Separation in Feature Alignment Between Muon and SGD

Working Draft

1 Introduction

This note studies why a Muon-style update aligns features faster than vanilla SGD. We use a minimal "Simplified Muon" that discards gradient magnitudes and steps with the orthogonal factor (right polar) of the batch gradient, and we compare it to SGD; see, e.g., Bernstein and Newhouse [2024] for a related Newton–Schulz approach to extracting polar factors.

Scope and contributions (informal):

- Warm-up: single-neuron (rank-1) teacher-student. We show a separation in alignment drift: Simplified Muon yields linear-in-angle drift near alignment, whereas SGD yields only quadratic drift.
- Multi-index teacher (rank $r \geq 1$) and width-m student under Gaussian mini-batches. We formalize Muon as the right polar step of the batch gradient and study it alongside mini-batch SGD.
- Batch-size and loss sensitivity: MUON benefits from larger mini-batches (B) and smaller training loss (cleaner polar orientation), yielding stronger alignment drift; by contrast, SGD is essentially insensitive to B and to the residual scale.
- Under an AGOP condition, the student covariance is axially structured and aligns toward the teacher subspace; we analyze alignment via trace-centered covariances and show stronger drift for Muon.

We work with the following optimizer, used throughout the paper. In the single-neuron model analyzed in Section 2, it reduces to the normalized, sign-based update in Definition "Update Rules." We will compare three update rules: *(i)* Simplified Muon (Alg. 1), which orthogonalizes the gradient via an SVD and steps with the factor $U_t V_t^\top$, effectively discarding magnitude information; *(ii)* Vanilla SGD (Alg. 3), which uses the full gradient $\nabla f(W_t)$; and *(iii)* Stochastic/mini-batch SGD (Alg. 2), which updates using a mini-batch gradient of size B: $g_t = \frac{1}{B}\sum_{i \in S_t} \nabla \ell(W_t; x_i, y_i)$. In the teacher–student analysis below, Simplified Muon further reduces to a sign-normalized step along the input direction.

Algorithm 1 Simplified Muon

Require: Initial weights W_0, learning rate schedule $\{\eta_t\}$
1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: $\quad G_t \leftarrow \nabla f(W_t)$
3: $\quad (U_t, S_t, V_t) \leftarrow \mathrm{SVD}(G_t)$
4: $\quad W_{t+1} \leftarrow W_t - \eta_t U_t V_t^\top$
5: **end for**

In words, Alg. 1 keeps only the orientation of the gradient by stepping with $U_t V_t^\top$; Alg. 3 follows the full (batch) gradient; and Alg. 2 uses a mini-batch stochastic gradient of size B. This trio lets us separate two effects: orthogonalization/normalization (Muon) versus stochasticity (SGD). The rest of the paper analyzes their alignment dynamics in the teacher–student model.

Algorithm 2 Stochastic SGD (mini-batch size B)

Require: Initial weights W_0, learning rate schedule $\{\eta_t\}$, batch size B, data distribution $\mathcal{D}$

for $t = 0, 1, \ldots, T-1$ **do**
  Sample mini-batch $S_t = \{(x_i, y_i)\}_{i=1}^{B} \overset{\text{i.i.d.}}{\sim} \mathcal{D}$
  $g_t \leftarrow \frac{1}{B} \sum_{i=1}^{B} \nabla_W \ell(W_t; x_i, y_i)$
  $W_{t+1} \leftarrow W_t - \eta_t \, g_t$
end for

Algorithm 3 Vanilla SGD

Require: Initial weights W_0, learning rate schedule $\{\eta_t\}$

for $t = 0, 1, \ldots, T-1$ **do**
  $G_t \leftarrow \nabla f(W_t)$
  $W_{t+1} \leftarrow W_t - \eta_t G_t$
end for

2 Problem Setting: One-neuron Teacher–Student

We begin by analyzing the simplest case: a single-neuron teacher-student model, which provides clear intuition for feature alignment dynamics.

Let the input data be drawn from a standard Gaussian distribution, $x \sim \mathcal{N}(0, I_p)$. The *target function* is a single neuron with a fixed, unknown weight vector $w_* \in \mathbb{R}^p$ with $\|w_*\| = 1$. Its output for an input x is: where $\sigma : \mathbb{R} \to \mathbb{R}$ is an *odd*, twice-differentiable activation function, with $\sigma'(0) > 0$. Recall that an odd function satisfies $\sigma(-x) = -\sigma(x)$, which implies its first derivative σ' is even ($\sigma'(-x) = \sigma'(x)$), and its second derivative σ'' is odd ($\sigma''(-x) = -\sigma''(x)$). where $\sigma : \mathbb{R} \to \mathbb{R}$ is an *odd*, twice-differentiable activation function, with $\sigma'(0) > 0$. The oddness of σ implies its first derivative σ' is even and its second derivative σ'' is odd.

The learner is another neuron with a trainable weight vector $w_t \in \mathbb{R}^p$. Its prediction is $f(x; w_t) = \sigma(w_t^\top x)$. At each step t, the student receives a single sample (x_t, y_t) and updates its weights to minimize the squared loss $\ell_t = \frac{1}{2}(f(x_t; w_t) - y_t)^2$.

We define the following quantities for convenience: The key measure of success is the **alignment** between the target and learnt weights, quantified by the angle θ between them. We define $\rho_t = \cos\theta = w_t^\top w_* / \|w_t\|$.

We assume $\|w_t\| \approx 1$, which is justified since the update rules either normalize the weights or use small learning rates, so the norm remains close to 1 throughout training; this allows us to interpret ρ_t directly as the cosine of the angle between w_t and w_*. But we note that the analysis can be adapted to the general case by considering the normalized direction $u_t = w_t / \|w_t\|$.

Using the algorithms defined above, the update at step t for each method is as follows:

Definition 1 (Update Rules)**.** *The weights are updated via $w_{t+1} = w_t - \eta \Delta w_t$ for a small step size $\eta > 0$. We analyze two algorithms defined by their update directions Δw_t:*

$$\Delta w_t^{\text{SGD}} := r_t \sigma'(u_t) x_t$$
$$\Delta w_t^{\text{MUON}} := \text{sign}(r_t) \frac{x_t}{\|x_t\|}$$

where $u_t = w_t^\top x_t$, $v_t = w_*^\top x_t$, and $r_t = \sigma(u_t) - \sigma(v_t)$ is the residual.

[**F:** add simplified Adam]

Contents

3 Single neuron: Separation in alignment drift

Theorem 1 (Alignment Gap)**.** *Under the stated assumptions, for a small angle θ, the expected one-step drift of the alignment $\rho_t = \cos\theta$ for* SGD *and* MUON *is given by:*

$$\mathbb{E}[\rho_{t+1} - \rho_t \mid \theta] = \begin{cases} C_{\text{MUON}} \frac{\eta}{\sqrt{p}} \theta + O(\eta^2) & \text{for MUON} \\ C_{\text{SGD}}\, \eta\, \theta^2 + O(\eta\theta^3) + O(\eta^2) & \text{for SGD} \end{cases}$$

where C_{MUON} and C_{SGD} are positive constants.

To prove this, we first establish a general formula for the alignment drift.

Lemma 1 (General Drift Equation)**.** *For an update $w_{t+1} = w_t - \eta\Delta w_t$ and assuming $\|w_t\| = 1$, the expected change in $\rho_t = w_t^\top w_*$ is:*

$$\mathbb{E}[\rho_{t+1} - \rho_t] = -\eta\mathbb{E}[(\Delta w_t)^\top w_*] + \eta\rho_t\mathbb{E}[w_t^\top \Delta w_t] + O(\eta^2)$$

The $O(\eta^2)$ term arises from higher-order terms in the Taylor expansion of the denominator and numerator when expanding $\|w_t - \eta\Delta w_t\|$ and $(w_t - \eta\Delta w_t)^\top w_*$; it can be neglected when the learning rate η is sufficiently small, so that second and higher-order terms have minimal effect on the alignment drift.

Remark 1 (On the unit-norm assumption)**.** *Our analysis states $\rho_t = \cos\theta_t = w_t^\top w_\star$ assuming $\|w_t\| \approx 1$. One can drop this assumption and work with the normalized direction $u_t := w_t/\|w_t\|$ and $\rho_t := u_t^\top w_\star$. A first-order expansion on the sphere yields*

$$u_{t+1} = u_t - \eta\,(I - u_t u_t^\top)\,\Delta w_t + O(\eta^2), \quad \Rightarrow \quad \rho_{t+1} - \rho_t = -\eta\big(\Delta w_t^\top w_\star - \rho_t\,\Delta w_t^\top u_t\big) + O(\eta^2).$$

This coincides with Lemma 1 after replacing w_t by u_t. In particular, for Muon's right-polar step the tangentiality condition $u_t^\top \Delta w_t = 0$ holds (ideally) at first order, preserving the norm up to $O(\eta^2)$ and simplifying the drift.

Proof. By definition we first note that

$$\rho_{t+1} = \frac{(w_t - \eta\Delta w_t)^\top w_*}{\|w_t - \eta\Delta w_t\|}.$$

Next, we expand the denominator:

$$\|w_t - \eta\Delta w_t\|^2 = \|w_t\|^2 + \eta^2\|\Delta w_t\|^2 - 2w_t^\top \Delta w_t = \|w_t\|^2 \left(1 + \eta^2 \frac{\|\Delta w_t\|^2}{\|w_t\|^2} - \eta\frac{2w_t^\top \Delta w_t}{\|w_t\|^2}\right)$$

which can be expanded, using Taylor series as follows

$$\begin{aligned}\sqrt{1 + \eta^2 \frac{\|\Delta w_t\|^2}{\|w_t\|^2} - \eta\frac{2w_t^\top \Delta w_t}{\|w_t\|^2}} &= 1 + \frac{1}{2}\left(\eta^2 \frac{\|\Delta w_t\|^2}{\|w_t\|^2} - \eta\frac{2w_t^\top \Delta w_t}{\|w_t\|^2}\right) - \frac{1}{8}\left(\eta^2 \frac{\|\Delta w_t\|^2}{\|w_t\|^2} - \eta\frac{2w_t^\top \Delta w_t}{\|w_t\|^2}\right)^2 + \dots \\ &= 1 + \frac{1}{2}\left(\eta^2 \frac{\|\Delta w_t\|^2}{\|w_t\|^2} - \eta\frac{2w_t^\top \Delta w_t}{\|w_t\|^2}\right) + O(\eta^2)\end{aligned}$$

where we we approximate the higher order term as $O(\eta^2)$. Thus, we can write

$$\rho_{t+1} = \frac{(w_t - \eta\Delta w_t)^\top w_*}{\|w_t - \eta\Delta w_t\|} = \frac{w_t^\top w_* - \eta(\Delta w_t)^\top w_*}{\|w_t\|\,(1 - \eta w_t^\top \Delta w_t/\,\|w_t\|^2 + O(\eta^2))}$$

Furthermore, using the series

$$\frac{1}{1-x} = 1 + x + x^2 + \dots$$

we get the following simplification

$$\rho_{t+1} = \rho_t - \eta(\Delta w_t)^\top w_* + \eta\rho_t(w_t^\top \Delta w_t) + O(\eta^2)$$

□

We now apply this to prove Theorem 1 for each algorithm.

3.1 Linear Drift for Muon

Proof for MUON. The update is $\Delta w_t^{\text{MUON}} = \text{sign}(r_t)\frac{x_t}{\|x_t\|}$. Let $u = x_t/\|x_t\|$ be a vector uniformly distributed on the unit sphere S^{p-1}. Since σ is odd and increasing near zero, $\text{sign}(r_t) = \text{sign}(\sigma(w_t^\top x_t) - \sigma(w_*^\top x_t)) = \text{sign}((w_t - w_*)^\top x_t)$.

We need to compute the expectation:

$$\mathbb{E}[\Delta w_t^{\text{MUON}}] = \mathbb{E}\left[u \cdot \text{sign}\left((w_t - w_*)^\top u\right)\right] = \mathbb{E}\left[\frac{x_t}{\|x_t\|} \cdot \text{sign}\left((w_t - w_*)^\top \frac{x_t}{\|x_t\|}\right)\right]$$

Radial input model. Assume $x_t = R_t U_t$ with $U_t \sim \text{Unif}(S^{p-1})$ and $R_t \geq 0$ independent of U_t (e.g., $x_t \sim \mathcal{N}(0, I_p)$). Then $x_t/\|x_t\|_2 = U_t$ and $\text{sign}(a^\top x_t) = \text{sign}(a^\top U_t)$ a.s.

Lemma 2 (Spherical sign–mean)**.** *Let $U \sim \text{Unif}(S^{p-1})$ and $a \in \mathbb{R}^p \setminus \{0\}$. Then*

$$\mathbb{E}\big[U\,\text{sign}(a^\top U)\big] = \kappa_p \frac{a}{\|a\|_2}, \qquad \kappa_p = \mathbb{E}[|U_1|] = \frac{\Gamma\left(\frac{p}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{p+1}{2}\right)} = \frac{2}{p-1}\frac{\Gamma\left(\frac{p}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{p-1}{2}\right)}.$$

Moreover, $\kappa_p \sim \sqrt{2/(\pi p)}$ as $p \to \infty$.

Proof. By rotational invariance the expectation is parallel to a. Let Q be orthogonal with $Qa = \|a\|_2 e_1$ and set $V = QU \sim \text{Unif}(S^{p-1})$. Then

$$\mathbb{E}\big[U\,\text{sign}(a^\top U)\big] = Q^\top \mathbb{E}[V\,\text{sign}(V_1)]\,,$$

whose $j \geq 2$ coordinates vanish by symmetry and whose first coordinate is

$$\mathbb{E}[V_1\,\text{sign}(V_1)] = \mathbb{E}[|V_1|] =: \kappa_p.$$

Indeed, the marginal density of V_1 is

$$f_{V_1}(t) = C_p\,(1-t^2)^{\frac{p-3}{2}}, \quad -1 < t < 1, \qquad C_p = \frac{\Gamma\left(\frac{p}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{p-1}{2}\right)}.$$

Thus

$$\mathbb{E}[|V_1|] = 2\int_0^1 t f_{V_1}(t)\,dt = 2C_p\int_0^1 t(1-t^2)^{\frac{p-3}{2}}\,dt = C_p\int_0^1 (1-s)^{\frac{p-3}{2}}\,ds = \frac{2C_p}{p-1}.$$

Equivalently, since $V_1^2 \sim \text{Beta}\left(\frac{1}{2}, \frac{p-1}{2}\right)$,

$$\mathbb{E}[|V_1|] = \mathbb{E}\left[(V_1^2)^{1/2}\right] = \frac{\Gamma\left(\frac{1}{2}+\frac{1}{2}\right)\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{p+1}{2}\right)} = \frac{\Gamma\left(\frac{p}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{p+1}{2}\right)} = \frac{2}{p-1}\frac{\Gamma\left(\frac{p}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{p-1}{2}\right)}.$$

By Stirling's formula, $\kappa_p = \mathbb{E}[|V_1|] \sim \sqrt{2/(\pi p)}$ as $p \to \infty$. Consequently,

$$\mathbb{E}[V \operatorname{sign}(V_1)] = \kappa_p e_1 \quad \text{and} \quad \mathbb{E}\big[U \operatorname{sign}(a^\top U)\big] = Q^\top(\kappa_p e_1) = \kappa_p \frac{a}{\|a\|_2}.$$

□

Proposition 1 (Mean Muon step under radial inputs)**.** *Under the radial model and strictly increasing σ,*

$$\mathbb{E}\big[\Delta w_t^{\text{Muon}}\big] = \mathbb{E}\left[\frac{x_t}{\|x_t\|_2} \operatorname{sign}(a^\top x_t)\right] = \mathbb{E}\big[U_t \operatorname{sign}(a^\top U_t)\big] = \kappa_p \frac{a}{\|a\|_2},$$

with κ_p from Lemma 2. Thus the expectation is exactly *parallel to $a = w_t - w_*$ and its magnitude is $\Theta(1/\sqrt{p})$ for isotropic data.*

Remark 2. *The $1/\sqrt{p}$ factor is a geometric property of the* uniform *sphere: for $U \sim \mathrm{Unif}(S^{p-1})$, a random unit vector is almost orthogonal to any fixed direction, and $\mathbb{E}|U_1| = \Theta(1/\sqrt{p})$. There are three standard ways to avoid it:*

(A) Use anisotropic (elliptical) inputs. *Let $x \sim \mathcal{N}(0, \Sigma)$, where Σ is a highly anisotropic covariance matrix. Then, using concentration of $\|x\|_2$,*

$$\mathbb{E}\left[\frac{x}{\|x\|_2} \operatorname{sign}(a^\top x)\right] = \frac{\sqrt{2/\pi} + o(1)}{\sqrt{\operatorname{tr}(\Sigma)}} \frac{\Sigma a}{\sqrt{a^\top \Sigma a}} \qquad (p \to \infty).$$

If $\operatorname{tr}(\Sigma)$ does not *grow with p (e.g., the data have low intrinsic dimension or a dominant eigenvalue aligned with a), the prefactor is $O(1)$. In particular, if Σ has one large eigenvalue $\lambda_1 \asymp \operatorname{tr}(\Sigma)$ with eigenvector v and a has a nontrivial component along v, the scaling can be constant rather than $1/\sqrt{p}$.*

(B) Use a non-uniform angular law. *If the direction $U = x/\|x\|_2$ is* not *uniform on S^{p-1} (e.g., the angular central Gaussian law induced by $x \sim \mathcal{N}(0, \Sigma)$ with highly anisotropic Σ), the analogue of Lemma 2 yields an expectation parallel to the principal directions, and the leading constant can be $O(1)$ if mass concentrates in a spherical cap around $a/\|a\|_2$.*

Now we provide the proof of linear drift for MUON. The norm in the denominator is

$$\|w_t - w_*\|^2 = \|w_t\|^2 + \|w_*\|^2 - 2w_t^\top w_* = 2(1 - \rho_t) \approx 2(\theta_t^2/2) = \theta_t^2.$$

Thus, $\|w_t - w_*\| \approx \theta_t$. Substituting this in gives:

$$\mathbb{E}[\Delta w_t^{\text{MUON}}] \approx \sqrt{\frac{2}{\pi p}} \frac{w_t - w_*}{\theta_t}$$

Now, we find the projection onto w_* as required by Lemma 1:

$$\mathbb{E}[(\Delta w_t^{\text{MUON}})^\top w_*] \approx \sqrt{\frac{2}{\pi p}} \frac{(w_t - w_*)^\top w_*}{\theta_t} = \sqrt{\frac{2}{\pi p}} \frac{\rho_t - 1}{\theta_t} \approx \sqrt{\frac{2}{\pi p}} \frac{-\theta_t^2/2}{\theta_t} = -\sqrt{\frac{2}{\pi p}} \frac{\theta_t}{2}$$

Similarly, we can compute

$$\mathbb{E}\left[(w_t^\top \Delta w_t^{\text{MUON}})\right] \approx \sqrt{\frac{2}{\pi p}} \frac{(w_t - w_*)^\top w_t}{\theta_t} = \sqrt{\frac{2}{\pi p}} \frac{1 - \rho_t}{\theta_t} \approx \sqrt{\frac{2}{\pi p}} \frac{\theta_t^2/2}{\theta_t} = \sqrt{\frac{2}{\pi p}} \frac{\theta_t}{2}$$

Plugging this into the drift equation gives the final result for MUON:

$$\mathbb{E}[\rho_{t+1} - \rho_t \,|\, \theta_t] \approx \eta \rho_t \left(\sqrt{\frac{2}{\pi p}} \frac{\theta_t}{2}\right) - \eta \left(-\sqrt{\frac{2}{\pi p}} \frac{\theta_t}{2}\right) + O(\eta^2) \approx 2\eta \sqrt{\frac{1}{2\pi p}} \theta_t + O(\eta^2) = \frac{c\eta}{\sqrt{p}} \theta_t + O(\eta^2)$$

This proves the linear drift for MUON. □

3.2 Quadratic Drift of SGD

We aim to prove the following result for the SGD algorithm. The SGD update is given by $w_{t+1} = w_t - \eta \Delta w_t^{\text{SGD}}$, where $\Delta w_t^{\text{SGD}} = r_t \sigma'(u_t) x_t$ and $r_t = \sigma(u_t) - \sigma(v_t)$. Our goal is to compute the drift by analyzing the expectation of the update $\mathbb{E}[\Delta w_t^{\text{SGD}}]$.

First, we state Stein's lemma in vector form.

Lemma 3 (Gaussian Stein; vector form)**.** *For $x \sim N(0, I_p)$ and differentiable $g : \mathbb{R}^p \to \mathbb{R}$ with suitable integrability, $\mathbb{E}[x\, g(x)] = \mathbb{E}[\nabla g(x)]$.*

Now, we apply this to our update for SGD:

$$\mathbb{E}[\Delta w_t^{\text{SGD}}] = \mathbb{E}[r_t \sigma'(u_t) x_t] = \mathbb{E}[\nabla_x \{r_t \sigma'(u_t)\}].$$

Next, we compute the gradient inside the expectation. Let $u = w_t^\top x$ and $v = w_*^\top x$. Using the product and chain rules, we have:

$$\nabla_x \{r \sigma'(u)\} = \nabla_x \{(\sigma(u) - \sigma(v))\sigma'(u)\}.$$

Using this we can derive the following lemma, where we define the constants A, $B(\rho)$, and $C(\rho)$ as follows:

$$A = \mathbb{E}[\sigma'(u)^2], \quad B(\rho) = \mathbb{E}[\sigma'(u)\sigma'(v)], \quad C(\rho) = \mathbb{E}[(\sigma(u) - \sigma(v))\, \sigma''(u)].$$

where ρ is the cosine of the angle between w (or w_t) and w_*.

Lemma 4 (Mean SGD step)**.** *Let $A := \mathbb{E}[\sigma'(u)^2]$ (note: A is constant, independent of ρ); $B(\rho) := \mathbb{E}[\sigma'(u)\sigma'(v)]$; $C(\rho) := \mathbb{E}[(\sigma(u) - \sigma(v))\, \sigma''(u)]$. Then*

$$\mathbb{E}[\Delta w_t^{\text{sgd}}] = \mathbb{E}\big[\nabla_x \{r\, \sigma'(u)\}\big] = \big(A + C(\rho)\big)\, w_t \; - \; B(\rho)\, w_*.$$

Proof. Using $u = w_t^\top x$, $v = w_*^\top x$ and product/chain rules,

$$\nabla_x \{r \sigma'(u)\} = \big(\sigma'(u)^2 + (\sigma(u) - \sigma(v))\sigma''(u)\big) w_t \; - \; \sigma'(v)\sigma'(u)\, w_*.$$

Taking expectations gives the claim; $A = \mathbb{E}[\sigma'(u)^2] = \mathbb{E}[\sigma'(Z)^2]$ since $u \sim N(0, 1)$. □

Now, we can prove the quadratic drift for SGD.

Proposition 2 (Quadratic drift for sgd)**.** *For all angles θ with $\rho_t = \cos\theta$, the expected one-step drift of the alignment for SGD is*

$$\mathbb{E}[\rho_{t+1} - \rho_t] = \eta\, B(\rho_t)\, \big(1 - \rho_t^2\big) + O(\eta^2).$$

In particular, for small θ with $\rho_t = \cos\theta$,

$$\mathbb{E}[\rho_{t+1} - \rho_t] = \eta\, \mathbb{E}[\sigma'(Z)^2]\, \theta^2 + O(\eta\, \theta^4) + O(\eta^2),$$

so $C_{\text{sgd}} = \mathbb{E}[\sigma'(Z)^2] > 0$.

Proof. From Lemma 4,

$$\mathbb{E}[\Delta w_t^{\text{sgd}\,\top} w_*] = (A + C)\rho_t - B, \quad \mathbb{E}[w_t^\top \Delta w_t^{\text{sgd}}] = (A + C) - B\rho_t.$$

Insert these into Lemma 1; the A and C terms cancel exactly, leaving

$$-\eta\big((A + C)\rho_t - B\big) + \eta\rho_t\big((A + C) - B\rho_t\big) = \eta\, B\, (1 - \rho_t^2). \tag{1}$$

Write $u_t = \rho_t z_1 + \sqrt{1-\rho_t^2}\, z_2$, $v_t = z_1$ with $z_1, z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and set $\delta := u_t - v_t = (\rho_t - 1)z_1 + \sqrt{1-\rho_t^2}\, z_2 = -\frac{\theta_t^2}{2} z_1 + \theta_t z_2 + O(\theta_t^3)$.

Using Taylor series expansion,

$$\sigma'(u_t) = \sigma'(v_t) + \sigma''(v_t)\delta + \tfrac{1}{2}\sigma'''(v_t)\delta^2 + O(\delta^3).$$

Then

$$B(\rho_t) = \mathbb{E}[\sigma'(u_t)\sigma'(v_t)] = \underbrace{\mathbb{E}[\sigma'(v_t)^2]}_{=:C_1} + \underbrace{\mathbb{E}[\sigma'(v_t)\sigma''(v_t)\delta]}_{=:T_2} + \underbrace{\tfrac{1}{2}\,\mathbb{E}[\sigma'(v_t)\sigma'''(v_t)\delta^2]}_{=:T_3} + O(\theta_t^3).$$

Now $C_1 = \mathbb{E}[\sigma'(Z)^2]$. For T_2, the $\theta_t z_2$ part vanishes by independence/oddness, while the $-(\theta_t^2/2)z_1$ part remains:

$$T_2 = -\tfrac{\theta_t^2}{2}\,\mathbb{E}[\sigma'(Z)\sigma''(Z)\,Z] = -\tfrac{\theta_t^2}{2}\,\mathbb{E}\big[\sigma''(Z)^2 + \sigma'(Z)\sigma'''(Z)\big]\,,$$

using $\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)]$ with $f(Z) = \sigma'(Z)\sigma''(Z)$. For T_3, to order θ_t^2 we have $\delta^2 = \theta_t^2 z_2^2 + O(\theta_t^3)$, so

$$T_3 = \tfrac{\theta_t^2}{2}\,\mathbb{E}[\sigma'(Z)\sigma'''(Z)] + O(\theta_t^3).$$

Hence the $\sigma'(Z)\sigma'''(Z)$ contributions cancel between T_2 and T_3, yielding

$$B(\rho_t) = \mathbb{E}[\sigma'(Z)^2] \;-\; \tfrac{\theta_t^2}{2}\,\mathbb{E}[\sigma''(Z)^2] \;+\; O(\theta_t^3).$$

Plugging this into Eq. (1) gives the final claim of the proposition.

□

4 Feature learning in multi-index models: Muon vs. SGD

First, we set up the model and notation, then define the optimizers (mini-batch updates).

4.1 Setup/Model/Notation

Ambient dimensions and notation. Let $p \in \mathbb{N}$ be the input dimension and $m \in \mathbb{N}$ the hidden width. Let $r \in \{1, \dots, p-1\}$ be the teacher rank. For $W \in \mathbb{R}^{m \times p}$ write its rows as $W = (w_1, \dots, w_m)^\top$, $w_j \in \mathbb{R}^p$. For nonzero $v \in \mathbb{R}^p$, set $\hat{v} := v/\|v\|_2$. All norms are Euclidean or their operator/Frobenius counterparts. We index iterations by $t \in \mathbb{N}_0$ and mini-batch elements by $b \in \{1, \dots, B\}$.

Data model and mini-batch sampling. At each iteration t, draw an i.i.d. mini-batch $\{(x_{b,t}, y_{b,t})\}_{b=1}^B$, independent of W_t, with

$$x_{b,t} \sim \mathcal{N}(0, I_p) \;\text{ i.i.d.}, \qquad y_{b,t} = y(x_{b,t}).$$

When no confusion arises, we suppress t and write (x_b, y_b).

Teacher (multi-index) model. Let $U = [u_1, \dots, u_r] \in \mathbb{R}^{p \times r}$ have orthonormal columns ($U^\top U = I_r$), and denote the orthogonal projector onto the teacher subspace by $P_\star := UU^\top$. Let $\sigma_\star : \mathbb{R}^r \to \mathbb{R}$ be odd and C^1 (with Lipschitz gradient). The teacher output is

$$y(x) = \sigma_\star\big(U^\top x\big) \in \mathbb{R}.$$

Student network. Let $\phi : \mathbb{R} \to \mathbb{R}$ be odd and C^2, and put $g := \phi'$ (even). For weights $W = (w_1, \dots, w_m)^\top \in \mathbb{R}^{m \times p}$ define

$$f_W(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^m \phi(w_j^\top x).$$

We write $f_t(\cdot) := f_{W_t}(\cdot)$ for the model at iteration t.

Population and mini-batch losses. The population loss is

$$\mathcal{L}_t \;=\; \frac{1}{2}\,\mathbb{E}_x\big[(f_t(x) - y(x))^2\big]\,.$$

The per-iteration mini-batch loss is

$$\widehat{\mathcal{L}}_t \;=\; \frac{1}{2B} \sum_{b=1}^B \big(f_t(x_{b,t}) - y_{b,t}\big)^2.$$

Per-sample residuals and scalar row-gradients. For a mini-batch element $(x_{b,t}, y_{b,t})$, define the residual

$$e_{b,t} := f_t(x_{b,t}) - y_{b,t},$$

and, for each row $j \in \{1, \dots, m\}$, the scalar

$$g_{j,b,t} := \frac{e_{b,t}}{\sqrt{m}}\, \phi'\big(w_{j,t}^\top x_{b,t}\big).$$

Collect $c_{b,t} := (g_{1,b,t}, \dots, g_{m,b,t})^\top \in \mathbb{R}^m$.

Per-sample and batch gradient matrices. Each sample contributes a rank-≤ 1 matrix

$$G_{b,t} := c_{b,t}\, x_{b,t}^\top \in \mathbb{R}^{m\times p}.$$

We use the *batch average* gradient matrix

$$\overline{G}_t := \frac{1}{B}\sum_{b=1}^{B} G_{b,t} \in \mathbb{R}^{m\times p}.$$

Note that the true parameter gradient satisfies

$$\nabla_{w_j}\widehat{\mathcal{L}}_t \;=\; \frac{1}{B}\sum_{b=1}^{B} g_{j,b,t}\, x_{b,t} \quad\Longleftrightarrow\quad (\nabla\widehat{\mathcal{L}}_t)_{j,*} = (\overline{G}_t)_{j,*}.$$

Row mean (for later alignment descriptors). Define

$$\bar{w}_t := \frac{1}{m}\sum_{j=1}^{m} w_{j,t}, \qquad \hat{\bar{w}}_t := \frac{\bar{w}_t}{\|\bar{w}_t\|_2} \;\; (\bar{w}_t \neq 0).$$

4.2 Optimizers (Mini-batch Updates)

Now, we define the mini-batch updates for two optimizers: (i) mini-batch SGD Algorithm 2 and (ii) Muon (true right polar step) Algorithm 1.

Fix a step size $\eta > 0$. All updates are written row-wise as $w_{j,t+1} = \mathrm{Update}(w_{j,t})$.

(i) Mini-batch SGD (no momentum). Using the mini-batch average gradient,

$$w_{j,t+1} \;=\; w_{j,t} \;-\; \eta\,\frac{1}{B}\sum_{b=1}^{B} g_{j,b,t}\, x_{b,t}, \qquad j = 1,\dots,m.$$

(ii) Muon (true right polar step; unprojected by default). Compute the *right polar factor* of the batch gradient matrix $\overline{G}_t$:

$$Q_t \;:=\; \mathrm{Polar}_R(\overline{G}_t) \;=\; \overline{G}_t\,\big(\overline{G}_t^\top \overline{G}_t\big)^{-1/2},$$

where $(\cdot)^{-1/2}$ denotes the positive-semidefinite inverse square root on $\mathrm{im}(\overline{G}_t^\top)$ (Moore–Penrose inverse on the orthogonal complement if needed). Update each row by

$$w_{j,t+1} \;=\; w_{j,t} \;-\; \eta\,(Q_t)_{j,*}, \qquad j = 1,\dots,m.$$

Remark. With $B = 1$, $\overline{G}_t = c_{1,t}x_{1,t}^\top$ and the right polar step moves along the *input direction* $x_{1,t}/\|x_{1,t}\|_2$ with row-dependent left singular weights absorbed by the polar factor.

(iii) Projected (row-normalized) variant (optional). After either SGD or Muon step, one may renormalize rows to unit norm:

$$w_{j,t+1} \;\leftarrow\; \frac{w_{j,t+1}}{\|w_{j,t+1}\|_2}, \qquad j = 1,\dots,m.$$

If used, we will explicitly refer to the *projected* (row-normalized) version.

(iv) Optional left-polar variant (for reference). For completeness, the left polar factor is

$$\widetilde{Q}_t \;:=\; \big(\overline{G}_t\overline{G}_t^\top\big)^{-1/2}\,\overline{G}_t \in \mathbb{R}^{m\times p},$$

leading to the update $w_{j,t+1} = w_{j,t} - \eta\,(\widetilde{Q}_t)_{j,*}$. Unless stated otherwise, our "Muon" refers to the *right* polar in (ii).

Right polar factor: notation and rank-zero convention. For any batch gradient matrix $\overline{G}_t$, take an SVD $\overline{G}_t = \mathcal{U}_t\Sigma_t\mathcal{V}_t^\top$ and define the right polar factor by $Q_t := \mathcal{U}_t\mathcal{V}_t^\top$ (a partial isometry). If $\overline{G}_t = 0$, set $Q_t := 0$. (We reserve U exclusively for the teacher subspace; $\mathcal{U}_t, \mathcal{V}_t$ refer to batch SVD factors.)

4.3 Assumptions (finite horizon, mini-batch feature learning)

In this section we list the assumptions used in our analysis.

Fix a finite horizon $t = 0, 1, \ldots, T$ with T independent of (p, m). All probabilities are with respect to the data draws; $O_{\mathbb{P}}(\cdot)$ denotes bounds in probability. We write "with high probability" to mean probability at least $1 - Cp^{-c}$ for some $c, C > 0$.

Assumption 1 (High-dimensional scaling)**.** *As $p \to \infty$, $m/p \to \gamma \in (0, \infty)$ and the teacher rank r is fixed.*

Assumption 2 (Fresh Gaussian mini-batches, independent of W_t)**.** *Conditionally on W_t, the batch $\{(x_{b,t}, y_{b,t})\}_{b=1}^B$ is i.i.d., with $x_{b,t} \sim \mathcal{N}(0, I_p)$ and $y_{b,t} = y(x_{b,t})$, independent of W_t.*

Assumption 3 (Teacher (multi-index) regularity)**.** *$U \in \mathbb{R}^{p\times r}$ has orthonormal columns and $y(x) = \sigma_\star(U^\top x)$ with $\sigma_\star : \mathbb{R}^r \to \mathbb{R}$ odd, C^1, and with Lipschitz gradient.*

Assumption 4 (Student activation regularity)**.** *ϕ is odd and C^2 with $g := \phi'$ even and $\|g\|_\infty < \infty$, $\|g'\|_\infty < \infty$. For the Gaussian correlation kernel $\kappa(\rho) := \mathbb{E}[g(Z_1)g(Z_2)]$ with $\mathrm{Corr}(Z_1, Z_2) = \rho$ one has by Price's theorem*

$$\kappa'(\rho) = \mathbb{E}[g'(Z_1)g'(Z_2)] \qquad \Rightarrow \qquad |\kappa(\rho) - \kappa(\rho')| \le \|g'\|_\infty^2\,|\rho - \rho'|.$$

Assumption 5 (Initialization: near-isotropic row ensemble)**.** *Let $\bar{w}_0 = \frac{1}{m}\sum_j w_{j,0}$, $\delta_{j,0} = w_{j,0} - \bar{w}_0$, and $S_0 = \frac{1}{m}\sum_{j=1}^m \delta_{j,0}\delta_{j,0}^\top$. Set $\alpha_0 = \mathrm{tr}(S_0)/p$ and $S_0' = S_0 - \alpha_0 I_p$. Assume $\|S_0'\|_F = O_{\mathbb{P}}(p^{-1/2})$. (This holds, e.g., for i.i.d. rotation-invariant rows with covariance $\alpha_0 I_p/p$ and $m \asymp p$.)*

Assumption 6 (Small steps; finite horizon)**.** *There exists $\eta_0 > 0$ such that the step size satisfies $\eta \le \eta_0 p^{-1/2}$. We consider $t \le T$ with T fixed in p.*

Assumption 7 (Mean–isotropy persistence (trace-centered))**.** *Let $\bar{w}_t = \frac{1}{m}\sum_j w_{j,t}$, $\delta_{j,t} = w_{j,t} - \bar{w}_t$, $S_t = \frac{1}{m}\sum_j \delta_{j,t}\delta_{j,t}^\top$, $\alpha_t = \mathrm{tr}(S_t)/p$, $S_t' := S_t - \alpha_t I_p$. Then, with high probability,*

$$\sup_{0\le t\le T} \|S_t'\|_F \;=\; O\big(p^{-1/2}\big).$$

(See Theorem 2.)

Assumption 8 (Row-norm control (unprojected case))**.** *There exist constants $0 < c_- < c_+ < \infty$ such that with high probability,*

$$c_- \;\le\; \|w_{j,t}\|_2 \;\le\; c_+, \qquad \forall j \le m,\ \forall t \le T.$$

(Automatic in the projected/row-normalized variant; in the unprojected case this follows from small steps plus bounded gradients; see the proof after Lemma 5.)

Assumption 9 (Mini-batch size)**.** *Either B is fixed (independent of p), or $B = o(p^{1/2})$.*

Assumption 10 (Nondegenerate Hermite coupling)**.** *With*

$$C^{(r)}_{\sigma_\star,\phi} \;:=\; \mathbb{E}_{Z\sim\mathcal{N}(0,I_r)}\big[\partial_{z_1}\sigma_\star(Z)\,\phi'(Z_1)\big],$$

assume $C^{(r)}_{\sigma_\star,\phi} \neq 0$ *for linear small-angle drift. (If it vanishes, the linear term cancels and the first nonzero drift enters at the next Hermite order—our later statements will note the cubic fallback.)*

Assumption 11 (Not at stationarity on the horizon)**.** *With high probability,* $\overline{G}_t \neq 0$ *for all* $t \le T$ *(equivalently,* $\mathbb{E}[e_{b,t}^2] > 0$*), so the right polar factor is well-defined on the batch span.*

4.4 Mean-isotropy on fixed horizons

Here, we show that mean–isotropy Assumption 7 holds on fixed horizons under the other assumptions.

Standing notation. Let $\bar{w}_t = \frac{1}{m}\sum_{j=1}^m w_{j,t}$, $\delta_{j,t} = w_{j,t} - \bar{w}_t$, and $\Delta_t \in \mathbb{R}^{m\times p}$ collect rows $\delta_{j,t}^\top$. Define the centered row second moment

$$S_t \;:=\; \frac{1}{m}\,\Delta_t^\top\Delta_t \;=\; \frac{1}{m}\sum_{j=1}^m \delta_{j,t}\delta_{j,t}^\top.$$

Lemma 5 (Residual moment bound)**.** *Let* $x \sim \mathcal{N}(0, I_p)$*, and suppose* $\sup_{t\le T}\mathbb{E}[|e_t(x)|^4] \le C_4 < \infty$*, where* $e_t(x) = f_t(x) - y(x)$*. Then, uniformly on the fixed horizon* $t \le T$*,*

$$\mathbb{E}\big[e_t(x)^2\,\|x\|_2^2\big] \;\le\; C\,p,$$

for a constant C *depending only on* C_4*.*

Proof. Fix $t \le T$. By Cauchy–Schwarz,

$$\mathbb{E}\big[e_t(x)^2\,\|x\|_2^2\big] \;\le\; \Big(\mathbb{E}[e_t(x)^4]\Big)^{1/2}\Big(\mathbb{E}[\|x\|_2^4]\Big)^{1/2}.$$

By assumption, $\big(\mathbb{E}[e_t(x)^4]\big)^{1/2} \le C_4^{1/2}$. Since $x \sim \mathcal{N}(0, I_p)$, $\|x\|_2^2 \sim \chi_p^2$ and

$$\mathbb{E}[\|x\|_2^4] = \mathrm{Var}(\chi_p^2) + \big(\mathbb{E}\chi_p^2\big)^2 = 2p + p^2.$$

Hence $\mathbb{E}[e_t(x)^2\|x\|_2^2] \le C_4^{1/2}(p^2+2p)^{1/2} \le C\,p$, for a constant C depending only on C_4. This is uniform over $t \le T$, and conditioning on W_t yields the same bound for $\mathbb{E}[\cdot \mid W_t]$ by the tower property. □

Lemma 6 (Batch step energy: SGD (variance-only, centered))**.** *For SGD, define the centered (fluctuation) step*

$$\widetilde{h}_{j,t} := h_{j,t} - \mathbb{E}[h_{j,t} \mid W_t] = -\eta\,\frac{1}{B}\sum_{b=1}^B \big(g_{j,b,t}x_{b,t} - \mu_j\big), \qquad \mu_j := \mathbb{E}[g_{j,1,t}x_{1,t} \mid W_t],$$

and let $\widetilde{H}_t$ *be the matrix with rows* $\widetilde{h}_{j,t}^\top$*, and* $\widetilde{\bar{h}}_t := \bar{h}_t - \mathbb{E}[\bar{h}_t \mid W_t]$*. Under Assumption 2-Assumption 4 and Lemma 5,*

$$\mathbb{E}\big[\|\widetilde{H}_t\|_F^2 \,\big|\, W_t\big] \;\le\; \eta^2\,\frac{C\,p}{B}, \qquad \mathbb{E}\big[\sqrt{m}\,\|\widetilde{\bar{h}}_t\|_2 \,\big|\, W_t\big] \;\le\; \eta\,C\sqrt{\frac{p}{B}}.$$

(High-probability versions follow by Gaussian/vector Bernstein concentration.)

Remark. *Without centering, an additional bias term* $\eta^2\,\|\nabla\mathcal{L}_t\|_F^2$ *appears; see Lemma 7.*

Proof. Condition on W_t and suppress t. Set $\xi_{j,b} := g_{j,b}x_b \in \mathbb{R}^p$ and $\mu_j = \mathbb{E}[\xi_{j,1} \mid W_t]$. Then $\widetilde{h}_j = -\eta \frac{1}{B}\sum_{b=1}^{B}(\xi_{j,b} - \mu_j)$ and, by independence across b,

$$\mathbb{E}\Big[\Big\|\frac{1}{B}\sum_{b=1}^{B}(\xi_{j,b}-\mu_j)\Big\|_2^2 \,\Big|\, W_t\Big] = \frac{1}{B}\operatorname{Var}(\xi_{j,1} \mid W_t) \;\le\; \frac{1}{B}\,\mathbb{E}\big[\|\xi_{j,1}\|_2^2 \mid W_t\big].$$

Summing in j and using $|\phi'| \le M$,

$$\mathbb{E}\big[\|\widetilde{H}_t\|_F^2 \mid W_t\big] \le \frac{\eta^2}{B}\sum_{j=1}^{m}\mathbb{E}\big[\|\xi_{j,1}\|_2^2 \mid W_t\big] \le \frac{\eta^2 M^2}{B}\,\mathbb{E}\big[e_1^2\,\|x_1\|_2^2 \mid W_t\big] \le \eta^2\,\frac{C\,p}{B},$$

by Lemma 5. For the centered mean step,

$$\widetilde{\bar{h}}_t = -\,\eta\,\frac{1}{B}\sum_{b=1}^{B}\big(\bar{g}_b\,x_b - \mathbb{E}[\bar{g}_1\,x_1 \mid W_t]\big), \qquad \bar{g}_b := \frac{1}{m}\sum_{j=1}^{m} g_{j,b}, \quad |\bar{g}_b| \le \frac{M}{\sqrt{m}}|e_b|,$$

so independence and the same bound yield

$$\mathbb{E}\big[\|\widetilde{\bar{h}}_t\|_2^2 \mid W_t\big] \le \frac{\eta^2}{B}\,\mathbb{E}\big[\bar{g}_1^2\,\|x_1\|_2^2 \mid W_t\big] \le \frac{\eta^2 M^2}{mB}\,\mathbb{E}\big[e_1^2\,\|x_1\|_2^2 \mid W_t\big] \le \frac{\eta^2 C\,p}{mB},$$

and Jensen gives $\mathbb{E}[\sqrt{m}\,\|\widetilde{\bar{h}}_t\|_2 \mid W_t] \le \eta\,C\,\sqrt{p/B}$. □

Lemma 7 (Batch step energy: SGD (bias–variance form))**.** *Fix t and condition on W_t. For SGD,*

$$h_{j,t} \;=\; -\eta\,\frac{1}{B}\sum_{b=1}^{B} g_{j,b,t}\,x_{b,t}, \qquad g_{j,b,t} := \frac{e_{b,t}}{\sqrt{m}}\,\phi'(w_{j,t}^{\top}x_{b,t}).$$

Let H_t be the matrix with rows $h_{j,t}^{\top}$, and write the population gradient $\mu_j := \mathbb{E}[g_{j,1,t}x_{1,t} \mid W_t] = \nabla_{w_j}\mathcal{L}_t$ so that $\|\nabla\mathcal{L}_t\|_F^2 = \sum_j \|\mu_j\|_2^2$. Under Assumption 2-Assumption 4 and Lemma 5,

$$\mathbb{E}\big[\|H_t\|_F^2 \mid W_t\big] \;\le\; \eta^2\,\|\nabla\mathcal{L}_t\|_F^2 \;+\; \eta^2\,\frac{C\,p}{B},$$

$$\mathbb{E}\big[\sqrt{m}\,\|\bar{h}_t\|_2 \mid W_t\big] \;\le\; \eta\,C\,\sqrt{\frac{p}{B}}\,.$$

(High-probability versions follow by Gaussian/vector Bernstein concentration.)

Proof. Let $\xi_{j,b} := g_{j,b,t}x_{b,t} \in \mathbb{R}^p$ and $\mu_j = \mathbb{E}[\xi_{j,1} \mid W_t]$. Independence across b gives the bias–variance decomposition

$$\mathbb{E}\Big[\Big\|\frac{1}{B}\sum_{b=1}^{B}\xi_{j,b}\Big\|_2^2 \,\Big|\, W_t\Big] = \frac{1}{B}\,\mathbb{E}\big[\|\xi_{j,1}\|_2^2 \mid W_t\big] \;+\; \Big(1-\frac{1}{B}\Big)\,\|\mu_j\|_2^2.$$

Summing in j and multiplying by η^2,

$$\mathbb{E}\big[\|H_t\|_F^2 \mid W_t\big] = \eta^2\,\|\nabla\mathcal{L}_t\|_F^2 + \frac{\eta^2}{B}\sum_{j=1}^{m}\mathbb{E}\big[\|\xi_{j,1}\|_2^2 \mid W_t\big] - \frac{\eta^2}{B}\,\|\nabla\mathcal{L}_t\|_F^2 \;\le\; \eta^2\,\|\nabla\mathcal{L}_t\|_F^2 + \frac{\eta^2}{B}\sum_{j}\mathbb{E}\|\xi_{j,1}\|^2.$$

Since $|\phi'| \le \|\phi'\|_\infty =: M$,

$$\|\xi_{j,1}\|_2^2 = \frac{e_{1,t}^2}{m}\,\phi'(w_{j,t}^{\top}x_{1,t})^2\,\|x_{1,t}\|_2^2 \le \frac{M^2}{m}\,e_{1,t}^2\,\|x_{1,t}\|_2^2,$$

hence $\sum_{j=1}^{m}\mathbb{E}\|\xi_{j,1}\|_2^2 \le M^2\,\mathbb{E}[e_{1,t}^2\|x_{1,t}\|_2^2] \le C\,p$ by Lemma 5. This yields the first display.

For $\bar{h}_t = \frac{1}{m}\sum_j h_{j,t} = -\eta\,\frac{1}{B}\sum_b \bar{g}_{b,t}\,x_{b,t}$ with $\bar{g}_{b,t} = \frac{1}{m}\sum_j g_{j,b,t}$, we have $|\bar{g}_{b,t}| \le \frac{M}{\sqrt{m}}\,|e_{b,t}|$, whence

$$\mathbb{E}\big[\|\bar{h}_t\|_2^2 \mid W_t\big] \le \eta^2\,\frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\big[\bar{g}_{b,t}^2\,\|x_{b,t}\|_2^2 \mid W_t\big] \le \eta^2\,\frac{M^2}{mB}\,\mathbb{E}\big[e_{1,t}^2\,\|x_{1,t}\|_2^2 \mid W_t\big] \le \eta^2\,\frac{C\,p}{mB}.$$

Apply Jensen to get $\mathbb{E}[\sqrt{m}\,\|\bar{h}_t\||W_t] \le \eta\,C\,\sqrt{p/B}$. □

Lemma 8 (Batch step energy: Muon (right polar)). *Let $\overline{G}_t = \frac{1}{B}\sum_{b=1}^{B} c_{b,t}x_{b,t}^\top$ and take an SVD $\overline{G}_t = \mathcal{U}_t\Sigma_t\mathcal{V}_t^\top$. Define the right polar factor $Q_t := \mathcal{U}_t\mathcal{V}_t^\top$ (partial isometry), and set $Q_t := 0$ if $\overline{G}_t = 0$. For the Muon step $h_{j,t} = -\eta(Q_t)_{j,*}$ and H_t the matrix of rows $h_{j,t}^\top$,*

$$\|H_t\|_F \;=\; \eta\|Q_t\|_F \;=\; \eta\sqrt{\operatorname{rank}(\overline{G}_t)} \;\le\; \eta\sqrt{B}, \qquad \sqrt{m}\,\|\bar{h}_t\|_2 \;\le\; \eta.$$

(The second inequality uses $\|Q_t^\top\mathbf{1}\|_2 \le \|Q_t\|_{\mathrm{op}}\|\mathbf{1}\|_2 = \sqrt{m}$.)

Proof. By definition, $\|H_t\|_F = \eta\|Q_t\|_F = \eta\sqrt{\operatorname{rank}(\overline{G}_t)} \le \eta\sqrt{B}$. For the mean step, $\bar{h}_t = -(\eta/m)Q_t^\top\mathbf{1}$, so $\sqrt{m}\,\|\bar{h}_t\|_2 \le (\eta/\sqrt{m})\|Q_t^\top\mathbf{1}\|_2 \le \eta$. □

Lemma 9 (Population gradient energy). *Under Assumption 2, Assumption 4 and Lemma 5,*

$$\|\nabla\mathcal{L}_t\|_F^2 = \sum_{j=1}^{m}\Big\|\mathbb{E}\big[g_{j,1,t}x_{1,t} \mid W_t\big]\Big\|_2^2 \;\le\; \sum_{j=1}^{m}\mathbb{E}\big[g_{j,1,t}^2\,\|x_{1,t}\|_2^2 \mid W_t\big].$$

Since $g_{j,1,t} = \frac{e_{1,t}}{\sqrt{m}}\phi'(w_{j,t}^\top x_{1,t})$ and $|\phi'| \le M$,

$$\mathbb{E}\big[g_{j,1,t}^2\,\|x_{1,t}\|_2^2 \mid W_t\big] \;\le\; \frac{M^2}{m}\,\mathbb{E}\big[e_{1,t}^2\,\|x_{1,t}\|_2^2 \mid W_t\big].$$

Summing over j cancels the factor $1/m$:

$$\|\nabla\mathcal{L}_t\|_F^2 \;\le\; M^2\,\mathbb{E}\big[e_{1,t}^2\,\|x_{1,t}\|_2^2 \mid W_t\big] \;\le\; C\,p,$$

by Lemma 5.

Theorem 2 (Mean–isotropy persists on fixed horizons, trace-centered). *Let Δ_t collect rows $\delta_{j,t}^\top$, and define*

$$S_t \;=\; \frac{1}{m}\,\Delta_t^\top\Delta_t, \qquad \alpha_t := \frac{\operatorname{tr} S_t}{p}, \qquad S_t' := S_t - \alpha_t I_p.$$

Assume: (i) initialization as in Assumption 5 with $S_0 = \alpha_0 I_p + E_{p,0}$, $\|E_{p,0}\|_F = O_{\mathbb{P}}(p^{-1/2})$; (ii) scaling $m/p \to \gamma \in (0,\infty)$ as in Assumption 1; (iii) small steps $\eta \le C_0 p^{-1}$ and fixed horizon T; (iv) Gaussian mini-batches with B fixed; (v) activation regularity as in Assumption 4; (vi) row-norm control as in (Assumption 8) (projected: $\|w_{j,t}\|_2 = 1$; unprojected: $c_- \le \|w_{j,t}\| \le c_+$). Then, with high probability (at least $1 - Cp^{-c}$),

$$\max_{0\le t\le T}\;\|S_t'\|_F \;\le\; \frac{C'}{p}.$$

Proof. Write $h_{j,t} := w_{j,t+1} - w_{j,t}$, let H_t be the $m\times p$ matrix whose jth row is $h_{j,t}^\top$, and set $\bar{h}_t := \frac{1}{m}\sum_{j=1}^{m} h_{j,t}$. Since $\delta_{j,t+1} = \delta_{j,t} + h_{j,t} - \bar{h}_t$, collecting rows into $\Delta_t \in \mathbb{R}^{m\times p}$ gives

$$S_{t+1} = \frac{1}{m}\,(\Delta_t + H_t - \mathbf{1}\bar{h}_t^\top)^\top(\Delta_t + H_t - \mathbf{1}\bar{h}_t^\top).$$

In the bounds below, $\lesssim$ hides absolute constants independent of (p, m, B, η) and, in the unprojected case, independent of p, m, B, η but possibly depending on c_+ from Assumption 8.

Now, subtracting $S_t = \frac{1}{m}\Delta_t^\top \Delta_t$ across two steps yields

$$S_{t+1} - S_t = \frac{1}{m}\Big(\Delta_t^\top \widetilde{H}_t + \widetilde{H}_t^\top \Delta_t\Big) + \frac{1}{m}\,\widetilde{H}_t^\top \widetilde{H}_t, \qquad \widetilde{H}_t := H_t - \mathbf{1}\bar{h}_t^\top. \tag{2}$$

Using $\|AB\|_F \le \|A\|_{\mathrm{op}}\|B\|_F$ and $\|\widetilde{H}_t\|_F \le \|H_t\|_F + \sqrt{m}\,\|\bar{h}_t\|_2$, we obtain

$$\|S_{t+1} - S_t\|_F \le \frac{2}{m}\,\|\Delta_t\|_{\mathrm{op}}\left(\|H_t\|_F + \sqrt{m}\,\|\bar{h}_t\|_2\right) + \frac{1}{m}\left(\|H_t\|_F + \sqrt{m}\,\|\bar{h}_t\|_2\right)^2. \tag{3}$$

In the *projected case*, $\|w_{j,t}\|_2 = 1$ and $\|\bar{w}_t\|_2 \le 1$, so $\|\delta_{j,t}\|_2 \le 2$ and $\|\Delta_t\|_{\mathrm{op}} \le 2\sqrt{m}$ deterministically. In the *unprojected case*, by Assumption 8 $\|w_{j,t}\| \le c_+$ implies $\|\Delta_t\|_{\mathrm{op}} \le 2c_+\sqrt{m}$ with high probability.

For SGD, decompose $H_t = \mathbb{E}[H_t \mid W_t] + \widetilde{H}_t^{(\mathrm{sgd})}$ and $\bar{h}_t = \mathbb{E}[\bar{h}_t \mid W_t] + \widetilde{\bar{h}}_t^{(\mathrm{sgd})}$. By Lemma 7 and its centered variant (Lemma 6), with high probability (at least $1 - Cp^{-c}$),

$$\|\widetilde{H}_t^{(\mathrm{sgd})}\|_F \lesssim \eta\sqrt{\tfrac{p}{B}}, \qquad \sqrt{m}\,\|\widetilde{\bar{h}}_t^{(\mathrm{sgd})}\|_2 \lesssim \eta\sqrt{\tfrac{p}{B}}, \qquad \|\mathbb{E}[H_t \mid W_t]\|_F = \eta\,\|\nabla\mathcal{L}_t\|_F.$$

where the last bound follows from the definition of the population gradient: since the mini-batch gradient is conditionally unbiased, $\mathbb{E}[H_t \mid W_t] = -\eta\,\nabla\mathcal{L}_t$.

Therefore $\|H_t\|_F + \sqrt{m}\,\|\bar{h}_t\|_2 \lesssim \eta\sqrt{\frac{p}{B}} + \eta\,\|\nabla\mathcal{L}_t\|_F$. Invoking Assumption 1 ($m/p \to \gamma$) gives

$$\|S_{t+1} - S_t\|_F \;\lesssim\; \eta\,B^{-1/2} + \eta^2/B + \eta\,\frac{\|\nabla\mathcal{L}_t\|_F}{\sqrt{p}} + \eta^2\,\frac{\|\nabla\mathcal{L}_t\|_F^2}{p}. \tag{4}$$

By Lemma 9, which bounds $\|\nabla\mathcal{L}_t\|_F/\sqrt{p} = O(1)$, the last two (gradient–energy) terms gives the following.

$$\|S_{t+1} - S_t\|_F \;\lesssim\; \eta\,B^{-1/2} \;+\; \eta^2/B \;+\; \eta \;+\; \eta^2, \tag{5}$$

and we will carry all four terms through the trace-centering step (no pointwise domination is asserted).

For Muon, Lemma 8 gives deterministically $\|H_t\|_F \le \eta\sqrt{B}$ and $\sqrt{m}\,\|\bar{h}_t\|_2 \le \eta$; plugging these into the Eq. (3) with $\|\Delta_t\|_{\mathrm{op}} \lesssim \sqrt{m}$ and $m \asymp p$ yields

$$\|S_{t+1} - S_t\|_F \;\lesssim\; \eta\,\sqrt{\tfrac{B}{p}} + \frac{\eta^2 B}{p}. \tag{6}$$

Now, taking the trace in Eq. (2) and using $\mathrm{tr}(A^\top B) = \langle A, B\rangle_F$ and $\mathrm{tr}(\widetilde{H}_t^\top \widetilde{H}_t) = \|\widetilde{H}_t\|_F^2$, we get

$$\alpha_{t+1} - \alpha_t = \frac{1}{p}\,\mathrm{tr}(S_{t+1} - S_t) = \frac{2}{mp}\,\langle \Delta_t, \widetilde{H}_t\rangle_F + \frac{1}{mp}\,\|\widetilde{H}_t\|_F^2.$$

By Assumption 1, Assumption 8: in the projected case we have $\|\delta_{j,t}\|_2 \le 2$, in the unprojected case we note $\|\delta_{j,t}\|_2 \le 2c_+$, so $\mathrm{tr}\,S_t = \frac{1}{m}\sum_j \|\delta_{j,t}\|_2^2 \le 4c_+^2$ and thus $\alpha_t = \mathrm{tr}\,S_t/p = O(p^{-1})$ and $\|\Delta_t\|_F^2 = m\ \mathrm{tr}\,S_t = O(m)$.

Hence

$$|\alpha_{t+1} - \alpha_t| \;\le\; \frac{2}{mp}\,\|\Delta_t\|_F\,\|\widetilde{H}_t\|_F + \frac{1}{mp}\,\|\widetilde{H}_t\|_F^2 \;\lesssim\; \frac{\|\widetilde{H}_t\|_F}{p\sqrt{m}} + \frac{\|\widetilde{H}_t\|_F^2}{mp}.$$

For SGD, using Lemma 6, $\|\widetilde{H}_t\|_F \lesssim C\eta\sqrt{p/B}$ yields

$$|\alpha_{t+1} - \alpha_t| \;\lesssim\; \frac{\eta}{\sqrt{pB}} + \frac{\eta^2}{pB}, \qquad \sqrt{p}\,|\alpha_{t+1} - \alpha_t| \;\lesssim\; \frac{\eta}{\sqrt{B}} + \frac{\eta^2}{\sqrt{p}\,B}. \tag{7}$$

For Muon, $\|\widetilde{H}_t\|_F \le \|H_t\|_F + \sqrt{m}\,\|\bar{h}_t\|_2 \lesssim \eta\sqrt{B} + \eta$ gives

$$|\alpha_{t+1} - \alpha_t| \;\lesssim\; \frac{\eta\sqrt{B}}{p^{3/2}} + \frac{\eta^2 B}{p^2}, \qquad \sqrt{p}\,|\alpha_{t+1} - \alpha_t| \;\lesssim\; \frac{\eta\sqrt{B}}{p} + \frac{\eta^2 B}{p^{3/2}}. \tag{8}$$

Since $S'_t = S_t - \alpha_t I_p$,

$$\|S'_{t+1} - S'_t\|_F \;\le\; \|S_{t+1} - S_t\|_F + \sqrt{p}\,|\alpha_{t+1} - \alpha_t|.$$

Combining Eq. (5) with Eq. (7) and simplifying gives

$$\|S'_{t+1} - S'_t\|_F \;\lesssim\; \Big(\frac{\eta}{\sqrt{B}} + \frac{\eta^2}{B}\Big) \;+\; (\eta + \eta^2) \;+\; \Big(\frac{\eta}{\sqrt{B}} + \frac{\eta^2}{\sqrt{p}\,B}\Big) \;\lesssim\; \eta\Big(1 + \tfrac{1}{\sqrt{B}}\Big) \;+\; O(\eta^2) \qquad \text{(SGD)}.$$

Hence, with B fixed and $\eta \le C_0 p^{-1}$, $\|S'_{t+1} - S'_t\|_F = O(p^{-1/2})$. For Muon, combining Eq. (6) with Eq. (8) yields

$$\|S'_{t+1} - S'_t\|_F \;\lesssim\; \eta\sqrt{\tfrac{B}{p}} + \frac{\eta^2 B}{p} \;+\; \frac{\eta\sqrt{B}}{p} + \frac{\eta^2 B}{p^{3/2}} \;\lesssim\; \eta\sqrt{\tfrac{B}{p}} \;+\; \frac{\eta^2 B}{p} \qquad \text{(Muon)}.$$

With B fixed and $\eta \le C_0 p^{-1}$ this is $O(p^{-1})$.

Now, for fixed B in both cases each per-step trace-centered increment is $O(p^{-1/2})$. To conclude a uniform-in-time bound on $\max_{0\le t\le T} \|S'_t\|_F$, we use (i) anchor at $t = 0$ via Assumption 5, and (ii) sum the per-step bounds over $t = 0, \dots, T-1$ (triangle inequality) and apply a union bound to keep the high-probability event across the fixed horizon. Thus,

$$\|S'_t\|_F \;\le\; \|S'_0\|_F + \sum_{s=0}^{t-1} \|S'_{s+1} - S'_s\|_F \;\le\; C_0/p + T \cdot O(1/p) \;=\; O(1/p),$$

whence

$$\max_{0\le t\le T} \|S'_t\|_F \;\le\; \frac{C'}{p}$$

with high probability. □

Here we show how to leverage mean–isotropy plus row-norm control to obtain the AGOP condition that is used in the later sections.

Lemma 10 (AGOP from mean–isotropy + row-norm control)**.** *Let $W_t = (w_{1,t}, \dots, w_{m,t})$, $\bar{w}_t = \frac{1}{m}\sum_j w_{j,t}$, and $S_t = \frac{1}{m}\sum_j (w_{j,t} - \bar{w}_t)(w_{j,t} - \bar{w}_t)^\top$. ¡¡¡¡¡¡¡ HEAD Assume Assumption 1 ($m/p \to \gamma \in (0,\infty)$), Assumption 8 row-norm control $c_- \le \|w_{j,t}\| \le c_+$, and the ======= Assume $m/p \to \gamma \in (0,\infty)$ (Assumption 1), row-norm control $c_- \le \|w_{j,t}\| \le c_+$ (Assumption 8), and the ¿¿¿¿¿¿¿ 8f83143 (message) trace-centered mean-isotropy bound from Theorem 2:*

$$\|S'_t\|_F \;=\; \Big\|S_t - \frac{\operatorname{tr} S_t}{p} I_p\Big\|_F \;\le\; \frac{C}{p} \quad \textit{uniformly for } t \le T.$$

Suppose in addition that along the fixed horizon T,

$$\|\bar{w}_t\|^2 \;=\; O\Big(\frac{1}{m}\Big) \;=\; O\Big(\frac{1}{p}\Big).$$

Then the unit–normalized AGOP condition holds (uniformly for $t \le T$):

$$\Big\|\frac{1}{m}\sum_{j=1}^{m} \widehat{w}_{j,t}\widehat{w}_{j,t}^\top - \frac{I_p}{p}\Big\|_{\mathrm{op}} \;\le\; \frac{C_{\mathrm{ag}}}{p}, \qquad \widehat{w}_{j,t} := \frac{w_{j,t}}{\|w_{j,t}\|},$$

for a constant C_{ag} depending only on (c_-, c_+) and the constant in Theorem 2.

Proof. Write the unnormalized second moment as

$$\frac{1}{m}\sum_{j=1}^{m} w_{j,t} w_{j,t}^\top \;=\; S_t \;+\; \bar{w}_t \bar{w}_t^\top.$$

Now, decompose and trace-center:

$$\frac{1}{m}\sum_{j=1}^{m} w_{j,t}w_{j,t}^\top - \frac{\operatorname{tr} S_t + \|\bar{w}_t\|^2}{p}\, I_p \;=\; S_t' \;+\; \Big(\bar{w}_t\bar{w}_t^\top - \frac{\|\bar{w}_t\|^2}{p} I_p\Big).$$

Now unit normalization using row–norm control gives, noting that $c_-^2 I \preceq \|w_{j,t}\|^2 I \preceq c_+^2 I$,

$$\frac{1}{c_+^2}\,\frac{1}{m}\sum_j w_{j,t}w_{j,t}^\top \;\preceq\; \frac{1}{m}\sum_j \widehat{w}_{j,t}\widehat{w}_{j,t}^\top \;\preceq\; \frac{1}{c_-^2}\,\frac{1}{m}\sum_j w_{j,t}w_{j,t}^\top.$$

Let $M := \frac{1}{m}\sum_j w_{j,t}w_{j,t}^\top$ and $\widehat{M} := \frac{1}{m}\sum_j \widehat{w}_{j,t}\widehat{w}_{j,t}^\top$. From the sandwich,

$$\frac{1}{c_+^2} M \;\preceq\; \widehat{M} \;\preceq\; \frac{1}{c_-^2} M.$$

For any $\beta \in \mathbb{R}$, if $A \preceq B \preceq C$ then

$$\|B - \beta I\|_{\mathrm{op}} \;\le\; \max\{\|A - \beta I\|_{\mathrm{op}},\; \|C - \beta I\|_{\mathrm{op}}\},$$

since $\lambda_{\min}(A) \le \lambda_{\min}(B) \le \lambda_{\min}(C)$ and $\lambda_{\max}(A) \le \lambda_{\max}(B) \le \lambda_{\max}(C)$, so the spectrum of B lies between those of A and C. Applying this with $\beta = \frac{1}{p}$ and $(A, B, C) = \big(\frac{1}{c_+^2}M,\; \widehat{M},\; \frac{1}{c_-^2}M\big)$ gives

$$\big\|\widehat{M} - \tfrac{1}{p} I\big\|_{\mathrm{op}} \;\le\; \max\Big\{\Big\|\tfrac{1}{c_+^2}M - \tfrac{1}{p}I\Big\|_{\mathrm{op}},\; \Big\|\tfrac{1}{c_-^2}M - \tfrac{1}{p}I\Big\|_{\mathrm{op}}\Big\}.$$

For any $a > 0$ and scalar μ, the triangle inequality yields

$$\Big\|aM - \tfrac{1}{p}I\Big\|_{\mathrm{op}} \;\le\; a\|M - \mu I\|_{\mathrm{op}} + \big|a\mu - \tfrac{1}{p}\big|.$$

Use the previous bound $\|M - \mu I\|_{\mathrm{op}} \le \frac{C}{p} + O(\frac{1}{p})$ with $\mu = \frac{\operatorname{tr} S_t + \|\bar{w}_t\|^2}{p}$, and note that

$$\Big|\mu - \tfrac{1}{p}\Big| = \frac{\big|\operatorname{tr} S_t + \|\bar{w}_t\|^2 - 1\big|}{p} = \frac{\big|\frac{1}{m}\sum_{j=1}^m \|w_{j,t}\|^2 - 1\big|}{p} = O\Big(\frac{1}{p}\Big).$$

so $\big|a\mu - \frac{1}{p}\big| = O(\frac{1}{p})$ for $a \in \{c_-^{-2}, c_+^{-2}\}$. Combining these with $a \le \max\{c_-^{-2}, c_+^{-2}\}$ yields

$$\big\|\widehat{M} - \tfrac{1}{p} I\big\|_{\mathrm{op}} \;\le\; \max\{c_-^{-2}, c_+^{-2}\}\Big(\frac{C}{p} + O\Big(\frac{1}{p}\Big)\Big),$$

which is the stated $\frac{C_{\mathrm{ag}}}{p}$ bound.

□

4.5 AGOP structure of student covariance

In this section we show that in the high-dimensional limit $p, m \to \infty$ with $m/p = \Theta(1)$, the student covariance $\Sigma_{\mathrm{s}} = \mathbb{E}[\nabla f_W(x) \nabla f_W(x)^\top]$ has an *axial* structure. We use this to compare to the teacher AGOP structure and provide a clean correlation bound between student and teacher AGOPs.

Standing setup. Throughout, $x \sim \mathcal{N}(0, I_p)$, $U \in \mathbb{R}^{p \times r}$ with $U^\top U = I_r$, $P_\star := UU^\top$, $y(x) = \sigma_\star(U^\top x)$, and $f_W(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^m \phi(w_j^\top x)$ with $g = \phi'$. Define

$$\Sigma_{\mathrm{t}} := \mathbb{E}[\nabla y(x) \nabla y(x)^\top], \qquad \Sigma_{\mathrm{s}} := \mathbb{E}[\nabla f_W(x) \nabla f_W(x)^\top],$$

and the trace-centered versions $\Sigma'_{\mathrm{t}} := \Sigma_{\mathrm{t}} - \frac{\operatorname{tr} \Sigma_{\mathrm{t}}}{p} I_p$, $\Sigma'_{\mathrm{s}} := \Sigma_{\mathrm{s}} - \frac{\operatorname{tr} \Sigma_{\mathrm{s}}}{p} I_p$.

In the previous section we showed a result on centering of mean covariance of unit normalized weights. For this section and the subsequent analysis, we state the centerting result as an assumption (Assumption 12) since it is a property of the weights w_j at a given time, and we will use it as a black-box in the following analysis. The assumption is satisfied at initialization (Assumption 5) and is approximately preserved during training (Lemma 10).

Assumption 12 (AGOP (unit-normalized, operator))**.** *There is $C_{\mathrm{ag}} > 0$ such that*

$$\Big\| \frac{1}{m} \sum_{j=1}^m \widehat{w}_j \widehat{w}_j^\top - \frac{I_p}{p} \Big\|_{\mathrm{op}} \le \frac{C_{\mathrm{ag}}}{p}, \qquad \widehat{w}_j := \frac{w_j}{\|w_j\|}.$$

Lemma 11 (Student covariance: axial structure under AGOP, axis along s)**.** *Let $x \sim \mathcal{N}(0, I_p)$, $f_W(x) = m^{-1/2} \sum_{j=1}^m \phi(w_j^\top x)$, and $g = \phi'$. Assume row–norm control $c_- \le \|w_j\| \le c_+$ and the AGOP condition in Assumption 12.*

Let $a_j := \mathbb{E}[g(\|w_j\| Z)]$ for $Z \sim \mathcal{N}(0,1)$ and $s := \sum_{j=1}^m a_j w_j$. Then there exist scalars $\alpha_{\mathrm{s}}, \beta_{\mathrm{s}}$ (depending only on $(c_\pm, \|g\|_\infty, \|g'\|_\infty, \|g''\|_\infty)$ and the empirical row–norm statistics) such that

$$\Sigma_{\mathrm{s}} = \alpha_{\mathrm{s}} I_p \; + \; \beta_{\mathrm{s}} \frac{s\, s^\top}{m} \; + \; R_p, \qquad \|R_p\|_{\mathrm{op}} \lesssim \frac{1}{p}.$$

If moreover $m/p = \Theta(1)$, the bound $\|R_p\|_{\mathrm{op}} \lesssim p^{-1}$ is uniform in m, p.

Proof. Recall

$$\nabla f_W(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^m g(w_j^\top x)\, w_j, \qquad \Sigma_{\mathrm{s}} = \mathbb{E}\big[\nabla f_W(x) \nabla f_W(x)^\top\big] = \frac{1}{m} \sum_{j,k=1}^m \mathbb{E}\big[g(T_j) g(T_k)\big]\; w_j w_k^\top,$$

where $T_j := w_j^\top x \sim \mathcal{N}(0, \|w_j\|^2)$ and $a_j := \mathbb{E}[g(T_j)]$. Using the mean–covariance decomposition,

$$\mathbb{E}[g(T_j) g(T_k)] = a_j a_k + \mathbf{1}_{j=k} \operatorname{Var}(g(T_j)) + \mathbf{1}_{j \ne k} \operatorname{Cov}(g(T_j), g(T_k)).$$

Let $s := \sum_{j=1}^m a_j w_j$. Then,

$$\Sigma_{\mathrm{s}} = \frac{1}{m} \sum_{j,k=1}^m \mathbb{E}\big[g(T_j) g(T_k)\big]\, w_j w_k^\top = \frac{1}{m} \sum_{j=1}^m \mathbb{E}\big[g(T_j)^2\big]\, w_j w_j^\top \; + \; \frac{1}{m} \sum_{j \ne k} \mathbb{E}\big[g(T_j) g(T_k)\big]\, w_j w_k^\top.$$

For $j \ne k$, use $\mathbb{E}[XY] = \operatorname{Cov}(X, Y) + \mathbb{E}[X]\mathbb{E}[Y]$ to get

$$\sum_{j \ne k} \mathbb{E}\big[g(T_j) g(T_k)\big]\, w_j w_k^\top = \sum_{j \ne k} \operatorname{Cov}\big(g(T_j), g(T_k)\big)\, w_j w_k^\top \; + \; \sum_{j \ne k} a_j a_k\, w_j w_k^\top.$$

The off-diagonal mean term re-expresses as

$$\sum_{j\neq k} a_j a_k\, w_j w_k^\top = \Big(\sum_{j=1}^m a_j w_j\Big)\Big(\sum_{k=1}^m a_k w_k\Big)^\top \;-\; \sum_{j=1}^m a_j^2\, w_j w_j^\top = ss^\top - \sum_{j=1}^m a_j^2\, w_j w_j^\top,$$

which yields

$$\Sigma_{\mathrm{s}} = \underbrace{\frac{1}{m}\sum_{j=1}^m \mathbb{E}[g(T_j)^2]\, w_j w_j^\top}_{\mathbf{D}} \;+\; \underbrace{\frac{1}{m} ss^\top - \frac{1}{m}\sum_{j=1}^m a_j^2\, w_j w_j^\top}_{\mathbf{A}} \;+\; \underbrace{\frac{1}{m}\sum_{j\neq k} \mathrm{Cov}(g(T_j), g(T_k))\, w_j w_k^\top}_{\mathbf{C}}.$$

We show: (i) $\mathbf{D} = \alpha_{\mathrm{s}} I_p + O_{\mathrm{op}}(p^{-1})$, (ii) $\mathbf{A} = \frac{1}{m} ss^\top + O_{\mathrm{op}}(p^{-1})$, (iii) $\mathbf{C} = O_{\mathrm{op}}(p^{-1})$.

(i) Diagonal part $\mathbf{D}$. Write $w_j w_j^\top = \|w_j\|^2 \widehat{w}_j \widehat{w}_j^\top$ and set $b_j := \|w_j\|^2\, \mathbb{E}[g(\|w_j\| G)^2]$, $B := \frac{1}{m}\sum_j b_j$. For unit u,

$$u^\top \mathbf{D}\, u = \frac{1}{m}\sum_{j=1}^m b_j\, (u^\top \widehat{w}_j)^2 = \frac{B}{m}\sum_{j=1}^m (u^\top \widehat{w}_j)^2 \;+\; \frac{1}{m}\sum_{j=1}^m (b_j - B)\, (u^\top \widehat{w}_j)^2,$$

where $B := \frac{1}{m}\sum_{j=1}^m b_j$. Set $q_j := (u^\top \widehat{w}_j)^2$ and $\bar{q} := \frac{1}{m}\sum_j q_j$. Then

$$u^\top \mathbf{D}\, u \;=\; B\,\bar{q} \;+\; \frac{1}{m}\sum_{j=1}^m (b_j - B)\, q_j.$$

By the AGOP assumption, $\bar{q} \;=\; \frac{1}{p} + O\big(\frac{1}{p}\big)$, so

$$B\,\bar{q} \;=\; \frac{B}{p} \;+\; O\Big(\frac{1}{p}\Big).$$

Moreover, row–norm control and bounded g imply $|b_j| \le C$, hence $|b_j - B| \le 2C$ and $0 \le q_j \le 1$, so

$$\Big|\frac{1}{m}\sum_{j=1}^m (b_j - B)\, q_j\Big| \;\le\; \max_j |b_j - B| \cdot \frac{1}{m}\sum_{j=1}^m q_j \;=\; O(1)\cdot \bar{q} \;=\; O\Big(\frac{1}{p}\Big).$$

Combining,

$$u^\top \mathbf{D}\, u \;=\; \frac{B}{p} \;+\; O\Big(\frac{1}{p}\Big),$$

so $\mathbf{D} = \alpha_{\mathrm{s}} I_p + O_{\mathrm{op}}(p^{-1})$ with $\alpha_{\mathrm{s}} = B/p$.

(ii) Mean part $\mathbf{A}$. Exactly the same argument applies to $\frac{1}{m}\sum_{j=1}^m a_j^2 w_j w_j^\top$ (the weights a_j^2 are uniformly bounded), giving

$$\frac{1}{m}\sum_{j=1}^m a_j^2\, w_j w_j^\top = \alpha'_{\mathrm{s}} I_p + O_{\mathrm{op}}(p^{-1}), \qquad \alpha'_{\mathrm{s}} := \frac{1}{p}\cdot\frac{1}{m}\sum_{j=1}^m a_j^2 \|w_j\|^2.$$

Absorb $-\alpha'_{\mathrm{s}} I_p$ into α_{s}. Thus $\mathbf{A} = \frac{1}{m} ss^\top + O_{\mathrm{op}}(p^{-1})$.

(iii) Off–diagonal part $\mathbf{C}$*: split and bounds.* Let $r_j := \|w_j\| \in [c_-, c_+]$, $\widehat{w}_j := w_j / r_j$, $A := [\widehat{w}_1, \ldots, \widehat{w}_m] \in \mathbb{R}^{p\times m}$, and $\mathsf{G} := A^\top A \in \mathbb{R}^{m\times m}$, so $\mathsf{G}_{jk} = \rho_{jk} := \widehat{w}_j^\top \widehat{w}_k$ and $\|A\|_{\mathrm{op}}^2 = \|\mathsf{G}\|_{\mathrm{op}}$. For $x \sim \mathcal{N}(0, I_p)$ put $T_j := w_j^\top x \sim \mathcal{N}(0, r_j^2)$. Define, for $j \neq k$,

$$F_{jk}(\rho) \;:=\; \mathbb{E}\big[g(r_j Z_1)\, g(r_k Z_2)\big], \quad (Z_1, Z_2) \sim \mathcal{N}\Big(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\Big),$$

so that $\mathrm{Cov}(g(T_j), g(T_k)) = F_{jk}(\rho_{jk}) - F_{jk}(0)$.

By Price's theorem,

$$F'_{jk}(\rho) = r_j r_k \, \mathbb{E}\big[g'(r_j Z_1) g'(r_k Z_2)\big], \qquad F''_{jk}(\rho) = r_j^2 r_k^2 \, \mathbb{E}\big[g''(r_j Z_1) g''(r_k Z_2)\big].$$

Thus at $\rho = 0$, using independence of Z_1, Z_2,

$$F'_{jk}(0) = r_j r_k \, \alpha_j \alpha_k, \quad \alpha_j := \mathbb{E}[g'(r_j Z)], \qquad \big|F''_{jk}(\rho)\big| \le r_j^2 r_k^2 \, \|g''\|_\infty^2.$$

Taylor's theorem with mean-value remainder yields, for some ξ_{jk} between 0 and ρ_{jk},

$$F_{jk}(\rho_{jk}) - F_{jk}(0) = \underbrace{\rho_{jk} \, r_j r_k \, \alpha_j \alpha_k}_{\text{linear part}} + \underbrace{\frac{1}{2} \, \rho_{jk}^2 \, r_j^2 r_k^2 \, \mathbb{E}[g''(r_j Z_1) g''(r_k Z_2)]_{\rho = \xi_{jk}}}_{\text{quadratic remainder}}. \tag{9}$$

Consequently,

$$\mathbf{C} = \frac{1}{m} \sum_{j \neq k} \big(F_{jk}(\rho_{jk}) - F_{jk}(0)\big) \, w_j w_k^\top =: \mathbf{C}_{\mathrm{lin}} + \mathbf{C}_{\mathrm{quad}},$$

with $\mathbf{C}_{\mathrm{lin}}$ the linear Price part and $\mathbf{C}_{\mathrm{quad}}$ the quadratic remainder.

Linear part $\mathbf{C}_{\mathrm{lin}}$. Using Eq. (9) and $w_j w_k^\top = (r_j \widehat{w}_j)(r_k \widehat{w}_k)^\top$,

$$\mathbf{C}_{\mathrm{lin}} = \frac{1}{m} \sum_{j \neq k} \rho_{jk} \, (r_j r_k) \, \alpha_j \alpha_k \, (r_j \widehat{w}_j)(r_k \widehat{w}_k)^\top = \frac{1}{m} \sum_{j \neq k} \rho_{jk} \, v_j v_k \, \widehat{w}_j \widehat{w}_k^\top,$$

where $v_j := r_j^2 \alpha_j$. Writing $V := \mathrm{diag}(v)$ and noting the zero diagonal, this is

$$\mathbf{C}_{\mathrm{lin}} = \frac{1}{m} \, A\big(V(\mathsf{G} - I)V\big)A^\top.$$

Hence

$$\|\mathbf{C}_{\mathrm{lin}}\|_{\mathrm{op}} \le \frac{1}{m} \, \|A\|_{\mathrm{op}}^2 \, \|V\|_{\mathrm{op}}^2 \, \|\mathsf{G} - I\|_{\mathrm{op}} \le \frac{1}{m} \, \|\mathsf{G}\|_{\mathrm{op}} \, \big(c_+^2 \|g'\|_\infty\big)^2 \, \big(\|\mathsf{G}\|_{\mathrm{op}} + 1\big).$$

Under AGOP condition, $\|\mathsf{G}\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}^2 = \big\| \sum_j \widehat{w}_j \widehat{w}_j^\top \big\|_{\mathrm{op}} \lesssim m/p$, so

$$\|\mathbf{C}_{\mathrm{lin}}\|_{\mathrm{op}} \;\lesssim\; \frac{m}{p^2} \;+\; \frac{1}{p} \;=\; O\Big(\frac{1}{p}\Big) \quad \text{when } m/p = \Theta(1).$$

Quadratic remainder $\mathbf{C}_{\mathrm{quad}}$. First note that we can write

$$\mathbf{C}_{\mathrm{quad}} = \frac{1}{m} \sum_{j \neq k} \Big(\tfrac{1}{2} \, \rho_{jk}^2 \, r_j^3 r_k^3 \, \mathbb{E}_{\xi_{jk}}[g''(r_j Z_1) g''(r_k Z_2)]\Big) \widehat{w}_j \widehat{w}_k^\top,$$

Using the uniform bound $\big|\mathbb{E}[g''(\cdot) \, g''(\cdot)]\big| \le \|g''\|_\infty^2$,

$$\Big|\tfrac{1}{2} \, \rho_{jk}^2 \, r_j^2 r_k^2 \, \mathbb{E}_{\xi_{jk}}\big[g''(r_j Z_1) g''(r_k Z_2)\big]\Big| \;\le\; \tfrac{1}{2} \, \rho_{jk}^2 \, r_j^2 r_k^2 \, \|g''\|_\infty^2.$$

Multiplying by $w_j w_k^\top = (r_j \widehat{w}_j)(r_k \widehat{w}_k)^\top$ yields the coefficient $\frac{1}{2} \, \rho_{jk}^2 \, r_j^3 r_k^3 \, \|g''\|_\infty^2$ in front of $\widehat{w}_j \widehat{w}_k^\top$. Hence we can rewrite $\mathbf{C}_{\mathrm{quad}}$ as

$$\mathbf{C}_{\mathrm{quad}} = \frac{1}{m} \, A\Big(\big(\mathsf{G} \circ \mathsf{G} - I\big) \circ B''\Big)A^\top, \qquad B''_{jk} = \tfrac{1}{2} \, r_j^3 r_k^3 \, \mathbb{E}_{\xi_{jk}}\big[g''(r_j Z_1) g''(r_k Z_2)\big] \;\; (j \neq k), \quad B''_{jj} = 0,$$

where $A = [\widehat{w}_1, \dots, \widehat{w}_m] \in \mathbb{R}^{p\times m}$, $\mathsf{G} = A^\top A$, and $\circ$ denotes the Hadamard product. Row–norm control implies $\|B''\|_\infty \le \frac{1}{2}\, c_+^6\, \|g''\|_\infty^2$.

For any symmetric X and any mask M,

$$\|X \circ M\|_{\mathrm{op}} \;\le\; \|M\|_\infty \,\big\||X|\big\|_{\mathrm{op}} \;\le\; \|M\|_\infty \max_i \sum_j |X_{ij}|,$$

(apply the Rayleigh quotient to $|X|$ and use the Perron–Frobenius bound for nonnegative symmetric matrices). With $X = \mathsf{G}\circ\mathsf{G} - I$ (nonnegative off–diagonal, zero diagonal after masking),

$$\big\|(\mathsf{G}\circ\mathsf{G} - I)\circ B''\big\|_{\mathrm{op}} \;\le\; \|B''\|_\infty \max_i \sum_j (\mathsf{G}\circ\mathsf{G})_{ij} \;=\; \|B''\|_\infty \max_i \sum_j \rho_{ij}^2.$$

But

$$\sum_j \rho_{ij}^2 = \widehat{w}_i^\top \Big(\sum_{\ell=1}^m \widehat{w}_\ell \widehat{w}_\ell^\top\Big)\widehat{w}_i \;\le\; \Big\|\sum_{\ell=1}^m \widehat{w}_\ell \widehat{w}_\ell^\top\Big\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}^2 = \|\mathsf{G}\|_{\mathrm{op}}.$$

Therefore

$$\|\mathbf{C}_{\mathrm{quad}}\|_{\mathrm{op}} \;\le\; \frac{1}{m}\,\|A\|_{\mathrm{op}}^2\,\big\|(\mathsf{G}\circ\mathsf{G} - I)\circ B''\big\|_{\mathrm{op}} \;\le\; \frac{1}{m}\,\|A\|_{\mathrm{op}}^2\,\|B''\|_\infty\,\|\mathsf{G}\|_{\mathrm{op}}.$$

Under AGOP condition, $\|A\|_{\mathrm{op}}^2 = \|\mathsf{G}\|_{\mathrm{op}} \lesssim m/p$. Hence

$$\|\mathbf{C}_{\mathrm{quad}}\|_{\mathrm{op}} \;\lesssim\; \frac{1}{m}\cdot\frac{m}{p}\cdot\frac{m}{p} = \frac{m^2}{p^3} \;=\; O\Big(\frac{1}{p}\Big) \quad \text{when } m/p = \Theta(1).$$

Combining the bounds for $\mathbf{C}_{\mathrm{lin}}$ and $\mathbf{C}_{\mathrm{quad}}$ gives

$$\|\mathbf{C}\|_{\mathrm{op}} \;\le\; \|\mathbf{C}_{\mathrm{lin}}\|_{\mathrm{op}} + \|\mathbf{C}_{\mathrm{quad}}\|_{\mathrm{op}} \;=\; O\Big(\frac{1}{p}\Big) \quad \text{when } m/p = \Theta(1),$$

which completes part (iii) under AGOP condition.

Thus,

$$\Sigma_s = \mathbf{D} + \mathbf{A} + \mathbf{C} = (\alpha_{\mathrm{s}} - \alpha_{\mathrm{s}'})I_p + \frac{ss^\top}{m} + R_p,$$

where $\|R_p\|_{\mathrm{op}} = O\Big(\frac{1}{p}\Big)$ when $m/p = \Theta(1)$.

This completes the proof of the lemma.

□

Now we can specialize the axis to be along the mean weight vector $\bar{w}$ under small coefficient dispersion.

Corollary 1 (Axis along $\bar{w}$ under small coefficient dispersion)**.** *Under the hypotheses of the Lemma 11, assume additionally*

$$\frac{1}{m}\sum_{j=1}^m \big(a_j - \bar{a}\big)^2 \;\lesssim\; \frac{1}{p^2}, \qquad \bar{a} := \frac{1}{m}\sum_{j=1}^m a_j.$$

Then, there exist scalars $\alpha_{\mathrm{s}}, \beta_{\mathrm{s}}$ such that

$$\Sigma_{\mathrm{s}} = \alpha_{\mathrm{s}} I_p \;+\; \beta_{\mathrm{s}}\,\bar{w}\bar{w}^\top \;+\; R'_p, \qquad \|R'_p\|_{\mathrm{op}} \;\lesssim\; \frac{1}{p}.$$

Proof. Write

$$s = \sum_{j=1}^{m} a_j w_j = \bar{a} \sum_{j=1}^{m} w_j + \sum_{j=1}^{m} (a_j - \bar{a}) w_j = \bar{a}\, \sqrt{m}\, \bar{w} + \delta, \qquad \delta := \sum_{j=1}^{m} (a_j - \bar{a}) w_j.$$

Let $A = [\widehat{w}_1, \dots, \widehat{w}_m] \in \mathbb{R}^{p \times m}$ and $r = (\|w_1\|, \dots, \|w_m\|)^\top$. Then $\delta = Ab$ with $b_j = r_j(a_j - \bar{a})$. By row–norm control and the dispersion bound,

$$\|b\|_2^2 = \sum_{j=1}^{m} r_j^2 (a_j - \bar{a})^2 \;\le\; c_+^2 \sum_{j=1}^{m} (a_j - \bar{a})^2 \;\lesssim\; c_+^2\, \frac{m}{p^2}, \quad \Rightarrow \quad \|b\|_2 \;\lesssim\; c_+\, \frac{\sqrt{m}}{p}.$$

AGOP condition implies

$$\Big\| \frac{1}{m} A A^\top - \frac{I_p}{p} \Big\|_{\mathrm{op}} \le \frac{C_{\mathrm{ag}}}{p} \implies \|A\|_{\mathrm{op}}^2 = \|A A^\top\|_{\mathrm{op}} \;\lesssim\; \frac{m}{p}.$$

Hence

$$\|\delta\| = \|Ab\| \;\le\; \|A\|_{\mathrm{op}} \|b\|_2 \;\lesssim\; \sqrt{\frac{m}{p}} \cdot \frac{\sqrt{m}}{p} \;=\; \frac{m}{p^{3/2}}.$$

Moreover,

$$\|\bar{w}\| = \frac{1}{m} \|Ar\| \le \frac{1}{m} \|A\|_{\mathrm{op}} \|r\|_2 \lesssim \frac{1}{m} \sqrt{\frac{m}{p}} \cdot c_+ \sqrt{m} = O\Big(\frac{1}{\sqrt{p}}\Big),$$

using $\|r\|_2 \le c_+ \sqrt{m}$.

Now expand

$$\frac{1}{m} s s^\top = \bar{a}^{\,2}\, \bar{w} \bar{w}^\top \;+\; \underbrace{\frac{\bar{a}}{\sqrt{m}} \left(\bar{w} \delta^\top + \delta \bar{w}^\top \right)}_{=: E_1} \;+\; \underbrace{\frac{1}{m}\, \delta \delta^\top}_{=: E_2}.$$

With $|\bar{a}| \le \|g\|_\infty$, $\|\bar{w}\| = O(p^{-1/2})$, and $\|\delta\| \lesssim m/p^{3/2}$,

$$\|E_1\|_{\mathrm{op}} \;\le\; \frac{2|\bar{a}|}{\sqrt{m}}\, \|\bar{w}\|\, \|\delta\| \;\lesssim\; \frac{1}{\sqrt{m}} \cdot \frac{1}{\sqrt{p}} \cdot \frac{m}{p^{3/2}} \;=\; \frac{\sqrt{m}}{p^2} \;=\; O\Big(\frac{1}{p^{3/2}}\Big) \quad \text{if } m/p = \Theta(1),$$

and

$$\|E_2\|_{\mathrm{op}} = \frac{1}{m} \|\delta\|^2 \;\lesssim\; \frac{1}{m} \cdot \frac{m^2}{p^3} \;=\; \frac{m}{p^3} \;=\; O\Big(\frac{1}{p^2}\Big) \quad \text{if } m/p = \Theta(1).$$

Thus

$$\frac{1}{m} s s^\top \;=\; \bar{a}^{\,2}\, \bar{w} \bar{w}^\top \;+\; O_{\mathrm{op}}\Big(\frac{1}{p}\Big) \quad \text{(in fact } O(p^{-3/2}) \text{ here)}.$$

Using this, we can rewrite the axial term in the decomposition from Lemma 13. Absorbing $\frac{1}{m} s s^\top - \bar{a}^{\,2} \bar{w} \bar{w}^\top$ into the remainder yields

$$\Sigma_{\mathrm{s}} \;=\; \alpha_{\mathrm{s}} I_p \;+\; \beta_{\mathrm{s}}\, \bar{w} \bar{w}^\top \;+\; R'_p, \qquad \|R'_p\|_{\mathrm{op}} \;\lesssim\; \frac{1}{p}.$$

□

Now, we will state the analogous result for the teacher AGOP.

Lemma 12 (Teacher AGOP)**.** *Let $Z \sim \mathcal{N}(0, I_r)$ and define $A_\star := \mathbb{E}[\nabla \sigma_\star(Z) \nabla \sigma_\star(Z)^\top] \succeq 0$. Then*

$$\Sigma_{\mathrm{t}} \;=\; U\, A_\star\, U^\top, \qquad \Sigma'_{\mathrm{t}} = \; U \Big(A_\star - \tfrac{\operatorname{tr} A_\star}{p} I_r \Big) U^\top \oplus \; \Big(- \tfrac{\operatorname{tr} A_\star}{p} I_{p-r} \Big).$$

In the isotropic case $A_\star = \tau_\star I_r$, $\Sigma_{\mathrm{t}} = \tau_\star P_\star$ and $\Sigma'_{\mathrm{t}} = \tau_\star \big(P_\star - \frac{r}{p} I_p \big)$.

Lemma 13 (Student AGOP: axial structure along s). *Let $x \sim \mathcal{N}(0, I_p)$, $f_W(x) = m^{-1/2} \sum_{j=1}^m \phi(w_j^\top x)$ with $g = \phi'$, and assume row–norm control $c_- \le \|w_j\| \le c_+$, bounded smoothness $\|g\|_\infty, \|g'\|_\infty, \|g''\|_\infty < \infty$, and the unit–normalized strong AGOP $\big\|\frac{1}{m}\sum_j \widehat{w}_j \widehat{w}_j^\top - \frac{I_p}{p}\big\|_{\mathrm{op}} \le C_{\mathrm{ag}}/p$, where $\widehat{w}_j = w_j/\|w_j\|$. Let $a_j := \mathbb{E}[g(\|w_j\|G)]$ for $G \sim \mathcal{N}(0,1)$ and set $s := \sum_{j=1}^m a_j w_j$. Then there exist scalars $\alpha_{\mathrm{s}}, \beta_{\mathrm{s}}$ depending only on $(c_\pm, \|g\|_\infty, \|g'\|_\infty, \|g''\|_\infty)$ and $\{\|w_j\|\}$ such that*

$$\Sigma_{\mathrm{s}} = \alpha_{\mathrm{s}} I_p + \beta_{\mathrm{s}} \frac{ss^\top}{m} + R_p, \qquad \|R_p\|_{\mathrm{op}} \lesssim p^{-1}.$$

Consequently,

$$\Sigma'_{\mathrm{s}} := \Sigma_{\mathrm{s}} - \frac{\operatorname{tr}\Sigma_{\mathrm{s}}}{p} I_p = \kappa_{\mathrm{s}}\Big(\widehat{s}\widehat{s}^\top - \tfrac{1}{p} I_p\Big) + R'_p, \quad \widehat{s} := s/\|s\|, \quad \kappa_{\mathrm{s}} := \beta_{\mathrm{s}} \frac{\|s\|^2}{m}, \quad \|R'_p\|_{\mathrm{op}} \lesssim p^{-1}.$$

Proof. Same proof as in Lemma 11 (decomposition into diagonal/mean/correlation pieces and AGOP control), noting that

$$\frac{s\,s^\top}{m} - \frac{\|s\|^2}{m}\frac{I_p}{p} = \frac{\|s\|^2}{m}\Big(\widehat{s}\widehat{s}^\top - \frac{I_p}{p}\Big).$$

The $O_{op}(p^{-1})$ remainder persists after trace centering. □

Now we can state the main AGOP–Subspace Alignment theorem when the student axis is along s. This establishes a direct connection between the alignment and the projection of s onto the teacher subspace. Now, in order to understand the alignment behavior, it suffices to analyze the projection of s onto the teacher subspace, which is often more tractable. We use this for analyzing the feature learning behavior for both MUON and SGD.

Theorem 3 (Main AGOP–Subspace Alignment (student AGOP, axis along s)). *Adopt the Standing Setup, Assumption 12, and Lemma 13, and suppose $s \neq 0$. Let $\widehat{s} := s/\|s\|$, set $c_{\mathrm{s}} := \|P_\star \widehat{s}\| \in [0,1]$, and let $A_\star$ be as in Lemma 12. Then:*

(i) Isotropic teacher. *If $A_\star = \tau_\star I_r$ (so $\Sigma'_{\mathrm{t}} = \tau_\star(P_\star - \frac{r}{p} I_p)$), then*

$$\frac{\langle \Sigma'_{\mathrm{s}}, \Sigma'_{\mathrm{t}} \rangle_F}{\|\Sigma'_{\mathrm{s}}\|_F \, \|\Sigma'_{\mathrm{t}}\|_F} = \frac{c_{\mathrm{s}}^2 - \frac{r}{p}}{\sqrt{1-\frac{1}{p}}\sqrt{r\big(1-\frac{r}{p}\big)}} + O\Big(\frac{1}{\sqrt{p}} \cdot \frac{1}{\|\Sigma'_{\mathrm{s}}\|_F}\Big).$$

In particular, if the axial strength κ_{s} in Lemma 13 is bounded below along the horizon (so $\|\Sigma'_{\mathrm{s}}\|_F \asymp 1$), the remainder is $O(p^{-1/2})$.

(ii) General anisotropic teacher. *Let $\lambda_1, \dots, \lambda_r$ be the eigenvalues of $A_\star$, and put $u_{\mathrm{s}} := U^\top \widehat{s} \in \mathbb{R}^r$. Then*

$$\frac{\langle \Sigma'_{\mathrm{s}}, \Sigma'_{\mathrm{t}} \rangle_F}{\|\Sigma'_{\mathrm{s}}\|_F \, \|\Sigma'_{\mathrm{t}}\|_F} = \frac{u_{\mathrm{s}}^\top A_\star u_{\mathrm{s}} - \frac{\operatorname{tr} A_\star}{p}}{\sqrt{1-\frac{1}{p}}\;\|\Sigma'_{\mathrm{t}}\|_F} + O\Big(\frac{1}{\sqrt{p}} \cdot \frac{1}{\|\Sigma'_{\mathrm{s}}\|_F}\Big),$$

with exact teacher normalization $\|\Sigma'_{\mathrm{t}}\|_F^2 = \sum_{i=1}^r \big(\lambda_i - \frac{\operatorname{tr} A_\star}{p}\big)^2 + (p-r)\big(\frac{\operatorname{tr} A_\star}{p}\big)^2$.

Proof. From Lemma 13, write

$$\Sigma'_{\mathrm{s}} = \kappa_{\mathrm{s}}\Big(\widehat{s}\widehat{s}^\top - \tfrac{1}{p} I_p\Big) + R'_p, \qquad \|R'_p\|_{op} \lesssim p^{-1}.$$

Since $\operatorname{tr}\Sigma'_{\mathrm{t}} = 0$, we have

$$\langle \widehat{s}\widehat{s}^\top - \tfrac{1}{p} I_p,\ \Sigma'_{\mathrm{t}} \rangle_F = \widehat{s}^\top \Sigma'_{\mathrm{t}}\, \widehat{s}.$$

(i) Isotropic case. With $\Sigma'_{\mathrm{t}} = \tau_\star(P_\star - \frac{r}{p} I_p)$,

$$\widehat{s}^\top \Sigma'_{\mathrm{t}}\, \widehat{s} = \tau_\star\Big(\widehat{s}^\top P_\star \widehat{s} - \tfrac{r}{p}\Big) = \tau_\star\Big(c_{\mathrm{s}}^2 - \tfrac{r}{p}\Big).$$

Furthermore,

$$\|\widehat{s}\widehat{s}^\top - \tfrac{1}{p} I_p\|_F^2 = 1 - \frac{1}{p}, \qquad \|\Sigma'_{\mathrm{t}}\|_F^2 = \tau_\star^2\, r\Big(1 - \frac{r}{p}\Big).$$

Therefore,

$$\frac{\langle \Sigma'_{\mathrm{s}}, \Sigma'_{\mathrm{t}} \rangle_F}{\|\Sigma'_{\mathrm{s}}\|_F\, \|\Sigma'_{\mathrm{t}}\|_F} = \frac{\kappa_{\mathrm{s}} \tau_\star\, (c_{\mathrm{s}}^2 - \frac{r}{p}) + \langle R'_p, \Sigma'_{\mathrm{t}} \rangle_F}{\|\Sigma'_{\mathrm{t}}\|_F \sqrt{\kappa_{\mathrm{s}}^2 (1 - \frac{1}{p}) + \|R'_p\|_F^2 + 2\kappa_{\mathrm{s}} \langle \widehat{s}\widehat{s}^\top - \frac{I}{p}, R'_p \rangle_F}}.$$

Use $|\langle R'_p, \Sigma'_{\mathrm{t}} \rangle_F| \le \|R'_p\|_{op}\, \|\Sigma'_{\mathrm{t}}\|_F \lesssim p^{-1} \|\Sigma'_{\mathrm{t}}\|_F$ and $|\langle \widehat{s}\widehat{s}^\top - \frac{I}{p}, R'_p \rangle_F| \le \|R'_p\|_F \lesssim p^{-1/2}$ to obtain the displayed $O\big(p^{-1/2}/\|\Sigma'_{\mathrm{s}}\|_F\big)$ remainder. If $\kappa_{\mathrm{s}} \gtrsim 1$, then $\|\Sigma'_{\mathrm{s}}\|_F \asymp 1$ and the error is $O(p^{-1/2})$.

(ii) Anisotropic case. Write $a := \mathrm{tr}(A_\star)/p$ and use the teacher block form:

$$\widehat{s}^\top \Sigma'_{\mathrm{t}}\, \widehat{s} = (U^\top \widehat{s})^\top (A_\star - a I_r)(U^\top \widehat{s}) - a\, \|(I - P_\star)\widehat{s}\|^2 = u_{\mathrm{s}}^\top A_\star u_{\mathrm{s}} - a\, \|u_{\mathrm{s}}\|^2 - a(1 - \|u_{\mathrm{s}}\|^2) = u_{\mathrm{s}}^\top A_\star u_{\mathrm{s}} - a.$$

Using similar ideas as in (i) for the normalization and the remainder proves the general anisotropic setting as well. □

Corollary 2 (Axis change to $\bar{w}$ under small dispersion)**.** *Assume in addition that the coefficient dispersion is small:*

$$\frac{1}{m} \sum_{j=1}^m \big(a_j - \bar{a}\big)^2 \;\lesssim\; \frac{1}{p^2}, \qquad \bar{a} := \frac{1}{m} \sum_{j=1}^m a_j.$$

Then, the conclusions of Theorem 3 remain valid with $\widehat{s}$ replaced by $\hat{\bar{w}} := \bar{w}/\|\bar{w}\|$ and the same $O(p^{-1/2})$ remainder (provided $\kappa_{\mathrm{s}} \gtrsim 1$).

Proof. From Corollary 1, we achieve the axis change in Theorem 3 at the cost of an additional assumption on the coefficient dispersion. □

5 SGD

In this section, we provide rates for feature learning in terms of the expected change in the angle between the projector $P_\star$ and the mean direction $\bar{w}_t$ (see Theorem 4).

We restate the standing assumptions for clarity: we assume the following for the setup: $x \sim \mathcal{N}(0, I_p)$, $U \in \mathbb{R}^{p\times r}$ with $U^\top U = I_r$, $P_\star = UU^\top$, $y(x) = \sigma_\star(U^\top x)$, $f_W(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^m \phi(w_j^\top x)$, $g = \phi'$. Row-norm control: $\|w_j\| \in [c_-, c_+]$. We denote $R_t := \sqrt{\mathbb{E}[e(x)^2]}$. For clarity, we distinguish between the *realized minibatch step*

$$\Delta\bar{w}_t := \bar{w}_{t+1} - \bar{w}_t,$$

and its *conditional mean step*

$$\bar{h}_t := \mathbb{E}[\Delta\bar{w}_t \mid W_t].$$

Unless stated otherwise, $\bar{h}_t$ will refer to the conditional mean update.

In the analysis below, we remove the subscript/superscipt for the step t for the weights to avoid overloading the notations. Weights are almost always timed unless the results state otherwise.

First, we state the Stein's identity for vectorial setting.

Lemma 14 (Stein's identity)**.** *Let $x \sim \mathcal{N}(0, I_p)$ and let $F : \mathbb{R}^p \to \mathbb{R}^r$ be C^1 with $\mathbb{E}[\|F(x)\|]+\mathbb{E}[\|\nabla F(x)\|_F] < \infty$. Then, entrywise,*

$$\mathbb{E}[x_i F_j(x)] = \mathbb{E}[\partial_{x_i} F_j(x)], \qquad 1 \le i \le p,\ 1 \le j \le r,$$

or equivalently in matrix form,

$$\mathbb{E}[x\, F(x)^\top] = \mathbb{E}[\nabla F(x)].$$

The scalar case $r = 1$ yields $\mathbb{E}[x\,\varphi(x)] = \mathbb{E}[\nabla\varphi(x)]$.

Proof. Let $\gamma_p(x) = (2\pi)^{-p/2} e^{-\|x\|^2/2}$ be the standard Gaussian density, so $\partial_{x_i}\gamma_p(x) = -x_i\gamma_p(x)$. Fix i and write $x = (t, x_{-i})$ with $t = x_i$. For $\varphi \in C^1$ with $\mathbb{E}[|\varphi(x)|] + \mathbb{E}[\|\nabla\varphi(x)\|] < \infty$,

$$\begin{aligned}\mathbb{E}[x_i\varphi(x)] &= \int_{\mathbb{R}^p} x_i\varphi(x)\,\gamma_p(x)\,dx = -\int_{\mathbb{R}^p} \varphi(x)\,\partial_{x_i}\gamma_p(x)\,dx \\ &= -\int_{\mathbb{R}^{p-1}}\int_{\mathbb{R}} \varphi(t, x_{-i})\,\partial_t\gamma_p(t, x_{-i})\,dt\,dx_{-i}.\end{aligned}$$

Integrating by parts in t gives

$$\int_{\mathbb{R}} \varphi\,(-\partial_t\gamma_p)\,dt = \big[\varphi\,\gamma_p\big]_{t=-\infty}^{t=+\infty} - \int_{\mathbb{R}} (\partial_t\varphi)\,\gamma_p\,dt.$$

The boundary term vanishes since $\gamma_p(t, x_{-i}) = \gamma_1(t)\gamma_{p-1}(x_{-i})$ with $\gamma_1(t) \to 0$ for $|t| \to \infty$ and $|\varphi|$ is integrable under γ_p, dominated convergence gives $\varphi(t, x_{-i})\gamma_p(t, x_{-i}) \to 0$ for a.e. x_{-i}, and the boundary contribution integrates to 0 over x_{-i}. Hence

$$\mathbb{E}[x_i\varphi(x)] = \int_{\mathbb{R}^p} \partial_{x_i}\varphi(x)\,\gamma_p(x)\,dx = \mathbb{E}[\partial_{x_i}\varphi(x)].$$

Stacking over i yields $\mathbb{E}[x\,\varphi(x)] = \mathbb{E}[\nabla\varphi(x)]$. If $F = (F_1, \dots, F_r)$ with $\mathbb{E}[\|F(x)\|] + \mathbb{E}[\|\nabla F(x)\|_F] < \infty$, then applying the scalar case to each component F_j provides the stated claim of the lemma, i.e., $\mathbb{E}[x_i F_j(x)] = \mathbb{E}[\partial_{x_i} F_j(x)]$. □

Now, we provide a general result on a relation between the projector $P_\star$ and the mean direction $\bar{w}_t$ in terms of the change in angle for a time step. This result will be used in the analysis of rate of feature learning for both SGDand MUON.

Lemma 15 (Angle differential). *Let* $u_t := \bar{w}_t/\|\bar{w}_t\|$, $c_t = \|P_\star u_t\| = \cos\theta_t$, *and*

$$v_t := \frac{(I - u_t u_t^\top)P_\star u_t}{\|(I - u_t u_t^\top)P_\star u_t\|}.$$

Then, with $\Delta\bar{w}_t = \bar{w}_{t+1} - \bar{w}_t$,

$$\theta_{t+1} - \theta_t = -\frac{v_t^\top \Delta\bar{w}_t}{\|\bar{w}_t\|} + O\Big(\frac{\|\Delta\bar{w}_t\|^2}{\|\bar{w}_t\|^2}\Big), \qquad v_t^\top P_\star \bar{w}_t = \|\bar{w}_t\|\cos\theta_t \sin\theta_t.$$

Proof. This result follows from the observation that the map $w \mapsto u = w/\|w\|$ has differential $(I - uu^\top)\mathrm{d}w/\|w\| + O(\|\mathrm{d}w\|^2)$. For $f(u) = \|P_\star u\|$, $\nabla f(u) = P_\star u/\|P_\star u\|$. Since $c = \cos\theta$, $\mathrm{d}\theta = -(\mathrm{d}c)/\sin\theta$. Combining the differentials give the first result.

The second result is a straight-forward application of the definition of v_t. □

Lemma 16 (Axial under mean-isotropy). *Assume row-norm control* $\|w_j\| \in [c_-, c_+]$ *and*

$$\Big\|\frac{1}{m}\sum_{j=1}^m \widehat{w}_j\widehat{w}_j^\top - \frac{I_p}{p}\Big\|_{op} = O\Big(\frac{1}{p}\Big), \qquad \widehat{w}_j := \frac{w_j}{\|w_j\|}.$$

Let $(\beta_j)_{j\le m}$ *be deterministic with* $|\beta_j| \le C$ *(for some constant* $C > 0$*), and set* $u_t := \bar{w}_t/\|\bar{w}_t\|$. *Then*

$$\sum_{j=1}^m \beta_j \widehat{w}_j = \Big(\sum_{j=1}^m \beta_j\, u_t^\top \widehat{w}_j\Big)u_t \;+\; r_\perp, \qquad \Big|\sum_{j=1}^m \beta_j\, u_t^\top \widehat{w}_j\Big| = O\Big(\frac{m}{\sqrt{p}}\Big), \quad \|r_\perp\| = O\Big(\frac{m}{\sqrt{p}}\Big),$$

and, consequently, $\big\|\sum_{j=1}^m \beta_j\, U^\top \widehat{w}_j\big\| = O(m/\sqrt{p})$.

Proof. For unit $s \perp u_t$,

$$\Big|\, s^\top \sum_{j=1}^m \beta_j \widehat{w}_j \,\Big| \le \Big(\sum_j \beta_j^2\Big)^{1/2}\Big(\sum_{j=1}^m (s^\top \widehat{w}_j)^2\Big)^{1/2} \le C\sqrt{m}\Big(\sum_{j=1}^m (s^\top \widehat{w}_j)^2\Big)^{1/2}.$$

AGOP condition (Assumption 12) gives $\frac{1}{m}\sum_j (s^\top \widehat{w}_j)^2 = \frac{1}{p} + O(\frac{1}{p})$, hence $\sum_j (s^\top \widehat{w}_j)^2 = O(m/p)$ and $\|r_\perp\| = O(m/\sqrt{p})$. With $s = u_t$ the same estimate yields $\big|\sum_j \beta_j\, u_t^\top \widehat{w}_j\big| = O(m/\sqrt{p})$. Finally, $\|U^\top\|_{op} \le 1$. □

Lemma 17 (Teacher–projected identity for SGD). *Let* $Z := U^\top x$ *and fix* $g := \phi'$. *Define*

$$S_0'(x) := \sum_{j=1}^m g(w_j^\top x), \qquad S_0''(x) := \sum_{j=1}^m g'(w_j^\top x)\, U^\top w_j.$$

With $e(x) := f_W(x) - \sigma_\star(U^\top x)$, *the conditional-mean minibatch step projected to the teacher space satisfies*

$$U^\top \bar{h}_t = -\frac{\eta}{m\sqrt{m}}\,\mathbb{E}\Big[\big(\nabla_Z f_W(Z) - \nabla\sigma_\star(Z)\big)\, S_0'(x)\Big] \;-\; \frac{\eta}{m\sqrt{m}}\,\mathbb{E}\big[e(x)\, S_0''(x)\big].$$

Here all expectations are over the minibatch randomness x *conditional on* W_t.

Proof. The realized averaged SGD step is

$$\Delta\bar{w}_t = -\frac{\eta}{m\sqrt{m}}\sum_{j=1}^m e(x)\, g(w_j^\top x)\, x.$$

Taking conditional expectation given W_t yields

$$\bar{h}_t = -\frac{\eta}{m\sqrt{m}} \sum_{j=1}^{m} \mathbb{E}\big[e(x)\, g(w_j^\top x)\, x \mid W_t\big].$$

Projecting with $U^\top$ and writing $Z = U^\top x$,

$$U^\top \bar{h}_t = -\frac{\eta}{m\sqrt{m}} \sum_{j=1}^{m} \mathbb{E}\big[Z\, e(x)\, g(w_j^\top x)\big].$$

Apply Gaussian Stein in Z (Lemma 14) to the map $\psi(Z) = e(x)\, g(w_j^\top x)$:

$$\nabla_Z \big(e\, g(w_j^\top x)\big) = (\nabla_Z e)\, g(w_j^\top x) + e(x)\, g'(w_j^\top x)\, \nabla_Z (w_j^\top x).$$

Since $\nabla_Z e = \nabla_Z f_W(Z) - \nabla \sigma_\star(Z)$ and $\nabla_Z (w_j^\top x) = U^\top w_j$, summing over j gives the claim. □

Lemma 18 (Centered Price expansion at small cross–covariance)**.** *Let $Z \sim \mathcal{N}(0, I_r)$ and $G \sim \mathcal{N}(0,1)$ be independent, fix $w \in \mathbb{R}^p \setminus \{0\}$, set $T := \|w\|\, G$, and let $a \in \mathbb{R}^r$ be unit. For $|\rho| \le \rho_0 < 1$, let (Z_ρ, T) be a centered Gaussian pair with*

$$\mathrm{Var}(Z_\rho) = I_r, \qquad \mathrm{Var}(T) = \|w\|^2, \qquad \mathrm{Cov}(Z_\rho, T) = \rho\, \|w\|\, a.$$

Assume $\sigma_\star \in C^3(\mathbb{R}^r)$ with $\mathbb{E}\|\nabla^k \sigma_\star(Z)\| < \infty$ for $k = 2, 3$, and $g \in C^2(\mathbb{R})$ with $\|g'\|_\infty, \|g''\|_\infty < \infty$. Let $m_g := \mathbb{E}[g(T)]$ and write $H_\star := \mathbb{E}[\nabla^2 \sigma_\star(Z)]$. Then, as $\rho \to 0$,

$$\mathbb{E}\big[\nabla \sigma_\star(Z_\rho)\, (g(T) - m_g)\big] \;=\; \rho\, \|w\|\, H_\star a\, \mathbb{E}\big[g'(\|w\| G)\big] \;+\; O(\rho^2),$$

where the $O(\rho^2)$ bound is uniform for $\|w\|$ in compact sets.

Proof. Define the vector–valued function

$$F(\rho) := \mathbb{E}\big[\nabla \sigma_\star(Z_\rho)\, (g(T) - m_g)\big] \in \mathbb{R}^r.$$

Note $F(0) = \mathbb{E}[\nabla \sigma_\star(Z)]\, \mathbb{E}[g(T) - m_g] = 0$.

Consider the Gaussian path (Z_ρ, T) whose covariance increments with respect to ρ are

$$\Delta = \frac{\mathrm{d}}{\mathrm{d}\rho} \mathrm{Cov}\begin{pmatrix} Z_\rho \\ T \end{pmatrix} = \begin{bmatrix} 0 & \|w\| a \\ \|w\| a^\top & 0 \end{bmatrix}.$$

For any scalar $h \in C^2(\mathbb{R}^{r+1})$ with suitable integrability, the (matrix) Price identity (see the appendix in Rezende et al. [2014]) yields

$$\frac{\mathrm{d}}{\mathrm{d}\rho} \mathbb{E}[h(Z_\rho, T)] = \frac{1}{2} \Big\langle \Delta,\ \mathbb{E}\big[\nabla^2 h(Z_\rho, T)\big] \Big\rangle_F = \sum_{k=1}^{r} \|w\| a_k\, \mathbb{E}\big[\partial^2_{z_k t} h(Z_\rho, T)\big],$$

where we used that Δ is purely off–diagonal.

Apply this componentwise to $h_i(z, t) := \partial_{z_i} \sigma_\star(z)\, (g(t) - m_g)$ so that $F_i(\rho) = \mathbb{E}[h_i(Z_\rho, T)]$. Differentiating once,

$$F_i'(\rho) = \sum_{k=1}^{r} \|w\| a_k\, \mathbb{E}\big[\partial^2_{z_k t} h_i(Z_\rho, T)\big] = \|w\|\, a^\top\, \mathbb{E}\big[\nabla^2 \sigma_\star(Z_\rho)\, e_i\, g'(T)\big].$$

Vectorially,

$$F'(\rho) = \|w\|\, \mathbb{E}\big[\nabla^2 \sigma_\star(Z_\rho)\, a\, g'(T)\big].$$

At $\rho = 0$, $Z_0 \overset{d}{=} Z$ is independent of T, hence the factorization

$$F'(0) = \|w\|\, H_\star a\, \mathbb{E}\big[g'(\|w\|G)\big].$$

Differentiating once more along the same path. Using Price again on the scalar function $\tilde{h}_i(z,t) := a^\top \big(\nabla^2 \sigma_\star(z)\, e_i\big)\, g'(t)$, only the mixed derivatives contribute:

$$\partial^2_{z_k t} \tilde{h}_i(z,t) = a^\top \big(\nabla^3 \sigma_\star(z)[e_k, \cdot, e_i]\big)\, g''(t).$$

Therefore,

$$F_i''(\rho) = \sum_{k=1}^{r} \|w\| a_k\, \mathbb{E}\big[\partial^2_{z_k t} \tilde{h}_i(Z_\rho, T)\big] = \|w\|^2\, a^\top\, \mathbb{E}\big[\nabla^3 \sigma_\star(Z_\rho)[a, \cdot, e_i]\; g''(T)\big],$$

and, in vector form,

$$F''(\rho) = \|w\|^2\, \mathbb{E}\big[\nabla^3 \sigma_\star(Z_\rho)[a,a]\; g''(T)\big].$$

Taking norms and using $\|g''\|_\infty < \infty$,

$$\sup_{|\rho| \le \rho_0} \|F''(\rho)\| \le \|w\|^2\, \|g''\|_\infty\, \mathbb{E}\|\nabla^3 \sigma_\star(Z)\|,$$

since the marginals of Z_ρ and T do not depend on ρ. (The centered term $m_g\, \mathbb{E}[\nabla \sigma_\star(Z_\rho)]$ contributes no derivative because Δ is off–diagonal and $h(z,t) = \partial_{z_i} \sigma_\star(z)$ has zero z–t mixed Hessian.)

By Taylor with integral remainder,

$$F(\rho) = F(0) + \rho F'(0) + \frac{1}{2}\rho^2 F''(\xi_\rho) = \rho\, \|w\|\, H_\star a\, \mathbb{E}\big[g'(\|w\|G)\big] \;+\; O(\rho^2),$$

with the $O(\rho^2)$ uniform for $\|w\|$ in compact sets by the bound above. □

Lemma 19 (Centered Price expansion)**.** *Let $Z \sim \mathcal{N}(0, I_r)$, $G \sim \mathcal{N}(0,1)$ independent. Fix $w \neq 0$, put $T := \|w\|\, G$, and let $a \in \mathbb{R}^r$ be unit. For $|\rho| \le \rho_0 < 1$, let (Z_ρ, T) be jointly Gaussian with $\mathrm{Var}(Z_\rho) = I_r$, $\mathrm{Var}(T) = \|w\|^2$, and $\mathrm{Cov}(Z_\rho, T) = \rho\, \|w\|\, a$. Assume $\sigma_\star \in C^3(\mathbb{R}^r)$ with $\mathbb{E}\|\nabla^k \sigma_\star(Z)\| < \infty$ for $k = 1,2,3$, and $g \in C^2(\mathbb{R})$ with g', g'' bounded. Write*

$$H_\star := \mathbb{E}[\nabla^2 \sigma_\star(Z)], \qquad a_j := \mathbb{E}[g(T)].$$

Then, as $\rho \to 0$,

$$\mathbb{E}\big[\nabla \sigma_\star(Z_\rho)\, \big(g(T) - a_j\big)\big] = \rho\, \|w\|\, H_\star a\, \mathbb{E}\big[g'(\|w\|G)\big] \;+\; O(\rho^2),$$

uniformly for $\|w\|$ in compact sets.

Proof. Let $F(\rho) := \mathbb{E}[\nabla \sigma_\star(Z_\rho)\, g(T)]$. Then $F(0) = \mathbb{E}[\nabla \sigma_\star(Z)]\, \mathbb{E}[g(T)] =: \mu_\star\, a_j$. Matrix Price's identity along the covariance path with off–diagonal increment $\Delta = \left(\begin{smallmatrix} 0 & \|w\| a \\ \|w\| a^\top & 0 \end{smallmatrix}\right)$ gives

$$F'(\rho) = \|w\|\, \mathbb{E}[\nabla^2 \sigma_\star(Z_\rho)]\, a\; \mathbb{E}[g'(T)] \quad \Rightarrow \quad F'(0) = \|w\|\, H_\star a\, \mathbb{E}[g'(\|w\|G)].$$

A second Price differentiation shows $F''(\rho)$ is bounded by a constant times $\|w\|^2 \|g''\|_\infty\, \mathbb{E}\|\nabla^3 \sigma_\star(Z)\|$, hence Taylor:

$$F(\rho) - F(0) = \rho F'(0) + O(\rho^2).$$

Since $F(\rho) - F(0) = \mathbb{E}[\nabla \sigma_\star(Z_\rho)\, (g(T) - a_j)]$, the claim follows. □

Lemma 20 (Axial collapse with exchangeable weights)**.** *Let $h_k(x)$ be bounded, exchangeable weights and $\bar{h}(x) := \frac{1}{m} \sum_{k=1}^m h_k(x)$. Then, conditional on W_t,*

$$\mathbb{E}\left[\frac{1}{\sqrt{m}} \sum_{k=1}^{m} h_k(x)\, w_k \,\middle|\, W_t\right] = \sqrt{m}\, \mathbb{E}[\bar{h}(x) \mid W_t]\; \bar{w}.$$

Proof. By exchangeability, $\mathbb{E}[h_k(x) - \bar{h}(x) \mid W_t] = 0$ for each k. Hence

$$\mathbb{E}\left[\frac{1}{\sqrt{m}}\sum_k h_k(x)w_k \,\middle|\, W_t\right] = \frac{1}{\sqrt{m}}\sum_k \mathbb{E}[\bar{h}(x) \mid W_t]\, w_k = \sqrt{m}\,\mathbb{E}[\bar{h}(x) \mid W_t]\ \bar{w}. \qquad \square$$

Theorem 4 (SGD small-angle drift under AGOP)**.** *Assume the standing regime (Gaussian batches, $\eta \le \eta_0 p^{-1/2}$, fixed horizon), row-norm control, and AGOP. Let $u_t = \bar{w}_t/\|\bar{w}_t\|$, $c_t = \|P_\star u_t\| = \cos\theta_t$, and define*

$$S_0'(x) := \sum_{j=1}^m g(w_j^\top x), \qquad S_0''(x) := \sum_{j=1}^m g'(w_j^\top x)\, U^\top w_j.$$

Let $Z \sim \mathcal{N}(0, I_r)$, $G \sim \mathcal{N}(0,1)$, $H_\star := \mathbb{E}[\nabla^2 \sigma_\star(Z)]$, and set

$$\alpha_j := \|w_j\|\,\mathbb{E}\big[g'(\|w_j\|G)\big], \qquad a_\parallel := \frac{U^\top u_t}{\|U^\top u_t\|} \quad \text{when } \|U^\top u_t\| > 0.$$

Define the signed coefficients

$$\kappa_t^{\mathrm{S}} := a_\parallel^\top H_\star \Big(\sum_{j=1}^m \alpha_j\, U^\top \widehat{w}_j\Big), \qquad \kappa_t^{\mathrm{R}} := a_\parallel^\top\, \mathbb{E}\big[e(x)\, S_0''(x) \mid W_t\big], \qquad \kappa_t^{\mathrm{M}} := \Big(\sum_{j=1}^m g(w_j^\top x)\Big) a_\parallel^\top \mathbb{E}[\nabla\sigma_\star(Z)],$$

and put $\kappa_t := \kappa_t^{\mathrm{S}} + \kappa_t^{\mathrm{M}} + \kappa_t^{\mathrm{R}}$. Then, uniformly on the fixed horizon,

$$\mathbb{E}[\theta_{t+1} - \theta_t \mid W_t] \;=\; -\,\eta\,\frac{\kappa_t}{m\sqrt{m}}\,\theta_t \;+\; O\big(\eta\,\theta_t^3 + \eta^2 + p^{-1/2}\big),$$

with size bounds

$$|\kappa_t^{\mathrm{S}}| \;=\; O\Big(\frac{m}{\sqrt{p}}\Big), \qquad |\kappa_t^{\mathrm{R}}| \;=\; O\Big(\frac{m\,R_t}{\sqrt{p}}\Big), \qquad |\kappa_t^{\mathrm{M}}| \;=\; O(m).$$

In particular, the linear coefficient has magnitude $O(1/\sqrt{m}) + O\big((1+R_t)/(\sqrt{m}\,\sqrt{p})\big)$ under mean-isotropy condition, where $O(1/\sqrt{m})$ can be dropped whene $\mathbb{E}[\sigma_\star] = 0$.

(Fix t and condition on W_t. We write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid W_t]$ and, when no subscript is shown, $\mathbb{E}[\cdot]$ also denotes $\mathbb{E}_t[\cdot]$.)

Proof. In the following, we provide the proof of the main theorem on the rate of feature learning for SGD. We show the following steps:

Step 1 (Angle differential and measurability). Let $\Delta\bar{w}_t := \bar{w}_{t+1} - \bar{w}_t$. By Lemma 15,

$$\theta_{t+1} - \theta_t = -\frac{v_t^\top \Delta\bar{w}_t}{\|\bar{w}_t\|} + O\Big(\frac{\|\Delta\bar{w}_t\|^2}{\|\bar{w}_t\|^2}\Big).$$

Both $\bar{w}_t$ and v_t are W_t-measurable (they are deterministic functions of the weights at time t), hence taking conditional expectation,

$$\mathbb{E}[\theta_{t+1} - \theta_t \mid W_t] = -\frac{v_t^\top \mathbb{E}[\Delta\bar{w}_t \mid W_t]}{\|\bar{w}_t\|} + O\Big(\mathbb{E}[\|\Delta\bar{w}_t\|^2 \mid W_t]/\|\bar{w}_t\|^2\Big).$$

With the realized SGD step $\Delta\bar{w}_t = -(\eta/(m\sqrt{m}))\sum_{j=1}^m e(x)\, g(w_j^\top x)\, x$ and $|g| \le G_1$,

$$\mathbb{E}[\|\Delta\bar{w}_t\|^2 \mid W_t] \;\le\; \frac{\eta^2 G_1^2}{m}\,\mathbb{E}\big[e(x)^2\,\|x\|^2 \mid W_t\big].$$

By Lemma 5, there is a constant C (depending only on bounded moments/derivatives in Assumption 4) with $\mathbb{E}[e(x)^2\|x\|^2 \mid W_t] \le Cp$, hence

$$\mathbb{E}[\|\Delta\bar{w}_t\|^2 \mid W_t] \;\le\; \eta^2\,\frac{CG_1^2}{m}\,p.$$

If $m/p \to \gamma \in (0,\infty)$, this is $O(\eta^2)$. Since $\mathbb{E}[\|\Delta\bar{w}_t\|] \lesssim \eta$, we also have $\|\bar{w}_t\| \geq \|\bar{w}_0\| - O(T\eta) \geq c_w > 0$. Thus the Step 1 remainder is $O(\eta^2)$.

Step 2 (Teacher–projected identity; student–student piece). By Lemma 17,

$$U^\top \bar{h}_t = -\frac{\eta}{m\sqrt{m}}\,\mathbb{E}\Big[(\nabla_Z f_W(Z) - \nabla\sigma_\star(Z))\,S_0'(x)\Big] - \frac{\eta}{m\sqrt{m}}\,\mathbb{E}\big[e(x)\,S_0''(x)\big]. \tag{10}$$

Moreover, since $\nabla_Z(w_k^\top x) = U^\top w_k$,

$$\nabla_Z f_W(Z) = \frac{1}{\sqrt{m}}\sum_{k=1}^m g(w_k^\top x)\,U^\top w_k = U^\top\left(\frac{1}{\sqrt{m}}\sum_{k=1}^m g(w_k^\top x)\,w_k\right),$$

so

$$\mathbb{E}[\nabla_Z f_W(Z)\,S_0'(x)] = U^\top\,\mathbb{E}\Big[\Big(\tfrac{1}{\sqrt{m}}\sum_{k=1}^m g(w_k^\top x)\,w_k\Big)\,S_0'(x)\Big].$$

Define

$$\kappa_t^{\mathrm{SS}} := a_\parallel^\top\,\mathbb{E}[\nabla_Z f_W(Z)\,S_0'(x)] = a_\parallel^\top U^\top\,\mathbb{E}\Big[\Big(\tfrac{1}{\sqrt{m}}\sum_{k=1}^m g(w_k^\top x)\,w_k\Big)\,S_0'(x)\Big].$$

Since $|g|_\infty \leq G_1$ and $|S_0'(x)| \leq mG_1$, the vector $\mathbb{E}[(\frac{1}{\sqrt{m}}\sum_k g(w_k^\top x)w_k)S_0'(x)]$ is a bounded deterministic linear combination of $\{w_k\}$. Using Lemma 16 with deterministic bounded coefficients to conclude

$$\begin{aligned}\big\|U^\top\mathbb{E}[(\frac{1}{\sqrt{m}}\sum_k g(w_k^\top x)w_k)S_0'(x)]\big\| &= \big\|U^\top\mathbb{E}[(\frac{1}{\sqrt{m}}\sum_k g(w_k^\top x)S_0'(x)||w_k||\widehat{w}_k)]\big\| \\ &= O(m/\sqrt{p}),\end{aligned}$$

hence $|\kappa_t^{\mathrm{SS}}| = O(m/\sqrt{p})$ (this could be improved to $O(\sqrt{m}/\sqrt{p})$ under additional dispersion assumptions). Its contribution to the angle drift is thus $O(\eta/\sqrt{mp})$.

Step 3 (Teacher–student Price term and residual bound). Fix $j \in \{1,\dots,m\}$. Write

$$r_j := \|w_j\| \in [c_-, c_+], \qquad \widehat{w}_j := \frac{w_j}{r_j} \in \mathbb{S}^{p-1}.$$

Project to the teacher space via $U \in \mathbb{R}^{p\times r}$, and define the (dimensionless) alignment

$$\rho_j := \|U^\top\widehat{w}_j\| \in [0,1], \qquad u_j := \begin{cases} \dfrac{U^\top\widehat{w}_j}{\rho_j} \in \mathbb{S}^{r-1}, & \rho_j > 0, \\ \text{any fixed unit vector in } \mathbb{R}^r, & \rho_j = 0. \end{cases}$$

Thus $U^\top w_j = r_j\,U^\top\widehat{w}_j = \rho_j\,r_j\,u_j$. For $x \sim \mathcal{N}(0, I_p)$ set

$$Z := U^\top x \in \mathbb{R}^r, \qquad T_j := w_j^\top x = r_j\,(\widehat{w}_j^\top x) \in \mathbb{R}.$$

Then (Z, T_j) is centered Gaussian with

$$\mathrm{Var}(Z) = I_r, \qquad \mathrm{Var}(T_j) = r_j^2, \qquad \mathrm{Cov}(Z, T_j) = U^\top w_j = \rho_j\,r_j\,u_j.$$

Now, define

$$m_{g,j} := \mathbb{E}\big[g(r_j G)\big], \qquad \alpha_j := r_j\,\mathbb{E}\big[g'(r_j G)\big],$$

so that, by Lemma 18,

$$\mathbb{E}\big[\nabla\sigma_\star(Z)\,g(T_j)\big] = m_{g,j}\,\mathbb{E}[\nabla\sigma_\star(Z)] \;+\; \alpha_j\,H_\star\,(U^\top\widehat{w}_j) \;+\; r_j^{(\mathrm{cen})}, \qquad \|r_j^{(\mathrm{cen})}\| \leq C\,\rho_j^2,$$

with C depending only on the bounds in Lemma 18, uniformly for $r_j \in [c_-, c_+]$. Summing over j,

$$\mathbb{E}\big[\nabla\sigma_\star(Z)\, S_0'(x)\big] = \mathbb{E}[\nabla\sigma_\star(Z)]\,\Big(\sum_{j=1}^m m_{g,j}\Big) + \sum_{j=1}^m \alpha_j\, H_\star(U^\top \widehat{w}_j) + O\Big(\sum_{j=1}^m \rho_j^2\Big), \tag{11}$$

where the $O(\sum_{j=1}^m \rho_j^2)$ is shorthand for the vector sum of the residuals $r_j^{(\mathrm{cen})}$. Now, using mean–isotropy from Assumption 12,

$$\sum_{j=1}^m \rho_j^2 = \sum_{j=1}^m \widehat{w}_j^\top P_\star \widehat{w}_j = \Big\langle \sum_{j=1}^m \widehat{w}_j \widehat{w}_j^\top,\ P_\star \Big\rangle_F = O\Big(\frac{mr}{p}\Big),$$

so $\big\| \sum_{j=1}^m r_j^{(\mathrm{cen})} \big\| = O(mr/p)$.

Furthermore, Lemma 16 with bounded deterministic weights α_j gives

$$\Big\| \sum_{j=1}^m \alpha_j\, U^\top \widehat{w}_j \Big\| = O\Big(\frac{m}{\sqrt{p}}\Big).$$

Now, projecting onto $a_\parallel$ gives

$$\kappa_t^{\mathrm{S}} := a_\parallel^\top H_\star\Big(\sum_{j=1}^m \alpha_j\, U^\top \widehat{w}_j\Big) = O\Big(\frac{m}{\sqrt{p}}\Big), \quad \text{since } \|a_\parallel\| = 1,\ \|H_\star\|_{op} < \infty.$$

Similarly, denote the first term in Eq. (11) by

$$\kappa_t^{\mathrm{M}} := \Big(\sum_{j=1}^m m_{g,j}\Big) a_\parallel^\top \mathbb{E}[\nabla\sigma_\star(Z)]$$

Since $g = \phi'$ is bounded and $\|w_j\| \in [c_-, c_+]$, we have

$$|m_{g,j}| \le \|g\|_\infty \quad \Longrightarrow \quad \Big|\sum_{j=1}^m m_{g,j}\Big| \le m\,\|g\|_\infty.$$

By our teacher smoothness assumptions, set $C_\star := \|\mathbb{E}[\nabla\sigma_\star(Z)]\| < \infty$ and note $\|a_\parallel\| = 1$. Hence

$$|\kappa_t^{\mathrm{M}}| = \Big|\sum_{j=1}^m m_{g,j}\Big|\,\big|a_\parallel^\top \mathbb{E}[\nabla\sigma_\star(Z)]\big| \ \le\ m\,\|g\|_\infty\,\|\mathbb{E}[\nabla\sigma_\star(Z)]\| = O(m).$$

For the residual in Eq. (10),

$$\kappa_t^{\mathrm{R}} = a_\parallel^\top\, \mathbb{E}\big[e(x)\, S_0''(x)\big] = \sum_{j=1}^m \mathbb{E}[e(x)\, g'(w_j^\top x)] \cdot a_\parallel^\top U^\top w_j.$$

Now, observe that $|a_\parallel^\top U^\top w_j| \le r_j\,|(Ua_\parallel)^\top \widehat{w}_j| \le c_+\,|(Ua_\parallel)^\top \widehat{w}_j|$. By Cauchy-Schwarz inequality and bounded g', $|\mathbb{E}[e\, g'(w_j^\top x)]| \le \|g'\|_\infty\, R_t$, where $R_t = \sqrt{\mathbb{E}[e(x)^2]}$. Alternately,

$$|\kappa_t^{\mathrm{R}}| \le c_+\|g'\|_\infty\, R_t \sum_{j=1}^m |(Ua_\parallel)^\top \widehat{w}_j| \le c_+\|g'\|_\infty\, R_t\, \sqrt{m}\Big(\sum_{j=1}^m ((Ua_\parallel)^\top \widehat{w}_j)^2\Big)^{1/2}.$$

Now, using mean-isotropy condition (cf. Assumption 12 or Lemma 10, Corollary 1) we have $\frac{1}{m}\sum_j ((Ua_\parallel)^\top \widehat{w}_j)^2 = \frac{1}{p} + O(1/p)$, and hence $\sum_j ((Ua_\parallel)^\top \widehat{w}_j)^2 = O(m/p)$ and $|\kappa_t^{\mathrm{R}}| = O(mR_t/\sqrt{p})$.

Step 4 (Final rate). From Step 1,

$$\mathbb{E}[\theta_{t+1} - \theta_t \mid W_t] = -\frac{1}{\|\bar{w}_t\|}\, v_t^\top \mathbb{E}[\Delta \bar{w}_t \mid W_t] \;+\; O(\eta^2).$$

Since $v_t \in \mathrm{im}(P_\star)$, $v_t^\top \mathbb{E}[\Delta \bar{w}_t \mid W_t] = v_t^\top P_\star\, \bar{h}_t$ where $\bar{h}_t := \mathbb{E}[\Delta \bar{w}_t \mid W_t]$. We now relate $v_t^\top P_\star h$ to $a_\parallel^\top U^\top h$ (for any $h \in \mathbb{R}^p$). First, we note the following for ease of clarity:

$$v_t = \frac{(I - u_t u_t^\top) P_\star u_t}{\|(I - u_t u_t^\top) P_\star u_t\|}, \qquad P_\star u_t = \|U^\top u_t\|\, U a_\parallel = \cos\theta_t\, U a_\parallel,$$

so a direct calculation gives the identity

$$v_t^\top P_\star h \;=\; \sin\theta_t\; a_\parallel^\top U^\top h \quad \text{for all } h \in \mathbb{R}^p.$$

Therefore,

$$v_t^\top P_\star \bar{h}_t = \sin\theta_t\; a_\parallel^\top U^\top \bar{h}_t.$$

Insert the teacher-projected mean-step identity from Step 2,

$$U^\top \bar{h}_t = -\frac{\eta}{m\sqrt{m}}\Big(\mathbb{E}[\nabla_Z f_W(Z)\, S_0'(x)] - \mathbb{E}[\nabla \sigma_\star(Z)\, S_0'(x)] + \mathbb{E}[e(x)\, S_0''(x)]\Big),$$

to obtain

$$-\frac{1}{\|\bar{w}_t\|}\, v_t^\top P_\star \bar{h}_t = \frac{\eta \sin\theta_t}{\|\bar{w}_t\|\, m\sqrt{m}} \Big(\underbrace{a_\parallel^\top \mathbb{E}[\nabla_Z f_W\, S_0']}_{\text{(SS)}} - \underbrace{a_\parallel^\top \mathbb{E}[\nabla \sigma_\star\, S_0']}_{\text{(S)}} + \underbrace{a_\parallel^\top \mathbb{E}[e\, S_0'']}_{\text{(R)}} \Big).$$

We now plug in the bounds from Step 3 and the definition of κ_t:

(S) By the Price expansion in Lemma 18 and mean-isotropy (Lemma 10),

$$a_\parallel^\top \mathbb{E}[\nabla \sigma_\star\, S_0'] = a_\parallel^\top \Big(\sum_{j=1}^m m_{g,j}\Big)\mathbb{E}[\nabla \sigma_\star(Z)] + a_\parallel^\top H_\star \Big(\sum_{j=1}^m \alpha_j\, U^\top \widehat{w}_j\Big) + O(mr/p) = \kappa_t^{\mathrm{S}} + \kappa_t^{\mathrm{M}} + O(mr/p).$$

(R) By Cauchy–Schwarz and mean-isotropy, $a_\parallel^\top \mathbb{E}[e\, S_0''] = \kappa_t^{\mathrm{R}}, \quad |\kappa_t^{\mathrm{R}}| = O(mR_t/\sqrt{p})$.

(SS) The purely student piece satisfies $a_\parallel^\top \mathbb{E}[\nabla_Z f_W\, S_0'] = O(m/\sqrt{p})$ (by Lemma 16 with bounded deterministic coefficients).

Putting these together,

$$-\frac{1}{\|\bar{w}_t\|}\, v_t^\top P_\star \bar{h}_t = \frac{\eta \sin\theta_t}{\|\bar{w}_t\|\, m\sqrt{m}} \Big(\underbrace{-\kappa_t^{\mathrm{S}} - \kappa_t^{\mathrm{M}} - \kappa_t^{\mathrm{R}}}_{=-\kappa_t} + O(mr/p) \;+\; O(m/\sqrt{p}) \;+\; O(m/p) \Big).$$

On the fixed horizon we have $\|\bar{w}_t\| \geq c_w > 0$, so $\|\bar{w}_t\|^{-1} = O(1)$. Moreover, $\sin\theta_t = \theta_t + O(\theta_t^3)$. Hence,

$$\mathbb{E}[\theta_{t+1} - \theta_t \mid W_t] = -\,\eta\, \frac{\kappa_t}{m\sqrt{m}}\, \theta_t \;+\; O\Big(\eta\, \theta_t^3\Big) \;+\; O\Big(\frac{\eta\, \theta_t}{m\sqrt{m}}\, \frac{m}{\sqrt{p}}\Big) \;+\; O\Big(\frac{\eta\, \theta_t}{m\sqrt{m}}\, \frac{mr}{p}\Big) \;+\; O(\eta^2).$$

Since $\eta \leq \eta_0 p^{-1/2}$ and (if assumed) $m/p \to \gamma \in (0, \infty)$,

$$\frac{\eta}{m\sqrt{m}}\, \frac{m}{\sqrt{p}} = O\Big(\frac{1}{p}\Big), \qquad \frac{\eta}{m\sqrt{m}}\, \frac{mr}{p} = O\Big(\frac{1}{p^{1/2}}\Big),$$

so both error terms above are $\leq O(p^{-1/2})$ (uniformly in $\theta_t \leq \pi/2$). Therefore

$$\mathbb{E}[\theta_{t+1} - \theta_t \mid W_t] \;=\; -\,\eta\, \frac{\kappa_t}{m\sqrt{m}}\, \theta_t \;+\; O\big(\eta\, \theta_t^3 + \eta^2 + p^{-1/2}\big),$$

where $\kappa_t = \kappa_t^{\mathrm{S}} + \kappa_t^{\mathrm{M}} + \kappa_t^{\mathrm{R}}$. This proves the theorem and the size bounds on $\kappa_t^{\mathrm{S}}, \kappa_t^{\mathrm{R}}$, and κ_t^{M}. □

6 MUON

In this section, we analyze the rate of feature learning for the Muon algorithm, providing a counterpart to the SGD analysis in Theorem 4. We operate under the same standing assumptions as the previous section, including Gaussian data, row-norm control for the weights, and smoothness of the activation's derivative $g = \phi'$.

For the Muon-specific setup, we consider a minibatch of size B drawn i.i.d. from $\mathcal{N}(0, I_p)$, denoted by the matrix $X = [x_1, \dots, x_B] \in \mathbb{R}^{p\times B}$. The algorithm's update depends on the batch entries $c_{jb} := e_t(x_b)\, g(w_j^\top x_b)$, where $e_t(x)$ is the prediction error at step t. We work in the proportional scaling regime where $m/p \to \gamma \in (0, \infty)$, with a stepsize[1] $\eta \lesssim p^{-1/2}$.

Lemma 21 (Rayleigh with batch–span overlap)**.** *Let Π be the orthogonal projector onto $\operatorname{span}\{x_b\}_{b\le B}$ and assume the batch–span sandwich $\kappa_-\,\lambda\,\Pi \preceq M \preceq \kappa_+\,\lambda\,\Pi$ on $\operatorname{im}(\Pi)$. Then for any unit $a \in \mathbb{R}^p$,*

$$(\kappa_+\,\lambda)^{-1/2}\,\|\Pi a\|^2 \;\le\; a^\top M^{-1/2} a \;\le\; (\kappa_-\,\lambda)^{-1/2}\,\|\Pi a\|^2.$$

Proof. Since $M^{-1/2} = (M|_{\operatorname{im}(\Pi)})^{-1/2}\Pi$, $a^\top M^{-1/2} a = (\Pi a)^\top (M|_{\operatorname{im}(\Pi)})^{-1/2}(\Pi a)$. The eigenvalues of $M|_{\operatorname{im}(\Pi)}$ lie in $[\kappa_-\lambda,\ \kappa_+\lambda]$. Applying the Rayleigh quotient bounds yields the desired inequalities. □

Lemma 22 (Random subspace overlap; lower tail)**.** *Let $x_1, \dots, x_B \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p)$, let Π be the orthogonal projector onto $\operatorname{span}\{x_b\}_{b=1}^B \subset \mathbb{R}^p$, and fix any unit vector $a \in \mathbb{R}^p$. Then for any $\varepsilon \in (0, 1)$, there exists a universal $c > 0$ such that*

$$\mathrm{P}\big(\|\Pi a\|^2 \ge (1-\varepsilon)\frac{B}{p}\big) \ge 1 - 2\exp(-c\varepsilon^2 B).$$

(An analogous upper tail holds with the same form.)

Proof. By rotational invariance of the Gaussian distribution, the random B-dimensional subspace $\operatorname{span}\{x_b\}$ is uniformly distributed over the Grassmannian manifold. For any fixed unit vector a, the squared projection length $\|\Pi a\|^2$ follows a Beta distribution:

$$\|\Pi a\|^2 \sim \operatorname{Beta}\left(\frac{B}{2}, \frac{p-B}{2}\right).$$

The mean is $\mathbb{E}[\|\Pi a\|^2] = B/p$.

Using known concentration results for Beta distributions (see Vershynin [2018]), there exists $c > 0$ such that for any $\varepsilon \in (0, 1)$:

$$\mathrm{P}\big(\|\Pi a\|^2 \le (1-\varepsilon)\frac{B}{p}\big) \le \exp(-c\varepsilon^2 B).$$

Taking complements gives the lower tail bound. The upper tail follows similarly from

$$\mathrm{P}\big(\|\Pi a\|^2 \ge (1+\varepsilon)\frac{B}{p}\big) \le \exp(-c\varepsilon^2 B).$$

A union bound gives the two-sided version with constant 4. □

Lemma 23 (Numerator alignment / teacher–overlap)**.** *Let $x \sim \mathcal{N}(0, I_p)$. Fix $W_t = (w_1, \dots, w_m)$ with row–norm control $\|w_j\| \in [c_-, c_+]$, and assume the AGOP/mean-isotropy condition as stated in Assumption 12:*

$$\Big\|\frac{1}{m}\sum_{j=1}^m \widehat{w}_j \widehat{w}_j^\top - \frac{I_p}{p}\Big\|_{op} \;\le\; \frac{C_{\text{ag}}}{p}, \qquad \widehat{w}_j := \frac{w_j}{\|w_j\|}.$$

[1] Note that the rate is strictly smaller than one appeared for mean-isotropy Theorem 2.

Define $f_W(x) = m^{-1/2}\sum_{j=1}^m \phi(w_j^\top x)$, $g = \phi'$, $e(x) := f_W(x) - \sigma_\star(U^\top x)$,

$$R_t := \sqrt{\mathbb{E}[e(x)^2 \mid W_t]}, \qquad S_0(x) := \frac{1}{\sqrt{m}}\sum_{j=1}^m g(w_j^\top x), \qquad u := \mathbb{E}\big[x\, e(x)\, S_0(x) \mid W_t\big] \in \mathbb{R}^p.$$

Assume $g \in C^2(\mathbb{R})$ *with* $\|g\|_\infty, \|g'\|_\infty, \|g''\|_\infty < \infty$. *Let* $u_t := \bar{w}_t/\|\bar{w}_t\|$ *and* $a_\parallel := \frac{U^\top u_t}{\|U^\top u_t\|}$ *when* $\|U^\top u_t\| > 0$. *Then there exist scalars* A_t', B_t' *and a remainder* $r_t \in \mathbb{R}^p$ *such that*

$$u \;=\; m\, A_t'\, \bar{w}_t \;-\; m\, B_t'\, P_\star \bar{w}_t \;+\; r_t,$$

with the bounds

$$|A_t'| \;\le\; C_A\, \frac{1+R_t}{\sqrt{p}}, \qquad |B_t'| \;\le\; C_B\, \frac{1}{\sqrt{p}}, \qquad \|r_t\| \;\le\; C_r\, \frac{m}{\sqrt{p}}.$$

where the constants C_A, C_B, C_r *depend only on* $\big(\|g\|_\infty, \|g'\|_\infty, \|g''\|_\infty, c_\pm, \|H_\star\|_{op}, C_{\mathrm{ag}}\big)$ *and* $H_\star := \mathbb{E}[\nabla^2 \sigma_\star(Z)]$ *for* $Z \sim \mathcal{N}(0, I_r)$. *Moreover,* B_t' *can be chosen as the* signed teacher–overlap

$$m\, B_t'\, \|U^\top \bar{w}_t\| \;:=\; a_\parallel^\top U^\top u,$$

so that the teacher contribution in the decomposition is exactly the component of u *along* $P_\star \bar{w}_t$.

Proof. All expectations are conditional on W_t. Put $Z := U^\top x$. Applying Gaussian Stein identity in Lemma 14 to the scalar $\psi(x) := e(x)S_0(x)$:

$$u = \mathbb{E}[x\psi(x)] = \mathbb{E}[\nabla_x \psi(x)] = \mathbb{E}[(\nabla_x e) S_0(x)] + \mathbb{E}[e(x)\, \nabla_x S_0(x)].$$

Using $\nabla_x e = \frac{1}{\sqrt{m}}\sum_k g(w_k^\top x)\, w_k - U\, \nabla_Z \sigma_\star(Z)$ and $\nabla_x S_0 = \frac{1}{\sqrt{m}}\sum_j g'(w_j^\top x)\, w_j$, we obtain

$$u \;=\; \underbrace{\mathbb{E}\Big[\Big(\tfrac{1}{\sqrt{m}}\sum_k g(w_k^\top x)\, w_k\Big) S_0(x)\Big]}_{\text{(S1)}} \;-\; \underbrace{U\, \mathbb{E}[\nabla_Z \sigma_\star(Z)\, S_0(x)]}_{\text{(T)}} \;+\; \underbrace{\tfrac{1}{\sqrt{m}}\sum_{j=1}^m \mathbb{E}[e(x)\, g'(w_j^\top x)]\, w_j}_{\text{(S2)}}. \tag{12}$$

A. Student blocks (S1) + (S2)**.** Define bounded coefficients

$$\psi_j \;:=\; \frac{1}{\sqrt{m}}\Big(\mathbb{E}\big[g(w_j^\top x)\, S_0(x)\big] \;+\; \mathbb{E}\big[e(x)\, g'(w_j^\top x)\big]\Big) \quad\Rightarrow\quad \text{(S1)} + \text{(S2)} \;=\; \sum_{j=1}^m \psi_j\, w_j.$$

Uniformly, $|\psi_j| \le \frac{1}{\sqrt{m}}\Big(C_0 + \|g'\|_\infty R_t\Big) =: C_\psi/\sqrt{m}$ by Cauchy–Schwarz and boundedness of g, g'. Write $w_j = r_j \widehat{w}_j$. By the AGOP axial Lemma 16 (deterministic bounded coefficients),

$$\sum_{j=1}^m \psi_j r_j\, \widehat{w}_j = \Big(\sum_{j=1}^m \psi_j r_j\, u_t^\top \widehat{w}_j\Big) u_t + r_\perp, \qquad \|r_\perp\| \;\le\; C\, \frac{m}{\sqrt{p}}.$$

Define

$$m\, A_t'\, \bar{w}_t \;:=\; \Big(\sum_{j=1}^m \psi_j r_j\, u_t^\top \widehat{w}_j\Big) u_t \quad\Longleftrightarrow\quad A_t' \;=\; \frac{1}{m\|\bar{w}_t\|}\sum_{j=1}^m \psi_j\, u_t^\top w_j.$$

Then (S1) + (S2) $= mA_t'\bar{w}_t + r_\perp$. Using Cauchy–Schwarz and $\frac{1}{m}\sum_j (u_t^\top \widehat{w}_j)^2 = \frac{1}{p} + O(\frac{1}{p})$ (mean-isotropy),

$$|A_t'| \;\le\; \frac{1}{m\|\bar{w}_t\|}\Big(\sum_j \psi_j^2 r_j^2\Big)^{1/2}\Big(\sum_j (u_t^\top \widehat{w}_j)^2\Big)^{1/2} \;\lesssim\; \frac{\sqrt{1+mR_t^2}}{m}\cdot\sqrt{\frac{m}{p}} \;\le\; C_A\, \frac{1+R_t}{\sqrt{p}}.$$

Set $r_{\mathrm{S}} := r_\perp$; then

$$(\mathrm{S1}) + (\mathrm{S2}) \;=\; m A'_t \bar{w}_t \;+\; r_{\mathrm{S}}, \qquad |A'_t| \le C_A \,\frac{1+R_t}{\sqrt{p}}, \quad \|r_{\mathrm{S}}\| \le C\,\frac{m}{\sqrt{p}}. \tag{13}$$

B. Teacher block (T)**.** We have

$$(\mathrm{T}) \;=\; U\,\mathbb{E}[\nabla_Z \sigma_\star(Z)\, S_0(x)] = \;\frac{1}{\sqrt{m}}\, U \sum_{j=1}^{m} \mathbb{E}\big[\nabla_Z \sigma_\star(Z)\, g(w_j^\top x)\big].$$

For a fixed j, consdier the following notation:

$$r_j := \|w_j\| \in [c_-, c_+], \quad \widehat{w}_j := \frac{w_j}{r_j}, \quad Z := U^\top x, \quad T_j := w_j^\top x, \quad \rho_j := \|U^\top \widehat{w}_j\|, \quad u_j := \begin{cases} \dfrac{U^\top \widehat{w}_j}{\rho_j}, & \rho_j > 0, \\ \text{any unit in } \mathbb{R}^r, & \rho_j = 0. \end{cases}$$

Let

$$m_{g,j} := \mathbb{E}\big[g(r_j G)\big], \qquad \alpha_j := r_j\,\mathbb{E}\big[g'(r_j G)\big], \qquad H_\star := \mathbb{E}[\nabla^2 \sigma_\star(Z)], \qquad \mu_\star := \mathbb{E}[\nabla \sigma_\star(Z)].$$

The centered two–variable Price expansion (Lemma 18) gives

$$\mathbb{E}\big[\nabla_Z \sigma_\star(Z)\, g(T_j)\big] \;=\; \underbrace{m_{g,j}\,\mu_\star}_{\text{zeroth order}} \;+\; \underbrace{\alpha_j\, H_\star\,(U^\top \widehat{w}_j)}_{\text{linear in } \rho_j} \;+\; \underbrace{r_j^{(\mathrm{cen})}}_{\|r_j^{(\mathrm{cen})}\| \le C\,\rho_j^2}.$$

Therefore

$$(\mathrm{T}) = U\,\mathbb{E}[\nabla_Z \sigma_\star(Z)\, S_0] = \frac{1}{\sqrt{m}}\, U \sum_{j=1}^{m} \Big(m_{g,j}\mu_\star + \alpha_j H_\star(U^\top \widehat{w}_j) + r_j^{(\mathrm{cen})} \Big).$$

Now, given that U is an isometry for left multiplication, we can drop it in norm bounds:

$$\Big\| \sum_{j=1}^{m} m_{g,j}\mu_\star \Big\| \le m\,\|g\|_\infty\,\|\mu_\star\|, \quad \Big\| \sum_{j=1}^{m} \alpha_j\, U^\top \widehat{w}_j \Big\| = O\Big(\frac{m}{\sqrt{p}}\Big), \quad \sum_{j=1}^{m} \|r_j^{(\mathrm{cen})}\| = O\Big(\frac{mr}{p}\Big),$$

hence

$$\|(\mathrm{T})\| \;\le\; C_0\,\sqrt{m} \;+\; C_1\,\frac{\sqrt{m}}{\sqrt{p}} \;+\; C_2\,\frac{\sqrt{m}\,r}{p} \;=\; O(\sqrt{m}).$$

Decompose the teacher block along the teacher direction of $\bar{w}_t$:

$$\begin{aligned} (\mathrm{T}) &= m\,B'_t\,P_\star \bar{w}_t + r_{\mathrm{T}}, \\ r_{\mathrm{T}} &\perp P_\star \bar{w}_t. \end{aligned} \tag{14}$$

where the signed coefficient B'_t is defined by

$$m\,B'_t\,\|U^\top \bar{w}_t\| \;:=\; a_\parallel^\top U^\top(\mathrm{T}), \qquad a_\parallel := \frac{U^\top u_t}{\|U^\top u_t\|}.$$

By construction $r_{\mathrm{T}} := (I - P_{P_\star \bar{w}_t})(\mathrm{T}) \in \mathrm{im}(U)$ and $r_{\mathrm{T}} \perp P_\star \bar{w}_t$. From the bound on (T) above we get

$$|B'_t| = \frac{|a_\parallel^\top U^\top(\mathrm{T})|}{m\,\|U^\top \bar{w}_t\|} \;\le\; \frac{\|(\mathrm{T})\|}{m\,\|U^\top \bar{w}_t\|} \;\le\; \frac{C}{\sqrt{m}} \quad \text{(and hence } |B'_t| \le C_\gamma/\sqrt{p} \text{ if } m/p \to \gamma),$$

and

$$\|r_{\mathrm{T}}\| = \|(I - P_{P_\star \bar{w}_t})(\mathrm{T})\| \;\le\; \|(\mathrm{T})\| \;\le\; O(\sqrt{m}).$$

C. Assemble. From the student decomposition Eq. (13) we have $(\mathrm{S1}) + (\mathrm{S2}) = mA'_t\bar{w}_t + r_{\mathrm{S}}$ with $|A'_t| \le C_A\frac{1+R_t}{\sqrt{p}}$ and $\|r_{\mathrm{S}}\| \le C\frac{\sqrt{m}}{\sqrt{p}}$. Subtract the teacher decomposition Eq. (14) to get

$$u \;=\; \big((\mathrm{S1}) + (\mathrm{S2})\big) \;-\; (\mathrm{T}) \;=\; mA'_t\bar{w}_t \;-\; mB'_tP_\star\bar{w}_t \;+\; (r_{\mathrm{S}} - r_{\mathrm{T}}).$$

Define the total remainder $r_t := r_{\mathrm{S}} - r_{\mathrm{T}}$. Then

$$|A'_t| \;\le\; C_A\,\frac{1+R_t}{\sqrt{p}}, \qquad |B'_t| \;\le\; \frac{C}{\sqrt{m}} \;\; (\le C_\gamma/\sqrt{p} \text{ if } m/p \to \gamma), \qquad \|r_t\| \;\le\; C\Big(\frac{\sqrt{m}}{\sqrt{p}} + \sqrt{m}\Big).$$

In the proportional regime $m/p \to \gamma \in (0,\infty)$ the bound simplifies to $\|r_t\| \le C_\gamma\,\frac{m}{\sqrt{p}}$, which is the form stated in the lemma.

□

Theorem 5 (Muon/right–polar small-angle drift). *Let* $C = [c_1, \dots, c_B] \in \mathbb{R}^{m\times B}$, $\overline{G} = \frac{1}{B} C X^\top \in \mathbb{R}^{m\times p}$, $M = \overline{G}^\top \overline{G}$, *and* $u = \overline{G}^\top \mathbf{1}_m$. *With* $S_0(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^m g(w_j^\top x)$ *we have* $\mathbb{E}[u \mid W_t] = \sqrt{m}\, u^{\mathrm{num}}$ *with* $u^{\mathrm{num}} := \mathbb{E}[x\, e_t(x)\, S_0(x) \mid W_t]$. *Let* Π *be the projector onto* $\mathrm{span}\{x_b\}_{b\le B}$.

Batch–span conditioning. *On a good event* $\mathcal{E}$ *with* $\mathbb{P}(\mathcal{E} \mid W_t) = 1 - o(1)$,

$$\kappa_-\, \lambda\, \Pi \;\preceq\; M \;\preceq\; \kappa_+\, \lambda\, \Pi, \qquad \lambda := \frac{p R_t^2}{mB}.$$

AGOP/mean-isotropy. *Assume* $\big\| \frac{1}{m} \sum_{j=1}^m \widehat{w}_j \widehat{w}_j^\top - \frac{I_p}{p} \big\|_{op} \le C_{\mathrm{ag}}/p$, $\widehat{w}_j = w_j/\|w_j\|$.

Let $\bar{w}_t = \frac{1}{\sqrt{m}} \sum_{j=1}^m w_{j,t}$, $u_t = \bar{w}_t/\|\bar{w}_t\|$, θ_t *the angle with* $P_\star = UU^\top$ *via* $\|P_\star u_t\| = \cos\theta_t$, *and* $a_\parallel := \frac{P_\star \bar{w}_t}{\|P_\star \bar{w}_t\|}$ *when* $\|P_\star \bar{w}_t\| > 0$.

Then for the (conditional mean) Muon/right–polar step

$$\bar{h}_t \;:=\; -\frac{\eta}{m}\, Q^\top \mathbf{1}_m \;=\; -\frac{\eta}{m}\, M^{-1/2} u, \qquad Q := \overline{G}\, M^{-1/2},$$

we have, uniformly on the fixed horizon,

$$\Big| \mathbb{E}[\theta_{t+1} - \theta_t \mid W_t] \Big| \;\le\; \eta\, K_{\mathrm{M}}^{(\mathrm{size})}\, \theta_t \;+\; O\big(\eta\, \theta_t^3 + \eta^2\big),$$

with the (unsigned) linear coefficient

$$K_{\mathrm{M}}^{(\mathrm{size})} \;\lesssim\; \frac{B^{3/2}}{pR_t} + \frac{\sqrt{B}}{R_t},$$

where the hidden constant depends only on $(\phi, \sigma_\star, \gamma, c_\pm, \kappa_\pm, C_{\mathrm{ag}})$.

Proof. In the following, we provide the proof of the main theorem on the rate of feature learning for MUON. We show the following steps:

Step 1 (Angle differential). With $v_t = \frac{(I - u_t u_t^\top) P_\star u_t}{\|(I - u_t u_t^\top) P_\star u_t\|}$, Lemma 15 gives

$$\theta_{t+1} - \theta_t = -\frac{v_t^\top \bar{h}_t}{\|\bar{w}_t\|} + O\Big(\frac{\|\bar{h}_t\|^2}{\|\bar{w}_t\|^2}\Big).$$

For the MUON/right–polar step, recall

$$\bar{h}_t \;=\; -\frac{\eta}{m}\, Q^\top \mathbf{1}_m, \qquad Q := \overline{G}\, M^{-1/2}, \quad M := \overline{G}^\top \overline{G}.$$

On good event $\mathcal{E}$, we have $Q^\top Q = P_{im(M)}\Pi$ (the projector onto $\mathrm{span}\{x_b\}_{b\le B}$), hence $\|Q\| = \|Q^\top\| \le 1$. Therefore, deterministically,

$$\|\bar{h}_t\| \;\le\; \frac{\eta}{m}\, \|Q^\top\|\, \|\mathbf{1}_m\| \;=\; \frac{\eta}{\sqrt{m}}.$$

On the fixed horizon, $\|\bar{w}_t\| \ge c_w > 0$, so the quadratic remainder is

$$O\Big(\frac{\|\bar{h}_t\|^2}{\|\bar{w}_t\|^2}\Big) \;=\; O\Big(\frac{\eta^2}{m}\Big) \;=\; O(\eta^2),$$

which is even sharper than $O(\eta^2)$ and requires *no* column-energy comparability.

Step 2 (Numerator identity and AGOP-only split). We will apply the numerator alignment (Lemma 23) to u^{num}, which depends only on (W_t, U) (no batch randomness). Under AGOP condition of Assumption 12, row-norm control, and bounded g, g', g'', there exist $A_t', B_t' \in \mathbb{R}$ and $r_t \in \mathbb{R}^p$ such that

$$u^{\mathrm{num}} = m A_t'\, \bar{w}_t - m B_t'\, P_\star \bar{w}_t + r_t,$$

with bounds

$$|A'_t| \;\le\; C_A\,\frac{1+R_t}{\sqrt{p}}, \qquad |B'_t| \;\le\; C_B\,\frac{1}{\sqrt{p}}, \qquad \|r_t\| \;\le\; C_r\,\frac{m}{\sqrt{p}}.$$

Signed teacher–overlap. Let $u_t = \bar{w}_t/\|\bar{w}_t\|$ and $a_\| := \frac{U^\top u_t}{\|U^\top u_t\|}$ whenever $\|U^\top u_t\| > 0$. Let $u_t = \bar{w}_t/\|\bar{w}_t\|$ and $a_\| := \frac{U^\top u_t}{\|U^\top u_t\|}$ whenever $\|U^\top u_t\| > 0$. We define B'_t by the projection of u^{num} onto the teacher direction of $\bar{w}_t$:

$$m\,B'_t\,\|U^\top \bar{w}_t\| \;:=\; a_\|^\top U^\top u^{\mathrm{num}}.$$

With this choice, the teacher contribution in the decomposition is *exactly* the component of u^{num} along $P_\star \bar{w}_t$, and any teacher-plane component orthogonal to $P_\star \bar{w}_t$ is absorbed into r_t (which is already accounted for in the $\|r_t\|$ bound above).

By Lemma 23, we have the decomposition

$$u^{\mathrm{num}} = mA'_t\,\bar{w}_t - mB'_t\,P_\star \bar{w}_t + r_t, \quad |A'_t| \le C_A\frac{1+R_t}{\sqrt{p}},\ |B'_t| \le C_B\frac{1}{\sqrt{p}},\ \|r_t\| \le C_r\frac{m}{\sqrt{p}}.$$

Define the signed overlap coefficient

$$\widetilde{B}'_t := A'_t - B'_t \quad \Rightarrow \quad a_\|^\top U^\top u^{\mathrm{num}} = m\,\widetilde{B}'_t\,\|U^\top \bar{w}_t\|.$$

Moreover $|\widetilde{B}'_t| \le (C_A(1+R_t)+C_B)/\sqrt{p}$.

Step 3 (Project the Muon step). On the good event $\mathcal{E}$,

$$\bar{h}_t \;=\; -\,\frac{\eta}{m}\,M^{-1/2}u, \qquad \mathbb{E}[u \mid W_t] = \sqrt{m}\,u^{\mathrm{num}}.$$

Hence, using $v_t^\top P_\star h = \sin\theta_t\, a_\|^\top U^\top h$ and conditioning on W_t,

$$\mathbb{E}\left[\frac{v_t^\top \bar{h}_t}{\|\bar{w}_t\|}\,\middle|\,W_t\right] = -\,\frac{\eta}{\sqrt{m}}\,\sin\theta_t\,\frac{a_\|^\top U^\top M^{-1/2}\,u^{\mathrm{num}}}{\|\bar{w}_t\|}.$$

Since $U^\top(I-P_\star)=0$ and $P_\star\bar{w}_t = \|\bar{w}_t\|\cos\theta_t\,Ua_\|$,

$$a_\|^\top U^\top M^{-1/2}(mA'_t\bar{w}_t - mB'_tP_\star\bar{w}_t) = m(A'_t-B'_t)\,\|\bar{w}_t\|\cos\theta_t\;y^\top M^{-1/2}y, \quad y := Ua_\|,\ \|y\|=1.$$

Therefore

$$\mathbb{E}\left[\frac{v_t^\top \bar{h}_t}{\|\bar{w}_t\|}\,\middle|\,W_t\right] = -\,\eta\,\sqrt{m}\,(A'_t-B'_t)\,\sin\theta_t\cos\theta_t\;y^\top M^{-1/2}y \tag{15}$$

$$-\;\frac{\eta}{\sqrt{m}}\,\sin\theta_t\,\frac{a_\|^\top U^\top M^{-1/2}r_t}{\|\bar{w}_t\|}. \tag{16}$$

By Lemma 21, on $\mathrm{im}(\Pi)$

$$(\kappa_+\lambda)^{-1/2}\,\|\Pi y\|^2 \;\le\; y^\top M^{-1/2}y \;\le\; (\kappa_-\lambda)^{-1/2}\,\|\Pi y\|^2, \qquad \|M^{-1/2}\Pi\| \;\le\; \kappa_-^{-1/2}\lambda^{-1/2}.$$

Using $\cos\theta_t = 1+O(\theta_t^2)$ and $\sin\theta_t = \theta_t + O(\theta_t^3)$, and $\|\bar{w}_t\| \ge c_w > 0$ on the fixed horizon,

$$\Big|\mathbb{E}\Big[\theta_{t+1}-\theta_t \mid W_t\Big]\Big| \;\le\; \eta\,\sqrt{m}\,|A'_t-B'_t|\,(\kappa_-\lambda)^{-1/2}\,\|\Pi U a_\|\|^2\,\theta_t \;+\; \eta\,\kappa_-^{-1/2}\lambda^{-1/2}\,\frac{\|\Pi r_t\|}{\sqrt{m}\,\|\bar{w}_t\|}\,\theta_t \;+\; O(\eta\,\theta_t^3+\eta^2).$$

Equivalently,

$$\Big|\mathbb{E}\Big[\theta_{t+1}-\theta_t \mid W_t\Big]\Big| \;\le\; \eta\,\lambda^{-1/2}\,\sqrt{m}\Big(\,|A'_t-B'_t|\,\|\Pi U a_\|\|^2 + \frac{\|r_t\|}{m}\,\Big)\theta_t \;+\; O(\eta\,\theta_t^3+\eta^2),$$

where the constants absorb $\kappa_\pm$ and c_w. Since $|A'_t| \lesssim (1+R_t)/\sqrt{p}$, $|B'_t| \lesssim 1/\sqrt{p}$, and $\|r_t\| \lesssim m/\sqrt{p}$, the parenthesis is $O(1/\sqrt{p})$, yielding the stated $K_{\mathrm{M}}^{(\mathrm{size})}$ bound with $\|\Pi U a_\parallel\|^2$.

Step 4 (Substitute scales.) With $\lambda^{-1/2} = \sqrt{\frac{mB}{p}}\,\frac{1}{R_t}$ (from $\lambda = \frac{p\,R_t^2}{mB}$) and $|A'_t - B'_t| \lesssim \frac{1+R_t}{\sqrt{p}}$,

$$\lambda^{-1/2}\sqrt{m}\,|A'_t - B'_t| \;\lesssim\; \frac{\sqrt{mB}}{\sqrt{p}\,R_t}\cdot\sqrt{m}\,\frac{1+R_t}{\sqrt{p}} \;=\; \frac{m\sqrt{B}}{p\,R_t}\,(1+R_t) \;\le\; C\,\frac{\sqrt{B}}{R_t}\,(1+R_t),$$

using $m/p \to \gamma$. Thus, $\lambda^{-1/2}\sqrt{m}\,|A'_t - B'_t|\|\Pi U a_\parallel\|^2 \le O(\frac{B^{3/2}}{pR_t})$, where we have used the assumption (see Lemma 22) that $\|\Pi U a_\parallel\| \le \frac{\sqrt{B}}{\sqrt{p}}$ (since $a_\parallel$ is a unit vector and Π projects onto a B-dimensional space).

For the unstructured remainder, using $\lambda^{-1/2} = \sqrt{mB/p}\,/R_t$ and $\|r_t\| \lesssim m/\sqrt{p}$,

$$\lambda^{-1/2}\sqrt{m}\,\frac{\|\Pi r_t\|}{m} \;\lesssim\; \frac{\sqrt{mB/p}}{R_t}\cdot\sqrt{m}\cdot\frac{1}{\sqrt{p}} \;=\; \frac{\sqrt{B}}{R_t}$$

where the last step uses $m/p = \Theta(1)$ (proportional regime). Under AGOP-only, this remainder term dominates asymptotically. Therefore

$$\Big|\mathbb{E}[\theta_{t+1} - \theta_t \mid W_t]\Big| \;\le\; \eta\,K_{\mathrm{M}}^{(\mathrm{size})}\,\theta_t \;+\; O\big(\eta\,\theta_t^3 + \eta^2\big), \qquad K_{\mathrm{M}}^{(\mathrm{size})} \;\lesssim\; \frac{B^{3/2}}{pR_t} + \frac{\sqrt{B}}{R_t}$$

Step 5 (Good-event bookkeeping and final rate.) On $\mathcal{E}^c$, $\|\bar{h}_t\| \le \eta/\sqrt{m}$ (right–polar bound), so

$$|v_t^\top \mathbb{E}[\bar{h}_t \mathbf{1}_{\mathcal{E}^c} \mid W_t]| \le \eta/\sqrt{m}\cdot\mathbb{P}(\mathcal{E}^c \mid W_t) = o(\eta/\sqrt{m}),$$

which is dominated by $O(\eta^2 + p^{-1/2})$ in the stated regime. □

Corollary 3. *Under the additional assumptionof numerator sign structure $A'_t - B'_t \le -c_0/\sqrt{p}$ along the horizon for some $c_0 > 0$ and batch span overlap (i.e., $\|\Pi U a_\parallel\|^2 \gtrsim B/p$), we have a signed rate with*

$$\mathbb{E}[\theta_{t+1} - \theta_t \mid W_t] \;\le\; -\,\eta\,\underbrace{\Big(\sqrt{m}\,|A'_t - B'_t|\;\cdot\; y^\top M^{-1/2} y\Big)}_{K_{\mathrm{M}}}\,\theta_t \;+\; \textit{(smaller remainders).}$$

such that

$$K_{\mathrm{M}} \;\gtrsim\; \frac{\sqrt{m}\,B^{3/2}}{p^2\,R_t}\cdot\frac{1}{\sqrt{\log(2B/\delta)}} \qquad \textit{with high probability.}$$

Proof. On the truncation event $\mathcal{G} := \{\max_b \|c_b\|^2 \le R\}$ with $R \asymp R_t^2\log(2B/\delta)$ (bounded g and fixed–horizon tails of e) in the proof of Lemma 24,

$$s_{\max}(K) \;\le\; \mathrm{tr}(K) \;=\; \sum_b \|c_b\|^2 \;\le\; BR, \qquad s_{\max}(X) \;\le\; \big(\sqrt{p} + \sqrt{B} + \tau\big)$$

with high probability. Hence

$$\lambda \;=\; \frac{s_{\min}(K)\,s_{\min}(X)^2}{B^2} \;\le\; \frac{s_{\max}(K)\,s_{\max}(X)^2}{B^2} \;\le\; C_\lambda\,\frac{p}{B}\,R_t^2\,\log\frac{2B}{\delta},$$

so the Rayleigh term obeys (see proof of Theorem 5)

$$y^\top M^{-1/2} y \;\ge\; (\kappa_+\,\lambda)^{-1/2}\,\|\Pi y\|^2 \;\ge\; \frac{c}{\sqrt{\log(2B/\delta)}}\cdot\frac{\sqrt{B}}{\sqrt{p}\,R_t}\,\|\Pi y\|^2 \qquad \text{with high probability.}$$

Plugging this into Eq. (16), and using the signed numerator structure $A'_t - B'_t \le -c_0/\sqrt{p}$ and batch-span overlap $\|\Pi U a_\parallel\|^2 \gtrsim B/p$,

$$K_{\mathrm{M}} \;\gtrsim\; \sqrt{m}\cdot\frac{1}{\sqrt{p}}\cdot\frac{1}{\sqrt{\log(2B/\delta)}}\cdot\frac{\sqrt{B}}{\sqrt{p}\,R_t}\cdot\frac{B}{p} \;=\; \frac{\sqrt{m}\,B^{3/2}}{p^2\,R_t}\cdot\frac{1}{\sqrt{\log(2B/\delta)}} \gtrsim \frac{\sqrt{m}\,B^{3/2}}{p^2\,R_t}\cdot\frac{1}{\sqrt{\log(2B/\delta)}}$$

as claimed. □

Lemma 24 (High-probability lower bound for λ tied to R_t)**.** *Let $X = [x_1, \dots, x_B] \in \mathbb{R}^{p\times B}$ with $x_b \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p)$, and let $C = [c_1, \dots, c_B] \in \mathbb{R}^{m\times B}$ with columns*

$$(c_b)_j = \frac{e(x_b)}{\sqrt{m}}\, g(w_j^\top x_b), \qquad e(x) := f_W(x) - \sigma_\star(U^\top x), \quad g = \phi'.$$

Set $K := C^\top C$ and $M := \overline{G}^\top \overline{G} = \frac{1}{B^2} X K X^\top$. Let Π be the orthogonal projector onto $\mathrm{span}\{x_b\}_{b\le B}$.

Assume the standing conditions of the paper: fresh Gaussian mini-batches (Assumption 2), smooth bounded g and g' (Assumption 4), row-norm control (Assumption 8), mean–isotropy/AGOP (Assumption 7), finite horizon/small steps (Assumption 6). In addition, assume the following weak coupling holds uniformly on the horizon:

$$\textbf{(WC)} \qquad \mathrm{Cov}\big(e(x)^2,\; S_v(x)^2 \;\big|\; W_t\big) \;\ge\; -\,\rho_0\, R_t^2\, \mathbb{E}[S_v(x)^2 \mid W_t] \quad \text{for all unit } v \in \mathbb{R}^m,$$

where $S_v(x) := \frac{1}{\sqrt{m}} \sum_{j=1}^m v_j\, g(w_j^\top x)$, $R_t^2 := \mathbb{E}[e(x)^2 \mid W_t]$, and some fixed $\rho_0 \in [0, 1)$ depending only on $(\phi, c_\pm)$ and the horizon.

Define

$$\kappa_2^- \;:=\; \inf_{r\in[c_-,c_+]} \mathbb{E}\big[g(rG)^2\big] \;>\; 0, \qquad G \sim \mathcal{N}(0,1).$$

Then, for any $\varepsilon_X \in (0, 1 - \sqrt{B/p})$ and $\varepsilon_B \in (0, 1)$, with probability at least

$$1 \;-\; 2e^{-c_X \varepsilon_X^2 p} \;-\; \delta \;-\; B\exp\Big(-\,\tfrac{\varepsilon_B^2\, B\, c_g\, R_t^2}{2C_g m \log(2B/\delta)}\Big),$$

where c_g is the constant from part (iv) and C_g depends on $\|g\|_\infty$ and the sub-exponential norm of $e(x)$ (Assumption 4, Assumption 6), the following hold simultaneously:

(i) ***Batch–span sandwich:*** *On $\mathrm{im}(\Pi)$,*

$$M \;\succeq\; \lambda\,\Pi, \qquad \lambda \;:=\; \frac{s_{\min}(K)\, s_{\min}(X)^2}{B^2}.$$

(ii) ***Wishart lower bound for $s_{\min}(X)$:***

$$s_{\min}(X) \;\ge\; \sqrt{p} - \sqrt{B} - \varepsilon_X\sqrt{p} \quad\Longrightarrow\quad s_{\min}(X)^2 \;\ge\; p\Big(1 - \sqrt{B/p} - \varepsilon_X\Big)^2,$$

by fresh Gaussian mini-batches (Assumption 2).

(iii) ***Matrix Chernoff lower bound for $s_{\min}(K)$:***

$$s_{\min}(K) \;\ge\; B(1-\varepsilon_B)\, s_{\min}(\Sigma_c), \qquad \Sigma_c := \mathbb{E}[c_b c_b^\top \mid W_t],$$

where the truncation event uses bounded g and fixed-horizon sub-exponential tails of $e(x_b)$ (Assumption 4, Assumption 6).

(iv) ***Population bound $s_{\min}(\Sigma_c) \gtrsim R_t^2/m$:*** *There exists $p_0 < \infty$ and a constant $c_g > 0$ (depending only on ϕ, $c_\pm$, the AGOP constant, and ρ_0) such that for all $p \ge p_0$,*

$$s_{\min}(\Sigma_c) \;\ge\; c_g\,\frac{R_t^2}{m}.$$

Consequently, on this event,

$$\lambda \;\geq\; c_g\,(1-\varepsilon_B)\left(1-\sqrt{B/p}-\varepsilon_X\right)^2 \;\cdot\; \frac{p}{mB}\,R_t^2.$$

Proof. (i) Sandwich on the batch span. For any $y \in \mathrm{im}(\Pi)$, write $y = Xu$ for some $u \in \mathbb{R}^B$. Then

$$y^\top M y = \frac{1}{B^2}\,y^\top XKX^\top y = \frac{1}{B^2}\,u^\top X^\top XKX^\top Xu \;\geq\; \frac{s_{\min}(X^\top X)\,s_{\min}(K)}{B^2}\,\|Xu\|^2 = \frac{s_{\min}(K)\,s_{\min}(X)^2}{B^2}\,\|y\|^2,$$

which is $M \succeq \lambda\,\Pi$ with the displayed λ.

(ii) Wishart lower tail for X. Since X has i.i.d. $\mathcal{N}(0,1)$ entries (by Assumption 2), standard Gaussian/Wishart concentration implies $s_{\min}(X) \geq \sqrt{p} - \sqrt{B} - \tau_X$ with probability $\geq 1 - 2e^{-c_X \tau_X^2}$. Setting $\tau_X = \varepsilon_X\sqrt{p}$ yields the stated bound.

(iii) Matrix Chernoff via $CC^\top$. Let $Y_b := c_b c_b^\top \succeq 0$ so $CC^\top = \sum_{b=1}^B Y_b$ and $\mathbb{E}[Y_b \mid W_t] = \Sigma_c$. On the truncation event $\mathcal{G} := \{\max_b \|Y_b\|_{\mathrm{op}} \leq R\}$ with $R = C_g\,R_t^2\log(2B/\delta)$ (by Assumption 4, Assumption 6), matrix Chernoff yields, for $\varepsilon_B \in (0,1)$,

$$\mathbb{P}\Big(\lambda_{\min}(CC^\top) \leq (1-\varepsilon_B)\,B\,s_{\min}(\Sigma_c) \;\Big|\; W_t, \mathcal{G}\Big) \;\leq\; m\,\exp\Big(-\tfrac{\varepsilon_B^2\,B\,s_{\min}(\Sigma_c)}{2R}\Big).$$

As $K = C^\top C$ and $CC^\top$ have the same nonzero eigenvalues, the same bound holds for $s_{\min}(K)$. Unconditioning adds the $+\delta$ term.

Thus,

$$\mathbb{P}\Big(s_{\min}(K) \leq (1-\varepsilon_B)\,B\,s_{\min}(\Sigma_c) \;\Big|\; W_t, \mathcal{G}\Big) \;\leq\; B\,\exp\Big(-\tfrac{\varepsilon_B^2\,B\,s_{\min}(\Sigma_c)}{2R}\Big).$$

Unconditioning (union bound with $\mathbb{P}(\mathcal{G}^c \mid W_t) \leq \delta$) gives

$$\mathbb{P}\Big(s_{\min}(K) \leq (1-\varepsilon_B)\,B\,s_{\min}(\Sigma_c) \;\Big|\; W_t\Big) \;\leq\; \delta \;+\; B\,\exp\Big(-\tfrac{\varepsilon_B^2\,B\,s_{\min}(\Sigma_c)}{2R}\Big).$$

In particular, with probability at least $1 - \delta - B\exp\big(-\varepsilon_B^2 B\,s_{\min}(\Sigma_c)/(2R)\big)$,

$$s_{\min}(K) \;\geq\; (1-\varepsilon_B)\,B\,s_{\min}(\Sigma_c).$$

(iv) Population lower bound $s_{\min}(\Sigma_c) \gtrsim R_t^2/m$. Recall

$$c_b = \frac{e(x_b)}{\sqrt{m}}\,g(x_b), \qquad \Sigma_c = \mathbb{E}[c_b c_b^\top \mid W_t] = \frac{1}{m}\,\mathbb{E}\big[e(x)^2\,g(x)g(x)^\top \mid W_t\big],$$

and for any unit $v \in \mathbb{R}^m$,

$$v^\top \Sigma_c v = \mathbb{E}\big[e(x)^2\,S_v(x)^2 \mid W_t\big], \qquad S_v(x) := \frac{1}{\sqrt{m}}\sum_{j=1}^m v_j\,g(w_j^\top x).$$

By **(WC)**,

$$\mathbb{E}\big[e(x)^2\,S_v(x)^2 \mid W_t\big] \;\geq\; (1-\rho_0)\,R_t^2\,\mathbb{E}\big[S_v(x)^2 \mid W_t\big]. \tag{17}$$

Let $\Sigma_g := \mathbb{E}[g(x)g(x)^\top \mid W_t]$. Then

$$\mathbb{E}\big[S_v(x)^2 \mid W_t\big] = \frac{1}{m}\,v^\top \Sigma_g v \;\geq\; \frac{1}{m}\,s_{\min}(\Sigma_g).$$

Hence

$$s_{\min}(\Sigma_c) \;\geq\; (1-\rho_0)\,\frac{R_t^2}{m}\,s_{\min}(\Sigma_g). \tag{18}$$

Variant (A): monotone g and nonnegative pre-activation correlations. Assume $g = \phi' \geq 0$ pointwise and, in addition, that

$$\rho_{ij} := \frac{\langle w_i, w_j \rangle}{\|w_i\| \, \|w_j\|} \;\geq\; 0 \qquad \text{for all } i \neq j.$$

Then (by association for jointly Gaussian inputs with nonnegative correlations and monotone g),

$$\mathbb{E}\big[g(w_i^\top x)\, g(w_j^\top x) \mid W_t\big] \;\geq\; 0 \qquad (i \neq j).$$

Writing out

$$\mathbb{E}\big[S_v(x)^2 \mid W_t\big] = \frac{1}{m} \sum_{j=1}^{m} v_j^2 \, \mathbb{E}\big[g(w_j^\top x)^2 \mid W_t\big] + \frac{2}{m} \sum_{i<j} v_i v_j \, \mathbb{E}\big[g(w_i^\top x)\, g(w_j^\top x) \mid W_t\big],$$

the cross-term is nonnegative; therefore

$$\mathbb{E}\big[S_v(x)^2 \mid W_t\big] \;\geq\; \frac{1}{m} \sum_{j=1}^{m} v_j^2 \, \mathbb{E}\big[g(\|w_j\| G)^2\big] \;\geq\; \frac{\kappa_2^-}{m},$$

where $G \sim \mathcal{N}(0,1)$ and

$$\kappa_2^- := \inf_{r \in [c_-, c_+]} \mathbb{E}\big[g(rG)^2\big] \;>\; 0 \qquad \text{(by row-norm control, Assumption 8).}$$

Combining with Eq. (17)–(18) gives

$$s_{\min}(\Sigma_c) \;\geq\; (1 - \rho_0)\, \kappa_2^- \, \frac{R_t^2}{m}.$$

Variant (B): general case (no sign assumption on g). Let $D := \operatorname{diag}(\mathbb{E}[g(w_j^\top x)^2 \mid W_t])$ and $E := \Sigma_g - D$. By row-norm control (Assumption 8),

$$s_{\min}(D) \;=\; \min_j \mathbb{E}[g(\|w_j\| G)^2] \;\geq\; \kappa_2^-.$$

For $i \neq j$, Price/Stein with bounded g' (Assumption 4) yields

$$|E_{ij}| \;=\; \big|\mathbb{E}[g(w_i^\top x)\, g(w_j^\top x) \mid W_t]\big| \;\leq\; L_\kappa \, |\rho_{ij}|, \qquad L_\kappa \;\lesssim\; \|g'\|_\infty^2 \, c_+^2,$$

hence

$$\|E\|_{\mathrm{op}} \;\leq\; L_\kappa \, \|R_{\mathrm{off}}\|_{\mathrm{op}}, \qquad R_{\mathrm{off}} := (\rho_{ij} \mathbf{1}_{i \neq j}).$$

Under AGOP condition (Assumption 7) there exists $C_{\mathrm{ag}} < \infty$ such that for all large p,

$$\|R_{\mathrm{off}}\|_{\mathrm{op}} \;\leq\; C_{\mathrm{ag}} \big(1 + |m/p - 1|\big).$$

By Weyl's inequality,

$$s_{\min}(\Sigma_g) \;\geq\; s_{\min}(D) - \|E\|_{\mathrm{op}} \;\geq\; \kappa_2^- \;-\; L_\kappa \, C_{\mathrm{ag}} \big(1 + |m/p - 1|\big).$$

Choose $p_0 < \infty$ so that for all $p \geq p_0$, $L_\kappa \, C_{\mathrm{ag}} (1 + |m/p - 1|) \leq \frac{1}{2} \kappa_2^-$. Then $s_{\min}(\Sigma_g) \geq \frac{1}{2} \kappa_2^-$, and by Eq. (18),

$$s_{\min}(\Sigma_c) \;\geq\; \frac{1}{2} (1 - \rho_0)\, \kappa_2^- \, \frac{R_t^2}{m}.$$

Both variants establish (iv) with $c_g = (1 - \rho_0)\kappa_2^-$ in (A) and $c_g = \frac{1}{2}(1 - \rho_0)\kappa_2^-$ in (B).

Conclusion. Let E_X be the event in (ii), $\mathcal{G}$ the "good event" in (iii), and E_K the Chernoff tail event in (iii). Then by a union bound,

$$\mathbb{P}(E_X \cap \mathcal{G} \cap E_K) \;\geq\; 1 - \mathbb{P}(E_X^c) - \mathbb{P}(\mathcal{G}^c) - \mathbb{P}(E_K^c \mid \mathcal{G}),$$

which equals the probability displayed in the lemma statement (after substituting $s_{\min}(\Sigma_c) \geq c_g R_t^2/m$ and $R = C_g R_t^2 \log(2B/\delta)$; see Assumption 4, Assumption 6). On $E_X \cap \mathcal{G} \cap E_K$, parts (ii)–(iii) hold, and (iv) holds deterministically for $p \geq p_0$. Plugging these into (i) yields

$$\lambda = \frac{s_{\min}(K) s_{\min}(X)^2}{B^2} \;\geq\; (1-\varepsilon_B)\, B\, s_{\min}(\Sigma_c) \cdot \frac{p}{B^2}\Big(1-\sqrt{B/p}-\varepsilon_X\Big)^2 \;\geq\; c_g(1-\varepsilon_B)\Big(1-\sqrt{B/p}-\varepsilon_X\Big)^2 \frac{p}{mB}\, R_t^2.$$

□

References

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning, 2024. URL https://arxiv.org/abs/2410.21265.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Beijing, China, June 2014. PMLR. URL https://proceedings.mlr.press/v32/rezende14.html.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018.