# SYMMETRY-AWARE FEATURE LEARNING: A POLYNOMIAL SEPARATION FOR MULTI-INDEX MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We establish a polynomial sample complexity separation between symmetry-aware and symmetry-agnostic feature learning. We study growing-rank multi-index models with high-dimensional Gaussian covariates in $\mathbb{R}^d$ and $r = \Theta(d^\delta)$ teacher directions forming a cyclic symmetry orbit, where $0 < \delta < 1/2$. We compare three ways of exploiting this structure: architectural weight sharing, data augmentation over the full symmetry group, and learning without access to the symmetry. In particular, we analyze a symmetry-tied convolutional network, an untied network, and the same untied network trained with full-group data augmentation, using spherical online SGD with correlation loss. For a class of polynomial links with information exponent $p \geq 3$, we prove matching sample complexity bounds up to logarithmic factors: the tied and augmented learners achieve weak directional recovery in $\widetilde{\Theta}(d^{p-1})$ samples, whereas the symmetry-agnostic learner requires $\widetilde{\Theta}(rd^{p-1})$. For the pure quadratic Hermite link, the same separation holds for weak recovery of the teacher subspace, with sample complexities $\widetilde{\Theta}(d)$ and $\widetilde{\Theta}(rd)$, respectively. Thus, full-group data augmentation matches the sample efficiency of architectural weight sharing, and both provide a polynomial advantage over training without symmetry. For $p \geq 3$, the proof reveals a two-stage mechanism: fluctuations at initialization select one direction in the teacher orbit, after which localized growth amplifies its overlap to the weak recovery scale while competing overlaps remain near their initialization scale.

## 1 INTRODUCTION

Symmetry is one of the most effective inductive biases in machine learning. When transformations of the input preserve the task, the same feature need not be learned independently at every transformed location. Convolutional neural networks exploit this principle through weight sharing, while group-equivariant architectures extend it to more general transformations (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Bronstein et al., 2021). More broadly, geometric deep learning views symmetry and geometry as organizing principles for designing neural architectures (Bronstein et al., 2021).

There are two common ways to exploit this prior knowledge. One is *architectural*: constrain the model through weight sharing or equivariance. The other is *algorithmic*: keep the model unconstrained but train on transformed copies of each observation, as in data augmentation. Group averaging provides a natural mathematical description of the latter (Chen et al., 2020; Lyle et al., 2020). A substantial literature has shown that invariance can improve variance, complexity, and generalization (Elesedy & Zaidi, 2021; Mei et al., 2021), and has quantified advantages of convolutional structure in infinite-width regimes (Cagnetta et al., 2023). Here we study how symmetry changes the sample complexity of feature learning by stochastic gradient descent (SGD). In particular, we ask

> *How much does symmetry reduce the number of samples that SGD needs to discover a relevant feature, and how do weight sharing and data augmentation compare?*

Answering this question mathematically requires analyzing the high-dimensional, nonconvex SGD dynamics from random initialization, rather than comparing hypothesis classes or estimators after optimization.



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

**Feature learning in single- and multi-index models —** To study this question, we focus on single- and multi-index models of the form

$$f^*(x) = g\left(\langle w_1^*, x\rangle, \dots, \langle w_r^*, x\rangle\right),$$

where $g\colon \mathbb{R}^r \to \mathbb{R}$ is the link function and $w_1^*, \dots, w_r^*$ are the latent directions. These models have been used to study feature recovery, computational barriers, and the dynamics of gradient-based learning, see e.g. Dudeja & Hsu (2018); Barbier et al. (2019); Ben Arous et al. (2021); Arnaboldi et al. (2023); Abbé et al. (2023); Dandi et al. (2024b); Troiani et al. (2025); Şimşek et al. (2025). In particular, for Gaussian single-index models, the degree of the first nonzero Hermite coefficient of the link function–its information exponent–determines the time scale on which online SGD escapes the nearly uninformative initialization regime (Ben Arous et al., 2021).

Interactions between several latent directions produce richer dynamics. Finite-rank analyses of online SGD show that competing correlations can suppress or select one another, leading to sequential feature recovery. In multi-spiked tensor PCA, Ben Arous et al. (2026) characterize this through a greedy maximum selection mechanism, which determines the order in which latent directions are recovered. Related selection mechanisms have also been identified in shallow multi-index models with growing numbers of features (Ren et al., 2025), while other recent works study growing collections of approximately independent or diverse features (Oko et al., 2024; Ren & Lee, 2025). Our setting combines the difficulty of interacting directions with that of growing rank: the number of competing directions grows with the dimension, while the directions are strongly dependent because they are symmetry-related copies of a single hidden feature. This structure raises a question that is absent from fixed-rank models: can a learner avoid paying separately for every direction by exploiting their common origin? A more detailed literature review is given in Appendix A.

**A symmetric growing-rank model —** We study a growing-rank model designed to isolate this effect:

$$f^*(x) = \frac{1}{\sqrt{r}} \sum_{k=0}^{r-1} \sigma\left(\langle \Pi^k w^*, x\rangle\right),$$

where $\Pi$ generates a group action of size $r$. Thus the $r$ latent directions $w^*, \Pi w^*, \dots, \Pi^{r-1} w^*$ are not independent parameters: they form the orbit of a single direction $w^*$. For cyclic shifts, this is a convolutional teacher. We consider the growing-rank regime $r \asymp d^\delta$ with $0 < \delta < \frac{1}{2}$.

We analyze two regimes for the link function $\sigma$. Our main results concern polynomial links with information exponent $p \geq 3$. We also treat the pure quadratic Hermite link $\sigma = h_2$, for which the individual orbit directions are generally not identifiable and the natural recovery target is instead the subspace spanned by the teacher orbit.

We compare three ways of learning this target with spherical online SGD. The *tied* procedure (T) uses a single trainable weight shared across its entire orbit. The *augmented* procedure (A) uses an untied two-layer model but averages each sample loss over the full group orbit. The *untied* procedure (U) uses the same untied model without augmentation. The latter two therefore have the same parametrization; their only difference is whether the symmetry is used during training or not.

**An $r$-fold separation —** Our main result shows that exploiting the symmetry reduces the sample complexity of feature recovery by a factor $r$. For polynomial links with information exponent $p \geq 3$, we consider *weak directional recovery*: some learned weight must reach a fixed, nonvanishing correlation with an element of the identifiable teacher orbit. For the dimension-dependent step sizes prescribed in our analysis, we prove matching upper and lower bounds (up to logarithmic factors):

$$n_{\mathsf{T}} = \widetilde{\Theta}(d^{p-1}), \qquad n_{\mathsf{A}} = \widetilde{\Theta}(d^{p-1}), \qquad n_{\mathsf{U}} = \widetilde{\Theta}(r d^{p-1}).$$

The scale $\widetilde{\Theta}(d^{p-1})$ is the weak-recovery scale of vanilla online SGD for Gaussian single-index models with information exponent $p$ (Ben Arous et al., 2021; Damian et al., 2023). Thus, despite the teacher rank growing as $r \asymp d^\delta$, weight sharing and exact data augmentation recover a teacher feature at the single-index scale, whereas the symmetry-agnostic untied learner requires an additional factor equal to the number of orbit directions. Since $r \asymp d^\delta$, this yields a polynomial separation in the dimension. We further show that, with width $s \gtrsim r \log r$, the untied and augmented procedures cover the entire teacher orbit within their respective weak recovery sample complexity scales.

For the pure Hermite-2 link, individual orbit directions are generally unidentifiable. Thus, we instead consider *weak subspace recovery*: some learned weight must acquire a fixed, nonvanishing fraction of its squared norm in the subspace spanned by the teacher orbit. We obtain the same $r$-fold separation: the tied and augmented procedures require $\widetilde{\Theta}(d)$ samples, whereas the untied learner requires $\widetilde{\Theta}(rd)$.

**Proof mechanism —** For $p \geq 3$, the proof reveals a two-step *selection-and-growth mechanism*. At random initialization, one of the $r \asymp d^{\delta}$ correlations with the teacher orbit is slightly larger than the others. During an initial microscopic phase, the advantage of this leading correlation over the competing correlations is amplified. In a subsequent macroscopic phase, the selected correlation grows to a macroscopic level while the competitors remain microscopic. For the pure quadratic Hermite link, individual orbit directions are not identifiable. Rather than tracking a single leading correlation, we track the squared norm of the projection onto the teacher subspace, which undergoes multiplicative growth.

The three procedures differ in how symmetry affects these dynamics. Weight sharing ties all transformed copies of the feature to a single trainable parameter, whereas data augmentation averages the stochastic gradient over the orbit. Although augmentation leaves the population objective unchanged, orbit averaging reduces stochastic fluctuations. With the corresponding learning-rate rescaling, the tied and augmented procedures have the same effective correlation dynamics. In the symmetry-agnostic untied procedure, the corresponding effective dynamics are slower by a factor $r$, yielding the $r$-fold separation in both recovery regimes. See Section 4 for a fuller proof overview.

## 2 SETTING

**Teacher model —** Let $d \geq 1$, and let $r \geq 1$ be an integer dividing $d$. Define $\Delta \coloneqq d/r$, and let $\Pi \in \mathbb{R}^{d\times d}$ denote the permutation matrix corresponding to a cyclic shift by $\Delta$ coordinates. Thus $\Pi^r = \mathrm{I}_d$, and the cyclic group generated by $\Pi$ is

$$\mathcal{G} \coloneqq \left\{\mathrm{I}_d, \Pi, \ldots, \Pi^{r-1}\right\}.$$

We observe a sequence of samples $(x_i, y_i)_{i=1}^n$, where $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathrm{I}_d)$ and $y_i = f^*(x_i)$, where

$$f^*(x) \coloneqq \frac{1}{\sqrt{r}} \sum_{k=0}^{r-1} \sigma\left(\langle \Pi^k w^*, x\rangle\right). \tag{1}$$

Here $w^* \in \mathbb{S}^{d-1}$ is the teacher direction and $\sigma\colon \mathbb{R} \to \mathbb{R}$ is a known nonlinear link function. The teacher depends on $w^*$ only through its orbit $\{\Pi^k w^*\colon 0 \leq k < r\}$. Thus the statistical model is invariant under $w^* \mapsto \Pi^k w^*$.

We analyze two regimes for the link function.

**Assumption 1** (Link function)**.** *The link $\sigma$ belongs to one of the following two classes.*

*(i)* ***Polynomial links with information exponent*** *$p \geq 3$. The link is a fixed polynomial with finite Hermite expansion*

$$\sigma(z) = \sum_{j=p}^{P} a_j h_j(z), \qquad 3 \leq p \leq P < \infty, \qquad a_p \neq 0,$$

*where $(h_j)_{j\geq 0}$ is the orthonormal Hermite basis of $L^2(\mathcal{N}(0,1))$. The integers $p, P$ and the coefficients $(a_j)_{j=p}^P$ are fixed independently of $d$. If $p$ is even, we aslo assume that $\sigma$ is even.*

*(ii)* ***Pure quadratic link.***

$$\sigma(z) = h_2(z) = \frac{1}{\sqrt{2}}(z^2 - 1).$$

In case of Assumption 1(i), we call $p = p_\sigma \coloneqq \min\{j \geq 1\colon a_j \neq 0\}$ the *information exponent* of the link. Since $p \geq 3$, the Hermite coefficients of degrees $1, \ldots, p-1$ vanish.

*Remark* 2.1 (On the link assumptions). The finite Hermite expansion in Assumption 1(i) is mainly technical and could be relaxed under suitable regularity and moment-growth conditions on the link. The parity condition for $p$ ensures that the symmetry of the leading-order dynamics agrees with the exact symmetry of the statistical model. This condition can be removed by considering paired opposite initializations; we do not establish this extension here.

For $u \in \mathbb{S}^{d-1}$, define the correlations

$$m_k(u) \coloneqq \left\langle \Pi^k w^*, u \right\rangle, \qquad 0 \le k < r. \tag{2}$$

The teacher direction $w^*$ is identifiable only up to cyclic shifts. When $\sigma$ is even, there is an additional global sign ambiguity. We therefore define the *identifiable teacher orbit* by

$$\mathcal{O}_\sigma(w^*) \coloneqq \begin{cases} \{\pm \Pi^k w^* : 0 \le k < r\}, & \sigma \text{ is even}, \\ \{\Pi^k w^* : 0 \le k < r\}, & \text{otherwise}. \end{cases}$$

For $u \in \mathbb{S}^{d-1}$, define its correlation with the identifiable teacher orbit by

$$M_\sigma(u) \coloneqq \max_{v \in \mathcal{O}_\sigma(w^*)} \langle v, u \rangle. \tag{3}$$

The quadratic case in Assumption 1(ii) is qualitatively different: the individual directions in the teacher orbit are generally not identifiable. Accordingly, weak recovery in this case will instead be formulated in terms of the subspace spanned by the teacher orbit (see Definition 3.6).

**Learning procedures —** We compare three learning procedures: a tied convolutional student, an untied student, and the same untied student trained with data augmentation. Throughout, let

$$\ell(y, z) \coloneqq -yz$$

denote the correlation loss.

**Definition 2.2** (Tied convolutional student)**.** The tied student is parametrized by the weight $w \in \mathbb{S}^{d-1}$ and is defined by

$$f_w^{\mathsf{T}}(x) \coloneqq \frac{1}{\sqrt{r}} \sum_{k=0}^{r-1} \sigma\left(\langle \Pi^k w, x \rangle\right).$$

Its sample loss is

$$L^{\mathsf{T}}(w; x, y) \coloneqq \ell\left(y, f_w^{\mathsf{T}}(x)\right).$$

**Definition 2.3** (Untied student)**.** Let $W = (w^1, \dots, w^s) \in (\mathbb{S}^{d-1})^s$ with $s = s(d) \ge 1$. The untied student is parametrized by $W$ and defined by

$$f_W^{\mathsf{U}}(x) \coloneqq \frac{1}{\sqrt{s}} \sum_{j=1}^{s} \sigma\left(\langle w^j, x \rangle\right).$$

Its sample loss is

$$L^{\mathsf{U}}(W; x, y) \coloneqq \ell\left(y, f_W^{\mathsf{U}}(x)\right).$$

Unlike the tied student, $w^1, \dots, w^s$ are not constrained to belong to a common orbit under $\mathcal{G}$.

**Definition 2.4** (Untied student with data augmentation)**.** The augmented procedure uses the same model class $f_W^{\mathsf{U}}$, but replaces the sample loss by its group average:

$$L^{\mathsf{A}}(W; x, y) \coloneqq \frac{1}{r} \sum_{k=0}^{r-1} \ell\left(y, f_W^U(\Pi^k x)\right).$$

Set $s_{\mathsf{T}} \coloneqq 1$ and $s_{\mathsf{U}} = s_{\mathsf{A}} \coloneqq s \in \mathbb{N}$, with $s \ge 1$. For $\mathsf{k} \in \{\mathsf{T}, \mathsf{U}, \mathsf{A}\}$, define $\mathcal{M}_{\mathsf{k}} \coloneqq (\mathbb{S}^{d-1})^{s_{\mathsf{k}}}$, so that

$$\mathcal{M}_{\mathsf{T}} = \mathbb{S}^{d-1}, \qquad \mathcal{M}_{\mathsf{U}} = \mathcal{M}_{\mathsf{A}} = (\mathbb{S}^{d-1})^s.$$

For $\theta \in \mathcal{M}_{\mathsf{k}}$, let

$$\Phi^{\mathsf{k}}(\theta) \coloneqq \mathbb{E}\left[L^{\mathsf{k}}(\theta; x, y)\right] \tag{4}$$

denote the corresponding population objective. Since the Gaussian input distribution is invariant under $\Pi$ and the teacher satisfies $f^*(\Pi^k x) = f^*(x)$, the augmented and unaugmented untied procedures have the same population objective:

$$\Phi^{\mathsf{A}}(W) = \Phi^{\mathsf{U}}(W).$$

Thus data augmentation leaves the population drift unchanged.

**Online spherical SGD —** For $\theta = (w^1, \dots, w^{s_k}) \in \mathcal{M}_k$, write

$$\nabla_{w^j} L(\theta) \coloneqq (\mathrm{I}_d - w^j w^{j\top}) \widehat{\nabla}_{w^j} L(\theta)$$

for the spherical gradient, where $\widehat{\nabla}_{w^j}$ denotes the Euclidean gradient. Given an initial condition $\theta_0 \in \mathcal{M}_k$ and a step size $\eta_d^k > 0$, single-pass spherical SGD is defined componentwise by

$$\begin{cases} \widetilde{w}_t^j = w_{t-1}^j - \eta_d^k \nabla_{w^j} L^k(\theta_{t-1}; x_t, y_t), \\ w_t^j = \dfrac{\widetilde{w}_t^j}{\|\widetilde{w}_t^j\|}, \end{cases} \qquad 1 \le j \le s_k. \tag{5}$$

For the tied procedure we identify $\theta_t = w_t^{1,\mathsf{T}}$ with $w_t$. For the untied and augmented procedures, $\theta_t = W_t = (w_t^1, \dots, w_t^s)$. At each iteration, the algorithm uses one new independent observation $(x_t, y_t)$. In the augmented procedure, this same observation additionally generates the $r$ transformed inputs $\Pi^k x_t$ for $0 \le k < r$. Hence the number of SGD iterations equals the number of fresh samples, although augmentation incurs a larger per-iteration computational cost.

## 3 Main results

We now state our main results on weak recovery by online spherical SGD. We work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting the teacher direction $w^*$, all parameter initializations, and an i.i.d. sequence of Gaussian covariates $(x_i)_{i \ge 1}$ used by online SGD.

**Assumption 2** (Random teacher and initialization)**.** *The teacher direction is sampled according to* $w^* \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. *For each learning procedure* $\mathsf{k} \in \{\mathsf{T}, \mathsf{U}, \mathsf{A}\}$*, initialize*

$$w_0^{1,\mathsf{k}}, \dots, w_0^{s_k,\mathsf{k}} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1}),$$

*where* $s_\mathsf{T} = 1$ *and* $s_\mathsf{U} = s_\mathsf{A} = s$*. The teacher, all parameter initializations, and the Gaussian covariates* $(x_i)_{i \ge 1}$ *are mutually independent.*

Unless stated otherwise, $\mathbb{P}$ denotes probability with respect to all of these sources of randomness. We work throughout in the following high-dimensional regime.

**Assumption 3** (High-dimensional regime)**.** *Fix* $0 < \delta < \frac{1}{2}$ *and* $\kappa \ge 0$*. We work along a sequence of dimensions* $d \to \infty$ *for which*

$$r = r(d) = d^\delta, \qquad 1 \le s = s(d) \le d^\kappa,$$

*with* $r(d)$ *an integer divisor of* $d$*.*

**Main results for $p \ge 3$ —** We first consider Assumption 1(i), namely polynomial links with information exponent $p \ge 3$. For $\rho \in (0, 1)$ and $\mathsf{k} \in \{\mathsf{T}, \mathsf{U}, \mathsf{A}\}$, define the event

$$\mathcal{R}_n^k(\rho) \coloneqq \left\{ \max_{0 \le t \le n} \max_{1 \le j \le s_k} M_\sigma(w_t^{j,\mathsf{k}}) \ge \rho \right\},$$

where $s_\mathsf{T} = 1$, $s_\mathsf{U} = s_\mathsf{A} = s$, and $w_t^{1,\mathsf{T}} = w_t^\mathsf{T}$.

**Definition 3.1** (Weak recovery)**.** Fix $\rho \in (0, 1)$, independently of $d$. We say that procedure $\mathsf{k} \in \{\mathsf{T}, \mathsf{U}, \mathsf{A}\}$ achieves *weak recovery at level $\rho$ within $n = n(d)$ samples* if

$$\mathbb{P}\big(\mathcal{R}_n^k(\rho)\big) \to 1 \qquad \text{as } d \to \infty.$$

Thus, weak recovery requires the tied procedure's single weight, or at least one weight of the untied or augmented procedure, to attain a fixed nonvanishing correlation with the identifiable teacher orbit. When $p$ is even, recovery is only up to cyclic shift and global sign, since both are exact symmetries of the model. When $p$ is odd, only the cyclic shift remains unidentifiable, and recovery requires positive correlation with the teacher orbit.

For $\mathsf{k} \in \{\mathsf{T}, \mathsf{U}, \mathsf{A}\}$, let $n_\mathsf{k}(\rho)$ denote the sample complexity of achieving weak recovery at level $\rho$ in the sense of Definition 3.1. Throughout, $\widetilde{\Theta}$ denotes matching upper and lower sample complexity bounds up to logarithmic factors in $d$. The upper bounds give recovery with probability $1 - o(1)$ by the stated scale, whereas below the lower-bound scale, recovery does not occur with probability $1 - o(1)$. Our main results identify the sample complexity for the three procedures.

**Theorem 3.2** (Sample complexity of weak recovery, $p \geq 3$)**.** *Suppose that Assumptions 1(i), 2, and 3 hold. Fix $\rho \in (0,1)$, independently of $d$. Then there exists $A_* = A_*(p,P) > 0$ such that, for every fixed $A > A_*$, the following holds. Set*

$$\eta_d = d^{-p/2}(\log d)^{-A},$$

*and run the three online spherical SGD procedures* (5) *with step sizes*

$$\eta_d^{\mathsf{T}} = \eta_d, \qquad \eta_d^{\mathsf{U}} = \sqrt{\frac{s}{r}}\,\eta_d, \qquad \eta_d^{\mathsf{A}} = \sqrt{rs}\,\eta_d. \tag{6}$$

*Then the sample complexities for weak recovery at level $\rho$ satisfy*

$$n_{\mathsf{T}}(\rho),\ n_{\mathsf{A}}(\rho) = \widetilde{\Theta}\left(\frac{d^{(p-2)/2}}{\eta_d}\right) = \widetilde{\Theta}(d^{p-1}), \qquad n_{\mathsf{U}}(\rho) = \widetilde{\Theta}\left(\frac{r\,d^{(p-2)/2}}{\eta_d}\right) = \widetilde{\Theta}(rd^{p-1}).$$

Thus, up to logarithmic factors, weight tying and exact data augmentation achieve weak recovery at the same $d^{p-1}$ scale as vanilla Gaussian single-index online SGD (Ben Arous et al., 2021; Damian et al., 2023), while the unaugmented untied procedure requires an additional factor $r$. The upper bounds in Theorem 3.2 are stronger than Definition 3.1 requires: any fixed weight of the untied or augmented procedure achieves weak recovery at the corresponding sample scale, so the proof does not rely on selecting the best among $s$ initializations. Conversely, the lower bounds hold uniformly over all $s$ weights. The proof combines the upper bounds of Theorem B.1 with the lower bounds of Theorem B.4; see Appendix B.

Figure 1 illustrates the predicted separation empirically for the pure Hermite-4 link at $\delta = 1/2$. Since the upper bounds hold for any fixed weight, we set $s = 1$ and use the corresponding rescaled learning rates with a common initialization across procedures. This isolates the single-weight dynamics and controls for differences in initial alignment that can strongly affect recovery times at small dimensions but do not change the asymptotic sample complexity scales established above.

*Remark* 3.3 (Choice of learning rates). The reference rate $\eta_d = d^{-p/2}(\log d)^{-A}$ has the largest polynomial order supported by our perturbative analysis. Up to logarithmic factors, the accumulated stochastic and normalization errors, relative to the relevant growth scale, are controlled by $(\eta_d d^{p/2})^{1/2}$ and $\eta_d d^{p/2}$, respectively. Choosing the constant $A$ sufficiently large makes both contributions negligible and allows comparison with the population dynamics.

For the tied student, recovery of one orbit direction automatically represents the entire orbit through weight sharing. For the untied and augmented students, weak recovery guarantees only that one weight aligns with one orbit direction. We therefore ask the stronger question of whether the collection of $s$ weights can recover every orbit direction. Define the relevant correlation

$$q_k(w) := \begin{cases} |m_k(w)|, & p \text{ even}, \\ m_k(w), & p \text{ odd}. \end{cases} \tag{7}$$

For $\mathsf{k} \in \{\mathsf{U}, \mathsf{A}\}$, define the full-orbit coverage event

$$\mathcal{C}_n^{\mathsf{k}}(\rho) := \left\{ \min_{0 \leq k < r} \max_{1 \leq j \leq s_{\mathsf{k}}} \max_{0 \leq t \leq n} q_k(w_t^{j,\mathsf{k}}) \geq \rho \right\}.$$

**Definition 3.4** (Weak orbit coverage)**.** We say that procedure $\mathsf{k} \in \{\mathsf{U}, \mathsf{A}\}$ *weakly covers the full teacher orbit at level $\rho$ within $n = n(d)$ samples* if

$$\mathbb{P}\big(\mathcal{C}_n^{\mathsf{k}}(\rho)\big) \to 1.$$

Write $n_{\mathsf{k}}^{\mathrm{cov}}(\rho)$ for the corresponding sample complexity scale.

Equivalently, Definition 3.4 requires every identifiable teacher-orbit direction to be recovered by at least one weight at some time before $n$. With sufficiently large width $s$, this stronger requirement can be achieved without changing the sample complexity scale.

**Corollary 3.5** (Weak orbit coverage)**.** *Under the assumptions and learning rates of Theorem 3.2, fix $\varepsilon > 0$ and suppose $s \geq (1+\varepsilon) r \log r$. Then*

$$n_{\mathsf{A}}^{\mathrm{cov}}(\rho) = \widetilde{\Theta}(d^{p-1}), \qquad n_{\mathsf{U}}^{\mathrm{cov}}(\rho) = \widetilde{\Theta}(rd^{p-1}).$$

Thus, once the width is sufficient to cover the $r$ orbit directions, requiring recovery of the entire orbit does not increase the sample complexity scale relative to weak recovery of a single direction. The proof is given in Appendix F.4.

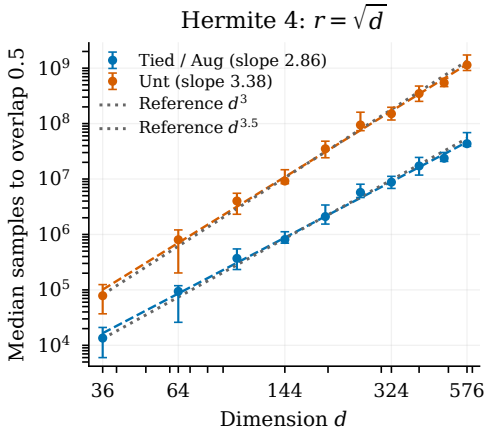


Figure 1: **Recovery-time scaling.** Spherical online SGD with correlation loss, Hermite-4 link, and $r = \sqrt{d}$ symmetries. Points show median samples to reach $M_\sigma(w) \geq 0.5$ over 10 paired seeds; bars are pointwise 95% bootstrap intervals. Colored dashed lines show power-law fits; gray dotted lines show reference powers. We report equivalent single-weight dynamics with $s_\mathsf{T} = s_\mathsf{U} = s_\mathsf{A} = 1$, shared initialization and samples, $\eta_d = 0.1d^{-2}$, and the rescalings in (6). Tied and augmented trajectories coincide.

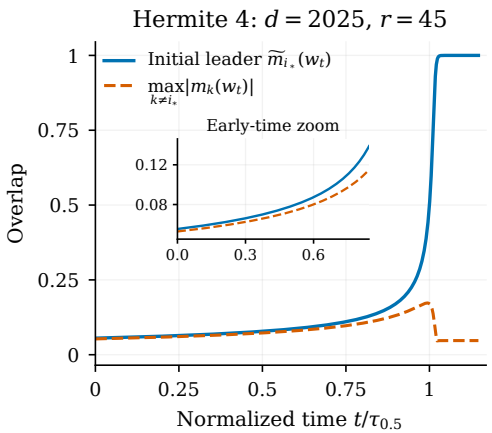


Figure 2: **Selection of an orbit direction in population flow.** Population gradient flow with correlation loss for the tied student with the Hermite-4 link, dimension $d = 2025$, and cyclic symmetries $r = \sqrt{d} = 45$. Solid blue follows the correlation $\widetilde{m}_{i_*}(w_t) := \mathrm{sgn}(m_{i_*}(w_0))m_{i_*}(w_t)$ with index $i_* = \arg\max_k |m_k(w_0)|$. Dashed orange shows $\max_{k \neq i_*} |m_k(w_t)|$. The inset magnifies the early gap amplification. Time is normalized by $\tau_{0.5}$, the first time the population flow reaches orbit correlation 0.5.

**Main results for $p = 2$ —** We next consider Assumption 1(ii), namely the pure quadratic link. In this case the individual directions in the teacher orbit are generally not identifiable, and the natural recovery target is instead the subspace

$$\mathcal{S}^* := \mathrm{span}\{\Pi^k w^* : 0 \leq k < r\}.$$

For $w \in \mathbb{S}^{d-1}$, let $w^\parallel$ denote its orthogonal projection onto $\mathcal{S}^*$. Since $\|w\| = 1$, the quantity $\|w^\parallel\|^2$ measures the fraction of the squared norm of $w$ contained in the teacher subspace.

**Definition 3.6** (Weak subspace recovery)**.** Fix $\rho \in (0, 1)$, independently of $d$. We say that procedure $\mathsf{k} \in \{\mathsf{T}, \mathsf{U}, \mathsf{A}\}$ achieves *weak subspace recovery at level $\rho$ within $n = n(d)$ samples* if

$$\mathbb{P}\left(\max_{0 \leq t \leq n} \max_{1 \leq j \leq s_\mathsf{k}} \left\|(w_t^{j,\mathsf{k}})^\parallel\right\|^2 \geq \rho\right) \to 1 \qquad \text{as } d \to \infty,$$

where $s_\mathsf{T} = 1$, $s_\mathsf{U} = s_\mathsf{A} = s$, and $w_t^{1,\mathsf{T}} = w_t^\mathsf{T}$.

Let $n_\mathsf{k}^{(2)}(\rho)$ denote the sample complexity of weak subspace recovery, with upper and lower bounds interpreted as for $n_\mathsf{k}(\rho)$.

**Theorem 3.7** (Sample complexity of weak subspace recovery)**.** *Suppose Assumption 1(ii), 2, and 3 hold. Fix $\rho \in (0, 1)$, independently of $d$. Then there exists a universal constant $A_0 > 0$ such that, for every fixed $A > A_0$, the following holds. Set*

$$\eta_d = d^{-1}(\log d)^{-A}, \tag{8}$$

*and use the same relative learning-rate scaling across the three procedures as in* (6)*. Then the sample complexities for weak subspace recovery at level $\rho$ satisfy*

$$n_\mathsf{T}^{(2)}(\rho),\ n_\mathsf{A}^{(2)}(\rho) = \widetilde{\Theta}\left(\frac{1}{\eta_d}\right) = \widetilde{\Theta}(d), \qquad n_\mathsf{U}^{(2)}(\rho) = \widetilde{\Theta}\left(\frac{r}{\eta_d}\right) = \widetilde{\Theta}(rd).$$

Thus, the pure quadratic case exhibits the same $r$-fold separation between symmetry-aware and symmetry-agnostic learning as for $p \geq 3$: weight tying and data augmentation achieve weak subspace recovery at scale $\widetilde{\Theta}(d)$, whereas the untied procedure requires $\widetilde{\Theta}(rd)$ samples. The proof is provided in Appendix B, combining the upper bounds of Theorem B.3 with the lower bounds of Theorem B.4.

*Remark* 3.8 (General links with information exponent 2). For general centered polynomial links with information exponent 2, our lower bound still applies, but our matching upper bound does not: we prove the upper bound only for the pure quadratic link. Higher-order Hermite components can change the appropriate recovery notion and the resulting dynamics. Related results for untied models have been obtained by Ren & Lee (2025).

## 4 Proofs overview

The proofs of our main results are based on tracking the correlations $m_k(w_t)$ defined in (2) between the SGD iterates and the teacher orbit. For $p \geq 3$, we track the relevant correlations $q_k(w_t)$ from (7), so that

$$M_\sigma(w_t) = \max_{0 \leq k < r} q_k(w_t).$$

Recall that $q_k = m_k$ when $p$ is odd and $q_k = |m_k|$ when $p$ is even. We show that random initialization produces, with high probability, a unique leading relevant correlation separated from its competitors by a gap large enough to dominate the error, and that SGD dynamics preserves and amplifies this gap.

**Initialization, interactions, and correlation growth —** At random initialization, each correlation $m_k(w_0)$ is typically of order $d^{-1/2}$. Since the teacher orbit is nearly orthogonal, the correlation vector is approximately Gaussian with covariance close to $d^{-1}\mathrm{I}_r$. In particular, its coordinates behave approximately as independent $N(0, d^{-1})$ variables, and their largest relevant value $q_k(w_0)$ is of order

$$\gamma_d \coloneqq \sqrt{\frac{\log r}{d}}.$$

Moreover, with high probability, there is a unique leading relevant correlation, separated from all competing correlations by a quantifiable gap. This initial gap breaks the symmetry between the teacher directions and is then amplified by the SGD dynamics. Hermite orthogonality gives

$$\Phi^{\mathsf{T}}(w) = -\sum_{k=0}^{r-1} \phi(m_k(w)), \qquad \phi(z) = \sum_{\ell=p}^{P} a_\ell^2 z^\ell,$$

with $\phi'(m) = p a_p^2 m^{p-1} + O(|m|^p)$ near the origin. Writing $m_k = m_k(w)$, the spherical population drift of the $k$th correlation is

$$-\langle \Pi^k w^*, \nabla \Phi^{\mathsf{T}}(w) \rangle = (1 - m_k^2)\phi'(m_k) + \sum_{i \neq k} \langle \Pi^k w^*, \Pi^i w^* \rangle \phi'(m_i) - m_k \sum_{i \neq k} m_i \phi'(m_i).$$

The first term gives the self-amplifying growth of $m_k$, while the remaining terms describe the interactions with the other teacher directions and the effect of the spherical constraint.

The main technical difficulty is to control these interaction terms uniformly when the number $r$ of teacher directions grows with $d$. Together with the stochastic fluctuations and normalization errors, we show that their cumulative effect remains smaller than the initialization gap throughout the microscopic phase. As a result, the relevant correlation that is largest at initialization remains the leader: no competitor overtakes it, and its advantage grows over time. The dynamics can therefore be reduced to an effectively 1D growth mechanism along the selected teacher direction (see Fig. 2).

**From microscopic selection to weak recovery —** Introduce the intermediate scales $\alpha_d, \beta_d$ satisfying

$$\gamma_d \ll \alpha_d \ll \beta_d \ll 1.$$

During the first, microscopic phase, all correlations grow but at different rates; thus, the initial gap between the leader and its competitors widens. As a result, the correlation that is largest at initialization reaches the threshold $\beta_d$, while all competing correlations remain at scale $O(\alpha_d)$. This phase takes $O((\eta_d \gamma_d^{p-2})^{-1})$ steps. During the second phase, the population drift drives the

selected correlation from $\beta_d$ to the weak recovery level $\rho \in (0,1)$, while the spherical dynamics keep the competing correlations small. This phase is faster and takes $O((\eta_d \beta_d^{p-2})^{-1})$ steps. Figure 3 illustrates the sketch of proof with this two-phase argument and Appendix E gives the proof for the tied procedure.

**Untied and augmented procedures —** For the untied and augmented learners, it is enough for the upper bound to follow one fixed weight $w^j$. Under the correlation loss, the population objective is additive across weights, so the gradient with respect to $w^j$ depends only on $w^j$:

$$\nabla_{w^j} \Phi^{\mathsf{U}}(W) = \nabla_{w^j} \Phi^{\mathsf{A}}(W) = \frac{1}{\sqrt{rs}} \nabla \Phi^{\mathsf{T}}(w^j).$$

Hence, the same correlation argument applies weight by weight. At the prescribed learning rates, an augmented weight has the same effective drift and fluctuation scales as the tied trajectory. For an untied weight, the effective drift is smaller by a factor $r$, while the stochastic fluctuations are smaller only by $\sqrt{r}$. The same selection-and-growth mechanism therefore takes $r$ times as many samples, producing the separation in Theorem 3.2. Appendix F makes these reductions precise.

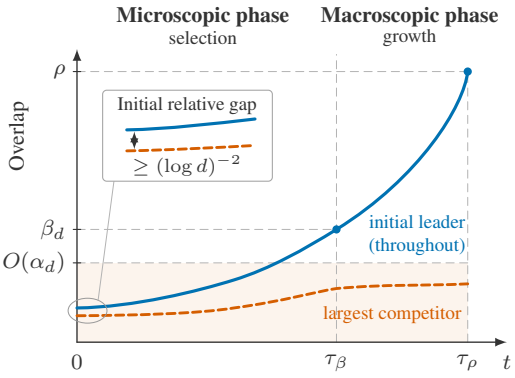


Figure 3: **Sketch proof: two phases of recovery.** Blue follows the initial leader; orange shows the largest competitor, whose index may change. Their initial gap widens during both phases. The shaded band bounds competing absolute overlaps. Axes and trajectories are schematic; the shorter macroscopic phase is expanded for visibility.

**Subspace recovery for the pure quadratic link —** For $\sigma = h_2$, we seek to grow $\|w_t^{\|}\|^2$ to a fixed positive level. Because the teacher orbit is nearly orthonormal, the squared norm of the projection onto its span is, up to a $1 + o(1)$ factor, the sum of the squared correlations:

$$\|w_t^{\|}\|^2 = (1 + o(1)) \sum_{k=0}^{r-1} m_k(w_t)^2.$$

We therefore track this sum, denoted by $R_t$, with $R_0 \asymp r/d$. Its population drift produces multiplicative growth. Controlling the accumulated stochastic and normalization errors shows that $\|w_t^{\|}\|^2$ reaches the weak recovery level $\rho \in (0,1)$ within $O(\eta_d^{-1} \log(d/r))$ samples for the tied and augmented procedures, and $r$ times as many for the untied procedure (see Appendix G).

**Matching lower bounds —** For every $p \geq 2$, our drift and noise estimates imply that, with high probability, all correlations remain $O(\gamma_d)$ for at least $c/(\eta_d \gamma_d^{p-2})$ steps for the tied and augmented procedures, and for $r$ times as many steps under the untied procedure. Here $c > 0$ is sufficiently small. Since $r\gamma_d^2 = o(1)$, this excludes both directional and subspace recovery, uniformly over polynomially many weights (see Appendix H).

## 5 Conclusion

We established an $r$-fold separation, up to logarithmic factors, in the sample complexity of symmetry-aware and symmetry-agnostic feature learning. For spherical online SGD with correlation loss, this separation holds for weak directional recovery when $p \geq 3$ and weak subspace recovery for the pure quadratic Hermite link. In this setting, full-group data augmentation matches the statistical benefit of weight sharing: exploiting symmetry through either the architecture or the training procedure improves the balance between informative drift and stochastic fluctuations. We conjecture that the same separation persists for $d^{1/2} \lesssim r \ll d$, even though the initial largest correlation no longer remains dominant until weak recovery, and a different proof mechanism is therefore needed.

## AI USE STATEMENT

We used generative AI tools to assist with formulating and refining mathematical statements, exploring proof strategies, and drafting and revising parts of the proofs. We also used AI assistance to identify relevant literature and references, improve the organization and exposition of the manuscript, and prepare and revise text, figure captions, and layouts. AI assistance was also used to develop and debug the experimental and analysis code and to prepare plots from the numerical results. The mathematical arguments and exposition were developed through iterative discussion, critical review, and substantive revision by the authors. The authors take responsibility for the final content of this work, including all AI-assisted mathematical claims, proofs, text, and artifacts.

## ETHICS STATEMENT

This work studies the sample complexity of learning symmetric multi-index models through theoretical analysis and numerical experiments on synthetic data. It involves no human participants or personal data. We do not identify specific ethical concerns or direct negative societal impacts arising from the results presented here.

## REPRODUCIBILITY STATEMENT

Proofs of our theoretical results are provided in the appendices. Upon acceptance, we will publicly release the code and configurations needed to reproduce our numerical experiments.

## REFERENCES


Emmanuel Abbé, Enric Boix-Adserà, Matthew S. Brennan, Guy Bresler, and Dheeraj Nagaraj. The staircase property: How hierarchical structure can guide deep learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26989–27002. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/e2db7186375992e729165726762cb4c1-Abstract.html.

Emmanuel Abbé, Enric Boix-Adserà, and Theodor Misiakiewicz. The merged-staircase property: A necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4782–4887. PMLR, 2022. URL https://proceedings.mlr.press/v178/abbe22a.html.

Emmanuel Abbé, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: Leap complexity and saddle-to-saddle dynamics. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2552–2623. PMLR, 2023. URL https://proceedings.mlr.press/v195/abbe23a.html.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1199–1227. PMLR, 2023.

Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Online learning and information exponents: The importance of batch size & Time/Complexity tradeoffs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1730–1762. PMLR, 2024. URL https://proceedings.mlr.press/v235/arnaboldi24a.html.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f7e7fabd73b3df96c54a320862afcb78-Abstract-Conference.html.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021. URL https://www.jmlr.org/papers/v22/20-1288.html.

Gérard Ben Arous, Murat A. Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d7ce06e9293c3d8e6cb3f80b4157f875-Abstract-Conference.html.

Gérard Ben Arous, Cédric Gerbelot, and Vanessa Piccolo. Stochastic gradient descent in high dimensions for multi-spiked tensor PCA. *Comm. Pure Appl. Math.*, 79(10):2291–2369, 2026. ISSN 0010-3640,1097-0312. doi: 10.1002/cpa.70056. URL https://doi.org/10.1002/cpa.70056.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow part I: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 78(12):2354–2435, 2025. doi: 10.1002/cpa.70006. URL https://doi.org/10.1002/cpa.70006.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. URL https://arxiv.org/abs/2104.13478.

Francesco Cagnetta, Alessandro Favero, and Matthieu Wyart. What can be learnt with wide convolutional neural networks? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3347–3379. PMLR, 2023. URL https://proceedings.mlr.press/v202/cagnetta23a.html.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. URL https://www.jmlr.org/papers/v21/20-163.html.

Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1161–1227. PMLR, 2020. URL https://proceedings.mlr.press/v125/chen20a.html.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Alex Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5413–5452. PMLR, 2022. URL https://proceedings.mlr.press/v178/damian22a.html.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 752–784. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/02763667a5761ff92bb15d8751bcd223-Abstract-Conference.html.

Alex Damian, Loucas Pillaud-Vivien, Jason D. Lee, and Joan Bruna. Computational-statistical gaps in Gaussian single-index models (extended abstract). In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1262. PMLR, 2024. URL https://proceedings.mlr.press/v247/damian24a.html.

Alex Damian, Jason D. Lee, and Joan Bruna. The generative leap: Tight sample complexity for efficiently learning Gaussian multi-index models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 28276–28311. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/290056cc784d4825eb8cf5c72f808ebb-Abstract-Conference.html.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024a. URL https://jmlr.org/papers/v25/23-1543.html.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9991–10016. PMLR, 2024b. URL https://proceedings.mlr.press/v235/dandi24a.html.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1887–1930. PMLR, 2018. URL https://proceedings.mlr.press/v75/dudeja18a.html.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969. PMLR, 2021. URL https://proceedings.mlr.press/v139/elesedy21a.html.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2747–2755. PMLR, 2018. URL https://proceedings.mlr.press/v80/kondor18a.html.

Aakash Sunil Lahoti, Stefani Karp, Ezra Winston, Aarti Singh, and Yuanzhi Li. Role of locality and weight sharing in image-based tasks: A sample complexity separation between CNNs, LCNs, and FCNs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/71b17f00017da0d73823ccf7fbce2d4f-Abstract-Conference.html.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020. URL https://arxiv.org/abs/2005.00178.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3351–3418. PMLR, 2021. URL https://proceedings.mlr.press/v134/mei21a.html.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A. Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/11421.

H. N. Nagaraja, Karthik Bharath, and Fangyuan Zhang. Spacings around an order statistic. *Ann. Inst. Statist. Math.*, 67(3):515–540, 2015. ISSN 0020-3157,1572-9052. doi: 10.1007/s10463-014-0466-9. URL https://doi.org/10.1007/s10463-014-0466-9.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Learning sum of diverse features: Computational hardness and efficient gradient-based training for ridge combinations. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4009–4081. PMLR, 2024. URL https://proceedings.mlr.press/v247/oko24a.html.

Yunwei Ren and Jason D. Lee. Learning orthogonal multi-index models: A fine-grained information exponent analysis. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 165869–165920. Curran Associates, Inc., 2025. doi: 10.52202/085713-5528. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f24777ce35e5a1ef111336dc85c5d9c0-Paper-Conference.pdf.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D. Lee. Emergence and scaling laws in sgd learning of shallow neural networks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 38227–38309. Curran Associates, Inc., 2025. doi: 10.52202/085713-1280. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/36d13b71487b29965a2d3f5fcc109afc-Paper-Conference.pdf.

Berfin Şimşek, Amire Bendjeddou, and Daniel Hsu. Learning gaussian multi-index models with gradient flow: Time complexity and directional convergence. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4204–4212. PMLR, 2025. URL https://proceedings.mlr.press/v258/simsek25a.html.

Emanuele Troiani, Yatin Dandi, Leonardo Defilippis, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Fundamental computational limits of weak learnability in high-dimensional multi-index models. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2467–2475. PMLR, 2025. URL https://proceedings.mlr.press/v258/troiani25a.html.

Konstantinos Christopher Tsiolis, Alireza Mousavi-Hosseini, and Murat A. Erdogdu. From information to generative exponent: Learning rate induces phase transitions in SGD. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/77b7d565f2370979b898d76d7ea27288-Abstract-Conference.html.

Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL https://doi.org/10.1017/9781108231596. An introduction with applications in data science, With a foreword by Sara van de Geer.

# CONTENTS

**A Related work** **16**

**B Detailed statements and proof roadmap** **17**

B.1 Weak recovery for $p \geq 3$ . . . . . 17

B.2 Subspace weak recovery for the quadratic link . . . . . 18

B.3 Matching lower bounds . . . . . 19

B.4 Organization of the proofs . . . . . 20

**C Teacher geometry and initialization** **20**

**D Correlation dynamics and auxiliary estimates** **25**

D.1 Correlation recursion and population drift . . . . . 25

D.2 Coordinate noise estimates . . . . . 27

D.3 Uniform control of the stochastic gradient . . . . . 30

**E Proof of weak recovery for the tied student** **31**

E.1 Proof strategy . . . . . 32

E.2 Microscopic selection . . . . . 33

E.3 Growth to a macroscopic level . . . . . 37

**F Proofs for the augmented and untied students** **40**

F.1 Componentwise reductions . . . . . 40

F.2 One-weight untied dynamics . . . . . 41

F.3 Proof of the untied weak recovery bound . . . . . 42

F.4 Proof of weak orbit coverage . . . . . 44

**G Quadratic link and weak subspace recovery** **46**

G.1 Proof strategy . . . . . 47

G.2 Total overlap and initialization . . . . . 47

G.3 Recursion and population growth . . . . . 48

G.4 Noise and gradient moments . . . . . 49

G.5 Growth and proof of the tied upper bound . . . . . 50

G.6 Augmented and untied upper bounds . . . . . 54

**H Lower bounds for weak recovery** **56**

H.1 Proof strategy . . . . . 56

H.2 Initialization and escape times . . . . . 57

H.3 An upper comparison for the correlations . . . . . 57

H.4 Noise before escape . . . . . 59

H.5 Proof of the non-escape theorem . . . . . 60

# A Related work

**Learning low-dimensional representations —** Multi-index models provide a tractable setting for studying how learning algorithms discover low-dimensional structure in high-dimensional data. Chen & Meka (2020) developed algorithms for Gaussian polynomial multi-index models using filtered PCA and gradient-based refinement. Neural-network representation learning has been established under second-order nondegeneracy conditions (Damian et al., 2022), while Mousavi-Hosseini et al. (2023) showed that SGD with weight decay suppresses weight components orthogonal to the teacher subspace. These approaches investigate how information about the index subspace can be extracted through moment-based methods or through the evolution of network weights.

**Staircase structure and leap complexity —** Hierarchical structure can allow previously learned features to facilitate the discovery of new ones. For Boolean inputs, Abbé et al. (2021) introduced the staircase property and established hierarchical learning guarantees under a layerwise training procedure. For Boolean functions depending on a fixed number of coordinates, Abbé et al. (2022) identified the merged-staircase property as necessary and nearly sufficient for learning with $O(d)$ samples in the two-layer mean-field setting. Abbé et al. (2023) extended this perspective through leap complexity, which quantifies how many new coordinates must be learned together, counting Hermite multiplicities for Gaussian inputs. They proved learning guarantees for a structured class of Gaussian targets under a modified SGD procedure, showing how the dynamics move between successive saddles as additional coordinates are recovered.

**Gradient dynamics and subspace discovery —** For Gaussian single-index models, Ben Arous et al. (2021) identified the information exponent governing the initial search phase of online SGD. Ba et al. (2022) quantified representation learning after one gradient step, highlighting the role of learning-rate scaling. For Gaussian multi-index models, Dandi et al. (2024a) showed that successive gradient steps with fresh batches of $O(d)$ samples can reveal additional directions through a staircase mechanism. Bietti et al. (2025) established global convergence and characterized successive learning phases for a two-timescale population gradient flow, with timescales determined by the target's Hermite decomposition. Related hierarchical subspace learning and weak-recovery thresholds were analyzed through approximate message passing by Troiani et al. (2025). For orthogonal additive teachers, Şimşek et al. (2025) studied directional selection under correlation-loss gradient flow. Competition between signal directions also appears in multi-spiked tensor PCA, where Ben Arous et al. (2026) characterize a greedy maximum selection mechanism leading to the sequential recovery of individual spikes.

**Dependence on the learning procedure —** Information-exponent rates depend on the updates being analyzed. Loss smoothing (Damian et al., 2023), batch reuse (Dandi et al., 2024b), and alternating layerwise updates with suitable learning-rate scales (Tsiolis et al., 2025) can improve upon the rates of the corresponding vanilla dynamics. Batch size and the loss also affect online learning times (Arnaboldi et al., 2024). Complementary algorithmic results introduce generative exponents and generative leap complexity for fixed-rank Gaussian models, accounting for information accessible through nonlinear label processing and sequential subspace estimation (Damian et al., 2024; 2025). These works establish lower bounds within statistical-query or low-degree frameworks and develop algorithms that attain the corresponding scales under their assumptions.

**Growing rank and subspace recovery —** When the number of teacher directions grows with dimension, the rank dependence of learning guarantees becomes essential. Oko et al. (2024) analyzed learning sums of nearly orthogonal ridge functions with growing rank, and Ren et al. (2025) related individual feature recovery times to prediction-risk scaling laws. In particular, the sample complexity of the untied learner in Theorem 3.2 coincides with the results in Oko et al. (2024), where they consider learning an untied teacher instead of a symmetry-tied one as in the present work. For quadratic targets, Ben Arous et al. (2025) studied growing-rank networks trained by orthogonality-constrained online SGD. Subspace and directional recovery can also occur at different stages: Ren & Lee (2025) showed how second-order information can reveal the index subspace before higher-order components distinguish individual teacher directions.

**Weight sharing and data augmentation —** Group-equivariant architectures (Cohen & Welling, 2016; Kondor & Trivedi, 2018) and group-averaged training (Chen et al., 2020; Lyle et al., 2020) provide different ways to exploit symmetry. Statistical benefits have been quantified for equivariant, random-feature, and kernel models (Elesedy & Zaidi, 2021; Mei et al., 2021; Cagnetta et al., 2023). Beyond these settings, Lahoti et al. (2024) established sample-complexity separations between convolutional, locally connected, and fully connected networks in a structured image model, distinguishing the benefits of locality and weight sharing.

Our work studies how a known symmetry changes online feature learning when the teacher rank grows with dimension. We compare weak directional and subspace recovery under weight sharing, full-group augmentation, and untied training without augmentation. The matching lower bounds are non-escape bounds for the specified spherical SGD updates and learning rates; they do not assert optimality over alternative learning algorithms.

# B Detailed statements and proof roadmap

The goal of the appendices B– H is to provide the proofs of the weak recovery results stated in Theorems 3.2 and 3.7, as well as the orbit-coverage result in Corollary 3.5. We begin by stating the precise recovery guarantees used in the proofs and fixing notation.

Given a reference step size $\eta_d > 0$, we use throughout the procedure-specific step sizes

$$\eta_d^{\mathsf{T}} \coloneqq \eta_d, \qquad \eta_d^{\mathsf{U}} \coloneqq \sqrt{\frac{s}{r}}\,\eta_d, \qquad \eta_d^{\mathsf{A}} \coloneqq \sqrt{rs}\,\eta_d. \tag{9}$$

We also write

$$\gamma_d \coloneqq \sqrt{\frac{\log r}{d}},$$

for the microscopic correlation scale appearing throughout the recovery analysis.

We treat separately polynomial links with information exponent $p \geq 3$, under Assumption 1(i), and the pure quadratic link $\sigma = h_2$, under Assumption 1(ii).

## B.1 Weak recovery for $p \geq 3$

Recall the orbit correlation $M_\sigma(w)$ defined in (3). For $\mathsf{k} \in \{\mathsf{T}, \mathsf{U}, \mathsf{A}\}$, $a \in (0,1)$, and $1 \leq j \leq s_{\mathsf{k}}$, define the hitting time

$$\tau_a^{j,\mathsf{k}} \coloneqq \inf\left\{t \geq 0 \colon M_\sigma(w_t^{j,\mathsf{k}}) \geq a\right\}, \tag{10}$$

with the convention $\inf \emptyset = \infty$. For the tied trajectory, we write $w_t^{\mathsf{T}} = w_t^{1,\mathsf{T}}$ and $\tau_a^{\mathsf{T}} = \tau_a^{1,\mathsf{T}}$.

**Theorem B.1** (Weak recovery for $p \geq 3$)**.** *Suppose that Assumptions 1(i), 2, and 3 hold. Fix $\rho \in (0,1)$. There exists $A_* = A_*(p, P) \geq (p-2)/2$ such that the following holds for every $A > A_*$. Set*

$$\eta_d = d^{-p/2}(\log d)^{-A},$$

*and use the procedure-specific step sizes* (9)*.*

(T) *If tied spherical SGD is run with $\eta_d^{\mathsf{T}}$, then there exists $C_T = C_T(\sigma, \delta, \rho) < \infty$ such that*

$$\mathbb{P}\left(\tau_\rho^{\mathsf{T}} \leq \frac{C_T}{\eta_d \gamma_d^{p-2}}\right) \to 1.$$

(U) *For any fixed $j \in \{1, \dots, s\}$, if untied spherical SGD is run with $\eta_d^{\mathsf{U}}$, then there exists $C_U = C_U(\sigma, \delta, \rho) < \infty$ such that*

$$\mathbb{P}\left(\tau_\rho^{j,\mathsf{U}} \leq \frac{C_U r}{\eta_d \gamma_d^{p-2}}\right) \to 1.$$

(A) *For any fixed $j \in \{1, \dots, s\}$, if augmented spherical SGD is run with $\eta_d^{\mathsf{A}}$, then there exists $C_A = C_A(\sigma, \delta, \rho) < \infty$ such that*

$$\mathbb{P}\left(\tau_\rho^{j,\mathsf{A}} \le \frac{C_A}{\eta_d \gamma_d^{p-2}}\right) \to 1.$$

Thus the tied and augmented procedures achieve weak recovery within

$$\frac{1}{\eta_d \gamma_d^{p-2}} = O\left(d^{p-1} \operatorname{polylog}(d)\right)$$

samples, whereas the untied procedure requires at most $O\left(rd^{p-1} \operatorname{polylog}(d)\right)$ samples. Theorem B.1 gives the upper bounds stated in Theorem 3.2. The matching lower bounds follow from Theorem B.4 below.

*Remark* B.2 (Fixed-weight upper bounds). For the untied and augmented procedures, the upper bound holds for any fixed weight index $j$. In particular, the proof does not exploit the best among the $s$ random initializations. The width $s$ enters only through the learning-rate rescaling in (9).

### B.2 Subspace weak recovery for the quadratic link

When $\sigma = h_2$, the individual teacher directions are generally not identifiable, and the natural recovery target is the teacher orbit subspace

$$\mathcal{S}^* := \operatorname{span}\{\Pi^k w^* : 0 \le k < r\}.$$

For $w \in \mathbb{S}^{d-1}$, let $w^{\parallel}$ denote its orthogonal projection onto $\mathcal{S}^*$. For $a \in (0,1)$ and $\mathsf{k} \in \{\mathsf{T}, \mathsf{U}, \mathsf{A}\}$, define

$$\tau_a^{(2),\mathsf{k}} := \inf\left\{t \ge 0 : \max_{1 \le j \le s_{\mathsf{k}}} \|(w_t^{j,\mathsf{k}})^{\parallel}\|^2 \ge a\right\}, \tag{11}$$

with $\inf \emptyset = \infty$. For the tied trajectory, we write $\tau_a^{(2)} = \tau_a^{(2),\mathsf{T}}$.

**Theorem B.3** (Weak subspace recovery for the quadratic link)**.** *Suppose that Assumptions 1(ii), 2, and 3 hold. Fix $\rho \in (0,1)$ independently of $d$. There exists a universal constant $A_0 > 0$ such that the following holds for every $A > A_0$. Set*

$$\eta_d = d^{-1}(\log d)^{-A},$$

*and use the procedure-specific step sizes* (9)*.*

(T) *If tied spherical SGD is run with $\eta_d^{\mathsf{T}}$, then there exists $C_T = C_T(\delta, \rho) < \infty$ such that*

$$\mathbb{P}\left(\tau_\rho^{(2),\mathsf{T}} \le \frac{C_T \log(d/r)}{\eta_d}\right) \to 1.$$

(U) *If untied spherical SGD is run with $\eta_d^{\mathsf{U}}$, then there exists $C_U = C_U(\delta, \rho) < \infty$ such that*

$$\mathbb{P}\left(\tau_\rho^{(2),\mathsf{U}} \le \frac{C_U r \log(d/r)}{\eta_d}\right) \to 1.$$

(A) *If augmented spherical SGD is run with $\eta_d^{\mathsf{A}}$, then there exists $C_A = C_A(\delta, \rho) < \infty$ such that*

$$\mathbb{P}\left(\tau_\rho^{(2),\mathsf{A}} \le \frac{C_A \log(d/r)}{\eta_d}\right) \to 1.$$

Since $\log(d/r) = (1-\delta)\log d$, these bounds are $\widetilde{O}(d)$ for the tied and augmented procedures, and $\widetilde{O}(rd)$ for the untied.

### B.3 MATCHING LOWER BOUNDS

We next give a non-escape estimate that yields the matching lower bounds for both recovery notions above. The result applies to any fixed centered polynomial link with information exponent $p \geq 2$. For $\mathsf{k} \in \{\mathsf{T}, \mathsf{U}, \mathsf{A}\}$, define

$$\tau_{\mathrm{esc}}^{\mathsf{k}}(C) := \inf\left\{ t \geq 0 \colon \max_{1 \leq j \leq s_{\mathsf{k}}} \max_{0 \leq \ell < r} |m_\ell(w_t^{j,\mathsf{k}})| \geq C\gamma_d \right\}, \tag{12}$$

with $\inf \emptyset = \infty$. Thus $\tau_{\mathrm{esc}}^{\mathsf{k}}$ is the first time that any weight of procedure $\mathsf{k}$ develops a correlation larger than a constant multiple of the initialization scale $\gamma_d$.

**Theorem B.4** (Non-escape for the three learning procedures)**.** *Suppose that Assumptions 2 and 3 hold, and let*

$$\sigma(z) = \sum_{j=p}^{P} a_j h_j(z), \qquad 2 \leq p \leq P < \infty, \qquad a_p \neq 0,$$

*where $P$ and the coefficients are fixed independently of $d$. Fix $A > 0$, set*

$$\eta_d = d^{-p/2}(\log d)^{-A},$$

*and use the procedure-specific step sizes* (9)*. Then there exist constants $c, C_{\mathrm{esc}} > 0$, depending only on $\sigma, \delta, \kappa$, such that, with*

$$T_d := \frac{c}{\eta_d \gamma_d^{p-2}},$$

*we have*

$$\mathbb{P}\left(\tau_{\mathrm{esc}}^{\mathsf{T}}(C_{\mathrm{esc}}) > T_d,\ \tau_{\mathrm{esc}}^{\mathsf{A}}(C_{\mathrm{esc}}) > T_d,\ \tau_{\mathrm{esc}}^{\mathsf{U}}(C_{\mathrm{esc}}) > rT_d\right) \to 1.$$

On the event in Theorem B.4, every weight remains at microscopic correlation with every teacher direction throughout the corresponding time interval. Hence, for $p \geq 3$, weak recovery is impossible on these intervals. For the quadratic link, the same event also rules out subspace weak recovery. Indeed, on the teacher-geometry event of Lemma C.1,

$$\|(w_t^{j,\mathsf{k}})^{\|}\|^2 \leq \frac{\sum_{\ell=0}^{r-1} m_\ell(w_t^{j,\mathsf{k}})^2}{1 - r\mu_*} \leq \frac{C_{\mathrm{esc}}^2 r \gamma_d^2}{1 - r\mu_*} = o(1),$$

uniformly over all weights and times in the corresponding intervals. Since each iteration uses one fresh observation, Theorem B.4 yields sample complexity lower bounds $\widetilde{\Omega}(d^{p-1})$ for the tied and augmented procedures and $\widetilde{\Omega}(rd^{p-1})$ for the untied procedure, for every polynomial link covered by Theorem B.4.

We now provide the proof of Theorems 3.2 and 3.7.

*Proof of Theorem 3.2.* Fix $A > A_*$, where $A_*$ is given by Theorem B.1. For the upper bounds, apply Theorem B.1 with $j = 1$. Since recovery of one weight is sufficient for Definition 3.1, weak recovery occurs with probability $1 - o(1)$ within $C/(\eta_d \gamma_d^{p-2})$ samples for the tied and augmented procedures, and within $Cr/(\eta_d \gamma_d^{p-2})$ samples for the untied procedure.

For the lower bounds, Theorem B.4 applies under the same assumptions and learning rates. Since $M_\sigma(w) \leq \max_{0 \leq \ell < r} |m_\ell(w)|$ and $C_{\mathrm{esc}}\gamma_d < \rho$ for all sufficiently large $d$, its non-escape event excludes weak recovery by every weight up to times $c/(\eta_d \gamma_d^{p-2})$ for the tied and augmented procedures, and $cr/(\eta_d \gamma_d^{p-2})$ for the untied procedure.

Finally,

$$\frac{1}{\eta_d \gamma_d^{p-2}} = \frac{d^{(p-2)/2}}{\eta_d (\log r)^{(p-2)/2}} = \widetilde{\Theta}\left(\frac{d^{(p-2)/2}}{\eta_d}\right) = \widetilde{\Theta}(d^{p-1}),$$

where we used that $r = d^\delta$ and $\eta_d = d^{-p/2}(\log d)^{-A}$. This proves the claim. □

*Proof of Theorem 3.7.* Fix $A > A_0$, where $A_0$ is given by Theorem B.3. By (11), the event in Definition 3.6 is exactly $\{\tau_\rho^{(2),\mathsf{k}} \le n\}$. Theorem B.3 therefore gives upper bounds of $C\log(d/r)/\eta_d$ samples for the tied and augmented procedures and of $Cr\log(d/r)/\eta_d$ samples for the untied procedure.

For the lower bounds, apply Theorem B.4 with $\sigma = h_2$ and $p = 2$. Then $T_d = c/\eta_d$, and the subspace estimate above shows that, with probability $1 - o(1)$, every weight satisfies $\|(w_t^{j,\mathsf{k}})^{\shortparallel}\|^2 = o(1)$ throughout the corresponding time interval. Since $\rho > 0$ is fixed, weak subspace recovery is therefore impossible before $c/\eta_d$ for the tied and augmented procedures, and before $cr/\eta_d$ for the untied procedure.

The upper and lower bounds differ only by the logarithmic factor $\log(d/r) = (1-\delta)\log d$. Since $\eta_d^{-1} = d(\log d)^A$, we obtain

$$n_{\mathsf{T}}^{(2)}(\rho),\ n_{\mathsf{A}}^{(2)}(\rho) = \widetilde{\Theta}(\eta_d^{-1}) = \widetilde{\Theta}(d), \qquad n_{\mathsf{U}}^{(2)}(\rho) = \widetilde{\Theta}(r\eta_d^{-1}) = \widetilde{\Theta}(rd).$$

□

### B.4 Organization of the proofs

The remainder of the appendices is organized as follows. Appendix C establishes the geometric properties of the random teacher orbit and the initialization estimates used throughout the analysis. Appendix D develops the correlation dynamics and stochastic gradient estimates needed for the recovery proofs. Appendix E proves the tied upper bound in Theorem B.1, while Appendix F proves the augmented and untied upper bounds. Section F.4 then proves Corollary 3.5. Appendix G treats the quadratic link and proves Theorem B.3. Finally, Appendix H proves the non-escape estimate in Theorem B.4, which yields the matching lower bounds.

## C Teacher geometry and initialization

The purpose of this section is to characterize the random geometry of the teacher orbit and the scale of the correlations at initialization. For a fixed unit vector $v \in \mathbb{S}^{d-1}$ and an independent vector $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, the correlation $\langle v, u\rangle$ is typically of order $d^{-1/2}$. In our model, however, we consider the correlations of a random direction with the $r$ shifted teacher directions $w^*, \Pi w^*, \dots, \Pi^{r-1}w^*$. Because these directions are nearly orthogonal, the corresponding correlations behave approximately as $r$ independent Gaussian coordinates. Their largest relevant value is therefore enhanced by the usual extreme-value factor and occurs at the scale

$$\gamma_d := \sqrt{\frac{\log r}{d}}.$$

We first show that the teacher orbit is nearly orthogonal. We then establish the initialization scale $\gamma_d$ and a quantitative gap between the largest relevant correlation and its closest competitor, where in the even-$p$ case correlations are compared in absolute value. This gap provides the initial asymmetry that is amplified by the nonlinear SGD dynamics.

Recall that the teacher direction is drawn uniformly at random, $w^* \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. We begin by controlling the correlations among its cyclic shifts. Define the maximum correlation within the teacher orbit by

$$\mu_* := \max_{1\le k\le r-1} \left|\left\langle w^*, \Pi^k w^*\right\rangle\right|.$$

**Lemma C.1** (Geometry of the teacher orbit)**.** *Under Assumptions 2 and 3, there exists a constant $C_* < \infty$ such that, for the event*

$$\mathcal{G}_d := \{\mu_* \le C_*\gamma_d\},$$

*it holds that*

$$\mathbb{P}(\mathcal{G}_d) \to 1.$$

*Moreover, on $\mathcal{G}_d$, $r\mu_* = o\left((\log d)^{-2}\right)$.*

*Proof.* Assumption 2 allows us to write $w^* = g/\|g\|$ with $g \sim \mathcal{N}(0, \mathrm{I}_d)$, so that

$$\mu_* = \max_{1\le a\le r-1} \frac{\left|g^\top \Pi^a g\right|}{\|g\|^2}.$$

Thus we need an upper bound on $\max_{1\le a\le r-1}\left|g^\top \Pi^a g\right|$ and a lower bound on $\|g\|^2$. By the Hanson–Wright inequality (see e.g. (Vershynin, 2018, Theorem 6.2.1)), for every symmetric matrix $A$,

$$\mathbb{P}\left(\left|g^\top A g - \mathbb{E}[g^\top A g]\right| > t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{\|A\|_{\mathrm{F}}^2}, \frac{t}{\|A\|_{\mathrm{op}}}\right\}\right), \qquad t > 0. \tag{13}$$

Note that $\mathbb{E}[g^\top A g] = \mathrm{Tr}A$. Fix $1 \le a \le r-1$. Since $a\Delta = ad/r$ is not a multiple of $d$, the cyclic shift $\Pi^a$ has no fixed coordinates and hence $\mathrm{Tr}(\Pi^a) = 0$. Since the matrix $\Pi^a$ is not necessarily symmetric, we consider its symmetric part

$$A_a \coloneqq \frac{\Pi^a + \Pi^{-a}}{2}.$$

Then

$$\mathrm{Tr}(A_a) = 0, \qquad \|A_a\|_{\mathrm{op}} \le 1, \qquad \|A_a\|_{\mathrm{F}}^2 \le d.$$

Moreover, since the skew-symmetric part does not contribute to a quadratic form, $g^\top \Pi^a g = g^\top A_a g$. Hence, setting $t = C\sqrt{d\log r}$ in (13) gives

$$\mathbb{P}\left(\left|g^\top \Pi^a g\right| > C\sqrt{d\log r}\right) \le 2\exp\left(-c\min\left\{C^2\log r, C\sqrt{d\log r}\right\}\right).$$

Since $\log r = \delta \log d = o(d)$, for all sufficiently large $d$,

$$\mathbb{P}\left(\left|g^\top \Pi^a g\right| > C\sqrt{d\log r}\right) \le 2e^{-cC^2\log r} = 2r^{-cC^2}.$$

Taking a union bound over $a = 1, \dots, r-1$, we obtain

$$\mathbb{P}\left(\max_{1\le a\le r-1}\left|g^\top \Pi^a g\right| > C\sqrt{d\log r}\right) \le 2r^{1-cC^2}. \tag{14}$$

Choosing $C$ sufficiently large, the right-hand side is $o(1)$. For the lower bound on $\|g\|^2$, we apply again (13) to $A = \mathrm{I}_d$, for which $\mathbb{E}\|g\|^2 = d$, with $t = d/2$. This gives

$$\mathbb{P}\left(\|g\|^2 < \frac{d}{2}\right) \le \mathbb{P}\left(\left|\|g\|^2 - d\right| > \frac{d}{2}\right) \le 2e^{-cd}. \tag{15}$$

Therefore, combining (14) and (15), with probability at least $1 - 2r^{1-cC^2} - 2e^{-cd} = 1 - o(1)$, we have

$$\mu_* = \max_{1\le a\le r-1} \frac{\left|g^\top \Pi^a g\right|}{\|g\|^2} \le 2C\sqrt{\frac{\log r}{d}}.$$

Finally, since $r = d^\delta$ with $\delta < 1/2$ by Assumption 3,

$$r\mu_* \le 2Cd^{\delta-\frac{1}{2}}\sqrt{\log r} = 2C\sqrt{\delta}d^{\delta-\frac{1}{2}}\sqrt{\log d} = o\left((\log d)^{-2}\right).$$

This completes the proof. □

We next introduce a generic initialization $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, independent of $w^*$, and study its correlations with the teacher orbit. They can therefore be applied to the tied initialization and, separately, to any fixed component of the untied or augmented initialization. Recall that

$$m_k(u) = \langle \Pi^k w^*, u\rangle, \qquad q_k(u) = \begin{cases} |m_k(u)|, & p \text{ even}, \\ m_k(u), & p \text{ odd}. \end{cases}$$

**Lemma C.2** (Initialization scale)**.** *Suppose that Assumption 3 holds. Let $w^*, u$ be independent and uniformly distributed on $\mathbb{S}^{d-1}$. Then*

$$\max_{0\le k<r} |m_k(u)| = \left(\sqrt{2} + o_{\mathbb{P}}(1)\right)\gamma_d.$$

*Moreover,*

$$\max_{0\le k<r} q_k(u) = \left(\sqrt{2} + o_{\mathbb{P}}(1)\right)\gamma_d.$$

*In particular, there exist constants $0 < c < C < \infty$ such that, with probability $1 - o(1)$,*

$$\max_{0\le k<r} |m_k(u)| \le C\gamma_d, \qquad \max_{0\le k<r} q_k(u) \ge c\gamma_d.$$

We next show that the largest relevant correlation is separated from all competing correlations by a relative gap of order $(\log d)^{-2}$.

**Lemma C.3** (Initialization gap)**.** *Suppose that Assumption 3 holds. Let $w^*, u$ be independent and uniformly distributed on $\mathbb{S}^{d-1}$, and set $\nu_d := (\log d)^{-2}$. With probability $1 - o(1)$, the maximizer*

$$i_* := \arg\max_{0 \le k < r} q_k(u)$$

*is unique and satisfies*

$$\max_{k \neq i_*} q_k(u) \le (1 - \nu_d) q_{i_*}(u).$$

We prove Lemmas C.2 and C.3 simultaneously.

*Proof of Lemmas C.2 and C.3.* We first record the extreme-value estimates used below. Let $Z_0, \dots, Z_{r-1}$ be i.i.d. $\mathcal{N}(0,1)$. Denote by $A_1 \ge A_2$ the two largest values among $|Z_0|, \dots, |Z_{r-1}|$, and by $P_1 \ge P_2$ the two largest values among $Z_0, \dots, Z_{r-1}$. Standard Gaussian tail estimates imply

$$\frac{A_1}{\sqrt{2\log r}} \xrightarrow{\mathbb{P}} 1, \qquad \frac{P_1}{\sqrt{2\log r}} \xrightarrow{\mathbb{P}} 1. \tag{16}$$

Moreover, applying the asymptotics of extreme spacings in the Gumbel domain (Nagaraja et al., 2015, Theorem 5) to the standard normal and half-normal distributions gives

$$\sqrt{2\log r}\,(A_1 - A_2) \xrightarrow{d} \mathrm{Exp}(1), \qquad \sqrt{2\log r}\,(P_1 - P_2) \xrightarrow{d} \mathrm{Exp}(1). \tag{17}$$

We now deduce the relative gap needed below. Recall that $\nu_d = (\log d)^{-2}$. Since $r = d^\delta$ by Assumption 3, $\nu_d \log r = \delta (\log d)^{-1} \to 0$. Moreover, by (16),

$$\nu_d \sqrt{2\log r} A_1 = (1 + o_{\mathbb{P}}(1)) 2\nu_d \log r \xrightarrow{\mathbb{P}} 0.$$

For every fixed $a > 0$, setting

$$U_d := \sqrt{2\log r}\,(A_1 - A_2), \qquad V_d := a\nu_d \sqrt{2\log r}\, A_1,$$

we have $U_d \xrightarrow{d} \mathrm{Exp}(1)$ and $V_d \xrightarrow{\mathbb{P}} 0$. For every $\varepsilon > 0$,

$$\mathbb{P}(U_d \le V_d) \le \mathbb{P}(U_d \le \varepsilon) + \mathbb{P}(V_d > \varepsilon).$$

Taking the limit superior and then letting $\varepsilon \downarrow 0$ gives

$$\mathbb{P}(U_d \le V_d) \to 0,$$

since an $\mathrm{Exp}(1)$ random variable is strictly positive almost surely. Hence, for every fixed $a > 0$,

$$\mathbb{P}\left(A_1 - A_2 \le a\nu_d A_1\right) \to 0.$$

Equivalently,

$$A_2 \le (1 - a\nu_d) A_1 \tag{18}$$

with probability $1 - o(1)$. The same argument yields

$$P_2 \le (1 - a\nu_d) P_1 \tag{19}$$

with probability $1 - o(1)$.

We now transfer the preceding extreme-value estimates from independent Gaussian coordinates to the actual correlations with the teacher orbit. Define the teacher-geometry event of Lemma C.1:

$$\mathcal{G}_d := \{\mu_* \le C\gamma_d\}.$$

By that lemma, $\mathbb{P}(\mathcal{G}_d) = 1 - o(1)$. Fix $w^* \in \mathcal{G}_d$. In what follows, probabilities are taken conditionally on this realization of $w^*$. Write

$$u = \frac{h}{\|h\|}, \qquad h \sim \mathcal{N}(0, \mathrm{I}_d),$$

where $h$ is independent of $w^*$, and define

$$X_k := \left\langle \Pi^k w^*, h \right\rangle, \qquad 0 \le k < r.$$

Indeed, for $0 \le i, j < r$,

$$\mathbb{E}[X_i X_j \mid w^*] = \left\langle \Pi^i w^*, \Pi^j w^* \right\rangle.$$

Conditionally on $w^*$, the vector $X := (X_0, \dots, X_{r-1})$ is centered Gaussian with covariance matrix

$$W^* := \left( \left\langle \Pi^i w^*, \Pi^j w^* \right\rangle \right)_{0 \le i,j \le r-1}.$$

For a matrix $A = (A_{ij})_{i,j=1}^r$, let

$$\|A\|_{\infty\to\infty} := \sup_{\|x\|_\infty \le 1} \|Ax\|_\infty = \max_{1 \le i \le r} \sum_{j=1}^r |A_{ij}|$$

denote the operator norm induced by the vector $\ell_\infty$-norm. The diagonal entries of $W^*$ are equal to one. Moreover, if $i \neq j$, then

$$\left|\left\langle \Pi^i w^*, \Pi^j w^* \right\rangle\right| = \left|\left\langle w^*, \Pi^{j-i} w^* \right\rangle\right| \le \mu_*,$$

where the exponent is understood modulo $r$. Hence, on $\mathcal{G}_d$,

$$\|W^* - \mathrm{I}_r\|_{\infty\to\infty} \le (r-1)\mu_* \le Cr\gamma_d.$$

Moreover, $b_d := Cr\gamma_d = o(\nu_d)$ according to Lemma C.1. Since $W^* - \mathrm{I}_r$ is symmetric,

$$\|W^* - \mathrm{I}_r\|_{\mathrm{op}} \le \|W^* - \mathrm{I}_r\|_{\infty\to\infty} \le b_d = o(1).$$

Therefore, for all sufficiently large $d$, every eigenvalue of $W^*$ belongs to $[1 - b_d, 1 + b_d] \subset (0, \infty)$, so $W^*$ is positive definite. Set $B := W^* - \mathrm{I}_r$. Then, for all sufficiently large $d$, $\|B\|_{\infty\to\infty} \le b_d < 1/2$. Hence the binomial series for the principal square root converges absolutely in the $\|\cdot\|_{\infty\to\infty}$-norm:

$$(W^*)^{1/2} = (\mathrm{I}_r + B)^{1/2} = \mathrm{I}_r + \sum_{m=1}^\infty \binom{1/2}{m} B^m.$$

By submultiplicativity,

$$\|(W^*)^{1/2} - \mathrm{I}_r\|_{\infty\to\infty} \le \sum_{m=1}^\infty \left|\binom{1/2}{m}\right| \|B\|_{\infty\to\infty}^m \le C\|B\|_{\infty\to\infty} \le Cb_d,$$

where $C < \infty$ is a universal constant. Therefore, if we set

$$\varepsilon_d(w^*) := \|(W^*)^{1/2} - \mathrm{I}_r\|_{\infty\to\infty},$$

then, uniformly over $w^* \in \mathcal{G}_d$,

$$\varepsilon_d(w^*) \le Cb_d = o(\nu_d). \tag{20}$$

Now define $Z := (W^*)^{-1/2} X$. Conditionally on $w^*$, $Z \sim \mathcal{N}(0, \mathrm{I}_r)$, so its coordinates are independent standard Gaussian variables. Let $A_1 \ge A_2$ and $P_1 \ge P_2$ denote the corresponding absolute and signed order statistics, as above. Since $X = (W^*)^{1/2} Z$, we have

$$\|X - Z\|_\infty \le \varepsilon_d(w^*)\|Z\|_\infty = \varepsilon_d(w^*) A_1. \tag{21}$$

**Absolute maximum.** Let $k_*$ be the almost surely unique index satisfying $|Z_{k_*}| = A_1$. Taking $a = 2$ in (18), with conditional probability $1 - o(1)$,

$$A_2 \le (1 - 2\nu_d) A_1.$$

On this event,

$$|X_{k_*}| \ge A_1 - \|X - Z\|_\infty \ge (1 - \varepsilon_d) A_1,$$

while

$$\max_{k \neq k_*} |X_k| \le A_2 + \|X - Z\|_\infty \le (1 - 2\nu_d + \varepsilon_d) A_1.$$

By (20), $\varepsilon_d = o(\nu_d)$ uniformly on $\mathcal{G}_d$. Hence, for all sufficiently large $d$,

$$1 - 2\nu_d + \varepsilon_d \le (1 - \nu_d)(1 - \varepsilon_d).$$

It follows that

$$\max_{k \neq k_*} |X_k| \le (1 - \nu_d) |X_{k_*}|. \tag{22}$$

In particular, $k_*$ is also the unique maximizer of $|X_k|$. Furthermore,

$$\left| \max_k |X_k| - A_1 \right| \le \|X - Z\|_\infty \le \varepsilon_d A_1.$$

Since $\varepsilon_d = o(1)$ uniformly on $\mathcal{G}_d$, it follows from (16) that, conditionally on $w^* \in \mathcal{G}_d$,

$$\max_{0 \le k < r} |X_k| = (1 + o_{\mathbb{P}}(1)) \sqrt{2 \log r}, \tag{23}$$

where the $o_{\mathbb{P}}(1)$ is uniform over $w^* \in \mathcal{G}_d$.

**Signed maximum.** Let $j_*$ be the almost surely unique index satisfying $Z_{j_*} = P_1$. By (16),

$$\frac{A_1}{P_1} \xrightarrow{\mathbb{P}} 1.$$

In particular, with probability $1 - o(1)$, $A_1 \le 2P_1$. Intersect this event with the spacing event $P_2 \le (1 - 2\nu_d)P_1$, which also has probability $1 - o(1)$ by (19). On their intersection,

$$X_{j_*} \ge P_1 - \varepsilon_d A_1 \ge (1 - 2\varepsilon_d)P_1,$$

while

$$\max_{k \ne j_*} X_k \le P_2 + \varepsilon_d A_1 \le (1 - 2\nu_d + 2\varepsilon_d)P_1.$$

Since $\varepsilon_d = o(\nu_d)$, for all sufficiently large $d$,

$$1 - 2\nu_d + 2\varepsilon_d \le (1 - \nu_d)(1 - 2\varepsilon_d).$$

Therefore

$$\max_{k \ne j_*} X_k \le (1 - \nu_d) X_{j_*}. \tag{24}$$

Thus $j_*$ is also the unique maximizer of the signed coordinates of $X$. Furthermore,

$$\left| \max_{0 \le k < r} X_k - P_1 \right| \le \|X - Z\|_\infty \le \varepsilon_d A_1.$$

Since $A_1/P_1 \to 1$ in probability and $\varepsilon_d = o(1)$, we obtain

$$\max_k X_k = (1 + o_{\mathbb{P}}(1))\sqrt{2 \log r}, \tag{25}$$

again uniformly over $w^* \in \mathcal{G}_d$.

**Normalization and conclusion.** Finally,

$$m_k(u) = \left\langle \Pi^k w^*, u \right\rangle = \frac{X_k}{\|h\|}.$$

Since $\|h\|/\sqrt{d} \xrightarrow{\mathbb{P}} 1$, combining this with (23) yields

$$\max_{0 \le k < r} |m_k(u)| = (1 + o_{\mathbb{P}}(1))\sqrt{\frac{2 \log r}{d}} = (\sqrt{2} + o_{\mathbb{P}}(1))\gamma_d.$$

Similarly, (25) gives

$$\max_{0 \le k < r} m_k(u) = (\sqrt{2} + o_{\mathbb{P}}(1))\gamma_d.$$

Consequently,

$$\max_{0 \le k < r} q_k(u) = (\sqrt{2} + o_{\mathbb{P}}(1))\gamma_d,$$

which proves Lemma C.2. The normalization by the common positive factor $\|h\|$ does not change either the maximizing index or any relative gap. Hence (22) gives, when $p$ is even,

$$\max_{k \ne i_*} q_k(u) = \max_{k \ne i_*} |m_k(u)| \le (1 - \nu_d)|m_{i_*}(u)| = (1 - \nu_d) q_{i_*}(u),$$

whereas (24) gives, when $p$ is odd,

$$\max_{k \ne i_*} q_k(u) = \max_{k \ne i_*} m_k(u) \le (1 - \nu_d) m_{i_*}(u) = (1 - \nu_d) q_{i_*}(u).$$

In both cases the maximizer is unique. All the preceding conditional estimates hold uniformly over $w^* \in \mathcal{G}_d$. Therefore, if $\mathcal{E}_d$ denotes the event on which the conclusions of Lemmas C.2 and C.3 hold, then

$$\mathbb{P}(\mathcal{E}_d^{\mathrm{c}}) \le \mathbb{P}(\mathcal{G}_d^{\mathrm{c}}) + \sup_{w^* \in \mathcal{G}_d} \mathbb{P}(\mathcal{E}_d^{\mathrm{c}} \mid w^*) = o(1).$$

This completes the proofs of Lemmas C.2 and C.3. □

# D CORRELATION DYNAMICS AND AUXILIARY ESTIMATES

The purpose of this section is to collect the dynamical identities and stochastic estimates used in the weak recovery proofs. We derive them for the tied student, for which the correlation dynamics take their cleanest form. Appendix E uses these estimates to prove weak recovery for the tied student, while Appendix F shows how the same structure changes under augmentation and untied training.

Throughout this section, we suppress the superscript T and write

$$w_t \coloneqq w_t^{\mathsf{T}}, \qquad L \coloneqq L^{\mathsf{T}}, \qquad \Phi \coloneqq \Phi^{\mathsf{T}}, \qquad \eta_d \coloneqq \eta_d^{\mathsf{T}}.$$

We also work on the teacher-geometry event $\mathcal{G}_d$ introduced in Lemma C.1. All estimates below that condition on the teacher are uniform over $w^* \in \mathcal{G}_d$.

## D.1 CORRELATION RECURSION AND POPULATION DRIFT

We begin by deriving the exact evolution of the correlations

$$m_k(w_t) = \langle \Pi^k w^*, w_t \rangle, \qquad 0 \le k < r,$$

under spherical SGD. We first identify the population contribution. For $k \in \{0, \dots, r-1\}$, define the teacher correlations

$$c_k \coloneqq \langle w^*, \Pi^k w^* \rangle,$$

where indices are understood modulo $r$. By cyclicity, $\langle \Pi^k w^*, \Pi^j w^* \rangle = c_{j-k}$.

**Lemma D.1** (Population drift)**.** *For every $w \in \mathbb{S}^{d-1}$ and $k \in \{0, \dots, r-1\}$,*

$$-\langle \Pi^k w^*, \nabla \Phi(w) \rangle = \phi'(m_k(w)) + \Gamma_k(w), \tag{26}$$

*where*

$$\Gamma_k(w) \coloneqq \sum_{j \neq k} c_{j-k} \phi'(m_j(w)) - m_k(w) \sum_{j=0}^{r-1} m_j(w) \phi'(m_j(w)). \tag{27}$$

*Proof.* On the unit sphere, Hermite orthogonality gives

$$\Phi(w) = -\sum_{j=0}^{r-1} \phi(m_j(w)).$$

The spherical gradient of the population objective is

$$\nabla \Phi(w) = (\mathrm{I}_d - w w^\top) \widehat{\nabla} \Phi(w) = -\sum_{j=0}^{r-1} \phi'(m_j(w)) \big( \Pi^j w^* - m_j(w) w \big).$$

Taking the scalar product with $\Pi^k w^*$ yields

$$-\langle \Pi^k w^*, \nabla \Phi(w) \rangle = \sum_{j=0}^{r-1} \phi'(m_j(w)) \left( \langle \Pi^k w^*, \Pi^j w^* \rangle - m_k(w) m_j(w) \right).$$

Separating the term $j = k$, using $\langle \Pi^k w^*, \Pi^k w^* \rangle = 1$, and recalling $\langle \Pi^k w^*, \Pi^j w^* \rangle = c_{j-k}$, gives (26). □

We now combine this population identity with the sample-level SGD update. Define

$$G_t \coloneqq \nabla L(w_{t-1}; x_t, y_t),$$

and introduce the sample-wise centered loss

$$H_t(w) \coloneqq L(w; x_t, y_t) - \Phi(w).$$

Then $G_t = \nabla \Phi(w_{t-1}) + \nabla H_t(w_{t-1})$. We also define the coordinate-wise stochastic fluctuation

$$\xi_{k,t} \coloneqq -\left\langle \Pi^k w^*, \nabla H_t(w_{t-1}) \right\rangle. \tag{28}$$

**Lemma D.2** (Exact correlation evolution)**.** *For every $0 \le k < r$ and $t \ge 1$,*

$$m_k(w_t) = m_k(w_0) + pa_p^2\eta_d \sum_{\ell=1}^{t} m_k(w_{\ell-1})^{p-1} + E_{k,t}, \tag{29}$$

*where*

$$\begin{aligned} E_{k,t} \coloneqq \eta_d \sum_{\ell=1}^{t} R_\sigma(m_k(w_{\ell-1})) + \eta_d \sum_{\ell=1}^{t} \Gamma_k(w_{\ell-1}) + \eta_d \sum_{\ell=1}^{t} \xi_{k,\ell} \\ - \sum_{\ell=1}^{t} \left( \sqrt{1+\eta_d^2\|G_\ell\|^2} - 1 \right) m_k(w_\ell), \end{aligned} \tag{30}$$

*with*

$$R_\sigma(m) \coloneqq \sum_{k=p+1}^{P} ka_k^2 m^{k-1}.$$

*Proof.* By the spherical SGD update (5),

$$w_t = \frac{w_{t-1} - \eta_d G_t}{\|w_{t-1} - \eta_d G_t\|}.$$

Since $G_t$ belongs to the tangent space of $\mathbb{S}^{d-1}$ at $w_{t-1}$, we have $\langle G_t, w_{t-1}\rangle = 0$, and therefore

$$\|w_{t-1} - \eta_d G_t\|^2 = 1 + \eta_d^2\|G_t\|^2.$$

Taking the scalar product of $w_t$ with $\Pi^k w^*$ gives

$$m_k(w_t) = \langle \Pi^k w^*, w_t\rangle = \frac{m_k(w_{t-1}) - \eta_d \langle \Pi^k w^*, G_t\rangle}{\|w_{t-1} - \eta_d G_t\|},$$

or equivalently,

$$\sqrt{1+\eta_d^2\|G_t\|^2}\, m_k(w_t) = m_k(w_{t-1}) - \eta_d \langle \Pi^k w^*, G_t\rangle.$$

Using $G_t = \nabla\Phi(w_{t-1}) + \nabla H_t(w_{t-1})$, Lemma D.1, and the definition of $\xi_{k,t}$, we obtain

$$\sqrt{1+\eta_d^2\|G_t\|^2}\, m_k(w_t) = m_k(w_{t-1}) + \eta_d\phi'(m_k(w_{t-1})) + \eta_d\Gamma_k(w_{t-1}) + \eta_d\xi_{k,t}.$$

Therefore

$$\begin{aligned} m_k(w_t) = m_k(w_{t-1}) + \eta_d\phi'(m_k(w_{t-1})) + \eta_d\Gamma_k(w_{t-1}) + \eta_d\xi_{k,t} \\ - \big(\sqrt{1+\eta_d^2\|G_t\|^2} - 1\big) m_k(w_t). \end{aligned} \tag{31}$$

Finally,

$$\phi'(m) = \sum_{k=p}^{P} ka_k^2 m^{k-1} = pa_p^2 m^{p-1} + R_\sigma(m).$$

Summing (31) over $\ell = 1, \dots, t$ and separating the leading Hermite contribution in (31) gives (29). □

Since $\phi(z) = \sum_{\ell=p}^{P} a_\ell^2 z^\ell$, termwise estimates give, for a constant $C_\sigma < \infty$,

$$|\phi'(z)| \le C_\sigma |z|^{p-1}, \qquad |R_\sigma(z)| \le C_\sigma |z|^p, \qquad |z| \le 1. \tag{32}$$

Moreover, the coefficients of $\phi$ are nonnegative, so

$$\phi'(z) \ge pa_p^2 z^{p-1}, \qquad 0 \le z \le 1. \tag{33}$$

The four terms in $E_{k,t}$ have distinct origins: they are, respectively, the higher-Hermite remainder, the interaction with the other teacher directions, the centered stochastic fluctuation, and the spherical normalization error. The deterministic remainder and interaction terms will be controlled directly from the correlation bounds and the geometry of the teacher orbit. The remaining two contributions require probabilistic estimates on $\xi_{k,t}$ and $\|G_t\|$, which we establish in the next two subsections.

### D.2 COORDINATE NOISE ESTIMATES

Define the filtration

$$\mathcal{F}_t \coloneqq \sigma(w^*, w_0, (x_\ell, y_\ell)_{1\le\ell\le t}).$$

Since $(x_t, y_t)$ is fresh conditionally on $\mathcal{F}_{t-1}$,

$$\mathbb{E}[G_t \mid \mathcal{F}_{t-1}] = \nabla\Phi(w_{t-1}).$$

Therefore,

$$\xi_{k,t} = -\left\langle \Pi^k w^*, G_t - \mathbb{E}[G_t \mid \mathcal{F}_{t-1}]\right\rangle$$

is a martingale difference with respect to $(\mathcal{F}_t)_{t\ge0}$.

We first derive an identity for the second moments of the feature gradient. Besides being used below to control $\xi_{k,t}$, this identity will also be useful in the proof of the uniform gradient bound.

**Lemma D.3** (Feature-gradient covariance)**.** *Let $w \in \mathbb{S}^{d-1}$ and $v \perp w$, and define $a_j(w) \coloneqq \langle w, \Pi^j w\rangle$. Then*

$$\mathbb{E}\left[\langle v, \widehat{\nabla} f_w(x)\rangle^2\right] = \sum_{j=0}^{r-1} \left(\phi'(a_j(w))\langle v, \Pi^j v\rangle + \phi''(a_j(w))\langle w, \Pi^j v\rangle\langle \Pi^j w, v\rangle\right).$$

*Proof.* Computing the Euclidean gradient of $f_w(x)$ defined in Definition 2.2 gives

$$\widehat{\nabla} f_w(x) = \frac{1}{\sqrt{r}} \sum_{i=0}^{r-1} \sigma'\left(\langle \Pi^i w, x\rangle\right) \Pi^{-i} x.$$

Set $Z_i = \langle \Pi^i w, x\rangle$ and $V_i = \langle \Pi^i v, x\rangle$. Then

$$\langle v, \widehat{\nabla} f_w(x)\rangle = \frac{1}{\sqrt{r}} \sum_{i=0}^{r-1} \sigma'(Z_i) V_i,$$

and therefore, by cyclic invariance,

$$\mathbb{E}\left[\langle v, \widehat{\nabla} f_w(x)\rangle^2\right] = \sum_{j=0}^{r-1} \mathbb{E}\left[\sigma'(Z_0)\sigma'(Z_j)V_0V_j\right].$$

For the jointly Gaussian vector $(Z_0, Z_j, V_0, V_j)$, since

$$\mathrm{Cov}(V_0, Z_0) = \mathrm{Cov}(V_j, Z_j) = \langle v, w\rangle = 0,$$

Gaussian integration by parts gives

$$\mathbb{E}\left[\sigma'(Z_0)\sigma'(Z_j)V_0V_j\right] = \langle v, \Pi^j v\rangle \mathbb{E}[\sigma'(Z_0)\sigma'(Z_j)] + \langle \Pi^j w, v\rangle\langle w, \Pi^j v\rangle \mathbb{E}[\sigma''(Z_0)\sigma''(Z_j)].$$

Using

$$\mathbb{E}[\sigma'(Z_0)\sigma'(Z_j)] = \phi'(a_j(w)), \qquad \mathbb{E}[\sigma''(Z_0)\sigma''(Z_j)] = \phi''(a_j(w)),$$

yields the announced statement. □

We now control the conditional moments of the coordinate noise.

**Lemma D.4** (Coordinate noise moments)**.** *On the teacher-geometry event $\mathcal{G}_d$, there exists $C_\sigma < \infty$ such that, for all sufficiently large $d$, all $0 \le k < r$, $t \ge 1$, and $\nu \ge 2$,*

$$\mathbb{E}[\xi_{k,t}^2 \mid \mathcal{F}_{t-1}] \le C_\sigma\left(1 + r m_k(w_{t-1})^2\right), \tag{34}$$

$$\left(\mathbb{E}[|\xi_{k,t}|^\nu \mid \mathcal{F}_{t-1}]\right)^{1/\nu} \le C_\sigma \nu^P\left(1 + \sqrt{r}|m_k(w_{t-1})|\right). \tag{35}$$

*The bounds hold uniformly over $w_{t-1} \in \mathbb{S}^{d-1}$ and over $w^* \in \mathcal{G}_d$.*

*Proof.* Fix a teacher $w^* \in \mathcal{G}_d$ and condition on $\mathcal{F}_{t-1}$. Then $w := w_{t-1}$ is fixed and the only remaining randomness comes from the fresh Gaussian input $x_t$. Throughout the proof, all expectations are with respect to this fresh sample. The constants below are uniform over $w \in \mathbb{S}^{d-1}$ and $w^* \in \mathcal{G}_d$.

We first claim that

$$\sum_{k=0}^{r-1} m_k(w)^2 \le 2. \tag{36}$$

Let

$$\boldsymbol{U}^* := [w^*, \Pi w^*, \dots, \Pi^{r-1} w^*] \in \mathbb{R}^{d\times r},$$

so that $(m_0(w), \dots, m_{r-1}(w)) = (\boldsymbol{U}^*)^\top w$. Its Gram matrix is $W^* = (\boldsymbol{U}^*)^\top \boldsymbol{U}^*$. On $\mathcal{G}_d$, the diagonal entries of $W^*$ are equal to one and the absolute sum of the off-diagonal entries in each row is at most

$$(r-1)\mu_* \le r\mu_* = o(1).$$

Hence, for all sufficiently large $d$,

$$\|W^*\|_{\mathrm{op}} \le \|W^*\|_{\infty\to\infty} \le 1 + (r-1)\mu_* \le 2.$$

This implies that

$$\sum_{k=0}^{r-1} m_k(w)^2 = \|(\boldsymbol{U}^*)^\top w\|^2 \le \|\boldsymbol{U}^*(\boldsymbol{U}^*)^\top\|_{\mathrm{op}} = \|W^*\|_{\mathrm{op}} \le 2,$$

thus proving (36).

Recall that $c_a = \langle w^*, \Pi^a w^*\rangle$ and $\phi(c) = \sum_{k=p}^P a_k^2 c^k$. By (1),

$$\mathbb{E}[y^2] = \frac{1}{r}\sum_{i,j=0}^{r-1} \phi(c_{j-i}) = \sum_{a=0}^{r-1} \phi(c_a),$$

because each difference $a = j - i(\mod r)$ occurs exactly $r$ times. For every $|c| \le 1$,

$$|\phi(c)| \le \sum_{k=p}^P a_k^2 |c|^k \le \left(\sum_{k=p}^P a_k^2\right)|c|^p = \|\sigma\|_{L^2}^2 |c|^p.$$

Therefore

$$\mathbb{E}[y^2] \le \phi(1) + \sum_{a=1}^{r-1} |\phi(c_a)| \le C_\sigma(1 + r\mu_*^p) \le C_\sigma,$$

where the last inequality holds uniformly on $\mathcal{G}_d$ for all sufficiently large $d$. Indeed, $\mu_* \le 1$ and $r\mu_* = o(1)$. Since $y$ is a polynomial of degree at most $P$ in the Gaussian input, Gaussian hypercontractivity gives

$$\|y\|_{L^4} \le C_\sigma. \tag{37}$$

Fix $0 \le k < r$, set $m := m_k(w)$, and define the tangent projection of the $k$th teacher direction by

$$v := (\mathrm{I}_d - ww^\top)\Pi^k w^* = \Pi^k w^* - mw.$$

Then $v \perp w$ and $\|v\|^2 = 1 - m^2 \le 1$. Moreover, since $G = -y\nabla f_w(x)$, we have

$$\langle \Pi^k w^*, G\rangle = -y\langle v, \widehat{\nabla} f_w(x)\rangle.$$

Set $Q := \langle v, \widehat{\nabla} f_w(x)\rangle$ and $a_j(w) := \langle w, \Pi^j w\rangle$. By Lemma D.3,

$$\mathbb{E}Q^2 = \sum_{j=0}^{r-1} \left(\phi'(a_j(w))\langle v, \Pi^j v\rangle + \phi''(a_j(w))\langle w, \Pi^j v\rangle\langle \Pi^j w, v\rangle\right). \tag{38}$$

For $j = 0$,

$$\langle v, \Pi^0 v\rangle = \|v\|^2 \le 1, \qquad \langle w, v\rangle = 0,$$

so the $j = 0$ contribution to (38) is bounded by a constant depending only on $\sigma$. For $j \neq 0$, with all indices understood modulo $r$,

$$\langle v, \Pi^j v\rangle = c_j - m\big(m_{k+j}(w) + m_{k-j}(w)\big) + m^2 a_j(w),$$
$$\langle w, \Pi^j v\rangle\langle \Pi^j w, v\rangle = \big(m_{k+j}(w) - m a_j(w)\big)\big(m_{k-j}(w) - m a_j(w)\big).$$

Using (36), the Cauchy–Schwarz inequality, and $|a_j(w)| \le 1$, we obtain

$$\sum_{j\neq 0} |\langle v, \Pi^j v\rangle| \le r\mu_* + 2|m| \sum_{j=0}^{r-1} |m_j(w)| + rm^2 \le r\mu_* + 2\sqrt{2r}|m| + rm^2, \tag{39}$$

and

$$\begin{aligned}\sum_{j\neq 0} \big|\langle w, \Pi^j v\rangle\langle \Pi^j w, v\rangle\big| &\le \sum_{j=0}^{r-1} |m_{k+j}(w) m_{k-j}(w)| + 2|m| \sum_{j=0}^{r-1} |m_j(w)| + rm^2 \\ &\le 2 + 2\sqrt{2r}|m| + rm^2.\end{aligned} \tag{40}$$

Here, for the first term in the last display, Cauchy–Schwarz and (36) give

$$\sum_{j=0}^{r-1} |m_{k+j}(w) m_{k-j}(w)| \le \left(\sum_j m_{k+j}(w)^2\right)^{1/2} \left(\sum_j m_{k-j}(w)^2\right)^{1/2} \le 2.$$

Since $\phi'$ and $\phi''$ are bounded on $[-1, 1]$, combining (38)–(40) gives

$$\mathbb{E}Q^2 \le C_\sigma \left(1 + r\mu_* + \sqrt{r}\,|m| + rm^2\right).$$

On $\mathcal{G}_d$, $r\mu_* = o(1)$, and $\sqrt{r}\,|m| \le (1 + rm^2)/2$. Therefore,

$$\mathbb{E}Q^2 \le C_\sigma(1 + rm^2). \tag{41}$$

The polynomial $Q$ has degree at most $P$ in the Gaussian input. Hence, by Gaussian hypercontractivity,

$$\|Q\|_{L^4} \le 3^{P/2} \|Q\|_{L^2}.$$

Using Cauchy–Schwarz, (37), and (41),

$$\mathbb{E}\left[\langle \Pi^k w^*, G\rangle^2\right] = \mathbb{E}[y^2 Q^2] \le \|y\|_{L^4}^2 \|Q\|_{L^4}^2 \le C_\sigma(1 + rm^2).$$

Now

$$\xi_{k,t} = -\left(\langle \Pi^k w^*, G_t\rangle - \mathbb{E}[\langle \Pi^k w^*, G_t\rangle \mid \mathcal{F}_{t-1}]\right).$$

Centering can only decrease the second moment, and therefore

$$\mathbb{E}[\xi_{k,t}^2 \mid \mathcal{F}_{t-1}] \le C_\sigma \left(1 + r\, m_k(w_{t-1})^2\right),$$

which proves (34).

Conditionally on $\mathcal{F}_{t-1}$, the random variable $\langle \Pi^k w^*, G_t\rangle$ is a polynomial of degree at most $2P$ in the fresh Gaussian input $x_t$. Subtracting its conditional expectation does not increase its degree, so $\xi_{k,t}$ is also a Gaussian polynomial of degree at most $2P$. Hence, for every $\nu \ge 2$, Gaussian hypercontractivity yields

$$\|\xi_{k,t}\|_{L^\nu(\cdot|\mathcal{F}_{t-1})} \le (\nu - 1)^P \|\xi_{k,t}\|_{L^2(\cdot|\mathcal{F}_{t-1})}.$$

Using (34),

$$(\mathbb{E}[|\xi_{k,t}|^\nu \mid \mathcal{F}_{t-1}])^{1/\nu} \le C_\sigma \nu^P \sqrt{1 + r\, m_k(w_{t-1})^2} \le C_\sigma \nu^P \left(1 + \sqrt{r}\,|m_k(w_{t-1})|\right),$$

which proves (35). □

### D.3 UNIFORM CONTROL OF THE STOCHASTIC GRADIENT

We finally establish a high-probability bound on the norm of the stochastic gradient, uniformly over polynomially many SGD iterations.

**Lemma D.5.** *For every fixed $C > 0$ there exists $C_{\sigma,C} < \infty$ such that, for every deterministic horizon $T \le d^C$,*

$$\mathbb{P}\left(\max_{1\le t\le T} \|G_t\| > C_{\sigma,C}(\log d)^P\sqrt{d}\right) = o(1).$$

*Proof.* We first work conditionally on a teacher $w^* \in \mathcal{G}_d$ and establish the uniform second-moment bound

$$\sup_{w\in\mathbb{S}^{d-1}} \mathbb{E}\|G(w;x,y)\|^2 \le C_\sigma d. \tag{42}$$

All expectations in this first part are with respect to a fresh input $x \sim \mathcal{N}(0, \mathrm{I}_d)$. Fix $w$ and write $\mathrm{P}_w \coloneqq \mathrm{I}_d - ww^\top$. Since

$$G(w;x,y) = -y\nabla_w f_w(x) = -y\,\mathrm{P}_w\widehat{\nabla}_w f_w(x),$$

Cauchy–Schwarz inequality gives

$$\mathbb{E}\|G\|^2 = \mathbb{E}\left[y^2\|\mathrm{P}_w\widehat{\nabla}_w f_w(x)\|^2\right] \le \|y\|_{L^4}^2\left(\mathbb{E}\|\mathrm{P}_w\widehat{\nabla}_w f_w(x)\|^4\right)^{1/2}. \tag{43}$$

By (37), $\|y\|_{L^4} \le C_\sigma$, uniformly over teachers in $\mathcal{G}_d$. Choose an orthonormal basis $(v_1,\dots,v_{d-1})$ of $w^\perp$, and set $Q_i \coloneqq \langle v_i, \widehat{\nabla}_w f_w(x)\rangle$. Then

$$\|\mathrm{P}_w\widehat{\nabla}_w f_w(x)\|^2 = \sum_{i=1}^{d-1} Q_i^2.$$

Hence, by Minkowski's inequality in $L^2$,

$$\left(\mathbb{E}\|\mathrm{P}_w\widehat{\nabla}_w f_w(x)\|^4\right)^{1/2} = \left\|\sum_{i=1}^{d-1} Q_i^2\right\|_{L^2} \le \sum_{i=1}^{d-1}\left\|Q_i^2\right\|_{L^2} = \sum_{i=1}^{d-1}\|Q_i\|_{L^4}^2.$$

Each $Q_i$ is a polynomial of degree at most $P$ in the Gaussian input $x$. Gaussian hypercontractivity therefore yields

$$\|Q_i\|_{L^4} \le 3^{P/2}\,\|Q_i\|_{L^2},$$

and consequently

$$\left(\mathbb{E}\|\mathrm{P}_w\widehat{\nabla}_w f_w(x)\|^4\right)^{1/2} \le 3^P\sum_{i=1}^{d-1}\mathbb{E}\left[Q_i^2\right]. \tag{44}$$

It remains to control the sum on the right-hand side of (44). By Lemma D.3,

$$\sum_{i=1}^{d-1}\mathbb{E}\left[Q_i^2\right] = \sum_{j=0}^{r-1}\left(\phi'(a_j(w))\sum_{i=1}^{d-1}\langle v_i,\Pi^j v_i\rangle + \phi''(a_j(w))\sum_{i=1}^{d-1}\langle w,\Pi^j v_i\rangle\langle\Pi^j w, v_i\rangle\right),$$

where $a_j(w) = \langle w, \Pi^j w\rangle$. For the first sum, since $\mathrm{P}_w = \sum_{i=1}^{d-1} v_i v_i^\top$, we have

$$\sum_{i=1}^{d-1}\langle v_i,\Pi^j v_i\rangle = \mathrm{Tr}(\mathrm{P}_w\Pi^j).$$

If $j = 0$, this equals $d-1$. If $1 \le j < r$, then $\mathrm{Tr}(\Pi^j) = 0$, and therefore

$$\mathrm{Tr}(\mathrm{P}_w\Pi^j) = -\langle w,\Pi^j w\rangle = -a_j(w).$$

As a result,

$$\sum_{j=0}^{r-1}\left|\sum_{i=1}^{d-1}\langle v_i,\Pi^j v_i\rangle\right| \le d + r.$$

For the second sum,

$$\sum_{i=1}^{d-1}\langle w,\Pi^j v_i\rangle\langle \Pi^j w, v_i\rangle = \langle \mathrm{P}_w\Pi^{-j}w, \mathrm{P}_w\Pi^j w\rangle,$$

and hence its absolute value is at most one. Since $\phi'$ and $\phi''$ are uniformly bounded on $[-1,1]$,

$$\sum_{i=1}^{d-1}\mathbb{E}\left[Q_i^2\right] \le C_\sigma(d+r) \le C_\sigma d,$$

where we used $r = d^\delta \le d$. Combining this estimate with (44) and (43) proves (42).

We now derive higher conditional moments. Fix $t \ge 1$ and condition on $\mathcal{F}_{t-1}$. Then $w_{t-1}$ is fixed and the only randomness comes from the fresh Gaussian input $x_t$. Choose an orthonormal basis $(v_1,\dots,v_{d-1})$ of $w_{t-1}^\perp$, and set $R_i := \langle v_i, G_t\rangle$. Each $R_i$ is a polynomial of degree at most $2P$ in $x_t$. Thus, for every $\nu \ge 2$, Gaussian hypercontractivity gives

$$\|R_i\|_{L^\nu(\,\cdot\,|\mathcal{F}_{t-1})} \le (\nu-1)^P \|R_i\|_{L^2(\,\cdot\,|\mathcal{F}_{t-1})}.$$

Using Minkowski in $L^{\nu/2}$,

$$\begin{aligned}
(\mathbb{E}[|G_t|^\nu \mid \mathcal{F}_{t-1}])^{2/\nu} &= \left\|\sum_{i=1}^{d-1} R_i^2\right\|_{L^{\nu/2}} \\
&\le \sum_{i=1}^{d-1}\|R_i\|_{L^\nu}^2 \\
&\le (\nu-1)^{2P}\sum_{i=1}^{d-1}\|R_i\|_{L^2}^2 \\
&= (\nu-1)^{2P}\mathbb{E}[|G_t|^2 \mid \mathcal{F}_{t-1}] \\
&\le C_\sigma d\,\nu^{2P},
\end{aligned}$$

where the last inequality follows from (42). Therefore

$$(\mathbb{E}[|G_t|^\nu \mid \mathcal{F}_{t-1}])^{1/\nu} \le C_\sigma\sqrt{d}\nu^P. \tag{45}$$

Take $\nu_d = \lceil c\log d\rceil$ for a fixed $c > 0$. By conditional Markov's inequality and (45), for every $D > 0$ one can choose $L = L(\sigma, D) < \infty$ sufficiently large so that

$$\mathbb{P}\left(\|G_t\| > L\sqrt{d}\,(\log d)^P \,\middle|\, \mathcal{F}_{t-1}\right) \le d^{-D}$$

uniformly in $t$, for all sufficiently large $d$, on the event $\mathcal{G}_d$. Choose $D > C+1$. Taking expectations and applying a union bound over $1 \le t \le T \le d^C$ gives

$$\mathbb{P}\left(\left\{\max_{1\le t\le T}\|G_t\| > L\sqrt{d}\,(\log d)^P\right\}\cap\mathcal{G}_d\right) \le d^{C-D} = o(1).$$

Finally, $\mathbb{P}(\mathcal{G}_d^{\mathrm{c}}) = o(1)$ by Lemma C.1, so that

$$\mathbb{P}\left(\max_{1\le t\le T}\|G_t\| > L\sqrt{d}\,(\log d)^P\right) = o(1).$$

This proves the lemma. □

## E Proof of weak recovery for the tied student

In this section we prove part (T) of Theorem B.1. The proof shows that the leading polynomial dependence of the recovery time is determined by the information exponent $p$: higher Hermite components affect the constants and logarithmic factors, but not the leading $d^{p-1}$ sample complexity exponent.

### E.1 PROOF STRATEGY

The proof proceeds in two main steps. The first step selects and amplifies one of the $r$ teacher correlations from its random initialization scale until the trajectory enters a suitable intermediate region. The second step starts from any point in this region and propagates the selected correlation to the fixed recovery level $\rho$.

Fix $B > \frac{2}{p-2}$, depending only on $p$, and define the two intermediate scales

$$\alpha_d := \gamma_d(\log d)^{2/(p-2)}, \qquad \beta_d := \gamma_d(\log d)^B.$$

Then

$$\gamma_d \ll \alpha_d \ll \beta_d \ll 1.$$

We next introduce notation for the correlation selected at initialization. Define

$$i_* := \min \arg\max_{0 \le k < r} q_k(w_0),$$

where $q_k(w)$ denotes the relevant correlation defined in (7). Define the corresponding oriented correlation by

$$\widetilde{m}_{i_*}(w_t) := \begin{cases} \operatorname{sgn}(m_{i_*}(w_0))m_{i_*}(w_t), & p \text{ even}, \\ m_{i_*}(w_t), & p \text{ odd}, \end{cases} \tag{46}$$

with the convention $\operatorname{sgn}(0) = 1$. Lemma C.3 shows that, with probability tending to one, the maximizer is unique, $q_{i_*}(w_0) \asymp \gamma_d$, and is separated from all competing relevant correlations by a quantitative relative gap.

Let

$$\mathcal{F}_t := \sigma\left(w^*, w_0, (x_\ell, y_\ell)_{1 \le \ell \le t}\right)$$

denote the natural filtration. Then $i_*$ is $\mathcal{F}_0$-measurable, $w_t$ is $\mathcal{F}_t$-measurable, and each hitting time $\tau_a$ is an $(\mathcal{F}_t)_{t \ge 0}$-stopping time.

The first phase shows that the small advantage present at initialization survives the stochastic fluctuations and interaction errors and is amplified until the selected correlation reaches the scale $\beta_d$, while all competing correlations remain at the smaller scale $\alpha_d$.

**Proposition E.1** (Microscopic selection)**.** *Suppose that Assumptions 1(i), 2, and 3 hold. There exists $A_* = A_*(p, P) \ge (p-2)/2$ such that the following holds for every fixed $A > A_*$. Run tied spherical SGD* (5) *with step size*

$$\eta_d = d^{-p/2}(\log d)^{-A}.$$

*Then there exist constants $C_\alpha, C_\beta < \infty$, depending only on the fixed model parameters, such that, with probability $1 - o(1)$,*

$$\tau_{\beta_d} \le \frac{C_\beta}{\eta_d \gamma_d^{p-2}},$$

*and*

$$\beta_d \le \widetilde{m}_{i_*}(w_{\tau_{\beta_d}}) \le 2\beta_d, \qquad \max_{k \ne i_*} |m_k(w_{\tau_{\beta_d}})| \le C_\alpha \alpha_d.$$

*Moreover, $i_*$ remains the unique relevant leader up to time $\tau_{\beta_d}$.*

The proof of Proposition E.1 is given in Appendix E.2.

Since $\tau_{\beta_d}$ is a stopping time and the SGD recursion is driven by i.i.d. observations, the strong Markov property implies that, conditionally on $\mathcal{F}_{\tau_{\beta_d}}$, the post-$\tau_{\beta_d}$ trajectory has the law of tied spherical SGD started from $w_{\tau_{\beta_d}}$ and driven by fresh observations, with $w^*$ fixed. This allows us to restart the analysis at $\tau_{\beta_d}$.

**Proposition E.2** (Growth to macroscopic size)**.** *Suppose that Assumptions 1(i), 2, and 3 hold. Fix $\rho \in (0, 1)$, independently of $d$. There exists $A_* = A_*(p, P) \ge (p-2)/2$ such that the following holds for every fixed $A > A_*$. Run tied spherical SGD* (5) *with step size*

$$\eta_d = d^{-p/2}(\log d)^{-A}.$$

*Then there exist $C_\rho < \infty$ and a deterministic sequence $\varepsilon_d \to 0$ such that, almost surely on*

$$\mathcal{G}_d \cap \left\{ \beta_d \le \widetilde{m}_{i_*}(w_{\tau_{\beta_d}}) \le 2\beta_d, \quad \max_{k \neq i_*} |m_k(w_{\tau_{\beta_d}})| \le C_\alpha \alpha_d \right\},$$

*it holds that*

$$\mathbb{P}\left( \tau_\rho - \tau_{\beta_d} \le \frac{C_\rho}{\eta_d \beta_d^{p-2}} \,\middle|\, \mathcal{F}_{\tau_{\beta_d}} \right) \ge 1 - \varepsilon_d,$$

*where $\mathcal{G}_d$ is the teacher-geometry event of Lemma C.1. Moreover, with the same conditional probability lower bound, $i_*$ remains the unique favorable leader up to time $\tau_\rho$, while all competing overlaps remain microscopic.*

The proof of Proposition E.2 is given in Appendix E.3. We now provide the proof of part (T) of Theorem B.1.

*Proof of part* (T) *of Theorem B.1.* Let

$$\mathcal{E}_d := \mathcal{G}_d \cap \left\{ \tau_{\beta_d} \le \frac{C_\beta}{\eta_d \gamma_d^{p-2}}, \quad \beta_d \le \widetilde{m}_{i_*}(w_{\tau_{\beta_d}}) \le 2\beta_d, \quad \max_{k \neq i_*} |m_k(w_{\tau_{\beta_d}})| \le C_\alpha \alpha_d \right\}.$$

By Proposition E.1 and Lemma C.1, $\mathbb{P}(\mathcal{E}_d) = 1 - o(1)$. Moreover, $\mathcal{E}_d \in \mathcal{F}_{\tau_{\beta_d}}$.

By Proposition E.2 and the tower property,

$$\mathbb{P}\left( \mathcal{E}_d \cap \left\{ \tau_\rho - \tau_{\beta_d} > \frac{C_\rho}{\eta_d \beta_d^{p-2}} \right\} \right) = \mathbb{E}\left[ \mathbf{1}_{\mathcal{E}_d} \mathbb{P}\left( \tau_\rho - \tau_{\beta_d} > \frac{C_\rho}{\eta_d \beta_d^{p-2}} \,\middle|\, \mathcal{F}_{\tau_{\beta_d}} \right) \right] \le \varepsilon_d.$$

Hence, with probability $1 - o(1)$,

$$\tau_\rho \le \frac{C_\beta}{\eta_d \gamma_d^{p-2}} + \frac{C_\rho}{\eta_d \beta_d^{p-2}}.$$

Since $\beta_d = \gamma_d (\log d)^B$, we have

$$\frac{1}{\eta_d \beta_d^{p-2}} = (\log d)^{-B(p-2)} \frac{1}{\eta_d \gamma_d^{p-2}} = o\left( \frac{1}{\eta_d \gamma_d^{p-2}} \right).$$

Therefore, after increasing the constant $C$ if necessary,

$$\tau_\rho \le \frac{C}{\eta_d \gamma_d^{p-2}}$$

with probability $1 - o(1)$. □

### E.2 Microscopic selection

We now study the dynamics while all correlations remain microscopic and prove Proposition E.1. According to (29), we have the evolution

$$m_k(w_t) = m_k(w_0) + p a_p^2 \eta_d \sum_{\ell=1}^{t} m_k(w_{\ell-1})^{p-1} + E_{k,t},$$

where $E_{k,t}$ is defined in (30). Set

$$T_0 := \left\lceil \frac{C_0}{\eta_d \gamma_d^{p-2}} \right\rceil, \tag{47}$$

where $C_0 > 0$ is fixed and sufficiently large, and define

$$\hat{\tau}_{\beta_d} := \inf \left\{ t \ge 0 \colon \max_{0 \le k \le r-1} |m_k(w_t)| \ge \beta_d \right\}.$$

For even $p$, $\hat{\tau}_{\beta_d} = \tau_{\beta_d}$. For odd $p$ we prove below that the first absolute crossing is necessarily positive, and hence the two stopping times agree up to $T_0$.

**Lemma E.3.** *Suppose $\beta_d = \gamma_d(\log d)^B$ with $B > 2/(p-2)$ and $\eta_d = d^{-p/2}(\log d)^{-A}$ with $A > A_*(p, P)$. Then, with probability $1 - o(1)$,*

$$\max_{0\le k\le r-1}\max_{0\le t\le \hat\tau_{\beta_d}\wedge T_0} |E_{k,t}| = o\big((\log d)^{-2}\gamma_d\big).$$

**Corollary E.4.** *With probability $1 - o(1)$,*

$$\hat\tau_{\beta_d} \wedge T_0 = \tau_{\beta_d} \wedge T_0.$$

*Consequently Lemma E.3 remains valid with $\hat\tau_{\beta_d}$ replaced by $\tau_{\beta_d}$.*

*Proof.* For even $p$ the two stopping times coincide by definition. Suppose $p$ is odd. Then $p-1$ is even, so the leading term in (29) is nonnegative. Hence, for $t \le \hat\tau_{\beta_d} \wedge T_0$,

$$m_k(w_t) \ge m_k(w_0) - |E_{k,t}|.$$

On the initialization event, $\max_k |m_k(w_0)| \le C\gamma_d$, while Lemma E.3 gives $\max_{k,t} |E_{k,t}| = o(\gamma_d)$. Thus

$$m_k(w_t) \ge -C\gamma_d - o(\gamma_d) > -\beta_d$$

for all sufficiently large $d$, uniformly in $k$ and in the stopped time interval. No coordinate can therefore exit through $-\beta_d$ before $T_0$, proving the claim. □

**Lemma E.5.** *With probability $1 - o(1)$,*

$$\max_{0\le k\le r-1}\max_{0\le t\le \hat\tau_{\beta_d}\wedge T_0}\left|\eta_d\sum_{\ell=1}^{t}\xi_{k,\ell}\right| \le C\eta_d\left(\sqrt{T_0\log d} + (\log d)^{P+1}\right).$$

*Proof.* For $1 \le \ell \le T_0$ set

$$Y_{k,\ell} = \mathbf{1}_{\{\ell\le\hat\tau_{\beta_d}\}}\xi_{k,\ell}.$$

Then $(Y_{k,\ell})_\ell$ is a martingale difference sequence and

$$\max_{t\le\hat\tau_{\beta_d}\wedge T_0}\left|\sum_{\ell=1}^{t}\xi_{k,\ell}\right| = \max_{t\le T_0}\left|\sum_{\ell=1}^{t}Y_{k,\ell}\right|.$$

On $\{\ell \le \hat\tau_{\beta_d}\}$, $|m_k(w_{\ell-1})| < \beta_d$. Since $r\beta_d^2 = o(1)$, Lemma D.4 gives, for all sufficiently large $d$,

$$\mathbb{E}[Y_{k,\ell}^2 \mid \mathcal{F}_{\ell-1}] \le C_\sigma, \qquad (\mathbb{E}[|Y_{k,\ell}|^q \mid \mathcal{F}_{\ell-1}])^{1/q} \le C_\sigma q^P. \tag{48}$$

Choose $q_d = \lceil c\log d\rceil$ and $R_d = C(\log d)^P$, with $c, C$ sufficiently large. Conditional Markov's inequality gives

$$\mathbb{P}(|Y_{k,\ell}| > R_d \mid \mathcal{F}_{\ell-1}) \le 4^{-q_d}.$$

Since $T_0 = d^{p-1}\,\mathrm{polylog}(d) \le d^p$ and $r \le d$, a union bound shows that, with probability $1 - o(1)$,

$$|Y_{k,\ell}| \le R_d \qquad \text{for all } k < r,\ \ell \le T_0.$$

Define the truncated and recentered increments

$$\widehat{Y}_{k,\ell} = Y_{k,\ell}\mathbf{1}_{\{|Y_{k,\ell}|\le R_d\}} - \mathbb{E}\big[Y_{k,\ell}\mathbf{1}_{\{|Y_{k,\ell}|\le R_d\}} \mid \mathcal{F}_{\ell-1}\big].$$

Then $|\widehat{Y}_{k,\ell}| \le 2R_d$ and the predictable quadratic variation up to $T_0$ is at most $C_\sigma T_0$. Freedman's maximal inequality therefore gives, after a union bound over $k < r$,

$$\max_{k<r}\max_{t\le T_0}\left|\sum_{\ell=1}^{t}\widehat{Y}_{k,\ell}\right| \le C\left(\sqrt{T_0\log d} + R_d\log d\right)$$

with probability $1 - o(1)$.

Finally, the truncation bias satisfies

$$\big|\mathbb{E}\big[Y_{k,\ell}\mathbf{1}_{\{|Y_{k,\ell}|\le R_d\}} \mid \mathcal{F}_{\ell-1}\big]\big| \le R_d 4^{-q_d},$$

so $T_0 R_d 4^{-q_d} = o(1)$ after increasing $c$. Since $R_d = C(\log d)^P$, the claim follows after multiplying by $\eta_d$. □

*Proof of Lemma E.3.* Set $\nu_d = (\log d)^{-2}$. We control the four terms in (30).

**Higher-Hermite remainder.** For $\ell \le \hat{\tau}_{\beta_d} \wedge T_0$, (32) gives

$$|R_\sigma(m_k(w_{\ell-1}))| \le C_\sigma \beta_d^p.$$

Therefore

$$\frac{\max_{k,t} \left| \eta_d \sum_{\ell=1}^{t} R_\sigma(m_k(w_{\ell-1})) \right|}{\nu_d \gamma_d} \le C_\sigma \, \frac{\eta_d T_0 \beta_d^p}{\nu_d \gamma_d} \lesssim \frac{\gamma_d}{\nu_d} \left( \frac{\beta_d}{\gamma_d} \right)^p = o(1).$$

**Interaction term.** On the same stopped interval, (27) and (32) imply

$$|\Gamma_k(w_{\ell-1})| \le C_\sigma \big( r\mu_* + r\beta_d^2 \big) \beta_d^{p-1}.$$

Hence

$$\frac{\max_{k,t} \left| \eta_d \sum_{\ell=1}^{t} \Gamma_k(w_{\ell-1}) \right|}{\nu_d \gamma_d} \le C_\sigma \big( r\mu_* + r\beta_d^2 \big) \nu_d^{-1} \left( \frac{\beta_d}{\gamma_d} \right)^{p-1} = o(1),$$

because $r\mu_*$ and $r\beta_d^2$ decay polynomially in $d$, whereas all remaining factors are polylogarithmic.

**Normalization term.** Since $T_0$ is polynomial in $d$, Lemma D.5 gives, with probability $1 - o(1)$,

$$\max_{1 \le \ell \le T_0} \|G_\ell\| \le C_\sigma \sqrt{d} \, (\log d)^P. \tag{49}$$

Our choice of $A$ ensures $C_\sigma \sqrt{d} \, (\log d)^P = o(\beta_d)$. Thus, also at the exit step,

$$|m_k(w_\ell)| \le \beta_d + C_\sigma \eta_d \sqrt{d} \, (\log d)^P \le 2\beta_d.$$

Using $\sqrt{1+x} - 1 \le x/2$,

$$\max_{k,t} \left| \sum_{\ell=1}^{t} \big( \sqrt{1 + \eta_d^2 \|G_\ell\|^2} - 1 \big) m_k(w_\ell) \right| \le C \eta_d^2 T_0 d (\log d)^{2P} \beta_d.$$

After division by $\nu_d \gamma_d$ this is at most

$$C_\sigma (\log d)^{-A-(p-2)/2+2P+B+2},$$

which is $o(1)$ for $A > A_*(p, P)$.

**Martingale term.** By Lemma E.5, after division by $\nu_d \gamma_d$ the first stochastic contribution is

$$\frac{\eta_d \sqrt{T_0 \log d}}{\nu_d \gamma_d} \lesssim (\log d)^{-A/2+5/2-p/4} = o(1)$$

for $A$ sufficiently large, while

$$\frac{\eta_d (\log d)^{P+1}}{\nu_d \gamma_d} \lesssim d^{-(p-1)/2} (\log d)^{-A+P+5/2} = o(1).$$

Combining the four estimates proves

$$\max_{k<r} \max_{t \le \hat{\tau}_{\beta_d} \wedge T_0} |E_{k,t}| = o(\nu_d \gamma_d)$$

with probability $1 - o(1)$, uniformly over admissible teachers and initializations. □

We now give the proof of Proposition E.1.

*Proof of Proposition E.1.* Set $\nu_d = (\log d)^{-2}$ and work on the intersection of the high-probability events from Lemma C.3, Lemma E.3, Corollary E.4, and (49). Let

$$m_0 \coloneqq M_\sigma(w_0), \qquad c\gamma_d \le m_0 \le C\gamma_d,$$

and

$$e_d \coloneqq \max_{k<r} \max_{t \le \tau_{\beta_d} \wedge T_0} |E_{k,t}| = o(\nu_d m_0).$$

Then, for every $t \le \tau_{\beta_d} \wedge T_0$,

$$m_k(w_0) + pa_p^2\eta_d \sum_{\ell=0}^{t-1} m_k(w_\ell)^{p-1} - e_d \le m_k(w_t) \le m_k(w_0) + pa_p^2\eta_d \sum_{\ell=0}^{t-1} m_k(w_\ell)^{p-1} + e_d. \quad (50)$$

Recall the oriented correlation $\widetilde{m}_{i_*}(w_t)$ defined in (46). Then $\widetilde{m}_{i_*}(w_0) = m_0$ and

$$\widetilde{m}_{i_*}(w_t) = m_0 + pa_p^2\eta_d \sum_{\ell=0}^{t-1} \widetilde{m}_{i_*}(w_\ell)^{p-1} + \overline{E}_{*,t}, \qquad |\overline{E}_{*,t}| \le e_d.$$

Define

$$L_t \coloneqq m_0 - e_d + pa_p^2\eta_d \sum_{\ell=0}^{t-1} \widetilde{m}_{i_*}(w_\ell)^{p-1}.$$

Then $\widetilde{m}_{i_*}(w_t) \ge L_t$, $L_0 > 0$, and

$$L_t - L_{t-1} = pa_p^2\eta_d \widetilde{m}_{i_*}(w_{t-1})^{p-1} \ge pa_p^2\eta_d L_{t-1}^{p-1}. \quad (51)$$

For a favorable competitor, set $b_{k,t} = |m_k(w_t)|$ when $p$ is even. When $p$ is odd, first consider competitors with $m_k(w_0) \ge 0$ and set $b_{k,t} = m_k(w_t)$ as long as this remains nonnegative. Define

$$U_{k,t} \coloneqq b_{k,0} + e_d + pa_p^2\eta_d \sum_{\ell=0}^{t-1} b_{k,\ell}^{p-1}.$$

Then $b_{k,t} \le U_{k,t}$ and

$$U_{k,t} - U_{k,t-1} \le pa_p^2\eta_d U_{k,t-1}^{p-1}.$$

The initialization gap gives

$$L_0 - U_{k,0} \ge \nu_d m_0 - 2e_d > 0,$$

so monotonicity of $x \mapsto x^{p-1}$ on $\mathbb{R}_+$ implies inductively that the selected coordinate remains strictly ahead of every favorable competitor.

**Time to reach $\beta_d$.** Set $q = p - 2$ and

$$t_\beta \coloneqq \inf\{t \ge 0 : L_t \ge \beta_d\}.$$

Before this time, the mean value theorem and (51) give

$$L_{t-1}^{-q} - L_t^{-q} \ge \frac{qpa_p^2\eta_d}{(1 + pa_p^2\eta_d\beta_d^q)^{q+1}}.$$

Hence

$$t_\beta \le 1 + \frac{(1 + pa_p^2\eta_d\beta_d^q)^{q+1}}{qpa_p^2\eta_d L_0^q}\left[1 - \left(\frac{L_0}{\beta_d}\right)^q\right]. \quad (52)$$

Since $pa_p^2\eta_d\beta_d^q = o(\nu_d)$ and $L_0 \asymp \gamma_d$,

$$t_\beta \lesssim \frac{1}{\eta_d \gamma_d^{p-2}}.$$

Taking $C_0$ sufficiently large gives $t_\beta \le T_0$, and therefore $\tau_{\beta_d} \le t_\beta$.

**Size of the competitors.** The initialization gap implies

$$\frac{U_{k,0}}{L_0} \le 1 - (1 - o(1))\nu_d.$$

Iterating the upper scalar recursion gives

$$U_{k,t} \le U_{k,0} \left(1 - qpa_p^2 \eta_d U_{k,0}^q t\right)^{-1/q}$$

whenever the denominator is positive. Combining this with (52), and using the extra $+1$ term only through $qpa_p^2 \eta_d U_{k,0}^q = o(\nu_d)$, yields

$$1 - qpa_p^2 \eta_d U_{k,0}^q t \ge c_p \nu_d, \qquad t \le \tau_{\beta_d},$$

for some $c_p > 0$. Consequently

$$\max_{t \le \tau_{\beta_d}} |m_k(w_t)| \lesssim \gamma_d \nu_d^{-1/(p-2)} = \alpha_d = o(\beta_d). \tag{53}$$

Suppose now that $p$ is odd and $m_k(w_0) < 0$. Let

$$\nu_k := \inf\{t \ge 0 \colon m_k(w_t) \ge 0\}.$$

Before $\nu_k$, (29) gives $|m_k(w_t)| \lesssim \gamma_d$. If $\nu_k \le \tau_{\beta_d}$ and $m_k(w_{\nu_k - 1}) = -u$, then $u < \beta_d$ and

$$-u + pa_p^2 \eta_d u^{p-1} = -u(1 - pa_p^2 \eta_d u^{p-2}) \le 0.$$

The crossing overshoot is therefore at most $2e_d$. Starting from this size, the positive scalar envelope remains $O(e_d)$ because

$$pa_p^2 \eta_d e_d^{p-2} \tau_{\beta_d} \lesssim \left(\frac{e_d}{\gamma_d}\right)^{p-2} = o(1).$$

This proves (53) for initially negative competitors as well. The same argument handles later sign changes.

Finally, the preceding comparison shows that $i_*$ is the coordinate that reaches the threshold. Immediately before the crossing its favorable magnitude is below $\beta_d$, and the exact normalized update gives

$$\beta_d \le M_\sigma(w_{\tau_{\beta_d}}) \le \beta_d + \eta_d \|G_{\tau_{\beta_d}}\|.$$

By (49), $\eta_d \|G_{\tau_{\beta_d}}\| = o(\beta_d)$, so

$$M_\sigma(w_{\tau_{\beta_d}}) = \beta_d(1 + o(1)).$$

This proves Proposition E.1. □

### E.3 Growth to a macroscopic level

We now prove Proposition E.2. By Proposition E.1, with probability $1 - o(1)$ there is a unique index $i_*$ such that

$$\beta_d \le \widetilde{m}_{i_*}(w_{\tau_{\beta_d}}) \le 2\beta_d, \qquad \max_{k \ne i_*} |m_k(w_{\tau_{\beta_d}})| \le C_\alpha \alpha_d,$$

where $\widetilde{m}_{i_*}(w_t)$ is the oriented correlation defined in (46). We may thus orient all correlations according to the selected direction and suppress the bars below, so that $m_{i_*}(w_t) = \widetilde{m}_{i_*}(w_t) > 0$.

The competitor dynamics contain a contractive factor generated by the selected overlap. We first record the deterministic estimate used to exploit this damping.

**Lemma E.6** (Damping preserves partial-sum bounds)**.** *Suppose real sequences satisfy*

$$d_t x_{t+1} = q_t x_t + F_t + N_{t+1} - N_t, \qquad 0 \le q_t \le 1, \quad d_t \ge 1, \quad N_0 = 0.$$

*Then, for every $t \ge 0$,*

$$|x_t| \le |x_0| + \sum_{\ell=0}^{t-1} |F_\ell| + 2 \max_{0 \le \ell \le t} |N_\ell|.$$

*Proof.* Iterating the recursion, an increment introduced at time $\ell$ is multiplied at time $t$ by

$$\omega_{\ell,t} \coloneqq \frac{1}{d_\ell} \prod_{j=\ell+1}^{t-1} \frac{q_j}{d_j}, \qquad 0 \le \ell < t.$$

The weights lie in $[0,1]$ and are nondecreasing in $\ell$. Summation by parts therefore gives

$$\left| \sum_{\ell=0}^{t-1} \omega_{\ell,t}(N_{\ell+1} - N_\ell) \right| \le 2 \max_{0 \le \ell \le t} |N_\ell|.$$

The coefficients multiplying $x_0$ and the forcing terms are at most one, which proves the claim. □

*Proof of Proposition E.2.* Condition on $\mathcal{F}_{\tau_{\beta_d}}$, then

$$\beta_d \le \widetilde{m}_{i_*}(w_0) \le 2\beta_d, \qquad \max_{k \ne i_*} |m_k(w_0)| \le C_\alpha \alpha_d.$$

Set

$$c \coloneqq \frac{p a_p^2 (1-\rho^2)}{4}, \qquad \varepsilon \coloneqq \frac{1}{4}, \qquad T \coloneqq \left\lceil \frac{C_T}{\eta_d \beta_d^{p-2}} \right\rceil,$$

where $C_T$ is chosen below. Define

$$D_t \coloneqq c\eta_d \sum_{\ell=0}^{t-1} m_{i_*}(w_\ell)^{p-1}$$

and

$$\tau \coloneqq T \wedge \inf \left\{ t \ge 0 : \begin{array}{l} m_{i_*}(w_t) \notin (\beta_d/2, \rho), \\ \text{or } \max_{k \ne i_*} |m_k(w_t)| \ge 2C_\alpha \alpha_d, \\ \text{or } D_t \ge 1 \end{array} \right\}.$$

For $t < \tau$, $0 < m_{i_*}(w_t) < \rho$ and $D_{t \wedge \tau} \le 1 + c\eta_d$.

**Step 1: Progress of the selected correlation.** From Lemma D.1,

$$\begin{aligned} -\langle \Pi^{i_*} w^*, \nabla \Phi(w_t) \rangle &= (1 - m_{i_*}(w_t)^2) \phi'(m_{i_*}(w_t)) \\ &\quad + \sum_{j \ne i_*} c_{j-i_*} \phi'(m_j(w_t)) - m_{i_*}(w_t) \sum_{j \ne i_*} m_j(w_t) \phi'(m_j(w_t)). \end{aligned}$$

For $m \in [0, \rho]$, (33) gives $\phi'(m) \ge p a_p^2 m^{p-1}$. For the competitors, (32) gives $|\phi'(m_j)| \le C_\sigma |m_j|^{p-1}$. Hence, for $t < \tau$,

$$-\langle \Pi^{i_*} w^*, \nabla \Phi(w_t) \rangle \ge p a_p^2 (1-\rho^2) m_{i_*}(w_t)^{p-1} - C_\sigma r \mu_* \alpha_d^{p-1} - C_\sigma r \alpha_d^p m_{i_*}(w_t).$$

Since $m_{i_*}(w_t) \ge \beta_d/2$, both error terms are $o(m_{i_*}(w_t)^{p-1})$. Thus, for all sufficiently large $d$,

$$-\langle \Pi^{i_*} w^*, \nabla \Phi(w_t) \rangle \ge 2c\, m_{i_*}(w_t)^{p-1}. \tag{54}$$

Lemma D.5 applies uniformly after the restart. Since $T$ is polynomial in $d$, with conditional probability $1 - o(1)$,

$$\max_{1 \le t \le T} \|G_t\| \le C_\sigma \sqrt{d}\, (\log d)^P. \tag{55}$$

Our choice of $A$ ensures

$$\frac{C_\sigma \eta_d \sqrt{d}\, (\log d)^P}{\beta_d} = o(1), \qquad \frac{C_\sigma^2 \eta_d d (\log d)^{2P}}{\beta_d^{p-2}} = o(1). \tag{56}$$

The normalization correction is therefore bounded by

$$\left| \big( \sqrt{1 + \eta_d^2 \|G_{t+1}\|^2} - 1 \big) m_{i_*}(w_{t+1}) \right| \le c\eta_d m_{i_*}(w_t)^{p-1}.$$

Combining with (54) gives

$$m_{i_*}(w_{t+1}) \geq m_{i_*}(w_t) + c\eta_d m_{i_*}(w_t)^{p-1} + \eta_d \xi_{i_*,t+1}, \qquad t < \tau. \tag{57}$$

Define the stopped martingales

$$N_{k,t} \coloneqq \eta_d \sum_{\ell=0}^{t-1} \mathbf{1}_{\{\ell<\tau\}} \xi_{k,\ell+1}.$$

For the selected coordinate,

$$\sum_{\ell<\tau} m_{i_*}(w_\ell)^2 \leq \frac{C}{\eta_d \beta_d^{p-3}},$$

because $m_{i_*}(w_\ell) \geq \beta_d/2$ and $D_\tau \leq 1 + c\eta_d$. Lemma D.4 yields

$$\mathbb{E}[N_{i_*,T}^2] \leq C_\sigma \eta_d^2 \left( T + \frac{r}{\eta_d \beta_d^{p-3}} \right) \leq C_\sigma \eta_d^2 T(1 + r\beta_d) = o(\beta_d^2).$$

Hence Doob's maximal inequality implies

$$\max_{0\leq n\leq T} |N_{i_*,n}| \leq \varepsilon \beta_d \tag{58}$$

with conditional probability $1 - o(1)$. Summing (57) gives, for every $t \leq \tau$,

$$m_{i_*}(w_t) \geq (1-\varepsilon)\beta_d + D_t. \tag{59}$$

Thus neither the lower boundary $\beta_d/2$ nor the cutoff $D_t = 1$ can trigger $\tau$.

**Step 2: Persistence of the microscopic separation.** For $k \neq i_*$ and $\ell < \tau$, $|m_k(w_\ell)| < 2C_\alpha \alpha_d$, so Lemma D.4 gives uniformly

$$\mathbb{E}[(\mathbf{1}_{\{\ell<\tau\}} \xi_{k,\ell+1})^2 \mid \mathcal{F}_\ell] \leq C_\sigma, \qquad \|\mathbf{1}_{\{\ell<\tau\}} \xi_{k,\ell+1}\|_{L^q(\,\cdot\,|\mathcal{F}_\ell)} \leq C_\sigma q^P.$$

Repeating the truncation–Freedman argument from Lemma E.5 gives, with conditional probability $1 - o(1)$,

$$\max_{k\neq i_*} \max_{0\leq t\leq T} |N_{k,t}| \leq C\eta_d \left( \sqrt{T \log d} + (\log d)^{P+1} \right) = o(\alpha_d). \tag{60}$$

For $t < \tau$ and $k \neq i_*$, isolate the damping generated by the selected coordinate in the exact recursion:

$$\sqrt{1 + \eta_d^2 \|G_{t+1}\|^2}\, m_k(w_{t+1}) = q_t m_k(w_t) + \eta_d F_{k,t} + \eta_d \xi_{k,t+1},$$

where

$$q_t \coloneqq 1 - \eta_d m_{i_*}(w_t) \phi'(m_{i_*}(w_t))$$

and

$$F_{k,t} \coloneqq \phi'(m_k(w_t)) + \sum_{j\neq k} c_{j-k} \phi'(m_j(w_t)) - m_k(w_t) \sum_{j\neq i_*} m_j(w_t) \phi'(m_j(w_t)).$$

Since $m_{i_*} > 0$ and $\phi'(m_{i_*}) > 0$, while $\eta_d \to 0$, we have $0 \leq q_t \leq 1$ for all sufficiently large $d$. Moreover, (32) implies

$$|F_{k,t}| \leq C_\sigma \left( \alpha_d^{p-1} + \mu_* m_{i_*}(w_t)^{p-1} + r\alpha_d^{p+1} \right) \leq C_\sigma \left( \alpha_d^{p-1} + \mu_* m_{i_*}(w_t)^{p-1} \right),$$

where we used $r\alpha_d^2 = o(1)$. Lemma E.6 therefore gives, for $t \leq \tau$,

$$|m_k(w_t)| \leq C_\alpha \alpha_d + C_\sigma \eta_d T \alpha_d^{p-1} + C_\sigma \mu_* \eta_d \sum_{\ell<t} m_{i_*}(w_\ell)^{p-1} + 2 \max_{\ell\leq t} |N_{k,\ell}|.$$

Now

$$\eta_d T \alpha_d^{p-2} = O\left( \left( \frac{\alpha_d}{\beta_d} \right)^{p-2} \right) = o(1),$$

$D_t \le 1 + c\eta_d$, $\mu_* = o(\alpha_d)$, and (60) is $o(\alpha_d)$. Hence

$$|m_k(w_t)| < 2C_\alpha \alpha_d$$

for every $k \neq i_*$ and $t \le \tau$, so no competitor can trigger the stopping time.

**Step 3: Reaching the target.** By (59),

$$m_{i_*}(w_t) \ge (1-\varepsilon)\beta_d + c\eta_d \sum_{\ell=0}^{t-1} m_{i_*}(w_\ell)^{p-1} =: L_t.$$

The standard discrete comparison for $x' = cx^{p-1}$ gives

$$t_\rho := \inf\{n : L_n \ge \rho\} \le \frac{C_{\sigma,\rho}}{\eta_d((1-\varepsilon)\beta_d)^{p-2}}.$$

Choosing $C_T$ sufficiently large ensures $t_\rho \le T$. Since the other stopping mechanisms have already been excluded, the selected coordinate reaches $\rho$ before $T$, and

$$\tau_\rho \le \left\lceil \frac{C_\rho}{\eta_d \beta_d^{p-2}} \right\rceil.$$

Throughout this interval,

$$m_{i_*}(w_t) \ge \frac{\beta_d}{2}, \qquad \max_{k \neq i_*} |m_k(w_t)| < 2C_\alpha \alpha_d.$$

Since $\alpha_d = o(\beta_d)$, $i_*$ remains the unique favorable leader. Therefore $M_\sigma(w_{\tau_\rho}) \ge \rho$. This proves Proposition E.2. □

# F Proofs for the augmented and untied students

We now prove parts (U) and (A) of Theorem B.1. The key observation is that, because the correlation loss is linear in the predictor, the gradient with respect to one untied weight does not involve the other weights. Consequently, both procedures reduce to one-vector dynamics.

## F.1 Componentwise reductions

For $u \in \mathbb{S}^{d-1}$, define the single-weight loss

$$L^{\mathrm{S}}(u; x, y) := -y\,\sigma(\langle u, x\rangle), \tag{61}$$

and its population counterpart

$$\Psi(u) := \mathbb{E}[L^{\mathrm{S}}(u; x, y)].$$

By linearity of the untied predictor,

$$\nabla_{w^j} L^{\mathsf{U}}(W; x, y) = \frac{1}{\sqrt{s}} \nabla L^{\mathrm{S}}(w^j; x, y), \qquad 1 \le j \le s. \tag{62}$$

For the augmented procedure, using $\langle w, \Pi^k x\rangle = \langle \Pi^{-k} w, x\rangle$ and the fact that $k \mapsto -k \pmod r$ permutes the orbit,

$$L^{\mathsf{A}}(W; x, y) = -\frac{y}{r\sqrt{s}} \sum_{j=1}^{s} \sum_{k=0}^{r-1} \sigma(\langle w^j, \Pi^k x\rangle) = \frac{1}{\sqrt{rs}} \sum_{j=1}^{s} L^{\mathsf{T}}(w^j; x, y). \tag{63}$$

Hence

$$\nabla_{w^j} L^{\mathsf{A}}(W; x, y) = \frac{1}{\sqrt{rs}} \nabla L^{\mathsf{T}}(w^j; x, y). \tag{64}$$

Using the step sizes in (9) and the identities in (62) and (64), we have

$$\widetilde{w}_t^{j,\mathsf{U}} = w_{t-1}^{j,\mathsf{U}} - \frac{\eta_d}{\sqrt{r}} \nabla L^{\mathrm{S}}(w_{t-1}^{j,\mathsf{U}}; x_t, y_t), \tag{65}$$

$$\widetilde{w}_t^{j,\mathsf{A}} = w_{t-1}^{j,\mathsf{A}} - \eta_d \nabla L^{\mathsf{T}}(w_{t-1}^{j,\mathsf{A}}; x_t, y_t). \tag{66}$$

The normalization to the sphere is then performed componentwise as in (5). Note that (66) is an exact identity: conditionally on its initialization, every augmented weight follows precisely the tied SGD trajectory with reference step size $\eta_d$, driven by the same stream of fresh samples.

*Proof of part* (A) *of Theorem B.1.* Fix any weight index $1 \le j \le s$. By (66), the trajectory of $w_t^{j,\mathsf{A}}$ has the same law as the tied trajectory analyzed in part (T), with an independent uniform initialization. Therefore, with probability $1 - o(1)$, there exists

$$t \le \frac{C}{\eta_d \gamma_d^{p-2}}$$

such that $M_\sigma(w_t^{j,\mathsf{A}}) \ge \rho$. This proves the required bound on $\tau_\rho^{j,\mathsf{A}}$. □

It remains to analyze the unaugmented untied student. Again it is enough to follow one weight, which we denote by $u_t \coloneqq w_t^{1,\mathsf{U}}$.

### F.2 One-weight untied dynamics

The population loss of the single-weight model is

$$\Psi(u) = -\mathbb{E}\big[y\,\sigma(\langle u, x\rangle)\big] = -\frac{1}{\sqrt{r}} \sum_{k=0}^{r-1} \phi(m_k(u)). \tag{67}$$

Thus the population landscape is the same scalar overlap landscape as for the tied model, but its magnitude is smaller by a factor $r^{-1/2}$.

Let

$$\mathcal{F}_t^{\mathrm{S}} \coloneqq \sigma\big(w^*, u_0, (x_\ell, y_\ell)_{1 \le \ell \le t}\big)$$

be the natural filtration of the one-weight process. Set

$$G_t^{\mathrm{S}} \coloneqq \nabla L^{\mathrm{S}}(u_{t-1}; x_t, y_t), \qquad H_t^{\mathrm{S}}(u) \coloneqq L^{\mathrm{S}}(u; x_t, y_t) - \Psi(u),$$

and define

$$\xi_{k,t}^{\mathrm{S}} \coloneqq -\langle \Pi^k w^*, \nabla H_t^{\mathrm{S}}(u_{t-1})\rangle.$$

Exactly as in Lemma D.1, differentiating (67) gives

$$-\langle \Pi^k w^*, \nabla \Psi(u)\rangle = \frac{1}{\sqrt{r}}\big(\phi'(m_k(u)) + \Gamma_k(u)\big), \tag{68}$$

where $\Gamma_k$ is the same function as in (27). Using the effective update (65), the exact overlap recursion is therefore

$$\sqrt{1 + \frac{\eta_d^2}{r}\|G_t^{\mathrm{S}}\|^2}\, m_k(u_t) = m_k(u_{t-1}) + \frac{\eta_d}{r}\big(\phi'(m_k(u_{t-1})) + \Gamma_k(u_{t-1})\big) + \frac{\eta_d}{\sqrt{r}}\,\xi_{k,t}^{\mathrm{S}}. \tag{69}$$

Consequently,

$$m_k(u_t) = m_k(u_0) + \frac{p a_p^2 \eta_d}{r} \sum_{\ell=1}^{t} m_k(u_{\ell-1})^{p-1} + E_{k,t}^{\mathsf{U}}, \tag{70}$$

where

$$\begin{aligned} E_{k,t}^{\mathsf{U}} \coloneqq{}& \frac{\eta_d}{r} \sum_{\ell=1}^{t} R_\sigma(m_k(u_{\ell-1})) + \frac{\eta_d}{r} \sum_{\ell=1}^{t} \Gamma_k(u_{\ell-1}) \\ &+ \frac{\eta_d}{\sqrt{r}} \sum_{\ell=1}^{t} \xi_{k,\ell}^{\mathrm{S}} - \sum_{\ell=1}^{t} \left( \sqrt{1 + \frac{\eta_d^2}{r}\|G_\ell^{\mathrm{S}}\|^2} - 1 \right) m_k(u_\ell). \end{aligned} \tag{71}$$

The stochastic estimates needed for (71) are simpler than in the tied model.

**Lemma F.1** (Single-weight stochastic bounds)**.** *Condition on a teacher satisfying the geometry event of Lemma C.1. There exists $C_\sigma < \infty$ such that, uniformly over $u \in \mathbb{S}^{d-1}$, $k < r$, and $t \ge 1$,*

$$\mathbb{E}[(\xi_{k,t}^{\mathrm{S}})^2 \mid \mathcal{F}_{t-1}^{\mathrm{S}}] \le C_\sigma, \tag{72}$$

$$\big(\mathbb{E}[|\xi_{k,t}^{\mathrm{S}}|^a \mid \mathcal{F}_{t-1}^{\mathrm{S}}]\big)^{1/a} \le C_\sigma a^P, \qquad a \ge 2. \tag{73}$$

*Moreover, for every fixed $K < \infty$, there exists $C_{\sigma,K} < \infty$ such that, for every deterministic horizon $T \le d^K$,*

$$\mathbb{P}\left( \max_{1 \le t \le T} \|G_t^{\mathrm{S}}\| > C_{\sigma,K} \sqrt{d}\,(\log d)^P \right) = o(1). \tag{74}$$

*Proof.* Fix $u \in \mathbb{S}^{d-1}$ and set $v = (\mathrm{I}_d - uu^\top)\Pi^k w^*$, so $\|v\| \le 1$. Then

$$\langle \Pi^k w^*, G^{\mathrm{S}} \rangle = -y\, \sigma'(\langle u, x\rangle)\langle v, x\rangle.$$

On the teacher-geometry event, $\|y\|_{L^a} \le C_\sigma a^{P/2}$ for every $a \ge 2$, by the same $L^2$ bound and Gaussian hypercontractivity used in Lemma D.4. The factor $\sigma'(\langle u, x\rangle)\langle v, x\rangle$ is a Gaussian polynomial of degree at most $P$ with uniformly bounded $L^2$-norm; hence hypercontractivity and Cauchy–Schwarz give

$$\|\langle \Pi^k w^*, G^{\mathrm{S}} \rangle\|_{L^a} \le C_\sigma a^P.$$

Centering can only decrease the variance and changes the $L^a$-norm by at most a factor two, proving (72)–(73).

For the gradient norm,

$$G^{\mathrm{S}} = -y\, \sigma'(\langle u, x\rangle)(\mathrm{I}_d - uu^\top)x.$$

Choosing an orthonormal basis of $u^\perp$, applying the preceding polynomial moment estimate coordinatewise, and using Minkowski's inequality yields

$$\left(\mathbb{E}[\|G^{\mathrm{S}}\|^a]\right)^{1/a} \le C_\sigma \sqrt{d}\, a^P, \qquad a \ge 2,$$

uniformly in $u$. Taking $a \asymp \log d$, applying Markov's inequality conditionally on the past, and then taking a union bound over $t \le T \le d^K$ proves (74). □

### F.3 Proof of the untied weak recovery bound

We now compare (70) with the tied decomposition (29). The deterministic drift has been reduced by a factor $r$, while the centered stochastic increment has been reduced by a factor $\sqrt{r}$. These two changes exactly balance when the time horizon is multiplied by $r$.

Fix the same scales $\alpha_d, \beta_d$ as in the tied proof, set $\nu_d = (\log d)^{-2}$, and define

$$T_0^{\mathsf{U}} := \left\lceil \frac{C_0 r}{\eta_d \gamma_d^{p-2}} \right\rceil, \qquad \widehat{\tau}^{\mathsf{U}}_{\beta_d} := \inf\left\{ t \ge 0 : \max_{k<r} |m_k(u_t)| \ge \beta_d \right\}.$$

The initialization Lemma C.3 applies unchanged to $u_0$.

We first record the analogue of Lemma E.3.

**Lemma F.2** (Untied microscopic error)**.** *For $A > A_*(p, P)$ sufficiently large,*

$$\max_{k<r} \max_{t \le \widehat{\tau}^{\mathsf{U}}_{\beta_d} \wedge T_0^{\mathsf{U}}} |E^{\mathsf{U}}_{k,t}| = o(\nu_d \gamma_d)$$

*with probability* $1 - o(1)$.

*Proof.* The higher-Hermite and interaction terms are identical in size to their tied counterparts because

$$\frac{\eta_d}{r} T_0^{\mathsf{U}} \asymp \gamma_d^{-(p-2)}.$$

Thus the calculations in the proof of Lemma E.3 apply verbatim.

For the martingale term, Lemma F.1 and the same truncation–Freedman argument as in Lemma E.5 give

$$\max_{k<r} \max_{t \le \widehat{\tau}^{\mathsf{U}}_{\beta_d} \wedge T_0^{\mathsf{U}}} \left| \frac{\eta_d}{\sqrt{r}} \sum_{\ell=1}^{t} \xi^{\mathrm{S}}_{k,\ell} \right| \le C \frac{\eta_d}{\sqrt{r}} \left( \sqrt{T_0^{\mathsf{U}} \log d} + (\log d)^{P+1} \right)$$

with probability $1 - o(1)$. Since

$$\frac{\eta_d}{\sqrt{r}} \sqrt{T_0^{\mathsf{U}} \log d} \asymp \sqrt{\frac{\eta_d \log d}{\gamma_d^{p-2}}},$$

this is exactly the leading martingale scale in the tied proof, while the truncation term is smaller by $r^{-1/2}$. Hence the whole term is $o(\nu_d \gamma_d)$.

Finally, on the gradient event (74),

$$\sqrt{1+\frac{\eta_d^2}{r}\|G_\ell^{\mathrm{S}}\|^2}-1\le\frac{\eta_d^2}{2r}\|G_\ell^{\mathrm{S}}\|^2.$$

Therefore the cumulative normalization error is bounded by

$$C\frac{\eta_d^2}{r}T_0^{\mathsf{U}}d(\log d)^{2P}\beta_d,$$

which coincides with the tied bound after using the definition of $T_0^{\mathsf{U}}$. This is again $o(\nu_d\gamma_d)$. Combining the four estimates proves the claim. □

If $p$ is odd, the analogue of Corollary E.4 also holds: before $T_0^{\mathsf{U}}$, no coordinate can be the first to leave the microscopic region through $-\beta_d$. Indeed, the leading drift in (70) is nonnegative and Lemma F.2 gives a cumulative error $o(\gamma_d)$, whereas $\gamma_d=o(\beta_d)$.

On this event, the scalar comparison proof of Proposition E.1 applies without any structural change, with $\eta_d$ in the leading scalar recursion replaced by $\eta_d/r$. Consequently, for a unique favorable index $i_*$,

$$\tau_{\beta_d}^{(1),\mathsf{U}}\le\frac{C_\beta r}{\eta_d\gamma_d^{p-2}},\qquad\beta_d\le\widetilde{m}_{i_*}(u_{\tau_{\beta_d}^{(1),\mathsf{U}}})\le2\beta_d,\tag{75}$$

and

$$\max_{k\neq i_*}|m_k(u_{\tau_{\beta_d}^{(1),\mathsf{U}}})|\le C_\alpha\alpha_d,$$

with probability $1-o(1)$. Here $\tau_a^{(1),\mathsf{U}}:=\inf\{t:M_\sigma(u_t)\ge a\}$.

The continuation argument rescales in the same way. Starting from the configuration in (75), define

$$T^{\mathsf{U}}:=\left\lceil\frac{C_T r}{\eta_d\beta_d^{p-2}}\right\rceil.$$

The selected population drift in (69) is $r^{-1}$ times the corresponding drift per tied SGD step, the martingale increment is multiplied by $r^{-1/2}$, and the normalization coefficient by $r^{-1}$. The cumulative scales over $T^{\mathsf{U}}$ are therefore the same as, or smaller than, those in the tied continuation proof. Indeed,

$$\frac{\eta_d}{r}T^{\mathsf{U}}\asymp\beta_d^{-(p-2)},\qquad\frac{\eta_d}{\sqrt{r}}\sqrt{T^{\mathsf{U}}\log d}\asymp\sqrt{\frac{\eta_d\log d}{\beta_d^{p-2}}},$$

and

$$\frac{\eta_d^2}{r}T^{\mathsf{U}}\asymp\frac{\eta_d}{\beta_d^{p-2}}.$$

Thus the deterministic forcing, martingale fluctuations, and normalization errors satisfy exactly the bounds used in Proposition E.2; the selected-coordinate noise is in fact simpler because (72) is uniform in the overlap. The damping argument for the competitors is unchanged after replacing each occurrence of $\eta_d$ in the population drift by $\eta_d/r$. Consequently, the selected overlap reaches $\rho$ within an additional

$$\frac{C_\rho r}{\eta_d\beta_d^{p-2}}=o\left(\frac{r}{\eta_d\gamma_d^{p-2}}\right)$$

steps, while all competitors remain microscopic. Therefore

$$\tau_\rho^{(1),\mathsf{U}}\le\frac{Cr}{\eta_d\gamma_d^{p-2}}\tag{76}$$

with probability $1-o(1)$.

*Proof of part* (U) *of Theorem B.1.* For any fixed $1\le j\le s$, the $j$th untied weight has the same trajectory law as the first weight analyzed above. Thus (76) gives

$$\tau_\rho^{j,\mathsf{U}}\le\frac{Cr}{\eta_d\gamma_d^{p-2}}$$

with probability $1 - o(1)$. Since $\gamma_d = \sqrt{\delta \log d/d}$, this is

$$O\left(rd^{p-1}(\log d)^{A-(p-2)/2}\right),$$

which proves the theorem. □

### F.4 Proof of weak orbit coverage

We now turn to the proof of Corollary 3.5, which describes how full weak orbit coverage is achieved by the augmented and untied students given enough weights.

Retain the assumptions and learning rates of Theorem B.1, and fix $\rho \in (0, 1)$. The preceding untied argument already shows that any fixed weight recovers its initial leader, by (75) and the continuation leading to (76). We now quantify its stochastic failure probability to apply this conclusion simultaneously to all weights satisfying the initialization conditions.

Recall the teacher-geometry event $\mathcal{G}_d$ from Lemma C.1. Moreover, recall the initialization conditions in Lemmas C.2 and C.3: for a weight $u$ with leader $i_* = \arg\max_{k<r} q_k(u)$,

$$\max_{k<r} |m_k(u)| \le C\gamma_d, \qquad q_{i_*}(u) \ge c\gamma_d, \qquad \max_{k\neq i_*} q_k(u) \le (1-\nu_d) q_{i_*}(u), \tag{77}$$

where $\nu_d = (\log d)^{-2}$. Let $\mathcal{E}_d^j$ denote the event that (77) holds for $u = w_0^{j,\mathsf{U}}$, and write $i_*(j)$ for its initial leader. We break ties by the smallest index; on $\mathcal{E}_d^j$ the leader is unique. For a constant $C_U > 0$, set

$$T_{\mathsf{U}} := \left\lceil \frac{C_U r}{\eta_d \gamma_d^{p-2}} \right\rceil.$$

**Lemma F.3** (Simultaneous untied recovery of initial leaders)**.** *There exists $C_U = C_U(\sigma, \delta, \rho) < \infty$ such that, for every fixed $D > 0$ and all sufficiently large $d$, uniformly over $w^* \in \mathcal{G}_d$,*

$$\mathbb{P}\left( \bigcup_{j=1}^{s} \left( \mathcal{E}_d^j \cap \left\{ \max_{0\le t\le T_{\mathsf{U}}} q_{i_*(j)}(w_t^{j,\mathsf{U}}) < \rho \right\} \right) \,\middle|\, w^* \right) \le s d^{-D}.$$

*In particular, taking $D > \kappa$ in Assumption 3, with conditional probability $1 - o(1)$ every weight satisfying* (77) *recovers its initial leader by $T_{\mathsf{U}}$.*

*Proof.* Fix a teacher in $\mathcal{G}_d$ and a weight index $j$, and condition further on an initialization $u_0 = w_0^{j,\mathsf{U}}$ satisfying (77). By (65), its trajectory $u_t = w_t^{j,\mathsf{U}}$ follows the one-weight dynamics. The samples remain fresh under this conditioning by Assumption 2. The conditional moment estimates of Lemma F.1 and its proof, together with the truncation argument in the proof of Lemma E.5, give an event of conditional probability at least $1 - d^{-D}$ on which

$$\frac{\eta_d}{\sqrt{r}} \max_{\ell<r} \max_{0\le n\le T_{\mathsf{U}}} \left| \sum_{t=1}^{n} \xi_{\ell,t}^{\mathrm{S}} \right| \le C_{\sigma,D} \frac{\eta_d}{\sqrt{r}} \left( \sqrt{T_{\mathsf{U}} \log d} + (\log d)^{P+1} \right) = o(\nu_d \gamma_d), \tag{78}$$

and

$$\max_{1\le t\le T_{\mathsf{U}}} \|G_t^{\mathrm{S}}\| \le C_{\sigma,D} \sqrt{d}\, (\log d)^P.$$

Indeed, choose the moment orders proportional to $\log d$ so that the truncation, concentration, and gradient estimates have combined failure probability at most

$$C_{\sigma,b} r (T_{\mathsf{U}} + 1) d^{-b} \le d^{-D},$$

where $b > D + p + 1$ is fixed. The last inequality holds for sufficiently large $d$, since $T_{\mathsf{U}} \le d^p$ and $r \le d$. The estimates hold over the full horizon because the untied conditional moments are uniform over the current weight. The little-$o$ bound in (78) follows from the same calculation as in the proof of Lemma F.2.

Any noise sum over an interval inside $[0, T_{\mathsf{U}}]$ is the difference of two partial sums, so its magnitude is at most twice the maximum in (78). Thus the same event also controls the noise after the microscopic

stopping time. The deterministic error bounds from Lemma F.2 and the comparisons leading to (75) and (76) now apply with $i_* = i_*(j)$. They bring this initial leader to $\rho$ within

$$\frac{C_\beta r}{\eta_d \gamma_d^{p-2}} + \frac{C_\rho r}{\eta_d \beta_d^{p-2}} \le T_{\mathsf{U}},$$

for sufficiently large $C_U$ and $d$, since $\beta_d/\gamma_d \to \infty$. Here $\beta_d$ is the microscopic threshold from Appendix E, also used in Lemma F.2. Increasing $D$ changes the concentration constants and the dimension threshold, but neither the little-$o$ comparisons nor the choice of $C_U$.

The estimates are uniform over the conditioned initialization. Integrating over $u_0$ therefore gives, for every $j$,

$$\mathbb{P}\left(\mathcal{E}_d^j \cap \left\{\max_{0\le t\le T_{\mathsf{U}}} q_{i_*(j)}(w_t^{j,\mathsf{U}}) < \rho\right\} \,\middle|\, w^*\right) \le d^{-D}.$$

A union bound over $j \le s$ proves the lemma. This step requires no independence between the trajectories, which share the sample sequence. □

Fix $\varepsilon > 0$ and suppose $s \ge (1+\varepsilon) r \log r$. Recall the relevant correlations $q_k$ from (7) and the coverage events $\mathcal{C}_n^{\mathsf{k}}(\rho)$ from Definition 3.4. Use the horizon $T_{\mathsf{U}}$ above, with $C_U$ chosen according to Lemma F.3. Choose $C_A$ as in part (A) of Theorem B.1, and set

$$T_{\mathsf{A}} \coloneqq \left\lceil \frac{C_A}{\eta_d \gamma_d^{p-2}} \right\rceil.$$

*Proof of Corollary 3.5.* We first prove the coverage upper bounds. According to (2) and (7), we have the shift identity

$$q_k(\Pi^a u) = q_{k-a}(u), \qquad 0 \le a, k < r, \tag{79}$$

where indices are understood modulo $r$.

**Untied procedure.** Condition on a teacher $w^* \in \mathcal{G}_d$. Using the initialization events $\mathcal{E}_d^j$, write

$$p_{\text{init}} \coloneqq \mathbb{P}(\mathcal{E}_d^1 \mid w^*) = 1 - o(1).$$

The last equality follows from the joint proof of Lemmas C.2 and C.3, uniformly over $w^* \in \mathcal{G}_d$.

By (79), shifting an initial weight by $\Pi^a$ preserves (77) and shifts its unique leader by $a$. Since uniform initialization on the sphere is invariant under these shifts, for every $j \le s$ and $k < r$,

$$\mathbb{P}\big(\mathcal{E}_d^j \cap \{i_*(j) = k\} \mid w^*\big) = \frac{p_{\text{init}}}{r}.$$

The initializations are independent by Assumption 2. Consequently, for a fixed teacher direction $k$, the probability that no initial weight satisfies (77) and has leader $k$ is $(1 - p_{\text{init}}/r)^s$. Taking a union bound over $k < r$, the probability that at least one teacher direction has no such initial weight is at most

$$r\left(1 - \frac{p_{\text{init}}}{r}\right)^s \le r \exp\left(-\frac{s p_{\text{init}}}{r}\right) \le r^{-\varepsilon + o(1)} = o(1).$$

If every teacher direction has such an initial weight and all these weights recover their initial leaders, full orbit coverage holds. Choose $D > \kappa$, so $s d^{-D} = o(1)$ by Assumption 3. Combining the preceding bound on missing initial leaders with the recovery failure bound in Lemma F.3 gives

$$\mathbb{P}\left((\mathcal{C}_{T_{\mathsf{U}}}^{\mathsf{U}}(\rho))^c\right) \le \mathbb{P}(\mathcal{G}_d^c) + r^{-\varepsilon + o(1)} + s d^{-D} = o(1),$$

The three terms account, respectively, for failure of the teacher geometry event, a teacher direction with no suitable initial weight, and a recovery failure among weights satisfying (77). The first term tends to zero by Lemma C.1.

**Augmented procedure.** Here we condition on the teacher and the samples through the recovery horizon. Define

$$\mathcal{D}_{\mathsf{A}} \coloneqq \sigma\big(w^*, (x_t, y_t)_{1 \le t \le T_{\mathsf{A}}}\big).$$

Under this conditioning, the initial weights remain independent and uniform by Assumption 2. The reduction (66) therefore makes their trajectories conditionally independent and identically distributed.

For every fixed sample, the tied loss from Definition 2.2 satisfies

$$L^{\mathsf{T}}(\Pi^a u; x, y) = L^{\mathsf{T}}(u; x, y), \qquad \nabla L^{\mathsf{T}}(\Pi^a u; x, y) = \Pi^a \nabla L^{\mathsf{T}}(u; x, y).$$

The normalization in (5) also commutes with $\Pi^a$. Hence, by (66), shifting an initial weight shifts its entire augmented trajectory by the same amount. Together with (79), this implies that the conditional probability of recovering direction $k$ by time $T_{\mathsf{A}}$ is the same for every $k$.

Using the orbit recovery time defined in (10), let

$$p_{\mathrm{rec}} \coloneqq \mathbb{P}\left(\tau_\rho^{1,\mathsf{A}} \le T_{\mathsf{A}} \,\middle|\, \mathcal{D}_{\mathsf{A}}\right).$$

By (3) and (7), $M_\sigma(u) = \max_{k<r} q_k(u)$. The event defining $p_{\mathrm{rec}}$ is therefore the union of the $r$ direction-recovery events. Since these events have equal conditional probabilities, a union bound gives

$$\mathbb{P}\left(\max_{0\le t\le T_{\mathsf{A}}} q_k(w_t^{1,\mathsf{A}}) \ge \rho \,\middle|\, \mathcal{D}_{\mathsf{A}}\right) \ge \frac{p_{\mathrm{rec}}}{r}, \qquad 0 \le k < r.$$

This bound allows a single weight to recover more than one direction.

For any fixed direction, conditional independence across weights bounds the probability that all $s$ trajectories miss it by $(1 - p_{\mathrm{rec}}/r)^s$. A union bound over directions consequently gives

$$\mathbb{P}\left(\left(\mathcal{C}_{T_{\mathsf{A}}}^{\mathsf{A}}(\rho)\right)^c \,\middle|\, \mathcal{D}_{\mathsf{A}}\right) \le r\left(1 - \frac{p_{\mathrm{rec}}}{r}\right)^s \le r\exp\left(-\frac{s p_{\mathrm{rec}}}{r}\right).$$

We now remove the conditioning. Part (A) of Theorem B.1 and the tower property imply

$$\mathbb{E}[1 - p_{\mathrm{rec}}] = \mathbb{P}\left(\tau_\rho^{1,\mathsf{A}} > T_{\mathsf{A}}\right) = o(1).$$

Fix any constant $\theta \in ((1+\varepsilon)^{-1}, 1)$. On $\{p_{\mathrm{rec}} \ge \theta\}$, the conditional coverage failure probability is at most $r^{1-(1+\varepsilon)\theta}$. On the complementary event, Markov's inequality gives

$$\mathbb{P}(p_{\mathrm{rec}} < \theta) \le \frac{\mathbb{E}[1 - p_{\mathrm{rec}}]}{1 - \theta}.$$

Averaging the conditional coverage estimate thus yields

$$\mathbb{P}\left(\left(\mathcal{C}_{T_{\mathsf{A}}}^{\mathsf{A}}(\rho)\right)^c\right) \le \frac{\mathbb{E}[1 - p_{\mathrm{rec}}]}{1 - \theta} + r^{1-(1+\varepsilon)\theta} = o(1).$$

**Sample complexity and matching lower bounds.** The preceding estimates prove coverage by $T_{\mathsf{U}}$ and $T_{\mathsf{A}}$ for the respective procedures. Different directions may be recovered at different times, as allowed by Definition 3.4. Each iteration of (5) uses one fresh observation. Using $\eta_d = d^{-p/2}(\log d)^{-A}$ from Theorem B.1 and $\gamma_d = \sqrt{\log r/d}$ from Appendix C, the horizons satisfy

$$T_{\mathsf{A}} = \widetilde{O}(d^{p-1}), \qquad T_{\mathsf{U}} = \widetilde{O}(r d^{p-1}).$$

For the lower bounds, note that by Definitions 3.1 and 3.4, we have

$$\mathcal{C}_n^{\mathsf{k}}(\rho) \subseteq \mathcal{R}_n^{\mathsf{k}}(\rho), \qquad \mathsf{k} \in \{\mathsf{U}, \mathsf{A}\}.$$

Theorem B.4 therefore supplies the matching lower bounds, giving

$$n_{\mathsf{A}}^{\mathrm{cov}}(\rho) = \widetilde{\Theta}(d^{p-1}), \qquad n_{\mathsf{U}}^{\mathrm{cov}}(\rho) = \widetilde{\Theta}(r d^{p-1}),$$

which proves the corollary. □

## G Quadratic link and weak subspace recovery

In this appendix, we prove Theorem B.3 with $\sigma = h_2$. We retain the assumptions, learning rates, and recovery times introduced in Appendix B. We first analyze the tied trajectory, using the notation of Appendix D, and then treat the augmented and untied procedures using the reductions from Appendix F.

### G.1 Proof strategy

For $p \geq 3$, the microscopic argument in Appendix E amplifies the initial gap between the leading correlation and its competitors. For the pure quadratic link, the leading coordinate drift is linear, and we instead study the growth of the total mass in the teacher orbit subspace.

Recall that $w^{\|}$ denotes the orthogonal projection of $w$ onto $\mathcal{S}^*$. We track its squared norm through the total overlap

$$R(w) \coloneqq \sum_{k=0}^{r-1} m_k(w)^2, \qquad R_t \coloneqq R(w_t).$$

Subsection G.2 shows that, on the teacher-geometry event $\mathcal{G}_d$ defined in Lemma C.1,

$$R(w) = (1 + o(1))\|w^{\|}\|^2$$

uniformly over unit weights, and $R_0 \asymp r/d$ with probability $1 - o(1)$. Thus weak subspace recovery follows once $R_t$ reaches a macroscopic level.

The identity $\Phi = -R$ gives a direct description of the population growth: the population contribution to the one-step increase of $R_t$ is $\eta_d\|\nabla\Phi(w_{t-1})\|^2$. Below any fixed threshold $a < 1$, this quantity is comparable to $\eta_d R_{t-1}$ on $\mathcal{G}_d$. Subsection G.3 establishes this comparison and the exact stochastic recursion.

To turn this proportional drift into a growth estimate, we study $\log R_t$. Its increments have a positive drift of order $\eta_d$, together with relative noise and discretization errors. The relative noise involves division by $R_{t-1}$, so we first stop the process when $R_t$ either reaches $a$ or falls below a fixed fraction of $R_0$. Before this exit, the denominator remains bounded below at the initialization scale. Using the moment estimates from Subsection G.4, Subsection G.5 shows that the accumulated relative noise, logarithmic remainders, and spherical normalization errors are $o(1)$ over a horizon of order $\eta_d^{-1}\log(d/r)$. The resulting lower bound on $\log(R_t/R_0)$ rules out the lower exit and forces $R_t$ to reach $a$ within this horizon. Choosing $a \in (\rho, 1)$ then yields weak subspace recovery.

Finally, Subsection G.6 treats the other procedures. Under the prescribed learning rates, each augmented weight follows the tied update exactly. For an untied weight, the drift carries a factor $\eta_d/r$ and the centered stochastic increment a factor $\eta_d/\sqrt{r}$. The same comparison therefore applies over a horizon longer by a factor $r$, with the accumulated errors still negligible.

### G.2 Total overlap and initialization

We work on the teacher-geometry event $\mathcal{G}_d$. Recall the teacher Gram matrix $W^*$ from the proof of Lemma C.3. On this event,

$$\|W^* - \mathrm{I}_r\|_{\mathrm{op}} \leq r\mu_* = o(1).$$

To relate total overlap to projection mass, define

$$B^* \coloneqq \sum_{k=0}^{r-1} (\Pi^k w^*)(\Pi^k w^*)^\top.$$

This matrix vanishes on $(\mathcal{S}^*)^\perp$, and

$$R(w) = \langle w, B^* w\rangle = \langle w^{\|}, B^* w^{\|}\rangle.$$

For sufficiently large $d$, $W^*$ is positive definite. The restriction of $B^*$ to $\mathcal{S}^*$ has the same eigenvalues as $W^*$, all lying in $[1 - r\mu_*, 1 + r\mu_*]$. Consequently, uniformly over $w \in \mathbb{S}^{d-1}$,

$$(1 - r\mu_*)\|w^{\|}\|^2 \leq R(w) \leq (1 + r\mu_*)\|w^{\|}\|^2. \tag{80}$$

Thus it suffices to control the total overlap. We next establish its initial scale.

**Lemma G.1** (Initial total overlap)**.** *Condition on a teacher $w^* \in \mathcal{G}_d$, and let $w_0$ be uniform on $\mathbb{S}^{d-1}$, independently of the teacher. There exist constants $c, C > 0$ such that, for all sufficiently large $d$ and every $\varepsilon \in (0, 1)$,*

$$\mathbb{P}\left(\left|R_0 - \frac{r}{d}\right| > \varepsilon\frac{r}{d}\,\middle|\, w^*\right) \leq C\exp(-c\varepsilon^2 r).$$

*The constants are uniform over $w^* \in \mathcal{G}_d$. In particular, $R_0 \asymp r/d$ with conditional probability $1 - o(1)$.*

*Proof.* Write $w_0 = h/\|h\|$, where $h \sim \mathcal{N}(0, \mathrm{I}_d)$ is independent of the teacher. Then

$$R_0 = \frac{h^\top B^* h}{\|h\|^2}.$$

The matrix $B^*$ is positive semidefinite and satisfies

$$\mathrm{Tr}(B^*) = r, \qquad \|B^*\|_{\mathrm{op}} = \|W^*\|_{\mathrm{op}} \le 1 + r\mu_* \le C,$$

and hence

$$\|B^*\|_{\mathrm{F}}^2 = \mathrm{Tr}((B^*)^2) \le \|B^*\|_{\mathrm{op}} \mathrm{Tr}(B^*) \le Cr.$$

For $\theta \in (0,1)$, apply the Hanson–Wright inequality (13) to $B^*$ and $\mathrm{I}_d$. A union bound gives

$$\begin{aligned}\mathbb{P}\Big(|h^\top B^* h - r| > \theta r \text{ or } \big|\|h\|^2 - d\big| > \theta d \;\Big|\; w^*\Big) \\ \le Ce^{-c\theta^2 r} + Ce^{-c\theta^2 d} \le Ce^{-c\theta^2 r},\end{aligned}$$

where we used $r \le d$. Outside this event,

$$\frac{1-\theta}{1+\theta}\frac{r}{d} \le R_0 \le \frac{1+\theta}{1-\theta}\frac{r}{d}.$$

Taking $\theta$ to be a sufficiently small universal multiple of $\varepsilon$ proves the concentration bound. The final assertion follows by fixing $\varepsilon \in (0,1)$ and using $r = d^\delta \to \infty$. □

Combining Lemma G.1, (80), and $\mathbb{P}(\mathcal{G}_d) = 1 - o(1)$ yields

$$\|w_0^{\|}\|^2 \asymp R_0 \asymp \frac{r}{d}$$

with probability $1 - o(1)$. The same initialization estimate applies to any fixed weight of the untied or augmented student.

### G.3 Recursion and population growth

Recall that $R(w) = \langle w, B^* w\rangle$ and $\Phi = -R$. The spherical gradient calculation in Lemma D.1, specialized to $\phi(z) = z^2$, gives

$$\nabla\Phi(w) = -2\big(B^* w - R(w)w\big).$$

This identity allows us to express the population contribution to the growth of $R$ as the squared norm of the spherical gradient.

Using the coordinate increments from (28), define

$$\Xi_t \coloneqq \sum_{k=0}^{r-1} m_k(w_{t-1})\xi_{k,t} = -\langle B^* w_{t-1}, \nabla H_t(w_{t-1})\rangle.$$

The coefficients $m_k(w_{t-1})$ are $\mathcal{F}_{t-1}$-measurable, so

$$\mathbb{E}[\Xi_t \mid \mathcal{F}_{t-1}] = 0.$$

**Lemma G.2** (Exact recursion for the total overlap)**.** *For every $t \ge 1$,*

$$\begin{aligned}(1 + \eta_d^2\|G_t\|^2)R_t = R_{t-1} + \eta_d\|\nabla\Phi(w_{t-1})\|^2 \\ + 2\eta_d\Xi_t + \eta_d^2\langle G_t, B^* G_t\rangle.\end{aligned} \tag{81}$$

*Proof.* Since $G_t \perp w_{t-1}$, the spherical update (5) yields

$$(1 + \eta_d^2\|G_t\|^2)R_t = R_{t-1} - 2\eta_d\langle B^* w_{t-1}, G_t\rangle + \eta_d^2\langle G_t, B^* G_t\rangle.$$

Using $G_t = \nabla\Phi(w_{t-1}) + \nabla H_t(w_{t-1})$ and the spherical gradient formula above,

$$\begin{aligned}-2\langle B^* w_{t-1}, G_t\rangle &= \langle \nabla\Phi(w_{t-1}), G_t\rangle \\ &= \|\nabla\Phi(w_{t-1})\|^2 + 2\Xi_t.\end{aligned}$$

Substitution proves the identity. □

The population contribution is therefore nonnegative. Teacher geometry gives the quantitative lower bound needed to amplify a small total overlap.

**Proposition G.3** (Population growth of the total overlap)**.** *On the teacher-geometry event $\mathcal{G}_d$, for every fixed $a \in (0,1)$ there exist constants $c_a, C_a > 0$, depending only on $a$, such that, for all sufficiently large $d$, uniformly over $w \in \mathbb{S}^{d-1}$ with $0 < R(w) \le a$,*

$$c_a R(w) \le \|\nabla\Phi(w)\|^2 \le C_a R(w).$$

*Proof.* Since $\langle w, B^* w\rangle = R(w)$ and $\|w\| = 1$,

$$\|\nabla\Phi(w)\|^2 = 4\big(\|B^* w\|^2 - R(w)^2\big). \tag{82}$$

By the spectral comparison established in Subsection G.2, the nonzero eigenvalues of $B^*$ belong to $[1 - r\mu_*, 1 + r\mu_*]$. Hence

$$(1 - r\mu_*)R(w) \le \|B^* w\|^2 \le (1 + r\mu_*)R(w).$$

Combining these bounds with (82) gives

$$4R(w)(1 - R(w) - r\mu_*) \le \|\nabla\Phi(w)\|^2 \le 4R(w)(1 - R(w) + r\mu_*).$$

On $\mathcal{G}_d$, we have $r\mu_* = o(1)$ uniformly over the teacher. Since $R(w) \le a < 1$, the claimed bounds follow for sufficiently large $d$. □

The matrix $B^*$ is positive semidefinite, so the last term in (81) is nonnegative. Discarding it and applying Proposition G.3, we obtain, on $\mathcal{G}_d$ and for sufficiently large $d$,

$$(1 + \eta_d^2 \|G_t\|^2) R_t \ge (1 + c_a \eta_d) R_{t-1} + 2\eta_d \Xi_t \tag{83}$$

whenever $0 < R_{t-1} < a$. The next subsection controls the centered increment $\Xi_t$ and the normalization term $\eta_d^2 \|G_t\|^2$.

### G.4 Noise and gradient moments

We work on the teacher-geometry event $\mathcal{G}_d$. The coordinate and gradient estimates from Appendices D and F, specialized to $p = P = 2$, give the stochastic bounds needed for (83).

For the untied procedure, recall the single-weight trajectory $u_t = w_t^{1,\mathsf{U}}$ and the quantities $G_t^{\mathrm{S}}$, $\xi_{k,t}^{\mathrm{S}}$, and $\mathcal{F}_t^{\mathrm{S}}$ from Appendix F. Define

$$\Xi_t^{\mathrm{S}} := \sum_{k=0}^{r-1} m_k(u_{t-1}) \xi_{k,t}^{\mathrm{S}}.$$

As for the tied process, $\mathbb{E}[\Xi_t^{\mathrm{S}} \mid \mathcal{F}_{t-1}^{\mathrm{S}}] = 0$.

**Corollary G.4** (Noise and gradient moments)**.** *There exists a constant $C > 0$ such that, for all sufficiently large $d$, every $t \ge 1$, and every $\nu \ge 2$,*

$$\mathbb{E}[\Xi_t^2 \mid \mathcal{F}_{t-1}] \le C r R_{t-1}, \tag{84}$$

$$(\mathbb{E}[|\Xi_t|^\nu \mid \mathcal{F}_{t-1}])^{1/\nu} \le C\nu^2 \sqrt{r R_{t-1}}, \tag{85}$$

$$(\mathbb{E}[\|G_t\|^\nu \mid \mathcal{F}_{t-1}])^{1/\nu} \le C\nu^2 \sqrt{d}. \tag{86}$$

*The estimates are uniform over the current weight and over teachers in $\mathcal{G}_d$. They also hold for the untied single-weight process, with*

$$(G_t, \Xi_t, R_{t-1}, \mathcal{F}_{t-1}) \quad \textit{replaced by} \quad (G_t^{\mathrm{S}}, \Xi_t^{\mathrm{S}}, R(u_{t-1}), \mathcal{F}_{t-1}^{\mathrm{S}}).$$

*Proof.* By (80),

$$R(w) \le 1 + r\mu_* \le C$$

uniformly over unit weights on $\mathcal{G}_d$, for sufficiently large $d$.

Condition on $\mathcal{F}_{t-1}$. Conditional Minkowski's inequality and Lemma D.4, with $P = 2$, give

$$\begin{aligned}(\mathbb{E}[|\Xi_t|^\nu \mid \mathcal{F}_{t-1}])^{1/\nu} &\le C\nu^2 \sum_{k=0}^{r-1} |m_k(w_{t-1})|\big(1 + \sqrt{r}\,|m_k(w_{t-1})|\big) \\ &\le C\nu^2\big(\sqrt{rR_{t-1}} + \sqrt{r}\,R_{t-1}\big) \\ &\le C\nu^2\sqrt{rR_{t-1}}.\end{aligned}$$

Here the second inequality uses Cauchy–Schwarz, and the last uses the uniform bound on $R$. This proves (85); taking $\nu = 2$ also gives (84). The gradient estimate (86) is the conditional moment bound established in the proof of Lemma D.5, specialized to $P = 2$.

For the untied process, Lemma F.1 gives

$$\big(\mathbb{E}[|\xi^{\mathrm{S}}_{k,t}|^\nu \mid \mathcal{F}^{\mathrm{S}}_{t-1}]\big)^{1/\nu} \le C\nu^2.$$

Consequently, the same conditional Minkowski and Cauchy–Schwarz inequalities yield

$$\big(\mathbb{E}[|\Xi^{\mathrm{S}}_t|^\nu \mid \mathcal{F}^{\mathrm{S}}_{t-1}]\big)^{1/\nu} \le C\nu^2 \sum_{k=0}^{r-1} |m_k(u_{t-1})| \le C\nu^2\sqrt{rR(u_{t-1})}.$$

Taking $\nu = 2$ gives the corresponding variance bound. The gradient-moment calculation in the proof of Lemma F.1, with $P = 2$, gives the asserted estimate for $G^{\mathrm{S}}_t$. All these bounds are uniform over the current weight and the conditioned teacher. □

We record two consequences for the growth argument. First, for every fixed $c_0 > 0$, whenever $R_{t-1} \ge c_0 r/d$,

$$\left(\mathbb{E}\left[\left|\frac{\Xi_t}{R_{t-1}}\right|^\nu \,\middle|\, \mathcal{F}_{t-1}\right]\right)^{1/\nu} \le C_{c_0}\nu^2\sqrt{d}, \qquad \nu \ge 2. \tag{87}$$

Indeed, $R_{t-1}$ is measurable with respect to the conditioning, and $\sqrt{r/R_{t-1}} \le \sqrt{d/c_0}$. This estimate controls the relative fluctuations of the total overlap.

Second, for every fixed $B, D > 0$, there exists $C_{B,D} < \infty$ such that, for every deterministic integer horizon $1 \le T \le d^B$,

$$\mathbb{P}\left(\max_{1\le t\le T} \|G_t\|^2 > C_{B,D}\, d(\log d)^4 \,\middle|\, w^*, w_0\right) \le d^{-D}. \tag{88}$$

This holds for sufficiently large $d$, uniformly over $w^* \in \mathcal{G}_d$ and $w_0 \in \mathbb{S}^{d-1}$. To see this, take $\nu$ to be a sufficiently large multiple of $\log d$ in (86). Conditional Markov's inequality makes the failure probability at each step at most $d^{-B-D-1}$, and a union bound over $t \le T$ proves the claim.

Both (87) and (88) hold for the untied process with the substitutions in Corollary G.4, conditioning on $w^*, u_0$ in the latter bound.

### G.5 Growth and proof of the tied upper bound

We combine the preceding estimates to prove exponential growth of the total overlap until it reaches a fixed threshold. Fix $a, \theta, c_{\mathrm{init}} \in (0, 1)$ and $C_0 > 0$, and define

$$\tau^R_a := \inf\{t \ge 0 : R_t \ge a\}, \qquad T_2 := \left\lceil \frac{C_0 \log(d/r)}{\eta_d} \right\rceil.$$

We first condition on a teacher $w^* \in \mathcal{G}_d$ and an initialization satisfying

$$R_0 \ge c_{\mathrm{init}} \frac{r}{d}. \tag{89}$$

By Lemma G.1, a uniform independent initialization satisfies this condition with conditional probability at least $1 - Ce^{-cr}$, uniformly over $w^* \in \mathcal{G}_d$.

**Proposition G.5** (Growth of the total overlap)**.** *There exist a universal constant $A_0 > 0$ and a constant $c_a > 0$, depending only on $a$, such that, for every fixed $A > A_0$ and $D > 0$, the following holds for sufficiently large $d$.*

*For every fixed teacher $w^* \in \mathcal{G}_d$ and initialization satisfying* (89)*, with probability at least $1 - d^{-D}$ over the SGD samples, simultaneously for every integer $0 \le n \le T_2$,*

$$R_{n\wedge\tau_a^R} \ge \theta R_0 \exp\big(c_a \eta_d (n \wedge \tau_a^R)\big). \tag{90}$$

To apply the relative-noise estimate, fix $\beta \in (0, \theta)$ and introduce

$$\widehat{\tau}_a^R := \inf\{t \ge 0 : R_t \le \beta R_0 \text{ or } R_t \ge a\}.$$

For $1 \le t \le \widehat{\tau}_a^R$,

$$R_{t-1} > \beta R_0 \ge \beta c_{\text{init}} \frac{r}{d},$$

so (87) applies. We will show that the lower boundary cannot be reached before $T_2$.

**Lemma G.6** (Stopped relative-noise bounds)**.** *Fix a teacher $w^* \in \mathcal{G}_d$ and an initialization satisfying* (89)*. For every fixed $D > 0$, there exists $C_D < \infty$ such that, for sufficiently large $d$, with conditional probability at least $1 - d^{-D}$ over the SGD samples,*

$$\max_{0\le n\le T_2} \left| 2\eta_d \sum_{t=1}^{n\wedge\widehat{\tau}_a^R} \frac{\Xi_t}{R_{t-1}} \right| \le C_D \eta_d \sqrt{d T_2 \log d},$$

*and, simultaneously,*

$$\left| \frac{\Xi_t}{R_{t-1}} \right| \le C_D \sqrt{d} (\log d)^2, \qquad 1 \le t \le \widehat{\tau}_a^R \wedge T_2.$$

*The constant may depend on the fixed parameters, but is uniform over the conditioned teachers and initializations.*

*Proof.* Fix the teacher and initialization as in the statement. All probabilities below are conditional on this pair.

**Step 1: Stopping and conditional moments.**

For $1 \le t \le T_2$, define

$$Y_t := \begin{cases} \dfrac{\Xi_t}{\sqrt{d}\, R_{t-1}}, & t \le \widehat{\tau}_a^R, \\ 0, & t > \widehat{\tau}_a^R. \end{cases}$$

The event $\{t \le \widehat{\tau}_a^R\}$ belongs to $\mathcal{F}_{t-1}$, and on this event,

$$R_{t-1} > \beta R_0 \ge \beta c_{\text{init}} \frac{r}{d}.$$

Thus the denominator is positive and $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = 0$. By (87), there exists $K < \infty$, uniform over the conditioned teacher and initialization, such that

$$\mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}] \le K, \qquad (\mathbb{E}[|Y_t|^\nu \mid \mathcal{F}_{t-1}])^{1/\nu} \le K\nu^2, \qquad \nu \ge 2.$$

**Step 2: Truncation and recentering.**

Fix $b > D + 2$ and set

$$q_d := \lceil b \log d \rceil, \qquad L_d := eKq_d^2.$$

Conditional Markov's inequality gives

$$\mathbb{P}\big(|Y_t| > L_d \mid \mathcal{F}_{t-1}\big) \le \frac{\mathbb{E}[|Y_t|^{q_d} \mid \mathcal{F}_{t-1}]}{L_d^{q_d}} \le e^{-q_d}.$$

Since $T_2 \asymp d(\log d)^{A+1} \le d^2$ for sufficiently large $d$, a union bound yields

$$\mathbb{P}\left(\max_{1\le t\le T_2} |Y_t| > L_d\right) \le T_2 e^{-q_d}.$$

Define the truncated and recentered increments

$$\widehat{Y}_t := Y_t \mathbf{1}_{\{|Y_t|\le L_d\}} - \mathbb{E}\big[Y_t \mathbf{1}_{\{|Y_t|\le L_d\}} \mid \mathcal{F}_{t-1}\big].$$

They are martingale differences and satisfy $|\widehat{Y}_t| \le 2L_d$. To control the centering correction, use $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = 0$ and the same moment bound:

$$\begin{aligned}\big|\mathbb{E}\big[Y_t \mathbf{1}_{\{|Y_t|\le L_d\}} \mid \mathcal{F}_{t-1}\big]\big| &= \big|\mathbb{E}\big[Y_t \mathbf{1}_{\{|Y_t|> L_d\}} \mid \mathcal{F}_{t-1}\big]\big| \\ &\le \frac{\mathbb{E}[|Y_t|^{q_d} \mid \mathcal{F}_{t-1}]}{L_d^{q_d-1}} \le L_d e^{-q_d}.\end{aligned}$$

Consequently, on the event that no increment is truncated,

$$\max_{0\le n\le T_2}\left|\sum_{t=1}^{n} Y_t\right| \le \max_{0\le n\le T_2}\left|\sum_{t=1}^{n} \widehat{Y}_t\right| + T_2 L_d e^{-q_d}.$$

**Step 3: Maximal concentration.**

Centering decreases the second moment, so

$$\mathbb{E}[\widehat{Y}_t^2 \mid \mathcal{F}_{t-1}] \le \mathbb{E}\big[Y_t^2 \mathbf{1}_{\{|Y_t|\le L_d\}} \mid \mathcal{F}_{t-1}\big] \le K.$$

The predictable quadratic variation is therefore at most $KT_2$. Applying Freedman's maximal inequality (Freedman, 1975) to both signs of the martingale gives, for every $z > 0$,

$$\mathbb{P}\left(\max_{0\le n\le T_2}\left|\sum_{t=1}^{n} \widehat{Y}_t\right| \ge z\right) \le 2\exp\left(-\frac{z^2}{2(KT_2 + \frac{2}{3}L_d z)}\right).$$

Take

$$z_d := \sqrt{2KT_2 q_d} + \frac{4}{3}L_d q_d.$$

This choice satisfies

$$z_d^2 \ge 2q_d\left(KT_2 + \frac{2}{3}L_d z_d\right),$$

so the preceding failure probability is at most $2e^{-q_d}$.

Combining this event with the absence of truncation, with probability at least

$$1 - (T_2+2)e^{-q_d} \ge 1 - d^{-D}$$

for sufficiently large $d$, we have simultaneously

$$\max_{0\le n\le T_2}\left|\sum_{t=1}^{n} Y_t\right| \le z_d + T_2 L_d e^{-q_d}, \qquad \max_{1\le t\le T_2}|Y_t| \le L_d.$$

Here

$$z_d \le C_D\big(\sqrt{T_2 \log d} + (\log d)^3\big), \qquad T_2 L_d e^{-q_d} \le C_D(\log d)^2 d^{2-b} = o(1).$$

Since $(\log d)^3 = o(\sqrt{T_2 \log d})$, multiplying the partial-sum bound by $2\eta_d\sqrt{d}$ proves the first assertion. The second follows from $|\Xi_t/R_{t-1}| = \sqrt{d}\,|Y_t|$ before and including the stopped update, and $L_d \le C_D(\log d)^2$. □

*Proof of Proposition G.5.* Fix a teacher $w^* \in \mathcal{G}_d$ and an initialization satisfying (89). All probabilities below are conditional on this pair. If $R_0 \ge a$, then $\tau_a^R = 0$ and the conclusion is immediate. Assume henceforth that $R_0 < a$.

**Step 1: Control of the stochastic terms.**

Fix $D' > D$. Apply Lemma G.6 and (88), both with failure exponent $D'$. With probability at least $1 - 2d^{-D'}$, the following bounds hold simultaneously:

$$\max_{0\le n\le T_2}\left|2\eta_d\sum_{t=1}^{n\wedge\widehat{\tau}_a^R}\frac{\Xi_t}{R_{t-1}}\right| \le C\eta_d\sqrt{dT_2\log d}, \tag{91}$$

$$\max_{1\le t\le\widehat{\tau}_a^R\wedge T_2}\left|\frac{\Xi_t}{R_{t-1}}\right| \le C\sqrt{d}(\log d)^2, \tag{92}$$

$$\max_{1\le t\le T_2}\|G_t\|^2 \le Cd(\log d)^4. \tag{93}$$

We work on this intersection. The first estimate controls the accumulated linear noise term. The second ensures that each logarithmic increment is well defined and bounds its quadratic remainder. The third controls the spherical normalization.

**Step 2: Logarithmic comparison.**

For $1 \le t \le \widehat{\tau}_a^R \wedge T_2$, we have $0 < R_{t-1} < a$. Therefore (83) gives

$$\frac{R_t}{R_{t-1}} \ge \frac{1 + c_a\eta_d + 2\eta_d\Xi_t/R_{t-1}}{1+\eta_d^2\|G_t\|^2},$$

where $c_a$ is the lower-bound constant from Proposition G.3. By (92),

$$\left|c_a\eta_d + 2\eta_d\frac{\Xi_t}{R_{t-1}}\right| \le C\eta_d\sqrt{d}(\log d)^2 = o(1).$$

Thus the numerator is positive for sufficiently large $d$, including at the update reaching either stopping boundary. In particular, $R_t > 0$ throughout these updates, so we may take logarithms.

Using $\log(1+x) \ge x - Cx^2$ for sufficiently small $|x|$ and $\log(1+y) \le y$ for $y \ge 0$, we obtain

$$\log\frac{R_t}{R_{t-1}} \ge c_a\eta_d + 2\eta_d\frac{\Xi_t}{R_{t-1}} - C\eta_d^2\left(1+\left|\frac{\Xi_t}{R_{t-1}}\right|^2\right) - \eta_d^2\|G_t\|^2.$$

Applying (92) and (93) therefore gives

$$\log\frac{R_t}{R_{t-1}} \ge c_a\eta_d + 2\eta_d\frac{\Xi_t}{R_{t-1}} - C\eta_d^2 d(\log d)^4. \tag{94}$$

Summing the logarithmic increments up to $n \wedge \widehat{\tau}_a^R$ and using (91), we find, simultaneously for every $0 \le n \le T_2$,

$$\log\frac{R_{n\wedge\widehat{\tau}_a^R}}{R_0} \ge c_a\eta_d(n\wedge\widehat{\tau}_a^R) - C\left[\eta_d\sqrt{dT_2\log d} + \eta_d^2 dT_2(\log d)^4\right].$$

Since $T_2 \asymp \eta_d^{-1}\log d$,

$$\eta_d\sqrt{dT_2\log d} = O\big(\sqrt{\eta_d d}\,\log d\big) = O\big((\log d)^{1-A/2}\big),$$
$$\eta_d^2 dT_2(\log d)^4 = O\big(\eta_d d(\log d)^5\big) = O\big((\log d)^{5-A}\big).$$

Both errors tend to zero for a sufficiently large universal choice of $A_0$. For sufficiently large $d$, their total contribution, including the constant $C$, is at most $-\log\theta$. Exponentiating yields

$$R_{n\wedge\widehat{\tau}_a^R} \ge \theta R_0\exp\big(c_a\eta_d(n\wedge\widehat{\tau}_a^R)\big), \qquad 0 \le n \le T_2.$$

**Step 3: Exclusion of the lower exit.**

Suppose that the auxiliary process exits through its lower boundary at some $\widehat{\tau}_a^R \le T_2$. Evaluating the preceding inequality at this time gives

$$R_{\widehat{\tau}_a^R} \ge \theta R_0\exp(c_a\eta_d\widehat{\tau}_a^R) \ge \theta R_0 > \beta R_0.$$

This contradicts the condition $R_{\widehat{\tau}_a^R} \le \beta R_0$ defining a lower exit. Therefore the auxiliary stopping time can only record an upper exit before $T_2$, and

$$\widehat{\tau}_a^R \wedge T_2 = \tau_a^R \wedge T_2.$$

Replacing the auxiliary stopping time by $\tau_a^R$ in the growth estimate proves (90).

The event used in the proof has conditional probability at least $1 - 2d^{-D'} \ge 1 - d^{-D}$ for sufficiently large $d$. All constants are uniform over the fixed teacher and initialization under consideration. □

We now translate total-overlap growth into subspace recovery.

*Proof of part* (T) *of Theorem B.3.* Choose $a = a(\rho) \in (\rho, 1)$ and fix $\theta, c_{\mathrm{init}} \in (0,1)$. Let $c_a$ be the constant from Proposition G.5, and choose $C_0 = C_0(\rho)$ such that $c_a C_0 > 1$.

By Lemmas C.1 and G.1, the teacher belongs to $\mathcal{G}_d$ and $R_0 \ge c_{\mathrm{init}} r/d$ with probability $1 - o(1)$. The uniform conditional estimate in Proposition G.5 then shows that (90) also holds up to $T_2$ with probability $1 - o(1)$. Work on this intersection.

If $\tau_a^R > T_2$, then

$$\begin{aligned} a > R_{T_2} &\ge \theta c_{\mathrm{init}} \frac{r}{d} \exp(c_a \eta_d T_2) \\ &\ge \theta c_{\mathrm{init}} \left(\frac{d}{r}\right)^{c_a C_0 - 1}. \end{aligned}$$

The last expression diverges because $d/r \to \infty$ and $c_a C_0 > 1$, a contradiction. Thus $\tau_a^R \le T_2$.

At that time, (80) gives

$$\|w^{\|}_{\tau_a^R}\|^2 \ge \frac{R_{\tau_a^R}}{1 + r\mu_*} \ge \frac{a}{1 + r\mu_*} > \rho$$

for sufficiently large $d$. Consequently,

$$\tau_\rho^{(2),\mathsf{T}} \le \tau_a^R \le \left\lceil \frac{C_0 \log(d/r)}{\eta_d} \right\rceil.$$

Absorbing the ceiling into a constant $C_T(\rho)$ proves part (T). □

### G.6 Augmented and untied upper bounds

We complete the proof of Theorem B.3 using the componentwise reductions from Appendix F. For each procedure, it suffices to follow one fixed weight.

*Proof of part* (A) *of Theorem B.3.* By (66), the first augmented weight follows the tied update with reference step size $\eta_d$. Its initialization has the same distribution as the tied initialization, so part (T) gives, with probability $1 - o(1)$,

$$\max_{0 \le t \le C_T \log(d/r)/\eta_d} \|(w_t^{1,\mathsf{A}})^{\|}\|^2 \ge \rho.$$

Recovery of this weight implies recovery of the student. Thus part (A) holds with $C_A = C_T$. □

*Proof of part* (U) *of Theorem B.3.* **Step 1: Effective recursion.** Retain the single-weight trajectory $u_t = w_t^{1,\mathsf{U}}$ and the quantities $G_t^{\mathrm{S}}$ and $\Xi_t^{\mathrm{S}}$ from Subsection G.4, and write

$$R_t^{\mathsf{U}} := R(u_t).$$

By (67), $\Psi = \Phi/\sqrt{r}$. Applying the quadratic-form calculation of Lemma G.2 to the effective update (65) therefore gives

$$\begin{aligned} \left(1 + \frac{\eta_d^2}{r}\|G_t^{\mathrm{S}}\|^2\right) R_t^{\mathsf{U}} = R_{t-1}^{\mathsf{U}} + \frac{\eta_d}{r}\|\nabla\Phi(u_{t-1})\|^2 \\ + \frac{2\eta_d}{\sqrt{r}}\Xi_t^{\mathrm{S}} + \frac{\eta_d^2}{r}\langle G_t^{\mathrm{S}}, B^* G_t^{\mathrm{S}}\rangle. \end{aligned} \tag{95}$$

Since $B^*$ is positive semidefinite, Proposition G.3 yields

$$\left(1+\frac{\eta_d^2}{r}\|G_t^{\mathrm{S}}\|^2\right)R_t^{\mathsf{U}} \geq \left(1+\frac{c_a\eta_d}{r}\right)R_{t-1}^{\mathsf{U}} + \frac{2\eta_d}{\sqrt{r}}\Xi_t^{\mathrm{S}} \tag{96}$$

whenever $0 < R_{t-1}^{\mathsf{U}} < a$, on $\mathcal{G}_d$ and for sufficiently large $d$. Thus the drift and normalization coefficients are divided by $r$, while the noise coefficient is divided by $\sqrt{r}$.

**Step 2: Stopped logarithmic growth.** Choose $a = a(\rho) \in (\rho, 1)$, fix $\theta, c_{\mathrm{init}} \in (0,1)$ and $\beta \in (0,\theta)$, and choose $C_0 = C_0(\rho)$ such that $c_a C_0 > 1$. Set

$$T_2^{\mathsf{U}} := \left\lceil \frac{C_0 r \log(d/r)}{\eta_d} \right\rceil.$$

Condition on a teacher $w^* \in \mathcal{G}_d$ and an initialization satisfying $R_0^{\mathsf{U}} \geq c_{\mathrm{init}} r/d$. If $R_0^{\mathsf{U}} \geq a$, recovery already holds for sufficiently large $d$ by (80). Otherwise, define

$$\widehat{\tau}_a^{R,\mathsf{U}} := \inf\{t \geq 0 : R_t^{\mathsf{U}} \leq \beta R_0^{\mathsf{U}} \text{ or } R_t^{\mathsf{U}} \geq a\}.$$

Before and including the update reaching this stopping time,

$$R_{t-1}^{\mathsf{U}} > \beta c_{\mathrm{init}} \frac{r}{d}.$$

Hence the untied relative-noise estimate from Subsection G.4 applies.

The variables $\Xi_t^{\mathrm{S}}/(\sqrt{d}\, R_{t-1}^{\mathsf{U}})$, set to zero after $\widehat{\tau}_a^{R,\mathsf{U}}$, are martingale differences with conditional $L^\nu$-norm at most $C\nu^2$. The stopping indicator includes the exit update and is measurable with respect to $\mathcal{F}_{t-1}^{\mathrm{S}}$. Since $r = d^\delta$ with $\delta < 1/2$ and $A$ is fixed,

$$T_2^{\mathsf{U}} \asymp d^{1+\delta}(\log d)^{A+1} \leq d^2$$

for sufficiently large $d$. Thus the truncation and concentration argument in the proof of Lemma G.6 applies with $T_2$ replaced by $T_2^{\mathsf{U}}$. It controls both the cumulative stopped relative noise and the individual increments over this horizon. Together with the untied version of (88), these estimates give the logarithmic comparison below.

More precisely, the one-step perturbation of the numerator in (96), after division by $R_{t-1}^{\mathsf{U}}$, satisfies

$$\left|\frac{c_a\eta_d}{r} + \frac{2\eta_d}{\sqrt{r}}\frac{\Xi_t^{\mathrm{S}}}{R_{t-1}^{\mathsf{U}}}\right| \leq C_D\frac{\eta_d}{\sqrt{r}}\sqrt{d}(\log d)^2 = o(1).$$

Thus the numerator remains positive through the exit update. Expanding the logarithm and summing, as in Subsection G.5, gives, for every fixed $D > 0$, with conditional probability at least $1 - d^{-D}$, simultaneously for $0 \leq n \leq T_2^{\mathsf{U}}$,

$$\log\frac{R_{n\wedge\widehat{\tau}_a^{R,\mathsf{U}}}^{\mathsf{U}}}{R_0^{\mathsf{U}}} \geq \frac{c_a\eta_d}{r}(n\wedge\widehat{\tau}_a^{R,\mathsf{U}}) - C_D\left[\frac{\eta_d}{\sqrt{r}}\sqrt{dT_2^{\mathsf{U}}\log d} + \frac{\eta_d^2}{r}dT_2^{\mathsf{U}}(\log d)^4\right].$$

All conditional estimates are uniform over the teacher and initialization under consideration.

Since $T_2^{\mathsf{U}} \asymp r\eta_d^{-1}\log d$, the accumulated errors satisfy

$$\frac{\eta_d}{\sqrt{r}}\sqrt{dT_2^{\mathsf{U}}\log d} = O\big((\log d)^{1-A/2}\big), \qquad \frac{\eta_d^2}{r}dT_2^{\mathsf{U}}(\log d)^4 = O\big((\log d)^{5-A}\big).$$

These are the same orders as in the tied proof. For the same sufficiently large universal choice of $A_0$, their total contribution tends to zero. Consequently, for sufficiently large $d$,

$$R_{n\wedge\widehat{\tau}_a^{R,\mathsf{U}}}^{\mathsf{U}} \geq \theta R_0^{\mathsf{U}}\exp\left(\frac{c_a\eta_d}{r}(n\wedge\widehat{\tau}_a^{R,\mathsf{U}})\right), \qquad 0 \leq n \leq T_2^{\mathsf{U}}.$$

**Step 3: Reaching the recovery threshold.**

Since $\theta > \beta$, this excludes an exit through the lower boundary before $T_2^{\mathsf{U}}$, exactly as in the tied proof. Moreover, if $\widehat{\tau}_a^{R,\mathsf{U}} > T_2^{\mathsf{U}}$, then

$$a > R^{\mathsf{U}}_{T_2^{\mathsf{U}}} \geq \theta c_{\mathrm{init}} \left(\frac{d}{r}\right)^{c_a C_0 - 1},$$

which is impossible for sufficiently large $d$. Therefore the process reaches $R_t^{\mathsf{U}} \geq a$ at some $t \leq T_2^{\mathsf{U}}$. At that time,

$$\|u_t^{\|}\|^2 \geq \frac{R_t^{\mathsf{U}}}{1 + r\mu_*} \geq \frac{a}{1 + r\mu_*} > \rho.$$

Finally, Lemma G.1 bounds the excluded initialization probability by $Ce^{-cr}$, uniformly over $w^* \in \mathcal{G}_d$. Averaging the conditional estimates and using Lemma C.1 shows that, with probability $1 - o(1)$,

$$\tau_\rho^{(2),\mathsf{U}} \leq T_2^{\mathsf{U}} \leq \frac{C_U(\rho) r \log(d/r)}{\eta_d},$$

after absorbing the ceiling into $C_U(\rho)$. This proves part (U). □

Together with part (T), a finite union bound gives the three recovery guarantees simultaneously with probability $1 - o(1)$.

*Remark* G.7 (Shifted tied weights). The subspaces $\mathcal{S}^*$ and $(\mathcal{S}^*)^\perp$ are invariant under $\Pi$. Consequently,

$$(\Pi^k w)^{\|} = \Pi^k w^{\|}, \qquad \|(\Pi^k w)^{\|}\|^2 = \|w^{\|}\|^2.$$

Thus, when the tied weight achieves subspace recovery, all its cyclic shifts have the same projection mass in the teacher subspace.

## H Lower bounds for weak recovery

In this appendix, we prove Theorem B.4 under its assumptions on the teacher, initialization, and scaling. The link is a fixed centered polynomial of degree $P$ with information exponent $p \geq 2$. The population identities and componentwise reductions from Appendices D and F apply to this entire class: their derivations use the finite Hermite expansion and do not require a parity assumption. Throughout, $A > 0$ is fixed, the reference learning rate is

$$\eta_d = d^{-p/2} (\log d)^{-A},$$

and the three procedures use the step sizes in (6). We retain the notation $w_t^{j,\mathsf{k}}$ and $s_{\mathsf{k}}$ from Appendix B.

### H.1 Proof strategy

We show that all absolute correlations remain at their initialization scale $O(\gamma_d)$ for a sufficiently small constant multiple of the microscopic growth time. This excludes both weak directional recovery and weak subspace recovery. The argument applies simultaneously to every weight, including when the procedures share their observations.

The proof has two ingredients. First, we bound the absolute population drift by the largest absolute correlation raised to the power $p - 1$. Lemma E.6 then gives a pathwise upper comparison that retains the exact spherical normalization. Second, we control the accumulated coordinate noise until the first escape from the initialization scale, uniformly over all weights and teacher directions.

For the tied and augmented procedures, the relevant horizon is of order $(\eta_d \gamma_d^{p-2})^{-1}$. For the untied procedure, it is longer by a factor $r$. On these horizons, the stochastic contribution is $o(\gamma_d)$, while a sufficiently small time constant makes the accumulated drift smaller than the distance to the escape threshold. Evaluating the comparison at a hypothetical first exit gives the desired contradiction.

### H.2 Initialization and escape times

Define the largest absolute correlation of a weight and of a procedure by

$$\widehat{M}(w) \coloneqq \max_{0\le k<r} |m_k(w)|, \qquad \widehat{M}_t^{\mathsf{k}} \coloneqq \max_{1\le j\le s_{\mathsf{k}}} \widehat{M}(w_t^{j,\mathsf{k}}).$$

We control absolute correlations for every $p \ge 2$.

Lemma C.2 establishes that $\widehat{M}(w_0) = O(\gamma_d)$ with probability $1 - o(1)$ for a single random weight. Here we need this upper bound simultaneously over all initial weights, so we use a quantitative tail estimate before taking the union bound.

Fix a teacher $w^*$. By Assumption 2, each initial weight is conditionally uniform on $\mathbb{S}^{d-1}$. The standard spherical-cap bound therefore gives, for every procedure, weight, and teacher direction,

$$\mathbb{P}\left(|m_k(w_0^{j,\mathsf{k}})| > z \,\middle|\, w^*\right) \le Ce^{-cdz^2}, \qquad 0 < z < 1,$$

with universal constants $c, C > 0$. Since $\gamma_d = o(1)$, we may take $z = C_{\mathrm{init}}\gamma_d$ for any fixed $C_{\mathrm{init}} > 0$ and sufficiently large $d$. A union bound over the $(1+2s)r$ initial correlations then yields

$$\mathbb{P}\left(\max_{\mathsf{k}\in\{\mathsf{T},\mathsf{U},\mathsf{A}\}} \widehat{M}_0^{\mathsf{k}} > C_{\mathrm{init}}\gamma_d \,\middle|\, w^*\right) \le C(1+2s)re^{-cC_{\mathrm{init}}^2 \log r} \le Cd^{\kappa+\delta-cC_{\mathrm{init}}^2\delta}. \tag{97}$$

The last inequality uses $s \le d^\kappa$ and $r = d^\delta$ from Assumption 3. Choose $C_{\mathrm{init}} = C_{\mathrm{init}}(\delta,\kappa)$ large enough that $cC_{\mathrm{init}}^2\delta > \kappa + \delta$. Then (97) is $o(1)$, uniformly over the teacher.

Combining this estimate with Lemma C.1, we obtain, with probability $1 - o(1)$,

$$w^* \in \mathcal{G}_d, \qquad \max_{\mathsf{k}\in\{\mathsf{T},\mathsf{U},\mathsf{A}\}} \widehat{M}_0^{\mathsf{k}} \le C_{\mathrm{init}}\gamma_d.$$

We will first establish the subsequent estimates conditionally on a teacher and initial weights satisfying these conditions.

Fix any constant $C_{\mathrm{esc}} > C_{\mathrm{init}}$ and write $\tau_{\mathrm{esc}}^{\mathsf{k}} \coloneqq \tau_{\mathrm{esc}}^{\mathsf{k}}(C_{\mathrm{esc}})$ for the escape times defined in (12). On the initialization event above, $\tau_{\mathrm{esc}}^{\mathsf{k}} \ge 1$. For every integer $1 \le t \le \tau_{\mathrm{esc}}^{\mathsf{k}}$,

$$\max_{1\le j\le s_{\mathsf{k}}} \widehat{M}(w_{t-1}^{j,\mathsf{k}}) < C_{\mathrm{esc}}\gamma_d.$$

Thus bounds evaluated at the preceding iterate remain valid through the update producing the first exit. We will use this observation to apply the upper comparison at a hypothetical exit time and obtain a contradiction.

### H.3 An upper comparison for the correlations

Throughout this subsection, $\Phi$ denotes the tied population objective. The finite Hermite expansion implies

$$|\phi'(z)| \le C_\sigma |z|^{p-1}, \qquad |z| \le 1.$$

We combine this bound with Lemma D.1 and the overlap-energy estimate (36).

Let $\mathcal{F}_t$ denote the joint filtration generated by the teacher, all initial weights, and all observations used through time $t$. For $\mathsf{k} \in \{\mathsf{T},\mathsf{A}\}$, let $G_t^{j,\mathsf{k}}$ be the tied sample gradient evaluated at $w_{t-1}^{j,\mathsf{k}}$. For $\mathsf{k} = \mathsf{U}$, let it be the single-weight sample gradient associated with (61). Define

$$\xi_{k,t}^{j,\mathsf{k}} \coloneqq -\left\langle \Pi^k w^*,\, G_t^{j,\mathsf{k}} - \mathbb{E}[G_t^{j,\mathsf{k}} \mid \mathcal{F}_{t-1}] \right\rangle.$$

These are the coordinate fluctuations already used in Appendices D and F, now indexed by the procedure and weight. Fresh-sample independence gives

$$\mathbb{E}[G_t^{j,\mathsf{k}} \mid \mathcal{F}_{t-1}] = \begin{cases} \nabla\Phi(w_{t-1}^{j,\mathsf{k}}), & \mathsf{k} \in \{\mathsf{T},\mathsf{A}\}, \\ \nabla\Psi(w_{t-1}^{j,\mathsf{U}}), & \mathsf{k} = \mathsf{U}, \end{cases}$$

where $\Psi$ is the single-weight population objective from (67). By the definition of the centered fluctuations,

$$\mathbb{E}[\xi_{k,t}^{j,\mathsf{k}} \mid \mathcal{F}_{t-1}] = 0.$$

**Lemma H.1** (Upper comparison before escape)**.** *On $\mathcal{G}_d$, there exists $C_\sigma < \infty$ such that, for all sufficiently large $d$ and every $w \in \mathbb{S}^{d-1}$,*

$$\max_{0\le k<r} \left|\left\langle \Pi^k w^*, \nabla\Phi(w)\right\rangle\right| \le C_\sigma \widehat{M}(w)^{p-1}.$$

*Moreover, for $\mathsf{k} \in \{\mathsf{T}, \mathsf{A}\}$ and every integer $0 \le n \le \tau_{\mathrm{esc}}^{\mathsf{k}}$,*

$$\widehat{M}_n^{\mathsf{k}} \le \widehat{M}_0^{\mathsf{k}} + C_\sigma \eta_d n (C_{\mathrm{esc}}\gamma_d)^{p-1} + 2\eta_d \max_{\substack{1\le j\le s_{\mathsf{k}}\\ 0\le k<r}} \max_{0\le h\le n} \left|\sum_{t=1}^{h} \xi_{k,t}^{j,\mathsf{k}}\right|. \tag{98}$$

*For the untied procedure and every integer $0 \le n \le \tau_{\mathrm{esc}}^{\mathsf{U}}$,*

$$\widehat{M}_n^{\mathsf{U}} \le \widehat{M}_0^{\mathsf{U}} + C_\sigma \frac{\eta_d}{r} n (C_{\mathrm{esc}}\gamma_d)^{p-1} + \frac{2\eta_d}{\sqrt{r}} \max_{\substack{1\le j\le s\\ 0\le k<r}} \max_{0\le h\le n} \left|\sum_{t=1}^{h} \xi_{k,t}^{j,\mathsf{U}}\right|. \tag{99}$$

*Both comparisons are pathwise and include the first exit step.*

*Proof.* **Population drift.** Fix $w \in \mathbb{S}^{d-1}$ and write $M = \widehat{M}(w)$. If $M = 0$, then every correlation vanishes and $\phi'(0) = 0$, so Lemma D.1 gives $\nabla\Phi(w) = 0$. Suppose henceforth that $M > 0$. Since $p \ge 2$,

$$\sum_{\ell=0}^{r-1} |m_\ell(w)|^p \le M^{p-2} \sum_{\ell=0}^{r-1} m_\ell(w)^2.$$

By Lemma D.1, the derivative bound above, and the definition of $\mu_*$,

$$\begin{aligned}
\left|\left\langle \Pi^k w^*, \nabla\Phi(w)\right\rangle\right| &\le C_\sigma \left[(1 + r\mu_*)M^{p-1} + M \sum_{\ell=0}^{r-1} |m_\ell(w)|^p\right] \\
&\le C_\sigma M^{p-1} \left[1 + r\mu_* + \sum_{\ell=0}^{r-1} m_\ell(w)^2\right] \\
&\le C_\sigma M^{p-1}.
\end{aligned}$$

The last inequality uses $r\mu_* = o(1)$ on $\mathcal{G}_d$ and the uniform overlap-energy bound (36). After enlarging $C_\sigma$, this proves the first assertion, including the case $p = 2$.

**Tied and augmented weights.** Fix $\mathsf{k} \in \{\mathsf{T}, \mathsf{A}\}$, a weight index $j$, and a teacher direction $k$. The exact coordinate update from the proof of Lemma D.2, together with (66), has the form

$$d_t x_{t+1} = x_t + F_t + N_{t+1} - N_t,$$

where

$$x_t = m_k(w_t^{j,\mathsf{k}}), \qquad d_t = \sqrt{1 + \eta_d^2 \|G_{t+1}^{j,\mathsf{k}}\|^2},$$

and

$$F_t = -\eta_d \left\langle \Pi^k w^*, \nabla\Phi(w_t^{j,\mathsf{k}})\right\rangle,$$

$$N_t = \eta_d \sum_{\ell=1}^{t} \xi_{k,\ell}^{j,\mathsf{k}}.$$

Thus $d_t \ge 1$ and $N_0 = 0$, so Lemma E.6 applies with $q_t = 1$.

For every integer $n \le \tau_{\mathrm{esc}}^{\mathsf{k}}$, each $t < n$ lies strictly before escape. The population bound just proved therefore gives

$$\sum_{t=0}^{n-1} |F_t| \le C_\sigma \eta_d \sum_{t=0}^{n-1} \widehat{M}(w_t^{j,\mathsf{k}})^{p-1} \le C_\sigma \eta_d n (C_{\mathrm{esc}}\gamma_d)^{p-1}.$$

Consequently, Lemma E.6 yields

$$|m_k(w_n^{j,\mathsf{k}})| \le |m_k(w_0^{j,\mathsf{k}})| + C_\sigma \eta_d n (C_{\mathrm{esc}}\gamma_d)^{p-1} + 2\eta_d \max_{0\le h\le n} \left| \sum_{\ell=1}^{h} \xi_{k,\ell}^{j,\mathsf{k}} \right|.$$

Taking the maximum over $j$ and $k$ proves (98).

**Untied weights.** By (67), $\nabla\Psi = \nabla\Phi/\sqrt{r}$. For a fixed untied weight $j$ and teacher direction $k$, the exact update (69) therefore has the same damping form, now with

$$x_t = m_k(w_t^{j,\mathsf{U}}), \qquad d_t = \sqrt{1 + \frac{\eta_d^2}{r}\|G_{t+1}^{j,\mathsf{U}}\|^2},$$

and

$$F_t = -\frac{\eta_d}{r}\left\langle \Pi^k w^*, \nabla\Phi(w_t^{j,\mathsf{U}})\right\rangle,$$
$$N_t = \frac{\eta_d}{\sqrt{r}} \sum_{\ell=1}^{t} \xi_{k,\ell}^{j,\mathsf{U}}.$$

Again $d_t \ge 1$, and for $n \le \tau_{\mathrm{esc}}^{\mathsf{U}}$,

$$\sum_{t=0}^{n-1} |F_t| \le C_\sigma \frac{\eta_d}{r} n (C_{\mathrm{esc}}\gamma_d)^{p-1}.$$

Applying Lemma E.6 with $q_t = 1$ and taking the maximum over all weights and teacher directions gives (99).

Both applications are pathwise. In particular, all drift terms used to bound the correlation at time $n$ are evaluated at times $t < n$. The comparisons therefore remain valid when $n$ is the first exit time. □

### H.4 Noise before escape

The moment estimates of Lemma D.4 apply to the tied and augmented weights, while Lemma F.1 applies to the untied weights. Their proofs use only the finite Hermite expansion and $p \ge 2$, so the estimates hold for every polynomial link considered in Theorem B.4. Before escape,

$$r\, m_k(w_{t-1}^{j,\mathsf{k}})^2 \le C_{\mathrm{esc}}^2 r \gamma_d^2 = o(1).$$

Thus the stopped coordinate fluctuations have bounded conditional variance and conditional $L^\nu$-norm of order $\nu^P$.

**Lemma H.2** (Noise before escape)**.** *Fix a teacher $w^* \in \mathcal{G}_d$ and initial weights satisfying*

$$\max_{\mathsf{k}\in\{\mathsf{T},\mathsf{U},\mathsf{A}\}} \widehat{M}_0^{\mathsf{k}} \le C_{\mathrm{init}}\gamma_d.$$

*For every fixed $D > 0$, there exists $C_{\sigma,D} < \infty$ such that the following holds for all sufficiently large $d$. For any deterministic integer horizons $1 \le T_{\mathsf{k}} \le d^p$, with probability at least $1 - d^{-D}$ over the SGD samples, simultaneously for $\mathsf{k} \in \{\mathsf{T},\mathsf{U},\mathsf{A}\}$,*

$$\max_{\substack{1\le j\le s_{\mathsf{k}}\\ 0\le k<r}} \max_{0\le n\le T_{\mathsf{k}}} \left| \sum_{t=1}^{n\wedge\tau_{\mathrm{esc}}^{\mathsf{k}}} \xi_{k,t}^{j,\mathsf{k}} \right| \le C_{\sigma,D}\left(\sqrt{T_{\mathsf{k}}\log d} + (\log d)^{P+1}\right). \tag{100}$$

*The constant may also depend on $\delta, \kappa, C_{\mathrm{init}}, C_{\mathrm{esc}}$. The constants and dimension threshold are uniform over the conditioned teacher and initial weights.*

*Proof.* Fix the teacher and initial weights as in the statement, and work under the corresponding conditional law.

For each procedure, weight, and teacher direction, set

$$Y_t := \mathbf{1}_{\{t\le\tau_{\mathrm{esc}}^{\mathsf{k}}\}}\xi_{k,t}^{j,\mathsf{k}}, \qquad 1 \le t \le T_{\mathsf{k}},$$

suppressing the coordinate indices in $Y_t$. The stopping indicator is $\mathcal{F}_{t-1}$-measurable, so $(Y_t)$ is a martingale difference sequence. On $\{t \le \tau_{\mathrm{esc}}^{\mathsf{k}}\}$,

$$r\, m_k (w_{t-1}^{j,\mathsf{k}})^2 \le C_{\mathrm{esc}}^2 r \gamma_d^2 = o(1).$$

Consequently, Lemmas D.4 and F.1 give, uniformly over all coordinate processes,

$$\mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}] \le C_\sigma, \qquad (\mathbb{E}[|Y_t|^\nu \mid \mathcal{F}_{t-1}])^{1/\nu} \le C_\sigma \nu^P, \qquad \nu \ge 2.$$

The earlier moment calculations apply to the joint filtration because the observation at time $t$ is fresh conditionally on $\mathcal{F}_{t-1}$.

These are exactly the conditional moment bounds used in the proof of Lemma E.5. Apply its truncation, recentering, and maximal-concentration argument with moment order proportional to $\log d$ and cutoff proportional to $(\log d)^P$. For any fixed $b > p$, choosing the proportionality constants sufficiently large gives

$$\mathbb{P}\left( \max_{0 \le n \le T_{\mathsf{k}}} \left| \sum_{t=1}^{n} Y_t \right| > C_{\sigma,b} \left[ \sqrt{T_{\mathsf{k}} \log d} + (\log d)^{P+1} \right] \right) \le C(T_{\mathsf{k}} + 1) d^{-b}.$$

Here the factor $T_{\mathsf{k}}$ accounts for the truncation events, while the maximal martingale bound contributes $Cd^{-b}$. The accumulated centering correction is at most

$$C_{\sigma,b} T_{\mathsf{k}} (\log d)^P d^{-b} \le C_{\sigma,b} d^{p-b} (\log d)^P = o(1),$$

and is absorbed into the displayed threshold.

Choose $b > D + \kappa + \delta + p$. A union bound over the $(1+2s)r$ coordinate processes bounds the total failure probability by

$$C(1+2s)r(d^p+1)d^{-b} \le C d^{\kappa+\delta+p-b} \le d^{-D}$$

for sufficiently large $d$. No independence between these processes is needed.

Finally,

$$\sum_{t=1}^{n} Y_t = \sum_{t=1}^{n \wedge \tau_{\mathrm{esc}}^{\mathsf{k}}} \xi_{k,t}^{j,\mathsf{k}},$$

so the simultaneous partial-sum bounds give (100). The stopping convention includes the exit-producing increment, and all estimates are uniform over the conditioned teacher and initial weights. □

### H.5 Proof of the non-escape theorem

We now combine Lemmas H.1 and H.2 to prove Theorem B.4.

*Proof of Theorem B.4.* Take $C_{\mathrm{init}}$ as in Subsection H.2 and fix $C_{\mathrm{esc}} > C_{\mathrm{init}}$ depending only on $\delta, \kappa$. Let $C_\sigma$ be the constant in Lemma H.1, and choose $c > 0$ so that

$$C_{\mathrm{init}} + C_\sigma c C_{\mathrm{esc}}^{p-1} < C_{\mathrm{esc}}.$$

With $T_d$ as in the theorem, define

$$T_{\mathsf{T}} = T_{\mathsf{A}} \coloneqq \lfloor T_d \rfloor, \qquad T_{\mathsf{U}} \coloneqq \lfloor r T_d \rfloor.$$

Since

$$\frac{1}{\eta_d \gamma_d^{p-2}} = \frac{d^{p-1}(\log d)^A}{(\log r)^{(p-2)/2}} \asymp d^{p-1}(\log d)^{A-(p-2)/2},$$

all three horizons diverge. Moreover,

$$T_{\mathsf{U}} = O\Big( d^{p-1+\delta} (\log d)^{A-(p-2)/2} \Big) = o(d^p).$$

Thus $1 \le T_{\mathsf{k}} \le d^p$ for sufficiently large $d$, and Lemma H.2 applies.

**Step 1: The accumulated noise is smaller than initialization.** Fix a teacher $w^* \in \mathcal{G}_d$ and initial weights satisfying $\max_{\mathsf{k}} \widehat{M}_0^{\mathsf{k}} \le C_{\mathrm{init}}\gamma_d$. Apply Lemma H.2 with the horizons above and any fixed $D > 0$. On its event, which has conditional probability at least $1 - d^{-D}$, the stochastic terms in (98) and (99) are $o(\gamma_d)$ simultaneously for all procedures. Indeed, the definitions of the horizons give

$$\max\left\{\frac{\eta_d\sqrt{T_{\mathsf{T}}\log d}}{\gamma_d}, \frac{\eta_d\sqrt{T_{\mathsf{U}}\log d}}{\sqrt{r}\,\gamma_d}\right\} \le \sqrt{c\eta_d\gamma_d^{-p}\log d} = O\Big((\log d)^{-A/2-(p-2)/4}\Big) = o(1).$$

For the remaining term,

$$\frac{\eta_d(\log d)^{P+1}}{\gamma_d} = O\Big(d^{-(p-1)/2}(\log d)^{P+1/2-A}\Big) = o(1).$$

The corresponding untied term is smaller by $r^{-1/2}$. Both estimates hold for every fixed $A > 0$ and $2 \le p \le P < \infty$.

**Step 2: Excluding the first exit.** Work on this common noise event. For $n \le \tau_{\mathrm{esc}}^{\mathsf{k}} \wedge T_{\mathsf{k}}$, every partial sum with $h \le n$ agrees with its stopped counterpart in (100). For $\mathsf{k} \in \{\mathsf{T}, \mathsf{A}\}$, Lemma H.1 therefore gives

$$\widehat{M}_n^{\mathsf{k}} \le C_{\mathrm{init}}\gamma_d + C_\sigma\eta_d T_{\mathsf{k}}(C_{\mathrm{esc}}\gamma_d)^{p-1} + o(\gamma_d) \le \big(C_{\mathrm{init}} + C_\sigma c C_{\mathrm{esc}}^{p-1} + o(1)\big)\,\gamma_d.$$

For the untied procedure, the same comparison yields

$$\widehat{M}_n^{\mathsf{U}} \le C_{\mathrm{init}}\gamma_d + C_\sigma\frac{\eta_d}{r}T_{\mathsf{U}}(C_{\mathrm{esc}}\gamma_d)^{p-1} + o(\gamma_d) \le \big(C_{\mathrm{init}} + C_\sigma c C_{\mathrm{esc}}^{p-1} + o(1)\big)\,\gamma_d.$$

By the choice of $c$, these upper bounds are strictly smaller than $C_{\mathrm{esc}}\gamma_d$ for sufficiently large $d$. If any procedure had $\tau_{\mathrm{esc}}^{\mathsf{k}} \le T_{\mathsf{k}}$, applying its comparison at $n = \tau_{\mathrm{esc}}^{\mathsf{k}}$ would contradict the definition of the exit time. Consequently, simultaneously for all three procedures,

$$\tau_{\mathrm{esc}}^{\mathsf{k}} > T_{\mathsf{k}}, \qquad \max_{0\le t\le T_{\mathsf{k}}} \widehat{M}_t^{\mathsf{k}} < C_{\mathrm{esc}}\gamma_d.$$

The conditional estimates are uniform over the teacher and initial weights under consideration. Lemma C.1 and (97) show that these conditioning requirements hold with probability $1 - o(1)$. Averaging gives the same unconditional probability for the simultaneous non-escape bounds. Since the escape times are integer-valued, they exceed the unrounded horizon $T_d$ for the tied and augmented procedures and $rT_d$ for the untied procedure. This proves the theorem, with $c$ and $C_{\mathrm{esc}}$ depending only on $\sigma, \delta, \kappa$. □