho_t < 0 \text{ or } \chi_t < 0\},$$

and set

$$\sigma := \tau_A \wedge \tau_B \wedge \tau_+ \wedge T.$$

Suppose that, for some constant $C_\eta > 0$,

$$\eta_t \le C_\eta \frac{\sqrt{\log d} + 1 + t\delta}{\sqrt{d}} \qquad (0 \le t \le \sigma). \tag{136}$$

Then, for all sufficiently large d,

$$\Omega_t \le C_\Omega \frac{1 + t\delta}{\sqrt{d}}, \qquad \omega_t \le C_\omega \frac{1 + t\delta}{\sqrt{d}}, \qquad 0 \le t \le \sigma. \tag{137}$$

The constants $K_\Omega, C_\Omega, C_\omega$ depend only on model constants, A, C_η, C_T, and the step-size regime.

Proof. Write $g_t := g_t^{(\theta)}$, $\bar g_t := \mathbb{E}[g_t \mid \mathcal{F}_t]$, and $\xi_t := g_t - \bar g_t$. Since $g_t \perp \theta_t$, the normalized spherical update is

$$\theta_{t+1} = a_t(\theta_t - \delta g_t), \qquad a_t = (1 + \delta^2 \|g_t\|_2^2)^{-1/2} \le 1.$$

Hence

$$\Omega_{t+1} = \|\theta_{t+1}\|_3^3 \le \|\theta_t - \delta g_t\|_3^3.$$

We use the elementary inequality

$$|x - y|^3 \le |x|^3 - 3|x|x\,y + C\big(|x|y^2 + |y|^3\big), \qquad x, y \in \mathbb{R}.$$

Applying it coordinatewise with $x = \theta_{t,i}$ and $y = \delta g_{t,i}$, and letting $v_t := \big(|\theta_{t,i}|\theta_{t,i}\big)_{i=1}^d$, gives

$$\Omega_{t+1} \le \Omega_t - 3\delta\langle v_t, \bar g_t\rangle - 3\delta\langle v_t, \xi_t\rangle + C\delta^2 \sum_{i=1}^d |\theta_{t,i}| g_{t,i}^2 + C\delta^3 \|g_t\|_3^3. \tag{138}$$

Since $\|v_t\|_1 = \sum_i |\theta_{t,i}|^2 = 1$,

$$-3\delta\langle v_t, \bar g_t\rangle \le 3\delta \|\bar g_t\|_\infty.$$

Summing (138) gives

$$\Omega_t \le \Omega_0 + 3\delta \sum_{k<t} \|\bar g_k\|_\infty + \mathcal{M}_t^{(\Omega)} + C\mathcal{Q}_t^{(\Omega)} + C\mathcal{C}_t^{(\Omega)}, \tag{139}$$

where

$$\mathcal{M}_t^{(\Omega)} := -3\delta \sum_{k<t} \langle v_k, \xi_k\rangle,$$

$$\mathcal{Q}_t^{(\Omega)} := \delta^2 \sum_{k<t} \sum_{i=1}^d |\theta_{k,i}| g_{k,i}^2, \qquad \mathcal{C}_t^{(\Omega)} := \delta^3 \sum_{k<t} \|g_k\|_3^3.$$

We first define the good event $E_T^{(\Omega)}$. Let

$$\lambda_{\Omega,M} := \frac{C_{\Omega,M}}{\sqrt{d}}, \qquad \lambda_{\Omega,Q} := \frac{C_{\Omega,Q}}{\sqrt{d}}, \qquad \lambda_{\Omega,C} := \frac{C_{\Omega,C}}{\sqrt{d}},$$

where the constants will be chosen large. Set

$$E_T^{(\Omega)} := \left\{ \sup_{t\le T} |\mathcal{M}_t^{(\Omega)}| \le \lambda_{\Omega,M} \right\} \cap \left\{ \sup_{t\le T} \mathcal{Q}_t^{(\Omega)} \le \lambda_{\Omega,Q} \right\} \cap \left\{ \sup_{t\le T} \mathcal{C}_t^{(\Omega)} \le \lambda_{\Omega,C} \right\}. \tag{140}$$

We verify that this event has high probability. The process $\mathcal{M}_t^{(\Omega)}$ is a martingale. Moreover,

$$\|v_k\|_2^2 = \sum_i |\theta_{k,i}|^4 \le \sum_i |\theta_{k,i}|^2 = 1.$$

By Lemma E.7,

$$\mathbb{E}\big[\langle v_k, \xi_k\rangle^2 \mid \mathcal{F}_k\big] \le K.$$

Thus Doob's L^2 inequality gives

$$\mathbb{P}\left(\sup_{t\le T} |\mathcal{M}_t^{(\Omega)}| \ge \lambda_{\Omega,M} \right) \le \frac{KT\delta^2}{\lambda_{\Omega,M}^2} = \frac{KT\delta^2 d}{C_{\Omega,M}^2} \le \frac{KC_T c_0}{C_{\Omega,M}^2}\varepsilon.$$

Choosing $C_{\Omega,M}$ large makes this contribution $O(\varepsilon)$.

Next, by the coordinate moment bound proved in Corollary D.10 and (118), we have

$$\mathbb{E}[g_{k,i}^2 \mid \mathcal{F}_k] \le K \qquad \text{uniformly in } i,k.$$

Therefore

$$\mathbb{E}\left[\sum_i |\theta_{k,i}| g_{k,i}^2 \,\middle|\, \mathcal{F}_k \right] \le K \sum_i |\theta_{k,i}| \le K\sqrt{d}.$$

Hence

$$\mathbb{E}[\mathcal{Q}_T^{(\Omega)}] \le KT\delta^2\sqrt{d}.$$

By Markov's inequality,

$$\mathbb{P}\left(\sup_{t\le T} \mathcal{Q}_t^{(\Omega)} \ge \lambda_{\Omega,Q} \right) \le \frac{KT\delta^2\sqrt{d}}{C_{\Omega,Q}/\sqrt{d}} = \frac{KT\delta^2 d}{C_{\Omega,Q}} \le \frac{KC_T c_0}{C_{\Omega,Q}}\varepsilon.$$

Choosing $C_{\Omega,Q}$ large makes this contribution $O(\varepsilon)$.

Similarly, the same coordinate moment bounds imply

$$\mathbb{E}[|g_{k,i}|^3 \mid \mathcal{F}_k] \le K, \qquad \mathbb{E}[\|g_k\|_3^3 \mid \mathcal{F}_k] \le Kd.$$

Thus

$$\mathbb{E}[\mathcal{C}_T^{(\Omega)}] \le KT\delta^3 d.$$

Again by Markov's inequality,

$$\mathbb{P}\left(\sup_{t\le T} \mathcal{C}_t^{(\Omega)} \ge \lambda_{\Omega,C} \right) \le \frac{KT\delta^3 d}{C_{\Omega,C}/\sqrt{d}} = \frac{KT\delta^3 d^{3/2}}{C_{\Omega,C}}.$$

Since $T\delta \le C_T \log d$ and $\delta \le c_0\varepsilon/(d\log d)$, the last display is $o(\varepsilon)$ for d large. Therefore, after increasing the constants if necessary,

$$\mathbb{P}\big((E_T^{(\Omega)})^c\big) \le K_\Omega \varepsilon. \tag{141}$$

It remains to prove the pathwise bound on Ω_t. Work on $\mathcal{E}_0 \cap E_T^{(\Omega)}$. From (139) and (140),

$$\Omega_t \le \Omega_0 + 3\delta \sum_{k<t} \|\bar{g}_k\|_\infty + \frac{C}{\sqrt{d}}, \qquad t \le T. \tag{142}$$

We now bound the mean-field term for $t \le \sigma$. Since $k < \sigma$ implies $k < \tau_+$, we have $\rho_k, \chi_k \ge 0$. Since $k < \sigma$ also implies $k < \tau_A$, we have $Z_k^2 < A\eta_k$. By (48) in Lemma C.4,

$$\|\bar{g}_k\|_\infty \le CZ_k\|\theta^*\|_\infty + CZ_k^2\eta_k + C\eta_k^2.$$

On $\mathcal{E}_0$, $\|\theta^*\|_\infty \le C\sqrt{\log d/d}$. Using $Z_k^2 < A\eta_k$ and the assumed envelope for η_k, we obtain

$$Z_k\|\theta^*\|_\infty \le \sqrt{A\eta_k}\,\|\theta^*\|_\infty \le C\sqrt{\frac{\sqrt{\log d}+1+k\delta}{\sqrt{d}}}\sqrt{\frac{\log d}{d}}.$$

Since $k\delta \le T\delta \le C_T \log d$, the right-hand side is bounded by $C(\log d)d^{-3/4} \le Cd^{-1/2}$ for all sufficiently large d. Also, $Z_k^2\eta_k \le A\eta_k^2$. Hence, for $k < \sigma$,

$$\|\bar{g}_k\|_\infty \le \frac{C}{\sqrt{d}} + C\eta_k^2. \tag{143}$$

Therefore, for $t \le \sigma$,

$$\delta\sum_{k<t}\|\bar{g}_k\|_\infty \le C\frac{t\delta}{\sqrt{d}} + C\delta\sum_{k<t}\eta_k^2.$$

Using the assumed η-envelope,

$$\delta\sum_{k<t}\eta_k^2 \le \frac{C}{d}\delta\sum_{k<t}\big(\sqrt{\log d}+1+k\delta\big)^2 \le \frac{C(\log d)^3}{d} = o(d^{-1/2}).$$

Plugging this into (142), and using $\Omega_0 \le C/\sqrt{d}$ on $\mathcal{E}_0$, gives

$$\Omega_t \le C\frac{1+t\delta}{\sqrt{d}}, \qquad 0 \le t \le \sigma.$$

This proves the desired Ω_t-bound.

Finally,

$$\omega_t = \sum_i\big(|\theta_{t,i}| + |\theta_i^*|\big)^3 \le 4\sum_i|\theta_{t,i}|^3 + 4\sum_i|\theta_i^*|^3 = 4\Omega_t + 4\|\theta^*\|_3^3.$$

On $\mathcal{E}_0$, $\|\theta^*\|_3^3 \le C/\sqrt{d}$. Therefore

$$\omega_t \le C_\omega\frac{1+t\delta}{\sqrt{d}}, \qquad 0 \le t \le \sigma.$$

This completes the proof. □

D.2 Weak Recovery Dynamics

Good Events for Weak Recovery Now, we define the good event for weak recovery by taking the intersection of $\mathcal{E}_{\mathrm{deloc}}$ from B.2 $E_T^{(T)}$ from (78), $E_T^{(Z)}$ from (101), $E_T^{(\rho,\chi)}$ from (104), $E_T^{(\eta)}$ from (131), and $E_T^{(\Omega)}$ from (140):

$$\mathcal{G}_T^{\mathrm{weak}} := \mathcal{E}_0 \cap E_T^{(T)} \cap E_T^{(Z)} \cap E_T^{(\rho,\chi)} \cap E_T^{(\eta)} \cap E_T^{(\Omega)}. \tag{144}$$

With their respective probabilistic controls made in (21), (89), (103), (105), (135), (141), they are true with probability $(1 - K_{\mathrm{alg}}\varepsilon)(1 - d^{-K_{\mathrm{ini}}}.)$

D.2.1 Phase I Dynamics, Weak Recovery

Lemma D.13 (Phase I: pull-up to the signal-dominant scale)**.** *Work on $\mathcal{G}_T^{\mathrm{weak}}$. Then, up to $\tau_A \wedge \tau_B$,*

$$\eta_t \le C_\eta\frac{\sqrt{\log d}+1+t\delta}{\sqrt{d}}, \qquad \omega_t \le C_\omega\frac{1+t\delta}{\sqrt{d}}. \tag{145}$$

Moreover,

$$\tau_A < \tau_B, \qquad \tau_+ > \tau_A, \qquad \tau_A\delta \le C\log d, \qquad Z_{\tau_A} \ge c_A d^{-1/4}. \tag{146}$$

Equivalently,

$$\rho_t, \chi_t > 0 \qquad \textit{for all } 0 \le t \le \tau_A. \tag{147}$$

Finally,

$$Z_{\tau_A} \le C_A\sqrt{\log d}\,d^{-1/4} + o(1). \tag{148}$$

Proof. Work throughout on $\mathcal{G}_T^{\mathrm{weak}}$. We use the favorable initialization event in the form

$$\rho_0, \chi_0 \geq \frac{z_-}{\sqrt{d}}, \qquad \eta_0 \leq C\sqrt{\frac{\log d}{d}},$$

where $z_- > 0$ is fixed large enough below. Recall that

$$\tau_+ := \inf\{t : \rho_t < 0 \text{ or } \chi_t < 0\}.$$

Set

$$\sigma := \tau_A \wedge \tau_B \wedge \tau_+ \wedge T.$$

It is enough to prove the desired estimates up to σ, and then show that $\sigma = \tau_A$.

Step 1: Phase-I η-envelope. Let

$$\Theta_t := \max_{0 \leq s \leq t} \eta_s, \qquad H_I(t) := K_I \frac{\sqrt{\log d} + 1 + t\delta}{\sqrt{d}},$$

where K_I will be chosen large. By the discrete η-estimate (130), on $E_T^{(\eta)}$,

$$\eta_t \leq C\left(\eta_0 + \delta \sum_{k<t} \|\bar{g}_k^{(\theta)}\|_\infty + \Lambda_\eta + \frac{C_{\mathrm{cub}}}{\sqrt{d}}\right), \qquad t \leq T.$$

The definition of $E_T^{(\eta)}$ gives

$$\Lambda_\eta \leq C\sqrt{\frac{\log(40d/\varepsilon)}{d}} \leq C\frac{\sqrt{\log d} + 1 + t\delta}{\sqrt{d}}, \qquad t \leq T,$$

For $k < \sigma$, we have $k < \tau_+$, hence $\rho_k, \chi_k \geq 0$, and also $k < \tau_A$, hence $Z_k^2 < A\eta_k$. The population ℓ_∞-gradient bound (48) gives

$$\|\bar{g}_k^{(\theta)}\|_\infty \leq CZ_k\|\theta^*\|_\infty + CZ_k^2\eta_k + C\eta_k^2.$$

Using $Z_k^2 < A\eta_k$, this becomes

$$\|\bar{g}_k^{(\theta)}\|_\infty \leq C\sqrt{\eta_k}\,\|\theta^*\|_\infty + C\eta_k^2.$$

Since $\|\theta^*\|_\infty \leq C\sqrt{\log d/d}$, and since throughout the bootstrap $k\delta \leq T\delta \leq C_T \log d$, the implication $\eta_k \leq H_I(k)$ gives

$$\sqrt{\eta_k}\,\|\theta^*\|_\infty \leq C\frac{\log d}{d^{3/4}} \leq \frac{C}{\sqrt{d}}$$

for all sufficiently large d. Thus, under the bootstrap,

$$\|\bar{g}_k^{(\theta)}\|_\infty \leq \frac{C}{\sqrt{d}} + C\eta_k^2.$$

Let

$$\tau_H := \inf\{t \leq \sigma : \Theta_t > H_I(t)\}.$$

Suppose for contradiction that $\tau_H \leq \sigma$. Then for every $k < \tau_H$, $\eta_k \leq \Theta_k \leq H_I(k)$, and therefore

$$\delta \sum_{k<\tau_H} \|\bar{g}_k^{(\theta)}\|_\infty \leq C\frac{\tau_H\delta}{\sqrt{d}} + C\delta \sum_{k<\tau_H} H_I(k)^2.$$

Moreover,

$$\delta \sum_{k<\tau_H} H_I(k)^2 \leq \frac{CK_I^2}{d}\delta \sum_{k<\tau_H} \left(\sqrt{\log d} + 1 + k\delta\right)^2 \leq \frac{C(\log d)^3}{d} = o\left(\frac{\sqrt{\log d} + 1 + \tau_H\delta}{\sqrt{d}}\right).$$

Combining the previous estimates with the bound on η_t, we get

$$\Theta_{\tau_H} \leq C\frac{\sqrt{\log d} + 1 + \tau_H\delta}{\sqrt{d}} + o\left(\frac{\sqrt{\log d} + 1 + \tau_H\delta}{\sqrt{d}}\right).$$

Choosing K_I large enough, and then d large enough, gives

$$\Theta_{\tau_H} \le \frac{1}{2} H_I(\tau_H),$$

contradicting the definition of τ_H. Hence $\tau_H > \sigma$, and therefore

$$\eta_t \le C_\eta \frac{\sqrt{\log d} + 1 + t\delta}{\sqrt{d}}, \qquad 0 \le t \le \sigma, \tag{149}$$

which is exactly the assumption (136) made in Lemma D.12. Alongside the condition $Z_t^2 < A\eta_t$ for $t < \tau_A$, we can apply Lemma D.12 to obtain

$$\omega_t \le C_\omega \frac{1 + t\delta}{\sqrt{d}}, \qquad 0 \le t \le \sigma. \tag{150}$$

Step 2: Positivity and scalar Z-pull-up. For $t < \sigma$, we have $t < \tau_A$, hence $Z_t^2 < A\eta_t$. By Step 1,

$$\eta_t \le C_\eta \frac{\sqrt{\log d} + 1 + t\delta}{\sqrt{d}} \le C \frac{\log d}{\sqrt{d}}.$$

Thus

$$Z_t \le C\sqrt{\log d}\, d^{-1/4} = o(1).$$

For d large, $Z_t \le c_{\mathrm{sm}}$, so we can apply Lemma C.6

$$\dot\rho_t \ge a\chi_t - b\omega_t, \qquad \dot\chi_t \ge a\rho_t - b\omega_t.$$

Combining this with Lemma D.2, the Taylor event $E_T^{(T)}$, and Lemma E.7, we obtain

$$u_{t+1} \ge \left[(1 - C\delta^2 d)I + a\delta J\right] u_t - C\delta\omega_t \mathbf{1} + \varepsilon_t^{(u)}, \qquad u_t := (\rho_t, \chi_t)^\top,$$

where

$$J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \mathbf{1} = (1,1)^\top, \qquad \varepsilon_t^{(u)} = \left(N_{t+1}^{(\rho)} - N_t^{(\rho)}, N_{t+1}^{(\chi)} - N_t^{(\chi)}\right)^\top.$$

Let

$$B := (1 - C\delta^2 d)I + a\delta J.$$

For d large, B has nonnegative entries. Its eigenvalues in the $\mathbf{1}$- and $(1,-1)^\top$-directions are

$$\gamma_+ := 1 + a\delta - C\delta^2 d, \qquad \gamma_- := 1 - a\delta - C\delta^2 d.$$

After decreasing a, we may assume

$$\gamma_+ \ge 1 + \beta\delta, \qquad 0 < \gamma_- < 1, \qquad \beta > 0.$$

Unrolling the previous cooperative recursion and applying the summation-by-parts bounds to the Z-noise and D-noise gives, componentwise,

$$\rho_t, \chi_t \ge \gamma_+^t \left[\frac{z_-}{\sqrt{d}} - C\sum_{k<t} \gamma_+^{-(k+1)} \delta\omega_k - C(\lambda_Z + \lambda_\rho + \lambda_\chi)\right], \qquad 0 \le t \le \sigma. \tag{151}$$

Indeed, the Z-noise is controlled by Corollary E.9, while the contracting D-noise is controlled by Corollary E.10.

By (150),

$$\sum_{k<t} \gamma_+^{-(k+1)} \delta\omega_k \le \frac{C}{\sqrt{d}} \left(\delta \sum_{k\ge 0} \gamma_+^{-(k+1)} + \delta^2 \sum_{k\ge 0} k\gamma_+^{-(k+1)}\right) \le \frac{C}{\sqrt{d}}.$$

Since

$$\lambda_Z + \lambda_\rho + \lambda_\chi = O(d^{-1/2}),$$

choosing z_- larger than the constants in the bracket yields

$$\rho_t, \chi_t > 0, \qquad 0 \le t \le \sigma. \tag{152}$$

In particular, $\sigma \neq \tau_+$.

Now sum the two one-step inequalities. For

$$\varepsilon_t^{(Z)} := N_{t+1}^{(Z)} - N_t^{(Z)},$$

we get

$$Z_{t+1} \geq (1 + \beta\delta) Z_t - C\delta\omega_t + \varepsilon_t^{(Z)}, \qquad t < \sigma.$$

Set $\gamma := 1 + \beta\delta$. Multiplying by $\gamma^{-(t+1)}$ and summing,

$$\gamma^{-t} Z_t \geq Z_0 - C \sum_{k<t} \gamma^{-(k+1)} \delta\omega_k + \sum_{k<t} \gamma^{-(k+1)} \varepsilon_k^{(Z)}.$$

The weighted ω-sum is $O(d^{-1/2})$ by (150). Also, on $E_T^{(Z)}$,

$$\sup_{u \leq T} |N_u^{(Z)}| \leq 2\lambda_Z,$$

so Corollary E.9 gives

$$\left| \sum_{k<t} \gamma^{-(k+1)} \varepsilon_k^{(Z)} \right| \leq 4\lambda_Z.$$

Since $Z_0 \geq 2z_- d^{-1/2}$, increasing z_- if necessary gives a constant $c_- > 0$ such that

$$Z_t \geq \frac{c_-}{\sqrt{d}} \gamma^t, \qquad 0 \leq t \leq \sigma. \tag{153}$$

Step 3: Some stopping event occurs by $O(\delta^{-1} \log d)$. Let

$$t_0 := \left\lceil c_6 \delta^{-1} \log d \right\rceil,$$

where c_6 will be chosen large. Choose the weak-recovery horizon constant C_T so that $t_0 < T$. Suppose, for contradiction, that $\sigma > t_0$. Then $t_0 < \tau_A$, hence

$$Z_{t_0}^2 < A\eta_{t_0}.$$

Using (153) and (149),

$$\frac{c_-^2}{d} \gamma^{2t_0} \leq Z_{t_0}^2 < A\eta_{t_0} \leq C \frac{\sqrt{\log d} + 1 + t_0\delta}{\sqrt{d}} \leq C \frac{\log d}{\sqrt{d}}.$$

Since $\log \gamma \geq c\delta$ for some $c > 0$, we have

$$\gamma^{2t_0} \geq d^{2cc_6 + o(1)}.$$

Choosing c_6 so large that $2cc_6 > 1/2$ gives a contradiction for all large d. Therefore

$$\sigma \leq t_0, \qquad \sigma\delta \leq c_6 \log d + O(\delta). \tag{154}$$

Step 4: The first stopping event is τ_A. Since $t_0 < T$, (154) implies $\sigma \neq T$. By (152), $\sigma \neq \tau_+$. Thus σ is either τ_A or τ_B.

Suppose for contradiction that $\sigma = \tau_B \leq \tau_A$. Since $\tau_B \leq t_0$, at time $\tau_B - 1$ we have

$$Z_{\tau_B - 1}^2 < A\eta_{\tau_B - 1}.$$

Thus, by (149),

$$Z_{\tau_B - 1} \leq C \frac{\sqrt{\log d}}{d^{1/4}} = o(1).$$

The one-step expansion gives

$$Z_{\tau_B} = Z_{\tau_B - 1} + \delta F_{\tau_B - 1}^{(Z)} + \varepsilon_{\tau_B - 1}^{(Z)}.$$

By Lemma E.7,

$$|F_t^{(Z)}| \leq C\sqrt{d} + C\delta d,$$

and hence

$$\delta|F_t^{(Z)}| \le C\delta\sqrt{d} + C\delta^2 d = o(1).$$

Also,

$$|\varepsilon_{\tau_B-1}^{(Z)}| \le |N_{\tau_B}^{(Z)}| + |N_{\tau_B-1}^{(Z)}| \le 4\lambda_Z = o(1).$$

Therefore $Z_{\tau_B} = o(1)$. But (152) gives $\rho_{\tau_B}, \chi_{\tau_B} > 0$, and by definition of τ_B, one of $\rho_{\tau_B}, \chi_{\tau_B}$ is at least $1/2$. Hence $Z_{\tau_B} \ge 1/2$, a contradiction. Therefore

$$\sigma = \tau_A, \qquad \tau_A < \tau_B, \qquad \tau_A\delta \le C\log d.$$

The bounds

$$\eta_t \le C_\eta \frac{\sqrt{\log d} + 1 + t\delta}{\sqrt{d}}, \qquad \omega_t \le C_\omega \frac{1 + t\delta}{\sqrt{d}}$$

now hold for all $t \le \tau_A$, and since $\tau_A < \tau_B$, also up to $\tau_A \wedge \tau_B$.

Step 5: Size of Z_{τ_A}. By definition of τ_A,

$$Z_{\tau_A}^2 \ge A\eta_{\tau_A}.$$

Since $\theta_{\tau_A} \in \mathbb{S}^{d-1}$,

$$\|\theta_{\tau_A}\|_\infty \ge d^{-1/2},$$

and therefore $\eta_{\tau_A} \ge d^{-1/2}$. Hence

$$Z_{\tau_A} \ge \sqrt{A}\, d^{-1/4}.$$

This proves the lower bound with $c_A := \sqrt{A}$.

For the upper bound, first note that $\tau_A \ge 1$ for large d, since $Z_0^2 = O(d^{-1})$ while $\eta_0 \ge d^{-1/2}$. Hence $\tau_A - 1 < \tau_A$, and so

$$Z_{\tau_A-1}^2 < A\eta_{\tau_A-1}.$$

Using the η-bound and $\tau_A\delta \le C\log d$,

$$Z_{\tau_A-1} \le C\frac{\sqrt{\log d}}{d^{1/4}}.$$

The one-step expansion gives

$$Z_{\tau_A} \le Z_{\tau_A-1} + \delta|F_{\tau_A-1}^{(Z)}| + |\varepsilon_{\tau_A-1}^{(Z)}|.$$

As above,

$$\delta|F_{\tau_A-1}^{(Z)}| \le C\delta\sqrt{d} + C\delta^2 d = o(1), \qquad |\varepsilon_{\tau_A-1}^{(Z)}| \le 4\lambda_Z = o(1).$$

Therefore

$$Z_{\tau_A} \le C_A\sqrt{\log d}\, d^{-1/4} + o(1).$$

This completes the proof. □

Lemma D.14 (Entry balance at τ_A)**.** *Work on $\mathcal{G}_T^{\mathrm{weak}}$, and assume the conclusions of Lemma D.13. Let*

$$D_t := \rho_t - \chi_t, \qquad \lambda_D := \lambda_\rho + \lambda_\chi.$$

Then

$$|D_{\tau_A}| \le C\frac{(\log d)^2}{\sqrt{d}} + C\lambda_D + C\delta d\,\frac{\sqrt{\log d}}{d^{1/4}}.$$

In particular,

$$|D_{\tau_A}| = o(Z_{\tau_A}).$$

Hence, for all sufficiently large d,

$$|D_{\tau_A}| \le \frac{1}{8} Z_{\tau_A}. \tag{155}$$

Proof. By Lemma D.13, on $0 \le t \le \tau_A$ we have

$$\tau_A < \tau_B, \qquad \tau_+ > \tau_A, \qquad \tau_A \delta \le C \log d,$$

and for $t \le \tau_A$,

$$\eta_t \le C_\eta \frac{\sqrt{\log d} + 1 + t\delta}{\sqrt{d}}.$$

Moreover, for $t < \tau_A$,

$$Z_t^2 < A\eta_t.$$

Therefore, for $t < \tau_A$,

$$Z_t \le C \frac{\sqrt{\log d}}{d^{1/4}} = o(1).$$

Thus, throughout the interval $0 \le t < \tau_A$, the overlaps are nonnegative and lie in the small-overlap regime. Hence, by (52) from Lemma C.6,

$$\dot{D}_t = -\widetilde{C}(1 + \rho_t \chi_t) D_t + E_t, \qquad |E_t| \le C(Z_t^2 + \omega_t).$$

Now use the one-step SGD expansion. Let

$$a_t := \mathbb{E}[\|g_t^{(\beta)}\|^2 \mid \mathcal{F}_t], \qquad b_t := \mathbb{E}[\|g_t^{(\theta)}\|^2 \mid \mathcal{F}_t].$$

Subtracting the one-step overlap expansions gives

$$D_{t+1} = D_t + \delta(\dot{\rho}_t - \dot{\chi}_t) - \frac{\delta^2}{2}(a_t \rho_t - b_t \chi_t) + \varepsilon_t^{(D)},$$

where

$$\varepsilon_t^{(D)} := N_{t+1}^{(D)} - N_t^{(D)}, \qquad N_t^{(D)} := N_t^{(\rho)} - N_t^{(\chi)}.$$

Decompose

$$a_t \rho_t - b_t \chi_t = \frac{a_t + b_t}{2} D_t + \frac{a_t - b_t}{2} Z_t.$$

Therefore

$$D_{t+1} = (1 - \alpha_t \delta) D_t + \delta e_t + \varepsilon_t^{(D)}, \tag{1}$$

where

$$\alpha_t := \widetilde{C}(1 + \rho_t \chi_t) + \frac{\delta}{4}(a_t + b_t),$$

and

$$e_t := E_t - \frac{\delta}{4}(a_t - b_t) Z_t.$$

By Lemma E.7, $a_t, b_t \le Kd$. Hence

$$|e_t| \le C(Z_t^2 + \eta_t) + C\delta d\, Z_t. \tag{2}$$

Also, since $\rho_t, \chi_t \ge 0$,

$$\alpha_t \ge \widetilde{C}.$$

After reducing the step-size constant, we may assume

$$0 \le 1 - \alpha_t \delta \le 1 - \alpha_0 \delta < 1$$

for a fixed $\alpha_0 > 0$.

Unroll (1) up to $t = \tau_A$. Define

$$P_{m,t} := \prod_{\ell=m}^{t-1} (1 - \alpha_\ell \delta), \qquad P_{t,t} := 1.$$

Then

$$D_{\tau_A} = P_{0,\tau_A} D_0 + \sum_{k=0}^{\tau_A - 1} P_{k+1,\tau_A} \delta e_k + \sum_{k=0}^{\tau_A - 1} P_{k+1,\tau_A} \varepsilon_k^{(D)}.$$

Since $0 \le P_{k+1,\tau_A} \le 1$,

$$|P_{0,\tau_A} D_0| \le |D_0| \le C d^{-1/2},$$

by the favorable initialization event.

For the first forcing term, using $Z_k^2 \le A\eta_k$ for $k < \tau_A$,

$$\delta \sum_{k<\tau_A} (Z_k^2 + \eta_k) \le C\delta \sum_{k<\tau_A} \eta_k.$$

By the Phase-I η-bound and $\tau_A \delta \le C \log d$,

$$\delta \sum_{k<\tau_A} \eta_k \le \frac{C}{\sqrt{d}} \delta \sum_{k<\tau_A} \left(\sqrt{\log d} + 1 + k\delta\right) \le C \frac{(\log d)^2}{\sqrt{d}}. \tag{3}$$

For the discretization forcing term, since

$$P_{k+1,\tau_A} \le (1 - \alpha_0 \delta)^{\tau_A - 1 - k},$$

we have

$$\delta \sum_{k<\tau_A} P_{k+1,\tau_A} Z_k \le \left(\sup_{k<\tau_A} Z_k\right) \delta \sum_{j \ge 0} (1 - \alpha_0 \delta)^j \le C \sup_{k<\tau_A} Z_k.$$

As observed above,

$$\sup_{k<\tau_A} Z_k \le C \frac{\sqrt{\log d}}{d^{1/4}}.$$

Therefore

$$\delta \sum_{k<\tau_A} P_{k+1,\tau_A}\, \delta d\, Z_k \le C\delta d\, \frac{\sqrt{\log d}}{d^{1/4}}. \tag{4}$$

It remains to control the noise term. For fixed τ_A, the weights

$$P_{k+1,\tau_A}, \qquad 0 \le k \le \tau_A - 1,$$

lie in $[0, 1]$ and are nondecreasing in k. On $E_T^{(\rho,\chi)}$,

$$\sup_{u \le T} |N_u^{(D)}| \le 2(\lambda_\rho + \lambda_\chi) = 2\lambda_D.$$

Thus Lemma E.8 gives

$$\left| \sum_{k=0}^{\tau_A - 1} P_{k+1,\tau_A} \varepsilon_k^{(D)} \right| \le 4\lambda_D. \tag{5}$$

Combining (2)–(5), and absorbing $|D_0| \le C d^{-1/2}$ into the first term, gives

$$|D_{\tau_A}| \le C \frac{(\log d)^2}{\sqrt{d}} + C\lambda_D + C\delta d\, \frac{\sqrt{\log d}}{d^{1/4}}.$$

Finally, Lemma D.13 gives

$$Z_{\tau_A} \ge c_A d^{-1/4}.$$

Hence

$$\frac{|D_{\tau_A}|}{Z_{\tau_A}} \le C \frac{(\log d)^2}{d^{1/4}} + C \frac{\lambda_D}{d^{-1/4}} + C\delta d \sqrt{\log d}.$$

Since $\lambda_D = O(d^{-1/2})$ and

$$\delta d \le \frac{c_0 \varepsilon}{\log d},$$

the right-hand side tends to zero. Therefore

$$|D_{\tau_A}| = o(Z_{\tau_A}).$$

In particular, for all sufficiently large d,

$$|D_{\tau_A}| \le \frac{1}{8} Z_{\tau_A}.$$

□

Before Phase II, we provide another lemma below that controls the difference of drift

Lemma D.15 (Balanced D-drift on the Phase-II band)**.** *Assume Lemma C.2. Let $D := \rho - \chi$, $Z := \rho + \chi$, and define*

$$\widetilde{C} := 2\alpha_1^2 a_1^2 > 0.$$

There exist constants $c_{\mathrm{bd}} \in (0, 1/8)$ and $C_E > 0$, depending only on the model, such that the following holds. Suppose

$$0 < Z \le 2c_{\mathrm{bd}}, \qquad |D| \le \frac{1}{2} Z. \tag{156}$$

Then $\rho, \chi \ge 0$, $|\rho|, |\chi| \le 3c_{\mathrm{bd}}/2$, and the population drifts satisfy

$$\dot{\rho} - \dot{\chi} = -\widetilde{C}(1 + \rho\chi)D + E, \tag{157}$$

where the error obeys the sharper bound

$$|E| \le C_E(Z^2 + \eta). \tag{158}$$

In particular, since $Z \le 2c_{\mathrm{bd}}$, one also has the weaker bound

$$|E| \le C_E(c_{\mathrm{bd}}^2 + \eta). \tag{159}$$

Proof. From $Z = \rho + \chi$ and $D = \rho - \chi$,

$$\rho = \frac{Z + D}{2}, \qquad \chi = \frac{Z - D}{2}.$$

The balance condition $|D| \le Z/2$ gives

$$\frac{1}{4} Z \le \rho, \chi \le \frac{3}{4} Z.$$

Thus $\rho, \chi \ge 0$, and since $Z \le 2c_{\mathrm{bd}}$, we also have

$$|\rho|, |\chi| \le \frac{3}{4} Z \le \frac{3}{2} c_{\mathrm{bd}} < \frac{1}{2}. \tag{160}$$

Let L_G denote the Gaussian-equivalent population loss. As in the proof of Lemma C.3,

$$L_G(\theta, \beta) = \Phi(\chi, \rho), \qquad \Phi(\chi, \rho) = 2 \sum_{k \ge 1} \alpha_k^2 \left[1 - (\kappa(\rho)\chi)^k\right],$$

where $\kappa(\rho) = \sum_{\ell \ge 1} a_\ell^2 \rho^\ell$. Set

$$q := \kappa(\rho)\chi, \qquad A(q) := \sum_{k \ge 1} k \alpha_k^2 q^{k-1}.$$

Then $A(0) = \alpha_1^2$. By the Gaussian β-gradient identity (30), equivalently by differentiating Φ in ρ, and by the corresponding χ-derivative displayed in the proof of Lemma C.3, the Gaussian-model overlap drifts are

$$\dot{\rho}_G = 2\chi\, \kappa'(\rho) A(q)(1 - \rho^2), \qquad \dot{\chi}_G = 2\kappa(\rho) A(q)(1 - \chi^2). \tag{161}$$

We now expand these identities on the small band (156). Since $|\rho| \le 3c_{\mathrm{bd}}/2 < 1/2$, the Hermite covariance function satisfies

$$\kappa(\rho) = a_1^2 \rho + O(\rho^2), \qquad \kappa'(\rho) = a_1^2 + O(|\rho|). \tag{162}$$

Moreover, by (160), $q = \kappa(\rho)\chi = O(Z^2)$, so

$$A(q) = \alpha_1^2 + O(Z^2). \tag{163}$$

Using (161)–(163),

$$\dot{\rho}_G = 2\alpha_1^2 a_1^2\, \chi(1 - \rho^2) + O(Z^2),$$

because the error from $\kappa'(\rho) - a_1^2$ is $O(|\chi||\rho|) = O(Z^2)$, and the error from $A(q) - \alpha_1^2$ is $O(|\chi| Z^2) = O(Z^3)$. Similarly,

$$\dot{\chi}_G = 2\alpha_1^2 a_1^2\, \rho(1 - \chi^2) + O(Z^2),$$

because $\kappa(\rho) - a_1^2\rho = O(\rho^2) = O(Z^2)$.

Thus, with $\widetilde{C} := 2\alpha_1^2 a_1^2$,

$$\dot{\rho}_G = \widetilde{C}\,\chi(1-\rho^2) + O(Z^2), \qquad \dot{\chi}_G = \widetilde{C}\,\rho(1-\chi^2) + O(Z^2). \tag{164}$$

It remains to transfer the Gaussian drift estimates to the true population loss. By (24) and (25) from Lemma C.2,

$$|\dot{\rho} - \dot{\rho}_G| \le C\eta, \qquad |\dot{\chi} - \dot{\chi}_G| \le C\eta.$$

Combining this with (164), we get

$$\dot{\rho} - \dot{\chi} = \widetilde{C}\big[\chi(1-\rho^2) - \rho(1-\chi^2)\big] + O(Z^2 + \eta).$$

Finally,

$$\chi(1-\rho^2) - \rho(1-\chi^2) = (\chi-\rho)(1+\rho\chi) = -D(1+\rho\chi). \tag{165}$$

Therefore

$$\dot{\rho} - \dot{\chi} = -\widetilde{C}(1+\rho\chi)D + E, \qquad |E| \le C_E(Z^2+\eta),$$

which proves (157) and (158). Since $Z \le 2c_{\mathrm{bd}}$, (159) follows immediately. □

D.2.2 Phase II Dynamics, Weak Recovery

Lemma D.16 (Phase II: signal-dominated amplification and balance)**.** *Assume Lemmas D.13, D.14, and D.15. Work on $\mathcal{G}_T^{\mathrm{weak}}$. There exist constants*

$$c_Z \in (0, c_{\mathrm{bd}} \wedge 1/8), \qquad C_2, C_{\eta,\mathrm{II}}, C_{\mathrm{II}} > 0,$$

depending only on the model constants, such that the following holds for all sufficiently large d. If

$$C_{\mathrm{II}} \log d \le T\delta \le C_T \log d,$$

and

$$t^* := \inf\{t \ge \tau_A : Z_t \ge c_Z\},$$

then

$$t^* \le \tau_A + \left\lceil C_2\delta^{-1}\log d \right\rceil, \qquad Z_{t^*} \in [c_Z, 2c_Z].$$

Moreover, for every $\tau_A \le t \le t^*$,

$$\eta_t \le C_{\eta,\mathrm{II}}\frac{\log d}{\sqrt{d}}, \qquad |D_t| \le \frac{1}{2}Z_t.$$

Consequently,

$$\rho_t, \chi_t \ge \frac{1}{4}Z_t \qquad (\tau_A \le t \le t^*), \tag{166}$$

with

$$t^* < \tau_B, \qquad \min\{\rho_{t^*}, \chi_{t^*}\} \ge \frac{c_Z}{4}. \tag{167}$$

Proof. Work on $\mathcal{G}_T^{\mathrm{weak}}$. By Lemma D.13 and Lemma D.14,

$$\tau_A < \tau_B, \qquad \tau_A\delta \le C\log d, \qquad Z_{\tau_A} \ge c_A d^{-1/4}, \tag{168}$$

and

$$\eta_{\tau_A} \le C\frac{\log d}{\sqrt{d}}, \qquad Z_{\tau_A} = o(1), \qquad |D_{\tau_A}| \le \frac{1}{8}Z_{\tau_A}. \tag{169}$$

Also, on $\mathcal{G}_T^{\mathrm{weak}}$,

$$\lambda_Z, \lambda_D = O(d^{-1/2}) = o(Z_{\tau_A}).$$

Choose

$$H := C_{\eta,\mathrm{II}}\frac{\log d}{\sqrt{d}}, \qquad T_2 := \tau_A + \left\lceil C_2\delta^{-1}\log d \right\rceil.$$

The constants $C_{\eta,\mathrm{II}}$, c_Z, and C_2 will be fixed below, and C_{II} is chosen large enough so that $T_2 \le T$, using $\tau_A\delta \le C\log d$.

Define the bootstrap stopping time

$$\sigma := \inf\left\{t \ge \tau_A : \eta_t > H \text{ or } Z_t > 2c_Z \text{ or } |D_t| > \frac{1}{2}Z_t \text{ or } t = T_2\right\}.$$

At $t = \tau_A$, none of the first three clauses holds for large d. For $t < \sigma$,

$$0 < Z_t \le 2c_Z, \qquad |D_t| \le \frac{1}{2}Z_t.$$

Hence

$$\rho_t = \frac{Z_t + D_t}{2}, \qquad \chi_t = \frac{Z_t - D_t}{2}, \qquad \frac{1}{4}Z_t \le \rho_t, \chi_t \le \frac{3}{4}Z_t. \tag{170}$$

Taking $c_Z < 1/3$, this implies $\rho_t, \chi_t < 1/2$, so all one-step expansions below are valid before σ.

Step 1: Z-growth and no-drop. Let

$$\varepsilon_t^{(Z)} := N_{t+1}^{(Z)} - N_t^{(Z)}.$$

For $t < \sigma$, the one-step overlap expansion and the aggregate drift bound give

$$Z_{t+1} \ge Z_t + \delta(C_9 Z_t - C_{11}\eta_t) - C\delta^2 d\, Z_t + \varepsilon_t^{(Z)}.$$

Since $\delta d = o(1)$, for large d,

$$C_9 - C\delta d \ge \frac{3}{4}C_9.$$

Set

$$\tau_- := \inf\left\{t \ge \tau_A : Z_t < \frac{1}{2}Z_{\tau_A}\right\}.$$

For $t < \sigma \wedge \tau_-$, we have $Z_t \ge Z_{\tau_A}/2$ and $\eta_t \le H$. Therefore

$$\frac{\eta_t}{Z_t} \le \frac{2H}{Z_{\tau_A}} \le C\frac{\log d}{d^{1/4}} = o(1).$$

Thus, for large d,

$$C_{11}\eta_t \le \frac{1}{4}C_9 Z_t,$$

and hence

$$Z_{t+1} \ge (1 + \beta\delta)Z_t + \varepsilon_t^{(Z)}, \qquad t < \sigma \wedge \tau_-, \tag{171}$$

where $\beta := C_9/2$. Let $\gamma := 1 + \beta\delta$.

Unrolling (171), for every $s \ge 0$ with $\tau_A + s < \sigma \wedge \tau_-$,

$$Z_{\tau_A+s} \ge \gamma^s\left(Z_{\tau_A} + \sum_{k=0}^{s-1}\gamma^{-(k+1)}\varepsilon_{\tau_A+k}^{(Z)}\right).$$

On $E_T^{(Z)}$, (102) gives

$$\sup_{u \le T}|N_u^{(Z)}| \le 2\lambda_Z.$$

By Corollary E.9 applied on the shifted interval starting at τ_A,

$$\left|\sum_{k=0}^{s-1}\gamma^{-(k+1)}\varepsilon_{\tau_A+k}^{(Z)}\right| \le 4\lambda_Z.$$

Therefore

$$Z_{\tau_A+s} \ge \big(Z_{\tau_A} - 4\lambda_Z\big)\gamma^s. \tag{172}$$

Since $\lambda_Z = o(Z_{\tau_A})$, the right side is at least $(3/4)Z_{\tau_A}\gamma^s$. This rules out $\tau_- \le \sigma$. Hence

$$Z_t \ge \frac{1}{2}Z_{\tau_A} \qquad (\tau_A \le t < \sigma),$$

and (172) holds for every $\tau_A + s < \sigma$.

We shall use the following summability estimates. Summing (171) from τ_A to $t-1$, for $t < \sigma$, gives

$$\beta\delta \sum_{k=\tau_A}^{t-1} Z_k \le Z_t - Z_{\tau_A} - \big(N_t^{(Z)} - N_{\tau_A}^{(Z)}\big).$$

Since $Z_t \le 2c_Z$ and $|N_t^{(Z)} - N_{\tau_A}^{(Z)}| \le 4\lambda_Z$,

$$\delta \sum_{k=\tau_A}^{t-1} Z_k \le Cc_Z. \tag{173}$$

Also, since $Z_k \le 2c_Z$,

$$\delta \sum_{k=\tau_A}^{t-1} Z_k^2 \le Cc_Z^2. \tag{174}$$

Finally, the backward form of (171) gives, for $\tau_A \le k < t < \sigma$,

$$Z_k \le \gamma^{-(t-k)} Z_t + C\lambda_Z. \tag{175}$$

Indeed, unroll (171) from k to t, divide by γ^{t-k}, and apply Corollary E.9 to the shifted weighted noise sum.

Step 2: Closing the η-bootstrap. Since $E_T^{(\eta)}$ controls the cumulative coordinate martingales from time 0, for every $t \ge \tau_A$ and every $i \le d$,

$$\left|\delta \sum_{k=\tau_A}^{t-1} \xi_{k,i}^{(\theta)}\right| \le \left|\delta \sum_{k<t} \xi_{k,i}^{(\theta)}\right| + \left|\delta \sum_{k<\tau_A} \xi_{k,i}^{(\theta)}\right| \le 2\Lambda_\eta.$$

Given that the quadratic and cubic remainder sums over $[\tau_A, t]$ are bounded by their global good-event bounds, therefore, applying Lemma D.9 on the interval $[\tau_A, t]$, for every $\tau_A \le t \le \sigma$,

$$\eta_t \le C\left(\eta_{\tau_A} + \delta \sum_{k=\tau_A}^{t-1} \|\bar{g}_k^{(\theta)}\|_\infty + 2\Lambda_\eta + \lambda_{\mathrm{cub}}\right).$$

After increasing C, we write this as

$$\eta_t \le C\left(\eta_{\tau_A} + \delta \sum_{k=\tau_A}^{t-1} \|\bar{g}_k^{(\theta)}\|_\infty + \Lambda_\eta + \lambda_{\mathrm{cub}}\right).$$

On the Phase-II band, the mean-field ℓ_∞-gradient bound gives

$$\|\bar{g}_k^{(\theta)}\|_\infty \le CZ_k\|\theta^*\|_\infty + CZ_k^2\eta_k + C\eta_k^2.$$

Using $\|\theta^*\|_\infty \le C\sqrt{\log d/d}$ and (173),

$$\delta \sum_{k=\tau_A}^{t-1} Z_k\|\theta^*\|_\infty \le Cc_Z\sqrt{\frac{\log d}{d}} = o(H).$$

By (174) and the bootstrap bound $\eta_k \le H$,

$$\delta \sum_{k=\tau_A}^{t-1} Z_k^2\eta_k \le H\delta \sum_{k=\tau_A}^{t-1} Z_k^2 \le Cc_Z^2 H.$$

Also,

$$\delta \sum_{k=\tau_A}^{t-1} \eta_k^2 \le H^2(t-\tau_A)\delta \le C\frac{(\log d)^3}{d} = o(H).$$

Together with

$$\eta_{\tau_A} + \lambda_\eta(t) + \lambda_{\mathrm{cub}} \le C\frac{\log d}{\sqrt{d}},$$

we obtain

$$\eta_t \le C\frac{\log d}{\sqrt{d}} + Cc_Z^2 H + o(H).$$

Choose c_Z small enough that $Cc_Z^2 \le 1/8$, then choose $C_{\eta,\mathrm{II}}$ large enough that $C(\log d)/\sqrt{d} \le H/4$. For large d,

$$\eta_t \le \frac{1}{2}H \qquad (\tau_A \le t \le \sigma).$$

Thus the stopping clause $\eta_t > H$ cannot be the first active clause before t^*.

Step 3: Closing the D/Z-bootstrap. For $t < \sigma$, choose $c_Z \le c_{\mathrm{bd}}$. Then (170) allows us to apply Lemma D.15:

$$\dot{D}_t = -\widetilde{C}(1+\rho_t\chi_t)D_t + E_t, \qquad |E_t| \le C(Z_t^2 + \eta_t).$$

The discrete one-step expansion for D_t has the form

$$D_{t+1} = (1-\alpha_t\delta)D_t + \delta e_t + \varepsilon_t^{(D)}, \tag{176}$$

where

$$\varepsilon_t^{(D)} := N_{t+1}^{(D)} - N_t^{(D)},$$

$$\alpha_t = \widetilde{C}(1+\rho_t\chi_t) + \frac{\delta}{4}\left(\mathbb{E}[\|g_t^{(\beta)}\|^2 \mid \mathcal{F}_t] + \mathbb{E}[\|g_t^{(\theta)}\|^2 \mid \mathcal{F}_t]\right),$$

and

$$|e_t| \le C(Z_t^2 + \eta_t) + C\delta d\, Z_t. \tag{177}$$

Here the last term comes from the difference of the two normalization corrections. Since c_Z is small, $\alpha_t \ge \alpha_0 > 0$, and after decreasing the step-size constant,

$$0 \le 1 - \alpha_t\delta \le 1 - \alpha_0\delta =: \gamma_D < 1.$$

Unroll (176). For $\tau_A \le t \le \sigma$, define

$$P_{k+1,t} := \prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\delta), \qquad \tau_A \le k \le t-1,$$

with $P_{t,t} = 1$. Then

$$D_t = P_{\tau_A,t}D_{\tau_A} + \sum_{k=\tau_A}^{t-1} P_{k+1,t}\delta e_k + \sum_{k=\tau_A}^{t-1} P_{k+1,t}\varepsilon_k^{(D)}.$$

The entry term is controlled by Lemma D.14 and the no-drop estimate:

$$|P_{\tau_A,t}D_{\tau_A}| \le |D_{\tau_A}| \le \frac{1}{8}Z_{\tau_A} \le \frac{1}{4}Z_t.$$

For the noise term, the weights $P_{k+1,t}$ lie in $[0,1]$ and are nondecreasing in k. On $E_T^{(\rho,\chi)}$,

$$\sup_{u\le T}|N_u^{(D)}| \le 2\lambda_D.$$

Therefore Lemma E.8 gives

$$\left|\sum_{k=\tau_A}^{t-1} P_{k+1,t}\varepsilon_k^{(D)}\right| \le 4\lambda_D = o(Z_{\tau_A}) = o(Z_t).$$

It remains to bound the forcing. Using (175),

$$Z_k \le \gamma^{-(t-k)}Z_t + C\lambda_Z.$$

Since $P_{k+1,t} \le \gamma_D^{t-1-k}$, we have

$$\delta \sum_{k=\tau_A}^{t-1} P_{k+1,t} Z_k \le C\delta \sum_{j\ge 0} \gamma_D^j \gamma^{-j} Z_t + C\lambda_Z \delta \sum_{j\ge 0} \gamma_D^j \le CZ_t + C\lambda_Z.$$

Since $\lambda_Z = o(Z_t)$, this is $CZ_t + o(Z_t)$. Similarly,

$$\delta \sum_{k=\tau_A}^{t-1} P_{k+1,t} Z_k^2 \le CZ_t^2 + C\lambda_Z^2 = CZ_t^2 + o(Z_t).$$

Also, by the η-bootstrap,

$$\delta \sum_{k=\tau_A}^{t-1} P_{k+1,t} \eta_k \le CH = o(Z_{\tau_A}) = o(Z_t).$$

Using (177), we conclude

$$\sum_{k=\tau_A}^{t-1} P_{k+1,t} \delta |e_k| \le Cc_Z Z_t + o(Z_t) + C\delta d\, Z_t.$$

Since $\delta d = o(1)$,

$$\sum_{k=\tau_A}^{t-1} P_{k+1,t} \delta |e_k| \le Cc_Z Z_t + o(Z_t).$$

Thus

$$|D_t| \le \frac{1}{4} Z_t + Cc_Z Z_t + o(Z_t).$$

Choose c_Z smaller, if necessary, so that $Cc_Z \le 1/16$, and then take d large enough so that the $o(Z_t)$-term is at most $Z_t/16$. Hence

$$|D_t| \le \frac{3}{8} Z_t < \frac{1}{2} Z_t \qquad (\tau_A \le t \le \sigma).$$

Therefore the stopping clause $|D_t| > Z_t/2$ cannot be the first active clause before t^*.

Step 4: Hitting c_Z before T_2. From (172),

$$Z_{\tau_A+s} \ge \big(Z_{\tau_A} - 4\lambda_Z\big)\gamma^s$$

as long as $\tau_A + s < \sigma$. Define

$$s_* := \left\lceil \frac{\log(c_Z/(Z_{\tau_A} - 4\lambda_Z))}{\log \gamma} \right\rceil.$$

Since $Z_{\tau_A} - 4\lambda_Z \ge cd^{-1/4}$ and $\log\gamma \ge c\delta$, we have

$$s_* \le C\delta^{-1} \log d.$$

Choose C_2 larger than this constant.

Suppose, for contradiction, that $t^* > \tau_A + s_*$. Then $Z_t < c_Z$ up to $\tau_A + s_*$, so the $Z_t > 2c_Z$ clause cannot trigger before that time. By Steps 2 and 3, the η- and D/Z-clauses cannot trigger before t^*. By the choice of C_2, the time clause $t = T_2$ also cannot trigger before $\tau_A + s_*$. Hence $\tau_A + s_* < \sigma$, and

$$Z_{\tau_A+s_*} \ge \big(Z_{\tau_A} - 4\lambda_Z\big)\gamma^{s_*} \ge c_Z,$$

contradicting the definition of t^*. Therefore

$$t^* \le \tau_A + s_* \le \tau_A + \left\lceil C_2\delta^{-1} \log d \right\rceil.$$

Step 5: Overshoot and conclusion. Since $Z_{\tau_A} = o(1)$, for large d, $Z_{\tau_A} < c_Z$, so $t^* > \tau_A$ and $Z_{t^*-1} < c_Z$. The one-step expansion gives

$$Z_{t^*} = Z_{t^*-1} + \delta F^{(Z)}_{t^*-1} + \varepsilon^{(Z)}_{t^*-1}.$$

By Lemma E.7,

$$|F^{(Z)}_t| \le C\sqrt{d} + C\delta d.$$

Hence

$$\delta|F^{(Z)}_t| \le C\delta\sqrt{d} + C\delta^2 d = o(1).$$

Also, on $E^{(Z)}_T$,

$$|\varepsilon^{(Z)}_t| \le |N^{(Z)}_{t+1}| + |N^{(Z)}_t| \le 4\lambda_Z = o(1).$$

Therefore

$$Z_{t^*} \le c_Z + o(1) \le 2c_Z$$

for all large d. By definition, $Z_{t^*} \ge c_Z$, so

$$Z_{t^*} \in [c_Z, 2c_Z].$$

The bootstrap conclusions give, for $\tau_A \le t \le t^*$,

$$\eta_t \le H = C_{\eta,\mathrm{II}}\frac{\log d}{\sqrt{d}}, \qquad |D_t| \le \frac{1}{2}Z_t.$$

Thus

$$\rho_t = \frac{Z_t + D_t}{2} \ge \frac{1}{4}Z_t, \qquad \chi_t = \frac{Z_t - D_t}{2} \ge \frac{1}{4}Z_t.$$

At $t = t^*$,

$$\min\{\rho_{t^*}, \chi_{t^*}\} \ge \frac{1}{4}Z_{t^*} \ge \frac{c_Z}{4}.$$

Moreover,

$$\max\{\rho_{t^*}, \chi_{t^*}\} \le \frac{Z_{t^*} + |D_{t^*}|}{2} \le \frac{3}{4}Z_{t^*} \le \frac{3}{2}c_Z < \frac{1}{2},$$

after choosing $c_Z < 1/3$. Hence $t^* < \tau_B$. This completes the proof. □

D.2.3 Weak Recovery Theorem

Now, we're prepared to derive the final theorem for weak recovery, which states:

Theorem D.17 (Escape from mediocrity under discrete-time SGD)**.** *Assume Assumption D.1 and the auxiliary estimates used in Lemmas D.13, D.14, and D.16. Let $\mathcal{E}_0$ be the favorable initialization event. Then there exist constants*

$$c_{\mathrm{esc}}, C_{\mathrm{time}}, K > 0$$

depending only on the model constants such that, for all sufficiently large d, the following holds. Let

$$T := \left\lceil C_{\mathrm{time}}\delta^{-1}\log d \right\rceil.$$

Conditional on $\mathcal{E}_0$, with probability at least $1 - K\varepsilon$, there exists a time

$$t_{\mathrm{esc}} \le C_{\mathrm{time}}\delta^{-1}\log d \tag{178}$$

such that

$$\min\{\rho_{t_{\mathrm{esc}}}, \chi_{t_{\mathrm{esc}}}\} \ge c_{\mathrm{esc}}. \tag{179}$$

Moreover, throughout the escape window,

$$\eta_t \le C\frac{\log d}{\sqrt{d}}, \qquad 0 \le t \le t_{\mathrm{esc}}. \tag{180}$$

Equivalently, if $\mathbb{P}(\mathcal{E}_0) \ge p_0 > 0$, then

$$\mathbb{P}\left(\exists\, t_{\mathrm{esc}} \le C_{\mathrm{time}}\delta^{-1}\log d : \min\{\rho_{t_{\mathrm{esc}}}, \chi_{t_{\mathrm{esc}}}\} \ge c_{\mathrm{esc}}\right) \ge p_0(1 - K\varepsilon). \tag{181}$$

Proof. Choose C_{time} large enough so that

$$C_{\mathrm{time}} > C_{\mathrm{I}} + C_{\mathrm{II}} + 2,$$

where C_{I} is the Phase-I time constant from Lemma D.13, and C_{II} is the Phase-II time constant from Lemma D.16. Increasing C_{time} if necessary, the horizon T satisfies all horizon assumptions required in the Phase-I, entry-balance, and Phase-II lemmas.

Define the weak-recovery good event, excluding initialization, by

$$\mathcal{H}_T^{\mathrm{weak}} := E_T^{(T)} \cap E_T^{(Z)} \cap E_T^{(\eta)} \cap E_T^{(\rho,\chi)} \cap E_T^{(\Omega)}.$$

By bounds of the Taylor (89), accumulative (103) and difference (105) martingale and discretization control, η-control (135), (141) and third-moment estimates,

$$\mathbb{P}\left((\mathcal{H}_T^{\mathrm{weak}})^c \mid \mathcal{E}_0\right) \le K\varepsilon. \tag{182}$$

Set

$$\mathcal{G}_T^{\mathrm{weak}} := \mathcal{E}_0 \cap \mathcal{H}_T^{\mathrm{weak}}.$$

We now work on $\mathcal{G}_T^{\mathrm{weak}}$.

By Lemma D.13,

$$\tau_A < \tau_B, \qquad \tau_A \delta \le C_{\mathrm{I}} \log d, \qquad Z_{\tau_A} \ge c_A d^{-1/4}. \tag{183}$$

The same lemma also gives, for $0 \le t \le \tau_A$,

$$\eta_t \le C_\eta \frac{\sqrt{\log d} + 1 + t\delta}{\sqrt{d}}. \tag{184}$$

Next, Lemma D.14 gives

$$|D_{\tau_A}| = o(Z_{\tau_A}).$$

Hence, for all sufficiently large d,

$$|D_{\tau_A}| \le \frac{1}{8} Z_{\tau_A}. \tag{185}$$

Now apply Lemma D.16. Define

$$t^* := \inf\{t \ge \tau_A : Z_t \ge c_Z\},$$

where $c_Z > 0$ is the constant from Lemma D.16. The Phase-II lemma gives

$$t^* \le \tau_A + \left\lceil C_{\mathrm{II}} \delta^{-1} \log d \right\rceil,$$

$$Z_{t^*} \in [c_Z, 2c_Z],$$

and, for every $\tau_A \le t \le t^*$,

$$\eta_t \le C_{\eta,\mathrm{II}} \frac{\log d}{\sqrt{d}}, \qquad |D_t| \le \frac{1}{2} Z_t. \tag{186}$$

Therefore,

$$t^* \le C_{\mathrm{time}} \delta^{-1} \log d$$

for all sufficiently large d.

Set

$$t_{\mathrm{esc}} := t^*, \qquad c_{\mathrm{esc}} := \frac{c_Z}{4}.$$

Since

$$\rho_t = \frac{Z_t + D_t}{2}, \qquad \chi_t = \frac{Z_t - D_t}{2},$$

the balance estimate at $t = t^*$ gives

$$\min\{\rho_{t_{\mathrm{esc}}}, \chi_{t_{\mathrm{esc}}}\} = \frac{Z_{t_{\mathrm{esc}}} - |D_{t_{\mathrm{esc}}}|}{2} \ge \frac{1}{4} Z_{t_{\mathrm{esc}}} \ge \frac{c_Z}{4} = c_{\mathrm{esc}}.$$

This proves weak recovery on $\mathcal{G}_T^{\mathrm{weak}}$.

It remains to verify the claimed η_t-bound. For $0 \le t \le \tau_A$, (184) and $\tau_A \delta \le C_{\mathrm{I}} \log d$ imply

$$\eta_t \le C \frac{\log d}{\sqrt{d}}.$$

For $\tau_A \le t \le t_{\mathrm{esc}}$, the same bound follows from (186). Hence

$$\eta_t \le C \frac{\log d}{\sqrt{d}}, \qquad 0 \le t \le t_{\mathrm{esc}}.$$

Finally, the conditional probability statement follows from (182). If $\mathbb{P}(\mathcal{E}_0) \ge p_0$, then

$$\mathbb{P}(\mathcal{E}_0 \cap \mathcal{H}_T^{\mathrm{weak}}) = \mathbb{P}(\mathcal{H}_T^{\mathrm{weak}} \mid \mathcal{E}_0)\mathbb{P}(\mathcal{E}_0) \ge p_0(1 - K\varepsilon),$$

which gives the unconditional statement. □

D.3 Strong Recovery Setups

Remark D.18 (Local strong-recovery step-size regime)**.** Entering strong recovery phase, we changed the step size to

$$\delta_{\mathrm{loc}} := \frac{c_{\mathrm{loc}}}{d},$$

where $c_{\mathrm{loc}} > 0$ is chosen sufficiently small so that $C_{\mathrm{ret}} C_{\mathrm{sq}}\, \delta_{\mathrm{loc}} d \le \mu_{\mathrm{loc}}$ for the optimum convergence rate: our one-step recursion is $Q_{t+1} \le Q_t + \delta\langle \bar{g}_t, \mathrm{Log}_t\rangle + \delta \Xi_t + C_{\mathrm{ret}}\delta^2 G_t$ With $\langle \bar{g}_t, \mathrm{Log}_t\rangle \le -2\mu_{\mathrm{loc}} Q_t$ and $\mathbb{E}[G_t \mid \mathcal{F}_t] \le C_{\mathrm{sq}} d Q_t$, we have in expectation $\mathbb{E}[Q_{t+1} \mid \mathcal{F}_t] \le (1 - 2\mu_{\mathrm{loc}} Q_t + Cd\delta^2) Q_t$, which means by taking $\delta \asymp \frac{1}{d}$, we will have the maximized contraction rate of $O(d \log \frac{1}{\varepsilon})$ samples.

Population inputs reused from the continuous-time analysis. Throughout this subsection, let $r_{\mathrm{loc}} > 0$ and $\mu_{\mathrm{loc}} > 0$ be the constants from Lemma C.14. In particular, for every

$$y \in \mathcal{U}_{r_{\mathrm{loc}}}, \qquad \mathcal{U}_{r_{\mathrm{loc}}} := \{y \in \mathcal{M} : \mathrm{dist}_{\mathcal{M}}(y, y^*) \le r_{\mathrm{loc}}\},$$

we have

$$\mathrm{Hess}_{\mathcal{M}} L(y) \succeq \mu_{\mathrm{loc}} I.$$

Discrete-time basin-entry. Mirroring Lemma C.9, the following lemmas establish that after an additional $O(d)$ iterations beyond weak recovery, the trajectory enters the local strongly-convex cap in discrete-time:

Lemma D.19 (Entry-window delocalization and overlap-noise control)**.** *Work on the weak-recovery event at t_{esc}. Assume the incoherence bound (A4). After t_{esc}, run SGD with the local step size $\delta_{\mathrm{loc}} = \frac{c_{\mathrm{loc}}}{d}$, where $c_{\mathrm{loc}} > 0$ is chosen sufficiently small depending only on the model constants, C_{in}, and the desired entry-window failure level. Fix $C_{\mathrm{in}} < \infty$, and set $N_{\mathrm{in}} := \lceil C_{\mathrm{in}} \delta_{\mathrm{loc}}^{-1} \rceil$. Then, conditionally on $\mathcal{F}_{t_{\mathrm{esc}}}$, there exists an event E_{in}, measurable with respect to the samples in $[t_{\mathrm{esc}}, t_{\mathrm{esc}} + N_{\mathrm{in}}]$, such that*

$$\mathbb{P}(E_{\mathrm{in}} \mid \mathcal{F}_{t_{\mathrm{esc}}}) \ge 1 - \nu_{\mathrm{in}}. \tag{187}$$

On E_{in},

$$\sup_{t_{\mathrm{esc}} \le t \le t_{\mathrm{esc}} + N_{\mathrm{in}}} \eta_t \le \bar{\eta}_d, \tag{188}$$

where, with $L_{\nu,d} := 1 + \log(d/\nu_{\mathrm{in}})$,

$$\bar{\eta}_d \le C \left(\frac{\log d}{\sqrt{d}} + \sqrt{\frac{\log d}{d}} + \sqrt{\delta_{\mathrm{loc}} L_{\nu,d}} + L_{\nu,d} \delta_{\mathrm{loc}}^2 d^{3/2} \right). \tag{189}$$

In particular, for fixed or polynomially small ν_{in},

$$\bar{\eta}_d = O\left(\frac{\log d}{\sqrt{d}} \right) = o_d(1). \tag{190}$$

Moreover, for $j \in \{\rho, \chi\}$,

$$\sup_{t_{\mathrm{esc}} \le s \le u \le t_{\mathrm{esc}}+N_{\mathrm{in}}} \left| \sum_{k=s}^{u-1} \Delta N_{k+1}^{(j)} \right| \le \Lambda_d, \tag{191}$$

where

$$\Lambda_d \le C \left(\sqrt{\delta_{\mathrm{loc}} \log(1/\nu_{\mathrm{in}})} + L_{\nu,d} \delta_{\mathrm{loc}}^2 d^{3/2} \right) = o_d(1) \tag{192}$$

for fixed or polynomially small ν_{in}.

Proof. Write

$$t_0 := t_{\mathrm{esc}}, \qquad \delta := \delta_{\mathrm{loc}}, \qquad N := N_{\mathrm{in}} = \left\lceil C_{\mathrm{in}} \delta^{-1} \right\rceil.$$

Throughout the proof we condition on $\mathcal{F}_{t_0}$ and on the weak-recovery event stated in the lemma. Since the samples after t_0 are fresh and independent of $\mathcal{F}_{t_0}$, the martingale and moment estimates from the weak-recovery section apply conditionally on $\mathcal{F}_{t_0}$, with the same constants.

We shall use the normalized spherical one-step expansion

$$\rho_{t+1} = \rho_t + \delta F_t^{(\rho)} + \Delta N_{t+1}^{(\rho)}, \qquad \chi_{t+1} = \chi_t + \delta F_t^{(\chi)} + \Delta N_{t+1}^{(\chi)},$$

where

$$\Delta N_{t+1}^{(j)} = \delta M_t^{(j)} + R_t^{(j)}, \qquad j \in \{\rho, \chi\}.$$

Here $M_t^{(j)}$ is a scalar martingale increment and $R_t^{(j)}$ is the normalization/retraction remainder.

Step 1: overlap-noise control.

By the scalar martingale moment bounds used in the weak-recovery proof, there is $K_M < \infty$, independent of d, such that

$$\mathbb{E}\Big[(M_t^{(j)})^2 \mid \mathcal{F}_t \Big] \le K_M (1 + \delta^2 d^2), \qquad j \in \{\rho, \chi\}.$$

Since $\delta d = c_{\mathrm{loc}}$, this is bounded by another constant K_M'.

For $u \in [t_0, t_0 + N]$, define

$$S_u^{(j)} := \delta \sum_{k=t_0}^{u-1} M_k^{(j)}.$$

Then

$$\langle S^{(j)} \rangle_{t_0+N} \le K_M' \delta^2 N \le C\delta,$$

because $N \le C\delta^{-1}$. By the same conditional Freedman/truncation argument used for the weak-recovery overlap martingales, with conditional probability at least $1 - \nu_{\mathrm{in}}/8$,

$$\sup_{t_0 \le u \le t_0+N} \left| \delta \sum_{k=t_0}^{u-1} M_k^{(j)} \right| \le C \sqrt{\delta \log(1/\nu_{\mathrm{in}})}.$$

Consequently,

$$\sup_{t_0 \le s \le u \le t_0+N} \left| \delta \sum_{k=s}^{u-1} M_k^{(j)} \right| \le C \sqrt{\delta \log(1/\nu_{\mathrm{in}})},$$

after increasing C.

Next consider the remainders. On the local Taylor event for the normalized spherical step,

$$|R_t^{(\rho)}| \le C\delta^3 \|g_t^{(\beta)}\|^3, \qquad |R_t^{(\chi)}| \le C\delta^3 \|g_t^{(\theta)}\|^3.$$

Moreover, the gradient moment bounds give

$$\mathbb{E}[\|g_t^{(\beta)}\|^3 \mid \mathcal{F}_t] + \mathbb{E}[\|g_t^{(\theta)}\|^3 \mid \mathcal{F}_t] \le C d^{3/2}.$$

Hence, for $j \in \{\rho, \chi\}$,

$$\mathbb{E}\left[\sum_{k=t_0}^{t_0+N-1} |R_k^{(j)}| \,\middle|\, \mathcal{F}_{t_0}\right] \le CN\delta^3 d^{3/2} \le C\delta^2 d^{3/2}.$$

Using the same high-moment/truncation estimate as in the weak-recovery discretization control, there is an event of conditional probability at least $1 - \nu_{\mathrm{in}}/8$ on which

$$\sup_{t_0 \le s \le u \le t_0+N} \left|\sum_{k=s}^{u-1} R_k^{(j)}\right| \le CL_{\nu,d}\delta^2 d^{3/2}, \qquad L_{\nu,d} := 1 + \log(d/\nu_{\mathrm{in}}).$$

Therefore, on an event E_{ov} of conditional probability at least $1 - \nu_{\mathrm{in}}/2$, for both $j \in \{\rho, \chi\}$,

$$\sup_{t_0 \le s \le u \le t_0+N} \left|\sum_{k=s}^{u-1} \Delta N_{k+1}^{(j)}\right| \le \Lambda_d,$$

where

$$\Lambda_d := C\left(\sqrt{\delta \log(1/\nu_{\mathrm{in}})} + L_{\nu,d}\delta^2 d^{3/2}\right).$$

Since $\delta = c_{\mathrm{loc}}/d$,

$$\Lambda_d \le C\sqrt{\frac{\log(1/\nu_{\mathrm{in}})}{d}} + CL_{\nu,d}\frac{1}{\sqrt{d}} = o_d(1)$$

for fixed or polynomially small ν_{in}.

Step 2: η_t-control on the entry window.

To control $\eta_t = \|\theta_t\|_\infty + \|\theta^*\|_\infty$, we use the same ℓ_∞ inequality as in the weak-recovery analysis:

$$\|\theta_t\|_\infty \le \|\theta_{t_0}\|_\infty + \delta \sum_{k=t_0}^{t-1} \|\bar{g}_k^{(\theta)}\|_\infty + \max_{i \le d}\left|\delta \sum_{k=t_0}^{t-1} \xi_{k,i}^{(\theta)}\right| + \frac{\delta^2}{2}\sum_{k=t_0}^{t-1} \|g_k^{(\theta)}\|^2\|\theta_k\|_\infty + \frac{\delta^3}{2}\sum_{k=t_0}^{t-1} \|g_k^{(\theta)}\|^3,$$

where

$$\bar{g}_k^{(\theta)} := \mathbb{E}[g_k^{(\theta)} \mid \mathcal{F}_k], \qquad \xi_k^{(\theta)} := g_k^{(\theta)} - \bar{g}_k^{(\theta)}.$$

Set

$$\Theta_u := \sup_{t_0 \le t \le u} \eta_t, \qquad t_0 \le u \le t_0 + N.$$

By the conditional coordinate Freedman bound, with conditional probability at least $1 - \nu_{\mathrm{in}}/8$,

$$\max_{i \le d} \sup_{t_0 \le u \le t_0+N} \left|\delta \sum_{k=t_0}^{u-1} \xi_{k,i}^{(\theta)}\right| \le C\sqrt{\delta L_{\nu,d}}.$$

The gradient moment estimates imply

$$\mathbb{E}\left[\delta^2 \sum_{k=t_0}^{t_0+N-1} \|g_k^{(\theta)}\|^2 \,\middle|\, \mathcal{F}_{t_0}\right] \le C\delta^2 Nd \le C\delta d = Cc_{\mathrm{loc}}.$$

Choosing $c_{\mathrm{loc}} > 0$ sufficiently small, the same concentration/Markov argument used in the weak phase gives, with conditional probability at least $1 - \nu_{\mathrm{in}}/8$,

$$\frac{\delta^2}{2} \sum_{k=t_0}^{t_0+N-1} \|g_k^{(\theta)}\|^2 \le \frac{1}{4}.$$

Similarly,

$$\mathbb{E}\left[\delta^3 \sum_{k=t_0}^{t_0+N-1} \|g_k^{(\theta)}\|^3 \,\middle|\, \mathcal{F}_{t_0}\right] \le C\delta^3 Nd^{3/2} \le C\delta^2 d^{3/2},$$

and hence, on an event of conditional probability at least $1 - \nu_{\mathrm{in}}/8$,

$$\delta^3 \sum_{k=t_0}^{t_0+N-1} \|g_k^{(\theta)}\|^3 \le C L_{\nu,d} \delta^2 d^{3/2}.$$

Let $E_{\eta,\mathrm{noise}}$ be the intersection of these three η-control events. On $E_{\eta,\mathrm{noise}}$, for every $u \in [t_0, t_0+N]$,

$$\Theta_u \le C\eta_{t_0} + C\delta \sum_{k=t_0}^{u-1} \|\bar{g}_k^{(\theta)}\|_\infty + C\sqrt{\delta L_{\nu,d}} + C L_{\nu,d}\delta^2 d^{3/2} + \frac{1}{4}\Theta_u.$$

Absorbing the last term,

$$\Theta_u \le C\eta_{t_0} + C\delta \sum_{k=t_0}^{u-1} \|\bar{g}_k^{(\theta)}\|_\infty + C\sqrt{\delta L_{\nu,d}} + C L_{\nu,d}\delta^2 d^{3/2}. \tag{193}$$

It remains to bound the mean-field term. The population ℓ_∞ gradient estimate gives, uniformly along the entry window,

$$\|\bar{g}_k^{(\theta)}\|_\infty \le C\|\theta^*\|_\infty + C\eta_k + C\eta_k^2.$$

Therefore, since $N\delta \le C_{\mathrm{in}} + 1$, (193) implies

$$\Theta_u \le A_d + C\delta \sum_{k=t_0}^{u-1} \Theta_k + C\delta \sum_{k=t_0}^{u-1} \Theta_k^2,$$

where

$$A_d := C\left(\eta_{t_0} + \|\theta^*\|_\infty + \sqrt{\delta L_{\nu,d}} + L_{\nu,d}\delta^2 d^{3/2}\right).$$

Now insert the scales. On the weak-recovery event,

$$\eta_{t_0} \le C_{\eta,\mathrm{esc}} \frac{\log d}{\sqrt{d}},$$

and by the teacher incoherence assumption,

$$\|\theta^*\|_\infty \le C_* \sqrt{\frac{\log d}{d}}.$$

Also $\delta = c_{\mathrm{loc}}/d$. Hence, for fixed or polynomially small ν_{in},

$$A_d \le C\left(\frac{\log d}{\sqrt{d}} + \sqrt{\frac{\log d}{d}} + \sqrt{\frac{L_{\nu,d}}{d}} + \frac{L_{\nu,d}}{\sqrt{d}}\right) = O\left(\frac{\log d}{\sqrt{d}}\right) = o_d(1).$$

We close the bootstrap. Define

$$\tau_\eta := \inf\{u \in [t_0, t_0+N] : \Theta_u \ge 1\} \wedge (t_0+N).$$

For $u \le \tau_\eta$, one has $\Theta_k^2 \le \Theta_k$, and therefore

$$\Theta_u \le A_d + C\delta \sum_{k=t_0}^{u-1} \Theta_k.$$

The discrete Gronwall inequality gives

$$\Theta_u \le A_d \exp(C(u-t_0)\delta) \le A_d \exp(C(C_{\mathrm{in}}+1)) \le C'_{\mathrm{in}} A_d, \qquad u \le \tau_\eta.$$

Since $A_d = o_d(1)$, for all sufficiently large d the right-hand side is strictly smaller than 1. Hence $\tau_\eta = t_0 + N$, and

$$\sup_{t_0 \le t \le t_0+N} \eta_t = \Theta_{t_0+N} \le \bar{\eta}_d,$$

where

$$\bar{\eta}_d := C'_{\text{in}} A_d \le C \left(\frac{\log d}{\sqrt{d}} + \sqrt{\frac{\log d}{d}} + \sqrt{\delta L_{\nu,d}} + L_{\nu,d} \delta^2 d^{3/2} \right).$$

This proves the claimed delocalization bound.

conclusion.

Define

$$E_{\text{in}} := E_{\text{ov}} \cap E_{\eta,\text{noise}}.$$

By the conditional probability bounds above and a union bound, after adjusting constants,

$$\mathbb{P}(E_{\text{in}} \mid \mathcal{F}_{t_0}) \ge 1 - \nu_{\text{in}}.$$

On E_{in}, both conclusions hold:

$$\sup_{t_0 \le t \le t_0+N} \eta_t \le \bar{\eta}_d$$

and, for $j \in \{\rho, \chi\}$,

$$\sup_{t_0 \le s \le u \le t_0+N} \left| \sum_{k=s}^{u-1} \Delta N_{k+1}^{(j)} \right| \le \Lambda_d.$$

Substituting back $t_0 = t_{\text{esc}}$, $N = N_{\text{in}}$, and $\delta = \delta_{\text{loc}}$ completes the proof. □

Lemma D.20 (One-step overlap recursion on the entry corridor)**.** *Assume Lemma C.3. After t_{esc}, run SGD with step size $\delta = \delta_{\text{loc}}$. On any time t in the entry window for which the one-step Taylor expansion is valid and*

$$\rho_t \ge 0, \qquad \chi_t \ge 0,$$

the overlap recursions can be written as

$$\rho_{t+1} = \rho_t + \delta F_t^{(\rho)} + \Delta N_{t+1}^{(\rho)}, \qquad \chi_{t+1} = \chi_t + \delta F_t^{(\chi)} + \Delta N_{t+1}^{(\chi)},$$

where

$$\Delta N_{t+1}^{(\rho)} = \delta M_t^{(\rho)} + R_t^{(\rho)}, \qquad \Delta N_{t+1}^{(\chi)} = \delta M_t^{(\chi)} + R_t^{(\chi)}.$$

Here $M_t^{(\rho)}, M_t^{(\chi)}$ are martingale increments and the remainders satisfy

$$|R_t^{(\rho)}| \le C\delta^3 \|g_t^{(\beta)}\|^3, \qquad |R_t^{(\chi)}| \le C\delta^3 \|g_t^{(\theta)}\|^3.$$

Moreover, the predictable drifts satisfy

$$F_t^{(\rho)} \ge a_\rho \chi_t (1 - \rho_t^2) - b_\rho \eta_t - C\delta d,$$

and

$$F_t^{(\chi)} \ge a_\chi \rho_t (1 - \chi_t^2) - b_\chi \eta_t - C\delta d.$$

Consequently, on the event E_{in} from Lemma D.19,

$$F_t^{(\rho)} \ge a_\rho \chi_t (1 - \rho_t^2) - B_d, \qquad F_t^{(\chi)} \ge a_\chi \rho_t (1 - \chi_t^2) - B_d,$$

where

$$B_d := C(\bar{\eta}_d + \delta d).$$

In particular, for $\delta = c_{\text{loc}}/d$,

$$B_d \le C \left(\frac{\log d}{\sqrt{d}} + c_{\text{loc}} \right)$$

for fixed or polynomially small ν_{in}.

Proof. We prove the ρ-recursion; the proof for χ is identical.

By the normalized spherical update,

$$\beta_{t+1} = \frac{\beta_t - \delta g_t^{(\beta)}}{\|\beta_t - \delta g_t^{(\beta)}\|}.$$

Since $g_t^{(\beta)} \perp \beta_t$, on the one-step Taylor event $\delta\|g_t^{(\beta)}\| \le 1/2$, we have

$$\beta_{t+1} = \beta_t - \delta g_t^{(\beta)} - \frac{\delta^2}{2}\|g_t^{(\beta)}\|^2\beta_t + O(\delta^3\|g_t^{(\beta)}\|^3).$$

Taking inner product with β^*, and using $\rho_t = \langle\beta_t, \beta^*\rangle$, gives

$$\rho_{t+1} = \rho_t - \delta\langle g_t^{(\beta)}, \beta^*\rangle - \frac{\delta^2}{2}\rho_t\|g_t^{(\beta)}\|^2 + R_t^{(\rho)},$$

where

$$|R_t^{(\rho)}| \le C\delta^3\|g_t^{(\beta)}\|^3.$$

Equivalently,

$$\rho_{t+1} = \rho_t - \delta\left(\langle g_t^{(\beta)}, \beta^*\rangle + \frac{\delta}{2}\rho_t\|g_t^{(\beta)}\|^2\right) + R_t^{(\rho)}.$$

Define

$$F_t^{(\rho)} := -\mathbb{E}\left[\langle g_t^{(\beta)}, \beta^*\rangle + \frac{\delta}{2}\rho_t\|g_t^{(\beta)}\|^2 \,\middle|\, \mathcal{F}_t\right],$$

and

$$M_t^{(\rho)} := -\left[\langle g_t^{(\beta)}, \beta^*\rangle + \frac{\delta}{2}\rho_t\|g_t^{(\beta)}\|^2 - \mathbb{E}\left[\langle g_t^{(\beta)}, \beta^*\rangle + \frac{\delta}{2}\rho_t\|g_t^{(\beta)}\|^2 \,\middle|\, \mathcal{F}_t\right]\right].$$

Then

$$\mathbb{E}[M_t^{(\rho)} \mid \mathcal{F}_t] = 0,$$

and

$$\rho_{t+1} = \rho_t + \delta F_t^{(\rho)} + \delta M_t^{(\rho)} + R_t^{(\rho)}.$$

Thus

$$\Delta N_{t+1}^{(\rho)} := \delta M_t^{(\rho)} + R_t^{(\rho)}.$$

It remains to lower-bound $F_t^{(\rho)}$. By definition,

$$F_t^{(\rho)} = -\mathbb{E}[\langle g_t^{(\beta)}, \beta^*\rangle \mid \mathcal{F}_t] - \frac{\delta}{2}\rho_t\mathbb{E}[\|g_t^{(\beta)}\|^2 \mid \mathcal{F}_t].$$

The first term is the population spherical-gradient drift:

$$-\mathbb{E}[\langle g_t^{(\beta)}, \beta^*\rangle \mid \mathcal{F}_t] = -\left\langle \nabla_{\mathbb{S},\beta} L(\theta_t, \beta_t), \beta^*\right\rangle.$$

Since $\rho_t, \chi_t \ge 0$, Lemma C.3 implies

$$-\left\langle \nabla_{\mathbb{S},\beta} L(\theta_t, \beta_t), \beta^*\right\rangle \ge a_\rho\chi_t(1-\rho_t^2) - b_\rho\eta_t.$$

For the normalization correction, use $|\rho_t| \le 1$ and the conditional second-moment gradient bound:

$$-\frac{\delta}{2}\rho_t\mathbb{E}[\|g_t^{(\beta)}\|^2 \mid \mathcal{F}_t] \ge -\frac{\delta}{2}\mathbb{E}[\|g_t^{(\beta)}\|^2 \mid \mathcal{F}_t] \ge -C\delta d.$$

Combining the two estimates gives

$$F_t^{(\rho)} \ge a_\rho\chi_t(1-\rho_t^2) - b_\rho\eta_t - C\delta d.$$

The χ-recursion follows from the same argument applied to

$$\theta_{t+1} = \frac{\theta_t - \delta g_t^{(\theta)}}{\|\theta_t - \delta g_t^{(\theta)}\|}.$$

Thus

$$\chi_{t+1} = \chi_t + \delta F_t^{(\chi)} + \Delta N_{t+1}^{(\chi)},$$

with

$$F_t^{(\chi)} \ge a_\chi\rho_t(1-\chi_t^2) - b_\chi\eta_t - C\delta d.$$

Finally, on E_{in},

$$\eta_t \le \bar{\eta}_d$$

throughout the entry window. Increasing C to absorb b_ρ, b_χ, the two drift bounds become

$$F_t^{(\rho)} \ge a_\rho \chi_t (1-\rho_t^2) - C(\bar{\eta}_d + \delta d),$$

and

$$F_t^{(\chi)} \ge a_\chi \rho_t (1-\chi_t^2) - C(\bar{\eta}_d + \delta d).$$

This proves the stated bounds with

$$B_d := C(\bar{\eta}_d + \delta d).$$

Since Lemma D.19 gives $\bar{\eta}_d = O(\log d/\sqrt{d})$ for fixed or polynomially small ν_{in}, and since $\delta = c_{\mathrm{loc}}/d$, we obtain

$$B_d \le C\left(\frac{\log d}{\sqrt{d}} + c_{\mathrm{loc}}\right).$$

The proof is complete. □

Proposition D.21 (Discrete entry into the local product basin)**.** *Let r_{loc} be the radius from Lemma C.14. After possibly shrinking r_{loc}, assume $0 < r_{\mathrm{loc}} < \pi$. Define*

$$s_b := \cos(r_{\mathrm{loc}}/8), \qquad s_+ := \frac{1+s_b}{2}.$$

Apply Lemma D.19 with a constant C_{in} at least as large as the fixed constant C_{ent} constructed in the proof below.

Then there exist constants $c_{\mathrm{loc},0} > 0$, $d_0 < \infty$, and $C_{\mathrm{ent}} < \infty$, depending only on r_{loc}, c_{esc}, and the model constants, such that the following holds. If

$$\delta_{\mathrm{loc}} = \frac{c_{\mathrm{loc}}}{d}, \qquad 0 < c_{\mathrm{loc}} \le c_{\mathrm{loc},0}, \qquad d \ge d_0,$$

then on the event E_{in} from Lemma D.19, there exists a time $t_1 \le t_{\mathrm{esc}} + C_{\mathrm{ent}}\delta_{\mathrm{loc}}^{-1}$ such that

$$\rho_{t_1} \ge s_b, \qquad \chi_{t_1} \ge s_b.$$

Consequently,

$$Q_{t_1} = \frac{1}{2}\operatorname{dist}_{\mathbb{S}}(\theta_{t_1}, \theta^*)^2 + \frac{1}{2}\operatorname{dist}_{\mathbb{S}}(\beta_{t_1}, \beta^*)^2 \le \frac{r_{\mathrm{loc}}^2}{64} \le \frac{r_{\mathrm{loc}}^2}{32}.$$

Proof. Write

$$t_0 := t_{\mathrm{esc}}, \qquad \delta := \delta_{\mathrm{loc}}, \qquad c_- := \frac{c_{\mathrm{esc}}}{2}.$$

We work deterministically on the event E_{in}, together with the weak recovery event at t_{esc}. Thus

$$\rho_{t_0} \ge c_{\mathrm{esc}}, \qquad \chi_{t_0} \ge c_{\mathrm{esc}},$$

and, throughout the entry window,

$$\eta_t \le \bar{\eta}_d, \qquad \sup_{t_0 \le s \le u \le t_0 + C_{\mathrm{in}}\delta^{-1}} \left| \sum_{k=s}^{u-1} \Delta N_{k+1}^{(j)} \right| \le \Lambda_d, \qquad j \in \{\rho, \chi\}.$$

By Lemma D.20, on E_{in},

$$F_t^{(\rho)} \ge a_\rho \chi_t (1-\rho_t^2) - B_d, \qquad F_t^{(\chi)} \ge a_\chi \rho_t (1-\chi_t^2) - B_d,$$

where

$$B_d := C(\bar{\eta}_d + \delta d).$$

Since

$$\bar{\eta}_d = O\left(\frac{\log d}{\sqrt{d}}\right), \qquad \delta d = c_{\mathrm{loc}},$$

we may choose $c_{\mathrm{loc}} > 0$ sufficiently small and then d sufficiently large so that all smallness conditions below hold.

Define

$$m_+ := 1 - s_+^2 > 0,$$

and set

$$\gamma_\rho := \frac{1}{2} a_\rho c_- m_+, \qquad \gamma_\chi := \frac{1}{2} a_\chi c_- m_+.$$

These are fixed positive constants independent of d. Also define

$$\alpha_* := \frac{1}{4} \min\{c_{\mathrm{esc}}, s_+ - s_b, \gamma_\rho, \gamma_\chi\} > 0.$$

Let

$$T_\rho := \frac{(s_+ - c_{\mathrm{esc}})_+ + \alpha_*}{\gamma_\rho}, \qquad T_\chi := \frac{(s_+ - c_{\mathrm{esc}})_+ + \alpha_*}{\gamma_\chi},$$

and choose

$$C_{\mathrm{ent}} := 2 + \max\{T_\rho, T_\chi\}.$$

We choose c_{loc} small enough and then d large enough so that

$$B_d \le \min\{\gamma_\rho, \gamma_\chi\}, \tag{194}$$

$$\Lambda_d \le \alpha_*, \tag{195}$$

$$C_{\mathrm{ent}} B_d + \Lambda_d \le \alpha_*. \tag{196}$$

These conditions are consistent because $B_d \le C(\log d/\sqrt{d} + c_{\mathrm{loc}})$ and $\Lambda_d = o_d(1)$.

Let

$$N_{\mathrm{ent}} := \left\lceil \max\{T_\rho, T_\chi\} \delta^{-1} \right\rceil.$$

Since $\delta \le 1$, we have

$$N_{\mathrm{ent}} \delta \le C_{\mathrm{ent}}.$$

By the assumption $C_{\mathrm{in}} \ge C_{\mathrm{ent}}$, all estimates from E_{in} hold on $[t_0, t_0 + N_{\mathrm{ent}}]$.

Step 1: the overlaps stay above the lower barrier.

We claim that

$$\rho_t \ge c_-, \qquad \chi_t \ge c_-, \qquad t_0 \le t \le t_0 + N_{\mathrm{ent}}.$$

Suppose not, and define the first lower-barrier exit time

$$\tau_- := \inf\{t \in [t_0, t_0 + N_{\mathrm{ent}}] : \min\{\rho_t, \chi_t\} < c_-\}.$$

Since $\rho_{t_0}, \chi_{t_0} \ge c_{\mathrm{esc}} > c_-$, one has $\tau_- > t_0$ if $\tau_- < \infty$.

For every $k < \tau_-$, both overlaps are at least c_-, and hence nonnegative. Therefore the crude drift bounds

$$F_k^{(\rho)} \ge -B_d, \qquad F_k^{(\chi)} \ge -B_d$$

hold. Summing the ρ-recursion from t_0 to τ_-, and using the cumulative noise bound from E_{in}, gives

$$\rho_{\tau_-} \ge \rho_{t_0} - (\tau_- - t_0)\delta B_d - \Lambda_d.$$

Since $(\tau_- - t_0)\delta \le C_{\mathrm{ent}}$, (196) implies

$$\rho_{\tau_-} \ge c_{\mathrm{esc}} - C_{\mathrm{ent}} B_d - \Lambda_d \ge c_{\mathrm{esc}} - \alpha_* \ge \frac{3}{4} c_{\mathrm{esc}} > c_-.$$

The same argument gives $\chi_{\tau_-} > c_-$. This contradicts the definition of τ_-. Hence no such exit occurs.

Step 2: each overlap reaches the buffer level s_+.

First consider ρ_t. If $\rho_{t_0} \ge s_+$, set $\tau_\rho := t_0$. Otherwise, suppose for contradiction that

$$\rho_t < s_+ \qquad \text{for all } t_0 \le t \le t_0 + \left\lceil T_\rho \delta^{-1} \right\rceil.$$

By Step 1, $\chi_t \ge c_-$ on this interval. Hence, for every such k,

$$F_k^{(\rho)} \ge a_\rho c_- (1 - \rho_k^2) - B_d.$$

Since $\rho_k < s_+$, we have $1 - \rho_k^2 \geq 1 - s_+^2 = m_+$, and therefore

$$F_k^{(\rho)} \geq a_\rho c_- m_+ - B_d = 2\gamma_\rho - B_d \geq \gamma_\rho$$

by (194). Summing the recursion gives

$$\rho_{t_0+\lceil T_\rho \delta^{-1} \rceil} \geq \rho_{t_0} + \gamma_\rho T_\rho - \Lambda_d.$$

Using $\rho_{t_0} \geq c_{\mathrm{esc}}$, the definition of T_ρ, and (195),

$$\rho_{t_0+\lceil T_\rho \delta^{-1} \rceil} \geq c_{\mathrm{esc}} + (s_+ - c_{\mathrm{esc}})_+ + \alpha_* - \Lambda_d \geq s_+.$$

This contradicts the assumption that $\rho_t < s_+$ throughout the interval. Thus there exists

$$\tau_\rho \leq t_0 + \lceil T_\rho \delta^{-1} \rceil$$

such that

$$\rho_{\tau_\rho} \geq s_+.$$

The proof for χ_t is identical. If $\chi_{t_0} \geq s_+$, set $\tau_\chi := t_0$. Otherwise, using Step 1 and the lower bound for $F_t^{(\chi)}$, one obtains a time

$$\tau_\chi \leq t_0 + \lceil T_\chi \delta^{-1} \rceil$$

such that

$$\chi_{\tau_\chi} \geq s_+.$$

Step 3: retain both overlaps above s_b.

Define

$$t_1 := \max\{\tau_\rho, \tau_\chi\}.$$

Then

$$t_1 \leq t_0 + N_{\mathrm{ent}} \leq t_{\mathrm{esc}} + C_{\mathrm{ent}} \delta^{-1}.$$

We show that both overlaps are at least s_b at time t_1. Since $\rho_{\tau_\rho} \geq s_+$, and since Step 1 gives the lower barrier on the whole window, the crude drift bound $F_k^{(\rho)} \geq -B_d$ holds on $[\tau_\rho, t_1]$. Therefore

$$\rho_{t_1} \geq \rho_{\tau_\rho} - (t_1 - \tau_\rho)\delta B_d - \Lambda_d.$$

As $(t_1 - \tau_\rho)\delta \leq C_{\mathrm{ent}}$, (196) gives

$$\rho_{t_1} \geq s_+ - C_{\mathrm{ent}} B_d - \Lambda_d \geq s_+ - \alpha_*.$$

By the definition of α_*,

$$\alpha_* \leq \frac{1}{4}(s_+ - s_b),$$

so

$$\rho_{t_1} \geq s_b.$$

The same argument gives

$$\chi_{t_1} \geq s_b.$$

Step 4: convert overlap entry into product-distance entry.

Since

$$\rho_{t_1} = \langle \beta_{t_1}, \beta^* \rangle \geq s_b = \cos(r_{\mathrm{loc}}/8),$$

and $0 < r_{\mathrm{loc}} < \pi$,

$$\mathrm{dist}_{\mathbb{S}}(\beta_{t_1}, \beta^*) = \arccos(\rho_{t_1}) \leq \frac{r_{\mathrm{loc}}}{8}.$$

Similarly,

$$\mathrm{dist}_{\mathbb{S}}(\theta_{t_1}, \theta^*) \leq \frac{r_{\mathrm{loc}}}{8}.$$

Therefore

$$Q_{t_1} = \frac{1}{2}\,\mathrm{dist}_{\mathbb{S}}(\theta_{t_1}, \theta^*)^2 + \frac{1}{2}\,\mathrm{dist}_{\mathbb{S}}(\beta_{t_1}, \beta^*)^2 \leq \frac{1}{2}\left(\frac{r_{\mathrm{loc}}}{8}\right)^2 + \frac{1}{2}\left(\frac{r_{\mathrm{loc}}}{8}\right)^2 = \frac{r_{\mathrm{loc}}^2}{64}.$$

This proves the proposition. □

Deriving Q_t-iteration and its controlling bounds.

Lemma D.22 (One-step local distance recursion)**.** *Let*

$$y_t = (\theta_t, \beta_t), \qquad y^* = (\theta^*, \beta^*), \qquad g_t = (g_t^{(\theta)}, g_t^{(\beta)}),$$

and define

$$Q_t := \frac{1}{2}\operatorname{dist}_{\mathcal{M}}(y_t, y^*)^2 = \frac{1}{2}\operatorname{dist}_{\mathbb{S}}(\theta_t, \theta^*)^2 + \frac{1}{2}\operatorname{dist}_{\mathbb{S}}(\beta_t, \beta^*)^2.$$

Assume $r_{\mathrm{loc}} < \pi/4$. There exists a constant $C_{\mathrm{ret}} < \infty$, depending only on r_{loc}, such that whenever $y_t \in \mathcal{U}_{r_{\mathrm{loc}}}$, the normalized SGD step satisfies

$$Q_{t+1} \le Q_t + \delta\langle g_t, \mathcal{L}_t\rangle + C_{\mathrm{ret}}\delta^2\|g_t\|^2,$$

where

$$\mathcal{L}_t := \left(\operatorname{Log}_{\theta_t}(\theta^*), \operatorname{Log}_{\beta_t}(\beta^*)\right).$$

Consequently, if

$$\bar{g}_t := \mathbb{E}[g_t \mid \mathcal{F}_t], \qquad \Xi_t := \langle g_t - \bar{g}_t, \mathcal{L}_t\rangle, \qquad G_t := \|g_t\|^2,$$

then

$$Q_{t+1} \le Q_t + \delta\langle \bar{g}_t, \mathcal{L}_t\rangle + \delta\Xi_t + C_{\mathrm{ret}}\delta^2 G_t. \tag{197}$$

Proof. We use the product metric on

$$\mathcal{M} = \mathbb{S}^{d-1} \times \mathbb{S}^{d-1},$$

namely

$$\langle (a_\theta, a_\beta), (b_\theta, b_\beta)\rangle := \langle a_\theta, b_\theta\rangle + \langle a_\beta, b_\beta\rangle,$$

and

$$\|(a_\theta, a_\beta)\|^2 := \|a_\theta\|^2 + \|a_\beta\|^2.$$

Fix t such that $y_t \in \mathcal{U}_{r_{\mathrm{loc}}}$. We prove the one-step estimate on one sphere and then add the two block estimates.

Let $u \in \mathbb{S}^{d-1}$, let $v \in \mathbb{S}^{d-1}$, and let $h \in T_u\mathbb{S}^{d-1}$. Consider the normalized spherical step

$$\widehat{u} := \frac{u - \delta h}{\|u - \delta h\|}.$$

Since $h \perp u$,

$$\|u - \delta h\|^2 = 1 + \delta^2\|h\|^2.$$

Define

$$\tau(h) := \begin{cases} \dfrac{\arctan(\delta\|h\|)}{\|h\|}, & h \neq 0, \\ 0, & h = 0. \end{cases}$$

Then

$$\widehat{u} = \operatorname{Exp}_u(-\tau(h)h).$$

Indeed, when $h \neq 0$, the normalized point lies on the great circle through u in direction $-h/\|h\|$, with angular displacement $\arctan(\delta\|h\|)$. The case $h = 0$ is immediate.

Now set

$$q_v(u) := \frac{1}{2}\operatorname{dist}_{\mathbb{S}}(u, v)^2.$$

Since u lies in the local cap around v, and since the normalized step has spherical length at most $\pi/2$, after shrinking r_{loc} if necessary the geodesic segment used above remains in a compact subset away from the cut locus of v. Therefore q_v has uniformly bounded Hessian on this region:

$$\|\operatorname{Hess}_{\mathbb{S}} q_v\| \le C_{\mathrm{sph}}.$$

Taylor expansion along

$$s \mapsto \operatorname{Exp}_u(-s\tau(h)h)$$

gives

$$q_v(\widehat{u}) \le q_v(u) + \langle \nabla_{\mathbb{S}} q_v(u), -\tau(h)h\rangle + \frac{C_{\mathrm{sph}}}{2}\tau(h)^2\|h\|^2.$$

Since

$$\nabla_{\mathbb{S}} q_v(u) = -\operatorname{Log}_u(v),$$

we get

$$q_v(\widehat{u}) \le q_v(u) + \tau(h)\langle h, \operatorname{Log}_u(v)\rangle + \frac{C_{\mathrm{sph}}}{2}\tau(h)^2\|h\|^2.$$

We compare $\tau(h)$ with δ. Since

$$0 \le \tau(h) \le \delta,$$

and, writing $x = \delta\|h\|$,

$$0 \le \delta\|h\| - \arctan(\delta\|h\|) = x - \arctan x \le Cx^2,$$

we have

$$|\delta - \tau(h)|\,\|h\| \le C\delta^2\|h\|^2.$$

Also u is within distance r_{loc} of v, so

$$\|\operatorname{Log}_u(v)\| = \operatorname{dist}_{\mathbb{S}}(u,v) \le r_{\mathrm{loc}}.$$

Therefore

$$\tau(h)\langle h, \operatorname{Log}_u(v)\rangle \le \delta\langle h, \operatorname{Log}_u(v)\rangle + Cr_{\mathrm{loc}}\delta^2\|h\|^2.$$

Moreover,

$$\tau(h)^2\|h\|^2 \le \delta^2\|h\|^2.$$

Thus

$$q_v(\widehat{u}) \le q_v(u) + \delta\langle h, \operatorname{Log}_u(v)\rangle + C^{(1)}_{\mathrm{ret}}\delta^2\|h\|^2,$$

where $C^{(1)}_{\mathrm{ret}}$ depends only on r_{loc} and C_{sph}.

Apply this estimate first with

$$u = \theta_t, \qquad v = \theta^*, \qquad h = g_t^{(\theta)},$$

and then with

$$u = \beta_t, \qquad v = \beta^*, \qquad h = g_t^{(\beta)}.$$

The normalized SGD updates are

$$\theta_{t+1} = \frac{\theta_t - \delta g_t^{(\theta)}}{\|\theta_t - \delta g_t^{(\theta)}\|}, \qquad \beta_{t+1} = \frac{\beta_t - \delta g_t^{(\beta)}}{\|\beta_t - \delta g_t^{(\beta)}\|}.$$

Hence

$$\frac{1}{2}\operatorname{dist}_{\mathbb{S}}(\theta_{t+1},\theta^*)^2 \le \frac{1}{2}\operatorname{dist}_{\mathbb{S}}(\theta_t,\theta^*)^2 + \delta\left\langle g_t^{(\theta)}, \operatorname{Log}_{\theta_t}(\theta^*)\right\rangle + C^{(1)}_{\mathrm{ret}}\delta^2\|g_t^{(\theta)}\|^2,$$

and

$$\frac{1}{2}\operatorname{dist}_{\mathbb{S}}(\beta_{t+1},\beta^*)^2 \le \frac{1}{2}\operatorname{dist}_{\mathbb{S}}(\beta_t,\beta^*)^2 + \delta\left\langle g_t^{(\beta)}, \operatorname{Log}_{\beta_t}(\beta^*)\right\rangle + C^{(1)}_{\mathrm{ret}}\delta^2\|g_t^{(\beta)}\|^2.$$

Adding the two estimates gives

$$Q_{t+1} \le Q_t + \delta\langle g_t, \mathcal{L}_t\rangle + C_{\mathrm{ret}}\delta^2\|g_t\|^2,$$

after increasing the constant C_{ret}.

Finally,

$$g_t = \bar{g}_t + (g_t - \bar{g}_t),$$

so

$$\langle g_t, \mathcal{L}_t\rangle = \langle \bar{g}_t, \mathcal{L}_t\rangle + \langle g_t - \bar{g}_t, \mathcal{L}_t\rangle = \langle \bar{g}_t, \mathcal{L}_t\rangle + \Xi_t.$$

With $G_t = \|g_t\|^2$, this yields

$$Q_{t+1} \le Q_t + \delta\langle \bar{g}_t, \mathcal{L}_t\rangle + \delta\Xi_t + C_{\mathrm{ret}}\delta^2 G_t,$$

which proves (197). □

Lemma D.23 (Sharp local self-quenching)**.** *Inside $\mathcal{U}_{r_{\mathrm{loc}}}$, there exists $C_{\mathrm{sq}} > 0$, independent of d, such that*

$$\mathbb{E}[G_t \mid \mathcal{F}_t] \le C_{\mathrm{sq}} d\, Q_t,$$
$$\mathbb{E}[\Xi_t^2 \mid \mathcal{F}_t] \le C_{\mathrm{sq}} Q_t^2,$$

and

$$\mathbb{E}\Big[\big(G_t - \mathbb{E}[G_t \mid \mathcal{F}_t]\big)^2 \mid \mathcal{F}_t\Big] \le C_{\mathrm{sq}} d^2 Q_t^2.$$

Proof. Condition on $\mathcal{F}_t$, so that (θ_t, β_t) are deterministic. To lighten notation, write

$$\theta := \theta_t, \qquad \beta := \beta_t,$$

and set

$$d_\theta := \mathrm{dist}_{\mathbb{S}}(\theta, \theta^*), \qquad d_\beta := \mathrm{dist}_{\mathbb{S}}(\beta, \beta^*).$$

Since

$$Q_t = \frac{1}{2}\big(d_\theta^2 + d_\beta^2\big),$$

we have

$$d_\theta^2 + d_\beta^2 = 2Q_t.$$

Let

$$a := \mathrm{Log}_\theta(\theta^*) \in T_\theta \mathbb{S}^{d-1}, \qquad b := \mathrm{Log}_\beta(\beta^*) \in T_\beta \mathbb{S}^{d-1},$$

so that

$$\|a\| = d_\theta, \qquad \|b\| = d_\beta.$$

Write the one-sample prediction as

$$f_{\beta,\theta}(x) = \psi(s), \qquad s := z_\beta(x) \cdot \theta = \sum_{i=1}^d \theta_i\, \phi(u_i), \qquad u_i := x^{(i)} \cdot \beta,$$

and the teacher prediction as

$$f_{\beta^*,\theta^*}(x) = \psi(s^*), \qquad s^* := z_{\beta^*}(x) \cdot \theta^* = \sum_{i=1}^d \theta_i^*\, \phi(u_i^*), \qquad u_i^* := x^{(i)} \cdot \beta^*.$$

Define the residual

$$R := f_{\beta,\theta}(x) - f_{\beta^*,\theta^*}(x) = \psi(s) - \psi(s^*).$$

The spherical one-sample gradients are

$$g_t^{(\theta)} = 2R\, \psi'(s)\, P_\theta z_\beta(x),$$

and

$$g_t^{(\beta)} = 2R\, \psi'(s)\, P_\beta\Big(\sum_{i=1}^d \theta_i \phi'(u_i) x^{(i)}\Big).$$

Set

$$g^{(\theta)} := g_t^{(\theta)}, \qquad g^{(\beta)} := g_t^{(\beta)}, \qquad g_t = (g^{(\theta)}, g^{(\beta)}),$$

and

$$G_t := \|g_t\|^2 = \|g^{(\theta)}\|^2 + \|g^{(\beta)}\|^2.$$

Also recall

$$\Xi_t = \langle g_t - \bar{g}_t, (a, b)\rangle, \qquad \bar{g}_t := \mathbb{E}[g_t \mid \mathcal{F}_t].$$

All moment estimates below follow from Assumption (A3) and Remark B.1.

residual self-quenching.

We claim that for $m = 2, 4, 8$,

$$\mathbb{E}[|R|^m \mid \mathcal{F}_t] \le C_m Q_t^{m/2}.$$

Since ψ' is bounded,

$$|R| \le \|\psi'\|_\infty \, |s - s^*|.$$

Decompose

$$s - s^* = A + B,$$

where

$$A := \sum_{i=1}^{d} (\theta_i - \theta_i^*)\phi(u_i),$$

and

$$B := \sum_{i=1}^{d} \theta_i^* \big(\phi(u_i) - \phi(u_i^*)\big).$$

For A: the variables $\phi(u_i)$ are iid, centered, and have finite moments up to order 8. Hence by Rosenthal's inequality, for $m \in \{2, 4, 8\}$,

$$\|A\|_{L^m(\mathbb{P}(\cdot|\mathcal{F}_t))} \le C_m \|\theta - \theta^*\|_2.$$

Since Euclidean distance is bounded by geodesic distance on the sphere,

$$\|\theta - \theta^*\|_2 \le d_\theta,$$

we get

$$\|A\|_{L^m(\mathbb{P}(\cdot|\mathcal{F}_t))} \le C_m d_\theta.$$

For B: the blocks $x^{(i)}$ are independent, so the summands

$$Y_i := \theta_i^* \big(\phi(u_i) - \phi(u_i^*)\big)$$

are independent across i. They are centered because $u_i \sim N(0,1)$ and $u_i^* \sim N(0,1)$, hence

$$\mathbb{E}[\phi(u_i) - \phi(u_i^*)] = 0.$$

Using the Lipschitzness of ϕ,

$$|Y_i| \le |\theta_i^*| \, \mathrm{Lip}(\phi) \, |u_i - u_i^*| = |\theta_i^*| \, \mathrm{Lip}(\phi) \, |x^{(i)} \cdot (\beta - \beta^*)|.$$

Thus for $m \in \{2, 4, 8\}$,

$$\sum_{i=1}^{d} \mathbb{E}[|Y_i|^m \mid \mathcal{F}_t] \le C_m \|\beta - \beta^*\|_2^m \sum_{i=1}^{d} |\theta_i^*|^m \le C_m d_\beta^m,$$

since $\sum_i |\theta_i^*|^m \le \|\theta^*\|_2^m = 1$ for $m \ge 2$. Applying Rosenthal again,

$$\|B\|_{L^m(\mathbb{P}(\cdot|\mathcal{F}_t))} \le C_m d_\beta, \qquad m = 2, 4, 8.$$

Combining the bounds for A and B,

$$\|s - s^*\|_{L^m(\mathbb{P}(\cdot|\mathcal{F}_t))} \le C_m (d_\theta + d_\beta) \le C_m \sqrt{d_\theta^2 + d_\beta^2} = C_m \sqrt{2 Q_t}.$$

Hence

$$\|R\|_{L^m(\mathbb{P}(\cdot|\mathcal{F}_t))} \le C_m \sqrt{Q_t}, \qquad m = 2, 4, 8,$$

which is equivalent to

$$\mathbb{E}[|R|^m \mid \mathcal{F}_t] \le C_m Q_t^{m/2}, \qquad m = 2, 4, 8.$$

feature-side moment bounds.

Define

$$z := z_\beta(x) = (\phi(u_1), \ldots, \phi(u_d)), \qquad h := P_\beta \sum_{i=1}^{d} \theta_i \phi'(u_i) x^{(i)},$$

so that

$$g^{(\theta)} = A \cdot P_\theta z, \qquad g^{(\beta)} = A \cdot h, \qquad A := 2\psi'(s) R,$$

and $G_t \le CR^2W$ where $W := \|z\|^2 + \|h\|^2$. The same Gaussian moment computations used in the proof of Lemma E.7 give, for $k = 1, 2, 4$,

$$\mathbb{E}[\|z\|^{2k} \mid \mathcal{F}_t] \le Cd^k, \qquad \mathbb{E}[\|h\|^{2k} \mid \mathcal{F}_t] \le Cd^k.$$

Hence $\mathbb{E}[W^k \mid \mathcal{F}_t] \le Cd^k$ for $k = 1, 2, 4$.

bound on $\mathbb{E}[G_t \mid \mathcal{F}_t]$.

Since $G_t \le CR^2W$, Cauchy–Schwarz gives

$$\mathbb{E}[G_t \mid \mathcal{F}_t] \le C \left(\mathbb{E}[R^4 \mid \mathcal{F}_t]\right)^{1/2} \left(\mathbb{E}[W^2 \mid \mathcal{F}_t]\right)^{1/2}.$$

Using Step 1 with $m = 4$ and Step 2,

$$\mathbb{E}[G_t \mid \mathcal{F}_t] \le C\,(Q_t^2)^{1/2}(d^2)^{1/2} = Cd\,Q_t.$$

bound on $\mathbb{E}[\Xi_t^2 \mid \mathcal{F}_t]$.

Since conditional variance is bounded by the second moment,

$$\mathbb{E}[\Xi_t^2 \mid \mathcal{F}_t] \le \mathbb{E}\Big[\big(\langle g^{(\theta)}, a\rangle + \langle g^{(\beta)}, b\rangle\big)^2 \Big| \mathcal{F}_t\Big].$$

Hence

$$\mathbb{E}[\Xi_t^2 \mid \mathcal{F}_t] \le 2\,\mathbb{E}[\langle g^{(\theta)}, a\rangle^2 \mid \mathcal{F}_t] + 2\,\mathbb{E}[\langle g^{(\beta)}, b\rangle^2 \mid \mathcal{F}_t].$$

For the θ-block, $a \in T_\theta \mathbb{S}^{d-1}$, so

$$\langle P_\theta z, a\rangle = \langle z, a\rangle.$$

Thus

$$\langle g^{(\theta)}, a\rangle = 2R\,\psi'(s)\,\langle z, a\rangle.$$

Therefore

$$\mathbb{E}[\langle g^{(\theta)}, a\rangle^2 \mid \mathcal{F}_t] \le C \left(\mathbb{E}[R^4 \mid \mathcal{F}_t]\right)^{1/2} \left(\mathbb{E}[\langle z, a\rangle^4 \mid \mathcal{F}_t]\right)^{1/2}.$$

Now $\langle z, a\rangle = \sum_i a_i \phi(u_i)$ is a linear statistic of iid centered variables with finite fourth moment, so Rosenthal yields

$$\mathbb{E}[\langle z, a\rangle^4 \mid \mathcal{F}_t] \le C\|a\|^4.$$

Hence

$$\mathbb{E}[\langle g^{(\theta)}, a\rangle^2 \mid \mathcal{F}_t] \le C\,Q_t\,\|a\|^2 \le C\,Q_t^2.$$

For the β-block,

$$\langle g^{(\beta)}, b\rangle = 2R\,\psi'(s)\,\langle h, b\rangle.$$

Conditionally on $(u_1, \dots, u_d)$, the scalar $\langle h, b\rangle$ is Gaussian with variance

$$\|b\|^2\sigma^2, \qquad \sigma^2 = \sum_{i=1}^{d} \theta_i^2 \phi'(u_i)^2.$$

Therefore

$$\mathbb{E}[\langle h, b\rangle^4 \mid u, \mathcal{F}_t] \le C\|b\|^4\sigma^4.$$

Averaging and using $\mathbb{E}[\sigma^4 \mid \mathcal{F}_t] \le C$,

$$\mathbb{E}[\langle h, b\rangle^4 \mid \mathcal{F}_t] \le C\|b\|^4.$$

Thus

$$\mathbb{E}[\langle g^{(\beta)}, b\rangle^2 \mid \mathcal{F}_t] \le C \left(\mathbb{E}[R^4 \mid \mathcal{F}_t]\right)^{1/2} \left(\mathbb{E}[\langle h, b\rangle^4 \mid \mathcal{F}_t]\right)^{1/2} \le C\,Q_t\,\|b\|^2 \le C\,Q_t^2.$$

Combining the two block estimates,

$$\mathbb{E}[\Xi_t^2 \mid \mathcal{F}_t] \le CQ_t^2.$$

centered normalization fluctuation.

Since

$$\mathbb{E}\Big[(G_t - \mathbb{E}[G_t \mid \mathcal{F}_t])^2 \Big| \mathcal{F}_t\Big] \le \mathbb{E}[G_t^2 \mid \mathcal{F}_t],$$

it suffices to bound $\mathbb{E}[G_t^2 \mid \mathcal{F}_t]$. Using $G_t \le CR^2W$,

$$G_t^2 \le CR^4W^2.$$

Hence, by Cauchy–Schwarz,

$$\mathbb{E}[G_t^2 \mid \mathcal{F}_t] \le C\left(\mathbb{E}[R^8 \mid \mathcal{F}_t]\right)^{1/2}\left(\mathbb{E}[W^4 \mid \mathcal{F}_t]\right)^{1/2}.$$

Applying Step 1 with $m = 8$ and Step 2,

$$\mathbb{E}[G_t^2 \mid \mathcal{F}_t] \le C\,(Q_t^4)^{1/2}(d^4)^{1/2} = Cd^2Q_t^2.$$

Therefore

$$\mathbb{E}\Big[(G_t - \mathbb{E}[G_t \mid \mathcal{F}_t])^2 \Big| \mathcal{F}_t\Big] \le Cd^2Q_t^2.$$

This proves the lemma. □

D.4 Strong Recovery Dynamics

With all the controlling lemmas needed, we now move to prove the strong recovery dynamics.

Proposition D.24 (One local SGD epoch contracts)**.** *Assume Lemma C.15, Lemma D.22, and Lemma D.23. Let the local step size satisfy*

$$\delta = \delta_{\mathrm{loc}} = \frac{c_{\mathrm{loc}}}{d}, \qquad C_{\mathrm{ret}}C_{\mathrm{sq}}\delta d \le \mu_{\mathrm{loc}}.$$

Let s be a stopping time, and suppose that, conditionally on $\mathcal{F}_s$,

$$Q_s \le B, \qquad 0 \le B \le \frac{r_{\mathrm{loc}}^2}{32}.$$

Define $\tau_B := \inf\{t \ge s : Q_t > 2B\}$. Then there exist constants $C_{\mathrm{ep}} < \infty$, $C < \infty$, and $d_0 < \infty$, depending only on the model constants, such that for all $d \ge d_0$, with

$$L_{\mathrm{ep}} := \left\lceil \frac{C_{\mathrm{ep}}}{\delta} \right\rceil,$$

one has

$$\mathbb{P}\left(Q_{s+L_{\mathrm{ep}}} \le \frac{B}{2} \text{ and } \tau_B > s + L_{\mathrm{ep}} \;\middle|\; \mathcal{F}_s\right) \ge 1 - C\delta.$$

Equivalently, with probability at least $1 - C\delta$,

$$Q_{s+L_{\mathrm{ep}}} \le \frac{B}{2} \qquad \text{and} \qquad Q_t \le 2B \quad \text{for every } s \le t \le s + L_{\mathrm{ep}}.$$

Proof. If $B = 0$, then $Q_s = 0$, so $y_s = y^*$. By Lemma D.23,

$$\mathbb{E}[G_s \mid \mathcal{F}_s] \le C_{\mathrm{sq}}dQ_s = 0,$$

hence $G_s = 0$ almost surely conditionally on $\mathcal{F}_s$. Thus the SGD step is trivial at y^*, and the same argument iterates. The conclusion is therefore immediate. Hence assume $B > 0$.

Throughout the proof we condition on $\mathcal{F}_s$. Since the samples after time s are fresh, all martingale and conditional moment estimates hold with the same constants.

Step 1: stopped multiplicative recursion.

For every $t < \tau_B$, one has $Q_t \le 2B$. Hence

$$\mathrm{dist}_{\mathcal{M}}(y_t, y^*)^2 = 2Q_t \le 4B \le \frac{r_{\mathrm{loc}}^2}{8} < r_{\mathrm{loc}}^2.$$

Thus $y_t \in \mathcal{U}_{r_{\mathrm{loc}}}$, and the local one-step recursion is valid. By Lemma D.22,

$$Q_{t+1} \le Q_t + \delta\langle \bar{g}_t, \mathcal{L}_t\rangle + \delta\Xi_t + C_{\mathrm{ret}}\delta^2 G_t.$$

Since $\bar{g}_t = \nabla_{\mathcal{M}} L(y_t)$, Lemma C.15 gives

$$\langle \bar{g}_t, \mathcal{L}_t\rangle = \langle \nabla_{\mathcal{M}} L(y_t), \mathcal{L}_t\rangle \le -2\mu_{\mathrm{loc}} Q_t.$$

Define

$$\zeta_t := G_t - \mathbb{E}[G_t \mid \mathcal{F}_t].$$

Then, for $t < \tau_B$,

$$Q_{t+1} \le Q_t - 2\mu_{\mathrm{loc}}\delta Q_t + \delta\Xi_t + C_{\mathrm{ret}}\delta^2\mathbb{E}[G_t \mid \mathcal{F}_t] + C_{\mathrm{ret}}\delta^2\zeta_t.$$

By Lemma D.23,

$$\mathbb{E}[G_t \mid \mathcal{F}_t] \le C_{\mathrm{sq}} d Q_t.$$

Therefore

$$Q_{t+1} \le \big(1 - \big(2\mu_{\mathrm{loc}} - C_{\mathrm{ret}}C_{\mathrm{sq}}\delta d\big)\delta\big)\, Q_t + \delta\Xi_t + C_{\mathrm{ret}}\delta^2\zeta_t.$$

The local step-size condition implies

$$2\mu_{\mathrm{loc}} - C_{\mathrm{ret}}C_{\mathrm{sq}}\delta d \ge \mu_{\mathrm{loc}}.$$

Set

$$\lambda := \mu_{\mathrm{loc}}, \qquad \alpha := 1 - \lambda\delta.$$

For d large enough, $\alpha \in (0,1)$. Thus, for every $t < \tau_B$,

$$Q_{t+1} \le \alpha Q_t + \Delta N_{t+1}, \tag{198}$$

where

$$\Delta N_{t+1} := \delta\Xi_t + C_{\mathrm{ret}}\delta^2\zeta_t.$$

Moreover,

$$\mathbb{E}[\Delta N_{t+1} \mid \mathcal{F}_t] = 0.$$

Step 2: martingale variance on the stopped epoch.

By Lemma D.23,

$$\mathbb{E}[\Xi_t^2 \mid \mathcal{F}_t] \le C_{\mathrm{sq}} Q_t^2,$$

and

$$\mathbb{E}[\zeta_t^2 \mid \mathcal{F}_t] = \mathbb{E}\Big[\big(G_t - \mathbb{E}[G_t \mid \mathcal{F}_t]\big)^2 \Big| \mathcal{F}_t\Big] \le C_{\mathrm{sq}} d^2 Q_t^2.$$

Hence, using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\mathbb{E}[(\Delta N_{t+1})^2 \mid \mathcal{F}_t] \le C\delta^2 Q_t^2 + C\delta^4 d^2 Q_t^2.$$

Since δd is bounded under the local step-size regime,

$$\mathbb{E}[(\Delta N_{t+1})^2 \mid \mathcal{F}_t] \le C\delta^2 Q_t^2.$$

On $\{t < \tau_B\}$, $Q_t \le 2B$, and therefore

$$\mathbb{E}[(\Delta N_{t+1})^2 \mid \mathcal{F}_t] \le C\delta^2 B^2.$$

Define the stopped martingale increments

$$\Delta\widetilde{N}_{t+1} := \mathbf{1}_{\{t<\tau_B\}}\Delta N_{t+1},$$

and for $u = s, \dots, s + L_{\mathrm{ep}}$,

$$\widetilde{N}_u := \sum_{k=s}^{u-1} \Delta\widetilde{N}_{k+1}, \qquad \widetilde{N}_s := 0.$$

Then $(\widetilde{N}_u)_{u=s}^{s+L_{\mathrm{ep}}}$ is a martingale, and

$$\mathbb{E}\Big[|\widetilde{N}_{s+L_{\mathrm{ep}}}|^2 \Big| \mathcal{F}_s\Big] = \mathbb{E}\left[\sum_{k=s}^{s+L_{\mathrm{ep}}-1} \mathbb{E}[(\Delta\widetilde{N}_{k+1})^2 \mid \mathcal{F}_k] \,\middle|\, \mathcal{F}_s\right] \le C L_{\mathrm{ep}}\delta^2 B^2.$$

Since

$$L_{\mathrm{ep}} = \left\lceil \frac{C_{\mathrm{ep}}}{\delta} \right\rceil \le \frac{C}{\delta}$$

for d large enough,

$$\mathbb{E}\Big[|\widetilde{N}_{s+L_{\mathrm{ep}}}|^2 \Big| \mathcal{F}_s\Big] \le C\delta B^2.$$

By Doob's L^2 maximal inequality,

$$\mathbb{E}\left[\sup_{s\le u\le s+L_{\mathrm{ep}}} |\widetilde{N}_u|^2 \middle| \mathcal{F}_s\right] \le C\delta B^2.$$

Therefore, by Markov's inequality,

$$\mathbb{P}\left(\sup_{s\le u\le s+L_{\mathrm{ep}}} |\widetilde{N}_u| > \frac{B}{8} \;\middle|\; \mathcal{F}_s\right) \le C\delta.$$

Step 3: weighted martingale control.

For $1 \le \ell \le L_{\mathrm{ep}}$, define the stopped geometric convolution

$$\widetilde{S}_\ell := \sum_{j=0}^{\ell-1} \alpha^{\ell-1-j} \Delta\widetilde{N}_{s+j+1}.$$

We claim that

$$\sup_{1\le\ell\le L_{\mathrm{ep}}} |\widetilde{S}_\ell| \le 2 \sup_{s\le u\le s+L_{\mathrm{ep}}} |\widetilde{N}_u|.$$

Indeed, for fixed ℓ, write

$$A_j := \widetilde{N}_{s+j}, \qquad j = 0, \dots, \ell.$$

Then $A_0 = 0$ and

$$\Delta\widetilde{N}_{s+j+1} = A_{j+1} - A_j.$$

By summation by parts,

$$\widetilde{S}_\ell = A_\ell - (1-\alpha)\sum_{j=1}^{\ell-1} \alpha^{\ell-1-j} A_j.$$

Thus

$$|\widetilde{S}_\ell| \le \sup_{0\le j\le\ell} |A_j| + (1-\alpha)\sum_{j=1}^{\ell-1} \alpha^{\ell-1-j} \sup_{0\le j\le\ell} |A_j| \le 2 \sup_{0\le j\le\ell} |A_j|.$$

Taking the supremum over ℓ proves the claim.

Define the epoch good event

$$\mathcal{E}_{\mathrm{ep}} := \left\{ \sup_{1\le\ell\le L_{\mathrm{ep}}} |\widetilde{S}_\ell| \le \frac{B}{4} \right\}.$$

By the preceding maximal bound,

$$\mathbb{P}(\mathcal{E}_{\mathrm{ep}} \mid \mathcal{F}_s) \ge 1 - C\delta.$$

Step 4: contraction on $\mathcal{E}_{\mathrm{ep}}$.

We prove by induction that, on $\mathcal{E}_{\mathrm{ep}}$,

$$Q_{s+\ell} \le \alpha^\ell B + \frac{B}{4} \qquad \text{for every } 0 \le \ell \le L_{\mathrm{ep}}.$$

For $\ell = 0$, this follows from $Q_s \le B$.

Assume the claim has been proved up to time $s+\ell-1$, where $1 \le \ell \le L_{\mathrm{ep}}$. Then, for every $0 \le j \le \ell-1$,

$$Q_{s+j} \le \alpha^j B + \frac{B}{4} \le B + \frac{B}{4} < 2B.$$

Hence $\tau_B > s+\ell-1$, so the original increments and the stopped increments agree on the interval $s, \dots, s+\ell-1$. Iterating (198) gives

$$Q_{s+\ell} \le \alpha^\ell Q_s + \sum_{j=0}^{\ell-1} \alpha^{\ell-1-j} \Delta N_{s+j+1} = \alpha^\ell Q_s + \widetilde{S}_\ell.$$

On $\mathcal{E}_{\mathrm{ep}}$, this yields

$$Q_{s+\ell} \le \alpha^\ell B + \frac{B}{4}.$$

This closes the induction. In particular,

$$Q_{s+\ell} < 2B \qquad \text{for every } 0 \le \ell \le L_{\mathrm{ep}},$$

and therefore

$$\tau_B > s + L_{\mathrm{ep}}.$$

It remains to choose C_{ep} so that the endpoint is at most $B/2$. Choose C_{ep} large enough that

$$e^{-\lambda C_{\mathrm{ep}}} \le \frac{1}{4}.$$

Since

$$L_{\mathrm{ep}} = \left\lceil \frac{C_{\mathrm{ep}}}{\delta} \right\rceil, \qquad \alpha = 1 - \lambda\delta,$$

we have, for d large enough,

$$\alpha^{L_{\mathrm{ep}}} \le e^{-\lambda\delta L_{\mathrm{ep}}} \le e^{-\lambda C_{\mathrm{ep}}} \le \frac{1}{4}.$$

Thus, on $\mathcal{E}_{\mathrm{ep}}$,

$$Q_{s+L_{\mathrm{ep}}} \le \alpha^{L_{\mathrm{ep}}} B + \frac{B}{4} \le \frac{B}{4} + \frac{B}{4} = \frac{B}{2}.$$

Together with $\tau_B > s + L_{\mathrm{ep}}$, this gives

$$\mathcal{E}_{\mathrm{ep}} \subseteq \left\{ Q_{s+L_{\mathrm{ep}}} \le \frac{B}{2} \text{ and } \tau_B > s + L_{\mathrm{ep}} \right\}.$$

Since

$$\mathbb{P}(\mathcal{E}_{\mathrm{ep}} \mid \mathcal{F}_s) \ge 1 - C\delta,$$

the proposition follows. □

Proposition D.25 (Iterated local contraction)**.** *Assume Proposition D.24. Let*

$$\delta = \delta_{\mathrm{loc}} = \frac{c_{\mathrm{loc}}}{d}$$

satisfy the local step-size condition, and let

$$L_{\mathrm{ep}} := \left\lceil \frac{C_{\mathrm{ep}}}{\delta} \right\rceil$$

be the epoch length from Proposition D.24. Let s be a stopping time and suppose that, conditionally on $\mathcal{F}_s$,

$$Q_s \le B_0, \qquad 0 < B_0 \le \frac{r_{\mathrm{loc}}^2}{32}.$$

For $\varepsilon \in (0, B_0]$, define

$$J_\varepsilon := \left\lceil \log_2 \frac{B_0}{\varepsilon} \right\rceil, \qquad N_\varepsilon := J_\varepsilon L_{\mathrm{ep}}.$$

Then

$$\mathbb{P}\left(Q_{s+N_\varepsilon} \le \varepsilon \text{ and } Q_t \le 2B_0 \text{ for every } s \le t \le s+N_\varepsilon \mid \mathcal{F}_s\right) \ge 1 - C\delta J_\varepsilon. \tag{199}$$

In particular, since $\delta = c_{\mathrm{loc}}/d$,

$$N_\varepsilon \le Cd\left(1 + \log\frac{B_0}{\varepsilon}\right), \tag{200}$$

and

$$\mathbb{P}\left(Q_{s+N_\varepsilon} \le \varepsilon \mid \mathcal{F}_s\right) \ge 1 - C\frac{\log(2B_0/\varepsilon)}{d}. \tag{201}$$

Proof. If $\varepsilon = B_0$, then $J_\varepsilon = 0$, $N_\varepsilon = 0$, and the claim follows from $Q_s \le B_0 = \varepsilon$. Hence assume $\varepsilon < B_0$, so $J_\varepsilon \ge 1$.

Set

$$s_j := s + jL_{\mathrm{ep}}, \qquad B_j := 2^{-j}B_0, \qquad j = 0, 1, \dots, J_\varepsilon.$$

Since s is a stopping time, each s_j is also a stopping time. Moreover,

$$B_j \le B_0 \le \frac{r_{\mathrm{loc}}^2}{32},$$

so Proposition D.24 is applicable at every epoch as soon as $Q_{s_j} \le B_j$.

For $j = 0, \dots, J_\varepsilon - 1$, define the epoch-success event

$$\mathcal{A}_j := \left\{Q_{s_{j+1}} \le B_{j+1} \text{ and } Q_t \le 2B_j \text{ for every } s_j \le t \le s_{j+1}\right\}.$$

Also define

$$\mathcal{G}_m := \bigcap_{j=0}^{m-1} \mathcal{A}_j, \qquad m = 1, \dots, J_\varepsilon,$$

with $\mathcal{G}_0$ equal to the whole probability space.

We claim that, for every $m \le J_\varepsilon$, on $\mathcal{G}_m$,

$$Q_{s_m} \le B_m$$

and

$$Q_t \le 2B_0 \qquad \text{for every } s \le t \le s_m.$$

Indeed, the first statement follows by induction from the endpoint conclusion in each $\mathcal{A}_j$. The second follows because, on the j-th epoch,

$$Q_t \le 2B_j \le 2B_0.$$

Now fix $j \in \{0, \dots, J_\varepsilon - 1\}$. On $\mathcal{G}_j$, we have $Q_{s_j} \le B_j$. Applying Proposition D.24 conditionally on $\mathcal{F}_{s_j}$ gives

$$\mathbb{P}\left(\mathcal{A}_j^c \mid \mathcal{F}_{s_j}\right) \le C\delta \qquad \text{on } \mathcal{G}_j.$$

Therefore,

$$\mathbb{P}\left(\mathcal{G}_j \cap \mathcal{A}_j^c \mid \mathcal{F}_s\right) = \mathbb{E}\left[\mathbf{1}_{\mathcal{G}_j}\mathbb{P}\left(\mathcal{A}_j^c \mid \mathcal{F}_{s_j}\right) \middle| \mathcal{F}_s\right] \le C\delta.$$

Taking a union bound over the J_ε epochs yields

$$\mathbb{P}\left(\mathcal{G}_{J_\varepsilon}^c \mid \mathcal{F}_s\right) \le C\delta J_\varepsilon.$$

Hence

$$\mathbb{P}\left(\mathcal{G}_{J_\varepsilon} \mid \mathcal{F}_s\right) \ge 1 - C\delta J_\varepsilon.$$

On $\mathcal{G}_{J_\varepsilon}$, we have

$$Q_{s+N_\varepsilon} = Q_{s_{J_\varepsilon}} \le B_{J_\varepsilon} = 2^{-J_\varepsilon}B_0 \le \varepsilon,$$

by the definition of J_ε. Also, as shown above,

$$Q_t \le 2B_0 \qquad \text{for every } s \le t \le s + N_\varepsilon.$$

This proves the probability estimate.

It remains only to record the sample complexity. Since

$$L_{\mathrm{ep}} = \left\lceil \frac{C_{\mathrm{ep}}}{\delta} \right\rceil \le \frac{C}{\delta}$$

for d sufficiently large, and

$$J_\varepsilon = \left\lceil \log_2 \frac{B_0}{\varepsilon} \right\rceil \le C \left(1 + \log \frac{B_0}{\varepsilon}\right),$$

we get

$$N_\varepsilon = J_\varepsilon L_{\mathrm{ep}} \le C\delta^{-1} \left(1 + \log \frac{B_0}{\varepsilon}\right).$$

With $\delta = c_{\mathrm{loc}}/d$, this becomes

$$N_\varepsilon \le Cd \left(1 + \log \frac{B_0}{\varepsilon}\right).$$

Finally,

$$\delta J_\varepsilon \le \frac{C}{d} \log \frac{2B_0}{\varepsilon},$$

after increasing C. This gives

$$\mathbb{P}\left(Q_{s+N_\varepsilon} \le \varepsilon \mid \mathcal{F}_s\right) \ge 1 - C\frac{\log(2B_0/\varepsilon)}{d}.$$

The proof is complete. □

Theorem D.26 (Discrete-time weak-to-strong recovery)**.** *Assume Theorem D.17 holds with event G_T^{weak}, stopping time t_{esc}, and constants $T_{\mathrm{esc}}, c_{\mathrm{esc}}, C_{\eta,\mathrm{esc}}$, such that on G_T^{weak},*

$$t_{\mathrm{esc}} \le T_{\mathrm{esc}}, \qquad \min\{\rho_{t_{\mathrm{esc}}}, \chi_{t_{\mathrm{esc}}}\} \ge c_{\mathrm{esc}}, \qquad \eta_{t_{\mathrm{esc}}} \le C_{\eta,\mathrm{esc}} \frac{\log d}{\sqrt{d}}.$$

Assume (A4). After t_{esc}, switch to the local step size $\delta_{\mathrm{loc}} = \frac{c_{\mathrm{loc}}}{d}$, where $c_{\mathrm{loc}} > 0$ is chosen sufficiently small so that the assumptions of Lemma D.19, Proposition D.21, and Proposition D.24 hold. Let

$$Q_{\mathrm{in}} := \frac{r_{\mathrm{loc}}^2}{64}.$$

Then, for every fixed or polynomially small ν_{in}, there exists an event E_{str} satisfying

$$\mathbb{P}(E_{\mathrm{str}}) \ge 1 - \mathbb{P}({G_T^{\mathrm{weak}}}^c) - \nu_{\mathrm{in}}, \tag{202}$$

on which there is a stopping time t_1 such that

$$t_1 \le t_{\mathrm{esc}} + C_{\mathrm{ent}}\delta_{\mathrm{loc}}^{-1} \le T_{\mathrm{esc}} + C_{\mathrm{ent}}\delta_{\mathrm{loc}}^{-1}, \tag{203}$$

and

$$Q_{t_1} \le Q_{\mathrm{in}} \le \frac{r_{\mathrm{loc}}^2}{32}. \tag{204}$$

Moreover, for every $\varepsilon \in (0, Q_{\mathrm{in}}]$, define

$$J_\varepsilon := \left\lceil \log_2 \frac{Q_{\mathrm{in}}}{\varepsilon} \right\rceil, \qquad L_{\mathrm{ep}} := \left\lceil \frac{C_{\mathrm{ep}}}{\delta_{\mathrm{loc}}} \right\rceil, \qquad N_\varepsilon := J_\varepsilon L_{\mathrm{ep}}.$$

Then

$$\mathbb{P}\left(Q_{t_1+N_\varepsilon} \le \varepsilon\right) \ge 1 - \mathbb{P}({G_T^{\mathrm{weak}}}^c) - \nu_{\mathrm{in}} - C\delta_{\mathrm{loc}} J_\varepsilon. \tag{205}$$

Equivalently, since $\delta_{\mathrm{loc}} = c_{\mathrm{loc}}/d$,

$$\mathbb{P}\left(Q_{t_1+N_\varepsilon} \le \varepsilon\right) \ge 1 - \mathbb{P}({G_T^{\mathrm{weak}}}^c) - \nu_{\mathrm{in}} - C\frac{\log(2Q_{\mathrm{in}}/\varepsilon)}{d}.$$

On the same event,

$$1 - \rho_{t_1+N_\varepsilon} \le \varepsilon, \qquad 1 - \chi_{t_1+N_\varepsilon} \le \varepsilon,$$

and hence

$$\min\{\rho_{t_1+N_\varepsilon}, \chi_{t_1+N_\varepsilon}\} \geq 1 - \varepsilon. \tag{206}$$

In particular,

$$N_\varepsilon \leq Cd\left(1 + \log\frac{Q_{\text{in}}}{\varepsilon}\right).$$

With the weak phase satisfies $T_{\text{esc}} = O(d\log^2 d)$, the total number of SGD samples needed to reach $Q \leq \varepsilon$ is

$$O(d\log^2 d) + O\left(d\log\frac{1}{\varepsilon}\right). \tag{207}$$

Proof. Let

$$t_0 := t_{\text{esc}}, \qquad \delta := \delta_{\text{loc}}.$$

Work first on the weak-recovery event G_T^{weak}. On this event,

$$\min\{\rho_{t_0}, \chi_{t_0}\} \geq c_{\text{esc}}, \qquad \eta_{t_0} \leq C_{\eta,\text{esc}}\frac{\log d}{\sqrt{d}}.$$

Together with the incoherence bound

$$\|\theta^*\|_\infty \leq C_*\sqrt{\frac{\log d}{d}},$$

these are exactly the assumptions of Lemma D.19.

Apply Lemma D.19 with $C_{\text{in}} \geq C_{\text{ent}}$. Conditionally on $\mathcal{F}_{t_0}$, there is an event E_{in} such that

$$\mathbb{P}(E_{\text{in}} \mid \mathcal{F}_{t_0}) \geq 1 - \nu_{\text{in}}.$$

Hence, by the tower property,

$$\mathbb{P}(G_T^{\text{weak}} \cap E_{\text{in}}) = \mathbb{E}\left[\mathbf{1}_{G_T^{\text{weak}}}\mathbb{P}(E_{\text{in}} \mid \mathcal{F}_{t_0})\right] \geq (1 - \nu_{\text{in}})\mathbb{P}(G_T^{\text{weak}}).$$

Therefore

$$\mathbb{P}(G_T^{\text{weak}} \cap E_{\text{in}}) \geq 1 - \mathbb{P}(G_T^{\text{weak}^c}) - \nu_{\text{in}}.$$

On $G_T^{\text{weak}} \cap E_{\text{in}}$, Proposition D.21 gives a time

$$t_1 \leq t_{\text{esc}} + C_{\text{ent}}\delta^{-1}$$

such that

$$\rho_{t_1} \geq \cos(r_{\text{loc}}/8), \qquad \chi_{t_1} \geq \cos(r_{\text{loc}}/8),$$

and consequently

$$Q_{t_1} \leq \frac{r_{\text{loc}}^2}{64} = Q_{\text{in}}.$$

Choosing t_1 as the stopping time constructed in Proposition D.21, this proves the first part. Since $t_{\text{esc}} \leq T_{\text{esc}}$ on G_T^{weak}, we also have

$$t_1 \leq T_{\text{esc}} + C_{\text{ent}}\delta^{-1}.$$

Set

$$E_{\text{str}}^{(0)} := G_T^{\text{weak}} \cap E_{\text{in}}.$$

Then

$$\mathbb{P}(E_{\text{str}}^{(0)}) \geq 1 - \mathbb{P}(G_T^{\text{weak}^c}) - \nu_{\text{in}}.$$

Now fix $\varepsilon \in (0, Q_{\text{in}}]$. On $E_{\text{str}}^{(0)}$, we may apply Proposition D.25 conditionally on $\mathcal{F}_{t_1}$, with

$$s = t_1, \qquad B_0 = Q_{\text{in}}.$$

Indeed,

$$Q_{t_1} \leq Q_{\text{in}} \leq \frac{r_{\text{loc}}^2}{32}.$$

Thus, conditionally on $\mathcal{F}_{t_1}$,

$$\mathbb{P}\left(Q_{t_1+N_\varepsilon} \le \varepsilon \mid \mathcal{F}_{t_1}\right) \ge 1 - C\delta J_\varepsilon.$$

Using the tower property again,

$$\mathbb{P}\left(E_{\mathrm{str}}^{(0)} \cap \{Q_{t_1+N_\varepsilon} > \varepsilon\}\right) \le C\delta J_\varepsilon.$$

Therefore

$$\mathbb{P}\left(Q_{t_1+N_\varepsilon} \le \varepsilon\right) \ge \mathbb{P}(E_{\mathrm{str}}^{(0)}) - C\delta J_\varepsilon.$$

Combining this with the lower bound on $\mathbb{P}(E_{\mathrm{str}}^{(0)})$ gives

$$\mathbb{P}\left(Q_{t_1+N_\varepsilon} \le \varepsilon\right) \ge 1 - \mathbb{P}(G_T^{\mathrm{weak}\,c}) - \nu_{\mathrm{in}} - C\delta J_\varepsilon.$$

Since

$$\delta = \frac{c_{\mathrm{loc}}}{d}$$

and

$$J_\varepsilon = \left\lceil \log_2 \frac{Q_{\mathrm{in}}}{\varepsilon} \right\rceil \le C \log \frac{2Q_{\mathrm{in}}}{\varepsilon},$$

we obtain

$$C\delta J_\varepsilon \le C\frac{\log(2Q_{\mathrm{in}}/\varepsilon)}{d}.$$

It remains to convert the Q-bound into the overlap bound. If

$$Q_{t_1+N_\varepsilon} \le \varepsilon,$$

then

$$\mathrm{dist}_{\mathbb{S}}(\theta_{t_1+N_\varepsilon}, \theta^*)^2 + \mathrm{dist}_{\mathbb{S}}(\beta_{t_1+N_\varepsilon}, \beta^*)^2 = 2Q_{t_1+N_\varepsilon} \le 2\varepsilon.$$

In particular,

$$\mathrm{dist}_{\mathbb{S}}(\theta_{t_1+N_\varepsilon}, \theta^*)^2 \le 2\varepsilon, \qquad \mathrm{dist}_{\mathbb{S}}(\beta_{t_1+N_\varepsilon}, \beta^*)^2 \le 2\varepsilon.$$

Using $1 - \cos x \le x^2/2$, we get

$$1 - \chi_{t_1+N_\varepsilon} = 1 - \cos(\mathrm{dist}_{\mathbb{S}}(\theta_{t_1+N_\varepsilon}, \theta^*)) \le \varepsilon,$$

and similarly

$$1 - \rho_{t_1+N_\varepsilon} \le \varepsilon.$$

Thus

$$\min\{\rho_{t_1+N_\varepsilon}, \chi_{t_1+N_\varepsilon}\} \ge 1 - \varepsilon.$$

Finally, Proposition D.25 gives

$$N_\varepsilon = J_\varepsilon L_{\mathrm{ep}} \le Cd\left(1 + \log \frac{Q_{\mathrm{in}}}{\varepsilon}\right).$$

Since

$$t_1 \le t_{\mathrm{esc}} + C_{\mathrm{ent}}\delta^{-1} = t_{\mathrm{esc}} + O(d),$$

the total number of samples after initialization is bounded by

$$t_{\mathrm{esc}} + O(d) + O\left(d \log \frac{Q_{\mathrm{in}}}{\varepsilon}\right).$$

Because $Q_{\mathrm{in}} = r_{\mathrm{loc}}^2/64$ is a fixed constant independent of d, this is

$$t_{\mathrm{esc}} + O\left(d \log \frac{1}{\varepsilon}\right).$$

If $T_{\mathrm{esc}} = O(d \log^2 d)$, then on G_T^{weak},

$$t_1 + N_\varepsilon \le O(d \log^2 d) + O\left(d \log \frac{1}{\varepsilon}\right).$$

The proof is complete. □

E Auxiliary Lemmas

E.1 Stein Bounds

Here, we present the comparison bounds between the Riemannian Hessian of the true loss and that of the Gaussian model via a Stein-type approximation, for θ-Hessian and β-Hessian respectively. They are used in the Lemma C.12. We prove our results based on this theorem from [8]:

Theorem E.1 (Stein coupling for product measures [8, Thm. D.1])**.** *Let $X = (X_1, \dots, X_d)$ be a random vector with independent coordinates such that $\mathbb{E}[X] = 0$, $\mathbb{E}[XX^\top] = I_d$, and $\mathbb{E}|X_i|^3 = \tau_i^3 < \infty$. Let $G \sim \mathcal{N}(0, I_d)$ be standard Gaussian. If $h : \mathbb{R}^d \to \mathbb{R}$ is three-times differentiable, then*

$$\big|\mathbb{E}[h(X)] - \mathbb{E}[h(G)]\big| \;\le\; \frac{5}{6}\sum_{i=1}^{d} \tau_i^3 \,\big\|\partial_{x_i}^3 h\big\|_\infty. \tag{208}$$

Stein Bound for θ-Hessian

Corollary E.2 (Block Stein coupling with moment remainder)**.** *Let $(X_i, X_i^*)_{i=1}^d$ be independent $\mathbb{R}^2$-valued random vectors with*

$$\mathbb{E}[X_i] = \mathbb{E}[X_i^*] = 0, \qquad \mathbb{E}[X_i^2] = \mathbb{E}[(X_i^*)^2] = 1, \qquad \mathbb{E}[X_i X_i^*] = m,$$

and $\tau := \max_i(\mathbb{E}|X_i|^5 + \mathbb{E}|X_i^*|^5) < \infty$. Let $(G_i, G_i^*)_{i=1}^d$ be iid $\mathcal{N}(0, \Sigma_m)$ with $\Sigma_m = \left(\begin{smallmatrix} 1 & m \\ m & 1 \end{smallmatrix}\right)$. If $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is three-times differentiable in each block (x_i, x_i^*), then

$$\big|\mathbb{E}[h(X, X^*)] - \mathbb{E}[h(G, G^*)]\big| \le C(\tau)\sum_{i=1}^{d}\sum_{\substack{a,b\ge 0\\ a+b=3}} \sup_{j\le d} \mathbb{E}\Big[\Big|\frac{\partial^{a+b}h}{\partial x_i^a\,\partial (x_i^*)^b}(W^{(j)}, W^{*(j)})\Big|\Big], \tag{209}$$

where $(W^{(j)}, W^{*(j)})$ is the hybrid that agrees with (G, G^*) on blocks $1, \dots, j$ and with (X, X^*) on blocks $j{+}1, \dots, d$.

Proof. We have $(W^{(0)}, W^{*(0)}) = (X, X^*)$ and $(W^{(d)}, W^{*(d)}) = (G, G^*)$, so

$$\mathbb{E}[h(X, X^*)] - \mathbb{E}[h(G, G^*)] = \sum_{j=1}^{d}\Big(\mathbb{E}\big[h(W^{(j-1)}, W^{*(j-1)})\big] - \mathbb{E}\big[h(W^{(j)}, W^{*(j)})\big]\Big). \tag{210}$$

Each summand differs only in block j. With all other blocks frozen, define $\varphi_j(s,t) := h\big(W^{(j)}\big|_{(s,t)\text{ in block } j}\big)$, so the j-th summand equals $\mathbb{E}[\varphi_j(X_j, X_j^*) - \varphi_j(G_j, G_j^*)]$. Taylor-expand φ_j around $(0,0)$ to second order with integral remainder:

$$\varphi_j(s,t) = \sum_{\substack{a,b\ge 0\\ a+b\le 2}} \frac{1}{a!\,b!}\frac{\partial^{a+b}\varphi_j}{\partial s^a \partial t^b}\Big|_{(0,0)} s^a t^b \;+\; R_j(s,t),$$

where the integral form of the remainder gives

$$|R_j(s,t)| \le C_0\,(|s| + |t|)^3 \sum_{\substack{a,b\ge 0\\ a+b=3}} \sup_{|\xi|\le|s|,\,|\zeta|\le|t|} \Big|\frac{\partial^{a+b}\varphi_j}{\partial s^a \partial t^b}(\xi,\zeta)\Big|.$$

Taking $\mathbb{E}[\varphi_j(X_j, X_j^*)] - \mathbb{E}[\varphi_j(G_j, G_j^*)]$: the zeroth-order terms cancel trivially; the first-order terms cancel since $\mathbb{E}[X_j] = \mathbb{E}[G_j] = 0$ and $\mathbb{E}[X_j^*] = \mathbb{E}[G_j^*] = 0$; the second-order terms cancel since

$$\mathbb{E}[X_j^2] = \mathbb{E}[G_j^2] = 1, \quad \mathbb{E}[(X_j^*)^2] = \mathbb{E}[(G_j^*)^2] = 1, \quad \mathbb{E}[X_j X_j^*] = \mathbb{E}[G_j G_j^*] = m.$$

Hence only the remainder contributes:

$$\big|\mathbb{E}\big[\varphi_j(X_j, X_j^*) - \varphi_j(G_j, G_j^*)\big]\big| \le C_0\big(\mathbb{E}(|X_j|{+}|X_j^*|)^3 + \mathbb{E}(|G_j|{+}|G_j^*|)^3\big) \sum_{\substack{a,b\ge 0\\ a+b=3}} \mathbb{E}\Big[\Big|\frac{\partial^{a+b}\varphi_j}{\partial s^a \partial t^b}\Big|\Big],$$

where the moment prefactors are bounded by $C'(\tau)$, and the expectations of the third-order partials use Hölder to separate the (s,t)-dependence from the frozen blocks. Summing over j and noting $\partial^{a+b}\varphi_j/\partial s^a \partial t^b = \partial^{a+b}h/\partial x_j^a \partial (x_j^*)^b$ evaluated at the hybrid yields (209). □

Corollary E.3 (Stein estimate for a weighted quadratic statistic)**.** *Let $X_1, \dots, X_d$ be independent, centered, unit-variance random variables with $\sup_i \mathbb{E}|X_i|^8 < \infty$, and let $\Gamma \sim N(0, I_d)$. Let $a, u \in \mathbb{R}^d$ satisfy $\|a\|_2 = \|u\|_2 = 1$. If $F \in C^3(\mathbb{R})$ and $\|F^{(k)}\|_\infty < \infty$ for $1 \le k \le 3$, then*

$$\left|\mathbb{E}\left[F(a \cdot X)(u \cdot X)^2\right] - \mathbb{E}\left[F(a \cdot \Gamma)(u \cdot \Gamma)^2\right]\right| \le C_F \left(\sum_{i=1}^d |a_i| u_i^2 + \sum_{i=1}^d a_i^2 |u_i| + \sum_{i=1}^d |a_i|^3\right),$$

where C_F depends only on F and the law of X_1, but not on d, a, u. In particular,

$$\left|\mathbb{E}\left[F(a \cdot X)(u \cdot X)^2\right] - \mathbb{E}\left[F(a \cdot \Gamma)(u \cdot \Gamma)^2\right]\right| \le C_F \|a\|_\infty.$$

Proof. Apply the scalar-block moment version of the Lindeberg–Stein comparison, as in Corollary E.2, to

$$h(x) := F(a \cdot x)(u \cdot x)^2.$$

For each coordinate i,

$$\partial_i^3 h(x) = a_i^3 F^{(3)}(a \cdot x)(u \cdot x)^2 + 6a_i^2 u_i F''(a \cdot x)(u \cdot x) + 6a_i u_i^2 F'(a \cdot x).$$

At every hybrid vector appearing in the Lindeberg interpolation,

$$\mathbb{E}|u \cdot W|^2 \le C\|u\|_2^2 = C, \qquad \mathbb{E}|u \cdot W| \le C,$$

because the coordinates have uniformly bounded second moments. Hence

$$\mathbb{E}|\partial_i^3 h(W)| \le C_F \left(|a_i|^3 + a_i^2 |u_i| + |a_i| u_i^2\right).$$

Summing over i gives the first estimate.

For the final bound, use $\|a\|_2 = \|u\|_2 = 1$:

$$\sum_i |a_i| u_i^2 \le \|a\|_\infty, \qquad \sum_i a_i^2 |u_i| \le \left(\sum_i a_i^4\right)^{1/2} \|u\|_2 \le \|a\|_\infty,$$

and

$$\sum_i |a_i|^3 \le \|a\|_\infty \sum_i a_i^2 = \|a\|_\infty.$$

This proves the claim. □

Lemma E.4 (Teacher-level θ-Hessian comparison)**.** *Let*

$$A_\psi := \mathbb{E}[\psi'(G)^2], \qquad G \sim N(0,1).$$

Assume $\mathbb{E}[\phi(G)] = 0$, $\mathbb{E}[\phi(G)^2] = 1$, $\mathbb{E}|\phi(G)|^8 < \infty$, and $\psi \in C^4(\mathbb{R})$ with bounded derivatives up to order 4. Then there exists $C_\theta > 0$, independent of d, such that for every $u \in T_{\theta^*}\mathbb{S}^{d-1}$,

$$\left|\mathrm{Hess}_{\mathbb{S},\theta}\, L(y^*)[u,u] - 2A_\psi \|u\|^2\right| \le C_\theta \|\theta^*\|_\infty \|u\|^2.$$

Equivalently,

$$|\mathrm{Hess}_{\mathbb{S},\theta}\, L(y^*)[u,u] - \mathrm{Hess}_{\mathbb{S},\theta}\, L_G(y^*)[u,u]| \le C_\theta \|\theta^*\|_\infty \|u\|^2.$$

Proof. By quadratic scaling, it is enough to prove the result for $\|u\| = 1$. Set

$$a := \theta^*, \qquad X_i := \phi(G_i), \qquad S := a \cdot X = \sum_{i=1}^d a_i X_i, \qquad A_u := u \cdot X = \sum_{i=1}^d u_i X_i,$$

where $G_i \overset{\text{iid}}{\sim} N(0,1)$. Also define

$$F := \psi'^2.$$

Since $u \in T_{\theta^*}\mathbb{S}^{d-1}$, we have $a \cdot u = 0$.

At the teacher, the residual vanishes pointwise. Therefore, along a spherical geodesic in the θ-variable with initial velocity u,

$$\mathrm{Hess}_{\mathbb{S},\theta}\, L(y^*)[u,u] = 2\mathbb{E}\left[F(S)A_u^2\right].$$

Let $\Gamma \sim N(0, I_d)$. By Corollary E.3,

$$\left|\mathbb{E}\left[F(a\cdot X)(u\cdot X)^2\right] - \mathbb{E}\left[F(a\cdot \Gamma)(u\cdot \Gamma)^2\right]\right| \le C_F\|a\|_\infty = C_F\|\theta^*\|_\infty.$$

Since $a \cdot u = 0$ and $\|a\| = \|u\| = 1$, the Gaussian random variables $a \cdot \Gamma$ and $u \cdot \Gamma$ are independent standard normals. Hence

$$\mathbb{E}\left[F(a\cdot \Gamma)(u\cdot \Gamma)^2\right] = \mathbb{E}[F(G)]\mathbb{E}[G^2] = A_\psi.$$

Thus

$$|\mathrm{Hess}_{\mathbb{S},\theta}\, L(y^*)[u,u] - 2A_\psi| \le C_\theta\|\theta^*\|_\infty$$

for every unit $u \in T_{\theta^*}\mathbb{S}^{d-1}$. Homogeneity gives the claim for arbitrary u.

Finally, the Gaussian teacher Hessian satisfies

$$\mathrm{Hess}_{\mathbb{S},\theta}\, L_G(y^*)[u,u] = 2A_\psi\|u\|^2,$$

because in the Gaussian model the same computation gives $2\mathbb{E}[F(a\cdot\Gamma)(u\cdot\Gamma)^2]$, and $a\cdot\Gamma$ is independent of $u\cdot\Gamma$. This proves the comparison with L_G. □

Stein Bound for β-Hessian To give a Stein bound for β-Hessian, we start by deriving this corollary from Theorem E.1.

Corollary E.5 (One-dimensional linear statistic)**.** *Assume the setting of Theorem E.1. Let $a \in \mathbb{R}^d$ and $f \in C^3(\mathbb{R})$ with $\|f^{(3)}\|_\infty < \infty$. Define $h(x) := f(\langle a, x\rangle)$. Then*

$$\left|\mathbb{E}[f(\langle a, X\rangle)] - \mathbb{E}[f(\langle a, G\rangle)]\right| \;\le\; \frac{5}{6}\|f^{(3)}\|_\infty \sum_{i=1}^d \tau_i^3|a_i|^3.$$

In particular, if $\sup_i \tau_i \le \tau$ and $\|a\|_2 = 1$, then

$$\left|\mathbb{E}[f(\langle a, X\rangle)] - \mathbb{E}[f(\langle a, G\rangle)]\right| \;\le\; C(\tau, f)\,\|a\|_\infty, \tag{211}$$

since $\sum_i |a_i|^3 \le \|a\|_\infty \sum_i a_i^2 = \|a\|_\infty$.

Proof. Define $h(x) := f(\langle a, x\rangle)$. By the chain rule, all partial derivatives act only through $\langle a, x\rangle$:

$$\frac{\partial^3 h}{\partial x_i^3}(x) = a_i^3 f'''(\langle a, x\rangle).$$

Therefore

$$\left\|\frac{\partial^3 h}{\partial x_i^3}\right\|_\infty = |a_i|^3\|f^{(3)}\|_\infty.$$

Applying Theorem E.1 directly gives

$$\left|\mathbb{E}[f(\langle a, X\rangle)] - \mathbb{E}[f(\langle a, G\rangle)]\right| \le \frac{5}{6}\sum_{i=1}^d \tau_i^3\left\|\partial_{x_i}^3 h\right\|_\infty = \frac{5}{6}\|f^{(3)}\|_\infty \sum_{i=1}^d \tau_i^3|a_i|^3.$$

If $\sup_i \tau_i \le \tau$ and $\|a\|_2 = 1$, then

$$\sum_{i=1}^d \tau_i^3|a_i|^3 \le \tau^3\sum_{i=1}^d |a_i|^3 \le \tau^3\|a\|_\infty \sum_{i=1}^d a_i^2 = \tau^3\|a\|_\infty,$$

giving $C(\tau, f) = \frac{5}{6}\tau^3|f^{(3)}|_\infty$, so we obtained (211). □

Lemma E.6 (Teacher-level β-Hessian comparison)**.** *Let*

$$A_\psi := \mathbb{E}[\psi'(G)^2], \qquad B_\phi := \mathbb{E}[\phi'(G)^2] = \kappa'(1), \qquad G \sim N(0,1).$$

Assume (A3). Then there exists a constant $C_\beta > 0$, independent of d, such that for every $v \in T_{\beta^*}\mathbb{S}^{d-1}$,

$$\left|H_{L,\beta\beta}(y^*)[v,v] - 2A_\psi B_\phi\|v\|^2\right| \le C_\beta\|\theta^*\|_\infty\|v\|^2.$$

Equivalently,

$$|H_{L,\beta\beta}(y^*)[v,v] - H_{L_G,\beta\beta}(y^*)[v,v]| \le C_\beta\|\theta^*\|_\infty\|v\|^2.$$

Proof. By quadratic scaling, it is enough to consider $\|v\| = 1$. Set

$$a_i := \theta_i^*, \qquad X_i := \phi(G_i), \qquad q_i := \phi'(G_i)^2, \qquad S := \sum_{i=1}^d a_i X_i, \qquad S_{-i} := S - a_i X_i,$$

and define

$$F(s) := \psi'(s)^2.$$

By Lemma C.11,

$$\frac{1}{2} H_{L,\beta\beta}(y^*)[v,v] = \sum_{i=1}^d a_i^2 \mathbb{E}[F(S) q_i].$$

We first decouple the i-th coordinate. Since S_{-i} is independent of (X_i, q_i),

$$\mathbb{E}[F(S_{-i}) q_i] = \mathbb{E}[F(S_{-i})]\mathbb{E}[q_i] = B_\phi \, \mathbb{E}[F(S_{-i})].$$

Moreover, by the boundedness of F' and the moment assumption,

$$|\mathbb{E}[F(S) q_i] - \mathbb{E}[F(S_{-i}) q_i]| \le \|F'\|_\infty |a_i| \mathbb{E}[|X_i| q_i] \le C |a_i|.$$

Multiplying by a_i^2 and summing gives

$$\sum_{i=1}^d a_i^2 \mathbb{E}[F(S) q_i] = B_\phi \sum_{i=1}^d a_i^2 \mathbb{E}[F(S_{-i})] + O\left(\sum_{i=1}^d |a_i|^3 \right).$$

Since $\sum_i a_i^2 = 1$ and

$$|\mathbb{E}[F(S)] - \mathbb{E}[F(S_{-i})]| \le \|F'\|_\infty |a_i| \mathbb{E}|X_i|,$$

we also have

$$\sum_{i=1}^d a_i^2 \mathbb{E}[F(S_{-i})] = \mathbb{E}[F(S)] + O\left(\sum_{i=1}^d |a_i|^3 \right).$$

Therefore,

$$\sum_{i=1}^d a_i^2 \mathbb{E}[F(S) q_i] = B_\phi \, \mathbb{E}[F(S)] + O\left(\sum_{i=1}^d |a_i|^3 \right).$$

Now apply the one-dimensional Stein comparison in Corollary E.5 with

$$X_i := \phi(G_i), \qquad a_i := \theta_i^*, \qquad f := F = \psi'^2.$$

By the assumptions on ϕ,

$$\mathbb{E}[X_i] = 0, \qquad \mathbb{E}[X_i^2] = 1, \qquad \tau_\phi^3 := \mathbb{E}|X_i|^3 < \infty,$$

and by the regularity assumptions on ψ, $F \in C^3(\mathbb{R})$ with $\|F^{(3)}\|_\infty < \infty$. Moreover $\|a\|_2 = \|\theta^*\|_2 = 1$. Let $\Gamma = (\Gamma_1, \dots, \Gamma_d) \sim N(0, I_d)$. Then $\langle a, \Gamma \rangle \sim N(0,1)$. Hence Corollary E.5 gives

$$\left| \mathbb{E} F\left(\sum_{i=1}^d a_i X_i \right) - \mathbb{E} F\left(\langle a, \Gamma \rangle \right) \right| \le \frac{5}{6} \|F^{(3)}\|_\infty \sum_{i=1}^d \tau_\phi^3 |a_i|^3.$$

Since $\langle a, \Gamma \rangle \sim G \sim N(0,1)$, this becomes

$$|\mathbb{E}[F(S)] - \mathbb{E}[F(G)]| \le C_F \sum_{i=1}^d |a_i|^3,$$

where $C_F := \frac{5}{6} \tau_\phi^3 \|F^{(3)}\|_\infty$. Finally,

$$\sum_{i=1}^d |a_i|^3 \le \|a\|_\infty \sum_{i=1}^d a_i^2 = \|\theta^*\|_\infty.$$

Therefore

$$|\mathbb{E}[F(S)] - \mathbb{E}[F(G)]| \le C_F \|\theta^*\|_\infty.$$

Combining this with the preceding leave-one-out estimate yields

$$\sum_{i=1}^{d} a_i^2 \mathbb{E}[F(S)q_i] = B_\phi \, \mathbb{E}[F(G)] + O(\|\theta^*\|_\infty) = B_\phi A_\psi + O(\|\theta^*\|_\infty).$$

Multiplying by $2\|v\|^2$ proves the true $\beta\beta$-block estimate.

For the Gaussian loss, since we have (61) in Lemma C.10,

$$L_G(\theta, \beta) = 2 \sum_{k \geq 1} \alpha_k^2 \left(1 - (\kappa(\rho)\chi)^k\right)$$

, gives, at y^*,

$$H_{L_G,\beta\beta}(y^*)[v, v] = 2\kappa'(1) \left(\sum_{k \geq 1} k\alpha_k^2 \right) \|v\|^2.$$

But

$$\sum_{k \geq 1} k\alpha_k^2 = \mathbb{E}[\psi'(G)^2] = A_\psi, \qquad \kappa'(1) = B_\phi.$$

Thus

$$H_{L_G,\beta\beta}(y^*)[v, v] = 2A_\psi B_\phi \|v\|^2,$$

and the comparison follows. □

E.2 Gradient Bounds

In the lemma below, we provide bounds for gradients $g^{(\theta)}, g^{(\beta)}$ to help us to control the evolution of Z_t and η_t in weak recovery, and that of Q_t in strong recovery.

Lemma E.7 (Technical gradient bounds for blockwise spherical SGD)**.** *Assume (A3). Then there exists a constant $K < \infty$, depending only on $L_\phi, L_\psi, \phi(0)$, such that if we draw a fresh sample x_t then with*

$$g^{(\beta)}(\beta, \theta; x_t) := \nabla_\beta^{\mathcal{S}} \ell(\beta, \theta; x_t), \qquad g^{(\theta)}(\beta, \theta; x_t) := \nabla_\theta^{\mathcal{S}} \ell(\beta, \theta; x_t),$$

we have the moment bounds

$$\begin{aligned}
\sup_{\beta,\theta} \mathbb{E}\big[\|g^{(\beta)}(\beta, \theta; x_t)\|^2\big] &\leq Kd, \\
\sup_{\beta,\theta} \mathbb{E}\big[\|g^{(\theta)}(\beta, \theta; x_t)\|^2\big] &\leq Kd, \\
\sup_{\beta,\theta} \mathbb{E}\big[\|g^{(\beta)}(\beta, \theta; x_t)\|^4\big] &\leq Kd^2, \\
\sup_{\beta,\theta} \mathbb{E}\big[\|g^{(\theta)}(\beta, \theta; x_t)\|^4\big] &\leq Kd^2.
\end{aligned}$$

Moreover, for the centered gradient-noise fields defined as

$$\Delta g^{(\beta)}(\beta, \theta; x_t) := g^{(\beta)}(\beta, \theta; x_t) - \mathbb{E}[g^{(\beta)}(\beta, \theta; x_t)], \qquad \Delta g^{(\theta)}(\beta, \theta; x_t) := g^{(\theta)}(\beta, \theta; x_t) - \mathbb{E}[g^{(\theta)}(\beta, \theta; x_t)],$$

we have the directional variance bounds

$$\begin{aligned}
\sup_{\beta,\theta} \sup_{\|u\|=1} \mathbb{E}\big[(\Delta g^{(\beta)}(\beta, \theta; x_t) \cdot u)^2\big] &\leq K, \\
\sup_{\beta,\theta} \sup_{\|u\|=1} \mathbb{E}\big[(\Delta g^{(\theta)}(\beta, \theta; x_t) \cdot u)^2\big] &\leq K.
\end{aligned}$$

Consequently, along the SGD trajectory (β_t, θ_t), since x_t is independent of $\mathcal{F}_t$, the same bounds hold conditionally on $\mathcal{F}_t$:

$$\mathbb{E}\big[\|g_t^{(\beta)}\|^2 \mid \mathcal{F}_t\big] \leq Kd, \quad \mathbb{E}\big[\|g_t^{(\beta)}\|^4 \mid \mathcal{F}_t\big] \leq Kd^2, \quad \sup_{\|u\|=1} \mathbb{E}\big[((\Delta g_t^{(\beta)}) \cdot u)^2 \mid \mathcal{F}_t\big] \leq K,$$

and similarly for the θ-block.

Proof. Since spherical gradient is the orthogonal projection of the Euclidean gradient onto the tangent space via $\nabla^{\mathcal{S}} F(x) = \nabla F(x) - (\nabla F(x) \cdot x)x$, we have $\|\nabla^{\mathcal{S}} F(x)\| \leq \|\nabla F(x)\|$ and $|u \cdot \nabla^{\mathcal{S}} F(x)| \leq |u \cdot \nabla F(x)| \leq \|\nabla F(x)\|$ for $|u| = 1$, it suffices to bound the Euclidean gradients.

Based on the model definition in Section 3.1 and define the per-sample loss as

$$l = (f - f^*)^2, \tag{212}$$

we have the per-sample Euclidean gradients w.r.t. each layer as:

$$\nabla_\theta l = 2(f - f^*) \cdot \psi'(s) \cdot z \tag{213}$$

$$\nabla_\beta l = 2(f - f^*) \cdot \psi'(s) \cdot \sum_{i=1}^{d} \theta_i \phi'(u_i) x^{(i)}. \tag{214}$$

We simplify the expression by defining

$$A(x; \beta, \theta) := 2(f - f^*) \cdot \psi'(s) \in \mathbb{R}. \tag{215}$$

θ bounds

$$\mathbb{E}[\|g^{(\theta)}\|^2] \leq \mathbb{E}[\|\nabla_\theta l\|^2] \tag{216}$$

$$= \mathbb{E}[|A|^2 \|z\|^2] \tag{217}$$

$$\leq (\mathbb{E}|A|^4)^{1/2} \cdot (\mathbb{E}\|z\|^4)^{1/2} \text{ by Cauchy-Schwarz} \tag{218}$$

For the first term of (218), with the bounded ψ, ψ' assumption, we have

$$|A| = 2|\psi(s) - \psi(s^*)||\psi'(s)| \leq 2(|\psi(s)| + |\psi(s^*)|)|\psi'(s)| \leq 4C_\psi C_{\psi'}. \tag{219}$$

For the second term, given z has i.i.d. entries, we have

$$\mathbb{E}[\|z\|^4] = \mathbb{E}\big[\sum_{i=1}^{d} z_i^2\big]^2 \tag{220}$$

$$= \mathbb{E}\big[\big(\sum_{i=1}^{d} z_i^2\big)\big(\sum_{i=1}^{d} z_i^2\big)\big] \tag{221}$$

$$= \mathbb{E}\big[\sum_{i=1}^{d} z_i^4 + \sum_{i \neq j} z_i^2 z_j^2\big] \tag{222}$$

$$= \sum_{i=1}^{d} \mathbb{E}[z_i^4] + \sum_{i \neq j} \mathbb{E}[z_i^2]\mathbb{E}[z_j^2] \tag{223}$$

$$= d\mu_4 + d(d-1)\mu_2^2 \tag{224}$$

$$= \Theta(d^2). \tag{225}$$

Therefore, there exists a unifying constant K such that

$$\mathbb{E}[\|g^{(\theta)}\|^2]| \leq \big[4C_\psi C_{\psi'}\big]^2 \big[\Theta(d^2)\big]^{1/2} = Kd. \tag{226}$$

Similarly, to bound the forth moment, using the Minkowski inequality in L^4 and the i.i.d. of $\{z_i\}$, we have

$$\|\sum_{i=1}^{d} z_i^2\|_{L^4} = \big(\mathbb{E}\big[\sum_{i=1}^{d} z_i^2\big]^4\big)^{1/4} \leq \sum_{i=1}^{d} \big(\mathbb{E}[z_i^8]\big)^{1/4} = d(\mathbb{E}[z_i^8])^{1/4}$$

Raise both sides to the 4th power, we have

$$\mathbb{E}[\|z\|^8] = \mathbb{E}\big(\sum_{i=1}^{d} z_i^2\big)^4 \leq d^4 \mathbb{E}[z_1^8] = O(d^4).$$

Thus, with the same, finite A, there exists a unifying constant K such that

$$\mathbb{E}[\|g^{(\theta)}\|^4] \leq \mathbb{E}[|A|^4 \|z\|^4] \leq \mathbb{E}[|A|^8]^{1/2} \mathbb{E}[\|z\|^8]^{1/2} \leq Kd^2 \tag{227}$$

θ-block directional variance. Fix $(\beta, \theta) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, and condition on $\mathcal{F}_t$, so that (β, θ) are deterministic. For a fresh sample x_t, write

$$u_i := x_t^{(i)} \cdot \beta, \qquad z_i := \phi(u_i), \qquad z := (z_1, \dots, z_d), \qquad s := z \cdot \theta.$$

Also write

$$A_t := 2(f_t - f_t^*)\psi'(s).$$

By the boundedness of ψ and ψ',

$$|A_t| \le C_A.$$

The Euclidean θ-gradient is $A_t z$, hence the spherical θ-gradient is

$$g_t^{(\theta)} = P_{\theta^\perp}(A_t z) = A_t P_{\theta^\perp} z, \qquad P_{\theta^\perp} := I - \theta\theta^\top.$$

Let $v \in \mathbb{R}^d$ be any unit vector. Since $P_{\theta^\perp}$ is self-adjoint,

$$\langle g_t^{(\theta)}, v\rangle = A_t \langle z, P_{\theta^\perp} v\rangle.$$

Set

$$\widetilde{v} := P_{\theta^\perp} v.$$

Then $\|\widetilde{v}\|_2 \le 1$. Since $\mathbb{E}[\phi(G)] = 0$ and $\mathbb{E}[\phi(G)^2] < \infty$, the coordinates $z_i = \phi(u_i)$ are iid centered variables with bounded second moment. Therefore

$$\mathbb{E}\big[\langle z, \widetilde{v}\rangle^2 \,\big|\, \mathcal{F}_t\big] = \mathbb{E}[\phi(G)^2]\|\widetilde{v}\|_2^2 \le C.$$

Using $|A_t| \le C_A$,

$$\mathbb{E}\Big[\langle g_t^{(\theta)}, v\rangle^2 \,\Big|\, \mathcal{F}_t\Big] \le C.$$

Hence

$$\mathbb{E}\Big[\big(\langle g_t^{(\theta)}, v\rangle - \mathbb{E}[\langle g_t^{(\theta)}, v\rangle \mid \mathcal{F}_t]\big)^2 \,\Big|\, \mathcal{F}_t\Big] \le \mathbb{E}\Big[\langle g_t^{(\theta)}, v\rangle^2 \,\Big|\, \mathcal{F}_t\Big] \le C.$$

Equivalently,

$$\sup_{\|v\|_2=1} \mathbb{E}\Big[\big((\Delta g_t^{(\theta)}) \cdot v\big)^2 \,\Big|\, \mathcal{F}_t\Big] \le C.$$

β bounds Recall

$$\nabla_\beta l = 2(f - f^*)\,\psi'(s) \sum_{i=1}^d \theta_i \phi'(u_i)\, x^{(i)}, \qquad u_i := x^{(i)} \cdot \beta, \qquad s := z \cdot \theta,\ z_i = \phi(u_i),$$

and define the scalar

$$A(x; \beta, \theta) := 2(f - f^*)\psi'(s).$$

By boundedness of ψ, ψ', we have the uniform bound

$$|A| \le 2\big(|\psi(s)| + |\psi(s^*)|\big)\,|\psi'(s)| \le 4C_\psi C_{\psi'} =: C_A.$$

Let $P_\beta := I - \beta\beta^\top$ be the orthogonal projector onto $\beta^\perp$. The β-spherical gradient is

$$g^{(\beta)}(\beta, \theta; x) = \nabla_\beta^{\mathcal{S}} l(\beta, \theta; x) = P_\beta \nabla_\beta l = A \sum_{i=1}^d \theta_i \phi'(u_i)\, P_\beta x^{(i)}.$$

Now exploit the Gaussian block structure: each block $x^{(i)} \sim \mathcal{N}(0, I_d)$ independently. Decompose

$$x^{(i)} = u_i \beta + v_i, \qquad u_i := x^{(i)} \cdot \beta \sim \mathcal{N}(0, 1), \qquad v_i := P_\beta x^{(i)} \sim \mathcal{N}(0, P_\beta),$$

and for each i, u_i is independent of v_i (and the pairs (u_i, v_i) are i.i.d. over i). Hence

$$g^{(\beta)} = A \sum_{i=1}^d \theta_i \phi'(u_i)\, v_i.$$

Conditioned on $(u_1, \dots, u_d)$, the vector

$$S := \sum_{i=1}^{d} \theta_i \phi'(u_i)\, v_i$$

is Gaussian with mean 0 and covariance

$$\mathrm{Cov}(S \mid u) = \Big(\sum_{i=1}^{d} \theta_i^2 \phi'(u_i)^2 \Big) P_\beta.$$

Write

$$\sigma^2(u) := \sum_{i=1}^{d} \theta_i^2 \phi'(u_i)^2.$$

Since $\mathrm{rank}(P_\beta) = d - 1$, standard Gaussian moment formulas give

$$\mathbb{E}\big[\|S\|^2 \mid u\big] = (d-1)\sigma^2(u), \qquad \mathbb{E}\big[\|S\|^4 \mid u\big] = (d-1)(d+1)\sigma^4(u).$$

At this point we use the regularity of ϕ. For instance, if ϕ is L_ϕ-Lipschitz (as in our regularity assumptions), then $|\phi'| \le L_\phi$ a.e., so

$$\sigma^2(u) \le L_\phi^2 \sum_{i=1}^{d} \theta_i^2 = L_\phi^2, \qquad \sigma^4(u) \le L_\phi^4.$$

Therefore,

$$\mathbb{E}\|g^{(\beta)}\|^2 = \mathbb{E}\big[|A|^2 \|S\|^2\big] \le C_A^2\, \mathbb{E}\|S\|^2 \le C_A^2 (d-1) L_\phi^2 \le K d,$$

and similarly

$$\mathbb{E}\|g^{(\beta)}\|^4 = \mathbb{E}\big[|A|^4 \|S\|^4\big] \le C_A^4\, \mathbb{E}\|S\|^4 \le C_A^4 (d-1)(d+1) L_\phi^4 \le K d^2.$$

Finally, we bound the directional variance of the centered noise field. Fix any unit vector $q \in \mathbb{R}^d$. Conditioned on u, $(q \cdot S)$ is Gaussian with variance

$$\mathrm{Var}(q \cdot S \mid u) = q^\top \mathrm{Cov}(S \mid u)\, q = \sigma^2(u)\, q^\top P_\beta q \le \sigma^2(u).$$

Thus,

$$\mathbb{E}[(q \cdot g^{(\beta)})^2] = \mathbb{E}\big[|A|^2 (q \cdot S)^2\big] \le C_A^2\, \mathbb{E}[\sigma^2(u)] \le C_A^2 L_\phi^2.$$

Since $\mathrm{Var}(X) \le \mathbb{E}[X^2]$, we get

$$\sup_{\|q\|=1} \mathbb{E}\big[((\Delta g^{(\beta)}) \cdot q)^2\big] \le \sup_{\|q\|=1} \mathbb{E}\big[(g^{(\beta)} \cdot q)^2\big] \le C_A^2 L_\phi^2 \le K,$$

which proves the claimed directional variance bound (absorbing constants into K).

Along the SGD trajectory (β_t, θ_t) is $\mathcal{F}_t$-measurable and x_t is independent of $\mathcal{F}_t$, hence the above bounds apply conditionally on $\mathcal{F}_t$ by evaluation at $(\beta, \theta) = (\beta_t, \theta_t)$. □

E.3 Summation by Parts

The following estimates are deterministic and pathwise; no martingale property is used. They are stated on subintervals because the SGD proof applies them after stopping times.

Lemma E.8 (Summation by parts for monotone weights)**.** *Let $(N_t)_{0 \le t \le T}$ be a real-valued sequence and set*

$$\varepsilon_t := N_{t+1} - N_t.$$

Fix integers $0 \le r < t \le T$, and let $(b_k)_{k=r}^{t-1}$ be monotone with $0 \le b_k \le 1$. If

$$\sup_{0 \le u \le T} |N_u| \le M,$$

then

$$\left| \sum_{k=r}^{t-1} b_k \varepsilon_k \right| \le 2M.$$

Proof. Abel summation gives

$$\sum_{k=r}^{t-1} b_k(N_{k+1}-N_k) = b_{t-1}N_t - b_rN_r - \sum_{k=r}^{t-2} N_{k+1}(b_{k+1}-b_k).$$

Hence

$$\left|\sum_{k=r}^{t-1} b_k\varepsilon_k\right| \le M\left(b_{t-1}+b_r+\sum_{k=r}^{t-2}|b_{k+1}-b_k|\right).$$

If (b_k) is increasing, then

$$\sum_{k=r}^{t-2}|b_{k+1}-b_k| = b_{t-1}-b_r,$$

so the coefficient is $2b_{t-1} \le 2$. If (b_k) is decreasing, then

$$\sum_{k=r}^{t-2}|b_{k+1}-b_k| = b_r - b_{t-1},$$

so the coefficient is $2b_r \le 2$. This proves the claim. □

Corollary E.9 (Summation by parts for decreasing geometric weights)**.** *Let $(N_t)_{0\le t\le T}$ be a real-valued sequence and set*

$$\varepsilon_t := N_{t+1} - N_t.$$

Let $\gamma > 1$. If

$$\sup_{0\le u\le T}|N_u| \le M,$$

then for every $0 \le r < T$ and every $s \in \{1,\dots,T-r\}$,

$$\left|\sum_{k=0}^{s-1}\gamma^{-(k+1)}\varepsilon_{r+k}\right| \le 2M.$$

In particular, if $M = 2\lambda$, the bound is 4λ. If moreover $r = 0$ and $N_0 = 0$, then the bound improves to M.

Proof. Apply Lemma E.8 with

$$b_{r+k} := \gamma^{-(k+1)}, \qquad 0 \le k \le s-1.$$

These weights are decreasing and lie in $[0,1]$, so the first bound follows.

If $r = 0$ and $N_0 = 0$, then the initial boundary term in Abel summation vanishes:

$$\sum_{k=0}^{s-1}\gamma^{-(k+1)}(N_{k+1}-N_k) = \gamma^{-s}N_s - \sum_{k=0}^{s-2}N_{k+1}\left(\gamma^{-(k+2)}-\gamma^{-(k+1)}\right).$$

Therefore

$$\left|\sum_{k=0}^{s-1}\gamma^{-(k+1)}\varepsilon_k\right| \le M\left(\gamma^{-s}+\sum_{k=0}^{s-2}\left(\gamma^{-(k+1)}-\gamma^{-(k+2)}\right)\right) = M\gamma^{-1} \le M.$$

□

Corollary E.10 (Summation by parts for increasing geometric weights)**.** *Let $(N_t)_{0\le t\le T}$ be a real-valued sequence and set*

$$\varepsilon_t := N_{t+1} - N_t.$$

Let $\alpha \in (0,1)$. If

$$\sup_{0\le u\le T}|N_u| \le M,$$

then for every $0 \le r < t \le T$,

$$\left|\sum_{k=r}^{t-1}\alpha^{t-1-k}\varepsilon_k\right| \le 2M.$$

In particular, if $M = \lambda$, the bound is 2λ, and if $M = 2\lambda$, the bound is 4λ.

Proof. Apply Lemma E.8 with

$$b_k := \alpha^{t-1-k}, \qquad r \le k \le t-1.$$

Since $\alpha \in (0,1)$, these weights are increasing in k and lie in $[0,1]$. The claim follows. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide proofs of the main claims and clearly distinguish between claims and open questions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the final conclusion section

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: N/A

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include an appendix with experimental details

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code used to run the experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.