An Analytical Model of Compute-limited Multistage Training Pipelines

Nishil Patel
Gatsby Computational Neuroscience Unit
University College London
nishil.patel@ucl.ac.uk

Jin Hwa Lee
Sainsbury Wellcome Centre
University College London

Basile Confavreux
Gatsby Computational Neuroscience Unit
University College London

Andrew Saxe
Gatsby Computational Neuroscience Unit
Sainsbury Wellcome Centre
University College London

Abstract

Modern foundation models are trained through multiple stages, yet how these stages interact and how compute should be allocated across them remain poorly understood. We introduce an analytically tractable high-dimensional model combining unsupervised pretraining with a linear autoencoder, supervised fine-tuning (SFT) with noisy stepwise labels, and reinforcement learning (RL) with sparse episodic rewards. Using random matrix theory and an operator-projection closure, we derive approximate low-dimensional learning dynamics that closely match simulations. We show that finite pretraining data imposes a representation-dependent performance limit, with Bayes classification error scaling as the inverse square root of the pretraining sample ratio, and observe the same exponent in asymptotic SFT and RL errors. Better pretraining also reduces sensitivity to noisy supervision. SFT initialization shortens the initial RL plateau induced by sparse rewards, an effect we also observe in OLMo-2 1B on GSM8K. Longer SFT warm starts additionally improve asymptotic RL performance. Optimizing compute across all three stages reveals that the optimal allocation depends on supervision quality and total budget, with a non-monotonic dependence of the RL compute share on budget. Finally, a complementary compositional model shows how RL selectively strengthens task-relevant pretrained skills while largely preserving unused skills. Together, these results identify mechanisms linking representation quality, feedback structure, and compute allocation in multistage training.

1 Introduction

Modern foundation models exhibit transformative capabilities from human-level language comprehension and production Radford et al. [2019], Brown et al. [2020], Touvron et al. [2023], to reasoning across modalities and mathematical problem-solving Jaech et al. [2024], Shao et al. [2024], Chervonyi et al. [2025], Team et al. [2025]. Multistage training is a critical aspect of the training paradigm of such models, which consists of unsupervised auto-regressive pretraining followed by several stages of supervised finetuning and reinforcement learning on smaller, domain-specific datasets Ouyang et al. [2022], Achiam et al. [2023], Liu et al. [2024], Yang et al. [2025], DeepSeek-AI et al. [2025]. Yet, the consequences of different training stages and their interactions remain poorly understood. Empirical investigation of these questions is difficult, as it requires substantial compute and costly

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

human annotation, constraining the scope of hypotheses that can be explored Zhao et al. [2025], Shao et al. [2025], Akter et al. [2025], Liu et al. [2025a].

Alongside the transition to multistage training pipelines, foundation models reflect a second shift from a sample-complexity–centric view of learning to a compute-centric one. Whereas classical theory treats data as the limiting resource and implicitly assumes unbounded compute, modern foundation models invert this assumption: data is abundant, compute is limited, and performance scales predictably with model size and training compute, as evidenced by large-scale empirical scaling laws Kaplan et al. [2020], Hoffmann et al. [2022]. In this compute-limited regime, models are rarely trained to convergence, and performance is instead determined by how a fixed computational budget is allocated across model capacity and multiple training stages.

Together, these two shifts—toward multistage pipelines and the compute-limited regime—pose a challenge for existing theory, which offers limited guidance for analyzing learning dynamics, performance trade-offs, or compute allocation in modern foundation model training. —

Contributions. —To this end, we develop an analytically tractable high-dimensional model based on statistical physics Seung et al. [1992], Saad and Solla [1995], Zdeborová and Krzakala [2016]. The model isolates the effects of pretrained representation quality, dense stepwise supervision, and sparse episodic rewards. We derive deterministic learning dynamics in the high-dimensional limit and obtain a tractable, approximately closed system of ODEs, allowing us to study how pretrained representations, supervision quality, and compute allocation jointly determine downstream performance. Our contributions are:

- We combine random matrix theory with an operator-projection closure to obtain a low-dimensional approximation of the SFT and RL dynamics, validated against simulations. We derive the $\alpha^{-1/2}$ scaling of the representation-dependent Bayes error with pretraining sample size and observe the same exponent in asymptotic finetuning error.
- We characterize how SFT and RL complement each other. SFT shortens the initial sparse-reward plateau, an effect also observed in OLMo-2 1B on GSM8K. Under spherical normalization and locally optimal learning rates, longer SFT warm starts additionally improve asymptotic RL performance. Better pretraining also reduces the excess SFT error caused by label noise.
- We optimize compute allocation across pretraining, SFT, and RL. The optimal split depends on both supervision quality and the total budget: noisier labels favour RL, while the optimal RL share can vary non-monotonically with total compute.
- We extend the model to compositional tasks, showing how prior skill proficiency accelerates RL and how finetuning selectively improves task-relevant skills while largely preserving unused skills.

2 Related Work

Empirical interactions across training stages. Controlled studies have investigated how pretraining shapes the effectiveness of downstream reinforcement learning. Shen et al. [2026] examine a complete pretraining-–SFT-–RL pipeline and relate pretraining loss and data scale to post-RL performance and learning speed, including a pretraining—RL compute frontier. Zhang et al. [2026a] isolate the effects of pretraining, intermediate training, and RL on compositional reasoning tasks. Zhao et al. [2024] investigate the amplification of pretrained behaviors under RL. Empirical work compares the generalization behavior of SFT and outcome-based RL [Chu et al., 2025] and demonstrates that higher standalone SFT accuracy need not yield better subsequent RL performance [Zhang et al., 2026b, Kang et al., 2026].

Closure of structured-data learning dynamics. The problem of anisotropic data covariance generating an infinite hierarchy of order parameters has been studied in the context of online learning with structured data Yoshida and Okada [2019]. In that work, the authors show that the ODE system can only be closed if the covariance has a finite number of unique eigenvalues, where there is a pair of order parameters associated to every unique eigenvalue. There are several works that address this feature: approaches include replacing the infinite discrete hierarchy by a closed functional object —

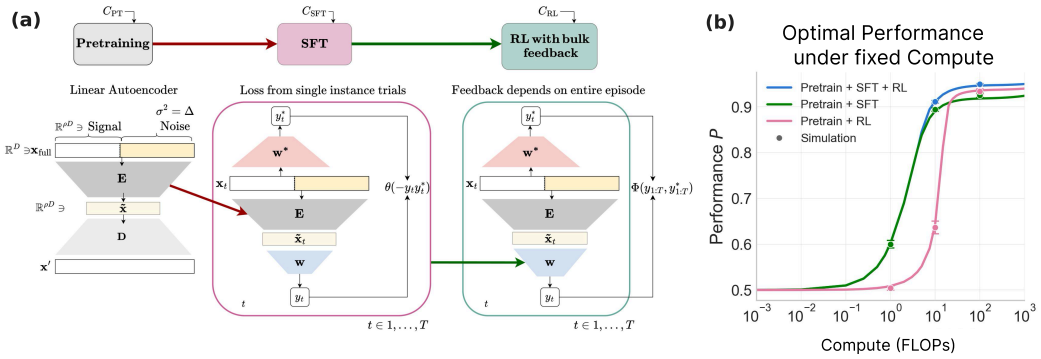


Figure 1: (a) **A Schematic of multistage training pipeline.** Each solid trapezoid indicates a single linear neural network layer. Pretraining is a Linear Autoencoder (LAE) that approximately learns the principal subspace of data $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$, where Σ is equal to the identity apart from a noise subspace, with identity covariance scaled by $\Delta < 1$. C_{PT} is the compute allocated to pretraining. In the subsequent finetuning (SFT and RL), a teacher network $\mathbf{w}^*$ gives labels y^* from the signal subspace and it is used to train a student network $\mathbf{w}$ learning in the subspace of the LAE encoder projection with feedback. The student predicts label y using the effective features $\tilde{\mathbf{x}}$. Importantly, SFT is modelled by every observation at each t updating the student, while RL is modelled by bulk feedback through reward at the end of a 'correct' sequence through the entire $t = 1, \ldots, T$ episode. We denote the compute allocations to SFT and RL, C_{SFT} and C_{RL} respectively. (b) Optimal generalisation performance against total Compute flops under three protocols: SFT only, RL only, and SFT with RL. Compute is allocated as described in 5.1, parameters are set as in 5d.

a spectral order parameter [Goldt et al., 2020, 2022] or a resolvent [Collins-Woodfin et al., 2024]. However, this is non-trivial to implement in the case of a random data covariance. Another more recent approach most similar in spirit to ours is Xu et al. [2026], where the authors show the infinite hierarchy can be exponentially well approximated by a finite truncated system. Ours is merely a different method of truncation, without a proof of exponential convergence.

3 Setup

Pretraining

We draw a fixed pretraining dataset $\mathbf{X} \in \mathbb{R}^{N \times D}$ with independent rows distributed as $\mathcal{N}(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{M}_s + \Delta \mathbf{M}_n$, $\mathbf{M}_s$ projects onto a signal subspace of dimension ρD, and $\mathbf{M}_n = \mathbf{I} - \mathbf{M}_s$. We take $0 < \rho < 1$, $0 < \Delta < 1$, and the high-dimensional limit $N, D \to \infty$ at fixed $\alpha = N/D > \rho$. A linear autoencoder (LAE) with hidden width ρD learns the empirical principal subspace of this dataset when trained to convergence [Baldi and Hornik, 1989, Saxe et al., 2019]. We represent its learned subspace by the orthogonal projector $\mathbf{P}$. Downstream observations are $\tilde{\mathbf{x}} = \mathbf{P}\mathbf{x}$, with fresh $\mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma})$. $\mathbf{P}$ remains fixed throughout finetuning. Thus α determines how much task-relevant signal is available for the downstream task. The LAE learns a compressed feature subspace from *unlabelled* data, abstracting the role of pretraining as representation learning: discovering reusable structure and organizing inputs into a feature space [Bengio et al., 2012, LeCun et al., 2015, 2022, Bommasani et al., 2021]. This frozen-feature assumption is motivated by evidence that finetuning preserves pretrained representations [Nadipalli, 2025] and the success of PEFT [Hu et al., 2022, Han et al., 2024]

Finetuning

The finetuning task is organized in episodes of length T, where a *student* perceptron $\mathbf{w} \in \text{Range}(\mathbf{P})$ makes a series of binary decisions $y_t = \text{sgn}\left(\frac{\tilde{\mathbf{x}}_t^\intercal \mathbf{w}}{\sqrt{D}}\right)$ given the raw observations $\mathbf{x}_{1\ldots T} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ projected onto the encoder subspace by $\mathbf{P}$. Data labels for supervision/reward $y_t^* = \text{sgn}\left(\frac{\mathbf{x}^\intercal \mathbf{w}^*}{\sqrt{D}}\right)$ are specified by a *teacher* perceptron $\mathbf{w}^* \in R^D$, which has access to the raw data $\mathbf{x}$, importantly, it sees the entire signal subspace. The teacher is a fixed-weight oracle that generates the target labels.

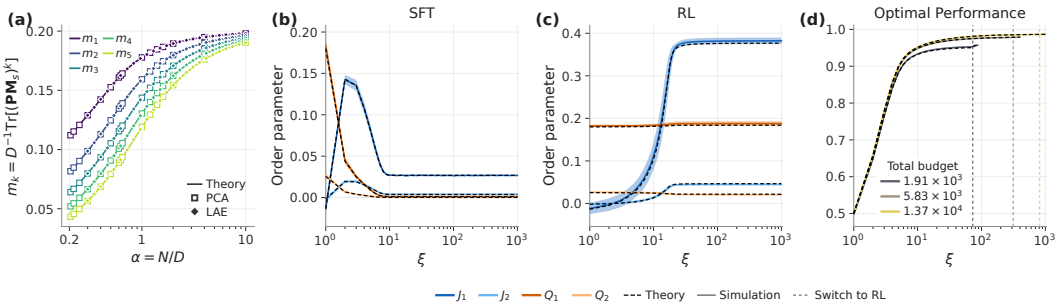


Figure 2: **ODEs accurately capture distinct learning dynamics of SFT and RL.** (a) Good agreement of normalised traces of powers of projection-signal overlap matrices for: RMT (solid), PCA (open square), and LAE (filled diamond). Good agreement of ODE theory (dashed) in time (ξ) and simulated evolution (solid) of order parameters for: (b) the SFT model, (c) the RL model. (d) The evolution of probability of correct classification (P) under optimal compute budget allocation for a range of total compute budgets in FLOPs. *Parameters*: $\Delta = 0.1$, $\rho = 0.2$. (a) $D = 2000$. (b) $\epsilon = 0.05$, spherically unconstrained weights, $\eta = 1$. (c) spherically constrained weights, locally optimal learning rate. (b), (c) and (d) $T = 6$, $D = 1000$. (d) $\epsilon = 0.2$.

During SFT and RL, the weights of the student are updated with feedback from the teacher, but the core difference lies in how the feedback from the output is given to the student. This setting is inspired by the teacher–student framework [Gardner and Derrida, 1989, Seung et al., 1992, Biehl and Schwarze, 1993]. We discuss related teacher–student models in Appendix .1.

SFT: Each individual sample in the episode contributes separately to the update by error feedback with the corresponding teacher label (Figure 1a, pink box). The update follows the Perceptron rule [Rosenblatt, 1958]. The update at the μ_{th} episode can be written as

$$\mathbf{w}^{(\mu+1)} = \mathbf{w}^{(\mu)} - \frac{\eta}{T\sqrt{D}} \sum_{t=1}^{T} y_t^{(\mu)} \tilde{\mathbf{x}}_t^{(\mu)} \theta(-y^*_{\mathrm{SFT}t}{}^{(\mu)} y_t^{(\mu)}), \tag{1}$$

where $\theta(\cdot)$ is the heaviside step function and η is the learning rate. This can be interpreted as the linearized gradient descent update on the mean squared loss $\Sigma_{t=1}^{T}(y_t - y^*_{\mathrm{SFT}t})^2$. The teacher labels for SFT are generated with noise, i.e. $y^*_{\mathrm{SFT}t} = y^*_t z_t$, where z_t takes the values of 1 or -1 with probability $(1-\epsilon)$ or ϵ respectively. This is a simple model of imperfect supervision, which is common in real-world datasets. N.B. this label noise is unique to SFT in our model. Importantly, the student receives an individual update for every t. For example, in multi-step math tasks, this is equivalent to getting a supervised learning signal at *every step* of a solution.

RL: We use the RL perceptron update of Patel et al. [2025], which extends the teacher-student framework to online policy learning under a sequential decision-making task. Given the same finetuning task, the critical difference with SFT lies in the bulk feedback: the student is updated using a delayed, sparse reward signal at the end of the episode, rather than at every step, as in SFT. The update at μ_{th} episode can be written as

$$\mathbf{w}^{(\mu+1)} = \mathbf{w}^{(\mu)} + \frac{\eta}{T} \left(\sum_{t=1}^{T} \frac{y_t^{(\mu)} \tilde{\mathbf{x}}_t^{(\mu)}}{\sqrt{D}} \right) R(y_{1:T}^{(\mu)}, y^{*(\mu)}_{1:T}), \tag{2}$$

where the reward rule is specified by $R(y_{1:T}, y^*_{1:T})$. In this work, we mainly use the reward condition, *every action in an episode must be correct*, (i.e $y_t = y^*_t$, $\forall t \in 1, \dots T$, or $R = \Pi_{t=1}^{T} \theta(y_t y^*_t)$). This is akin to RLVR where reward is received based on the correctness of a final result, where the final result requires multiple previous steps to be correct in order to reach it. The episode length T controls the difficulty of the task. Larger values of T make the task more difficult as the model needs to make a longer sequence of correct decisions, giving an idealized analogue of verifiable-reward methods such as GRPO [DeepSeek-AI et al., 2025]. This task structure reflects the nature of multi-step problems, for example, a math problem requiring multiple steps of reasoning to be correct. When we consider

multiple stages of finetuning (e.g. SFT followed by RL), each stage trains the same student and uses the final student state of the previous stage as the initial student state of the subsequent stage.

We use it not as a literal model of full finetuning, but to isolate how representation quality constrains later learning.

4 Learning Dynamics of SFT and RL

We can derive a set of ordinary differential equations (ODEs) that describe the learning dynamics of our model using the approach of Saad and Solla. In the thermodynamic limit ($D \to \infty$), the stochastic evolution of the high-dimensional student weights reduces to the deterministic evolution of a set of low-dimensional *order parameters*:

$$J_{s/n} = \frac{\mathbf{w}^\top \mathbf{P}\mathbf{M}_{s/n}\mathbf{P}\mathbf{w}^*}{D},\ Q_{s/n} = \frac{\mathbf{w}^\top \mathbf{P}\mathbf{M}_{s/n}\mathbf{P}\mathbf{w}}{D},\ S_{s/n} = \frac{\mathbf{w}^{*\top} \mathbf{P}\mathbf{M}_{s/n}\mathbf{P}\mathbf{w}^*}{D}\ S = \frac{\mathbf{w}^{*\top}\mathbf{w}^*}{D} \tag{3}$$

Intuitively, $Q_{s/n}$ is the size of the projected student in the signal/noise subspace, $J_{s/n}$ is the overlap between the projected student and the projected teacher in signal/noise subspace, and $(S_{s/n}, S)$ are constant throughout training.

By appropriate multiplication of 1 and 2, we can derive a set of ODEs that govern the evolution of order parameters in the thermodynamic limit.

For the SFT:

$$\begin{aligned}
\frac{dJ_{s/n}}{d\xi} &= -\eta \left\langle y \frac{\mathbf{w}^{*\top}\mathbf{P}\mathbf{M}_{s/n}\mathbf{P}\mathbf{x}}{\sqrt{D}} \theta(-yy^*) \right\rangle, \\
\frac{dQ_{s/n}}{d\xi} &= -2\eta \left\langle y \frac{\mathbf{w}^{\top}\mathbf{P}\mathbf{M}_{s/n}\mathbf{P}\mathbf{x}}{\sqrt{D}} \theta(-yy^*) \right\rangle + \frac{\eta^2}{T} \frac{\mathrm{Tr}(\mathbf{P}\mathbf{M}_{s/n}\mathbf{P}\Sigma)}{D} \langle \theta(-yy^*) \rangle,
\end{aligned} \tag{4}$$

For the RL:

$$\begin{aligned}
\frac{dJ_{s/n}}{d\xi} &= \frac{\eta}{T} \left\langle \sum_{t=1}^{T} y_t \frac{\mathbf{w}^{*\top}\mathbf{P}\mathbf{M}_{s/n}\mathbf{P}\mathbf{x}_t}{\sqrt{D}} R \right\rangle, \\
\frac{dQ_{s/n}}{d\xi} &= \frac{2\eta}{T} \left\langle \sum_{t=1}^{T} y_t \frac{\mathbf{w}^{\top}\mathbf{P}\mathbf{M}_{s/n}\mathbf{P}\mathbf{x}_t}{\sqrt{D}} R \right\rangle + \frac{\eta^2}{T} \frac{\mathrm{Tr}(\mathbf{P}\mathbf{M}_{s/n}\mathbf{P}\Sigma)}{D} \langle R \rangle.
\end{aligned} \tag{5}$$

Here, ξ denotes continuous training time and the angled bracket expectations are taken over the online samples $\mathbf{x}$, and can be approximated in terms of the order parameters, allowing approximate closure (see Section 4.1). The trace term is self-averaging in the proportional $D, N \to \infty$ limit, and can be expressed self-consistently (see Appendix A). The expected generalization ability P (i.e. the probability of a single decision being correct) can be expressed in terms of the order parameters:

$$P = \frac{1}{2} + \frac{1}{\pi} \sin^{-1} \left(\frac{J_s + J_n}{\sqrt{S(Q_s + \Delta Q_n)}} \right), \tag{6}$$

allowing us to compute generalization performance at any point during training. See Appendix A for the full derivation and explicit calculation of the expectations and traces in Equation (4) and Equation (5). In Figure 2, we see that the evolution of P found from the numerical integration of the ODEs (dashed) shows an excellent match with the simulation of the updates (solid). We consider both unconstrained dynamics and a spherical constraint fixing $Q_s + Q_n$. Where specified, the learning rate is chosen to maximize the instantaneous predicted increase in clean generalization probability, using the rule in Appendix A.9.

4.1 Closing the ODEs

Evaluating the expectations in Equations (4) and (5) introduces higher-order products of the pretraining and signal/noise projectors, such as $\mathbf{U}_{ij} = \mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\mathbf{P}$. Their associated order parameters in turn depend on higher-order products, generating a hierarchy of dynamical observables. We truncate this hierarchy by approximating higher-order operators through projection onto the retained lower-order operator basis. This is a polynomial least-squares approximation and can be applied at any chosen truncation order. Here, we retain the four order parameters (J_s, J_n, Q_s, Q_n), using

$$\mathbf{U}_{ij} \approx \Omega_{ij}^{(s)}\mathbf{P}\mathbf{M}_s\mathbf{P} + \Omega_{ij}^{(n)}\mathbf{P}\mathbf{M}_n\mathbf{P},$$

with coefficients chosen to minimize the squared Frobenius norm of the operator approximation error. A six-order-parameter closure is considered in the appendix.

The projection coefficients are completely determined by normalized trace moments

$$m_k = \frac{1}{D}\operatorname{Tr}\big[(\mathbf{P}\mathbf{M}_s)^k\big].$$

For the four-order-parameter closure, only m_1, m_2, and m_3 are required. Random matrix theory determines the exact high-dimensional limits of these moments, which agree with empirical PCA and trained-LAE measurements in Figure 2a. The resulting dynamical closure remains approximate because it truncates the order-parameter hierarchy. The derivation and random-matrix calculations are given in Appendices A and B. We empirically assess accuracy of the higher order closure in Appendix A.6.

5 Results

We examine how pretraining sample size $\alpha = N/D$ and SFT label noise ϵ affect learning dynamics, asymptotic performance, and the benefit of SFT before RL. We then optimize compute allocation across the three training stages and extend the model to study how RL modifies individual skills in compositional tasks.

SFT and RL exhibit distinct learning dynamics. The closed ODEs closely match simulated order-parameter trajectories for both SFT and RL (Figure 2b–c). SFT begins learning immediately from stepwise supervision, whereas RL exhibits an initial plateau followed by rapid improvement. Under the all-correct reward criterion, an episode produces an update with probability P^T. Rewards are therefore rare when initial accuracy is low, but become more frequent as learning progresses. This accounts for the longer plateaus at larger episode lengths T reported in Patel et al. [2025]. Figure 2d shows accuracy throughout SFT followed by RL, when the pretraining level and time durations of SFT/RL are jointly optimized for different total compute budgets. We examine the benefits of SFT initialization below and the allocation problem in Section 5.1.

Unless otherwise stated, subsequent experiments use a locally optimal learning rate η_{opt} that maximizes the instantaneous predicted increase in generalization accuracy (see Appendix A.9).

SFT and RL attain the Bayes scaling exponent. Figure 3a shows that the asymptotic SFT and RL errors decrease as $\alpha^{-1/2}$, matching the scaling of the representation-dependent Bayes error despite remaining above it. For an isotropic teacher in the signal subspace, at large α, Appendix C gives

$$e_B(\alpha) = \frac{\Delta\sqrt{1-\rho}}{\pi(1-\Delta)}\alpha^{-1/2} + O(\alpha^{-1}).$$

This result identifies a mechanism by which increasing pretraining data produces power-law improvements in downstream performance. The $\alpha^{-1/2}$ scaling follows from the accuracy of the pretrained subspace: finite-sample covariance fluctuations leave an unrecoverable fraction of teacher-field variance of order α^{-1}, and the Bayes classification error scales as the square root of this fraction.

SFT warm starts improve asymptotic RL performance. Under the spherical constraint and locally optimal learning rate used in Figure 3a–b, increasing the duration of SFT reduces the asymptotic error of subsequent RL. Figure 3b reports $e_{\mathrm{RL}}^{\infty}(\xi_{\mathrm{SFT}}) - e_{\mathrm{RL}}^{\infty}(\infty)$, the excess error relative to RL initialized from converged SFT. For short warm starts, this excess error depends strongly on pretraining quality. At longer SFT durations, the curves approach a common power-law decay, indicating similar convergence towards their respective fully warm-started RL asymptotes.

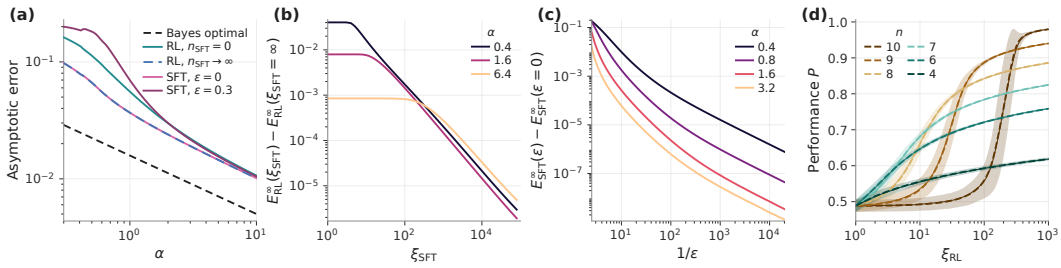


Figure 3: **Asymptotic performance predictably scales with batch size (a), degree of SFT warm start (b), label noise (c) and reward stringency (d).** (a) Power law scaling in asymptotic error reached by RL and SFT with higher/lower intermediate label noises ϵ, and Bayes optimal scaling (black dashed). (b) SFT warm start improves the RL asymptotic performance. The residual asymptotic RL error after infinite-time vs ξ_{SFT} SFT warm start is plotted against ξ_{SFT}, showing diminishing returns in accuracy. (c) The residual SFT asymptotic error for ϵ-noise vs 0-noise against inverse ϵ shows diminishing returns in label quality and robustness to label-noise. (d) Evolution of generalisation performance (P) is plotted for different levels of RL reward stringency, where 'n or more' actions in an episode are required correct for reward. We observe a speed accuracy tradeoff mediated by reward stringency. *Parameters*: (a), (b) and (c) Locally optimal learning rate, spherically constrained, $T = 6$, $\rho = 0.2$, $\Delta = 0.1$. (d) $T = 10$, $\alpha = 7$, $\eta = 1$, spherically unconstrained.

Pretraining improves SFT robustness to label noise. Figure 3c shows the excess asymptotic error due to label corruption, $\Delta e^{\infty}(\alpha, \epsilon) = e^{\infty}_{\mathrm{SFT}}(\alpha, \epsilon) - e^{\infty}_{\mathrm{SFT}}(\alpha, 0)$. At fixed ϵ, increasing α reduces this excess error, showing that better pretraining reduces sensitivity to corrupted supervision in addition to improving clean-label performance. For small ϵ, the approximately parallel curves indicate power-law scaling with a similar exponent across pretraining levels and a prefactor that decreases with α. For the unconstrained SFT dynamics, the Appendix C derives asymptotic SFT performance $e^{\infty}_{\mathrm{SFT}}(\alpha, \epsilon = 0) \sim \sqrt{2}\, e_B(\alpha)$. For fixed $0 < \epsilon < 1/2$, the SFT error instead approaches $e_\epsilon = \pi^{-1} \arccos(1 - 2\epsilon)$, with $e^{\infty}_{\mathrm{SFT}}(\alpha, \epsilon) - e_\epsilon = O(\alpha^{-1})$.

Reward stringency mediates a speed accuracy tradeoff. We consider the reward function that rewards an episode when at least n of its T decisions are correct, $R = \mathbf{1}(\sum_{t=1}^{T} \theta(y_t y_t^*) \geq n)$. As noted by Patel et al. [2025], Figure 3d shows that relaxing the all-correct criterion can accelerate early learning at the expense of asymptotic accuracy. With constant learning rate and unconstrained weights, the initial accuracy slope is larger by a factor $\binom{T-1}{n-1}$ than under the all-correct rule. However, the asymptotic error decays as $e_g(\xi) \propto (\log \xi)^{-1/(T-n)}$ for $n < T$, compared with $e_g(\xi) \propto \sqrt{\log \xi / \xi}$ for $n = T$. Less stringent reward therefore allows faster initial learning but slower asymptotic convergence. Intuitively, this is because early in learning rewards are more frequent with less stringent reward, but conversely near convergence episodes containing incorrect decisions are increasingly rewarded. With spherical normalization and constant learning rate, the small-error stationary solution instead scales as $e_g^{\infty} \propto (\eta/Q)^{1/(T-n+2)}$ as $\eta/Q \to 0$, where $Q = Q_s + Q_n$ is fixed. The derivations are provided in Appendix E.2.

SFT overcomes the RL plateau. Figure 4a compares RL trajectories initialized after different amounts of SFT. Increasing the duration of preceding SFT raises the initial probability of a successful episode to a higher degree, allowing subsequent RL to quickly improve performance. A corresponding qualitative effect appears in OLMo-2 1B [for AI , Ai2] finetuned on GSM8K [Cobbe et al., 2021a] (Figure 4b): models initialized from SFT checkpoints show faster reward improvement while the RL-only baseline initially remains near zero. Experimental details are in Appendix J. We next include the costs of all stages to evaluate the complete pipeline.

5.1 Optimal compute budget allocation

The previous section showed that fine-tuning performance is shaped by both the quality of the pretrained representation and the reliability of SFT supervision. We therefore extend our compute-allocation analysis to the full training pipeline, including pretraining, and ask how the optimal

allocation changes as the SFT data quality varies. The compute costs for SFT and RL are computed as defined in Appendix F.1.

Compute model and allocation objective. We count a multiply-add as two FLOPs. For an LAE of hidden width ρD, trained for M updates on a fixed batch of size αD, the assigned pretraining cost is $C_{\mathrm{PT}} = 10M\alpha\rho D^3$. The assigned SFT and RL cost per episode is $c_{\mathrm{ep}} = 2\rho TD^2$, including the frozen encoder projection and readout update. Thus a finetuning duration ξ costs $Dc_{\mathrm{ep}}\xi$ when $\xi = \mu/D$. The resulting budget constraint is

$$2\rho D^3(5M\alpha + T(\xi_{\mathrm{SFT}} + \xi_{\mathrm{RL}})) \leq \mathrm{B}_{\mathrm{FLOPs}}.$$

For convenience, we express compute in normalized units, $\widetilde{B} = B_{\mathrm{FLOPs}}/(2\rho D^3)$, so that the budget constraint becomes $5M\alpha + T(\xi_{\mathrm{SFT}} + \xi_{\mathrm{RL}}) \leq \widetilde{B}$. All compute budgets reported in the following figures are in these normalized units unless otherwise stated. We numerically optimize the final accuracy over the pretraining level and the SFT/RL durations, with SFT preceding RL. Costs and optimization details are given in Appendices F.1 and H.

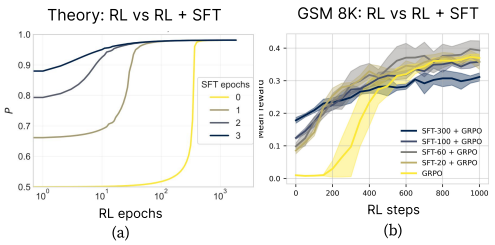


Figure 4: **SFT overcomes the RL Plateau.** (a) Pure RL exhibits a plateau due to reward sparsity (yellow). It is beneficial to first run SFT to overcome this plateau (navy) since SFT provides dense supervision without waiting for a successful episode. (b) We replicate that early SFT complementing RL by removing plateau in a realistic model and task using OlMo2 1B [OLMo et al., 2024] and GSM8K [Cobbe et al., 2021b]. We show training curves during RL after different amount of SFT over 3 seeds. *Parameters*: $\alpha = 4$, $T = 12$, $\rho = 0.2$, $\Delta = 0.1$.

Pretraining quality and budget determine the optimal SFT–RL split. Figure 5c shows the optimal fraction of finetuning compute allocated to RL at fixed pretraining level α. At a given budget, better pretraining increases the optimal RL fraction, whereas at fixed α, increasing the finetuning budget shifts the optimal allocation towards SFT. The optimality of increasing the proportion of preceding SFT compute is consistent with the warm-start dependence in Figure 3b: longer SFT can improve the asymptotic performance of subsequent RL, in addition to shortening its initial plateau.

Optimal budget allocation under varying conditions When compute is allocated jointly across all three stages, pretraining receives the largest share (Figure 5a–b). The optimal RL share initially increases with the total budget before decreasing (Figure 5a), demonstrating that additional compute need not favour a progressively larger RL allocation. Within the remaining fine-tuning budget, SFT remains important at low noise because it helps overcome the initial RL plateau through dense supervision (Figure 5a). At a fixed total budget, increasing intermediate-label noise increases the optimal RL share in Figure 5b. Intuitively, at zero label noise, SFT is compute optimal, and increasing noise reduces its effectiveness.

In Figure 5d, we compare the maximum achievable performance under three training paradigms: the full pretraining + SFT + RL pipeline (blue), pretraining + SFT only (green), and pretraining + RL only (pink). The full three-stage pipeline is compute-optimal the entire compute-limited regime, showing the benefit of combining SFT and RL as we demonstrated previously by showing SFT removes the early reward plateau from RL. At smaller compute budgets, SFT-only outperforms RL-only, and is compute optimal in the multi-stage case too, since it can overcome the RL plateau. But as compute budgets increase, the optimality of RL-only surpasses the SFT-only, since it avoids noisy SFT supervision. However, it still falls short of the full three-stage pipeline, due the increase in asymptotic performance afforded by SFT warm-start.

5.2 RL finetuning and skill amplification

So far, we have treated pretraining quality as a single scalar controlling the overall usefulness of the learned representation. In practice, however, pretrained models may contain many latent skills, only some of which are relevant to a given downstream task. Fine-tuning on a domain specific task, such as on math or coding, may therefore selectively amplify task-relevant skills rather than uniformly improve the full representation Zhao et al. [2025], Shao et al. [2025]. This motivates the question: how does the fine-tuning task specification influence the skills acquired during pretraining?

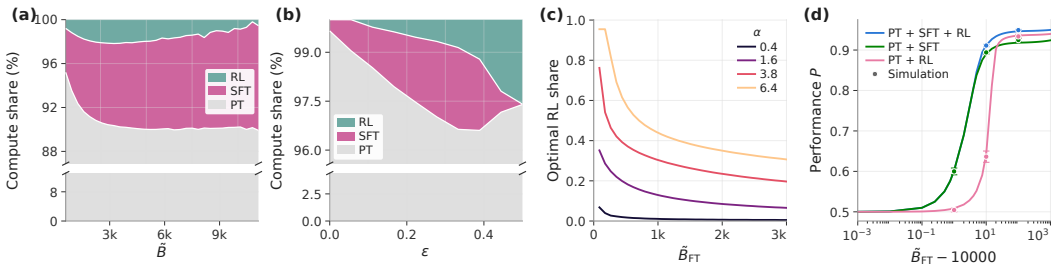


Figure 5: **Optimal allocation of compute across training stages**, (a) Optimal compute shares allocated to pretraining (gray), SFT (pink), and RL (teal) as the total budget varies. Most compute is allocated to pretraining; within the fine-tuning budget there is non-monotonicity in the optimal RL share. (b) Optimal compute shares at a fixed budget of 50000 flops as the SFT label noise ϵ varies. Higher SFT noise reduces the optimal share allocated to SFT and increases the share allocated to RL. (c) The optimal finetuning compute share to be allocated to RL under a given pretraining level α. With a strong pretrained model, more share of the RL compute is beneficial, but its optimal share decreases as the finetuning budget increases. (d) Best achievable performance under three training paradigms plotted against Total budget, offset by 18000 FLOPs: pretraining + SFT + RL (blue), pretraining + SFT only (green), and pretraining + RL only (pink). The full multistage pipeline performs best across all of the compute-limited regime, SFT only outperforms RL only for smaller budgets, reflecting the overcoming of the RL plateau, before RL overtakes at larger budgets. *Parameters*: $T = 6$, $\rho = 0.2$, $\Delta = 0.1$, 1000 pretraining update steps, spherically constrained, locally optimal learning rate. (a) $\epsilon = 0.2$. (c) $\epsilon = 0.3$. (d) $\epsilon = 0.3$, Simulation $D = 1000$.

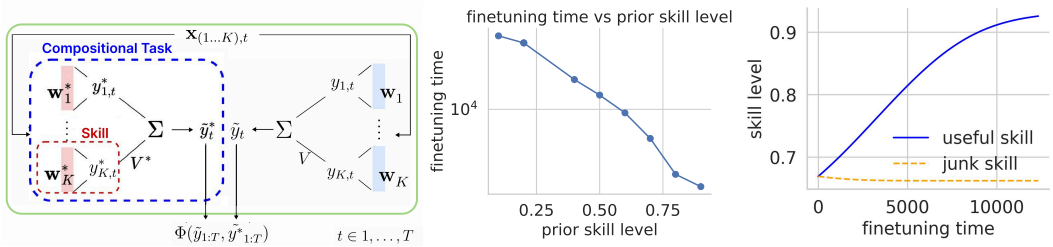
We address this question by extending our analytical model to study RL finetuning as learning to linearly compose multiple 'pretrained' skills into a more complex task [Lee et al., 2024]. Concretely, each skill is represented by one of K independent teacher-student pairs $(\mathbf{w}^*_{1\ldots K}, \mathbf{w}_{1\ldots K})$, and the downstream task is defined by a context teacher-student vector $(V^*, V \in \mathbb{R}^K)$ that selects and combines these skills. During fine-tuning, both the student skills and the context are updated. This setup allows some pretrained skills to be task-relevant, while others are present in the base model but unused by the downstream task. We provide the ODE derivation for this compositional RL setting in Appendix I. The notion of pretraining in this setup is how well each skill is learned before the downstream task.

Prior skill proficiency accelerates relevant RL finetuning. We first vary the initial proficiency of each skill and measure the fine-tuning time required to reach 99% of the maximal task performance. Fine-tuning converges exponentially faster as prior skill proficiency improves (Figure 6b), echoing the importance of a strong pretrained base model for RL efficiency.

RL amplifies task-relevant skills while preserving unused skills. We then study how task specification shapes individual skills during fine-tuning. We distinguish *useful skills*, which contribute to the downstream task, from *junk skills*, which are present in the base model but unused for finetuned task. As shown in Figure 6c, useful skills are further amplified, while junk skills initially deteriorate only mildly and are not catastrophically forgotten. This aligns with previous empirical observations that RL amplifies useful patterns learned during pretraining [Zhao et al., 2025, Shao et al., 2025], and predicts that unused skills may remain largely recoverable for future tasks where they become relevant.

6 Discussion

Our contribution is an analytically tractable starting point for studying multistage training pipelines in foundation models. Despite its simplicity, the model captures power-law error scaling, SFT–RL complementarity, warm-start dependence, and selective skill amplification. The dependence of downstream learning on pretraining quality is consistent with controlled experiments showing that stronger pretrained models improve faster and reach higher performance under RL [Shen et al., 2026], and that pretraining quality is a central bottleneck for downstream performance and compute efficiency [Akter et al., 2025, Yue et al., 2025, Ye et al., 2025]. The benefit of supervised preparation



(a) (b) (c)

Figure 6: **Skill composition in RL finetuning.** (a) Schematic of a compositional RL perceptron where individual skill is modeled as a single RL perceptron and compositional task is defined as a linear combination of those skills. (b) The fine-tuning time required to get the compositional task correct decreases as the pre-trained level of individual skill becomes better. (c) During the finetuning, the skill used in the finetuned compositional task further improves and is amplified. The junk skill previously acquired during pretraining but not used for the compositional task persists, not forgotten. *Parameters*: $\eta = 1$, $T = 8$, $K = 8$, $J_{\text{init}} = 0.5$.

is consistent with empirical findings that stronger SFT initialization can improve subsequent RL [Liu et al., 2025b].

Our allocation analysis excludes the cost of acquiring high-quality supervision, which can substantially improve post-training [Zhou et al., 2023, Ye et al., 2025] but is difficult to obtain at scale. In our model, proportional reductions in label noise yield diminishing absolute improvements in SFT error, while better pretraining reduces sensitivity to noise. Accounting for data acquisition and verification costs could provide more practical insights into when improving supervision is preferable to additional pretraining or RL.

Broadly, by making the interactions between pretraining, SFT, and RL analytically tractable, our model provides a starting point for understanding how these stages jointly impact learning performance.

Limitations: Our model is a simplified and abstract setup, where each component is modeled as a simple single-layer network. The real-world models rely on complex architectures and heavy engineering, which we do not account for. Our proposed model of RL captures an aspect of bulk feedback in RL, but it assumes independence between all observations in an episode. In practice, the cost of SFT is not just a compute one, but the cost of being able to source enough high quality supervision data we do not take into account the difficulty of obtaining the labels of the data for defining compute budget in our analysis. Using frozen encoder weights is a simplification, in practice all weights are updated during fine-tuning.

Acknowledgements

We thank the following funding sources: Gatsby Charitable Foundation (GAT4058) to AS; Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) to AS; Schmidt Science Polymath Award to AS. AS is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains program. The authors declare no competing interests.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Syeda Nahida Akter, Shrimai Prabhumoye, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, Yejin Choi, and Bryan Catanzaro. Front-loading reasoning: The synergy between pretraining and post-training data. *arXiv preprint arXiv:2510.03264*, 2025.

Haruka Asanuma, Shiro Takagi, Yoshihiro Nagano, Yuki Yoshida, Yasuhiko Igarashi, and Masato Okada. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *Journal of the Physical Society of Japan*, 90(10):104001, 2021.

Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual review of condensed matter physics*, 11(1):501–528, 2020.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. arxiv 2012. *arXiv preprint arXiv:1206.5538*, 10, 2012.

Michael Biehl and Holm Schwarze. Learning drifting concepts with neural networks. *Journal of Physics A: Mathematical and General*, 26(11):2651, 1993.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yuri Chervonyi, Trieu H Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang H Nguyen, Marcelo Menegali, Junehyuk Jung, Junsu Kim, Vikas Verma, Quoc V Le, et al. Gold-medalist performance in solving olympiad geometry with alphageometry2. *Journal of Machine Learning Research*, 26 (241):1–39, 2025.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10818–10838, 2025. URL `https://proceedings.mlr.press/v267/chu25c.html`.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021a.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021b.

Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,

Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pages 1339–1348. PMLR, 2018.

Andreas Engel. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Allen Institute for AI (Ai2) et al. 2 olmo 2 furious (colm's version). *arXiv preprint arXiv:2501.00656*, 2025. URL `https://arxiv.org/abs/2501.00656`.

Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

Federica Gerace, Luca Saglietti, Stefano Sarao Mannelli, Andrew Saxe, and Lenka Zdeborová. Probing transfer learning with a model of synthetic correlated datasets. *Machine Learning: Science and Technology*, 3(1):015030, 2022.

Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10 (4):041044, 2020.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. URL `https://arxiv.org/abs/2403.14608`.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in SFT-RL post-training: When high SFT scores mislead and what to use instead. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5d6ae8ba43ecb378030753c4408ef9bd-Abstract-Conference.html.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

Andrew K Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. *arXiv preprint arXiv:1809.10374*, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1):233–264, 2011.

Jin Hwa Lee, Stefano Sarao Mannelli, and Andrew Saxe. Why do animals need shaping? a theory of task composition and curriculum learning. *arXiv preprint arXiv:2402.18361*, 2024.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pages 6109–6119. PMLR, 2021.

Sebastian Lee, Stefano Sarao Mannelli, Claudia Clopath, Sebastian Goldt, and Andrew Saxe. Maslow's hammer for catastrophic forgetting: Node re-use vs node activation. *arXiv preprint arXiv:2205.09029*, 2022.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025a.

Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl synergy. *arXiv preprint arXiv:2506.13284*, 2025b.

Suneel Nadipalli. Layer-wise evolution of representations in fine-tuned transformers: Insights from sparse autoencoders. *arXiv preprint arXiv:2502.16722*, 2025.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Nishil Patel, Sebastian Lee, Stefano Sarao Mannelli, Sebastian Goldt, and Andrew Saxe. Rl perceptron: Generalization dynamics of policy learning in high dimensions. *Phys. Rev. X*, 15:021051, May 2025. doi: 10.1103/PhysRevX.15.021051. URL `https://link.aps.org/doi/10.1103/PhysRevX.15.021051`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519. URL `http://dx.doi.org/10.1037/h0042519`.

David Saad and Sara A. Solla. Exact solution for on-line learning in multilayer neural networks. *Phys. Rev. Lett.*, 74:4337–4340, May 1995. doi: 10.1103/PhysRevLett.74.4337. URL `https://link.aps.org/doi/10.1103/PhysRevLett.74.4337`.

Luca Saglietti, Stefano Mannelli, and Andrew Saxe. An analytical theory of curriculum learning in teacher-student networks. *Advances in Neural Information Processing Systems*, 35:21113–21127, 2022.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

Erwin Schrödinger. Quantisierung als eigenwertproblem. *Annalen der physik*, 385(13):437–490, 1926.

Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, et al. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jingyan Shen, Ang Li, Salman Rahman, Yifan Sun, Micah Goldblum, Matus Telgarsky, and Pavel Izmailov. Understanding reasoning from pretraining to post-training. *arXiv preprint arXiv:2607.16097*, 2026.

Jack W Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Yizhou Xu, Margarita Sagitova, Lenka Zdeborová, and Florent Krzakala. High-dimensional learning dynamics of attention-indexed models. *arXiv preprint arXiv:2609.03858*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

Yuki Yoshida and Masato Okada. Data-dependence of plateau phenomenon in learning with neural network—statistical mechanical analysis. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and RL on reasoning language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 157861–157894, 2026a. URL `https://proceedings.mlr.press/v306/zhang26en.html`.

Dylan Zhang, Yufeng Xu, Haojin Wang, Qingzhi Chen, and Hao Peng. Good SFT optimizes for SFT, better SFT prepares for reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 159617–159640, 2026b. URL `https://proceedings.mlr.press/v306/zhang26hd.html`.

Haoyu Zhao, Simran Kaur, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Can models learn skill composition from examples? *Advances in Neural Information Processing Systems*, 37: 102393–102427, 2024.

Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

All code and instructions to reproduce the figures are available at `https://github.com/nishp99/multistage_pipeline.git`.

.1 Teacher-student framework

There is a growing body of neural network theory of learning and generalization based on a teacher-student framework that applies tools from statistical physics, which allows us to characterize the typical behavior in high-dimensional stochastic systems Gardner and Derrida [1989], Seung et al. [1992], Engel [2001], Bahri et al. [2020], Zdeborová and Krzakala [2016]. Previous works studied diverse learning paradigms and setups, including vanilla supervised learning Du et al. [2018], Goldt et al. [2019], Aubin et al. [2018], Advani et al. [2020], continual learning Lee et al. [2021, 2022], Asanuma et al. [2021], curriculum learning Saglietti et al. [2022], Lee et al. [2024], transfer learning Gerace et al. [2022], Lampinen and Ganguli [2018] and online policy learning Patel et al. [2025].

A Derivation of ODEs

A.1 Thermodynamic Limit

We follow the standard procedure of Saad and Solla, but the main difference is the random projection which requires averaging over. It was shown by Yoshida and Okada [2019] that macroscopically describing the dynamics when data is sampled from a centred Gaussian generates an infinite hierarchy of order parameters to track, and the ODE system can only be closed if the covariance has a finite number of unique eigenvalues, where there is a pair of order parameters associated to every unique eigenvalue. In our case, the projection onto the empirical principal subspace introduces an effective data covariance with, in general, ρD unique eigenvalues. We prevent the infinite cascade of ODEs by approximating the higher-order order parameters in terms of lower order ones using the Hilbert Schmidt Operator projection.

We denote by $\mathbf{M}_s$ and $\mathbf{M}_n = \mathbf{I} - \mathbf{M}_s$ the orthogonal projectors onto the signal and noise subspaces, respectively. The data covariance is

$$\Sigma = \mathbf{M}_s + \Delta \mathbf{M}_n, \tag{7}$$

where $0 < \Delta < 1$. The teacher $\mathbf{w}^*$ is assumed to lie entirely in the signal subspace, so that $\mathbf{M}_n \mathbf{w}^* = 0$. The student does not observe $\mathbf{x}$ directly, but instead observes the projected input $\mathbf{P}\mathbf{x}$, where $\mathbf{P}$ is the orthogonal projector onto the PCA subspace learned during pretraining.

We define the following order parameters:

$$J_i = \frac{\mathbf{w}^\top \mathbf{P} \mathbf{M}_i \mathbf{P} \mathbf{w}^*}{D}, \qquad Q_i = \frac{\mathbf{w}^\top \mathbf{P} \mathbf{M}_i \mathbf{P} \mathbf{w}}{D}, \qquad S_i = \frac{\mathbf{w}^{*\top} \mathbf{P} \mathbf{M}_i \mathbf{P} \mathbf{w}^*}{D}, \tag{8}$$

for $i \in \{s, n\}$, together with

$$S = \frac{\mathbf{w}^{*\top} \mathbf{w}^*}{D}. \tag{9}$$

The total student-teacher overlap observed through the projected representation is

$$J = J_s + J_n = \frac{\mathbf{w}^\top \mathbf{P} \mathbf{w}^*}{D}, \tag{10}$$

and the variance of the student field is

$$\widehat{Q} = Q_s + \Delta Q_n. \tag{11}$$

We define the local fields

$$\lambda = \lambda_s + \lambda_n = \frac{\mathbf{w}^\top \mathbf{P} \mathbf{x}}{\sqrt{D}}, \qquad \nu = \frac{\mathbf{w}^{*\top} \mathbf{x}}{\sqrt{D}}, \tag{12}$$

and the projected subfields

$$\lambda_i = \frac{\mathbf{w}^\top \mathbf{P} \mathbf{M}_i \mathbf{P} \mathbf{x}}{\sqrt{D}}, \qquad \tilde{\nu}_i = \frac{\mathbf{w}^{*\top} \mathbf{P} \mathbf{M}_i \mathbf{P} \mathbf{x}}{\sqrt{D}}, \qquad i \in \{s, n\}. \tag{13}$$

The student and teacher actions are

$$y = \mathrm{sgn}(\lambda), \qquad y^* = \mathrm{sgn}(\nu), \tag{14}$$

with noisy labels in SFT obtained by flipping y^* with probability ϵ.

The fields are jointly Gaussian with zero mean. The covariances needed below are

$$\mathrm{Var}(\lambda) = \widehat{Q} = Q_s + \Delta Q_n, \qquad \mathrm{Var}(\nu) = S, \qquad \mathrm{Cov}(\lambda, \nu) = J_s + J_n. \tag{15}$$

We also need

$$\mathrm{Cov}(\nu, \lambda_i) = J_i, \qquad \mathrm{Cov}(\nu, \tilde{\nu}_i) = S_i. \tag{16}$$

Because $\mathbf{P}$ does not in general commute with $\mathbf{M}_s$ and $\mathbf{M}_n$, the remaining covariances involve higher-order operator products:

$$Q_{ij} = \frac{\mathbf{w}^\top \mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\mathbf{P}\mathbf{w}}{D}, \tag{17}$$

$$J_{ij} = \frac{\mathbf{w}^\top \mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\mathbf{P}\mathbf{w}^*}{D}. \tag{18}$$

Using these, we have

$$\mathrm{Cov}(\lambda, \lambda_s) = Q_{ss} + \Delta Q_{sn}, \qquad \mathrm{Cov}(\lambda, \lambda_n) = Q_{sn} + \Delta Q_{nn}, \tag{19}$$

$$\mathrm{Cov}(\lambda, \tilde{\nu}_s) = J_{ss} + \Delta J_{ns}, \qquad \mathrm{Cov}(\lambda, \tilde{\nu}_n) = J_{sn} + \Delta J_{nn}. \tag{20}$$

A.2 Closure of Higher-Order Projector Products

The non-commutativity of $\mathbf{P}$ with $\mathbf{M}_s$ and $\mathbf{M}_n$ generates terms such as

$$\mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\mathbf{P}.$$

To close the dynamics, we project these operators onto the two-dimensional operator basis

$$\mathbf{B}_s = \mathbf{P}\mathbf{M}_s\mathbf{P}, \qquad \mathbf{B}_n = \mathbf{P}\mathbf{M}_n\mathbf{P}.$$

Thus we approximate

$$\mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\mathbf{P} \approx \Omega_{ij}^{(s)}\mathbf{P}\mathbf{M}_s\mathbf{P} + \Omega_{ij}^{(n)}\mathbf{P}\mathbf{M}_n\mathbf{P}. \tag{21}$$

This gives the closed approximations

$$Q_{ij} \approx \Omega_{ij}^{(s)} Q_s + \Omega_{ij}^{(n)} Q_n, \tag{22}$$

$$J_{ij} \approx \Omega_{ij}^{(s)} J_s + \Omega_{ij}^{(n)} J_n. \tag{23}$$

The coefficients $\Omega_{ij}^{(s)}$ and $\Omega_{ij}^{(n)}$ are obtained by Hilbert-Schmidt projection, which we will detail now. The purpose of the Hilbert-Schmidt projection is not to assert that $\mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\mathbf{P}$ is exactly a linear combination of $\mathbf{P}\mathbf{M}_s\mathbf{P}$ and $\mathbf{P}\mathbf{M}_n\mathbf{P}$. Rather, it gives the unique best approximation to this operator within the space already represented by the order parameters. More precisely, define

$$\mathbf{U}_{ij} = \mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\mathbf{P}, \qquad \mathbf{V}_s = \mathbf{P}\mathbf{M}_s\mathbf{P}, \qquad \mathbf{V}_n = \mathbf{P}\mathbf{M}_n\mathbf{P}.$$

We choose $\Omega_{ij}^{(s)}$ and $\Omega_{ij}^{(n)}$ to minimize

$$\frac{1}{D}\mathrm{Tr}\left[\left(\mathbf{U}_{ij} - (\Omega_{ij}^{(s)}\mathbf{V}_s + \Omega_{ij}^{(n)}\mathbf{V}_n)\right)^2\right].$$

The discarded component, or residual operator, is therefore

$$\mathbf{\Delta}_{ij} = \mathbf{U}_{ij} - (\Omega_{ij}^{(s)}\mathbf{V}_s + \Omega_{ij}^{(n)}\mathbf{V}_n). \tag{24}$$

By construction, this residual is orthogonal, in trace average, to all tracked operators: $\frac{1}{D}\mathrm{Tr}(\mathbf{\Delta}_{ij}\mathbf{V}_s) = 0, \qquad \frac{1}{D}\mathrm{Tr}(\mathbf{\Delta}_{ij}\mathbf{V}_n) = 0$. This orthogonality does not ensure that the residual contractions with the student and teacher vanish: training can generate correlations with the eigendirections of $\mathbf{\Delta}_{ij}$.

Vanishing residual contractions would yield an exact quadratic-form closure, whereas residuals that remain $O(1)$ require additional macroscopic order parameters. We therefore treat the four-order-parameter dynamics as approximate. Appendix A.6 tests this approximation numerically, finding small relative residuals whose disorder-averaged magnitudes decrease with dimension over the tested range, without establishing self-averaging to zero.

Define the normalized trace constants

$$t = \frac{1}{D}\mathrm{Tr}(\mathbf{P}), \qquad t_s = \frac{1}{D}\mathrm{Tr}(\mathbf{P}\mathbf{M}_s), \tag{25}$$

$$T_{ss} = \frac{1}{D}\mathrm{Tr}(\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s), \qquad \tau_{sss} = \frac{1}{D}\mathrm{Tr}(\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s). \tag{26}$$

Since $\mathbf{M}_s + \mathbf{M}_n = \mathbf{I}$ and $\mathbf{P}^2 = \mathbf{P}$, all other second- and third-order trace constants can be written as

$$T_{sn} = t_s - T_{ss}, \tag{27}$$

$$T_{nn} = t - 2t_s + T_{ss}, \tag{28}$$

and

$$V_{ss} = \tau_{sss}, \tag{29}$$

$$V_{sn} = T_{ss} - \tau_{sss}, \tag{30}$$

$$V_{nn} = t_s - 2T_{ss} + \tau_{sss}, \tag{31}$$

where

$$T_{ij} = \frac{1}{D}\mathrm{Tr}\left(\mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\right), \tag{32}$$

$$V_{ij} = \frac{1}{D}\mathrm{Tr}\left(\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\right). \tag{33}$$

Let

$$\Delta_G = T_{ss}t - t_s^2. \tag{34}$$

Then the projection coefficients are

$$\Omega_{ij}^{(s)} = \frac{(t - t_s)V_{ij} - (t_s - T_{ss})T_{ij}}{\Delta_G}, \tag{35}$$

$$\Omega_{ij}^{(n)} = \frac{-t_s V_{ij} + T_{ss}T_{ij}}{\Delta_G}. \tag{36}$$

Explicitly,

$$\Omega_{ss}^{(s)} = \frac{(t - t_s)\tau_{sss} - (t_s - T_{ss})T_{ss}}{\Delta_G}, \tag{37}$$

$$\Omega_{sn}^{(s)} = \frac{(t - t_s)(T_{ss} - \tau_{sss}) - (t_s - T_{ss})(t_s - T_{ss})}{\Delta_G}, \tag{38}$$

$$\Omega_{nn}^{(s)} = \frac{(t - t_s)(t_s - 2T_{ss} + \tau_{sss}) - (t_s - T_{ss})(t - 2t_s + T_{ss})}{\Delta_G}, \tag{39}$$

and

$$\Omega_{ss}^{(n)} = \frac{-t_s\tau_{sss} + T_{ss}^2}{\Delta_G}, \tag{40}$$

$$\Omega_{sn}^{(n)} = \frac{-t_s(T_{ss} - \tau_{sss}) + T_{ss}(t_s - T_{ss})}{\Delta_G}, \tag{41}$$

$$\Omega_{nn}^{(n)} = \frac{-t_s(t_s - 2T_{ss} + \tau_{sss}) + T_{ss}(t - 2t_s + T_{ss})}{\Delta_G}. \tag{42}$$

A useful check is

$$\Omega_{nn}^{(s)} = \Omega_{ss}^{(s)} - 1, \qquad \Omega_{nn}^{(n)} = \Omega_{ss}^{(n)} + 1, \tag{43}$$

which indeed holds, and follows from

$$\mathbf{P}\mathbf{M}_n\mathbf{P}\mathbf{M}_n\mathbf{P} = \mathbf{P} - 2\mathbf{P}\mathbf{M}_s\mathbf{P} + \mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s\mathbf{P}.$$

The trace constants are static during finetuning. They depend only on the pretraining projector $\mathbf{P}$ and on the true signal/noise decomposition. In the typical-case theory, these constants are self-averaging and are averaged over the randomness in the pretraining batch.

A.3 Projected RL Dynamics

The RL update is

$$\mathbf{w}^{(\mu+1)} = \mathbf{w}^{(\mu)} + \frac{\eta}{T}\left(\sum_{t=1}^{T}\frac{y_t^{(\mu)}\mathbf{P}\mathbf{x}_t^{(\mu)}}{\sqrt{D}}\right)R(y_{1:T}^{(\mu)}, y^{*(\mu)}_{1:T}), \tag{44}$$

$$\mathbf{w}^{(\mu+1)} = \mathbf{w}^{(\mu)} + \frac{\eta}{TG}\sum_{g=1}^{G}\left[\left(\sum_{t=1}^{T}\frac{y_t^{g,(\mu)}\mathbf{P}\mathbf{x}_t^{g,(\mu)}}{\sqrt{D}}\right)(R(y_{1:T}^{g,(\mu)}, y^{*g,(\mu)}_{1:T}) - A^{(\mu)})\right], \tag{45}$$

with $A = \frac{1}{G}\sum_{g=1}^{G}(R(y_{1:T}^{g,(\mu)}, y^{*g,(\mu)}_{1:T}))$ and $R = 2\left(\prod_{t=1}^{T}\theta(\nu_t\lambda_t)\right) - 1$

where for the all-correct reward rule,

$$R = \prod_{t=1}^{T}\theta(\nu_t\lambda_t). \tag{46}$$

Taking the inner product of the update with $\mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{w}^*$, summing over a window of episodes, and taking the limit $D \to \infty$, gives

$$\frac{dJ_i}{d\xi} = \frac{\eta}{T}\left\langle\sum_{t=1}^{T}\mathrm{sgn}(\lambda_t)\tilde{\nu}_{i,t}R\right\rangle. \tag{47}$$

Similarly, squaring the update gives

$$\frac{dQ_i}{d\xi} = \frac{2\eta}{T}\left\langle\sum_{t=1}^{T}\mathrm{sgn}(\lambda_t)\lambda_{i,t}R\right\rangle + \frac{\eta^2}{T}\frac{1}{D}\mathrm{Tr}\left(\mathbf{P}\mathbf{M}_i\mathbf{P}\Sigma\right)\langle R\rangle + \mathcal{O}\left(\frac{1}{D}\right). \tag{48}$$

The trace terms are

$$G_s = \frac{1}{D}\mathrm{Tr}\left(\mathbf{P}\mathbf{M}_s\mathbf{P}\Sigma\right) = T_{ss} + \Delta T_{sn}, \tag{49}$$

$$G_n = \frac{1}{D}\mathrm{Tr}\left(\mathbf{P}\mathbf{M}_n\mathbf{P}\Sigma\right) = T_{sn} + \Delta T_{nn}. \tag{50}$$

Let

$$P = \langle\theta(\nu\lambda)\rangle = \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left(\frac{J_s + J_n}{\sqrt{S(Q_s + \Delta Q_n)}}\right). \tag{51}$$

This is the probability that a single step is correct (the generalisation performance). And is trivially computed from the Gaussian statistics of the fields. For the all-correct reward rule, $\langle R\rangle = P^T$. Also,

$$\mathrm{sgn}(\lambda)\theta(\nu\lambda) = \frac{1}{2}\left[\mathrm{sgn}(\lambda) + \mathrm{sgn}(\nu)\right]. \tag{52}$$

Therefore,

$$\frac{dJ_i}{d\xi} = \frac{\eta}{2}P^{T-1}\left(\langle\tilde{\nu}_i\mathrm{sgn}(\lambda)\rangle + \langle\tilde{\nu}_i\mathrm{sgn}(\nu)\rangle\right), \tag{53}$$

and

$$\frac{dQ_i}{d\xi} = \eta P^{T-1}\left(\langle\lambda_i\mathrm{sgn}(\lambda)\rangle + \langle\lambda_i\mathrm{sgn}(\nu)\rangle\right) + \frac{\eta^2}{T}G_iP^T. \tag{54}$$

A.4 Computing the Gaussian Averages

For zero-mean jointly Gaussian variables X, Y,

$$\langle X\mathrm{sgn}(Y)\rangle = \sqrt{\frac{2}{\pi}}\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(Y)}}. \tag{55}$$

Using the covariance structure above, define

$$C_s^Q = Q_{ss} + \Delta Q_{sn}, \qquad C_n^Q = Q_{sn} + \Delta Q_{nn}, \tag{56}$$

$$C_s^J = J_{ss} + \Delta J_{sn}, \qquad C_n^J = J_{sn} + \Delta J_{nn}. \tag{57}$$

Then

$$\langle \tilde{\nu}_i \mathrm{sgn}(\lambda)\rangle = \sqrt{\frac{2}{\pi}}\frac{C_i^J}{\sqrt{\widehat{Q}}}, \tag{58}$$

$$\langle \tilde{\nu}_i \mathrm{sgn}(\nu)\rangle = \sqrt{\frac{2}{\pi}}\frac{S_i}{\sqrt{S}}, \tag{59}$$

$$\langle \lambda_i \mathrm{sgn}(\lambda)\rangle = \sqrt{\frac{2}{\pi}}\frac{C_i^Q}{\sqrt{\widehat{Q}}}, \tag{60}$$

$$\langle \lambda_i \mathrm{sgn}(\nu)\rangle = \sqrt{\frac{2}{\pi}}\frac{J_i}{\sqrt{S}}. \tag{61}$$

Here Q_{ij} and J_{ij} are replaced by their closed forms from 21:

$$Q_{ij} = \Omega_{ij}^{(s)} Q_s + \Omega_{ij}^{(n)} Q_n, \tag{62}$$

$$J_{ij} = \Omega_{ij}^{(s)} J_s + \Omega_{ij}^{(n)} J_n. \tag{63}$$

Substituting these averages into 53 and 54 gives the closed RL ODEs:

$$\frac{dJ_i}{d\xi} = \frac{\eta}{2} P^{T-1}\sqrt{\frac{2}{\pi}}\left(\frac{C_i^J}{\sqrt{\widehat{Q}}} + \frac{S_i}{\sqrt{S}}\right), \tag{64}$$

$$\frac{dQ_i}{d\xi} = \eta P^{T-1}\sqrt{\frac{2}{\pi}}\left(\frac{C_i^Q}{\sqrt{\widehat{Q}}} + \frac{J_i}{\sqrt{S}}\right) + \frac{\eta^2}{T} G_i P^T, \qquad i \in \{s, n\}. \tag{65}$$

A.5 Projected SFT Dynamics

For SFT, the update is

$$\mathbf{w}^{(\mu+1)} = \mathbf{w}^{(\mu)} - \frac{\eta}{T\sqrt{D}} \sum_{t=1}^{T} y_t^{(\mu)} \mathbf{P}\mathbf{x}_t^{(\mu)} \theta\left(-y_t^{*(\mu)} y_t^{(\mu)}\right). \tag{66}$$

The label y^* is noisy: with probability $1-\epsilon$, $y^* = \mathrm{sgn}(\nu)$, and with probability ϵ, $y^* = -\mathrm{sgn}(\nu)$. Following the same thermodynamic-limit procedure,

$$\frac{dJ_i}{d\xi} = -\eta \left\langle \mathrm{sgn}(\lambda)\tilde{\nu}_i \theta(-yy^*)\right\rangle, \tag{67}$$

and

$$\frac{dQ_i}{d\xi} = -2\eta \left\langle \mathrm{sgn}(\lambda)\lambda_i \theta(-yy^*)\right\rangle + \frac{\eta^2}{T} G_i \left\langle \theta(-yy^*)\right\rangle. \tag{68}$$

Using

$$\mathrm{sgn}(\lambda)\theta(-yy^*) = \frac{1}{2}\left[\mathrm{sgn}(\lambda) - y^*\right], \tag{69}$$

and averaging over the label noise gives

$$\langle Xy^* \rangle = (1-2\epsilon)\,\langle X\mathrm{sgn}(\nu)\rangle\,. \tag{70}$$

The probability of matching the noisy label is

$$P_\epsilon = (1-2\epsilon)P + \epsilon. \tag{71}$$

Therefore,

$$\frac{dJ_i}{d\xi} = -\frac{\eta}{2}\left(\langle\tilde{\nu}_i\mathrm{sgn}(\lambda)\rangle - (1-2\epsilon)\,\langle\tilde{\nu}_i\mathrm{sgn}(\nu)\rangle\right), \tag{72}$$

$$\frac{dQ_i}{d\xi} = -\eta\left(\langle\lambda_i\mathrm{sgn}(\lambda)\rangle - (1-2\epsilon)\,\langle\lambda_i\mathrm{sgn}(\nu)\rangle\right) + \frac{\eta^2}{T}G_i(1-P_\epsilon). \tag{73}$$

Substituting the Gaussian averages gives the closed SFT ODEs:

$$\frac{dJ_i}{d\xi} = -\frac{\eta}{2}\sqrt{\frac{2}{\pi}}\left(\frac{C_i^J}{\sqrt{\widehat{Q}}} - (1-2\epsilon)\frac{S_i}{\sqrt{S}}\right), \tag{74}$$

$$\frac{dQ_i}{d\xi} = -\eta\sqrt{\frac{2}{\pi}}\left(\frac{C_i^Q}{\sqrt{\widehat{Q}}} - (1-2\epsilon)\frac{J_i}{\sqrt{S}}\right) + \frac{\eta^2}{T}G_i(1-P_\epsilon), \qquad i\in\{s,n\}. \tag{75}$$

A.6 Magnitude and dimension scaling of closure residuals

Only two residual contractions need to be tracked, one for J and one for Q. To see this, recall $\mathbf{U}_{ij} = \mathbf{P}\mathbf{M}_i\mathbf{P}\mathbf{M}_j\mathbf{P}$ and $\mathbf{V}_i = \mathbf{P}\mathbf{M}_i\mathbf{P}$. Since $\mathbf{M}_n = \mathbf{I} - \mathbf{M}_s$ and $\mathbf{P}^2 = \mathbf{P}$,

$$\mathbf{U}_{sn} = \mathbf{U}_{ns} = \mathbf{V}_s - \mathbf{U}_{ss}, \qquad \mathbf{U}_{nn} = \mathbf{V}_n - \mathbf{V}_s + \mathbf{U}_{ss}.$$

The Hilbert–Schmidt projection onto $\mathrm{span}\{\mathbf{V}_s, \mathbf{V}_n\}$ is linear and leaves both basis operators unchanged. Subtracting the projected operators therefore gives $\mathbf{\Delta}_{sn} = \mathbf{\Delta}_{ns} = -\mathbf{\Delta}_{ss}$ and $\mathbf{\Delta}_{nn} = \mathbf{\Delta}_{ss}$. Consequently, the signed contraction residuals $r_{ij}^{(J)} = D^{-1}\mathbf{w}^\top\mathbf{\Delta}_{ij}\mathbf{w}^*$ and $r_{ij}^{(Q)} = D^{-1}\mathbf{w}^\top\mathbf{\Delta}_{ij}\mathbf{w}$ satisfy

$$r_{ss}^{(C)} = r_{nn}^{(C)} = -r_{sn}^{(C)} = -r_{ns}^{(C)}, \qquad C\in\{J,Q\}.$$

Thus $r_{ss}^{(J)}$ and $r_{ss}^{(Q)}$ determine all eight contraction residuals exactly, at every training time. These identities concern the signed residuals before relative normalization; the relative residuals can differ because their denominators depend on the contraction.

To assess the Hilbert–Schmidt closure, we compare the exact signal–signal contractions $C_{\mathrm{exact}} \in \{J_{ss}, Q_{ss}\}$ with their closed approximations $C_{\mathrm{closed}} = \Omega_{ss}^{(s)}C_s + \Omega_{ss}^{(n)}C_n$, using the symmetric relative residual

$$r_{\mathrm{rel}}(\xi;D) = \frac{C_{\mathrm{exact}} - C_{\mathrm{closed}}}{|C_{\mathrm{exact}}| + |C_{\mathrm{closed}}| + 10^{-12}}. \tag{76}$$

Figure 7B shows that, after the initial transient, the mean absolute relative residual remains small and decreases with dimension. This motivates fitting $\langle|r_{\mathrm{rel}}(\xi;D)|\rangle \propto D^{\beta(\xi)}$ at each training time (panel A), with absolute values taken before averaging over realizations. Negative β indicates decreasing residual magnitude over the tested dimensions; $\beta = -1/2$ is the usual $D^{-1/2}$ self-averaging benchmark. The late-time Q scaling is broadly compatible with this benchmark, while J shows a persistently shallower decrease. These finite-size results support the accuracy of the approximate closure in the simulated regime, but do not establish that the residuals self-average to zero as $D\to\infty$.

A.7 Generalization Performance

At any point in training, the probability that the student agrees with the clean teacher on a single example is

$$P = \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left(\frac{J_s + J_n}{\sqrt{S(Q_s + \Delta Q_n)}}\right). \tag{77}$$

This expression depends only on the total projected student-teacher overlap $J_s + J_n$, the teacher norm S, and the effective student variance $Q_s + \Delta Q_n$.

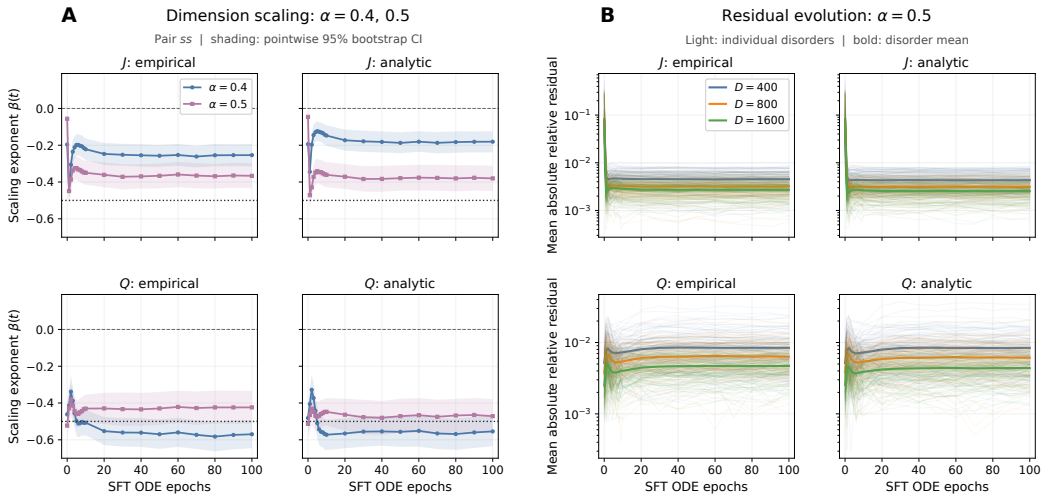


Figure 7: **Evolution and dimension scaling of closure residuals during spherical SFT in the projector limit.** Rows show the J_{ss} and Q_{ss} contractions; columns within each block use empirical Hilbert–Schmidt or analytic RMT coefficients. **(A)** Exponents fitted using $D = 400, 800, 1600$ at $\alpha = 0.4, 0.5$, with pointwise 95% bootstrap confidence intervals over disorders. Dotted and dashed lines mark $\beta = -1/2$ and 0. **(B)** Absolute relative residuals at $\alpha = 0.5$ (logarithmic axis). Faint curves average four PCA projections per fixed teacher/student disorder; bold curves additionally average over 128 disorders at each D. Parameters: $\epsilon = 0$, $\eta = 1$, $T = 5$, $\rho_{\text{noise}} = 0.8$, $\Delta = 0.1$. One ODE epoch comprises D microscopic updates; no finite-pretraining dynamics are included.

A.8 Optional Spherical Constraint

If the student norm in the projected feature space is constrained to remain fixed,

$$Q_s + Q_n = R,$$

then the unconstrained increments may be projected back onto the constraint manifold. If

$$dQ_{\text{unc}} = dQ_s + dQ_n,$$

then the constrained increments used in the implementation are

$$dJ_i \leftarrow dJ_i - \frac{J_i}{2R} dQ_{\text{unc}}, \tag{78}$$

$$dQ_i \leftarrow dQ_i - \frac{Q_i}{R} dQ_{\text{unc}}, \qquad i \in \{s, n\}. \tag{79}$$

This is a Lagrange-multiplier projection that removes the radial component of the update while preserving the projected order-parameter dynamics tangential to the fixed-norm surface.

A.9 Locally Optimal Learning Rate

In some experiments, rather than using a fixed learning rate, we choose a locally optimal learning rate at each integration step. The learning rate is chosen to maximize the instantaneous predicted increase in the clean single-step generalization probability

$$P = \frac{1}{2} + \frac{1}{\pi} \sin^{-1}\left(\frac{J_s + J_n}{\sqrt{S(Q_s + \Delta Q_n)}} \right). \tag{80}$$

For SFT with label noise, the dynamics use the noisy-label probability

$$P_\epsilon = (1 - 2\epsilon)P + \epsilon, \tag{81}$$

but the locally optimal learning rate is chosen to improve the clean probability P. This corresponds to choosing the learning rate that gives the greatest predicted improvement in true teacher-student alignment, rather than in the noisy labels.

Let

$$\widehat{Q} = Q_s + \Delta Q_n, \qquad J = J_s + J_n. \tag{82}$$

The derivatives of P with respect to the order parameters are

$$\frac{\partial P}{\partial J_s} = \frac{\partial P}{\partial J_n} = \frac{1}{\pi\sqrt{S\widehat{Q}}\sqrt{1-\frac{J^2}{S\widehat{Q}}}}, \tag{83}$$

$$\frac{\partial P}{\partial Q_s} = -\frac{J}{2\pi\sqrt{S}\widehat{Q}^{3/2}\sqrt{1-\frac{J^2}{S\widehat{Q}}}}, \tag{84}$$

$$\frac{\partial P}{\partial Q_n} = -\frac{\Delta J}{2\pi\sqrt{S}\widehat{Q}^{3/2}\sqrt{1-\frac{J^2}{S\widehat{Q}}}}. \tag{85}$$

For compactness, define

$$p_J = \frac{\partial P}{\partial J_s} = \frac{\partial P}{\partial J_n}, \qquad p_{Q_s} = \frac{\partial P}{\partial Q_s}, \qquad p_{Q_n} = \frac{\partial P}{\partial Q_n}. \tag{86}$$

For either SFT or RL, the order-parameter ODEs can be written in the form

$$\frac{dJ_i}{d\xi} = \eta A_{J_i}, \tag{87}$$

$$\frac{dQ_i}{d\xi} = \eta A_{Q_i} + \eta^2 B_{Q_i}, \qquad i \in \{s, n\}, \tag{88}$$

where A_{J_i}, A_{Q_i}, and B_{Q_i} are the drift and diffusion coefficients appropriate to the current finetuning rule. Explicitly, for SFT,

$$A_{J_i}^{\mathrm{SFT}} = -\frac{1}{2}\left(\langle\tilde{\nu}_i \mathrm{sgn}(\lambda)\rangle - (1-2\epsilon)\langle\tilde{\nu}_i \mathrm{sgn}(\nu)\rangle\right), \tag{89}$$

$$A_{Q_i}^{\mathrm{SFT}} = -\left(\langle\lambda_i \mathrm{sgn}(\lambda)\rangle - (1-2\epsilon)\langle\lambda_i \mathrm{sgn}(\nu)\rangle\right), \tag{90}$$

$$B_{Q_i}^{\mathrm{SFT}} = \frac{G_i}{T}(1-P_\epsilon), \tag{91}$$

whereas for RL,

$$A_{J_i}^{\mathrm{RL}} = \frac{1}{2}\left(\langle\tilde{\nu}_i \mathrm{sgn}(\lambda)\rangle + \langle\tilde{\nu}_i \mathrm{sgn}(\nu)\rangle\right) P^{T-1}, \tag{92}$$

$$A_{Q_i}^{\mathrm{RL}} = \left(\langle\lambda_i \mathrm{sgn}(\lambda)\rangle + \langle\lambda_i \mathrm{sgn}(\nu)\rangle\right) P^{T-1}, \tag{93}$$

$$B_{Q_i}^{\mathrm{RL}} = \frac{G_i}{T} P^T. \tag{94}$$

Here

$$G_s = T_{ss} + \Delta T_{sn}, \qquad G_n = T_{sn} + \Delta T_{nn}. \tag{95}$$

To choose the learning rate, we locally linearize p in the order parameters while retaining the quadratic dependence on η coming from the Q_i diffusion terms. For a small integration step $d\xi$,

$$\Delta P \approx d\xi\left[p_J(dJ_s + dJ_n) + p_{Q_s} dQ_s + p_{Q_n} dQ_n\right]. \tag{96}$$

Substituting the unconstrained increments gives

$$\Delta P(\eta) \approx d\xi\left[b\eta + a\eta^2\right], \tag{97}$$

where

$$b = p_J(A_{J_s} + A_{J_n}) + p_{Q_s} A_{Q_s} + p_{Q_n} A_{Q_n}, \tag{98}$$

$$a = p_{Q_s} B_{Q_s} + p_{Q_n} B_{Q_n}. \tag{99}$$

Thus the locally optimal learning rate is obtained by maximizing the quadratic

$$f(\eta) = b\eta + a\eta^2 \tag{100}$$

over the allowed interval $0 < \eta \leq \eta_{\max}$. In the implementation we take $\eta_{\max} = 1$, and use a small positive floor $\eta_{\min}$ to avoid exactly zero updates.

If $a < 0$ and $b > 0$, the quadratic is locally concave and initially increasing, so the unconstrained maximizer is

$$\eta_* = -\frac{b}{2a}. \tag{101}$$

If the quadratic is convex, or if the vertex is not a useful positive maximizer, the optimum over a bounded interval lies on the boundary. The implemented rule is

$$\eta_{\text{opt}} = \min\left\{\eta_{\max}, \begin{cases} -\dfrac{b}{2a}, & a < 0 \text{ and } b > 0, \\ \max\left\{-\dfrac{b}{a}, 0\right\} + \eta_{\min}, & a > 0 \text{ and } b < 0, \\ \eta_{\max}, & b > 0, \\ \eta_{\min}, & \text{otherwise.} \end{cases}\right. \tag{102}$$

This is the learning rate used for the next Euler step.

Spherical constraint. When the spherical constraint is enabled, the projected norm

$$Q_s + Q_n = R \tag{103}$$

is held fixed by removing the radial component of the update. Let the unconstrained Q-increment be

$$dQ_{\text{unc}} = dQ_s + dQ_n. \tag{104}$$

The projected increments are

$$dJ_i \leftarrow dJ_i - \frac{J_i}{2R} dQ_{\text{unc}}, \tag{105}$$

$$dQ_i \leftarrow dQ_i - \frac{Q_i}{R} dQ_{\text{unc}}, \qquad i \in \{s, n\}. \tag{106}$$

Since

$$dQ_{\text{unc}} = \eta(A_{Q_s} + A_{Q_n}) + \eta^2(B_{Q_s} + B_{Q_n}), \tag{107}$$

the spherical projection modifies both the linear and quadratic coefficients of the local objective.

Define

$$A_Q^{\text{tot}} = A_{Q_s} + A_{Q_n}, \qquad B_Q^{\text{tot}} = B_{Q_s} + B_{Q_n}. \tag{108}$$

The constrained linear coefficient is

$$\begin{aligned} b_{\text{con}} = {} & p_J \left[A_{J_s} - \frac{J_s}{2R} A_Q^{\text{tot}}\right] + p_J \left[A_{J_n} - \frac{J_n}{2R} A_Q^{\text{tot}}\right] \\ & + p_{Q_s} \left[A_{Q_s} - \frac{Q_s}{R} A_Q^{\text{tot}}\right] + p_{Q_n} \left[A_{Q_n} - \frac{Q_n}{R} A_Q^{\text{tot}}\right]. \end{aligned} \tag{109}$$

The constrained quadratic coefficient is

$$\begin{aligned} a_{\text{con}} = {} & -p_J \frac{J_s}{2R} B_Q^{\text{tot}} - p_J \frac{J_n}{2R} B_Q^{\text{tot}} \\ & + p_{Q_s} \left[B_{Q_s} - \frac{Q_s}{R} B_Q^{\text{tot}}\right] + p_{Q_n} \left[B_{Q_n} - \frac{Q_n}{R} B_Q^{\text{tot}}\right]. \end{aligned} \tag{110}$$

When the spherical constraint is used, the same quadratic maximization rule is applied with

$$b \leftarrow b_{\text{con}}, \qquad a \leftarrow a_{\text{con}}. \tag{111}$$

Otherwise, the unconstrained coefficients b and a are used.

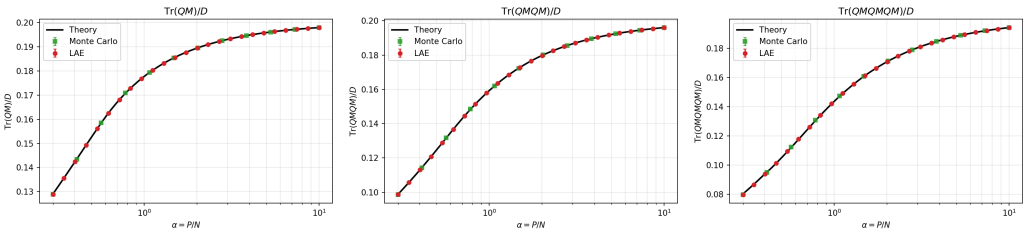


Figure 8: The expected normalised traces of powers of $\mathbf{PM}_s$, calculated by theory (solid), PCA Monte carlo simulation (orange) and trained LAE encoder projection for different batch sizes (blue). *Parameters*: $N = 2000, \quad (k = 0.2), \Delta = 0.1$

B Expected Trace Constants for the Random PCA Projector

The projected dynamics in Appendix A depend on the static trace constants

$$t, \qquad t_s, \qquad T_{ss}, \qquad \tau_{sss},$$

where

$$t = \frac{1}{D}\operatorname{Tr}(\mathbf{P}), \qquad t_s = \frac{1}{D}\operatorname{Tr}(\mathbf{PM}_s), \tag{112}$$

$$T_{ss} = \frac{1}{D}\operatorname{Tr}(\mathbf{PM}_s\mathbf{PM}_s), \qquad \tau_{sss} = \frac{1}{D}\operatorname{Tr}(\mathbf{PM}_s\mathbf{PM}_s\mathbf{PM}_s). \tag{113}$$

These quantities depend only on the empirical PCA projector $\mathbf{P}$ learned during pretraining. In the proportional high-dimensional limit they are self-averaging, so in the typical-case ODEs we replace them by their deterministic limits, or, equivalently, by their expectations over the pretraining dataset.

Let the ambient dimension be D. The signal and noise projectors satisfy

$$\operatorname{rank}(\mathbf{M}_s) = \rho D, \qquad \operatorname{rank}(\mathbf{M}_n) = (1-\rho)D, \qquad \mathbf{M}_s + \mathbf{M}_n = \mathbf{I}_D, \tag{114}$$

with $0 < \rho < 1$. The population covariance is

$$\Sigma = \mathbf{M}_s + \Delta\mathbf{M}_n, \qquad 0 < \Delta < 1. \tag{115}$$

The pretraining data matrix is

$$\mathbf{X} = \mathbf{Z}\Sigma^{1/2}, \qquad Z_{\mu i} \sim \mathcal{N}(0,1), \qquad \mu = 1, \ldots, N, \tag{116}$$

and the empirical covariance is

$$\widehat{\Sigma} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}. \tag{117}$$

The PCA projector $\mathbf{P}$ is the orthogonal projector onto the top ρD empirical eigenvectors of $\widehat{\Sigma}$. Hence

$$t = \frac{1}{D}\operatorname{Tr}(\mathbf{P}) = \rho \tag{118}$$

for every realization of the pretraining data. We work in the Marchenko–Pastur scaling

$$D, N \to \infty, \qquad \gamma = \frac{D}{N}, \text{ fixed.} \tag{119}$$

We assume that the selected top-ρD empirical eigenspace lies in the non-zero empirical spectrum. If $N < \rho D$, then the projector necessarily includes directions from the null space of $\widehat{\Sigma}$; in that case the trace constants acquire additional convention-dependent null-space contributions (we do not consider this regime).

B.1 Resolvent representation of the PCA projector

Define the resolvent as

$$\mathbf{R}_{\widehat{\Sigma}}(z) = (\widehat{\Sigma} - z\mathbf{I})^{-1}. \tag{120}$$

If

$$\widehat{\Sigma} = \sum_{a=1}^{D} \widehat{\lambda}_a \hat{\mathbf{u}}_a \hat{\mathbf{u}}_a^\top,$$

then we can write

$$\mathbf{R}_{\widehat{\Sigma}}(z) = \sum_{a=1}^{D} \frac{\hat{\mathbf{u}}_a \hat{\mathbf{u}}_a^\top}{\widehat{\lambda}_a - z}.$$

Let I denote the limiting empirical eigenvalue window selected by the PCA projector. Equivalently, I is the upper spectral set satisfying

$$\int_I \varrho(x)\, dx = \rho, \tag{121}$$

where ϱ is the limiting empirical spectral density. If the top spectral component is separated from the rest of the spectrum, then I is the support of that top component. If the spectrum is gapless, then I is the upper ρ-quantile of the limiting spectrum.

By Cauchy's functional calculus, For an eigenvalue $\widehat{\lambda}_a$ inside the PCA window I,

$$\oint_{\Gamma_a} \frac{dz}{\widehat{\lambda}_a - z} = -2\pi i,$$

where Γ_a is a positively oriented contour enclosing $\widehat{\lambda}_a$ and no other eigenvalues. Therefore, when the selected spectral set is separated by a macroscopic gap, this projector has the contour representation

$$-\frac{1}{2\pi i} \oint_{\Gamma_I} \mathbf{R}_{\widehat{\Sigma}}(z)\, dz = \sum_{\widehat{\lambda}_a \in I} \hat{\mathbf{u}}_a \hat{\mathbf{u}}_a^\top = \mathbf{P}, \tag{122}$$

where Γ_I is a positively oriented contour enclosing the spectral set I, and no other part of the spectrum. The minus sign appears because we use the convention $\mathbf{R}_{\widehat{\Sigma}}(z) = (\widehat{\Sigma} - z\mathbf{I})^{-1}$. This contour representation is useful because each trace constant can be written as a contour integral of a resolvent chain, as below:

$$t_s = -\frac{1}{2\pi i} \oint_{\Gamma_I} \frac{1}{D} \operatorname{Tr}\big(\mathbf{M}_s \mathbf{R}_{\widehat{\Sigma}}(z)\big)\, dz, \tag{123}$$

$$T_{ss} = \frac{1}{(2\pi i)^2} \oint_{\Gamma_I} \oint_{\Gamma_I} \frac{1}{D} \operatorname{Tr}\big(\mathbf{M}_s \mathbf{R}_{\widehat{\Sigma}}(z_1) \mathbf{M}_s \mathbf{R}_{\widehat{\Sigma}}(z_2)\big)\, dz_1 dz_2, \tag{124}$$

$$\tau_{sss} = -\frac{1}{(2\pi i)^3} \oint_{\Gamma_I} \oint_{\Gamma_I} \oint_{\Gamma_I} \frac{1}{D} \operatorname{Tr}\big(\mathbf{M}_s \mathbf{R}_{\widehat{\Sigma}}(z_1) \mathbf{M}_s \mathbf{R}_{\widehat{\Sigma}}(z_2) \mathbf{M}_s \mathbf{R}_{\widehat{\Sigma}}(z_3)\big)\, dz_1 dz_2 dz_3. \tag{125}$$

The contour integral is simply a way of replacing the discontinuous operation "keep only eigenvectors in the PCA window" by an analytic operation on the resolvent. Random matrix theory can compute deterministic limits of resolvent traces much more directly than deterministic limits of eigenvector projectors. Thus the problem reduces to evaluating deterministic equivalents for one-, two-, and three-resolvent traces with deterministic insertions $\mathbf{M}_s$. This is the standard route in random matrix theory for extracting eigenvector information from resolvents; see, for example, Silverstein and Choi [1995], Ledoit and Péché [2011]. N.B. **the contour formula is clean only in the separated case**, i.e. where the spectrum is gapped ($\gamma < 1$). If the selected window I lies inside a connected bulk component, there is no fixed macroscopic contour enclosing the selected eigenvalues and no other spectrum. At finite D, there is technically a tiny random gap between consecutive eigenvalues, but it vanishes as $D \to \infty$. In the gapless quantile-selection regime, the contour picture should be understood as the limiting boundary-value representation obtained by replacing the contour integral

by the spectral discontinuity of the resolvent ($\mathbf{R}_{\widehat{\Sigma}}(x+i0) - \mathbf{R}_{\widehat{\Sigma}}(x-i0)$) across the interval I, we detail this in the subsequent sections.

In the Marchenko–Pastur limit, the resolvent traces self-average, so the trace constants can be replaced by their deterministic limits. The purpose of the Matrix Dyson Equation (next section) is to compute these deterministic limits.

B.2 Matrix Dyson Equation

The random matrix

$$\widehat{\Sigma} = \frac{1}{N}\mathbf{X}^\top \mathbf{X} \tag{126}$$

depends quadratically on the Gaussian data matrix $\mathbf{X}$. Consequently, its resolvent

$$\mathbf{R}_{\widehat{\Sigma}}(z) = (\widehat{\Sigma} - z\mathbf{I})^{-1} \tag{127}$$

is a nonlinear function of the Gaussian randomness. The purpose of the Matrix Dyson Equation, or MDE, is to replace this random resolvent, in the proportional high-dimensional limit, by a deterministic matrix-valued function ($\mathbf{R}_{\widehat{\Sigma}}(z) \to \mathbf{\Pi}_{\widehat{\Sigma}}(z)$). (More general methods already exist for doing this Atanasov et al. [2024], but we require only a specific case of this for our simplified block structure). In this sense, the MDE plays the role of a law of large numbers for resolvents.

It is convenient not to apply the MDE directly to the quadratic matrix $\widehat{\Sigma}$, but instead to linearise it. Define

$$\mathcal{H} = \begin{pmatrix} 0 & \mathbf{X}^\top/\sqrt{N} \\ \mathbf{X}/\sqrt{N} & 0 \end{pmatrix} \in \mathbb{R}^{(D+N)\times(D+N)}. \tag{128}$$

The matrix $\mathcal{H}$ is linear in the Gaussian data. Moreover, $\mathcal{H}^2$ has upper-left block

$$\frac{1}{N}\mathbf{X}^\top \mathbf{X} = \widehat{\Sigma}. \tag{129}$$

Thus the nonzero eigenvalues of $\widehat{\Sigma}$ are the squares of the corresponding singular-value eigenvalues of the upper block of $\mathcal{H}$. This is the reason for introducing $\mathcal{H}$: it turns the quadratic covariance problem into a linear random-matrix problem.

Let

$$\mathbf{R}_{\mathcal{H}}(\zeta) = (\mathcal{H} - \zeta\mathbf{I})^{-1} \tag{130}$$

be the resolvent of the linearised matrix. We denote its deterministic equivalent by $\mathbf{\Pi}_{\mathcal{H}}(\zeta)$. Heuristically, this deterministic equivalent is obtained by averaging the resolvent identity and solving for the expected resolvent:

$$(\mathcal{H} - \zeta\mathbf{I})\mathbf{R}_{\mathcal{H}}(\zeta) = \mathbf{I} \tag{131}$$

$$\mathbb{E}[\mathcal{H}\mathbf{R}_{\mathcal{H}}(\zeta)] - \zeta\mathbb{E}[\mathbf{R}_{\mathcal{H}}(\zeta)] = \mathbf{I}. \tag{132}$$

If $\mathcal{H}$ were deterministic we could just solve for $\mathbf{\Pi}_{\mathcal{H}}(\zeta)$, but $\mathcal{H}$ is random and $\boldsymbol{R}_{\mathcal{H}}$ depends on $\mathcal{H}$. Hence, we are left with the problem of evaluating the correlation $\mathbb{E}[\mathcal{H}\mathbf{R}_{\mathcal{H}}]$.

B.3 Derivation of the Matrix Dyson Equation

Let

$$\mathbf{R}(\zeta) := \mathbf{R}_{\mathcal{H}}(\zeta) = \left(\mathcal{H} - \zeta\mathbf{I}_{D+N}\right)^{-1}, \tag{133}$$

$$\mathbf{m}(\zeta) := \mathbb{E}[\mathbf{R}(\zeta)]\,. \tag{134}$$

The deterministic equivalent $\mathbf{\Pi}(\zeta)$ is defined below as the solution of the corresponding Matrix Dyson Equation. In the proportional high-dimensional limit, one expects

$$\mathbf{m}(\zeta) - \mathbf{\Pi}(\zeta) \longrightarrow 0. \tag{135}$$

Define

$$A_{\mu i} := \frac{X_{\mu i}}{\sqrt{N}}, \tag{136}$$

$$\mathbf{B}_{\mu i} := \mathbf{E}_{i,D+\mu} + \mathbf{E}_{D+\mu,i}, \tag{137}$$

where $\mathbf{E}_{ab}$ denotes the matrix with a 1 in entry (a, b) and zeros elsewhere. Then

$$\mathcal{H} = \sum_{\mu=1}^{N} \sum_{i=1}^{D} A_{\mu i} \mathbf{B}_{\mu i}. \tag{138}$$

The random variables $A_{\mu i}$ are independent, centred Gaussian variables satisfying

$$\mathbb{E}\left[A_{\mu i}^2\right] = \frac{\sigma_i}{N}, \tag{139}$$

$$\sigma_i = \begin{cases} 1, & i \in \mathcal{I}_{\text{signal}}, \\ \Delta, & i \in \mathcal{I}_{\text{noise}}. \end{cases} \tag{140}$$

We begin by expanding

$$\begin{aligned} \mathbb{E}[\mathcal{H}\mathbf{R}(\zeta)] &= \sum_{\mu=1}^{N} \sum_{i=1}^{D} \mathbb{E}[A_{\mu i} \mathbf{B}_{\mu i} \mathbf{R}(\zeta)] \\ &= \sum_{\mu=1}^{N} \sum_{i=1}^{D} \mathbf{B}_{\mu i} \mathbb{E}[A_{\mu i} \mathbf{R}(\zeta)], \end{aligned} \tag{141}$$

where we used the fact that $\mathbf{B}_{\mu i}$ is deterministic.

Gaussian integration by parts gives

$$\mathbb{E}[A_{\mu i} \mathbf{R}(\zeta)] = \frac{\sigma_i}{N} \mathbb{E}\left[\frac{\partial \mathbf{R}(\zeta)}{\partial A_{\mu i}}\right]. \tag{142}$$

Since

$$\mathbf{R}(\zeta) = (\mathcal{H} - \zeta \mathbf{I}_{D+N})^{-1}, \tag{143}$$

the derivative of the inverse matrix gives

$$\begin{aligned} \frac{\partial \mathbf{R}(\zeta)}{\partial A_{\mu i}} &= -\mathbf{R}(\zeta) \frac{\partial \mathcal{H}}{\partial A_{\mu i}} \mathbf{R}(\zeta) \\ &= -\mathbf{R}(\zeta) \mathbf{B}_{\mu i} \mathbf{R}(\zeta). \end{aligned} \tag{144}$$

Therefore,

$$\mathbb{E}[A_{\mu i} \mathbf{R}(\zeta)] = -\frac{\sigma_i}{N} \mathbb{E}[\mathbf{R}(\zeta) \mathbf{B}_{\mu i} \mathbf{R}(\zeta)]. \tag{145}$$

Substituting 145 into 141, we obtain

$$\mathbb{E}[\mathcal{H}\mathbf{R}(\zeta)] = -\sum_{\mu=1}^{N} \sum_{i=1}^{D} \frac{\sigma_i}{N} \mathbf{B}_{\mu i} \mathbb{E}[\mathbf{R}(\zeta) \mathbf{B}_{\mu i} \mathbf{R}(\zeta)]. \tag{146}$$

Define the linear self-energy operator by

$$\mathcal{S}[\mathbf{T}] := \sum_{\mu=1}^{N} \sum_{i=1}^{D} \frac{\sigma_i}{N} \mathbf{B}_{\mu i} \mathbf{T} \mathbf{B}_{\mu i}, \tag{147}$$

for every $\mathbf{T} \in \mathbb{C}^{(D+N)\times(D+N)}$.

When $\mathbf{T}$ is deterministic, this can equivalently be written as

$$\mathcal{S}[\mathbf{T}] = \mathbb{E}[\mathcal{H}\mathbf{T}\mathcal{H}]\,. \tag{148}$$

Using the definition of $\mathcal{S}$, equation 146 becomes

$$\boxed{\mathbb{E}[\mathcal{H}\mathbf{R}(\zeta)] = -\mathbb{E}[\mathcal{S}[\mathbf{R}(\zeta)]\mathbf{R}(\zeta)]} \tag{149}$$

This is an exact identity for any centred Gaussian model. No deterministic-equivalent approximation has yet been made.

The resolvent identity

$$(\mathcal{H} - \zeta\mathbf{I}_{D+N})\,\mathbf{R}(\zeta) = \mathbf{I}_{D+N} \tag{150}$$

implies

$$\mathbb{E}[\mathcal{H}\mathbf{R}(\zeta)] - \zeta\mathbf{m}(\zeta) = \mathbf{I}_{D+N}. \tag{151}$$

Combining 149 and 151, we find

$$-\mathbb{E}[\mathcal{S}[\mathbf{R}(\zeta)]\mathbf{R}(\zeta)] - \zeta\mathbf{m}(\zeta) = \mathbf{I}_{D+N}. \tag{152}$$

To isolate the fluctuation term, write

$$\mathbf{R}(\zeta) = \mathbf{m}(\zeta) + \mathbf{Q}(\zeta), \tag{153}$$

$$\mathbb{E}[\mathbf{Q}(\zeta)] = \mathbf{0}. \tag{154}$$

Since $\mathcal{S}$ is deterministic and linear,

$$\mathbb{E}[\mathcal{S}[\mathbf{R}]\mathbf{R}] = \mathcal{S}[\mathbf{m}]\mathbf{m} + \mathbb{E}[\mathcal{S}[\mathbf{Q}]\mathbf{Q}]\,. \tag{155}$$

Consequently,

$$-\left(\zeta\mathbf{I}_{D+N} + \mathcal{S}[\mathbf{m}(\zeta)]\right)\mathbf{m}(\zeta) = \mathbf{I}_{D+N} + \mathbb{E}[\mathcal{S}[\mathbf{Q}(\zeta)]\mathbf{Q}(\zeta)]\,. \tag{156}$$

In the proportional high-dimensional limit, concentration and fluctuation averaging are used to show that the final term in 156 is negligible in an appropriate sense. This motivates defining $\mathbf{\Pi}(\zeta)$ as the solution of

$$-\left(\zeta\mathbf{I}_{D+N} + \mathcal{S}[\mathbf{\Pi}(\zeta)]\right)\mathbf{\Pi}(\zeta) = \mathbf{I}_{D+N}. \tag{157}$$

Multiplying 157 from the right by $\mathbf{\Pi}(\zeta)^{-1}$ gives

$$\boxed{-\mathbf{\Pi}(\zeta)^{-1} = \zeta\mathbf{I}_{D+N} + \mathcal{S}[\mathbf{\Pi}(\zeta)]} \tag{158}$$

This is the Matrix Dyson Equation (MDE), and we have just shown that its solution is the deterministic equivalent of the resolvent of the linearised covariance matrix $\mathcal{H}$.

B.3.1 Solving the Matrix Dyson Equation to find the deterministic Resolvent (In our block symmetric case)

We now compute the action of the self-energy operator $\mathcal{S}$ on block-constant diagonal matrices of the form

$$\mathrm{diag}\big(x_s\mathbf{I}_{\rho D}, x_n\mathbf{I}_{(1-\rho)D}, x_0\mathbf{I}_N\big), \tag{159}$$

the operator $\mathcal{S}$ acts as

$$\mathcal{S} : (x_s, x_n, x_0) \mapsto \big(x_0,\, \Delta x_0,\, \gamma\big[\rho x_s + (1-\rho)\Delta x_n\big]\big)\,. \tag{160}$$

This can be shown from the definition of the self-energy operator and the covariance structure of the problem. A signal coordinate has variance $1/N$ against each sample index, a noise coordinate

has variance Δ/N, and the sample block receives the average contribution from all signal and noise coordinates. Since $D/N \to \gamma$, the sample-block term is

$$\gamma\big[\rho x_s + (1-\rho)\Delta x_n\big]. \tag{161}$$

Because the model is invariant under rotations inside the signal subspace, inside the noise subspace, and inside the N-dimensional sample space, the deterministic resolvent cannot distinguish individual directions within each of these three blocks. Hence the MDE solution has the form

$$\mathbf{\Pi}_{\mathcal{H}}(\zeta) = \operatorname{diag}\left(a_s(\zeta)\mathbf{I}_{\rho D}, a_n(\zeta)\mathbf{I}_{(1-\rho)D}, b(\zeta)\mathbf{I}_N\right). \tag{162}$$

Substituting this ansatz and using the action of the self-energy operator 160 into the MDE 158 gives the three scalar self-consistency equations

$$-\frac{1}{a_s} = \zeta + b, \tag{163}$$

$$-\frac{1}{a_n} = \zeta + \Delta b, \tag{164}$$

$$-\frac{1}{b} = \zeta + \gamma\big[\rho a_s + (1-\rho)\Delta a_n\big]. \tag{165}$$

These equations determine $\mathbf{\Pi}_{\mathcal{H}}(\zeta)$, the deterministic resolvent of the linearised matrix $\mathcal{H}$. But all of this was in order to find the deterministic resolvent of the empirical covariance matrix $\widehat{\Sigma}$. So we now translate this back to the covariance resolvent

$$\mathbf{R}_{\widehat{\Sigma}}(\zeta) = (\widehat{\Sigma} - \zeta^2\mathbf{I})^{-1}. \tag{166}$$

The link between this quantity and $\mathbf{\Pi}(\zeta)$ is the upper-left block of the linearised resolvent. Applying the Schur complement to $\mathcal{H} - \zeta\mathbf{I}$ gives

$$\big[(\mathcal{H} - \zeta\mathbf{I})^{-1}\big]_{\text{upper-left}} = \zeta\left(\widehat{\Sigma} - \zeta^2\mathbf{I}\right)^{-1}. \tag{167}$$

Thus, with $z = \zeta^2$,

$$\big[(\mathcal{H} - \zeta\mathbf{I})^{-1}\big]_{\text{upper-left}} = \zeta\mathbf{R}_{\widehat{\Sigma}}(z). \tag{168}$$

The variables a_s and a_n are therefore not themselves the block values of the covariance resolvent. They are the signal and noise block values of the upper-left block of the linearised resolvent, which is ζ times the covariance resolvent.

Since the upper-left block of $\mathbf{\Pi}_{\mathcal{H}}(\zeta)$ is

$$a_s(\zeta)\mathbf{M}_s + a_n(\zeta)\mathbf{M}_n, \tag{169}$$

the deterministic equivalent of the covariance resolvent has the block form

$$\mathbf{\Pi}_{\widehat{\Sigma}}(z) = g_s(z)\mathbf{M}_s + g_n(z)\mathbf{M}_n, \tag{170}$$

where

$$g_s(z) = \frac{a_s(\zeta)}{\zeta}, \tag{171}$$

$$g_n(z) = \frac{a_n(\zeta)}{\zeta}, \qquad z = \zeta^2. \tag{172}$$

Finally, introduce the scalar variable

$$\xi(z) = -\zeta b(\zeta). \tag{173}$$

Using the first two scalar MDE equations, one obtains

$$a_s(\zeta) = \frac{\zeta}{\xi(z) - z}, \tag{174}$$

$$a_n(\zeta) = \frac{\zeta}{\Delta\xi(z) - z}, \tag{175}$$

$$b(\zeta) = -\frac{\xi(z)}{\zeta}. \tag{176}$$

Therefore

$$g_s(z) = \frac{1}{\xi(z) - z}, \tag{177}$$

$$g_n(z) = \frac{1}{\Delta\xi(z) - z}. \tag{178}$$

These are the signal and noise block values of the deterministic equivalent of the covariance resolvent. They are the one-resolvent quantities used in the subsequent sections to compute the trace constants

$$t_s, \qquad T_{ss}, \qquad \tau_{sss}. \tag{179}$$

B.4 One-resolvent quantities and the local signal overlap

We use the previously derived deterministic equivalent of the resolvent to compute the limiting empirical spectral density and the expected signal overlap of empirical eigenvectors. The local signal overlap is the squared projection of an empirical eigenvector onto the true signal subspace, as a function of the corresponding empirical eigenvalue. This is the continuum version of the Ledoit–Péché eigenvector-overlap law [Ledoit and Péché, 2011].

Using Equation (165) 172 173, we can write $\xi(z)$ defined as the solution of

$$\frac{1}{\xi(z)} = 1 + \gamma \left(\frac{\rho}{\xi(z) - z} + \frac{(1-\rho)\Delta}{\Delta\xi(z) - z} \right). \tag{180}$$

The physical branch is the solution satisfying $\xi(z) \to 1$ as $|z| \to \infty$, and such that the corresponding Stieltjes transform $m(z)$ (defined in Equation (182)) obeys $\operatorname{Im} m(z) > 0$ for $z \in \mathbb{C}^+$.

Equivalently, ξ is the appropriate root of the cubic

$$\begin{aligned} &\Delta\xi^3 - \left[z(1+\Delta) + \Delta(1-\gamma)\right]\xi^2 \\ &\quad + z\left[z + 1 + \Delta - \gamma\big(\rho + (1-\rho)\Delta\big)\right]\xi - z^2 = 0. \end{aligned} \tag{181}$$

Then, using Equation (170), the deterministic equivalent of the normalized resolvent trace (the Stieltjes transform) is

$$m(z) = \frac{1}{D}\operatorname{Tr}\mathbf{\Pi}_{\widehat{\Sigma}}(z) = g_s(z)\frac{\operatorname{Tr}\mathbf{M}_s}{D}\mathbf{M}_s + g_n(z)\frac{\operatorname{Tr}\mathbf{M}_n}{D} = \rho g_s(z) + (1-\rho)g_n(z). \tag{182}$$

The limiting empirical spectral density is recovered from the boundary value using the Stieltjes-Perron inversion formula

$$\varrho(x) = \frac{1}{\pi}\operatorname{Im} m(x + i0).$$

Similarly, we can compute the limiting signal-resolved trace constant from the deterministic equivalent of the resolvent:

$$\frac{1}{D}\operatorname{Tr}\left(\mathbf{M}_s\mathbf{R}_{\widehat{\Sigma}}(z)\right) \longrightarrow \frac{1}{D}\operatorname{Tr}\left(\mathbf{M}_s\mathbf{\Pi}_{\widehat{\Sigma}}(z)\right) = \rho g_s(z), \tag{183}$$

so we define the signal-resolved density

$$\varrho_s(x) = \frac{1}{\pi}\operatorname{Im} g_s(x + i0). \tag{184}$$

The density $\rho\varrho_s(x)$ is the empirical spectral density weighted by the squared projection of the corresponding empirical eigenvector onto the true signal subspace. Therefore the local signal overlap of an empirical eigenvector with eigenvalue x is

$$w_s(x) = \frac{\rho\varrho_s(x)}{\varrho(x)}. \tag{185}$$

This is the continuum version of the Ledoit–Péché eigenvector-overlap law: for an empirical eigenvector $\hat{\mathbf{u}}_a$ with eigenvalue $\widehat{\lambda}_a \approx x$,

$$\hat{\mathbf{u}}_a^\top \mathbf{M}_s \hat{\mathbf{u}}_a \approx w_s(x).$$

The first-order trace constant is therefore

$$\begin{aligned} t_s &= \frac{1}{D}\mathbb{E}\,\mathrm{Tr}(\mathbf{P}\mathbf{M}_s) \\ &= \int_I \varrho(x) w_s(x)\,dx \\ &= \rho \int_I \varrho_s(x)\,dx. \end{aligned} \tag{186}$$

Notice that t_s/ρ is the average fraction of the selected PCA subspace lying in the true signal subspace, whereas t_s is the normalization used in the ODE closure.

B.5 Two-body stability operator and the deterministic equivalent of a two-resolvent chain

The one-resolvent Matrix Dyson Equation determines the deterministic matrix $\mathbf{\Pi}(\zeta)$ that approximates a single linearised resolvent

$$\mathbf{R}_a := \mathbf{R}_{\mathcal{H}}(\zeta_a) = (\mathcal{H} - \zeta_a \mathbf{I})^{-1}, \qquad \mathbf{\Pi}_a := \mathbf{\Pi}(\zeta_a). \tag{187}$$

For a deterministic matrix $\mathbf{K}$, however, the deterministic equivalent of the chain

$$\mathbf{R}_1 \mathbf{K} \mathbf{R}_2 \tag{188}$$

is not, in general, obtained by replacing the two resolvents separately:

$$\mathbf{R}_1 \mathbf{K} \mathbf{R}_2 \not\sim \mathbf{\Pi}_1 \mathbf{K} \mathbf{\Pi}_2. \tag{189}$$

The reason is that $\mathbf{R}_1$ and $\mathbf{R}_2$ are functions of the same random matrix $\mathcal{H}$. Their fluctuations are therefore correlated. The deterministic approximation to 188 must retain the repeated contractions generated by this shared randomness. The operator that resums these contractions is the two-body stability operator.

We derive this operator directly from the MDE by introducing two replicas of the linearised problem and coupling them by a deterministic source. This gives, in one calculation, both the definition of the two-body stability operator and the formula for the deterministic equivalent of a two-resolvent chain.

B.5.1 Replicated resolvent with an off-diagonal source

Let

$$\mathbb{H} := \begin{pmatrix} \mathcal{H} & \mathbf{0} \\ \mathbf{0} & \mathcal{H} \end{pmatrix}, \tag{190}$$

and, for a deterministic matrix $\mathbf{K} \in \mathbb{C}^{(D+N)\times(D+N)}$, define the upper-triangular spectral-source matrix

$$\mathbb{Z}_\varepsilon = \begin{pmatrix} \zeta_1 \mathbf{I}_{D+N} & \varepsilon \mathbf{K} \\ \mathbf{0} & \zeta_2 \mathbf{I}_{D+N} \end{pmatrix}. \tag{191}$$

The corresponding replicated resolvent is

$$\mathbb{R}_\varepsilon = (\mathbb{H} - \mathbb{Z}_\varepsilon)^{-1}. \tag{192}$$

Since

$$\mathbb{H} - \mathbb{Z}_\varepsilon = \begin{pmatrix} \mathcal{H} - \zeta_1 \mathbf{I} & -\varepsilon \mathbf{K} \\ \mathbf{0} & \mathcal{H} - \zeta_2 \mathbf{I} \end{pmatrix}, \tag{193}$$

the inverse of this block-upper-triangular matrix is

$$\mathbb{R}_\varepsilon = \begin{pmatrix} \mathbf{R}_1 & \varepsilon \mathbf{R}_1 \mathbf{K} \mathbf{R}_2 \\ \mathbf{0} & \mathbf{R}_2 \end{pmatrix}. \tag{194}$$

In particular, the two-resolvent chain is exactly the linear response of the upper-right block with respect to the source:

$$\frac{\partial}{\partial \varepsilon}\mathbb{R}_{\varepsilon,12}\Big|_{\varepsilon=0} = \mathbf{R}_1\mathbf{K}\mathbf{R}_2. \tag{195}$$

This source construction is useful because the same derivative can be taken on the deterministic MDE solution.

The two diagonal blocks of $\mathbb{H}$ contain the same random matrix, not two independent copies. To make the resulting covariance structure explicit, recall that

$$\mathcal{H} = \sum_{\mu=1}^{N}\sum_{i=1}^{D} A_{\mu i}\mathbf{B}_{\mu i}. \tag{196}$$

Then

$$\mathbb{H} = \sum_{\mu=1}^{N}\sum_{i=1}^{D} A_{\mu i}\begin{pmatrix}\mathbf{B}_{\mu i} & \mathbf{0}\\ \mathbf{0} & \mathbf{B}_{\mu i}\end{pmatrix}. \tag{197}$$

The same Gaussian coefficient $A_{\mu i}$ appears in both replicas. Hence the self-energy operator of the replicated problem acts on a 2×2 block matrix $\mathbb{T} = (\mathbf{T}_{ab})_{a,b=1}^{2}$ as

$$\mathbb{S}[\mathbb{T}] = \begin{pmatrix}\mathcal{S}[\mathbf{T}_{11}] & \mathcal{S}[\mathbf{T}_{12}]\\ \mathcal{S}[\mathbf{T}_{21}] & \mathcal{S}[\mathbf{T}_{22}]\end{pmatrix}. \tag{198}$$

The off-diagonal action $\mathbb{S}[\mathbb{T}]_{12} = \mathcal{S}[\mathbf{T}_{12}]$ is precisely where the correlation between the two resolvents enters. Had the two diagonal blocks of $\mathbb{H}$ contained independent random matrices, this off-diagonal covariance term would vanish, and the leading deterministic approximation would reduce to the naive product $\mathbf{\Pi}_1\mathbf{K}\mathbf{\Pi}_2$.

B.5.2 Linearising the replicated Matrix Dyson Equation

Let $\mathbb{\Pi}_\varepsilon$ denote the deterministic MDE solution for the replicated resolvent $\mathbb{R}_\varepsilon$. With the same sign convention as in 158, it satisfies

$$-\mathbb{\Pi}_\varepsilon^{-1} = \mathbb{Z}_\varepsilon + \mathbb{S}[\mathbb{\Pi}_\varepsilon]. \tag{199}$$

At $\varepsilon = 0$, the two replicas decouple at the deterministic level and

$$\mathbb{\Pi}_0 = \begin{pmatrix}\mathbf{\Pi}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{\Pi}_2\end{pmatrix}. \tag{200}$$

We therefore expand the upper-right block to first order as

$$\mathbb{\Pi}_\varepsilon = \begin{pmatrix}\mathbf{\Pi}_1 & \varepsilon\mathbf{M}_{12}[\mathbf{K}]\\ \mathbf{0} & \mathbf{\Pi}_2\end{pmatrix} + O(\varepsilon^2). \tag{201}$$

The matrix $\mathbf{M}_{12}[\mathbf{K}]$ will be the deterministic equivalent of $\mathbf{R}_1\mathbf{K}\mathbf{R}_2$.

For an invertible matrix-valued function $\mathbb{Q}_\varepsilon$,

$$\frac{\partial}{\partial\varepsilon}\mathbb{Q}_\varepsilon^{-1} = -\mathbb{Q}_\varepsilon^{-1}\left(\frac{\partial\mathbb{Q}_\varepsilon}{\partial\varepsilon}\right)\mathbb{Q}_\varepsilon^{-1}. \tag{202}$$

Consequently, the upper-right block of $-\mathbb{\Pi}_\varepsilon^{-1}$, to first order in ε, is

$$\left[-\mathbb{\Pi}_\varepsilon^{-1}\right]_{12} = \varepsilon\mathbf{\Pi}_1^{-1}\mathbf{M}_{12}[\mathbf{K}]\mathbf{\Pi}_2^{-1} + O(\varepsilon^2). \tag{203}$$

On the right-hand side of 199, the upper-right block is

$$\left[\mathbb{Z}_\varepsilon + \mathbb{S}[\mathbb{\Pi}_\varepsilon]\right]_{12} = \varepsilon\left(\mathbf{K} + \mathcal{S}[\mathbf{M}_{12}[\mathbf{K}]]\right) + O(\varepsilon^2). \tag{204}$$

Equating 203 and 204 gives

$$\mathbf{\Pi}_1^{-1}\mathbf{M}_{12}[\mathbf{K}]\mathbf{\Pi}_2^{-1} = \mathbf{K} + \mathcal{S}[\mathbf{M}_{12}[\mathbf{K}]]. \tag{205}$$

Multiplying from the left by $\mathbf{\Pi}_1$ and from the right by $\mathbf{\Pi}_2$, we obtain the recursive Dyson equation

$$\mathbf{M}_{12}[\mathbf{K}] = \mathbf{\Pi}_1\mathbf{K}\mathbf{\Pi}_2 + \mathbf{\Pi}_1\mathcal{S}[\mathbf{M}_{12}[\mathbf{K}]]\mathbf{\Pi}_2. \tag{206}$$

This equation is linear in the unknown two-resolvent deterministic equivalent. It is therefore natural to define the two-body stability operator

$$\boxed{\mathcal{B}_{12}[\mathbf{T}] := \mathbf{T} - \mathbf{\Pi}_1\mathcal{S}[\mathbf{T}]\mathbf{\Pi}_2} \tag{207}$$

for every $\mathbf{T} \in \mathbb{C}^{(D+N)\times(D+N)}$. In this notation, 206 becomes

$$\mathcal{B}_{12}[\mathbf{M}_{12}[\mathbf{K}]] = \mathbf{\Pi}_1\mathbf{K}\mathbf{\Pi}_2. \tag{208}$$

Whenever $\mathcal{B}_{12}$ is invertible on the relevant matrix subspace,

$$\boxed{\mathbf{M}_{12}[\mathbf{K}] = \mathcal{B}_{12}^{-1}\left[\mathbf{\Pi}_1\mathbf{K}\mathbf{\Pi}_2\right]} \tag{209}$$

is the deterministic equivalent of the chain $\mathbf{R}_1\mathbf{K}\mathbf{R}_2$, in the same weak, anisotropic, or normalised-trace sense in which the corresponding multi-resolvent local law is used. For the global contour calculations considered here, the spectral parameters remain away from the real spectrum, and the inverse is bounded. If one subsequently takes boundary values close to the spectrum, the size of $\mathcal{B}_{12}^{-1}$ must be controlled as part of the local-law analysis.

The word "stability" has the same meaning as for the one-body MDE. The operator $\mathcal{B}_{12}$ is the linear operator that must be inverted to turn a perturbation of the two-replica MDE into the corresponding perturbation of its solution. When $\mathcal{B}_{12}^{-1}$ is large, a small source can produce a large two-resolvent response.

B.5.3 Solving to find the two-body deterministic equivalent

We now take the deterministic insertion to be the signal projector lifted to the linearised space,

$$\widehat{\mathbf{M}}_s = \operatorname{diag}\left(\mathbf{I}_{\rho D}, \mathbf{0}_{(1-\rho)D}, \mathbf{0}_N\right). \tag{210}$$

For $\zeta_a^2 = z_a$, write

$$\mathbf{\Pi}_a = \operatorname{diag}\left(a_{s,a}\mathbf{I}_{\rho D}, a_{n,a}\mathbf{I}_{(1-\rho)D}, b_a\mathbf{I}_N\right), \qquad a = 1, 2. \tag{211}$$

The deterministic equivalent of $\mathbf{R}_1\widehat{\mathbf{M}}_s\mathbf{R}_2$ is therefore

$$\boxed{\mathbf{A}_{12}^{(s)} := \mathcal{B}_{12}^{-1}\left[\mathbf{\Pi}_1\widehat{\mathbf{M}}_s\mathbf{\Pi}_2\right]} \tag{212}$$

with

$$\mathbf{\Pi}_1\widehat{\mathbf{M}}_s\mathbf{\Pi}_2 = \operatorname{diag}\left(a_{s,1}a_{s,2}\mathbf{I}_{\rho D}, \mathbf{0}_{(1-\rho)D}, \mathbf{0}_N\right). \tag{213}$$

Because both the MDE solution and the input are invariant under rotations within the signal, noise, and sample blocks, the solution has the form

$$\mathbf{A}_{12}^{(s)} = \operatorname{diag}\left(x_s\mathbf{I}_{\rho D}, x_n\mathbf{I}_{(1-\rho)D}, x_0\mathbf{I}_N\right). \tag{214}$$

Introduce the abbreviations

$$\alpha_s := a_{s,1}a_{s,2}, \qquad \alpha_n := a_{n,1}a_{n,2}, \qquad \beta := b_1b_2. \tag{215}$$

Using the block action of the self-energy operator from 160,

$$\mathcal{S} : (x_s, x_n, x_0) \mapsto \left(x_0, \Delta x_0, \gamma\left[\rho x_s + (1-\rho)\Delta x_n\right]\right), \tag{216}$$

we find

$$\mathbf{\Pi}_1\mathcal{S}[\mathbf{A}_{12}^{(s)}]\mathbf{\Pi}_2 = \operatorname{diag}\left(\alpha_s x_0\mathbf{I}_{\rho D}, \Delta\alpha_n x_0\mathbf{I}_{(1-\rho)D}, \beta\gamma\left[\rho x_s + (1-\rho)\Delta x_n\right]\mathbf{I}_N\right). \tag{217}$$

The stability equation

$$\mathcal{B}_{12}[\mathbf{A}_{12}^{(s)}] = \mathbf{\Pi}_1\widehat{\mathbf{M}}_s\mathbf{\Pi}_2 \tag{218}$$

therefore reduces to the three-dimensional linear system

$$x_s - \alpha_s x_0 = \alpha_s, \tag{219}$$
$$x_n - \Delta\alpha_n x_0 = 0, \tag{220}$$
$$x_0 - \beta\gamma\left[\rho x_s + (1-\rho)\Delta x_n\right] = 0. \tag{221}$$

Equivalently,

$$\begin{pmatrix} 1 & 0 & -\alpha_s \\ 0 & 1 & -\Delta\alpha_n \\ -\gamma\beta\rho & -\gamma\beta(1-\rho)\Delta & 1 \end{pmatrix} \begin{pmatrix} x_s \\ x_n \\ x_0 \end{pmatrix} = \begin{pmatrix} \alpha_s \\ 0 \\ 0 \end{pmatrix}. \tag{222}$$

The determinant controlling the inverse on this invariant subspace is

$$\mathfrak{d}_{12} := 1 - \gamma\beta\left[\rho\alpha_s + (1-\rho)\Delta^2\alpha_n\right]. \tag{223}$$

Solving 221 gives

$$x_0 = \frac{\gamma\beta\rho\alpha_s}{\mathfrak{d}_{12}}, \tag{224}$$
$$x_n = \frac{\gamma\beta\rho\Delta\alpha_s\alpha_n}{\mathfrak{d}_{12}}, \tag{225}$$
$$x_s = \alpha_s\left(1 + \frac{\gamma\beta\rho\alpha_s}{\mathfrak{d}_{12}}\right) \tag{226}$$
$$= \alpha_s\frac{1-\gamma\beta(1-\rho)\Delta^2\alpha_n}{\mathfrak{d}_{12}}. \tag{227}$$

B.5.4 Returning to the covariance resolvent

Let

$$\mathbf{G}_a := \mathbf{R}_{\widehat{\Sigma}}(z_a) = (\widehat{\Sigma} - z_a\mathbf{I})^{-1}, \qquad z_a = \zeta_a^2. \tag{228}$$

From 168, the upper-left block of the linearised resolvent is

$$[\mathbf{R}_a]_{DD} = \zeta_a\mathbf{G}_a. \tag{229}$$

Since $\widehat{\mathbf{M}}_s$ is supported only in the upper-left block,

$$\operatorname{Tr}\left(\widehat{\mathbf{M}}_s\mathbf{R}_1\widehat{\mathbf{M}}_s\mathbf{R}_2\right) = \zeta_1\zeta_2\operatorname{Tr}\left(\mathbf{M}_s\mathbf{G}_1\mathbf{M}_s\mathbf{G}_2\right). \tag{230}$$

Therefore the deterministic equivalent of the normalised covariance-resolvent chain is

$$\frac{1}{D}\operatorname{Tr}\left(\mathbf{M}_s\mathbf{G}_1\mathbf{M}_s\mathbf{G}_2\right) \sim \frac{1}{D\zeta_1\zeta_2}\operatorname{Tr}\left(\widehat{\mathbf{M}}_s\mathbf{A}_{12}^{(s)}\right) \tag{231}$$
$$= \frac{\rho x_s}{\zeta_1\zeta_2}. \tag{232}$$

Using

$$a_{s,a} = \zeta_a g_s(z_a), \qquad a_{n,a} = \zeta_a g_n(z_a), \qquad b_a = -\frac{\xi(z_a)}{\zeta_a}, \tag{233}$$

we have

$$\beta\alpha_s = \xi(z_1)\xi(z_2)g_s(z_1)g_s(z_2), \tag{234}$$
$$\beta\alpha_n = \xi(z_1)\xi(z_2)g_n(z_1)g_n(z_2). \tag{235}$$

It is convenient to define

$$\kappa(z_1,z_2) := \gamma\xi(z_1)\xi(z_2)\left[\rho g_s(z_1)g_s(z_2) + (1-\rho)\Delta^2 g_n(z_1)g_n(z_2)\right]. \tag{236}$$

Then 232 becomes

$$\boxed{\frac{1}{D}\operatorname{Tr}\left(\mathbf{M}_s\mathbf{R}_{\widehat{\Sigma}}(z_1)\mathbf{M}_s\mathbf{R}_{\widehat{\Sigma}}(z_2)\right) \sim \rho g_s(z_1)g_s(z_2)\frac{1-\gamma(1-\rho)\Delta^2\xi(z_1)\xi(z_2)g_n(z_1)g_n(z_2)}{1-\kappa(z_1,z_2)}} \tag{237}$$

B.5.5 Two-resolvent quantity T_{ss}

The previous calculated quantity is the two-body deterministic equivalent required in the double-contour representation of T_{ss}. In particular,

$$T_{ss} = \frac{1}{(2\pi i)^2} \oint_{\Gamma_I} \oint_{\Gamma_I} \frac{1}{D} \operatorname{Tr}\big(\mathbf{M}_s \mathbf{R}_{\widehat{\Sigma}}(z_1) \mathbf{M}_s \mathbf{R}_{\widehat{\Sigma}}(z_2)\big)\, dz_1 dz_2 \tag{238}$$

$$T_{ss} = \frac{1}{(2\pi i)^2} \oint_{\Gamma_I} \oint_{\Gamma_I} \frac{\rho x_s(z_1, z_2)}{\zeta_1 \zeta_2}\, dz_1\, dz_2, \tag{239}$$

where $x_s(z_1, z_2)$ is given by 227, and the square roots $\zeta_a^2 = z_a$ are chosen consistently along the contours. In a gapless regime the same formula is interpreted through the limiting upper and lower boundary values of the resolvents, as in the one-resolvent discussion. We detail this now.

The second-order trace constant is

$$T_{ss} = \frac{1}{D} \mathbb{E} \operatorname{Tr}(\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s)\,. \tag{240}$$

Expanding in empirical eigenvectors, with

$$m_{ab} = \hat{\mathbf{u}}_a^\top \mathbf{M}_s \hat{\mathbf{u}}_b,$$

gives

$$T_{ss} = \frac{1}{D} \sum_{a,b:\, \widehat{\lambda}_a, \widehat{\lambda}_b \in I} m_{ab}^2. \tag{241}$$

The diagonal terms ($a = b$) are controlled by the local overlap w_s, but the off-diagonal terms ($a \neq b$) measure correlations between distinct empirical eigenvectors through the fixed signal projector $\mathbf{M}_s$. Even if each m_{ab} is small, there are $\mathcal{O}(D^2)$ such terms. After normalisation by $1/D$, their total contribution remains $\mathcal{O}(1)$. Therefore we must keep track of the off-diagonal terms, and we require the deterministic equivalent of the two-resolvent chain:

$$\mathcal{G}_{ss}^{(D)}(z_1, z_2) = \frac{1}{D} \operatorname{Tr}(\mathbf{M}_s \mathbf{R}(z_1) \mathbf{M}_s \mathbf{R}(z_2))\,, \qquad \mathbf{R}(z) = (\widehat{\Sigma} - z\mathbf{I})^{-1}, \tag{242}$$

its spectral expansion is

$$\mathcal{G}_{ss}^{(D)}(z_1, z_2) = \frac{1}{D} \sum_{a,b=1}^{D} \frac{m_{ab}^2}{(\widehat{\lambda}_a - z_1)(\widehat{\lambda}_b - z_2),} \tag{243}$$

so the double discontinuity of $\mathcal{G}_{ss}^{(D)}$ extracts the limiting measure of m_{ab}^2 over pairs of eigenvalues. Its deterministic equivalent, $\mathcal{G}_{ss}(z_1, z_2) := \lim_{D\to\infty} \mathcal{G}_{ss}^{(D)}(z_1, z_2)$, is obtained from the *two-body stability operator* of the linearized MDE, detailed in the next section.

In the gapless bulk-selection regime the mixed boundary value has a simple diagonal singularity because, for x in the bulk,

$$\kappa_{+-}(x, x) = 1. \tag{244}$$

Indeed, subtracting the scalar equation for $\xi_+(x)$ from its conjugate gives

$$\frac{1}{\xi_+(x)\xi_-(x)} = \gamma \left[\rho |g_{s,+}(x)|^2 + (1-\rho)\Delta^2 |g_{n,+}(x)|^2\right], \tag{245}$$

which is precisely the identity $\kappa_{+-}(x, x) = 1$. Thus the singularity is a first-order pole of the mixed boundary value through the stability denominator, rather than a second-order pole in $x - y$.

To recover T_{ss}, take the double spectral discontinuity of $\mathcal{G}_{ss}$. For $x, y \in I$, define

$$\mathcal{G}_{ss}^{\sigma\tau}(x, y) = \mathcal{G}_{ss}(x + i\sigma 0, y + i\tau 0), \qquad \sigma, \tau \in \{+, -\}. \tag{246}$$

The two-point overlap measure is

$$K_2(x,y) = -\frac{1}{4\pi^2}\left(\mathcal{G}_{ss}^{++} - \mathcal{G}_{ss}^{+-} - \mathcal{G}_{ss}^{-+} + \mathcal{G}_{ss}^{--}\right)(x,y). \tag{247}$$

Equivalently,

$$K_2(x,y) = \lim_{D\to\infty} \frac{1}{D}\sum_{a,b} m_{ab}^2 \delta(x-\widehat{\lambda}_a)\delta(y-\widehat{\lambda}_b).$$

It decomposes as

$$K_2(x,y) = d_s(x)\delta(x-y) + k_{ss}^{\mathrm{reg}}(x,y), \tag{248}$$

with diagonal weight

$$d_s(x) = \varrho(x) w_s(x)^2 = \rho^2 \frac{\varrho_s(x)^2}{\varrho(x)}. \tag{249}$$

Because the mixed-boundary residue is purely imaginary, no real principal-value term remains in the overlap measure. For $x \neq y$, the smooth off-diagonal kernel is

$$k_{ss}^{\mathrm{reg}}(x,y) = -\frac{1}{2\pi^2}\left[\mathrm{Re}\,\mathcal{G}_{ss}^{++}(x,y) - \mathrm{Re}\,\mathcal{G}_{ss}^{+-}(x,y)\right]. \tag{250}$$

The diagonal $x = y$ of k_{ss}^{reg} is removable; numerically one should treat the analytic diagonal term 249 separately and then take the continuous limit of 250.

Combining the diagonal and off-diagonal contributions gives

$$\begin{aligned} T_{ss} &= \int_I d_s(x)\,dx + \iint_{I\times I} k_{ss}^{\mathrm{reg}}(x,y)\,dx\,dy \\ &= \rho^2 \int_I \frac{\varrho_s(x)^2}{\varrho(x)}\,dx + \iint_{I\times I} k_{ss}^{\mathrm{reg}}(x,y)\,dx\,dy. \end{aligned} \tag{251}$$

This formula uses the exact two-body stability denominator and separates the diagonal singularity analytically. It avoids any guessed pole expansion in $x - y$.

B.6 Three-resolvent quantity τ_{sss}

The third-order trace constant is

$$\tau_{sss} = \frac{1}{D}\mathbb{E}\,\mathrm{Tr}(\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s)\,. \tag{252}$$

Expanding again in empirical eigenvectors gives

$$\tau_{sss} = \frac{1}{D}\sum_{a,b,c:\,\widehat{\lambda}_a,\widehat{\lambda}_b,\widehat{\lambda}_c\in I} m_{ab}m_{bc}m_{ca}. \tag{253}$$

This expression separates into three types of contributions:

$$\begin{aligned} \tau_{sss} = &\underbrace{\frac{1}{D}\sum_{a\in I} m_{aa}^3}_{\text{all indices equal}} \\ &+ \underbrace{\frac{3}{D}\sum_{\substack{a,b\in I\\ a\neq b}} m_{aa}m_{ab}^2}_{\text{one pair of indices equal}} \\ &+ \underbrace{\frac{1}{D}\sum_{\substack{a,b,c\in I\\ a,b,c\ \text{distinct}}} m_{ab}m_{bc}m_{ca}}_{\text{all indices distinct}}. \end{aligned} \tag{254}$$

The factor of 3 in the second term accounts for the three possible coincident pairs $(a = b)$, $(b = c)$, and $(c = a)$.

The fully diagonal term is determined by the one-point overlap density:

$$\frac{1}{D}\sum_{a\in I} m_{aa}^3 \longrightarrow \int_I \varrho(x) w_s(x)^3\, dx = \rho^3 \int_I \frac{\varrho_s(x)^3}{\varrho(x)^2}\, dx. \tag{255}$$

The one-pair-equal contribution is determined by the diagonal overlap w_s and the off-diagonal part of the two-point kernel:

$$\frac{3}{D}\sum_{\substack{a,b\in I\\ a\neq b}} m_{aa} m_{ab}^2 \longrightarrow 3\iint_{I\times I} w_s(x) k_{ss}^{\mathrm{reg}}(x,y)\, dx\, dy. \tag{256}$$

The remaining all-distinct term is a new three-eigenvector contribution. It is not determined by the scalar one- and two-point trace constants alone. At the deterministic-equivalent level, however, it is obtained from the length-three multi-resolvent chain generated by the same pairwise stability operators used above.

The required covariance-side chain is

$$\mathcal{G}_{sss}^{(D)}(z_1,z_2,z_3) = \frac{1}{D}\operatorname{Tr}(\mathbf{M}_s\mathbf{R}(z_1)\mathbf{M}_s\mathbf{R}(z_2)\mathbf{M}_s\mathbf{R}(z_3))\,. \tag{257}$$

We use the linearized notation from the previous subsection. For $a, b \in \{1, 2, 3\}$, define

$$\mathcal{B}_{ab}[\mathbf{X}] = \mathbf{X} - \mathbf{\Pi}_a\mathcal{S}[\mathbf{X}]\mathbf{\Pi}_b,$$

and define the pair-dressed signal insertion

$$\mathbf{V}_{ab} = \mathcal{B}_{ab}^{-1}\left[\widehat{\mathbf{M}}_s\right]. \tag{258}$$

By block symmetry,

$$\mathbf{V}_{ab} = \operatorname{diag}\left(v_s^{ab}\mathbf{I}_{\rho D}, v_n^{ab}\mathbf{I}_{(1-\rho)D}, v_0^{ab}\mathbf{I}_N\right),$$

where

$$v_s^{ab} - a_{s,a}a_{s,b}v_0^{ab} = 1, \tag{259}$$

$$v_n^{ab} - \Delta a_{n,a}a_{n,b}v_0^{ab} = 0, \tag{260}$$

$$v_0^{ab} - \gamma b_a b_b\left[\rho v_s^{ab} + (1-\rho)\Delta v_n^{ab}\right] = 0. \tag{261}$$

Equivalently, with

$$\kappa_{ab} := \kappa(z_a, z_b),$$

one has

$$v_0^{ab} = \frac{\gamma\rho b_a b_b}{1-\kappa_{ab}}, \tag{262}$$

$$v_s^{ab} = 1 + a_{s,a}a_{s,b}v_0^{ab}, \tag{263}$$

$$v_n^{ab} = \Delta a_{n,a}a_{n,b}v_0^{ab}. \tag{264}$$

The length-three deterministic chain is obtained by dressing the product

$$\mathbf{\Pi}_1\mathbf{V}_{12}\mathbf{\Pi}_2\mathbf{V}_{23}\mathbf{\Pi}_3$$

with the outer stability operator $\mathcal{B}_{13}$. Define

$$\mathbf{Y}_{123} = \mathcal{B}_{13}^{-1}\left[\mathbf{\Pi}_1\mathbf{V}_{12}\mathbf{\Pi}_2\mathbf{V}_{23}\mathbf{\Pi}_3\right]. \tag{265}$$

Again,

$$\mathbf{Y}_{123} = \operatorname{diag}\left(y_s\mathbf{I}_{\rho D}, y_n\mathbf{I}_{(1-\rho)D}, y_0\mathbf{I}_N\right).$$

Introduce the source terms

$$s_s = a_{s,1}v_s^{12}a_{s,2}v_s^{23}a_{s,3}, \tag{266}$$

$$s_n = a_{n,1}v_n^{12}a_{n,2}v_n^{23}a_{n,3}, \tag{267}$$

$$s_0 = b_1 v_0^{12} b_2 v_0^{23} b_3. \tag{268}$$

Then the block equations for $\mathbf{Y}_{123}$ are

$$y_s - a_{s,1}a_{s,3}y_0 = s_s, \tag{269}$$
$$y_n - \Delta a_{n,1}a_{n,3}y_0 = s_n, \tag{270}$$
$$y_0 - \gamma b_1 b_3 \left[\rho y_s + (1-\rho)\Delta y_n\right] = s_0. \tag{271}$$

Solving gives

$$y_0 = \frac{s_0 + \gamma b_1 b_3 \left[\rho s_s + (1-\rho)\Delta s_n\right]}{1-\kappa_{13}}, \qquad y_s = s_s + a_{s,1}a_{s,3}y_0. \tag{272}$$

The deterministic equivalent of the three-resolvent covariance-side chain is therefore

$$\begin{aligned}\mathcal{G}_{sss}(z_1,z_2,z_3) &:= \lim_{D\to\infty} \mathcal{G}_{sss}^{(D)}(z_1,z_2,z_3) \\ &= \rho \frac{y_s}{\zeta_1\zeta_2\zeta_3}.\end{aligned} \tag{273}$$

The final expression is independent of the choice of square-root branches $\zeta_a^2 = z_a$. All denominators appearing in 273 are pairwise stability denominators:

$$1-\kappa_{12}, \qquad 1-\kappa_{23}, \qquad 1-\kappa_{13}.$$

Thus there is no additional cubic or triple stability pole to isolate.

The compact contour representation is

$$\tau_{sss} = -\frac{1}{(2\pi i)^3} \oint_{\Gamma_I}\oint_{\Gamma_I}\oint_{\Gamma_I} \mathcal{G}_{sss}(z_1,z_2,z_3)\, dz_1 dz_2 dz_3. \tag{274}$$

This formula automatically includes the diagonal, pair-coincident, and all-distinct contributions.

For completeness, we also give the equivalent real-axis decomposition. Away from the pair diagonals $x=y$, $y=z$, and $z=x$, define the all-distinct three-point kernel by

$$k_{sss}^{\mathrm{reg}}(x,y,z) = \frac{1}{(2\pi i)^3} \sum_{\sigma_1,\sigma_2,\sigma_3\in\{+,-\}} \sigma_1\sigma_2\sigma_3\, \mathcal{G}_{sss}^{\sigma_1\sigma_2\sigma_3}(x,y,z), \tag{275}$$

where

$$\mathcal{G}_{sss}^{\sigma_1\sigma_2\sigma_3}(x,y,z) = \mathcal{G}_{sss}(x+i\sigma_1 0, y+i\sigma_2 0, z+i\sigma_3 0).$$

The full triple spectral measure has the distributional decomposition

$$\begin{aligned}K_3(x,y,z) = {} & \varrho(x) w_s(x)^3 \delta(x-y)\delta(y-z) \\ & + w_s(x) k_{ss}^{\mathrm{reg}}(x,z)\delta(x-y) + w_s(y) k_{ss}^{\mathrm{reg}}(y,x)\delta(y-z) \\ & + w_s(z) k_{ss}^{\mathrm{reg}}(z,y)\delta(z-x) + k_{sss}^{\mathrm{reg}}(x,y,z).\end{aligned} \tag{276}$$

The only singular surfaces are therefore the pair diagonals, together with their triple intersection. These singular pieces are precisely the diagonal and one-pair-equal contributions already displayed in 255 and 256.

Integrating 276 over I^3 gives

$$\begin{aligned}\tau_{sss} = {} & \rho^3 \int_I \frac{\varrho_s(x)^3}{\varrho(x)^2}\, dx \\ & + 3 \iint_{I\times I} w_s(x) k_{ss}^{\mathrm{reg}}(x,y)\, dx\, dy \\ & + \iiint_{I^3} k_{sss}^{\mathrm{reg}}(x,y,z)\, dx\, dy\, dz.\end{aligned} \tag{277}$$

The first line is the contribution of a single empirical eigenvector appearing three times, the second line is the contribution of two distinct empirical eigenvectors, and the final line is the genuine all-distinct three-eigenvector contribution computed from the length-three deterministic equivalent 273.

B.7 Trace constants used in the ODE closure

Once

$$t, \qquad t_s, \qquad T_{ss}, \qquad \tau_{sss}$$

have been computed, all remaining trace constants needed in Appendix A follow from

$$\mathbf{M}_s + \mathbf{M}_n = \mathbf{I}, \qquad \mathbf{P}^2 = \mathbf{P}.$$

For the second-order traces,

$$T_{sn} = \frac{1}{D}\operatorname{Tr}(\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_n) = t_s - T_{ss}, \tag{278}$$

$$T_{nn} = \frac{1}{D}\operatorname{Tr}(\mathbf{P}\mathbf{M}_n\mathbf{P}\mathbf{M}_n) = t - 2t_s + T_{ss}. \tag{279}$$

For the third-order traces appearing in the Hilbert–Schmidt closure,

$$V_{ss} = \frac{1}{D}\operatorname{Tr}(\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s) = \tau_{sss}, \tag{280}$$

$$V_{sn} = \frac{1}{D}\operatorname{Tr}(\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_n) = T_{ss} - \tau_{sss}, \tag{281}$$

$$V_{nn} = \frac{1}{D}\operatorname{Tr}(\mathbf{P}\mathbf{M}_s\mathbf{P}\mathbf{M}_n\mathbf{P}\mathbf{M}_n) = t_s - 2T_{ss} + \tau_{sss}. \tag{282}$$

The stochastic variance terms in the Q_i dynamics are therefore

$$\begin{aligned} G_s &= T_{ss} + \Delta T_{sn} \\ &= T_{ss} + \Delta(t_s - T_{ss}), \end{aligned} \tag{283}$$

$$\begin{aligned} G_n &= T_{sn} + \Delta T_{nn} \\ &= (t_s - T_{ss}) + \Delta(t - 2t_s + T_{ss}). \end{aligned} \tag{284}$$

Finally, the Hilbert–Schmidt projection coefficients are obtained by defining

$$\Delta_G = T_{ss}t - t_s^2 \tag{285}$$

and substituting the deterministic trace limits into

$$\Omega_{ij}^{(s)} = \frac{(t - t_s)V_{ij} - (t_s - T_{ss})T_{ij}}{\Delta_G}, \tag{286}$$

$$\Omega_{ij}^{(n)} = \frac{-t_s V_{ij} + T_{ss}T_{ij}}{\Delta_G}. \tag{287}$$

These coefficients are fixed during fine-tuning. They encode the typical geometry of the random PCA projector relative to the true signal/noise decomposition learned during pretraining.

C Large-alpha scaling of the representation-limited error

We now derive the large-pretraining-sample scaling of the best possible generalisation error. As before, let

$$\alpha = \frac{N}{D}, \qquad \widehat{\mathbf{\Sigma}} = \mathbf{\Sigma} + \mathbf{E}, \qquad \mathbf{\Sigma} = \mathbf{M}_s + \Delta\mathbf{M}_n,$$

where $\mathbf{M}_s$ and $\mathbf{M}_n$ project onto subspaces of dimensions ρD and $(1-\rho)D$, respectively, and $\mathbf{E}$ is random matrix (capturing finite sample effects). The population signal and noise eigenvalues are therefore 1 and Δ, with spectral gap

$$g = 1 - \Delta.$$

We assume throughout that $0 < \Delta < 1$ is fixed as $\alpha \to \infty$, so that the population eigengap remains nonzero. The empirical PCA projector $\mathbf{P}$ is the rank-ρD spectral projector of $\widehat{\mathbf{\Sigma}}$ associated with its upper spectral component, as defined in the main theory. In coordinates adapted to the signal/noise decomposition, write

$$\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma} + \mathbf{E} \tag{288}$$

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} \mathbf{I}_{\rho D} & 0 \\ 0 & \Delta\mathbf{I}_{(1-\rho)D} \end{pmatrix} + \begin{pmatrix} \mathbf{E}_{ss} & \mathbf{E}_{sn} \\ \mathbf{E}_{ns} & \mathbf{E}_{nn} \end{pmatrix}. \tag{289}$$

Trivially, $\widehat{\mathbf{\Sigma}} \to \mathbf{\Sigma}$ as $\alpha \to \infty$. For the Gaussian pretraining model,

$$\mathbf{X} = \begin{pmatrix} \mathbf{Z}_s & \sqrt{\Delta}\,\mathbf{Z}_n \end{pmatrix},$$

where $\mathbf{Z}_s \in \mathbb{R}^{N \times \rho D}$ and $\mathbf{Z}_n \in \mathbb{R}^{N \times (1-\rho) D}$ have independent standard Gaussian entries. Therefore,

$$\mathbf{E}_{ns} = \mathbf{M}_n(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{M}_s = \frac{\sqrt{\Delta}}{N}\mathbf{Z}_n^\top \mathbf{Z}_s. \tag{290}$$

In the subsequent sections we will find the perturbative correction to the PCA projector $\mathbf{P}$ in terms of the off-diagonal block $\mathbf{E}_{ns}$. We will then relate it to the principal angles between the empirical PCA subspace and the true signal subspace, and finally compute the recoverable teacher variance in terms of these angles. This allows characterization of the representation-limited error in the large-α limit, in terms of α.

C.1 Perturbation of the PC projector

The first-order change in a spectral projector is found from standard Rayleigh-Schrödinger perturbation theory [Schrödinger, 1926]. Perturbation theory tells us how the eigenvectors of $\widehat{\mathbf{\Sigma}}$ change as a function of the perturbation $\mathbf{E}$. Since the projection $\mathbf{P}$ onto the top-ρD subspace is formed from summing the outer products of the top-ρD eigenvectors, we can use perturbation theory to compute the change in the projector $\mathbf{P}$ as a function of $\mathbf{E}$. This is done by expanding the projector in powers of the perturbation $\mathbf{E}$. The first-order term gives

$$\mathbf{P} = \mathbf{M}_s + \frac{\mathbf{M}_n \mathbf{E} \mathbf{M}_s + \mathbf{M}_s \mathbf{E} \mathbf{M}_n}{1 - \Delta} + O\left(\frac{\|\mathbf{E}\|_{\text{op}}^2}{(1-\Delta)^2}\right),$$

where $\|\cdot\|_{\text{op}}$ denotes the operator norm. In particular, the block that rotates the empirical signal subspace into the population noise subspace is

$$\mathbf{M}_n \mathbf{P} \mathbf{M}_s = \frac{\mathbf{E}_{ns}}{1 - \Delta} + O\left(\frac{\|\mathbf{E}\|_{\text{op}}^2}{(1-\Delta)^2}\right). \tag{291}$$

From Corollary 2 in Koltchinskii and Lounici [2017], $\|\mathbf{E}\|_{\text{op}}^2 = O_{\mathbb{P}}(\alpha^{-1})$. From the Gaussianity of 290, we have $\|\mathbf{E}_{ns}\|_F = O_{\mathbb{P}}(\sqrt{D\alpha^{-1}})$. Then applying Cauchy-Schwarz to 291, we have

$$\frac{1}{D}\|\mathbf{M}_n \mathbf{P} \mathbf{M}_s\|_F^2 = \frac{\|\mathbf{E}_{ns}\|_F^2}{D(1-\Delta)^2} + O_{\mathbb{P}}(\alpha^{-3/2}). \tag{292}$$

We will use this result in the next section.

C.2 Scaling of the First order trace overlap, t_s

Let $\theta_1, \ldots, \theta_{\rho D}$ denote the principal angles between the empirical PCA subspace and the true signal subspace. Then

$$t_s = \frac{1}{D}\operatorname{Tr}(\mathbf{P}\mathbf{M}_s) = \frac{1}{D}\operatorname{Tr}(\mathbf{M}_s \mathbf{P} \mathbf{M}_s) = \frac{1}{D}\sum_{a=1}^{\rho D} \cos^2\theta_a,$$

This can be seen by rewriting $\mathbf{P}$ and $\mathbf{M}_s$ in their respective eigenbases.

and hence

$$\rho - t_s = \frac{1}{D}\sum_{a=1}^{\rho D} \sin^2\theta_a. \tag{293}$$

By writing $\mathbf{A} = \mathbf{M}_n \mathbf{P} \mathbf{M}_s$, we can see that $\mathbf{A}^\top \mathbf{A} = \mathbf{M}_s \mathbf{P} \mathbf{M}_s - (\mathbf{M}_s \mathbf{P} \mathbf{M}_s)^2$. Hence $\mathbf{A}$ has singular values $\cos^2\theta_a - \cos^4\theta_a = \sin^2\theta_a \cos^2\theta_a$. Consequently,

$$\|\mathbf{M}_n \mathbf{P} \mathbf{M}_s\|_F^2 = \operatorname{Tr}\big(\mathbf{A}^\top \mathbf{A}\big) = \sum_{a=1}^{\rho D} \sin^2\theta_a \cos^2\theta_a = \sum_{a=1}^{\rho D} (\sin^2\theta_a - \sin^4\theta_a). \tag{294}$$

Using the Davis-Kahan theorem [Davis and Kahan, 1970], which bounds the principal angles in terms of the operator norm of the perturbation (for $\Delta \neq 1$), we have $\max_{1\leq a\leq \rho D} \sin\theta_a = O\left(\frac{\|\mathbf{E}\|_{\text{op}}}{1-\Delta}\right) = O_{\mathbb{P}}(\alpha^{-1/2})$. Allowing us to write

$$\frac{1}{D}\|\mathbf{M}_n\mathbf{P}\mathbf{M}_s\|_F^2 = \frac{1}{D}\sum_{a=1}^{\rho D}\sin^2\theta_a + O_{\mathbb{P}}(\alpha^{-2}) \tag{295}$$

Combining 292, 293 and 295 we get

$$\rho - t_s = \frac{1}{D(1-\Delta)^2}\|\mathbf{E}_{ns}\|_F^2 + O_{\mathbb{P}}(\alpha^{-3/2}). \tag{296}$$

Since the $O_{\mathbb{P}}(\alpha^{-3/2})$ term is uniform in the proportional $D, N \to \infty$ limit, and t_s and $\|\mathbf{E}_{ns}\|_F^2$ are self-averaging for fixed α we can easily take the deterministic limit of 296. First we now evaluate the Gaussian average. For each signal index a and noise index b,

$$(\mathbf{E}_{ns})_{ba} = \frac{\sqrt{\Delta}}{N}\sum_{\mu=1}^{N}(\mathbf{Z}_n)_{\mu b}(\mathbf{Z}_s)_{\mu a}.$$

This variable has zero mean and variance

$$\mathbb{E}(\mathbf{E}_{ns})_{ba}^2 = \frac{\Delta}{N}.$$

There are

$$\rho(1-\rho)D^2$$

entries in E_{ns}, so

$$\mathbb{E}\|\mathbf{E}_{ns}\|_F^2 = \rho(1-\rho)D^2\frac{\Delta}{N}.$$

Using $N = \alpha D$, this becomes

$$\mathbb{E}\|\mathbf{E}_{ns}\|_F^2 = \rho(1-\rho)D\frac{\Delta}{\alpha}.$$

Therefore,

$$\rho - t_s = \frac{\rho(1-\rho)\Delta}{(1-\Delta)^2}\frac{1}{\alpha} + O(\alpha^{-3/2}),$$

or equivalently,

$$\boxed{1 - \frac{t_s}{\rho} = \frac{(1-\rho)\Delta}{(1-\Delta)^2}\frac{1}{\alpha} + O(\alpha^{-3/2})}. \tag{297}$$

C.3 Bayes Optimal Classifier

Let $\mathbf{x}' = \mathbf{P}\mathbf{x}$ where $\mathbf{x}$ denote the projected observation seen by the student. Recall our definition for the teacher aligning field, $\nu = \frac{\mathbf{x}^\top\mathbf{w}^*}{\sqrt{D}}$, whose sign is the teacher label. We can characterise the Bayes optimal classifier by considering the conditional expectation of ν given the projected observation. $(\mathbf{x}', \nu)$ is jointly Gaussian, hence

$$\mathbb{E}[\nu|\mathbf{x}'] = \text{Cov}(\nu, \mathbf{x}')\text{Cov}(\mathbf{x}')^+\mathbf{x}' = \frac{1}{\sqrt{D}}\mathbf{w}^{*\top}\mathbf{P}(\mathbf{P}\boldsymbol{\Sigma}\mathbf{P})^+\mathbf{x}', \tag{298}$$

where $+$ denotes the Moore-Penrose pseudoinverse. Also, because the conditional variance of $\nu \mid z$ is independent of z,

$$\Pr(\nu > 0 \mid \mathbf{x}') > \frac{1}{2} \iff \mathbb{E}[\nu \mid \mathbf{x}'] > 0. \tag{299}$$

Hence the Bayes optimal classifier is given by

$$y_B(\mathbf{x}') = \operatorname{sgn}(\mathbf{w}_B \mathbf{x}') = \operatorname{sgn}({\mathbf{w}^*}^\top \mathbf{P}(\mathbf{P}\boldsymbol{\Sigma}\mathbf{P})^+ \mathbf{x}'), \tag{300}$$

where $\mathbf{w}_B = (\mathbf{P}\boldsymbol{\Sigma}\mathbf{P})^+ \mathbf{P}\mathbf{w}^*$. Let

$$q_a = \sin^2\theta_a.$$

The fraction of teacher-field variance recoverable from the projected observation is

$$\chi = \frac{(\mathbf{w}_B^\top \boldsymbol{\Sigma}\mathbf{w}^*)^2}{(\mathbf{w}_B^\top \boldsymbol{\Sigma}\mathbf{w}_B)({\mathbf{w}^*}^\top \boldsymbol{\Sigma}\mathbf{w}^*)}, \tag{301}$$

Defining $\lambda_B = \frac{\mathbf{w}_B^\top \mathbf{P}\mathbf{x}}{\sqrt{D}} = \frac{\mathbf{w}_B^\top \mathbf{x}}{\sqrt{D}}$ and noting that $\boldsymbol{\Sigma}\mathbf{w}^* = \mathbf{w}^*$ and $\mathbf{w}_B^\top \boldsymbol{\Sigma}\mathbf{w}_B = \mathbf{w}_B^\top \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}\mathbf{w}_B = \mathbf{w}_B^\top \mathbf{w}^*$ (since $\mathbf{w}_B$ lies in the range of $\mathbf{P}$), we show that $\operatorname{Cov}(\nu, \lambda_B) = \operatorname{Var}(\lambda_B)$. And 301 becomes

$$\chi = \frac{(\mathbf{w}_B^\top \mathbf{w}^*)^2}{(\mathbf{w}_B^\top \mathbf{w}^*)({\mathbf{w}^*}^\top \mathbf{w}^*)} = \frac{\mathbf{w}_B^\top \mathbf{w}^*}{{\mathbf{w}^*}^\top \mathbf{w}^*} = \frac{\operatorname{Var}(\lambda_B)}{\operatorname{Var}(\nu)}, \tag{302}$$

which is precisely the fraction of teacher-field variance recoverable from the projected observation. We can use 301 and the noted relationships above to compute χ in terms of the principal angles.

$$\chi = \frac{{\mathbf{w_B}}^\top \boldsymbol{\Sigma}\mathbf{w}_B}{{\mathbf{w}^*}^\top \mathbf{w}^*} = \frac{1}{SD}{\mathbf{w}^*}^\top (\mathbf{P}\boldsymbol{\Sigma}\mathbf{P})^+ \mathbf{w}^* = \frac{1}{SD}{\mathbf{w}^*}^\top (\Delta\mathbf{P} + (1-\Delta)\mathbf{P}\mathbf{M}_s\mathbf{P})^+ \mathbf{w}^* \tag{303}$$

$$= \frac{1}{\rho D}\sum_{a=1}^{\rho D} \frac{(\mathbf{p}_a^\top \mathbf{w}^*)^2}{\Delta + (1-\Delta)\cos^2\theta_a}, \tag{304}$$

where $\mathbf{p}_a$ is the set of orthonormal eigenvectors of $\mathbf{P}\mathbf{M_s}\mathbf{P}$. For an isotropically distributed teacher in signal space, this becomes

$$\chi = \frac{1}{\rho D}\sum_{a=1}^{\rho D} \frac{\cos^2\theta_a}{\Delta + (1-\Delta)\cos^2\theta_a}. \tag{305}$$

rewriting each term in the summation as

$$\frac{\cos^2\theta_a}{\Delta + (1-\Delta)\cos^2\theta_a} = 1 - \Delta\sin^2\theta_a + O(\sin^4\theta_a).$$

we can recall the use of the Davis-Kahan [Davis and Kahan, 1970] theorem to bound the principal angles in terms of the operator norm of the perturbation, which gives us $\sin^2\theta_a = O_{\mathbb{P}}(\alpha^{-1})$. Therefore, we can write

$$1 - \chi = \frac{\Delta}{\rho D}\sum_{a=1}^{\rho D} \sin^2\theta_a + O(\alpha^{-2}).$$

Using 293, we obtain

$$1 - \chi = \Delta\left(1 - \frac{t_s}{\rho}\right) + O(\alpha^{-2}).$$

Then using 297, we obtain

$$\boxed{1 - \chi = \frac{(1-\rho)\Delta^2}{(1-\Delta)^2}\frac{1}{\alpha} + O(\alpha^{-3/2})}. \tag{306}$$

C.4 Bayes optimal classification error

The Bayes optimal classifier based only on the projected observation has error

$$e_B = \frac{1}{\pi} \arccos \sqrt{\chi}.$$

Equivalently,

$$e_B = \frac{1}{\pi} \arcsin \sqrt{1-\chi}.$$

Substituting

$$\arcsin x = x + O(x^3) \qquad \text{as } x \to 0,$$

we obtain

$$e_B = \frac{1}{\pi}\sqrt{1-\chi} + O((1-\chi)^{3/2}) = \frac{1}{\pi}\sqrt{1-\chi} + O(\alpha^{-3/2}).$$

Substituting 306 yields

$$\boxed{e_B = \frac{\Delta\sqrt{1-\rho}}{\pi(1-\Delta)}\alpha^{-1/2} + O(\alpha^{-1})}.$$

Thus the recoverable-variance deficit scales as

$$1 - \chi \sim \alpha^{-1},$$

whereas the classification error scales as

$$e_B \sim \alpha^{-1/2}.$$

This result holds in the separated, perturbative regime with fixed $0 < \Delta < 1$. It is not uniform as $\Delta \to 1$, because the eigengap $1 - \Delta$ closes and the Kato expansion ceases to be controlled. A sufficient qualitative condition for the expansion is

$$\frac{\sqrt{\Delta/\alpha}}{1-\Delta} \ll 1, \qquad \text{or equivalently} \qquad \alpha\frac{(1-\Delta)^2}{\Delta} \gg 1.$$

D Large-alpha scaling of the Fixed points of the Dynamics

In this section we derive the asymptotic SFT generalization error directly from the projected stochastic dynamics. The central observation is that, although the microscopic student weight does not in general converge to a deterministic vector, one can identify a linear observable whose stationary drift determines the student–teacher field correlation exactly.

Recalling the SFT update 1

$$\mathbf{w}^{(\mu+1)} = \mathbf{w}^{(\mu)} - \frac{\eta}{T\sqrt{D}} \sum_{t=1}^{T} y_t^{(\mu)} \mathbf{P}\mathbf{x}_t^{(\mu)} \theta\left(-y_t^{*(\mu)} y_t^{(\mu)}\right), \tag{307}$$

The student observes only the projected input Px, where P is the empirical PCA projector learned during pretraining. Recall

$$S = \frac{\|\mathbf{w}^*\|^2}{D}, \qquad \widehat{Q} = \frac{1}{D}\mathbf{w}^\top \mathbf{P}\mathbf{\Sigma}\mathbf{P}\mathbf{w}, \qquad J = \frac{1}{D}\mathbf{w}^\top \mathbf{P}\mathbf{w}^*. \tag{308}$$

The student–teacher field correlation is

$$c = \frac{J}{\sqrt{S\widehat{Q}}}, \tag{309}$$

and the classification error is

$$E \equiv 1 - P_{\text{gen}} = \frac{1}{\pi} \arccos c. \tag{310}$$

Let

$$\beta = 1 - 2\epsilon, \qquad k = \sqrt{\frac{2}{\pi}}, \tag{311}$$

where ϵ is the probability of flipping the teacher label. For $y = \mathrm{sgn}(\lambda)$ and noisy teacher label y^*, the elementary identity

$$y\,\mathbf{1}_{\{y \neq y^*\}} = \frac{1}{2}(y - y^*) \tag{312}$$

together with

$$\mathbb{E}[y^*|\nu] = (1 - 2\epsilon)\mathrm{sgn}(\nu) = \beta\,\mathrm{sgn}(\nu) \tag{313}$$

implies that the conditional mean SFT update is proportional to

$$\mathbb{E}[\Delta\mathbf{w}\,|\,\mathbf{w}] \propto -\frac{\eta}{2}\left[\mathbb{E}(\mathbf{P}\mathbf{x}\,\mathrm{sgn}\lambda) - \beta\,\mathbb{E}(\mathbf{P}\mathbf{x}\,\mathrm{sgn}\nu)\right]. \tag{314}$$

For centred jointly Gaussian variables one has

$$\mathbb{E}[X\mathrm{sgn}(Y)] = \sqrt{\frac{2}{\pi}}\,\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(Y)}}. \tag{315}$$

Applying 315 to the Gaussian input x gives

$$\mathbb{E}[\mathbf{P}\mathbf{x}\,\mathrm{sgn}\lambda] = k\,\frac{\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}\mathbf{w}}{\sqrt{D\widehat{Q}}}, \tag{316}$$

$$\mathbb{E}[\mathbf{P}\mathbf{x}\,\mathrm{sgn}\nu] = k\,\frac{\mathbf{P}\boldsymbol{\Sigma}\mathbf{w}^*}{\sqrt{DS}} = k\,\frac{\mathbf{P}\mathbf{w}^*}{\sqrt{DS}}. \tag{317}$$

Thus, on the usual macroscopic time scale, the conditional drift is

$$\mathbf{b}(\mathbf{w}) = -\frac{\eta k}{2\sqrt{D}}\left[\frac{\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}\mathbf{w}}{\sqrt{\widehat{Q}}} - \beta\,\frac{\mathbf{P}\mathbf{w}^*}{\sqrt{S}}\right]. \tag{318}$$

Importantly, one should *not* set $b(w) = 0$: the stochastic SFT dynamics contain an $O(\eta^2)$ contribution to quadratic observables, so a stationary macroscopic state need not correspond to a deterministic fixed weight vector.

Instead, define the projected covariance on Ran$\mathbf{P}$,

$$\mathbf{A} := \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}\big|_{\mathrm{Ran}\mathbf{P}}. \tag{319}$$

Since $\Sigma \succeq \Delta\mathbf{I}$, $\mathbf{A}$ is positive definite on Ran$\mathbf{P}$. Let

$$\mathbf{p} := \mathbf{P}\mathbf{w}^* \tag{320}$$

and introduce an additional order-parameter (linear in $\mathbf{w}$)

$$K := \frac{1}{D}\mathbf{w}^\top\mathbf{A}^{-1}\mathbf{p}. \tag{321}$$

Because K is linear in $\mathbf{w}$, its macroscopic drift contains no quadratic diffusion correction. Using 318,

$$\begin{aligned}\dot{K} &= \frac{1}{D}(\mathbf{A}^{-1}\mathbf{p})^\top\mathbf{b}(\mathbf{w}) \\ &= -\frac{\eta k}{2\sqrt{D}}\left[\frac{J}{\sqrt{\widehat{Q}}} - \beta\,\frac{\mathbf{p}^\top\mathbf{A}^{-1}\mathbf{p}}{\sqrt{S}}\right].\end{aligned} \tag{322}$$

At a stationary macroscopic state, $\dot{K} = 0$, and hence

$$\boxed{c^*_{\mathrm{SFT}} = \frac{J^*}{\sqrt{S\widehat{Q}^*}} = (1 - 2\epsilon)H_\alpha} \tag{323}$$

with

$$\boxed{H_\alpha := \frac{{w^*}^\top P(P\Sigma P)^+ P w^*}{{w^*}^\top w^*} = \frac{{w_B}^\top P w^*}{{w^*}^\top w^*} = \frac{{w_B}^\top w^*}{{w^*}^\top w^*} = \chi.} \tag{324}$$

Recall from the definition of χ, we see that H_α is precisely the fraction of teacher-field variance recoverable from the projected observation. Recalling 306, we have

$$e^*_{\mathrm{SFT}}(\alpha, \epsilon) = \frac{1}{\pi} \arccos c^*_{\mathrm{SFT}} = \frac{1}{\pi} \arccos\left[(1-2\epsilon)\left(1 - A(\Delta,\rho)\frac{1}{\alpha} + O(\alpha^{-3/2})\right)\right], \tag{325}$$

with $A(\Delta,\rho) = \frac{(1-\rho)\Delta^2}{(1-\Delta)^2}$. For clean labels, $\epsilon = 0$, using $\arccos(1-x) = \sqrt{2x} + O(x^{3/2})$,

$$\boxed{e^*_{\mathrm{SFT}}(\epsilon = 0) = \frac{\sqrt{2}}{\pi}\sqrt{A(\Delta,\rho)}\alpha^{-1/2} + O(\alpha^{-1}) = \sqrt{2} e_B,} \tag{326}$$

where the leading $O(\alpha^{-1})$ term is inherited from the $O(\alpha^{-3/2})$ term in 306, not from the $O(\alpha^{-3/2})$ term in the inverse cosine Taylor expansion. Thus the SFT exponent is $-1/2$, identical to the Bayes error exponent. For any fixed $0 < \epsilon < 1/2$, and using $\arccos(a-x) = \arccos(a) - \frac{x}{\sqrt{1-a^2}} + O\left(\frac{a}{(1-a^2)^{3/2}}x^2\right)$,

$$e^*_{\mathrm{SFT}} = \frac{1}{\pi}\arccos(1-2\epsilon) + \frac{(1-2\epsilon)A(\Delta,\rho)}{\pi\sqrt{1-(1-2\epsilon)^2}}\alpha^{-1} + O\left(\epsilon^{-1/2}\alpha^{-3/2} + \epsilon^{-3/2}\alpha^{-2}\right). \tag{327}$$

Where the $O(\epsilon^{-1/2}\alpha^{-3/2}$ term is inherited from the $O(\alpha^{-3/2})$ term in 306, and the $O(\epsilon^{-3/2}\alpha^{-2})$ term is inherited from the $O\left(\frac{a}{(1-a^2)^{3/2}}x^2\right)$ term in the inverse cosine Taylor expansion.

$$e^*_{\mathrm{SFT}} - e_\epsilon = \frac{\beta A(\Delta,\rho)}{\pi\sqrt{1-\beta^2}}\alpha^{-1} + O\left(\epsilon^{-1/2}\alpha^{-3/2} + \epsilon^{-3/2}\alpha^{-2}\right), \tag{328}$$

where $e_\epsilon = \frac{1}{\pi}\arccos(1-2\epsilon)$. 328 is valid only for $\epsilon \gg \alpha^{-1}$, otherwise the $O(\epsilon^{-3/2}\alpha^{-2})$ term is not subleading. We see the raw error approaches a non-zero noise floor, while the excess error above that floor decays as α^{-1}. In the joint small-noise, large-α regime,

$$E^*_{\mathrm{SFT}} \simeq \frac{1}{\pi}\sqrt{4\epsilon + \frac{2\beta A(\Delta,\rho)}{\alpha}}, \tag{329}$$

which displays explicitly the crossover between the clean $\alpha^{-1/2}$ representation-limited regime and the noisy-label error floor.

E RL with n-or-more correct decisions

We now extend the all-correct reward rule to episodes in which at least n of the T decisions must be correct. The thermodynamic limit, Gaussian averages, and operator closure are the same as in Appendix A; only the reward-weighted expectations change. We first calculate these expectations and then derive the consequences for initial learning and asymptotic performance. We distinguish the isotropic, realizable model from the projected model, where information loss and anisotropy can both affect the limiting error.

E.1 Reward-weighted expectations and projected dynamics

For an episode of independent observations, define

$$K = \sum_{t=1}^{T} \mathbb{1}\{y_t = y_t^*\}, \qquad R_n = \mathbb{1}\{K \geq n\}, \qquad 1 \leq n \leq T. \tag{330}$$

The student weights are held fixed within an episode, and the update is

$$\mathbf{w}^{(\mu+1)} = \mathbf{w}^{(\mu)} + \frac{\eta}{T\sqrt{D}} R_n^{(\mu)} \sum_{t=1}^{T} y_t^{(\mu)} \mathbf{P}\mathbf{x}_t^{(\mu)}. \tag{331}$$

Thus unrewarded episodes give no update. We use clean teacher labels and a binary positive reward, without a reward baseline or a penalty. As before, $\xi = \mu/D$ denotes training time, and $D \to \infty$ is taken at fixed T and bounded covariance eigenvalues. The pretraining ratio $\alpha = N/D$ and projector $\mathbf{P}$ are fixed during finetuning.

Recall that the probability of one correct decision is

$$P = \frac{1}{2} + \frac{1}{\pi} \sin^{-1} c, \qquad c = \frac{J_s + J_n}{\sqrt{S\widehat{Q}}}, \qquad \widehat{Q} = Q_s + \Delta Q_n. \tag{332}$$

Condition on one tagged decision. If it is correct, the other $T-1$ decisions must contain at least $n-1$ correct answers; if it is incorrect, they must contain at least n. The corresponding conditional reward probabilities are

$$a_n(P) = \Pr\{\mathrm{Bin}(T-1, P) \geq n-1\}, \qquad b_n(P) = \Pr\{\mathrm{Bin}(T-1, P) \geq n\}. \tag{333}$$

We take a binomial tail to be one when its threshold is zero and zero when the threshold exceeds the number of trials. It is useful to define

$$H_n(P) = a_n(P) + b_n(P), \tag{334}$$

$$h_n(P) = a_n(P) - b_n(P) = \binom{T-1}{n-1} P^{n-1}(1-P)^{T-n}, \tag{335}$$

$$B_n(P) = \langle R_n \rangle = \sum_{k=n}^{T} \binom{T}{k} P^k (1-P)^{T-k} = P a_n(P) + (1-P) b_n(P). \tag{336}$$

The factor h_n is the probability that the other decisions contain exactly $n-1$ correct answers. Only in this event does changing the tagged decision change whether the episode is rewarded.

For any field X depending on the tagged observation, averaging over the remaining observations gives

$$\langle X y R_n \rangle = \langle X y \, [a_n \mathbb{1}\{y = y^*\} + b_n \mathbb{1}\{y \neq y^*\}] \rangle \tag{337}$$

$$= \frac{H_n}{2} \langle X \, \mathrm{sgn}(\lambda) \rangle + \frac{h_n}{2} \langle X \, \mathrm{sgn}(\nu) \rangle. \tag{338}$$

Here we used $\mathbb{1}\{y = y^*\} = (1 + y y^*)/2$ and $y^2 = 1$. This calculation does not assume that X is independent of correctness. The remaining expectations are precisely the Gaussian sign averages already calculated in Appendix A.

Taking $X = \tilde{\nu}_i$ and $X = \lambda_i$ in Equation (338), and substituting into Equation (5), gives

$$\frac{dJ_i}{d\xi} = \frac{\eta}{\sqrt{2\pi}} \left(H_n \frac{C_i^J}{\sqrt{\widehat{Q}}} + h_n \frac{S_i}{\sqrt{S}} \right), \tag{339}$$

$$\frac{dQ_i}{d\xi} = \eta \sqrt{\frac{2}{\pi}} \left(H_n \frac{C_i^Q}{\sqrt{\widehat{Q}}} + h_n \frac{J_i}{\sqrt{S}} \right) + \frac{\eta^2}{T} G_i B_n, \qquad i \in \{s, n\}. \tag{340}$$

All reward probabilities in these equations are evaluated at the current P. The quantities C_i^J, C_i^Q and G_i retain their earlier definitions. In particular, the quadratic contribution contains B_n, since $R_n^2 = R_n$. The T diagonal terms in the squared update supply $\eta^2 G_i B_n / T$; cross terms between distinct observations and the effect of conditioning on the extensive quadratic norm vanish at leading order in D. No additional reward-variance factor appears for this binary update.

The reward averages are exact in the thermodynamic limit. Closing Equation (340) still requires the operator approximation in Equation (21), with the same trace constants and coefficients as before.

The spherical constraint in Appendix A.8 and the local learning-rate calculation in Appendix A.9 apply with these new drift and quadratic coefficients. The scaling laws below, however, are derived for a constant learning rate.

For $n = T$, we have $a_T = P^{T-1}$, $b_T = 0$, and $B_T = P^T$. Thus $H_T = h_T = P^{T-1}$, and Equation (340) reduces to Equation (65). At the other endpoint, $n = 1$, $a_1 = 1$, $b_1 = 1-(1-P)^{T-1}$, and $B_1 = 1 - (1 - P)^T$.

E.2 Isotropic benchmark and initial learning

To isolate the effect of the reward threshold, first consider the unprojected, isotropic model $\mathbf{P} = \mathbf{I}$, $\Sigma = \mathbf{I}$, with $S = 1$, as in the RL perceptron setting of Patel et al. [2025]. This is a separate realizable benchmark, rather than an assumption about the empirical PCA projector. Define

$$J = \frac{\mathbf{w}^\top \mathbf{w}^*}{D}, \qquad Q = \frac{\mathbf{w}^\top \mathbf{w}}{D}, \qquad q = \sqrt{Q}, \qquad c = \frac{J}{q}, \qquad e = 1 - P = \frac{1}{\pi} \arccos c. \tag{341}$$

Using Equation (338) with the teacher and student fields yields

$$\dot{J} = \frac{\eta}{\sqrt{2\pi}}(H_n c + h_n), \tag{342}$$

$$\dot{Q} = \eta q \sqrt{\frac{2}{\pi}}(H_n + h_n c) + \frac{\eta^2}{T} B_n, \tag{343}$$

$$\dot{q} = \frac{\eta}{\sqrt{2\pi}}(H_n + h_n c) + \frac{\eta^2}{2Tq} B_n. \tag{344}$$

Dots denote derivatives with respect to ξ. Differentiating $c = J/q$, the terms proportional to H_n cancel, leaving

$$\boxed{\dot{c} = \frac{\eta}{q\sqrt{2\pi}} h_n (1 - c^2) - \frac{\eta^2}{2Tq^2} B_n c.} \tag{345}$$

The first term improves alignment with the teacher, while the second is the angular cost of the stochastic update. A radial rescaling does not change c, so Equation (345) also applies with a spherical constraint, with q held fixed. The norm equation itself describes only the unconstrained dynamics.

At independent random initialization, $c(0) = 0$ and $P(0) = 1/2$. Hence

$$\dot{c}_n(0) = \frac{\eta}{q_0\sqrt{2\pi}} \binom{T-1}{n-1} 2^{-(T-1)}, \tag{346}$$

$$\boxed{\frac{\dot{c}_n(0)}{\dot{c}_{n=T}(0)} = \binom{T-1}{n-1}.} \tag{347}$$

This compares equal initial norms and learning rates. It is an initial-slope multiplier, rather than an exact rescaling of learning time. In particular, the multiplier is largest near the middle of the binomial row and returns to one at $n = 1$.

The same multiplier holds for the projected model at independent random initialization. There, J_i and the covariance-weighted teacher overlaps C_i^J vanish initially. Since $c = 0$, changes in its denominator do not contribute to its initial derivative, and Equation (340) gives

$$\dot{c}_n(0) = \frac{\eta(S_s + S_n)}{S\sqrt{2\pi \widehat{Q}_0}} h_n(1/2). \tag{348}$$

The geometry changes the common prefactor, but not the threshold multiplier when that prefactor is nonzero. Merely choosing a specially constructed student with $J_s + J_n = 0$ is insufficient: its higher teacher-weighted overlaps need not vanish.

E.3 Scaling laws in the isotropic model

Write $m = T - n$ for the number of allowed mistakes and $C_{T,n} = \binom{T-1}{n-1}$. Transforming Equation (345) using $c = \cos(\pi e)$ gives

$$\dot{e} = -\frac{\eta}{\pi q\sqrt{2\pi}} h_n \sin(\pi e) + \frac{\eta^2}{2\pi T q^2} B_n \cot(\pi e). \tag{349}$$

As $e \to 0$ at fixed T, n, $h_n = C_{T,n}(1-e)^{n-1}e^m \sim C_{T,n}e^m$ and $B_n \to 1$. Expanding the trigonometric factors therefore gives

$$\dot{e} \simeq -\frac{C_{T,n}}{\sqrt{2\pi}} \frac{\eta}{q} e^{m+1} + \frac{1}{2\pi^2 T} \left(\frac{\eta}{q}\right)^2 \frac{1}{e}. \tag{350}$$

The asymptotic balance depends on whether the student norm is allowed to grow.

Unconstrained weights. For constant $\eta > 0$, on a trajectory approaching $c = 1$, $H_n + h_n c \to 2$. This holds both for $n < T$, where $a_n, b_n \to 1$, and for $n = T$, where $a_T \to 1$ and $b_T = 0$. Thus Equation (344) implies

$$q(\xi) \sim \eta\sqrt{\frac{2}{\pi}}\,\xi, \qquad \dot{e} \simeq -\frac{C_{T,n}}{2\xi} e^{m+1} + \frac{1}{4\pi T \xi^2 e}. \tag{351}$$

For $m \geq 1$, neglecting the second term gives

$$\frac{d(e^{-m})}{d\xi} \simeq \frac{mC_{T,n}}{2\xi}, \qquad e(\xi)^{-m} \simeq e(\xi_0)^{-m} + \frac{mC_{T,n}}{2} \log\frac{\xi}{\xi_0}. \tag{352}$$

Consequently,

$$\boxed{e(\xi) \sim \left[\frac{T-n}{2}\binom{T-1}{n-1}\log\frac{\xi}{\xi_0}\right]^{-1/(T-n)}, \qquad n < T.} \tag{353}$$

Here $\xi_0 > 0$ is a fixed reference time. The approximation is self-consistent: the ratio of the discarded stochastic term to the alignment term is proportional to

$$\frac{1}{\xi e^{m+2}} \sim \frac{(\log\xi)^{(m+2)/m}}{\xi} \longrightarrow 0.$$

Although almost every episode is rewarded late in learning, the reward discriminates between correct and incorrect decisions only with probability of order e^m. Together with the decreasing effective rate η/q, this produces inverse-logarithmic learning. The constant η cancels from the leading asymptotic coefficient, but still affects the transient.

The all-correct boundary $n = T$ must be treated separately. Here $m = 0$ and $C_{T,T} = 1$, and both terms in Equation (351) contribute. Setting $u = e^2$ gives

$$\dot{u} \simeq -\frac{u}{\xi} + \frac{1}{2\pi T\xi^2}, \qquad \frac{d(\xi u)}{d\xi} \simeq \frac{1}{2\pi T\xi}. \tag{354}$$

Integrating yields

$$\boxed{e(\xi) \sim \sqrt{\frac{\log(\xi/\xi_0)}{2\pi T\xi}}, \qquad n = T.} \tag{355}$$

Thus the all-correct rule has algebraic decay with a logarithmic correction.

Fixed norm. Now hold q fixed and define the effective learning rate $\gamma = \eta/q$. At a small-error equilibrium, Equation (350) gives

$$\frac{C_{T,n}}{\sqrt{2\pi}} \gamma e_\infty^{m+1} \simeq \frac{\gamma^2}{2\pi^2 T e_\infty}, \tag{356}$$

and hence

$$\boxed{e_\infty \sim \left[\frac{\eta/q}{\sqrt{2}\,\pi^{3/2}T\binom{T-1}{n-1}}\right]^{1/(T-n+2)}}. \tag{357}$$

This is a small-rate asymptotic, $\eta/q \to 0$, at fixed T, n. For a finite rate, setting Equation (349) to zero instead gives the implicit equilibrium equation on the positive-correlation branch,

$$\frac{\eta}{q} = \sqrt{\frac{2}{\pi}}\,T\,\frac{C_{T,n}(1-e_\infty)^{n-1}e_\infty^{T-n}\sin^2(\pi e_\infty)}{B_n(1-e_\infty)\cos(\pi e_\infty)}. \tag{358}$$

Unlike Equation (353), Equation (357) includes $n = T$, where the exponent is $1/2$; both all-correct results also include $T = n = 1$. Substituting a growing $q(\xi)$ into the fixed-norm floor gives an instantaneous drift-balance scale, not necessarily the unconstrained trajectory. These constant-rate results do not determine the asymptotics under the locally optimal schedule of Appendix A.9.

E.4 Information loss and anisotropy under projection

We return to a fixed projector $\mathbf{P}$. Let $\mathbf{U} \in \mathbb{R}^{D\times r}$ have orthonormal columns spanning its range, so $\mathbf{P} = \mathbf{U}\mathbf{U}^\top$, and define

$$\mathbf{z} = \mathbf{U}^\top\mathbf{x}, \qquad \mathbf{v} = \mathbf{U}^\top\mathbf{w}, \qquad \mathbf{C} = \mathbf{U}^\top\Sigma\mathbf{U}, \qquad \mathbf{g} = \mathbf{U}^\top\Sigma\mathbf{w}^*. \tag{359}$$

Then

$$\widehat{Q} = \frac{\mathbf{v}^\top\mathbf{C}\mathbf{v}}{D}, \qquad J = J_s + J_n = \frac{\mathbf{v}^\top\mathbf{g}}{D}, \qquad q^2 = Q_s + Q_n = \frac{\mathbf{v}^\top\mathbf{v}}{D}. \tag{360}$$

In general $\widehat{Q} \neq q^2$, so the isotropic angular cancellation cannot be assumed for this model.

The information bound. The conditional Gaussian representation from Appendix C.3 can be written in retained coordinates as

$$\nu = \frac{\mathbf{g}^\top\mathbf{C}^{-1}\mathbf{z}}{\sqrt{D}} + \zeta, \qquad \zeta \perp \mathbf{z}, \qquad \mathrm{Var}(\zeta) = S - \frac{\mathbf{g}^\top\mathbf{C}^{-1}\mathbf{g}}{D}. \tag{361}$$

The inverse acts on supported covariance directions; it is a pseudoinverse if the retained covariance is singular. The Bayes direction is $\mathbf{v}_B \propto \mathbf{C}^{-1}\mathbf{g}$. In the notation of Equation (301),

$$\chi = \frac{\mathbf{g}^\top\mathbf{C}^{-1}\mathbf{g}}{DS}, \qquad \kappa = \sqrt{\chi}, \qquad e_B = \frac{1}{\pi}\arccos\kappa. \tag{362}$$

This lower bound applies to any classifier using the projected observation. It is independent of n and of the learning-rate schedule. A positive conditional variance in Equation (361) gives $e_B > 0$, which reducing η/q cannot remove.

For the covariance and signal-supported teacher of our model,

$$\mathbf{C} = \Delta\mathbf{I}_r + (1-\Delta)\mathbf{U}^\top\mathbf{M}_s\mathbf{U}, \qquad \mathbf{g} = \mathbf{U}^\top\mathbf{w}^*. \tag{363}$$

If ℓ_a are the eigenvalues of $\mathbf{U}^\top\mathbf{M}_s\mathbf{U}$ and g_a are the teacher coordinates in this eigenbasis, then

$$\chi = \frac{1}{DS}\sum_{a=1}^{r}\frac{g_a^2}{\Delta + (1-\Delta)\ell_a}. \tag{364}$$

This is the same recoverable variance used in the large-α analysis. The residual variance, rather than subspace misalignment alone, determines whether the information floor is strictly positive.

The direction selected by RL. Applying Equation (338) to each retained coordinate gives the conditional mean drift per unit training time,

$$\left.\frac{d\mathbf{v}}{d\xi}\right|_{\mathrm{mean}} = \frac{\eta}{\sqrt{2\pi}}\left(H_n\frac{\mathbf{C}\mathbf{v}}{\sqrt{\widehat{Q}}} + h_n\frac{\mathbf{g}}{\sqrt{S}}\right). \tag{365}$$

This describes the mean update; the quadratic terms in the macroscopic dynamics remain necessary at finite effective learning rate. When $\mathbf{Cv}$ is not parallel to $\mathbf{v}$, the H_n term also changes direction, reinforcing components of larger retained variance. In the limit $\eta/q \to 0$, a stationary direction must therefore satisfy

$$H_n(P_0)\frac{\mathbf{Cv}_0}{\sqrt{\widehat{Q}_0}} + h_n(P_0)\frac{\mathbf{g}}{\sqrt{S}} = \Lambda \mathbf{v}_0, \tag{366}$$

where Λ accounts for radial growth or spherical normalization. At the Bayes direction $\mathbf{v}_0 \propto \mathbf{C}^{-1}\mathbf{g}$, the left-hand side is proportional to $\mathbf{g}$. It is generally not parallel to $\mathbf{C}^{-1}\mathbf{g}$ unless the teacher-bearing direction lies within a single covariance eigenspace. Thus the Bayes classifier need not be stationary under RL, even when the effective rate vanishes. If the dynamics converge to a stable stationary branch, its limiting error e_0 obeys $e_0 \geq e_B$ and may depend on n through H_n/h_n. This argument does not establish uniqueness or global stability of the branch.

Isotropic retained covariance with missing information. To separate information loss from anisotropic drift, consider an auxiliary projected Gaussian model with $\mathbf{C} = \mathbf{I}_r$ and $0 < \kappa < 1$. This is a solvable special case, not an assumption about the empirical PCA covariance in our model. Write $t = r/D = \mathrm{Tr}(\mathbf{P})/D$ for the retained fraction, preserving the notation for the trace constant, and $\gamma = \eta/q$. Here $\widehat{Q} = q^2$ and $\|\mathbf{g}\|^2/(DS) = \kappa^2$. Taking the teacher overlap and norm of the update gives

$$\dot{J} = \frac{\eta\sqrt{S}}{\sqrt{2\pi}}(H_n c + h_n \kappa^2), \tag{367}$$

$$\dot{q} = \frac{\eta}{\sqrt{2\pi}}(H_n + h_n c) + \frac{\eta^2 t}{2Tq}B_n, \tag{368}$$

$$\boxed{\dot{c} = \frac{\gamma}{\sqrt{2\pi}}h_n(\kappa^2 - c^2) - \frac{\gamma^2 t}{2T}B_n c.} \tag{369}$$

The quadratic term contains t because $\mathrm{Tr}(\mathbf{C})/D = r/D$. The deterministic angular target is $c = \kappa$, rather than $c = 1$.

At fixed norm, let $c_* = \kappa - \delta c$ and set $P_B = 1 - e_B$. Denote reward factors evaluated at P_B by $h_B = h_n(P_B)$, $H_B = H_n(P_B)$, and $B_B = B_n(P_B)$. Expanding Equation (369) at small γ gives

$$\frac{\gamma}{\sqrt{2\pi}}h_B\, 2\kappa\delta c \simeq \frac{\gamma^2 t}{2T}B_B\kappa, \qquad\qquad \delta c \simeq \frac{\gamma t B_B\sqrt{2\pi}}{4T h_B}. \tag{370}$$

Using $de/dc = -1/(\pi\sqrt{1-c^2})$, we obtain

$$\boxed{e_\infty = e_B + \frac{tB_B}{2T\sqrt{2\pi}\,h_B\sqrt{1-\kappa^2}}\frac{\eta}{q} + o(\eta/q).} \tag{371}$$

The excess error is linear in the effective rate at fixed $0 < \kappa < 1$ and fixed T, n. Its coefficient can be large, since

$$h_B = \binom{T-1}{n-1}(1-e_B)^{n-1}e_B^{T-n}. \tag{372}$$

Thus the small-rate expansion is not uniform as $\kappa \to 1$ or as T grows. In particular, a permissive threshold and a small information floor may require a very small effective rate before the linear regime is reached.

For unconstrained weights and constant η, a trajectory approaching $c = \kappa$ instead has

$$q(\xi) \sim \eta L \xi, \qquad L = \frac{H_B + \kappa h_B}{\sqrt{2\pi}}, \qquad \gamma \sim \frac{1}{L\xi}. \tag{373}$$

Writing $\delta c = \kappa - c$ and linearizing yields

$$\frac{d\delta c}{d\xi} \simeq -\frac{\beta_n}{\xi}\delta c + \frac{tB_B\kappa}{2TL^2\xi^2}, \qquad\qquad \beta_n = \frac{2\kappa h_B}{H_B + \kappa h_B}. \tag{374}$$

Since $H_B \geq h_B > 0$, we have $0 < \beta_n \leq 2\kappa/(1+\kappa) < 1$. Multiplication by ξ^{β_n} and integration show that the homogeneous term decays as $\xi^{-\beta_n}$, while the quadratic-update forcing in this linearized equation contributes a term proportional to ξ^{-1}. For a nonzero leading amplitude,

$$\boxed{e(\xi) - e_B \sim A\xi^{-\beta_n},} \tag{375}$$

where A depends on the trajectory. Nonlinear corrections can also contribute subleading powers. For $n = T$, $H_B = h_B$ and $\beta_T = 2\kappa/(1+\kappa)$. For $n < T$, a small h_B can make the relaxation extremely slow.

At a strictly positive information floor the angular fixed point can be linearized because $h_B > 0$. As $e_B \to 0$, $h_B \to 0$ for $n < T$ and the exponent β_n tends to zero. The nonlinear terms leading to Equation (353) then become essential. The long-time and vanishing-information-loss limits are therefore nonuniform: the realizable law may describe a crossover when $e \gg e_B$, but it is not the final relaxation law at fixed positive e_B in this special case.

E.5 Fixed points of the projected four-order-parameter dynamics

For the general covariance in our model, we use the same operator closure as in Equation (21). Introduce normalized order parameters

$$q^2 = Q_s + Q_n, \qquad j_i = J_i/q, \qquad u_i = Q_i/q^2, \qquad u_s + u_n = 1, \qquad \widehat{u} = u_s + \Delta u_n. \tag{376}$$

There are three independent coordinates, $\mathbf{x} = (j_s, j_n, u_s)^\top$, with $u_n = 1 - u_s$, and

$$e(\mathbf{x}) = \frac{1}{\pi}\arccos\left(\frac{j_s + j_n}{\sqrt{S\widehat{u}}}\right), \qquad P(\mathbf{x}) = 1 - e(\mathbf{x}). \tag{377}$$

To construct their dynamics explicitly, define the normalized closed moments

$$\mathcal{C}_i^J(\mathbf{x}) = \sum_{k\in\{s,n\}} (\Omega_{is}^{(k)} + \Delta\Omega_{in}^{(k)}) j_k, \tag{378}$$

$$\mathcal{C}_i^Q(\mathbf{x}) = \sum_{k\in\{s,n\}} (\Omega_{is}^{(k)} + \Delta\Omega_{in}^{(k)}) u_k. \tag{379}$$

Within the closure, $C_i^J = q\mathcal{C}_i^J$ and $C_i^Q = q^2\mathcal{C}_i^Q$. Consequently, Equation (340) takes the form

$$\dot{J}_i = \eta\mathcal{A}_i(\mathbf{x}), \qquad \dot{Q}_i = \eta q\mathcal{D}_i(\mathbf{x}) + \eta^2\mathcal{B}_i(\mathbf{x}), \tag{380}$$

where

$$\mathcal{A}_i = \frac{1}{\sqrt{2\pi}}\left(H_n\frac{\mathcal{C}_i^J}{\sqrt{\widehat{u}}} + h_n\frac{S_i}{\sqrt{S}}\right), \tag{381}$$

$$\mathcal{D}_i = \sqrt{\frac{2}{\pi}}\left(H_n\frac{\mathcal{C}_i^Q}{\sqrt{\widehat{u}}} + h_n\frac{j_i}{\sqrt{S}}\right), \qquad \mathcal{B}_i = \frac{G_i}{T}B_n. \tag{382}$$

Let $\mathcal{D} = \mathcal{D}_s + \mathcal{D}_n$ and $\mathcal{B} = \mathcal{B}_s + \mathcal{B}_n$. The unconstrained radial equation is

$$\dot{q} = \frac{\eta}{2}\mathcal{D} + \frac{\eta^2}{2q}\mathcal{B}. \tag{383}$$

Differentiating $j_i = J_i/q$ and $u_i = Q_i/q^2$ now gives

$$\dot{j}_i = \gamma\left(\mathcal{A}_i - \frac{j_i}{2}\mathcal{D}\right) - \gamma^2\frac{j_i}{2}\mathcal{B}, \tag{384}$$

$$\dot{u}_i = \gamma(\mathcal{D}_i - u_i\mathcal{D}) + \gamma^2(\mathcal{B}_i - u_i\mathcal{B}), \qquad \gamma = \eta/q. \tag{385}$$

Thus the normalized dynamics can be written as

$$\dot{\mathbf{x}} = \gamma\mathbf{F}_{T,n}(\mathbf{x}) + \gamma^2\mathbf{G}_{T,n}(\mathbf{x}), \tag{386}$$

$$\mathbf{F}_{T,n} = \begin{pmatrix} \mathcal{A}_s - j_s\mathcal{D}/2 \\ \mathcal{A}_n - j_n\mathcal{D}/2 \\ \mathcal{D}_s - u_s\mathcal{D} \end{pmatrix}, \qquad \mathbf{G}_{T,n} = \begin{pmatrix} -j_s\mathcal{B}/2 \\ -j_n\mathcal{B}/2 \\ \mathcal{B}_s - u_s\mathcal{B} \end{pmatrix}. \tag{387}$$

The same normalized equations apply with the spherical constraint, since the radial corrections in Appendix A.8 leave j_i, u_i unchanged. In that case q, and hence γ for constant η, is fixed.

Consider a fixed-norm stationary branch at small γ. Its limiting state solves $\mathbf{F}_{T,n}(\mathbf{x}_0) = 0$ on the physically admissible domain; we do not impose zero error. Suppose this state is locally asymptotically stable, its Jacobian $\mathbf{M} = \mathrm{D}\mathbf{F}_{T,n}(\mathbf{x}_0)$ is invertible in the three independent coordinates, and the coefficient functions are smooth nearby. Substituting $\mathbf{x}_*(\gamma) = \mathbf{x}_0 + \gamma\mathbf{x}_1 + O(\gamma^2)$ into $\mathbf{F}_{T,n}(\mathbf{x}_*) + \gamma\mathbf{G}_{T,n}(\mathbf{x}_*) = 0$ gives

$$\mathbf{M}\mathbf{x}_1 + \mathbf{G}_{T,n}(\mathbf{x}_0) = 0, \qquad \mathbf{x}_*(\gamma) = \mathbf{x}_0 - \gamma\mathbf{M}^{-1}\mathbf{G}_{T,n}(\mathbf{x}_0) + O(\gamma^2). \tag{388}$$

Provided the error is also smooth at $\mathbf{x}_0$,

$$\boxed{e_\infty(\gamma) = e_0 - \gamma\nabla e(\mathbf{x}_0)^\top\mathbf{M}^{-1}\mathbf{G}_{T,n}(\mathbf{x}_0) + O(\gamma^2), \qquad e_0 = e(\mathbf{x}_0).} \tag{389}$$

Thus a regular projected fixed point has an $O(\eta/q)$ correction, whose coefficient can vanish, about its limiting error. The conclusion is local to the selected stable branch; degenerate fixed points, branch changes, and the singular realizable boundary require separate analysis.

At fixed norm, local relaxation is determined by the eigenvalues of $\gamma\mathrm{D}\mathbf{F}_{T,n}(\mathbf{x}_*) + \gamma^2\mathrm{D}\mathbf{G}_{T,n}(\mathbf{x}_*)$. For unconstrained weights, suppose a trajectory approaches $\mathbf{x}_0$ with $\mathcal{D}(\mathbf{x}_0) > 0$. Then Equation (383) gives $q \sim \eta L_0\xi$, where $L_0 = \mathcal{D}(\mathbf{x}_0)/2$. To leading order, the perturbation satisfies

$$\frac{d\delta\mathbf{x}}{d\xi} \simeq \frac{\mathbf{M}}{L_0\xi}\delta\mathbf{x} + \frac{\mathbf{G}_{T,n}(\mathbf{x}_0)}{L_0^2\xi^2}. \tag{390}$$

A homogeneous eigenmode with Jacobian eigenvalue λ_a therefore scales as ξ^{λ_a/L_0}. Quadratic-update forcing, resonances, nonlinear corrections, and the projection of these modes onto the error can affect the leading observable relaxation. There is no universal late-time exponent depending only on $T - n$ for the anisotropic model.

The fixed points and relaxation rates in this subsection are predictions of the four-order-parameter closure. In contrast, the information bound $e \geq e_B$ is exact and independent of that approximation. A closed-ODE prediction below this bound would indicate a failure of the closure to represent the physical dynamics.

F Compute cost calculation details

F.1 Pretraining cost

For an LAE with encoder dimension K, data dimension D and batch size B, usingf the convention that one multiply-add costes 2 FLOPs, the cost of a single forward pass is $2KBD + 2KBD = 4KBD$. For the backward pass, the cost is decoder weight gradient + gradient into hidden layer + encoder weight gradient, which is $2KBD + 2KBD + 2KBD = 6KBD$. Therefore the total cost per step is $C_{\mathrm{LAE}} = 10KBD$. The total pretraining cost is then

$$C_{\mathrm{PT}} = 10N_{\mathrm{steps}}KBD = 10N_{\mathrm{steps}}\alpha\rho D^3 = \mathcal{O}(N_{\mathrm{steps}}\alpha\rho D^3),$$

where N_{steps} is the number of training steps taken to convergence.

F.2 SFT and RL step costs

Each step requires projecting observed data into the encoder subspace, which is a dense matrix-vector multiplication costing $2KD$ FLOPs. The SFT/RL steps then requires a forward pass through the student model, this costs $2K$ FLOPs + K for the update. This must be calculated for each observation in an episode of length T. Therefore the total cost per SFT/RL step is

$$C_{\mathrm{SFT/RL}} = 2\rho TD^2 + 3\rho TD = \mathcal{O}(\rho TD^2).$$

In simulation of the stochastic updates to obtain the deterministic ODEs limit, D subupdates per epoch are made, so we can write

$$C_{\mathrm{SFT/RL}} = \mathcal{O}(\rho TD^3).$$

G Experimental Details of LAE training

For each batchsize N 20 LAEs were trained to convergence, the overlap of the projection of the final encoders with the signal subspace was then calculated and averaged over runs. The simple LAE had an encoder $\mathbf{E} \in \mathbb{R}^{(D\times K)}$ and a decoder $\mathbf{D} \in \mathbb{R}^{(K\times D)}$, with $D = 2000$ and $K = 400$. Weights were initialised from zero mean Gaussians with variance $1/D$ and $1/K$ for the encoder and decoder respectively. Adam optimizer with learning rate 0.0001, and 10000 steps were used ensuring convergence. JAX with float32 precision was used.

H Optimal Compute Allocation Experimental details

In addition to solving the projected ODEs for fixed hyperparameters, we also optimize how a fixed computational budget is allocated between pretraining data selection, SFT updates, and RL updates. The optimization is performed over the deterministic ODE dynamics rather than over stochastic simulations. This makes it possible to evaluate many candidate allocations efficiently.

Let B_{tot} denote the total available budget. We allocate this budget between three stages:

$$C_{PT}, \qquad C_{\mathrm{SFT}}, \qquad C_{\mathrm{RL}},$$

where C_{PT} is the budget spent on choosing the pretraining level α, C_{SFT} is the budget spent on SFT updates, and C_{RL} is the budget spent on RL updates. The allocation satisfies

$$C_{PT} + C_{\mathrm{SFT}} + C_{\mathrm{RL}} = B_{\mathrm{tot}}.$$

The cost of choosing a pretraining level α is taken to be linear,

$$C_{PT} = c_\alpha \alpha,$$

where c_α is the cost per unit α. The remaining budget is

$$B_{\mathrm{ODE}}(\alpha) = B_{\mathrm{tot}} - c_\alpha \alpha.$$

For the combined SFT–RL setting, we parameterize the split of this remaining budget by a fraction $f_{\mathrm{SFT}} \in [0, 1]$:

$$C_{\mathrm{SFT}} = f_{\mathrm{SFT}} B_{\mathrm{ODE}}(\alpha), \qquad C_{\mathrm{RL}} = (1 - f_{\mathrm{SFT}}) B_{\mathrm{ODE}}(\alpha).$$

If the cost per SFT step is a, and the cost per RL step is b, then the corresponding numbers of ODE steps are

$$n_{\mathrm{SFT}} = \frac{C_{\mathrm{SFT}}}{a}, \qquad n_{\mathrm{RL}} = \frac{C_{\mathrm{RL}}}{b}.$$

The objective is to maximize the final clean teacher–student agreement probability after applying n_{SFT} SFT steps followed by n_{RL} RL steps:

$$\max_{\alpha, f_{\mathrm{SFT}}} P\left(R_{\mathrm{RL}}^{n_{\mathrm{RL}}} \circ R_{\mathrm{SFT}}^{n_{\mathrm{SFT}}}(\mathbf{q}_0(\alpha))\right),$$

where

$$\mathbf{q} = (J_s, J_n, Q_s, Q_n)$$

denotes the ODE state, $\mathbf{q}_0(\alpha)$ is the initial condition induced by the PCA projector learned at pretraining level α, and p_{gen} is the probability of agreement with the clean teacher.

H.1 Precomputation and interpolation over α

For each value of α, the initial condition and closure coefficients are computed from the random-matrix trace constants described in Appendix B. Recomputing these constants at every candidate value of α would be expensive. We therefore first evaluate them on a fixed interpolation grid

$$\alpha_1, \ldots, \alpha_M,$$

with $M =$ `interp_alpha_res`. In the experiments reported here, `interp_alpha_res` $= 200$ unless otherwise stated.

For an arbitrary candidate α, the initial condition

$$(J_s, J_n, Q_s, Q_n)$$

and the associated constants

$$S, \quad S_s, \quad S_n, \quad \Omega_{ij}^{(s)}, \quad \Omega_{ij}^{(n)}, \quad T_{ss}, \quad T_{sn}, \quad T_{nn}, \quad R$$

are obtained by linear interpolation between the precomputed values. Here

$$R = Q_s + Q_n$$

is the projected student norm used when the spherical constraint is imposed.

This interpolation step allows the allocation search to vary α continuously over the search interval while avoiding repeated evaluations of the random-matrix formulas.

H.2 ODE evaluation for a fixed allocation

For a fixed pair $(\alpha, f_{\mathrm{SFT}})$, the algorithm first computes the available ODE budget

$$B_{\mathrm{ODE}}(\alpha) = B_{\mathrm{tot}} - c_\alpha \alpha.$$

It then converts the SFT and RL budget shares into fractional step counts:

$$n_{\mathrm{SFT}} = \frac{f_{\mathrm{SFT}} B_{\mathrm{ODE}}(\alpha)}{a}, \qquad n_{\mathrm{RL}} = \frac{(1 - f_{\mathrm{SFT}}) B_{\mathrm{ODE}}(\alpha)}{b}.$$

The SFT trajectory is computed once per value of α. Starting from $\mathbf{q}_0(\alpha)$, the SFT ODE update is iterated up to a maximum static number of steps,

$$n_{\max} = \min\left\{ \left\lfloor \frac{B_{\mathrm{tot}} - c_\alpha \alpha_{\min}}{\min(a, b)} \right\rfloor + 10,\ n_{\mathrm{cap}} \right\},$$

where $n_{\mathrm{cap}} = \texttt{max_ode_steps}$. The additional $+10$ buffer prevents interpolation errors near the largest affordable step count. The code stores the full SFT prefix

$$\mathbf{q}_0, \mathbf{q}_1^{\mathrm{SFT}}, \ldots, \mathbf{q}_{n_{\max}}^{\mathrm{SFT}}.$$

For a fractional value of n_{SFT}, the state after SFT is obtained by linear interpolation between the two neighboring integer SFT iterates.

Starting from this interpolated SFT state, the RL ODE update is applied for $\lfloor n_{\mathrm{RL}} \rfloor$ integer steps. If n_{RL} is fractional, one additional RL step is computed and the final state is linearly interpolated between the integer-step and next-step states. The final performance is then computed using the analytic expression for the clean generalization probability,

$$P = \frac{1}{2} + \frac{1}{\pi} \sin^{-1}\left(\frac{J_s + J_n}{\sqrt{S(Q_s + \Delta Q_n)}} \right).$$

H.3 Coarse-to-fine grid search

The budget-allocation problem is solved by a coarse-to-fine grid search. For a given total budget, the largest affordable pretraining level is

$$\alpha_{\max}^{\mathrm{aff}} = \min\left\{ \alpha_{\max}, \frac{B_{\mathrm{tot}}}{c_\alpha} \right\}.$$

If $\alpha_{\max}^{\mathrm{aff}} < \alpha_{\min}$, the budget is declared infeasible.

The coarse search evaluates a grid of candidate α values and SFT fractions:

$$\alpha \in \{\alpha_1^{\mathrm{coarse}}, \ldots, \alpha_{N_\alpha}^{\mathrm{coarse}}\}, \qquad f_{\mathrm{SFT}} \in \{f_1^{\mathrm{coarse}}, \ldots, f_{N_f}^{\mathrm{coarse}}\}.$$

The α-grid may be linearly spaced or warped according to

$$\alpha(u) = \alpha_{\min} + (\alpha_{\max}^{\mathrm{aff}} - \alpha_{\min}) u^p, \qquad u \in [0, 1],$$

where $p = \texttt{alpha_grid_power}$. This allows more resolution near small α when desired.

After evaluating the full coarse grid, the algorithm identifies the best `top_k` coarse α-regions according to

$$\max_{f_{\mathrm{SFT}}} p_{\mathrm{gen}}(\alpha, f_{\mathrm{SFT}}).$$

Each selected region is then refined. The refined α-interval is formed by taking a small neighborhood around the selected coarse α-cell, with width controlled by `refine_alpha_half_width_cells`. Within each selected region, the algorithm evaluates a finer grid of `refine_alpha_res` values of α. The SFT fraction is refined over the full interval $f_{\mathrm{SFT}} \in [0, 1]$, rather than only near the coarse optimum. This is useful because the optimal SFT–RL split can shift substantially as α changes, especially near the boundary where the remaining ODE budget changes rapidly.

The best refined candidate over all selected regions is returned. The reported allocation is

$$(C_{PT}, C_{\mathrm{SFT}}, C_{\mathrm{RL}}) = (c_\alpha \alpha, f_{\mathrm{SFT}} B_{\mathrm{ODE}}(\alpha), (1 - f_{\mathrm{SFT}}) B_{\mathrm{ODE}}(\alpha)),$$

together with the corresponding final generalization performance.

H.4 Single-stage baselines

We also evaluate two single-stage baselines. In the SFT-only baseline,

$$f_{\mathrm{SFT}} = 1, \qquad C_{\mathrm{RL}} = 0.$$

The search is then one-dimensional over α, and the full remaining ODE budget is spent on SFT:

$$n_{\mathrm{SFT}} = \frac{B_{\mathrm{tot}} - c_\alpha \alpha}{a}.$$

In the RL-only baseline,

$$f_{\mathrm{SFT}} = 0, \qquad C_{\mathrm{SFT}} = 0,$$

and the remaining budget is spent entirely on RL:

$$n_{\mathrm{RL}} = \frac{B_{\mathrm{tot}} - c_\alpha \alpha}{b}.$$

Both baselines use the same coarse-to-fine search over α as the full SFT–RL allocation problem.

H.5 Implementation details

The allocation search is implemented in JAX and run on CPU with 64-bit floating point arithmetic enabled. The ODE update maps for SFT and RL are JIT-compiled. The coarse and refined grids are evaluated using vectorized JAX loops. For each candidate α, the SFT prefix trajectory is computed once and reused for all candidate values of f_{SFT}. This reuse is the main computational saving in the two-stage search, because each candidate fraction only requires interpolating along the stored SFT trajectory before running the RL segment.

Invalid numerical evaluations, such as candidates producing NaN values, are assigned performance $-\infty$ and are therefore excluded from the optimum. For each run, the code saves a JSON file containing the training mode, total budget, task parameters T and ϵ, runtime, maximum ODE step count, best allocation, best performance, selected α, selected f_{SFT}, refined search regions, and the full search configuration. The saved search configuration includes the coarse and refined grid resolutions, the number of top coarse regions refined, the α-grid warping parameter, and the interpolation resolution used for the precomputed random-matrix constants.

I Skill Composition

We adopt linear composition of multiple RL perceptrons following Lee et al. [2024]. In this section, we make a short summary of the framework to make the paper self-contained. Total K of RL perceptrons, which is an abstraction of K independent skills are linearly combined with compositional context vector $\mathbf{V} \in \mathbb{R}^K$. Each element V_k indicates the contribution of k_{th} skill. We We restrict our model with $\|\mathbf{V}\| = 1$. We notate $\mathbf{V}^*$ as teacher and $\mathbf{V}$ as a student, and $\mathbf{V}$ is updated as the agent gets learning signal. The skill composition task is defined as following:

$$\tilde{y}_t^* = \mathrm{sign}\left(\sum_{k=1}^{K} \mathbf{V}_k^* \frac{\mathbf{w}_k^{*\intercal} \mathbf{x}_{k,t}}{\sqrt{N}} \right), \tag{391}$$

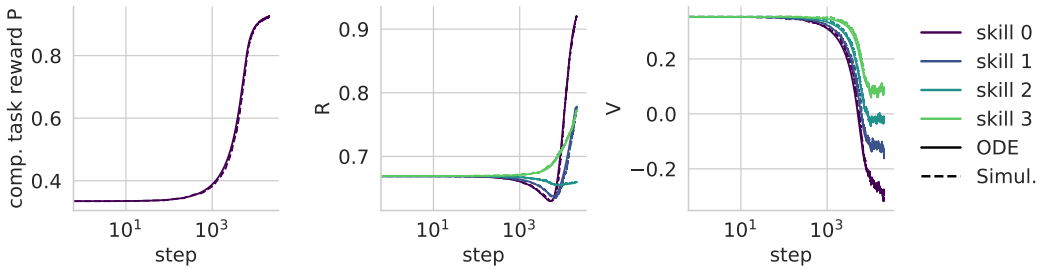


Figure 9: Comparison of ODEs and simulation from skill compositions. We show dynamics of compositional task reward probability (left), each skill level (R_{0-3}, middle) and composition context (V_{0-3}, right). The integration of ODEs and simulation show good match. ($T = 8$, $K = 8$ (we visualize first 4 skills out of 8 skills in total), $\eta_w = \eta v = 1.0$.)

where we use " ˜ " to notate skill composition task. The updates of the student $\mathbf{w}$ and $\mathbf{V}$ becomes

$$\Delta \mathbf{w}_k^{n+1} = \frac{\eta_{\mathbf{w}}}{\sqrt{N}} \left(\frac{1}{T} \sum_{t=1}^{T} \tilde{y}_t V_k \mathbf{x}_{k,t} \right) \mathbb{I}(R(\tilde{y}_{1:T}, \tilde{y}^*_{1:T})) \tag{392}$$

$$\Delta V_k^{n+1} = \frac{\eta_v}{N} \left(\frac{1}{T} \sum_{t=1}^{T} \tilde{y}_t \frac{\mathbf{w}_k{}^\intercal x_{k,t}}{\sqrt{N}} \right) \mathbb{I}(R(\tilde{y}_{1:T}, \tilde{y}^*_{1:T})). \tag{393}$$

In thermodynamic limit, the ODEs of the order parameters are as following,

$$\begin{aligned}
\frac{dQ_k}{d\xi} &= \frac{2\eta_w}{\sqrt{2\pi}} \left(\frac{V_k^2 Q_k}{\sqrt{\sum_{i=1}^K V_i^2 Q_i}} + \frac{V_k V_k^* R_k}{\sqrt{\sum_{i=1}^K V_i^{*2} S_i}} \right) \tilde{P}^{T-1} + \frac{\eta_w^2}{T} V_k^2 \tilde{P}^T \\
\frac{dR_k}{d\xi} &= \frac{\eta_w}{\sqrt{2\pi}} \left(\frac{V_k^2 R_k}{\sqrt{\sum_{i=1}^K V_i^2 Q_i}} + \frac{V_k V_k^* S_k}{\sqrt{\sum_{i=1}^K V_i^{*2} S_i}} \right) \tilde{P}^{T-1} \qquad (394) \\
\frac{dV_k}{d\xi} &= \frac{\eta_v}{\sqrt{2\pi}} \left(\frac{V_k Q_k}{\sqrt{\sum_{i=1}^K V_i^2 Q_i}} + \frac{V_k^* R_k}{\sqrt{\sum_{i=1}^K V_i^{*2} S_i}} \right) \tilde{P}^{T-1}, \qquad (395)
\end{aligned}$$

where η_w and η_v denote learning rates for $\mathbf{w}$ and $\mathbf{V}$, respectively.

We vary the initial R_k for different pre-trained skill level between 0 and maximum value that could be reach given T.

In Figure 9, we report the number of steps that reached the 99% of the minimum generalization error by numerical integration of ODEs. We used $K = 8, T = 8$ for the experiments and varied the number of junk skills among 8 skills. We used the same learning rate 1.0 for η_w and η_v.

J Details on LLM Experiments

We fine-tuned OLMo-2 1B base model [OLMo et al., 2024] on GSM8K [Cobbe et al., 2021b] math word problems across 3 random seeds for each setup. The GSM8K training set was split 90/10 into an RL partition and an SFT partition. In the SFT + RL conditions, the base model was first fine-tuned via supervised learning on the SFT partition (10% of training data, 747 examples) using the full chain-of-thought reasoning traces from GSM8K official dataset for 20, 60, 100, or 300 steps (lr = 2e-4, batch size 8), and the resulting checkpoint was then used to initialize GRPO training. In the RL-only baseline, GRPO training was applied directly to the base model with no SFF a priori. All GRPO runs trained for 1000 steps on the 90% RL partition (lr = 2e-5, batch size 8, 8 generations per

prompt, max completion length 768 tokens), using a binary accuracy reward over the GSM8K test format. All models used 4-bit NF4 quantization with LoRA adapters (rank 32) applied to all attention and MLP projection layers. Evaluation reward was computed every 50 steps on a held-out subset of 128 GSM8K test split examples.

Both SFT and RL training used the same plain text prompt format applied directly to the base model. Each prompt consisted of a brief instruction:

> "Solve the math problem. The last line must be exactly: Final answer: boxed{NUMBER}",

followed by the problem under a "Problem:" header and a "Solution:" prefix that cued the model to begin its response.

We used this format to work with a raw base model rather than an instruction-tuned one, providing just enough structure to elicit a parseable answer while keeping the format simple. We verified that before any finetuning, the model can often generate coherent answers with the correct format, but wrong answers.

For SFT, the full target sequence was appended. We used the GSM8K dataset's full answer trace, for example,

> Natalia sold 48/2 = «48/2=24»24 clips in May.
>
> Natalia sold 48+24 = «48+24=72»72 clips altogether in April and May.

followed by the formatted final answer. In this way, we expect the model learned to produce step-by-step solutions ending in the required boxed format.

For GRPO, only the prompt portion was provided and the model generated completions freely, with a binary accuracy reward assigned based on whether the final boxed answer matched the ground truth.

K Compute Resources

Only experiments that required a considerable compute resources was LLM experiment with OlMO 1B model and GSM8k dataset in Figure 3b. All of the LLM experiments were run on a single NVIDIA H100 GPU. A RL run with our setup took 8 hours, whereas SFT completed in under 1 hour.

L LLM Usage

We used LLM assistance during manuscript preparation to improve clarity and consistency. In particular, LLMs were used to help reorganize mathematical notation, double-check the presentation of proofs, and improve the manuscript for grammar, style, and readability.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract summarizes the contribution of our work: a simple analytically tractable model of multi-stage training.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a dedicated limitation section at the end of the Discussion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide a theory, its derivation steps and the simulation which directly tests the correctness of the theory.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the experimental details and hyperparameters used for generating the results. The code for the experiments will be publicly available upon acceptance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We provide the experimental details and hyperparameters used for generating the results. The code for the experiments will be publicly available upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We include the details of the hyperparameters for all the results we include in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Most of our experiment is for the validation of the theory that provides exact learning dynamics. For one experiment with LLM finetuning to compare the characteristics of learning curves, we used 3 seeds and notated mean and standard deviation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide additional section in appendix for compute resources providing the relevant details for the experiments that required considerable compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The goal of this work is to advance the theoretical understanding of the foundation model with a complex training pipeline. We do not foresee any direct societal impacts of the work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: N/A

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used OLMo-2 1B and GSM8K and cited them.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: NA

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We included section of LLM usage declaration in the appendix.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.