TO BIAS OR NOT TO BIAS: A PHASE TRANSITION IN LEARNING FROM SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world dataset curation is expensive. Pre-trained machine learning models provide a powerful means of generating synthetic data to augment limited real data for training downstream models. However, while synthetic-data augmentation has demonstrated empirical success, there is little theoretical understanding of when synthetic data actually helps. Here, we tackle fundamental questions: when can a model trained on synthetic data outperform its teacher, how much synthetic data is required, and how should the student be trained? We study these questions through a student-teacher framework in which both the teacher learns the distribution from the real data, and the student learns from the synthetic data generated by the teacher. When the hypothesis class for both teacher and student is the family of Gaussian distributions, we show if both teacher and student are maximum likelihood estimators (MLE), then the student can never outperform the teacher in terms of Kullback-Leibler (KL) divergence to the true distribution. Crucially, allowing the student to use an appropriately biased estimator changes this picture. We characterize a phase transition in the number of real and synthetic samples: the boundary nearly scales *quadratically* with the number of real samples; below which no student that is scalar multiple of MLE can outperform the teacher, and above the boundary a suitably biased student can achieve strictly lower KL divergence to the true distribution and consequently improved generalization. This phase transition provides exact practical guidance on how to train student models on synthetic data. A natural question is whether we can repeatedly use the student model as a teacher to train a new student model, and so on, to further boost the performance of the student model. We show that achieving an improvement at every stage requires a sample complexity that grows exponentially with the length of the chain. Moreover, we show that a two-stage mechanism can achieve the same asymptotic performance as arbitrarily long chains, revealing a fundamental *No Free Bootstrapping* phenomenon. Finally, we demonstrate the applicability of our theoretical framework in both discriminative and generative models, including linear discriminant analysis and linear flow matching.

1 INTRODUCTION

Recent advances in generative modeling have produced foundation models (Touvron et al., 2023; Liu et al., 2024a; Polyak et al., 2024) trained on internet-scale datasets. More data classically improves model performance (Kaplan et al., 2020), but curating new real datasets is expensive. Synthetic data generated by pre-trained generative models is therefore a promising way to augment real data when training downstream models. Recent study (Pew Research Center, 2026) showed that a non-trivial fraction of the common crawl dataset (Common Crawl Foundation, 2026) is estimated to be machine-generated. In practice using synthetic data is widespread – for example, knowledge distillation (Hinton et al., 2015) and self-distillation (Li et al., 2017; Furlanello et al., 2018; Zhang et al., 2021; Briesch et al., 2024) (see Liu et al. (2024b); Hu et al. (2025) for surveys) – where a student model is trained on the outputs of a teacher model, known as the *student-teacher setting*. Using naive training procedures in this setting can often lead to degradation of performance retraining on model output such as model collapse (Alemohammad et al., 2023; Shumailov et al., 2024a;b) and the related AI autophagy phenomenon (Xing et al., 2025). These results are largely empirical or asymptotic, and they diagnose collapse as a property of the student-teacher pipeline.

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

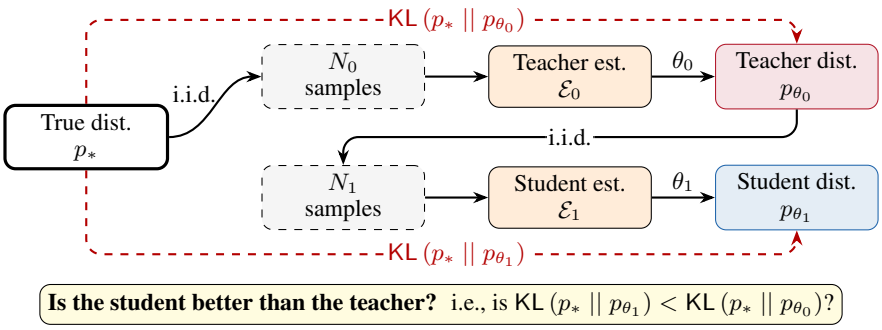


Figure 1: The student–teacher generative pipeline.

To overcome this degradation, a few lines of work (Burns et al., 2024; Li et al., 2025) use students that have more expressive power than their teacher model leading to the student outperforming their teacher, a phenomenon referred to as *weak-to-strong generalization*. A large algorithmic literature on weak-to-strong generalization (Burns et al., 2024; Jeon et al., 2025; Li et al., 2025; Shin et al., 2025; Somerstep et al., 2025; Uzunoglu et al., 2026; Jiang et al., 2026; Lang et al., 2026; Liu et al., 2026) refines *how* a student should consume teacher output for optimal performance. A parallel line of works in student-teacher learning theory which are predominantly discriminative and model-specific: ridge regression (Loureiro et al., 2021; Xue et al., 2025; Lecoiu et al., 2026), kernel regression (Xue et al., 2025), random feature models (Medvedev et al., 2025; Wu et al., 2026), single-layer and feed-forward networks in the statistical-physics tradition (Gardner & Derrida, 1989; Seung et al., 1992), two-layer networks (Oh et al., 2025), general convex function classes (Charikar et al., 2024), binary classification (Yao et al., 2025a;b), and reinforcement learning (Fu et al., 2024). These methods supply a training rule, but leave open the three quantitative questions a practitioner actually faces: how do the theory extend to generative settings?, how many synthetic samples are required?, and whether *any* gain is achievable at a given budget?

In this work, we formulate the problem of analyzing the effect of synthetic-data in generative setting. Mathematically, let p_* be the true data distribution supported in d-dimensional Euclidean space. To approximate the true distribution we consider distributions that are parameterized by the set Θ, where for each $\theta \in \Theta$, it is denoted as p_θ. We call the *Teacher model* as the distribution p_{θ_0} learned from the true data distribution using N_0 i.i.d. data points arranged as matrix $X_0 \in \mathbb{R}^{d\times N_0}$. Similarly, the *Student model* is the distribution, p_{θ_1}, learned from the *Teacher model* using N_1 i.i.d. data points arranged as matrix $X_1 \in \mathbb{R}^{d\times N_1}$ as depicted in Figure 1. The mapping between X_i and θ_i is called an estimator; formally, the ith estimator is a function $\mathcal{E}_i : \mathbb{R}^{d\times N_i} \to \Theta$. Here, we pose a fundamental question about the student-teacher setting:

> **Research Question.** Given finitely many samples, under what properties of the estimators is data generated by the student model assigns higher likelihood to real data than the teacher does , i.e. is p_{θ_1} closer to p_* than p_{θ_0}?

The posed question can be formalized in terms of the Kullback-Leibler (KL) divergence between the true distribution and the teacher and student models; mathematically,

$$\mathsf{KL}\left(p_* \,||\, p_{\theta_1}\right) \lesseqgtr \mathsf{KL}\left(p_* \,||\, p_{\theta_0}\right), \tag{1}$$

where $\mathsf{KL}\left(p \,||\, q\right) := \int p(x) \ln\left(p(x)/q(x)\right) dx$. Equivalently, we can compare the log-likelihood of the data generated: $\mathbb{E}_{\mathbf{x}\sim p_*}\left[\ln p_{\theta_1}(\mathbf{x})\right] \gtrless \mathbb{E}_{\mathbf{x}\sim p_*}\left[\ln p_{\theta_0}(\mathbf{x})\right]$. In this work, we characterize these conditions theoretically for models learned by maximum likelihood estimation (MLE), or variants of it, over a Gaussian hypothesis class. Our contributions are as follows.

1. **Necessary Conditions.** We prove an impossibility result: if the teacher is the MLE and the student receives fewer than quadratically many synthetic samples (relative to the real samples), no student that is a scalar multiple of the MLE can beat the teacher in the KL divergence sense.

2. **Sufficient Conditions.** Conversely, with quadratically more synthetic samples and an appropriate scaling, a biased student can outperform its teacher. Together, these conditions characterize the phase transition (Figure 2) and tell practitioners how to adjust the learning procedure.
3. **No Free Bootstrapping.** Since a single student can improve on its teacher, we next consider a chain of students, each trained on synthetic data from its predecessor. We show that successive students can keep improving, but the required sample size grows exponentially with the chain length. Moreover, a two-stage procedure achieves the same asymptotic performance as an arbitrarily long chain, so *bootstrapping yields no free gains*.
4. **Practical Implications.** We show that students can outperform their teachers in linear discriminant analysis and linear flow matching on real-world datasets, confirming the practical relevance.

Relation to prior work. Gaussian analyses of self-consuming models identify *shrinkage* of the sample-covariance eigenvalues as the failure mode of training on purely synthetic data (Alemohammad et al., 2023), and recover performance only by mixing real data back in (Alemohammad et al., 2024); related analyses cover multi-label classification (Wu & Sahai, 2025) and Gaussian mixtures (Somerstep et al., 2024; 2025; Wei et al., 2021). We show that deliberate *inflation* of the estimated covariance corrects this shrinkage, and pin down the exact inflation factor and the sample budget at which the student wins–without any real data. Analyses of iterative retraining (Bertrand et al., 2024; Ferbach et al., 2024; Kazdan et al., 2025; Lang et al., 2024; Amin et al., 2026) retain real data at every round and give stability, convergence, or asymptotic guarantees. In contrast, our chains use purely synthetic data, yield explicit sample-complexity thresholds, show that shorter chains dominate longer ones, and prove that a single teacher–student stage suffices; we are not aware of an analogous impossibility result for bootstrapped generative chains.

Outline. § 2 characterizes the phase transition, both in expectation and with high probability, and applies it to LDA and linear flow matching. § 3 analyzes chains of students, and § 4 concludes. The appendix contains all proofs, additional related work, and further experimental details.

2 Phase Transition in Student-Teacher Settings

We now establish necessary and sufficient conditions for the student to outperform its teacher, both in expectation and with high probability. We first state our assumptions and the class of estimators.

> **Assumption 1** (True distribution)**.** *The data are drawn i.i.d. from $\mathcal{N}(0, \Sigma_*)$, where $\Sigma_* \in \mathbb{S}^d_{++}$ is unknown and the zero mean is known.*

Gaussianity makes an exact characterization of the student–teacher setting possible. We take the mean to be zero and known and focus on covariance estimation; § B treats an unknown mean.
Next, both teacher and student use the following hypothesis class.

> **Hypothesis 1** (Gaussian model)**.** $p_\theta = \mathcal{N}(0, \Sigma)$ *on* $\mathbb{R}^d$ *with* $\theta = \Sigma \in \mathbb{S}^d_{++}$.

Gaussian universality results (Dudeja et al., 2024; Han, 2025a;b) suggest our conclusions may extend to broader distribution classes. Finally, we define the class of estimators our results cover.

> **Definition 1** ($(\overline{\beta}, \underline{\beta})$-MLE)**.** *For $0 < \underline{\beta} \le \overline{\beta}$, an estimator $\hat{\Sigma}$ computed from samples $X \in \mathbb{R}^{d\times N}$ is a $(\overline{\beta}, \underline{\beta})$-MLE if $\frac{\underline{\beta}}{N} XX^\top \preceq \hat{\Sigma} \preceq \frac{\overline{\beta}}{N} XX^\top$ almost surely.*

This class is broad: it contains every estimator that is sandwiched between two multiples of the sample second moment, including:

1. **MLE:** $\hat{\Sigma} = \frac{1}{N} XX^\top$, a $(1, 1)$-MLE.
2. **Minimax estimator (James & Stein, 1961):** $\hat{\Sigma} = LDL^\top$, where $XX^\top = LL^\top$ is the Cholesky factorization and $D_{ii} = (N + d + 1 - 2i)^{-1}$; for $N \ge d$, a $\left(\frac{N}{N-d+1}, \frac{N}{N+d-1}\right)$-MLE.

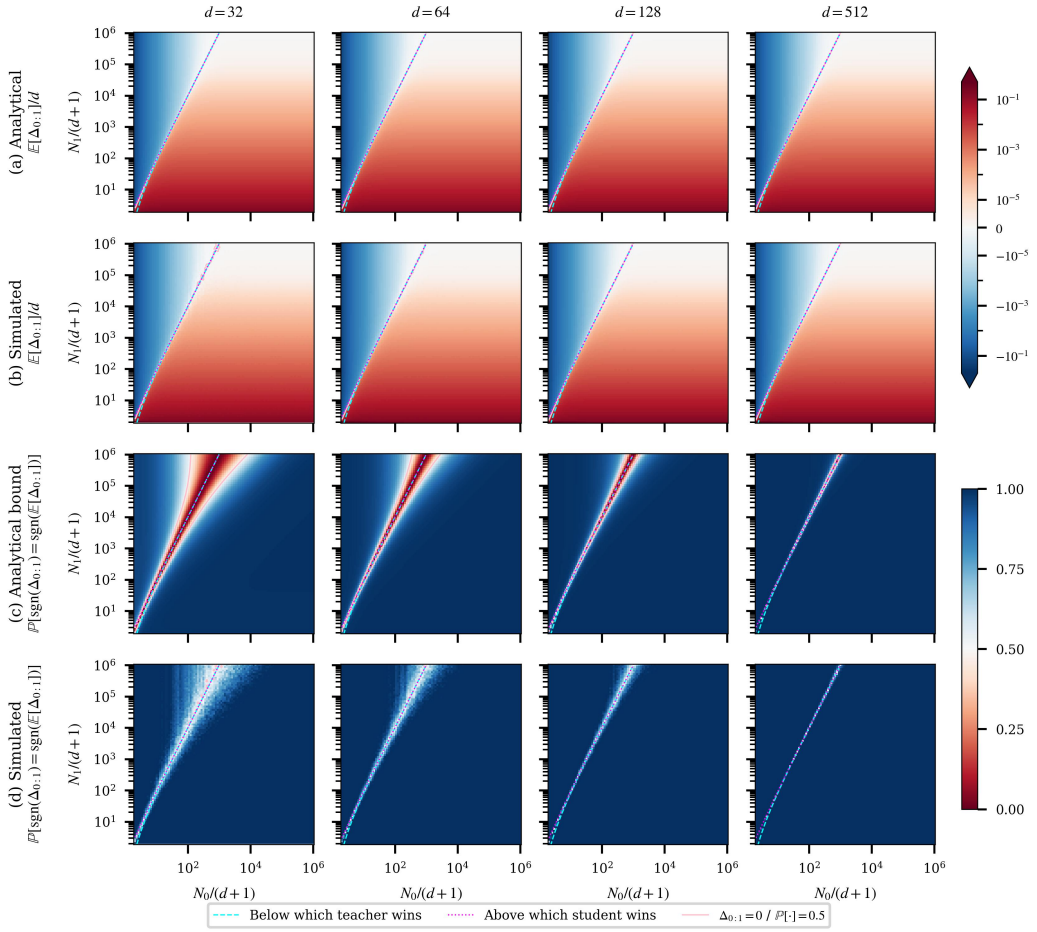


Figure 2: **A quadratic phase transition in learning with synthetic data.** A $(1,1)$-MLE teacher sees N_0 real samples; a $(\beta^*_{0:1}, \beta^*_{0:1})$-MLE student sees N_1 synthetic ones. Axes are in units of $d+1$. (a) Exact $\mathbb{E}[\Delta_{0:1}]/d$ (Theorem 1) and (b) its Monte Carlo estimate; blue: student wins, red: teacher wins. (c) Lower bound on $\mathbb{P}(\mathcal{A})$ (Theorem 2) and (d) its Monte Carlo estimate, where $\mathcal{A} = \{\operatorname{sgn}\Delta_{0:1} = \operatorname{sgn}\mathbb{E}[\Delta_{0:1}]\}$; blue: a single run matches the expected winner with high probability. Monte Carlo uses 100 trials. The pink line marks the transition ($\mathbb{E}[\Delta_{0:1}] = 0$ in (a,b); $\mathbb{P}(\mathcal{A}) = 0.5$ in (c,d)) and lies between the conditions of Corollary 1: the teacher wins below the cyan dashed curve, the student above the magenta dotted one.

3. **Haff estimator (Haff, 1980):** for $t : \mathbb{R} \to [0, T]$, $C \in \mathbb{S}^d_{++}$, and $u = \langle (XX^\top)^{-1}, C\rangle$, $\hat{\Sigma} = \frac{1}{N}\left(XX^\top + \frac{t(u)}{u}C\right)$, a $(1+T, 1)$-MLE.

The class does, however, exclude estimators designed for the high-dimensional ($d \geq N$) regime (Ledoit & Wolf, 2003; 2004; 2012; Ding & Yang, 2018; Ding, 2019; Ding et al., 2023; 2024).

2.1 Different phases in student-teacher setting

We now characterize $\mathbb{E}\left[\mathsf{KL}\left(p_* \,||\, p_{\theta_1}\right) - \mathsf{KL}\left(p_* \,||\, p_{\theta_0}\right)\right]$, where the expectation is over both the real samples and the synthetic samples. Our results involve transcendental functions, which we introduce first. Let $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$ be the gamma function, $\psi(z) = \frac{d}{dz}\ln\Gamma(z)$ the digamma function, and $\psi^{(n)}$ its n-th derivative. We define the d-digamma function $\psi_d(x) = \frac{1}{d}\sum_{j=1}^{d}\psi\left(\frac{x+1-j}{2}\right)$ and its derivatives $\psi_d^{(n)}$. We also define $\phi(x) := \ln x + 1/x - 1$, which is convex on $(0,\infty)$ with

unique minimum $\phi(1) = 0$. Hence for $y > 0$ the equation $\phi(x) = y$ has exactly two roots, $\phi_-^{-1}(y) < 1 < \phi_+^{-1}(y)$, and $\phi_-^{-1}(0) = \phi_+^{-1}(0) = 1$. The proof of the following theorem is in §D.

Theorem 1. *Under the assumption 1, and hypothesis 1, suppose $N_0, N_1 \geq d + 2$. Let the* teacher $\mathcal{E}_0$ *be a $(1, 1)$-MLE, and the* student $\mathcal{E}_1$ *be a $(\overline{\beta}, \underline{\beta})$-MLE. Define the KL gap*

$$\Delta_{0:1} := \mathsf{KL}\left(p_* \,||\, p_{\theta_1}\right) - \mathsf{KL}\left(p_* \,||\, p_{\theta_0}\right), \tag{2}$$

phase-transition function $h_d : \mathbb{R}_+ \to \mathbb{R}_+$ that satisfies the fixed point equation

$$\psi_d(h_d(M)) + \ln\left(\frac{2}{h_d(M) - d - 1}\right) = \phi\left(\frac{M - d - 1}{M}\right), \qquad \forall M \geq d + 2 \tag{3}$$

and the analytic proxy of $\Delta_{0:1}$:

$$\tilde{\Delta}_{0:1}(\beta) := \frac{d}{2}\left[\phi\left(\frac{\beta}{\beta^*_{0:1}}\right) + \psi_d(N_1) + \ln\left(\frac{2}{N_1 - d - 1}\right) - \phi\left(\frac{N_0 - d - 1}{N_0}\right)\right], \tag{4}$$

where $\beta^*_{0:1} := N_0 N_1 / ((N_0 - d - 1)(N_1 - d - 1))$. Then the following statements hold true:

1. ***Necessary sample complexity.*** *If synthetic data size is fewer than $h_d(N_0)$; i.e., $N_1 < h_d(N_0)$, then $\nexists \mathcal{E}_1 \in \{(\beta, \beta)\text{-MLE} : \forall \beta > 0\}$ s.t. $\mathbb{E}[\Delta_{0:1}] < 0$.*
2. ***Unbiased student performs more poorly than teacher.*** *If $\underline{\beta} = \overline{\beta} = 1$, then for any (d, N_0, N_1) it holds that $\mathbb{E}[\Delta_{0:1}] = \tilde{\Delta}_{0:1}(1) > 0$.*
3. ***Performance bounds for biased student.*** *For any pair $(\overline{\beta}, \underline{\beta})$ we have the inequality:*

$$-\frac{d}{2}\ln\left(\frac{\overline{\beta}}{\underline{\beta}}\right) + \tilde{\Delta}_{0:1}(\overline{\beta}) \leq \mathbb{E}[\Delta_{0:1}] \leq \tilde{\Delta}_{0:1}(\underline{\beta}) + \frac{d}{2}\ln\left(\frac{\overline{\beta}}{\underline{\beta}}\right). \tag{5}$$

Remarks. Theorem 1 characterizes the expected KL gap exactly in terms of N_0, N_1, and the student's bias $(\overline{\beta}, \underline{\beta})$. The phase-transition function $h_d(N_0)$ gives the minimum synthetic sample size a (β, β)-MLE student needs to have any chance of beating its teacher (Figure 2(a)). An unbiased student always loses in expectation; Shumailov et al. (2024b, Theorem. 3.1) give an analogous result in Wasserstein distance, so bias is *necessary*. For a biased student, Equation (5) lets us identify regimes where the upper bound is negative, i.e., the student provably wins.
Using standard digamma inequalities (Chen, 2005), we next make these regimes explicit with non-asymptotic rates (proof in § E).

Corollary 1 (Non-asymptotic KL gap)**.** *In the setting of Theorem 1, let $g(x) := \ln x + x - 1$. Suppose the student's scaling is within a factor $\zeta \geq 1$ of the optimum: $\overline{\beta} = \zeta\beta^*_{0:1}$ and $\underline{\beta} = \beta^*_{0:1}/\zeta$. Then*

$$\begin{aligned} \frac{4}{d}\mathbb{E}[\Delta_{0:1}] &\geq \underbrace{\frac{d+1}{N_1 - 1}}_{\text{student error}} - \underbrace{\frac{(d+1)^2}{(N_0 - d - 1)^2}}_{\text{teacher error}} + \underbrace{2g(1/\zeta)}_{\text{mis-scaling}}, \\ \frac{4}{d}\mathbb{E}[\Delta_{0:1}] &\leq \frac{d+1}{N_1 - d - 1} - \frac{(d+1)^2}{N_0(N_0 - d - 1)} + 2g(\zeta). \end{aligned} \tag{6}$$

In particular, with the optimal scaling ($\zeta = 1$) the mis-scaling term vanishes, and

$$\mathbb{E}[\Delta_{0:1}] \begin{cases} > 0 \text{ (teacher wins)} & \text{if } N_1 < L := 1 + \frac{(N_0 - d - 1)^2}{d+1}, \\ = 0 & \text{if } N_1 = h_d(N_0) \in [L, U], \\ < 0 \text{ (student wins)} & \text{if } N_1 > U := N_0 + \frac{(N_0 - d - 1)^2}{d+1}. \end{cases}$$

Remarks. With optimal scaling ($\zeta = 1$), Corollary 1 pins the phase transition at $h_d(N_0) = \Theta(N_0^2/d)$: to beat its teacher, the student needs quadratically more samples, ignoring the effect of dimension. To our knowledge, these are the first synthetic-data scaling laws for biased student estimators, addressing an open question on how much synthetic data is needed (Liu et al., 2024b). Figure 2(a,b) confirms these regimes and shows that the boundary is tight. Since these guarantees hold only in expectation, we next derive high-probability bounds.

2.2 PAC FOR STUDENT-TEACHER SETTING

We now give probably approximately correct (PAC) guarantees for $\Delta_{0:1}$. To determine whether the student wins or loses in a single run, it suffices to show that the random gap $\Delta_{0:1}$ has the same sign as $\mathbb{E}[\Delta_{0:1}]$, which follows directly from Cantelli's inequality (Cantelli, 1928).

Theorem 2 (PAC guarantees). *In the setting of Theorem 1 with $\underline{\beta} = \overline{\beta} = \beta^*_{0:1}$, let $\mathcal{A} := \{\operatorname{sgn}\Delta_{0:1} = \operatorname{sgn}\mathbb{E}[\Delta_{0:1}]\}$ be the event that a single run agrees with the expectation. Then*

$$\mathbb{P}(\mathcal{A}^c) \le \frac{\mathsf{Var}(\Delta_{0:1})}{\mathbb{E}[\Delta_{0:1}]^2 + \mathsf{Var}(\Delta_{0:1})}, \tag{7}$$

where $\mathsf{Var}(\Delta_{0:1})$ is given in closed form in (60). Moreover, with L, U as in Corollary 1 and any $\varepsilon \in (0,1)$:

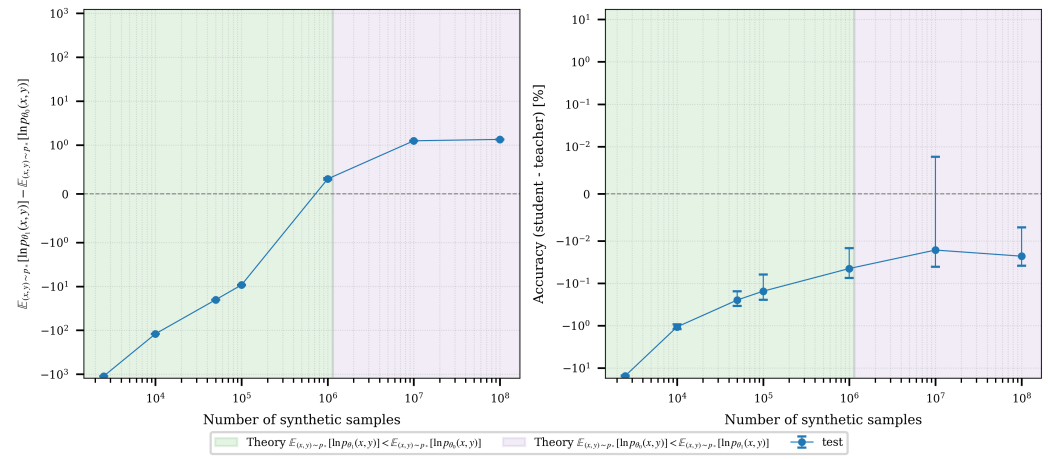
1. ***Student wins w.h.p.*** *If $2d+3 \le N_0 \le (d+1)\big(1+\frac{d^\varepsilon}{10\sqrt{6}}\big)$ and $N_1 > \max\{2d+3,\ (2+\sqrt{2})\,U - (1+\sqrt{2})(d+1)\}$, then $\mathbb{P}(\mathcal{A}^c) < \frac{169}{160}\,d^{-2-2\varepsilon}$.*
2. ***Teacher wins w.h.p.*** *If $N_0 > 2d+3$ and $2d+3 < N_1 < 1 + (1-2\sqrt{10}\,d^{-\varepsilon})(L-1)$, then $\mathbb{P}(\mathcal{A}^c) < d^{-2+2\varepsilon}\big[1+\frac{8}{15}\big(\frac{N_1}{N_0-d-1}\big)^2\big]$, which vanishes whenever $N_1 = o\big(d^{1-\varepsilon}N_0\big)$.*

Remarks. Items 1 and 2 of Theorem 2 cover the regimes where the student and the teacher, respectively, win in expectation. In both, the probability that a single run disagrees with the expectation vanishes polynomially in d: as $\mathcal{O}(d^{-2-2\varepsilon})$ in item 1, and as $\mathcal{O}(d^{-2+2\varepsilon})$ in item 2 when $N_1 = \mathcal{O}(N_0 - d - 1)$. Hence the guarantees in expectation hold w.h.p. for individual draws. Figure 2(c,d) compares the bound (7) with Monte Carlo estimates (see also Figure 5).

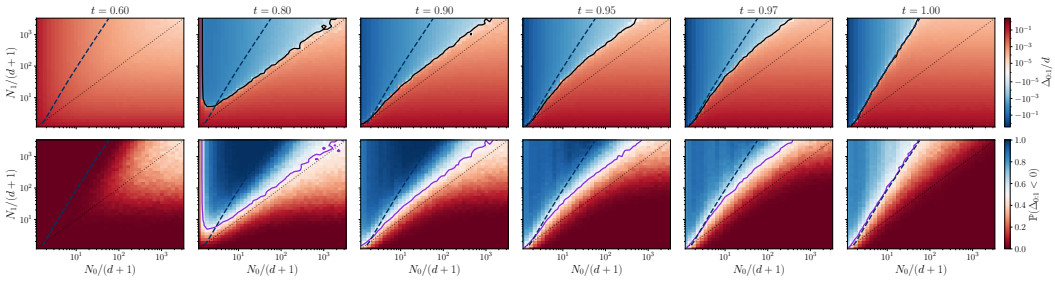
Implications for machine learning models So far, we have established that biasing the student appropriately and under sufficient synthetic data can overcome the teacher. As a natural extension, we will consider two machine learning models (i) linear discriminant analysis (LDA) (Fisher, 1936; Bishop, 2006), and (ii) linear flow matching (FM) (Lipman et al., 2023) we train their teacher counterparts using standard training procedure, and scale up the students via factor $\beta^*_{0:1}$ measure the performance difference. We perform this experiment on ResNet-50 features of CIFAR-10 dataset for LDA, and use Gaussian data for the linear FM. In Figure 3a shows: the student's log-likelihood crosses the teacher's near the budget predicted by Theorem 5, while its test accuracy stays within $\pm 0.05\%$ (or ± 5 misclassifications out of 10000). In Figure 3b sweeps the stopping time on all three schedules with an *unscaled* student using linear optimal transport schedule (more in Figure 6) towards ends the heatmap resembles Figure 2.

So far we have shown that, given enough synthetic data, an appropriately biased student outperforms its teacher. We now ask whether this carries over to two models used in practice (details in § B).

Linear discriminant analysis (LDA) (Fisher, 1936; Bishop, 2006) is a class-conditional Gaussian model with a shared covariance, so our analysis extends to it directly: the student should inflate the pooled covariance by a factor β^*_{LDA}, and it overtakes the teacher once the synthetic budget exceeds a threshold that, as before, grows quadratically in N_0 (Theorem 5, § B). *Setup.* We use 2048-dimensional ResNet-50 (He et al., 2016) features of CIFAR-10 (Krizhevsky, 2009) (10 classes). The teacher is the maximum-likelihood LDA fitted to $N_0 = 5 \times 10^4$ training points. The student receives N_1 labelled samples drawn from the teacher's model, with class proportions matched to the teacher's (19), and fits an LDA whose pooled covariance is scaled by β^*_{LDA}. We compare the two by mean log-likelihood and accuracy on the 10^4 held-out test points, averaged over 20 independent draws of the synthetic data. Figure 3a shows that the student's log-likelihood crosses the teacher's



(a) LDA on CIFAR-10 ResNet-50 features.



(b) Linear flow matching (linear OT path), $d = 5$.

Figure 3: **Student versus teacher in machine learning models.** a Student minus teacher test log-likelihood (left) and test accuracy (right) versus N_1; green/purple: region where Theorem 5 predicts the student loses/wins. b $\mathbb{E}[\Delta_{0:1}]/d$ (top) and $\mathbb{P}(\Delta_{0:1} < 0)$ (bottom) with both flows stopped at time t; dashed: boundary of Theorem 1.

close to the predicted budget, while test accuracy changes by at most $\pm 0.05\%$ (± 5 of 10^4 test points). Scaling therefore improves the generative fit without hurting the classifier.

Flow matching (FM) (Lipman et al., 2023) has no explicit covariance to rescale. Instead, we rescale the data: the teacher's flow is integrated to $t = 1$, and its samples are multiplied by $\sqrt{\beta^*_{0:1}}$ before the student is trained on them. For Gaussian data the flow-matching minimizer (27) depends on the data only through its covariance, so the fitted flow reproduces the empirical covariance and the rescaled student is a $(\beta^*_{0:1}, \beta^*_{0:1})$-MLE (Corollary 2). *Setup.* The target is $\mathcal{N}(0, 5I_d)$ with $d = 5$. The teacher's velocity field is the closed-form minimizer (27) fitted to N_0 real samples; the student's is fitted to N_1 rescaled samples from the teacher. Running either flow from $\mathcal{N}(0, I_d)$ to time t yields the Gaussian marginal $\mathcal{N}\big(0, (1-t)^2 I_d + t^2 \hat{\Sigma}\big)$ along the linear OT path, so the KL divergence to p_* is computed exactly at every t. We sweep a 40×40 logarithmic grid of (N_0, N_1) over $[7, 2 \times 10^4]$ with 1000 trials per cell. Figure 3b shows $\mathbb{E}[\Delta_{0:1}]/d$ (top) and $\mathbb{P}(\Delta_{0:1} < 0)$ (bottom) at increasing t: as $t \to 1$ the phase boundary converges to that of Theorem 1 (dashed), recovering Figure 2.

3 No Free Bootstrapping

In § 2, we established necessary and sufficient conditions for student to outperform its teacher. An immediate consequence of this phenomenon is to use the student to train another new student, and so on, with a hypothesis that each subsequent new student will improve upon its predecessor. Here, we show that such progressive improvement is possible, with certain trade-offs and limitations.

First, we provide condition under which progressive improvement is possible.

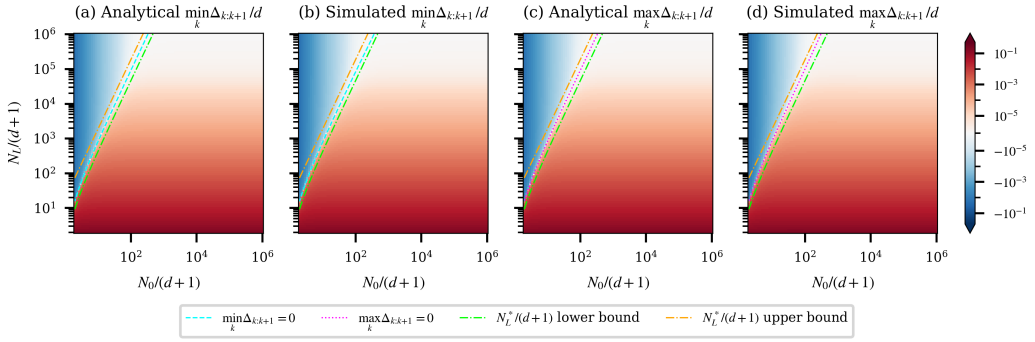


Figure 4: **Progressive improvements among the successive students.** A $(1,1)$-MLE teacher sees N_0 real samples, and each of the $L-1$ students sees $N_1 = \cdots = N_{L-1}$ synthetic samples from its predecessor, with the scalings $\beta^{(k)}$ of Theorem 3 ($d = 128$, $L = 10$, $\rho = 10^{-9}$); axes are in units of $d+1$. (a) Exact $\min_k \mathbb{E}[\Delta_{k:k+1}]/d$, the best single-stage gap, and (b) its Monte Carlo estimate (100 trials); blue: *all* student improves on its predecessor. (c) Exact $\max_k \mathbb{E}[\Delta_{k:k+1}]/d$, the worst single-stage gap, and (d) its Monte Carlo estimate; red: *every* student worsens compared to its predecessor. The cyan dashed and magenta dotted curves are the zero contours of the min and max gaps. The latter, the boundary for improvement at every stage, lies between the green and orange dash-dotted curves, the lower and upper bounds on the chaining threshold $N_L^*(\rho, N_0)$ from Theorem 3. Both boundaries have slope 2 on log-log axes, so improvement at every stage costs quadratically many synthetic samples in N_0.

Theorem 3 (Progressive improvements). *Under Assumption 1 and Hypothesis 1, let the teacher $\mathcal{E}_0$ be a $(1,1)$-MLE on $N_0 \geq d+2$ samples from p_*, and for $k \in [L-1]$ let student $\mathcal{E}_k$ be a (β_k, β_k)-MLE on N_k samples from $p_{\theta_{k-1}}$. Write $\Delta_{k:k'} := \mathsf{KL}\left(p_* \,||\, p_{\theta_{k'}}\right) - \mathsf{KL}\left(p_* \,||\, p_{\theta_k}\right)$ and, for $\rho > 0$,*

$$c := \phi\left(1 - \frac{d+1}{N_0}\right), \qquad S_m := \frac{(1+\rho)^{m-1} - 1}{\rho}, \qquad \xi_k := \phi_-^{-1}\left(\frac{S_{L-k}}{S_L}\, c\right). \tag{8}$$

If $N_1 = \cdots = N_{L-1} \geq N_L^* := h_d\left(\frac{d+1}{1-\xi_{L-2}}\right)$ and $\beta_k = \frac{\xi_k}{\xi_{k-1}} \frac{N_k}{N_k - d - 1}$, then every student improves on its predecessor:

$$\mathbb{E}[\Delta_{k:k+1}] \leq -\frac{d}{2}\,\rho\,\phi(\xi_{k+1}) \leq 0, \qquad k \in \{0\} \cup [L-2], \tag{9}$$

with $\mathbb{E}[\Delta_{k:k+1}] \to 0^-$ as $\rho \to 0^+$. Moreover,

$$\frac{S_L}{2}\,\frac{N_0(N_0-d-1)}{d+1} \;\leq\; N_L^* \;\leq\; 2S_L\,\frac{N_0^2}{d+1}. \tag{10}$$

Remarks. By Eq (9), the expected KL to p_* decreases along the chain: every student improves on its predecessor. The price is the sample size: by the sandwich bound, each student needs $N_L^* = \Theta\big((1+\rho)^{L-2} N_0^2/d\big)$ samples, which grows exponentially in the chain length. Figure 4 confirms this with Monte Carlo simulations. This raises a natural question: can a long chain ever beat the best *single* student? The next theorem shows it cannot.

Theorem 4 (No free bootstrapping). *In the setting of Theorem 3, let every student $\mathcal{E}_k$, $k \in [L-1]$, be a $(\overline{\beta}, \underline{\beta})$-MLE. For $m \in [L-1]$ let $P_m := \prod_{l=0}^{m}\left(1 - \frac{d+1}{N_l}\right)^{-1}$ and*

$$\tilde{\Delta}_{0:m}(\beta) := \frac{d}{2}\left[m \ln \beta + \sum_{l=1}^{m}\left(\ln\left(\frac{2}{N_l}\right) + \psi_d(N_l)\right) + \beta^{-m} P_m - P_0\right]. \tag{11}$$

Then,

$$\tilde{\Delta}_{0:L-1}(\overline{\beta}) - (L-1)\frac{d}{2}\ln\left(\frac{\overline{\beta}}{\underline{\beta}}\right) \;\le\; \mathbb{E}[\Delta_{0:L-1}] \;\le\; \tilde{\Delta}_{0:L-1}(\underline{\beta}) + (L-1)\frac{d}{2}\ln\left(\frac{\overline{\beta}}{\underline{\beta}}\right), \tag{12}$$

with equality $\mathbb{E}[\Delta_{0:L-1}] = \tilde{\Delta}_{0:L-1}(\beta)$ *when* $\overline{\beta} = \underline{\beta} = \beta$. *Moreover, each additional round strictly increases the optimal gap: for* $2 \le m \le L-1$,

$$\min_{\beta>0} \tilde{\Delta}_{0:m}(\beta) - \min_{\beta>0} \tilde{\Delta}_{0:m-1}(\beta) = \frac{d}{2}\left(\psi_d(N_m) + \ln\frac{2}{N_m - d - 1}\right) > 0. \tag{13}$$

Remarks. For simplicity, suppose we have the every stage is simple β-scaled MLE. Eq (13) shows that the gap between KL divergence gain of last student with the teacher of L-length and $(L-1)$-length chain is strictly positive. Therefore, we conclude longer chains only worsen the performance of the students, albeight, having the possibility of improving at each stage (see Theorem 3).

4 Conclusions

In this work, we asked when a student trained on synthetic data can outperform its teacher. For Gaussian models learned by (scaled) maximum likelihood, we gave matching necessary and sufficient conditions: the student must be appropriately biased, and it must receive quadratically more samples than its teacher, on the order of N_0^2/d. These conditions hold both in expectation and with high probability, and they carry over to LDA and flow matching. Finally, we showed that chaining students can keep improving performance, but only with a sample budget that grows exponentially in the chain length, and that the best achievable performance strictly degrades with every additional round. Bootstrapping from synthetic data therefore comes at a price: there is *no free bootstrapping*.

AI use statement

In this work, we used generative AI tools as assistants in three capacities: (i) *software development*, i.e., writing and debugging the code used for our experiments and for producing figures; (ii) *numerical simulations*, i.e., implementing and running the Monte Carlo simulations that we use to validate our theoretical results; and (iii) *writing*, i.e., improving the grammar, clarity, and presentation of the manuscript. We have not used generative AI tools for formulating the research questions or the problem formulation. We have reviewed all AI-assisted work. AI-assisted code was read and tested by the authors. We checked every simulation against the closed-form expressions derived in the paper (e.g., Figure 2(a) versus (b)). Every AI-suggested edit to the text was reviewed by the authors, who verified that it did not change the technical content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-Consuming Generative Models Go MAD. In *The Twelfth International Conference on Learning Representations*, October 2023. 1, 3, 15

Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse, and Richard Baraniuk. Self-Improving Diffusion Models with Synthetic Data, August 2024. 3, 15

Kareem Amin, Alex Bie, Weiwei Kong, Umar Syed, and Sergei Vassilvitskii. Learning from synthetic data: Limitations of erm. In Matus Telgarsky and Jonathan Ullman (eds.), *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–23. PMLR, 23–26 Feb 2026. URL https://proceedings.mlr.press/v313/amin26a.html. 3, 15

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JORAfH2xFd. 3, 15

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006. ISBN 978-0-387-31073-2. 6, 16

Christian Bongiorno and Lamia Lamrani. Quantifying the information lost in optimal covariance matrix cleaning. *Physica A: Statistical Mechanics and its Applications*, 657:130225, 2025. 40

Martin Briesch, Dominik Sobania, and Franz Rothlauf. Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop. *arXiv preprint arXiv:2311.16822*, October 2024. 1, 15

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4971–5012. PMLR, 2024. URL https://proceedings.mlr.press/v235/burns24b.html. 2, 15

Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020. 16

Francesco Paolo Cantelli. Sui confini della probabilità. In *Atti del Congresso Internazionale dei Matematici (Bologna, 1928)*, volume 6, pp. 47–59. Nicola Zanichelli, 1928. 6

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the Gain in Weak-to-Strong Generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024. 2, 15

Chaoping Chen. Inequalities for the polygamma functions with application. *Scientia Magna*, 1(2): 91–95, 2005. 5, 34

Common Crawl Foundation. Common crawl, 2026. URL https://commoncrawl.org/. Accessed: 2026-09-21. 1, 15

Xiucai Ding. Singular vector distribution of sample covariance matrices. *Advances in applied probability*, 51(1):236–267, 2019. 4

Xiucai Ding and Fan Yang. A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *The Annals of Applied Probability*, 28(3):1679–1738, 2018. 4

Xiucai Ding, Jiahui Xie, Long Yu, and Wang Zhou. Extreme eigenvalues of sample covariance matrices under generalized elliptical models with applications. *arXiv preprint arXiv:2303.03532*, 2023. 4

Xiucai Ding, Yun Li, and Fan Yang. Eigenvector distributions and optimal shrinkage estimators for large covariance and precision matrices. *arXiv preprint arXiv:2404.14751*, 2024. 4

Rishabh Dudeja, Subhabrata Sen, and Yue M Lu. Spectral universality in regularized linear regression with nearly deterministic sensing matrices. *IEEE Transactions on Information Theory*, 70 (11):7923–7951, 2024. 3

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385. 3, 15

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x. 6, 16

Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao. Towards Theoretical Understandings of Self-Consuming Generative Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14228–14255. PMLR, 2024. 2, 15

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pp. 1607–1616. PMLR, 2018. 1, 15

Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983–1994, 1989. 2, 15

L. R. Haff. Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix. *The Annals of Statistics*, 8(3):586–597, May 1980. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345010. 4

Qiyang Han. Long-time dynamics and universality of nonconvex gradient descent. *arXiv preprint arXiv:2509.11426*, 2025a. 3

Qiyang Han. Entrywise dynamics and universality of general first order methods. *The Annals of Statistics*, 53(4):1783–1807, 2025b. 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 6

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 15

Zixuan Hu, Yongxian Wei, Guozheng Ma, Jiahe Wang, Tianci Luo, Sinan Du, Runxi Cheng, Qianpu Sun, Haotian Luo, Xiaoyu Ma, Chun Yuan, Li Shen, and Dacheng Tao. Synthetic Data Beyond Generative Models: A Comprehensive Survey of How, Why, and Where. Preprint, 2025. Preprint. 1, 15

William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 361–379. University of California Press, 1961. 3

Myeongho Jeon, Jan Sobotka, Suhwan Choi, and Maria Brbić. Weak-to-Strong Generalization under Distribution Shifts, October 2025. 2, 15

Houcheng Jiang, Junfeng Fang, Jiaxin Wu, Tianyu Zhang, Chen Gao, Xiang Wang, Xiangnan He, and Yang Deng. Contrastive Weak-to-strong Generalization, July 2026. 2, 15

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. 1

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L. Donoho, and Sanmi Koyejo. Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World, March 2025. 3, 15

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. 6

Hao Lang, Fei Huang, and Yongbin Li. Selective weak-to-strong generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37556–37564. AAAI Press, 2026. ISBN 978-1-57735-906-7. URL https://doi.org/10.1609/aaai.v40i44.41089. 2, 15

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385. 3, 15

Radu Lecoiu, Debarghya Mukherjee, and Pragya Sur. Self-distillation is optimal among spectral shrinkage estimators in spiked covariance models. *arXiv preprint arXiv:2605.17778*, 2026. 2, 15

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003. 4

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004. 4

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012. 4

Jinhao Li, Sarah Monazam Erfani, Lei Feng, James Bailey, and Feng Liu. Exploring Weak-to-Strong Generalization for CLIP-based Classification. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=FwkYeLovHk. 2, 15

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1928–1936. IEEE, 2017. 1, 15

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t. 6, 7, 17

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*, 2024a. 1

Chenruo Liu, Yijun Dong, and Qi Lei. Does weak-to-strong generalization happen under spurious correlations? In *International Conference on Learning Representations*, volume 2026, pp. 32657–32687, 2026. 2, 15

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*, 2024b. URL https://openreview.net/forum?id=OJaWBhh61C. 1, 6, 15

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18137–18151. Curran Associates, Inc., 2021. 2, 15

Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-to-Strong Generalization Even in Random Feature Networks, Provably. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43519–43556. PMLR, 2025. URL https://proceedings.mlr.press/v267/medvedev25a.html. 2, 15

Junsoo Oh, Jerry Song, and Chulhee Yun. From Linear to Nonlinear: Provable Weak-to-Strong Generalization through Feature Learning. In *Advances in Neural Information Processing Systems*, June 2025. 2, 15

Pew Research Center. How much of the internet is written with AI?, August 2026. URL https://www.pewresearch.org/internet/2026/08/20/how-much-of-the-internet-is-written-with-ai/. 1, 15

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1

S. James Press. *Applied multivariate analysis*. Dover Books on Mathematics. Dover Publications, Mineola, NY, 2 edition, June 2005. 36

Christoph Schuhmann. LAION-Aesthetics. LAION Blog, August 2022. URL https://laion.ai/blog/laion-aesthetics/. Accessed: 2026-09-21. 15

Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992. 2, 15

Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-centric lens. In *International Conference on Learning Representations*, 2025. 2, 15

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The Curse of Recursion: Training on Generated Data Makes Models Forget, April 2024a. 1, 15

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024b. 1, 5, 15

Seamus Somerstep, Felipe Maia Polo, Moulinath Banerjee, Yaacov Ritov, Mikhail Yurochkin, and Yuekai Sun. A statistical framework for weak-to-strong generalization. In *ICML 2024 Next Generation of AI Safety Workshop*, July 2024. 3, 15

Seamus Somerstep, Ya'acov Ritov, Mikhail Yurochkin, Subha Maity, and Yuekai Sun. Limitations of refinement methods for weak to strong generalization. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=OKvSnV5Ar7. 2, 3, 15

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. 1

Arda Uzunoglu, Alvin Zhang, and Daniel Khashabi. Trust Functions: Near-Lossless Weak-to-Strong Generalization by Learning When to Trust the Weak Teacher, May 2026. 2, 15

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data. In *International Conference on Learning Representations*, October 2021. 3, 15

David Wu and Anant Sahai. Provable weak-to-strong generalization via benign overfitting. In *International Conference on Learning Representations*, volume 2025, pp. 88280–88316, 2025. 3, 15

Diyuan Wu, Lehan Chen, Theodor Misiakiewicz, and Marco Mondelli. Improved Scaling Laws via Weak-to-Strong Generalization in Random Feature Ridge Regression, June 2026. 2, 15

Xiaodan Xing, Fadong Shi, Jiahao Huang, Yinzhe Wu, Yang Nan, Sheng Zhang, Yingying Fang, Michael Roberts, Carola-Bibiane Schönlieb, Javier Del Ser, and Guang Yang. On the caveats of AI autophagy. *Nature Machine Intelligence*, 7(2):172–180, February 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-00984-1. 1, 15

Yihao Xue, Jiping Li, and Baharan Mirzasoleiman. Representations shape weak-to-strong generalization: Theoretical insights and empirical predictions. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70120–70159. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/xue25d.html. 2, 15

Wei Yao, Gengze Xu, Huayi Tang, Wenkai Yang, Donglin Di, Ziqiao Wang, and Yong Liu. On weak-to-strong generalization and f-divergence. *arXiv preprint arXiv:2506.03109*, 2025a. 2, 15

Wei Yao, Wenkai Yang, Ziqiao Wang, Yankai Lin, and Yong Liu. Revisiting Weak-to-Strong Generalization in Theory and Practice: Reverse KL vs. Forward KL, May 2025b. 2, 15

Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2021. 1, 15

A Related works

Synthetic data in practice. Large corpora are already partly synthetic: a non-trivial fraction of Common Crawl (Common Crawl Foundation, 2026) is estimated to be machine-generated (Pew Research Center, 2026), and the same concern applies to image corpora such as LAION-Aesthetics (Schuhmann, 2022). Deliberate reuse of model output is equally widespread, from knowledge distillation (Hinton et al., 2015) and self-distillation (Li et al., 2017; Furlanello et al., 2018; Zhang et al., 2021) to self-training of language models (Briesch et al., 2024); see Liu et al. (2024b) and Hu et al. (2025) for surveys. A large algorithmic literature on weak-to-strong generalization refines *how* a student should consume teacher output (Burns et al., 2024; Jeon et al., 2025; Li et al., 2025; Shin et al., 2025; Somerstep et al., 2025; Uzunoglu et al., 2026; Jiang et al., 2026; Lang et al., 2026; Liu et al., 2026). These methods supply a training rule, but leave open the two quantitative questions a practitioner actually faces: how many synthetic samples are required, and whether *any* gain is achievable at a given budget. Our phase transition answers both with an explicit threshold rather than a heuristic.

When synthetic data hurts. A parallel line of work documents degradation under repeated retraining on model output – model collapse (Alemohammad et al., 2023; Shumailov et al., 2024a;b) and the related AI autophagy phenomenon (Xing et al., 2025). These results are largely empirical or asymptotic, and they diagnose collapse as a property of the generative pipeline. Theorem 1 recasts it as a statement about the *estimator*: for a Gaussian hypothesis class the maximum likelihood student is strictly worse than its teacher for every (d, N_0, N_1), with an exact gap $\tilde{\Delta}_{0:1}(1) > 0$, and the obstruction is precisely the unbiasedness that Cramér–Rao rewards. This reframing is what makes the failure *reversible*: the same analysis shows that a suitably biased student escapes it, so collapse is not intrinsic to training on synthetic data but to training on it naively.

Theory of student–teacher learning. Existing analyses are predominantly discriminative and model-specific: ridge regression (Loureiro et al., 2021; Xue et al., 2025; Lecoiu et al., 2026), kernel regression (Xue et al., 2025), random feature models (Medvedev et al., 2025; Wu et al., 2026), single-layer and feed-forward networks in the statistical-physics tradition (Gardner & Derrida, 1989; Seung et al., 1992), two-layer networks (Oh et al., 2025), general convex function classes (Charikar et al., 2024), binary classification (Yao et al., 2025a;b), and reinforcement learning (Fu et al., 2024). These works typically deliver one-sided guarantees – conditions under which the student improves – or asymptotic rates. Because we obtain the KL gap in closed form, our conditions are two-sided and non-asymptotic: below $h_d(N_0)$ *no* scaled student improves on the teacher, and above it an explicitly constructed one does.

Gaussian and mixture analyses. Closest to our hypothesis class, Alemohammad et al. (2023) use a Gaussian example to exhibit shrinkage of the sample covariance eigenvalues, and show that purely synthetic data does not help; their follow-up (Alemohammad et al., 2024) restores performance by mixing in real data. Related treatments cover multi-label classification (Wu & Sahai, 2025), Gaussian mixtures in supervised learning (Somerstep et al., 2024; 2025), and mixtures in GANs (Wei et al., 2021). The distinction we draw is directional: this literature identifies eigenvalue *shrinkage* as the failure mode, whereas our analysis shows that deliberate *inflation* of the estimated covariance is the corrective, and pins down the exact factor and the sample budget at which it begins to pay off – without recourse to real data.

Iterative retraining and mixed corpora. For chains of models, Bertrand et al. (2024) establish that a neighborhood of parameters exists on which iterative retraining remains stable, though the neighborhood is not characterized even in the Gaussian case; Ferbach et al. (2024) obtain convergence under reward-model curation; Kazdan et al. (2025) give a finite-sample Gaussian treatment with asymptotic guarantees; and Lang et al. (2024) and Amin et al. (2026) analyze binary classification under corruption of real data by synthetic. A common feature is that real data is retained at every round, and that the guarantees are existence or limit statements. Our chain results differ on both counts. They concern the purely synthetic chain, they are finite-sample and explicit – the threshold $N_1^*(\rho, N_0)$ of Theorem 3 is sandwiched between multiples of $\big((1+\rho)^{L-1} - 1\big)\rho^{-1}N_0^2/(d+1)$, so the cost of insisting on improvement at *every* stage grows quadratically in the teacher's sample size and geometrically in the chain length – and they come with a matching negative result. § 3 shows the optimally tuned L-stage chain cannot beat the optimally tuned $(L-1)$-stage chain, so length

buys nothing that two stages do not already provide. We are not aware of an analogous impossibility statement for bootstrapped generative chains.

Phase transitions. Sharp transitions governed by a sample-size ratio are familiar in high-dimensional statistics, as in the existence of the logistic regression MLE (Candès & Sur, 2020). The transition we identify lives in the (N_0, N_1) plane of real versus synthetic budgets, and its boundary is the fixed point (3), which behaves nearly quadratically in N_0.

In conclusion, three features distinguish our treatment. First, *exactness*: the expected KL gap is available in closed form, so the transition is a curve rather than a pair of unmatched bounds. Second, *necessity*: we show not only that a biased student can win, but that below the threshold none can. Third, *scope*: the same Gaussian analysis yields concrete prescriptions in both a discriminative model (linear discriminant analysis, Theorem 5) and a generative one (linear flow matching, Corollary 2), where the optimal rescaling is given explicitly. The main limitation is the hypothesis class: our results are proved for Gaussian families and scaled maximum likelihood estimators.

B Applications

In this section we discuss the implications of our results on phase transitions in student-teacher setting for generative models that can provide us insights into design choices for machine learning models such as linear discriminant analysis (LDA) and flow matching (FM) model.

B.1 Linear Discriminant Analysis

We now instantiate the pipeline of Figure 1 in a *labelled, multi-class* setting, which is the setting closest to how synthetic data is actually used downstream. Let $\mathcal{Y} = [K]$ be the label alphabet. We work throughout with *linear* discriminant analysis (LDA) (Fisher, 1936; Bishop, 2006), i.e. a single covariance shared across classes with parameters

$$\theta = \left(\boldsymbol{\pi}, \{\mu_k\}_{k\in[K]}, \Sigma\right), \qquad \boldsymbol{\pi} \in \mathcal{S}_K, \;\; \mu_k \in \mathbb{R}^d, \;\; \Sigma \in \mathbb{S}^d_{++}, \tag{14}$$

where $\mathcal{S}_K := \{\boldsymbol{\pi} \in \mathbb{R}^K_{\geq 0} : \sum_k \pi_k = 1\}$ is the probability simplex on $[K]$. The LDA model is the joint density on $\mathbb{R}^d \times \mathcal{Y}$ taking the form:

$$p_\theta(\mathbf{x}, y) \;=\; \pi_y \, \mathcal{N}(\mathbf{x} \mid \mu_y, \Sigma)\,. \tag{15}$$

The optimal LDA classifier is obtained via the Bayes rule:

$$\hat{y}_\theta(\mathbf{x}) = \arg\max_{k\in[K]} \left(\ln \pi_k + \mu_k^\top \Sigma^{-1} \mathbf{x} - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k \right). \tag{16}$$

Given labelled samples $\{(\mathbf{x}_i, y_i)\}_{i\in[N]}$, write $N_k := |\{i : y_i = k\}|$ for the class counts, $\hat{\mu}_k := \frac{1}{N_k} \sum_{i:y_i=k} \mathbf{x}_i$ for the class means, and

$$\hat{\Sigma}_{\mathsf{LDA}} := \frac{1}{\sum_{k=1}^K N_k} \sum_{k=1}^K \sum_{i:y_i=k} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^\top, \tag{17}$$

for the combined covariance. The definition of a $(\overline{\beta}, \underline{\beta})$-MLE (see Definition 1) extends verbatim, with $\hat{\Sigma}_{\mathsf{LDA}}$ in place of the empirical covariance.

Definition 2 ($(\overline{\beta}, \underline{\beta})$ scaled LDA estimator)**.** *For positive numbers $\overline{\beta} \geq \underline{\beta}$, an LDA estimator that outputs $\theta = (\boldsymbol{\pi}, \{\mu_k\}, \Sigma)$ from labelled samples $\{(\mathbf{x}_i, y_i)\}_{i\in[N]}$ with class k having N_k samples is a $(\overline{\beta}, \underline{\beta})$ scaled LDA estimator if almost surely $\pi_k = N_k / \sum_{k=1}^K N_k$ and $\mu_k = \hat{\mu}_k$ for every $k \in [K]$, and*

$$\underline{\beta}\, \hat{\Sigma}_{\mathsf{LDA}} \;\preceq\; \Sigma \;\preceq\; \overline{\beta}\, \hat{\Sigma}_{\mathsf{LDA}}. \tag{18}$$

As before $\overline{\beta} = \underline{\beta} = 1$ recovers the maximum-likelihood LDA estimator.

With this definition we state the theorem that analyzes the regimes of the LDA estimator.

Theorem 5 (LDA). *Let $N_0, N_1 \geq d + K + 2$, and let p_{θ^*} be an LDA model (15) with non-singular Σ_*. Let the teacher $\mathcal{E}_0$ be a $(1,1)$-scaled LDA estimator using samples from p_{θ^*} with class counts N_{k0}, and let the student $\mathcal{E}_1$ be a $(\overline{\beta}, \underline{\beta})$-scaled LDA estimator using samples from p_{θ_0} with class counts N_{k1} satisfying the condition:*

$$\frac{N_{k1}}{\sum_{k=1}^{K} N_{k1}} = \frac{N_{k0}}{\sum_{k=1}^{K} N_{k0}} \qquad \text{for every } k \in [K], \tag{19}$$

For $l \in \{0,1\}$ define the constants

$$\gamma_l := \left(1 - \frac{d+K+1}{\sum_{k=1}^{K} N_l^{(k)}}\right)^{-1}, \qquad \rho_l := \sum_{k=1}^{K} \frac{\pi_*^{(k)}}{N_{kl}}, \tag{20}$$

$$c := \gamma_0 \left(1 + \rho_0\right), \qquad \beta^*_{\mathsf{LDA}} \;=\; \gamma_1 \left[c + \rho_1\right], \tag{21}$$

and the KL divergence approximation gap

$$\begin{aligned} \Delta_0^{\mathsf{LDA}}(\beta) := \frac{d}{2} \Bigg[\ln\left(\frac{2}{\sum_{k=1}^{K} N_{k1} - d - K - 1}\right) + \rho_1 \\ + \psi_d\left(\sum_{k=1}^{K} (N_1^{(k)} - 1)\right) + \phi\left(\frac{\beta}{\beta^*_{\mathsf{LDA}}}\right) - \phi\left(\frac{1}{c + \rho_1}\right) \Bigg]. \end{aligned} \tag{22}$$

Then the following statements hold true.

1. ***Unbiased student performs worse than its teacher.*** *If $\underline{\beta} = \overline{\beta} = 1$, then the ML student is strictly worse than the ML teacher, quantitatively*

$$\mathbb{E}\left[\mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_1}\right) - \mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_0}\right)\right] = \Delta_0^{\mathsf{LDA}}(1) \;>\; \frac{d(d+1)}{10\left(N_1 - d - K + 1\right)} \;>\; 0. \tag{23}$$

   Equivalently, the log-likelihood of the labelled data generated by the student under the true distribution is strictly less than that of the teacher.

2. ***Biased student case.*** *For general $\overline{\beta} \geq \underline{\beta} > 0$,*

$$-\frac{d}{2} \ln\left(\frac{\overline{\beta}}{\underline{\beta}}\right) + \Delta_0^{\mathsf{LDA}}(\overline{\beta}) \;\leq\; \mathbb{E}\left[\mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_1}\right) - \mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_0}\right)\right], \tag{24}$$

$$\mathbb{E}\left[\mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_1}\right) - \mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_0}\right)\right] \;\leq\; \frac{d}{2} \ln\left(\frac{\overline{\beta}}{\underline{\beta}}\right) + \Delta_0^{\mathsf{LDA}}(\underline{\beta}). \tag{25}$$

3. ***Optimally biased student performs better than teacher.*** *If $\underline{\beta} = \overline{\beta} = \beta^*_{\mathsf{LDA}}$, then there exists a function $h_{(d,K)}(N_0)$ such that for all $N_1 \geq h_{(d,K)}(N_0)$ it holds true that $\mathbb{E}\left[\mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_1}\right)\right] \leq \left[\mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_0}\right)\right]$.*

B.2 Flow Matching models

We consider a setting where we aim to learn $\mathcal{N}(\mu^*, \Sigma_*)$ using a FM (Lipman et al., 2023) with velocity field, at time t the distribution takes the form

$$p_t = \mathcal{N}\big(\alpha_t \mu^*,\; \alpha_t^2 \Sigma_* + \sigma_t^2 I_d\big)\,, \tag{26}$$

the law of the stochastic interpolant $x_t = \alpha_t x_1 + \sigma_t x_0$ built from $x_1 \sim \mathcal{N}(\mu^*, \Sigma_*)$ and an independent $x_0 \sim \mathcal{N}(0, I_d)$, where the schedule (α_t, σ_t) is differentiable and satisfies $(\alpha_0, \sigma_0) = (0, 1)$ and $(\alpha_1, \sigma_1) = (1, 0)$. Both endpoints being Gaussian, the minimizer of the flow-matching objective is available in closed form and is affine in x,

$$v_t(x) = \dot{\alpha}_t \mu^* + \left(\dot{\alpha}_t \alpha_t \Sigma_* + \dot{\sigma}_t \sigma_t I_d\right)\left(\alpha_t^2 \Sigma_* + \sigma_t^2 I_d\right)^{-1}(x - \alpha_t \mu^*), \tag{27}$$

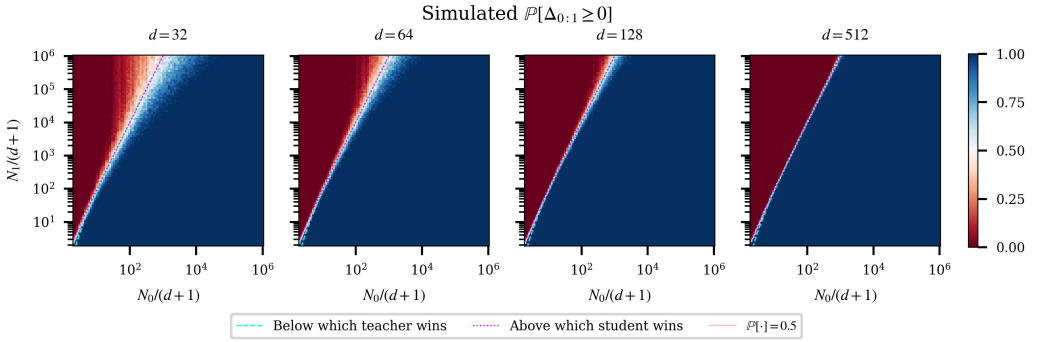


Figure 5: Monte carlo (100 trials) simulations to estimate $\mathbb{P}(\Delta_{0:1} \geq 0)$.

so that transporting $\mathcal{N}(0, I_d)$ along $\dot{x} = v_t(x)$ up to time t reproduces (26) exactly; the three schedules we consider are the linear (OT) path $(\alpha_t, \sigma_t) = (t, 1-t)$, the cosine path $(\sin \frac{\pi t}{2}, \cos \frac{\pi t}{2})$ and the variance-preserving SDE path $(t, \sqrt{1-t^2})$.

Now let the field be fitted not to Σ_* but to the teacher's estimate Σ_0 formed from N_0 samples, on which a student is subsequently trained using N_1 samples. By Corollary 1 the student should inflate its covariance by $\beta^*_{0:1} = ((1 - \frac{d+1}{N_0})(1 - \frac{d+1}{N_1}))^{-1} > 1$. A flow-matching sampler has no such scaling knob therefore before we learn a student model we scale up the data points obtained from teacher, this leads to the provable improvements:

Corollary 2 (Improved FM). *Under the setting of Theorem 1, integrate the flow to $t = 1$ and rescale its output,*

$$\tilde{\mathbf{x}} := \sqrt{\beta^*}\,\mathbf{x}, \qquad \mathbf{x} \sim \mathcal{N}(0, \Sigma_0)\,, \qquad \beta^*_{0:1} = \frac{N_0 N_1}{(N_0 - d - 1)(N_1 - d - 1)}, \tag{28}$$

and let the student be the ML estimator on M such samples. Then holds true that

$$\mathbb{E}[\Delta_{0:1}] \;=\; \tilde{\Delta}_{0:1}(\beta^*_{0:1}). \tag{29}$$

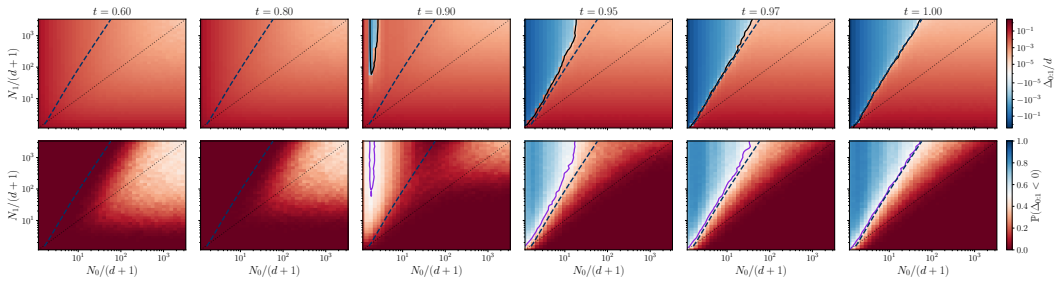
Moreover, the sample complexity for student to outperform the teacher or vice-versa are the same as Corollary 1.

Proof. Direct application of Theorem 1. □

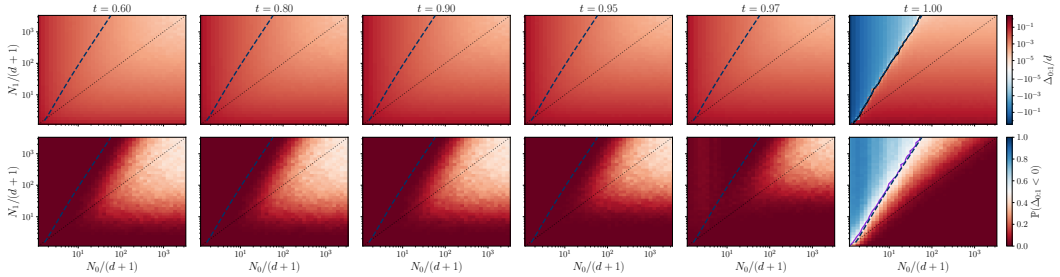
C Additional Numerical Simulations

PAC simulations. To complement the sign-agreement results of Figure 2(c,d), Figure 5 shows the Monte Carlo estimate of $\mathbb{P}(\Delta_{0:1} \geq 0)$, the probability that the teacher is at least as good as the student, over 100 trials. The transition occurs exactly at the theoretical boundary $\mathbb{E}[\Delta_{0:1}] = 0$. Near the boundary the probability is close to $1/2$, so the winner of a single run is essentially a fair coin toss. This uncertain band narrows rapidly as d grows, in line with the $\mathcal{O}(d^{-2})$ concentration of Theorem 2.

Flow matching. Figure 6 repeats the experiment of Figure 3b with the cosine and variance-preserving schedules. At $t = 1$ all three schedules produce the same phase diagram, which matches Theorem 1, since the endpoint of every flow is the fitted Gaussian. At intermediate times they differ: the student's winning region appears earliest for the linear OT path, around $t \approx 0.8$, later for the cosine path, around $t \approx 0.95$, and only at close to $t = 1$ for the variance-preserving path.



(a) Cosine, $(\alpha_t, \sigma_t) = (\sin \frac{\pi t}{2}, \cos \frac{\pi t}{2})$.



(b) Variance-preserving, $(\alpha_t, \sigma_t) = (t, \sqrt{1-t^2})$.

Figure 6: Flow-matching student versus teacher as the stopping time is swept, for the three schedules.

D Proof sketch of Theorem 1

Proof. We will majorly focus on proving the last statement, as the first two can be obtained as simple consequence of the third. Denote Σ_* as the true covariance matrix, Σ_0 as the teacher's estimated covariance matrix, and Σ_1 as the student's estimated covariance matrix. We now decompose the expectation into two stages

$$\mathbb{E}\left[\Delta_{0:1}\right] = \mathbb{E}_{\Sigma_0}\left[\mathbb{E}_{\Sigma_1|\Sigma_0}\left[\Delta_{0:1}\right]\right]. \tag{30}$$

The KL divergence between any two Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ is given by (see Proposition 3):

$$\mathsf{KL}\left(\mathcal{N}(\mu_1, \Sigma_1) \,||\, \mathcal{N}(\mu_2, \Sigma_2)\right) = \frac{1}{2}\left[\ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - d + \langle \Sigma_2^{-1}, \Sigma_1\rangle + (\mu_2 - \mu_1)^\top \Sigma_2^{-1}(\mu_2 - \mu_1)\right]. \tag{31}$$

The KL gap can be written as:

$$\Delta_{0:1} = \frac{1}{2}\left[\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \langle (\Sigma_1)^{-1} - (\Sigma_0)^{-1}, \Sigma_*\rangle\right]. \tag{32}$$

Recall that student estimator is a $(\overline{\beta}, \underline{\beta})$-MLE, that satisfies the Loewner order: $\underline{\beta}\Sigma_{\mathsf{ML}}(N_1, \Sigma_0) \preceq \Sigma_1 \preceq \overline{\beta}\Sigma_{\mathsf{ML}}(N_1, \Sigma_0)$, where $\Sigma_{\mathsf{ML}}(N_1, \Sigma_0)$ is the ML estimator of the covariance matrix based on N_1 i.i.d. samples from p_{θ_0}. Since Σ_* is a non-singular covariance matrix, we have

$$\begin{aligned}
&\frac{1}{2}\left[d\ln(\underline{\beta}) + \ln\left(\frac{|\Sigma_{\mathsf{ML}}(N_1, \Sigma_0)|}{|\Sigma_{\mathsf{ML}}(N_0, \Sigma_*)|}\right) + \langle \overline{\beta}^{-1}\left(\Sigma_{\mathsf{ML}}(N_1, \Sigma_0)\right)^{-1} - \left(\Sigma_{\mathsf{ML}}(N_0, \Sigma_*)\right)^{-1}, \Sigma_*\rangle\right] \\
&\quad \leq \Delta_{0:1} \leq \\
&\frac{1}{2}\left[d\ln(\overline{\beta}) + \ln\left(\frac{|\Sigma_{\mathsf{ML}}(N_1, \Sigma_0)|}{|\Sigma_{\mathsf{ML}}(N_0, \Sigma_*)|}\right) + \langle \underline{\beta}^{-1}\left(\Sigma_{\mathsf{ML}}(N_1, \Sigma_0)\right)^{-1} - \left(\Sigma_{\mathsf{ML}}(N_0, \Sigma_*)\right)^{-1}, \Sigma_*\rangle\right].
\end{aligned} \tag{33}$$

Next we will apply conditional expectation to the above bounds, and use the properties of Wishart distribution to obtain the desired bounds. Empirical covariance of N i.i.d. samples from $\mathcal{N}(0, \Sigma)$

follows a Wishart distribution $\mathcal{W}_n(\Sigma, N)$, and the inverse of a Wishart distribution $\mathcal{W}_n(\Sigma, N)^{-1}$ is an inverse-Wishart distribution $\mathcal{W}_n^{-1}(\Sigma^{-1}, N)$. Refer to Lemma 5 and Proposition 3 for the properties of KL divergence between Gaussian distributions with Wishart-distributed covariance matrices, and the moments of Wishart distribution, respectively. We have the following bounds on the conditional expectation of the KL gap:

$$\begin{aligned}
&\frac{1}{2}\left[d\ln(\underline{\beta}) + d\ln(2/N_1) + d\psi_d(N_1) + \left[\frac{\overline{\beta}^{-1}N_1}{N_1-d-1} - 1\right]\langle(\Sigma_{\mathsf{ML}}(N_0,\Sigma_*))^{-1}, \Sigma_*\rangle\right] \\
&\quad \leq \mathbb{E}_{\Sigma_1|\Sigma_0}\left[\Delta_{0:1}\right] \leq \\
&\frac{1}{2}\left[d\ln(\overline{\beta}) + d\ln(2/N_1) + d\psi_d(N_1) + \left[\frac{\underline{\beta}^{-1}N_1}{N_1-d-1} - 1\right]\langle(\Sigma_{\mathsf{ML}}(N_0,\Sigma_*))^{-1}, \Sigma_*\rangle\right].
\end{aligned} \tag{34}$$

We now take the outer expectation and use the properties of Wishart distribution to obtain the desired bounds.

$$\begin{aligned}
&\frac{d}{2}\left[\ln(\underline{\beta}) + \ln(2/N_1) + \psi_d(N_1) + \left[\frac{\overline{\beta}^{-1}N_1}{N_1-d-1} - 1\right]\left[\frac{N_0}{N_0-d-1}\right]\right] \\
&\quad \leq \mathbb{E}\left[\Delta_{0:1}\right] \leq \\
&\frac{d}{2}\left[\ln(\overline{\beta}) + \ln(2/N_1) + \psi_d(N_1) + \left[\frac{\underline{\beta}^{-1}N_1}{N_1-d-1} - 1\right]\left[\frac{N_0}{N_0-d-1}\right]\right].
\end{aligned} \tag{35}$$

Using the definition of $\tilde{\Delta}_{0:1}(\cdot)$ we have

$$-\frac{d}{2}\ln\left(\frac{\overline{\beta}}{\underline{\beta}}\right) + \tilde{\Delta}_{0:1}(\overline{\beta}) \leq \mathbb{E}\left[\Delta_{0:1}\right] \leq \frac{d}{2}\ln\left(\frac{\overline{\beta}}{\underline{\beta}}\right) + \tilde{\Delta}_{0:1}(\underline{\beta}). \tag{36}$$

Statement 2 can be proved setting $\underline{\beta} = \overline{\beta} = 1$ and noting that $\tilde{\Delta}_{0:1}(\underline{\beta}) = \frac{d}{2}\left[\psi_d(N_1) + \ln(2/N_1) + \frac{N_0}{N_0-d-1}\frac{d+1}{N_1-d-1}\right]$, and using the inequality developed in Eq (43) in the proof of Corollary 1 (see § E) we conclude that $\tilde{\Delta}_{0:1}(\underline{\beta}) > 0$.

Statement 1 can be proved by setting $\underline{\beta} = \overline{\beta} = \beta$, this gives us $\mathbb{E}[\Delta_{0:1}] = \frac{d}{2}\left[\phi\left(\frac{\beta}{\beta^*_{0:1}}\right) + \phi\left(1 - \frac{d+1}{h_d^{-1}(N_1)}\right) - \phi\left(1 - \frac{d+1}{N_0}\right)\right]$. The proof is completed by observing that $h_d^{-1}(\cdot)$ is increasing function, and $\phi(x)$ is decreasing and positive function when $x < 1$. □

E Proof of Corollary 1

Proof. When $\underline{\beta}\zeta = \overline{\beta}/\zeta = \beta^* = (1-(d+1)/(N_0))^{-1}(1-(d+1)/N_1)^{-1}$, Theorem 1 establishes the inequality:

$$-\frac{d}{2}\left[\ln(\zeta) - 1/\zeta + 1\right] \leq \mathbb{E}\left[\Delta_{0:1}\right] - \tilde{\Delta}_{0:1}(\beta^*_{0:1}) \leq \frac{d}{2}\left[\ln(\zeta) + \zeta - 1\right] \tag{37}$$

Since we have a concentration bound for $\mathbb{E}[\Delta_{0:1}]$ around $\tilde{\Delta}_{0:1}(\beta^*)$, we can now analyze the sign of $\tilde{\Delta}_{0:1}(\beta^*)$ from here on. Recall,

$$\begin{aligned}
\tilde{\Delta}_{0:1}(\beta^*) = \frac{d}{2}\Bigg[&-\ln\left(1 - \left[\frac{d+1}{N_0}\right]\right) - \frac{(d+1)/(N_0)}{(1-(d+1)/N_0)} \\
&+ \frac{1}{d}\sum_{j\in[d]}\left[\ln\left(\frac{2}{N_1-d-1}\right) + \psi\left(\frac{N_1+1-j}{2}\right)\right]\Bigg].
\end{aligned} \tag{38}$$

It is easy to verify that

$$-\frac{1}{2\left(1-\frac{d+1}{N_0}\right)^2}\left(\frac{d+1}{N_0}\right)^2 \leq -\ln\left(1 - \left[\frac{d+1}{N_0}\right]\right) - \frac{(d+1)/(N_0)}{(1-(d+1)/N_0)} \tag{39}$$

$$-\ln\left(1 - \left[\frac{d+1}{N_0}\right]\right) - \frac{(d+1)/(N_0)}{(1-(d+1)/N_0)} \leq -\frac{1}{2\left(1-\frac{d+1}{N_0}\right)}\left(\frac{d+1}{N_0}\right)^2, \tag{40}$$

and using the fact that $\ln(x) - \frac{1}{x} < \psi(x) < \ln(x) - \frac{1}{2x}$ for $x \geq 1$ (see Theorem 6) we obtain

$$\begin{aligned}
&\frac{1}{d}\sum_{j\in[d]}\left[\ln\left(1+\frac{d-j+2}{N_1-1-d}\right)-\frac{1}{N_1+1-j}\right] \\
&\leq \frac{1}{d}\sum_{j\in[d]}\left[\ln\left(\frac{2}{N_1-d-1}\right)+\psi\left(\frac{N_1+1-j}{2}\right)\right] \leq \\
&\frac{1}{d}\sum_{j\in[d]}\left[\ln\left(1+\frac{d-j+2}{N_1-1-d}\right)-\frac{1}{N_1+1-j}\right].
\end{aligned} \tag{41}$$

We can simply the above using the fact that $1 \leq j \leq d$ to obtain

$$\begin{aligned}
&\ln\left(1+\frac{2}{N_1-1-d}\right)-\frac{1}{N_1+1-d} \\
&< \frac{1}{d}\sum_{j\in[d]}\left[\ln\left(\frac{2}{N_1-d-1}\right)+\psi\left(\frac{N_1+1-j}{2}\right)\right] < \\
&\ln\left(1+\frac{d+1}{N_1-1-d}\right)-\frac{1}{N_1}.
\end{aligned} \tag{42}$$

Claim. The below inequality holds true:

$$\frac{1}{2\left(1-\frac{1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] < \ln\left(\frac{2}{N_1-d-1}\right)+\psi_d(N_1) < \frac{1}{2\left(1-\frac{d+1}{N_1}\right)}\left[\frac{d+1}{N_1}\right]. \tag{43}$$

Proof. Write $n := d$, $m := N_1 - n - 1 > 0$ and $x_j := \frac{N_1+1-j}{2}$, so the middle quantity is $M := \ln\left(\frac{2}{m}\right) + \frac{1}{n}\sum_{j\in[n]}\psi(x_j)$. We use two standard digamma inequalities,

$$\text{(A)}\quad \psi(x) < \ln(x) - \frac{1}{2x} \quad (x>0), \qquad \text{(B)}\quad \psi(x) > \ln\left(x-\tfrac{1}{2}\right) \quad \left(x>\tfrac{1}{2}\right), \tag{44}$$

together with the convexity of $t \mapsto 1/t$. In both directions the $\ln 2$ terms cancel, because $\ln(2/m) + \frac{1}{n}\sum_j \ln(a_j/2) = \frac{1}{n}\sum_j \ln(a_j/m)$.

Upper bound. By (A), $\psi(x_j) < \ln\frac{N_1+1-j}{2} - \frac{1}{N_1+1-j}$. Re-indexing with $i = n+2-j$ (so i runs over $2,\ldots,n+1$ and $N_1+1-j = m+i$),

$$M < \frac{1}{n}\sum_{i=2}^{n+1}\left[\ln\left(1+\frac{i}{m}\right)-\frac{1}{m+i}\right]. \tag{45}$$

As $t \mapsto 1/t$ is convex, the trapezoidal rule over-estimates the integral, $\ln\left(1+\frac{i}{m}\right) = \int_m^{m+i}\frac{dt}{t} \leq \frac{i}{2}\left(\frac{1}{m}+\frac{1}{m+i}\right)$, so each summand is at most $\frac{i}{2m} + \frac{i-2}{2(m+i)}$. Using $\sum_{i=2}^{n+1} i = \frac{n(n+3)}{2}$ for the first part, and $\frac{1}{m+i} < \frac{1}{m}$ together with $\sum_{i=2}^{n+1}(i-2) = \frac{n(n-1)}{2}$ for the second, we obtain

$$M < \frac{n+3}{4m} + \frac{n-1}{4m} = \frac{n+1}{2m} = \frac{1}{2\left(1-\frac{d+1}{N_1}\right)}\left[\frac{d+1}{N_1}\right]. \tag{46}$$

Lower bound. By (B), $\psi(x_j) > \ln\frac{N_1-j}{2}$, which is legitimate since $N_1 > n \geq j$. Re-indexing with $i = n+1-j$ (so i runs over $1,\ldots,n$ and $N_1 - j = m+i$),

$$M > \frac{1}{n}\sum_{i=1}^{n}\ln\left(1+\frac{i}{m}\right). \tag{47}$$

By convexity the midpoint rule under-estimates the integral, $\ln\left(1+\frac{i}{m}\right) = \int_m^{m+i}\frac{dt}{t} \geq \frac{i}{m+i/2} = \frac{2i}{2m+i} \geq \frac{2i}{2m+n}$, the last step because $i \leq n$. Summing with $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$,

$$M > \frac{n+1}{2m+n} = \frac{d+1}{2N_1-d-2} \geq \frac{d+1}{2\left(N_1-1\right)} = \frac{1}{2\left(1-\frac{1}{N_1}\right)}\left[\frac{d+1}{N_1}\right], \tag{48}$$

where the final inequality uses $2m+n = 2N_1 - n - 2 \leq 2(N_1 - 1)$. In particular the sharper lower bound $\frac{d+1}{2N_1-d-2}$ is also available. Both bounds hold for every $N_1 > d+1$ and $d \geq 1$, and each is tight: the ratio of M to either bound tends to 1 as $N_1 \to \infty$. □

This gives us the final bound

$$\begin{aligned}
&\frac{1}{2\left(1-\frac{1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{2\left(1-\frac{d+1}{N_0}\right)^2}\left[\frac{d+1}{N_0}\right]^2 \\
&\qquad \leq \frac{2}{d}\tilde{\Delta}_{0:1}(\beta^*_{0:1}) \leq \\
&\frac{1}{2\left(1-\frac{d+1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{2\left(1-\frac{d+1}{N_0}\right)}\left[\frac{d+1}{N_0}\right]^2 .
\end{aligned} \tag{49}$$

Combining Eq (37) and Eq (49) gives us

$$\begin{aligned}
&\frac{1}{2\left(1-\frac{1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{2\left(1-\frac{d+1}{N_0}\right)^2}\left[\frac{d+1}{N_0}\right]^2 - (\ln(\zeta) - 1/\zeta + 1) \\
&\qquad \leq \frac{2}{d}\mathbb{E}[\Delta_{0:1}] \leq \\
&\frac{1}{2\left(1-\frac{d+1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{2\left(1-\frac{d+1}{N_0}\right)}\left[\frac{d+1}{N_0}\right]^2 + (\ln(\zeta) + \zeta - 1) .
\end{aligned} \tag{50}$$

This concludes the proof of the first part of the Corollary. For the second part we set $\zeta = 1$, and we have the following:

$$\begin{aligned}
&\frac{1}{2\left(1-\frac{1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{2\left(1-\frac{d+1}{N_0}\right)^2}\left[\frac{d+1}{N_0}\right]^2 \\
&\qquad \leq \frac{2}{d}\mathbb{E}[\Delta_{0:1}] \leq \\
&\frac{1}{2\left(1-\frac{d+1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{2\left(1-\frac{d+1}{N_0}\right)}\left[\frac{d+1}{N_0}\right]^2 .
\end{aligned} \tag{51}$$

For the student to outperform the teacher, we require the upper bound to be negative, which is equivalent to the following condition on N_1:

$$\frac{N_1}{d+1} > \frac{N_0}{d+1} + \left(\frac{N_0}{d+1} - 1\right)^2 . \tag{52}$$

For the teacher to outperform the student, we require the lower bound to be positive, which is equivalent to the following condition on N_1:

$$\frac{N_1}{d+1} < \frac{1}{d+1} + \left(\frac{N_0}{d+1} - 1\right)^2 . \tag{53}$$

For the second statement recall the decomposition of the mean:

$$\begin{aligned}
\mathbb{E}\left[\Delta_{0:1}\right] = \frac{d}{2}\Bigg[&\underbrace{-\ln\left(1-\left[\frac{d+1}{N_0}\right]\right) - \frac{(d+1)/(N_0)}{(1-(d+1)/N_0)}}_{=:T_1(N_0)} \\
&+ \underbrace{\frac{1}{d}\sum_{j\in[d]}\left[\ln\left(\frac{2}{N_1-d-1}\right) + \psi\left(\frac{N_1+1-j}{2}\right)\right]}_{=:T_2(N_1)}\Bigg] .
\end{aligned} \tag{54}$$

Observe that clearly $T_1 < 0$. We will verify the monotonicity of function T_2. The first derivative of T_2 is given by

$$T_2'(N_1) = \frac{1}{2d} \sum_{j \in [d]} \left[\psi^{(1)} \left(\frac{N_1 + 1 - j}{2} \right) - \frac{2}{N_1 - d - 1} \right]. \tag{55}$$

Using Theorem 6 we have the bound:

$$\begin{aligned} &\frac{1}{d} \sum_{j \in [d]} \left[\frac{1}{N_1 + 1 - j} - \frac{1}{N_1 - d - 1} + \frac{1}{(N_1 + 1 - j)^2} \right] \\ &\quad \leq T_2'(N_1) \leq \\ &\frac{1}{d} \sum_{j \in [d]} \left[\frac{1}{N_1 + 1 - j} - \frac{1}{N_1 - d - 1} + \frac{1}{(N_1 + 1 - j)^2} + \frac{1}{3(N_1 + 1 - j)^3} \right]. \end{aligned} \tag{56}$$

Using the fact that $1 \leq j \leq d$ we obtain the upper bound:

$$\begin{aligned} T_2'(N_1) &< -\frac{2}{(N_1 - d)^2 - 1} + \frac{1}{(N_1 + 1 - d)^2} + \frac{1}{3(N_1 + 1 - d)^3}, \\ &= -\frac{3(N_1 - d)^2 + 11(N_1 - d) + 10}{3(N_1 - 1 - d)(N_1 + 1 - d)^3} < 0. \end{aligned} \tag{57}$$

Therefore, T_2 is strictly decreasing function of N_1. Using the first two statements and mean value theorem, the statement 2 is true. This concludes the proof. □

F Proof of Theorem 2

Proof. We apply Cantelli's Lemma on the random variable that is

$$\mathbb{P}\left(\operatorname{sgn}(\Delta) \neq \operatorname{sgn}(\mathbb{E}[\Delta])\right) \leq \frac{\mathsf{Var}(\Delta)}{\left(\mathbb{E}[\Delta]\right)^2 + \mathsf{Var}(\Delta)}, \tag{58}$$

We obtain the mean from Theorem 1 and variance from Proposition 1. This completes the first part of the proof. We now bound the failure probability in the interested regimes.

The mapping $x \to \frac{x}{1+x}$ is strictly increasing in x when $x \geq 0$. Therefore, we will focus on obtain a tight upper bound for variance-squared mean. First we will work on the variance term given by

$$\begin{aligned} \mathsf{Var}(\Delta) = &\frac{d}{2}\psi_d^{(1)}(N_1) + \frac{1}{(d+1)^2}\left[\frac{d+1}{N_1}\right]^3 - \frac{1}{(d+1)^2}\left[\frac{d+1}{N_1}\right]^2 \\ &+ \frac{1}{2(d+1)^4(1-d/N_1)(1-(d+3)/N_1)}\left[\frac{d+1}{N_1}\right]^4 \\ &+ \frac{dN_1(N_1-d-1)}{2(d+1)^3(N_1-d)(N_1-d-3)}\left(\frac{(N_0-1)\,(N_0-d-1)}{(N_0-d)(N_0-d-3)}\right)\left[\frac{d+1}{N_1}\right]^3 \\ &+ \frac{2d}{(d+1)}\left[\frac{1}{1-(d+1)/N_0}\right]^2 \frac{N_0(N_0-1)}{(N_0-d)(N_0-d-3)}\left[\frac{d+1}{N_0}\right]^3 \\ &+ \frac{d}{(d+1)^5}\frac{N_0(N_0-1)}{(N_0-d)(N_0-d-3)}\frac{(N_1)^2}{(N_1-d)(N_1-d-3)}\left[\frac{d+1}{N_0}\right]\left[\frac{d+1}{N_1}\right]^4. \end{aligned} \tag{59}$$

$$\begin{aligned} \mathsf{Var}(\Delta) = &\frac{d}{2}\psi_d^{(1)}(N_1) - \left(\frac{d}{d+1}\right)\left(\frac{N_1}{N_1-d-1}\right)\left[\frac{d+1}{N_1}\right] \\ &+ \left(\frac{d}{d+1}\right)\frac{g_1 g_2}{2\beta^*}\left[\frac{d+1}{N_1}\right] + \left(\frac{d}{d+1}\right)^2 \frac{g_2}{2}\left[\frac{d+1}{N_1}\right]^2 \\ &+ g_1 g_2 \frac{d}{(d+1)^3}\left[\frac{d+1}{N_0}\right]\left[\frac{d+1}{N_1}\right]^2 + \left(\frac{d}{d+1}\right)\frac{g_1}{2}\left[\frac{N_0}{N_0-d-1}\right]^2\left[\frac{d+1}{N_0}\right]^3, \end{aligned} \tag{60}$$

We will upper bound the poly-gamma function. Recall the inequalities $0 \leq \psi^{(1)}(x) - \frac{1}{x} \leq \frac{1}{x^2}$. The first term in the variance expression can be written as

$$\begin{aligned}\frac{d}{2}\psi_d^{(1)}(N_1) = \frac{1}{4}\sum_{i=1}^{d}\psi^{(1)}\left(\frac{N_1+1-i}{2}\right) &\leq \frac{1}{4}\sum_{i=1}^{d}\left[\frac{2}{N_1+1-i} + \frac{4}{(N_1+1-i)^2}\right], \\ &\leq \frac{d}{2}\left[\frac{1}{N_1+1-d} + \frac{2}{(N_1+1-d)^2}\right], \\ &< \frac{1}{2}\left(\frac{d}{d+1}\right)\left[\frac{N_1}{N_1-d-1}\right]\left[\frac{d+1}{N_1}\right].\end{aligned} \tag{61}$$

This gives us

$$\begin{aligned}\mathsf{Var}(\Delta) < \frac{1}{2}\left(\frac{d}{d+1}\right)\left[\frac{g_1g_2}{\beta^*} - \frac{N_1}{N_1-d-1}\right]\left[\frac{d+1}{N_1}\right] + \frac{g_2}{2}\left(\frac{d}{d+1}\right)^2\left[\frac{d+1}{N_1}\right]^2 \\ + \left(\frac{g_1g_2d}{d+1}\right)\frac{1}{(d+1)^2}\left[\frac{d+1}{N_0}\right]\left[\frac{d+1}{N_1}\right]^2 + \frac{g_1}{2}\left(\frac{d}{d+1}\right)\left[\frac{N_0}{N_0-d-1}\right]^2\left[\frac{d+1}{N_0}\right]^3,\end{aligned} \tag{62}$$

Using Cauchy-Schwarz for the cross term we obtain

$$\left[\frac{d+1}{N_0}\right]\left[\frac{d+1}{N_1}\right]^2 \leq \frac{1}{2}\left[\frac{d+1}{N_0}\right]^2 + \frac{1}{2}\left[\frac{d+1}{N_1}\right]^4, \tag{63}$$

this gives us the upper bound on variance

$$\begin{aligned}\mathsf{Var}(\Delta) < \frac{1}{2}\left(\frac{d}{d+1}\right)\left[\frac{g_1g_2}{\beta^*} - \frac{N_1}{N_1-d-1}\right]\left[\frac{d+1}{N_1}\right] \\ + \frac{g_2}{2}\left(\frac{d}{d+1}\right)^2\left[\frac{d+1}{N_1}\right]^2 + \left(\frac{g_1g_2d}{2(d+1)}\right)\frac{1}{(d+1)^2}\left[\frac{d+1}{N_1}\right]^4 \\ + \left(\frac{g_1g_2d}{2(d+1)}\right)\frac{1}{(d+1)^2}\left[\frac{d+1}{N_0}\right]^2 + \frac{g_1}{2}\left(\frac{d}{d+1}\right)\left[\frac{N_0}{N_0-d-1}\right]^2\left[\frac{d+1}{N_0}\right]^3,\end{aligned} \tag{64}$$

Taking the leading terms common we obtain

$$\mathsf{Var}(\Delta) \leq c_1\frac{1}{\left[\frac{N_0}{d+1}-1\right]^2} + c_2\left[\frac{d+1}{N_1}\right]^2, \tag{65}$$

where

$$c_1 = \left(\frac{g_1d}{2(d+1)}\right)\left[1 - \frac{2g_2}{(d+1)^2} + \frac{g_2}{(d+1)^2}\left[\frac{d+1}{N_0}\right]\right]\left[\frac{d+1}{N_0}\right] + \frac{g_1g_2d}{2(d+1)^3}, \tag{66}$$

$$c_2 = \frac{g_2}{2}\left(\frac{d}{d+1}\right)^2\left[1 + \frac{g_1}{d(d+1)}\left[\frac{d+1}{N_1}\right]^2\right] \tag{67}$$

$$+ \frac{1}{2}\left(\frac{d}{d+1}\right)\left[\frac{N_1}{d+1} - 1\right]\left[\frac{g_1g_2(N_1-d-1)}{\beta^*N_1} - 1\right]. \tag{68}$$

In the regime where $d \geq 1$, $N_0, N_1 \geq 2d+3$ we have that $\frac{3}{4} < c_1 \leq \frac{4}{3}$, and $\frac{3}{2} < c_2 \leq \frac{5}{2}$, both attained at $d = 1$, $N_0 = N_1 = 5$. Therefore, in the desired regime we have the bound

$$\mathsf{Var}(\Delta) \leq \frac{77}{64}\frac{1}{\left[\frac{N_0}{d+1}-1\right]^2} + \frac{17}{6}\left[\frac{d+1}{N_1}\right], \tag{69}$$

$$< \frac{17}{6}\left[\frac{1}{2\left[\frac{N_0}{d+1}-1\right]^2} + \left[\frac{d+1}{N_1}\right]\right]. \tag{70}$$

Next, we move onto lower bounding the squared mean. After setting $\varepsilon = 0$ in Corollary 1 we obtain the bound

$$\begin{aligned} \frac{d}{4}\left[\frac{1}{\left(1-\frac{1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{\left(\frac{N_0}{d+1}-1\right)^2}\right] \\ \leq \quad \mathbb{E}[\Delta] \quad \leq \\ \frac{d}{4}\left[\frac{1}{\left(1-\frac{d+1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{\left(\frac{N_0}{d+1}-1\right)^2 + \left(\frac{N_0}{d+1}-1\right)}\right]. \end{aligned} \tag{71}$$

Analyzing the positive and negative cases separately gives us

$$(\mathbb{E}[\Delta])^2 \geq \begin{cases} \frac{d^2}{16}\left[\frac{1}{\left(1-\frac{1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{\left(\frac{N_0}{d+1}-1\right)^2}\right]^2; & \text{if } \frac{N_1}{d+1} \leq \frac{1}{d+1} + \left(\frac{N_0}{d+1}-1\right)^2, \\ \frac{d^2}{16}\left[\frac{1}{\left(1-\frac{d+1}{N_1}\right)}\left[\frac{d+1}{N_1}\right] - \frac{1}{\left(\frac{N_0}{d+1}-1\right)^2 + \left(\frac{N_0}{d+1}-1\right)}\right]^2; & \text{if } \frac{N_1}{d+1} \geq \frac{N_0}{d+1} + \left(\frac{N_0}{d+1}-1\right)^2. \end{cases} \tag{72}$$

Taking the positive term common and lower-bounding the terms gives us

$$(\mathbb{E}[\Delta])^2 > \begin{cases} \frac{d^2}{16}\left[\frac{d+1}{N_1}\right]^2\left[1 - \frac{\frac{N_1-1}{d+1}}{\left(\frac{N_0}{d+1}-1\right)^2}\right]^2; & \text{if } 1 < \frac{N_1}{d+1} \leq \frac{1}{d+1} + \left(\frac{N_0}{d+1}-1\right)^2, \\ \frac{9d^2}{400}\frac{1}{\left[\frac{N_0}{d+1}-1\right]^4}\left[1 - \frac{\frac{N_0}{d+1}\left(\frac{N_0}{d+1}-1\right)}{\left(\frac{N_1}{d+1}-1\right)}\right]^2; & \text{if } \frac{N_1}{d+1} \geq \frac{N_0}{d+1} + \left(\frac{N_0}{d+1}-1\right)^2, \frac{N_0}{d+1} \geq 2 + \frac{1}{d+1}. \end{cases} \tag{73}$$

Now for any fixed $\overline{\kappa}, \underline{\kappa} \in (0,1)$ we have

$$(\mathbb{E}[\Delta])^2 > \begin{cases} \frac{d^2}{16}\left[\frac{d+1}{N_1}\right]^2 \underline{\kappa}; & \text{if } 1 < \frac{N_1}{d+1} < \frac{1}{d+1} + \left(1-\sqrt{\underline{\kappa}}\right)\left(\frac{N_0}{d+1}-1\right)^2, \\ \frac{9d^2}{400}\left[\frac{N_0}{d+1}-1\right]^{-4}\overline{\kappa}; & \text{if } \frac{N_1}{d+1} > \frac{1}{1-\sqrt{\overline{\kappa}}}\frac{N_0}{d+1} - \frac{\sqrt{\overline{\kappa}}}{1-\sqrt{\overline{\kappa}}} + \frac{1}{1-\sqrt{\overline{\kappa}}}\left(\frac{N_0}{d+1}-1\right)^2. \end{cases} \tag{74}$$

It is convenient to carry the two regions of (76) in parametric form: for $\kappa, \eta \in (0,1)$ and $T > 0$ put

$$\mathsf{T}_\kappa := \frac{\frac{N_0}{d+1} - \sqrt{\kappa} + \left(\frac{N_0}{d+1}-1\right)^2}{1-\sqrt{\kappa}}, \tag{75}$$

Define the sets:

$$\begin{aligned} \mathcal{R}_- := \Bigg\{ & (N_0, N_1) : 2d+3 \leq N_0 \leq 1 + \frac{d^\varepsilon}{10\sqrt{6}} + d + \frac{d^{1+\varepsilon}}{10\sqrt{6}}, \\ & N_1 > \max\left\{2d+3,\ \frac{\sqrt{2}}{\sqrt{2}-1}N_0 - \frac{(d+1)}{\sqrt{2}-1} + \frac{\sqrt{2}(d+1)}{\sqrt{2}-1}\left(\frac{N_0}{d+1}-1\right)^2\right\}\Bigg\}, \text{and} \\ \mathcal{R}_+ := \Bigg\{ & (N_0, N_1) : 2d+3 < N_0, 2d+3 < N_1 < 1 + \left(1 - \frac{2\sqrt{10}}{d^\varepsilon}\right)(d+1)\left(\frac{N_0}{d+1}-1\right)^2\Bigg\}. \end{aligned} \tag{76}$$

$$\mathcal{R}_1(\eta) := \left\{2 + \tfrac{1}{d+1} < \frac{N_1}{d+1} < \tfrac{1}{d+1} + (1-\eta)\left(\frac{N_0}{d+1}-1\right)^2\right\}, \tag{77}$$

$$\mathcal{R}_2(\kappa, T) := \left\{2 + \tfrac{1}{d+1} \leq \frac{N_0}{d+1} \leq 1 + T,\ \ \frac{N_1}{d+1} > \max\left\{2 + \tfrac{1}{d+1},\ \mathsf{T}_\kappa\right\}\right\},$$

so that the regions $\mathcal{R}_1, \mathcal{R}_2$ of (76) are $\mathcal{R}_1\big(4\sqrt{c_2}/d^{\varepsilon}\big)$ and $\mathcal{R}_2\Big(\frac{1}{2}, \frac{3\sqrt{c_1}}{20\sqrt{2}}d^{\varepsilon}\Big)$, the latter because $\mathsf{T}_{1/2} = \big(\sqrt{2}\frac{N_0}{d+1} - 1 + \sqrt{2}\big(\frac{N_0}{d+1} - 1\big)^2\big)/(\sqrt{2}-1)$. Combining the equations (69), and (74) in the regime where $N_0 \geq 2d+3$ we have, writing $\tilde{N}_l := N_l/(d+1)$,

$$\frac{\mathsf{Var}(\Delta)}{(\mathbb{E}[\Delta])^2} < \begin{cases} \frac{56542\,\tilde{N}_1}{3125\underline{\kappa}d^2}\left[1 + \frac{10\,\tilde{N}_1}{(\tilde{N}_0-1)^2}\right], & \text{on } \mathcal{R}_1\big(\sqrt{\underline{\kappa}}\big), \\ \frac{113084\,(\tilde{N}_0-1)^2}{225\overline{\kappa}d^2}\left[1 + \frac{(\tilde{N}_0-1)^2}{10\,\tilde{N}_1}\right], & \text{if } \tilde{N}_1 > \max\left\{2 + \frac{1}{d+1}, \mathsf{T}_{\overline{\kappa}}\right\}. \end{cases} \tag{78}$$

It can be easily shown for any $\overline{\delta}, \underline{\delta} \in (0,1]$ we have

$$\frac{\mathsf{Var}(\Delta)}{\mathsf{Var}(\Delta) + (\mathbb{E}[\Delta])^2} < \begin{cases} \underline{\delta}\left[1 + \frac{c_1}{c_2}\left(\frac{\frac{N_1}{d+1}}{\frac{N_0}{d+1}-1}\right)^2\right], & \text{on } \mathcal{R}_1\left(\frac{4}{d}\sqrt{\frac{c_2}{\delta}}\right), \\ \overline{\delta}\left[1 + \frac{c_2}{c_1}\left(\frac{\frac{N_1}{d+1}}{\frac{N_0}{d+1}-1}\right)^{-2}\right], & \text{on } \mathcal{R}_2\Big(\overline{\kappa}, \frac{3}{20}d\sqrt{c_1\overline{\kappa}\overline{\delta}}\Big). \end{cases} \tag{79}$$

For any $\varepsilon \geq \frac{\ln(4\sqrt{c_2})}{\ln(d)}$, we have

$$\frac{\mathsf{Var}(\Delta)}{\mathsf{Var}(\Delta) + (\mathbb{E}[\Delta])^2} < \begin{cases} \frac{1}{d^{2-2\varepsilon}}\left[1 + \frac{c_1}{c_2}\left(\frac{\frac{N_1}{d+1}}{\frac{N_0}{d+1}-1}\right)^2\right], & \text{on } \mathcal{R}_1\left(\frac{4\sqrt{c_2}}{d^{\varepsilon}}\right), \\ \overline{\delta}\left[1 + \frac{c_2}{c_1}\left(\frac{\frac{N_1}{d+1}}{\frac{N_0}{d+1}-1}\right)^{-2}\right], & \text{on } \mathcal{R}_2\Big(\overline{\kappa}, \frac{3}{20}d\sqrt{c_1\overline{\kappa}\overline{\delta}}\Big). \end{cases} \tag{80}$$

We now find the optimal parameter $\overline{\kappa}$ by posing an optimization problem. Specifically, we maximize the ratio of saturation point of sufficient condition and the minimum possible N_0. Similarly, we pose a similar optimization problem for $\overline{\kappa}$, before which we will write equivalent conditions:

$$2 + \frac{1}{d+1} \leq \frac{N_0}{d+1} \leq 1 + \frac{3}{20}d\sqrt{\overline{\kappa}\overline{\delta}}, \quad \frac{N_1}{d+1} > \max\left\{2 + \frac{1}{d+1}, \mathsf{T}_{\overline{\kappa}}\right\} \tag{81}$$

$$\equiv 2 + \frac{1}{d+1} \leq \frac{N_0}{d+1} \leq \min\left\{1 + \frac{3}{20}d\sqrt{c_1\overline{\kappa}\overline{\delta}}, \frac{1}{2}\left[1 + \sqrt{1 + 4(1-\sqrt{\overline{\kappa}})\left(\frac{N_1}{d+1} - 1\right)}\right]\right\}. \tag{82}$$

We have the optimization program

$$\begin{aligned} \overline{\kappa}^* = &\arg\max_{\overline{\kappa}\in(0,1]} \frac{1 + \frac{3}{20}d\sqrt{c_1\overline{\kappa}\overline{\delta}}}{\frac{1}{2}\left[1 + \sqrt{\frac{\left(3 + \frac{2}{d+1}\right)^2 - \sqrt{\overline{\kappa}}}{1-\sqrt{\overline{\kappa}}}}\right]}, \\ &\text{s.t. } \frac{3}{20}d\sqrt{c_1\overline{\kappa}\overline{\delta}} \geq 1 + \frac{1}{d+1}. \end{aligned} \tag{83}$$

Numerically, it is easy to verify that $\lim_{d\to\infty}\overline{\kappa}^* \to 0.50541$. Therefore, we take a good but a sub-optimal choice $\overline{\kappa} = \frac{1}{2}$ with $\overline{\delta} = d^{2+2\varepsilon}$. This gives us a final bound

$$\frac{\mathsf{Var}(\Delta)}{\mathsf{Var}(\Delta) + (\mathbb{E}[\Delta])^2} < \begin{cases} \frac{1}{d^{2-2\varepsilon}}\left[1 + \frac{c_1}{c_2}\left(\frac{\frac{N_1}{d+1}}{\frac{N_0}{d+1}-1}\right)^2\right], & \text{on } \mathcal{R}_1\left(\frac{4\sqrt{c_2}}{d^{\varepsilon}}\right), \\ \frac{1}{d^{2+2\varepsilon}}\left[1 + \frac{c_2}{c_1}\left(\frac{\frac{N_1}{d+1}}{\frac{N_0}{d+1}-1}\right)^{-2}\right], & \text{on } \mathcal{R}_2\Big(\frac{1}{2}, \frac{3\sqrt{c_1}}{20\sqrt{2}}d^{\varepsilon}\Big). \end{cases} \tag{84}$$

Observe that when $\frac{N_1}{d+1} > \max\left\{2 + \frac{1}{d+1},\ \mathsf{T}_{1/2}\right\}$, where by (75) $\mathsf{T}_{1/2} = \frac{\sqrt{2}\,\frac{N_0}{d+1} - 1 + \sqrt{2}\left(\frac{N_0}{d+1} - 1\right)^2}{\sqrt{2} - 1}$, we have that

$$\left(\frac{\frac{N_1}{d+1}}{\frac{N_0}{d+1} - 1}\right)^{-2} \leq \left(\sqrt{2} - 1\right)^4 < \frac{3}{100}. \tag{85}$$

Therefore we have

$$\frac{\mathsf{Var}(\Delta)}{\mathsf{Var}(\Delta) + (\mathbb{E}[\Delta])^2} < \begin{cases} \dfrac{1}{d^{2-2\varepsilon}}\left[1 + \dfrac{c_1}{c_2}\left(\dfrac{\frac{N_1}{d+1}}{\frac{N_0}{d+1} - 1}\right)^2\right], & \text{on } \mathcal{R}_1\left(\dfrac{4\sqrt{c_2}}{d^{\varepsilon}}\right), \\ \dfrac{1}{d^{2+2\varepsilon}}\left[1 + \dfrac{3c_2}{100c_1}\right], & \text{on } \mathcal{R}_2\left(\frac{1}{2},\ \dfrac{3\sqrt{c_1}}{20\sqrt{2}}d^{\varepsilon}\right). \end{cases} \tag{86}$$

The proof is concluding by setting the bounds on c_1 and c_2.

□

G Proof of Theorem 3

Proof. Let $\beta^{(k)}$ be the scaling factor of kth learner from its ML estimate, i.e, $\Sigma^{(k)} = \beta^{(k)}\Sigma_{\text{ML}}^{(k)}$. Recall the difference between ath learner and bth learner (w.l.o.g. assume $b > a$),

$$\begin{aligned} \Delta_{a:b} &= \frac{1}{2}\left[\ln\frac{|\Sigma^{(b)}|}{|\Sigma^{(a)}|} + \langle(\Sigma^{(b)})^{-1} - (\Sigma^{(a)})^{-1}, \Sigma_*\rangle\right], \\ &= \frac{1}{2}\left[d\ln\left(\beta^{(b)}\right) + \ln\frac{|\Sigma_{\text{ML}}^{(b)}|}{|\Sigma^{(a)}|} + \langle\frac{1}{\beta^{(b)}}(\Sigma_{\text{ML}}^{(b)})^{-1} - (\Sigma^{(a)})^{-1}, \Sigma_*\rangle\right], \end{aligned} \tag{87}$$

we now apply expectation to obtain

$$\begin{aligned} \mathbb{E}\left[\Delta_{a:b}\right] = \frac{d}{2}\Bigg[\sum_{j=a+1}^{b}\left(\ln\left(\beta^{(j)}\right) + \ln\left(\frac{2}{N^{(j)}}\right) + \psi_d(N^{(j)})\right) \\ + \prod_{j=0}^{a}\left(\frac{\gamma_j}{\beta^{(j)}}\right)\left[\prod_{j=a+1}^{b}\left(\frac{\gamma_j}{\beta^{(j)}}\right) - 1\right]\Bigg], \end{aligned} \tag{88}$$

where $\gamma_j = \left(1 - \frac{d+1}{N^{(j)}}\right)^{-1}$. Setting $a = k$, and $b = k+1$ we simplify the expression to

$$\mathbb{E}\left[\Delta_{k:k+1}\right] = \frac{d}{2}\left[\ln\left(\frac{2}{N^{(k+1)}}\right) + \psi_d(N^{(k+1)}) + \ln\left(\beta^{(k+1)}\right) + \prod_{j=0}^{k}\left(\frac{\gamma_j}{\beta^{(j)}}\right)\left[\left(\frac{\gamma_{k+1}}{\beta^{(k+1)}}\right) - 1\right]\right]. \tag{89}$$

For generic length chains we obtain the following, we have the bound:

$$\mathbb{E}\left[\Delta_{k:k+1}\right] = \frac{d}{2}\left[\ln\left(\frac{2}{N^{(k+1)} - d - 1}\right) + \psi_d\left(N^{(k+1)}\right) + \phi\left(\frac{\beta^{(k+1)}}{\beta_*^{(k+1)}}\right) - \phi\left(\prod_{l=0}^{k}\left(\frac{\beta^{(l)}}{\gamma_l}\right)\right)\right], \tag{90}$$

where $\beta_*^{(k+1)} = \gamma_{k+1}\prod_{l=0}^{k}\left(\frac{\gamma_l}{\beta^{(l)}}\right)$.

Claim 1. Suppose $\beta^{(k)} = \gamma_k\,\xi_k/\xi_{k-1}$, with the convention $\xi_{-1} := 1$, and fix $\rho \geq 0$. If the ξ_k satisfy

1. $\xi_0 = 1/\gamma_0$ and $\xi_{L-1} = 1$,
2. $(1+\rho)\,\phi(\xi_{k+1}) \leq \phi(\xi_k) - \phi(1/\gamma'_{k+1})$ for $k = 0, \ldots, L-2$,

where $\gamma'_k = \left(1 - \frac{d+1}{h_d^{-1}(N_k)}\right)^{-1}$. Then, $\mathbb{E}\left[\Delta_{k:k+1}\right] \le -\frac{d\rho}{2}\phi(\xi_{k+1})$.

Proof. Write $Q_k := \prod_{l=0}^{k} \left(\beta^{(l)}/\gamma_l\right)$, so that $\beta_*^{(k+1)} = \gamma_{k+1}/Q_k$.

Step 1 (the scaling telescopes). Since $\beta^{(l)}/\gamma_l = \xi_l/\xi_{l-1}$, the product collapses:

$$Q_k \;=\; \prod_{l=0}^{k} \frac{\xi_l}{\xi_{l-1}} \;=\; \frac{\xi_k}{\xi_{-1}} \;=\; \xi_k, \tag{91}$$

so the entire history is carried by the single number ξ_k. Two consequences follow. First, $\beta^{(0)} = \gamma_0\xi_0$, so the left half of point 1 is exactly the unscaled teacher $\beta^{(0)} = 1$. Second,

$$\frac{\beta^{(k+1)}}{\beta_*^{(k+1)}} \;=\; \frac{\gamma_{k+1}\xi_{k+1}/\xi_k}{\gamma_{k+1}/\xi_k} \;=\; \xi_{k+1}. \tag{92}$$

(This is what the ξ parametrization buys over writing $\beta^{(k)} = (1+\varepsilon_k)\beta_*^{(k)}$: there $Q_0 = 1/\gamma_0 \neq 1+\varepsilon_0$, so the $k=0$ step needs a separate hypothesis, whereas here $\xi_0 = 1/\gamma_0$ is merely the initial value of one recursion.)

Step 2 (the role of γ'_k). By the definition of h_d in Eq (3), $\psi_d(h_d(M)) + \ln\left(\frac{2}{h_d(M)-d-1}\right) = \phi\left(\frac{M-d-1}{M}\right) = \phi(1/\gamma(M))$ for every admissible M, where $\gamma(M) := (1 - \frac{d+1}{M})^{-1}$. Taking $M = h_d^{-1}(N^{(k+1)})$, so that $h_d(M) = N^{(k+1)}$ and $\gamma(M) = \gamma'_{k+1}$, gives

$$\ln\left(\frac{2}{N^{(k+1)} - d - 1}\right) + \psi_d\left(N^{(k+1)}\right) \;=\; \phi\left(1/\gamma'_{k+1}\right). \tag{93}$$

Step 3. Substituting Steps 1–2 into the displayed expression for $\mathbb{E}[\Delta_{k:k+1}]$, and using $\beta^{(k+1)}/\beta_*^{(k+1)} = \xi_{k+1}$ and $Q_k = \xi_k$,

$$\mathbb{E}\left[\Delta_{k:k+1}\right] = \frac{d}{2}\left[\phi\left(1/\gamma'_{k+1}\right) + \phi(\xi_{k+1}) - \phi(\xi_k)\right]. \tag{94}$$

Point 2 states exactly $\phi(1/\gamma'_{k+1}) - \phi(\xi_k) \le -(1+\rho)\,\phi(\xi_{k+1})$ for every $k = 0, \ldots, L-2$ — the case $k=0$ needing no separate treatment, since $\phi(\xi_0) = \phi(1/\gamma_0)$ by point 1. Hence

$$\mathbb{E}\left[\Delta_{k:k+1}\right] \le \frac{d}{2}\left[\phi(\xi_{k+1}) - (1+\rho)\,\phi(\xi_{k+1})\right] = -\frac{d\rho}{2}\,\phi(\xi_{k+1}), \tag{95}$$

as claimed. The index on the right-hand side matters: by point 2 the sequence $\phi(\xi_k)$ is non-increasing, so $\phi(\xi_{k+1}) \le \phi(\xi_k)$ and the variant with $\phi(\xi_k)$ in place of $\phi(\xi_{k+1})$ is strictly stronger. That variant is in fact false: substituted into (94) at equality in point 2, it would force $\phi(1/\gamma'_{k+1}) \le -\rho\,\phi(\xi_k) \le 0$, whereas $\phi(1/\gamma'_{k+1}) > 0$ whenever $N^{(k+1)}$ is finite. □

Claim 2. Suppose $N_l = N_1$ for all $l \in \{1, \cdots, L-1\}$, and recall the chaining threshold

$$N_1^*(\rho, N_0) := h_d\left(\frac{d+1}{1 - \phi_-^{-1}\left(\frac{\rho}{(1+\rho)^{L-1}-1}\,\phi\left(\frac{N_0-d-1}{N_0}\right)\right)}\right). \tag{96}$$

If $N_1 \ge N_1^*(\rho, N_0)$, then a sequence $\{\xi_k\}$ satisfying points 1–2 exists.

Proof. Since $N_l = N_1$ for every $l \in \{1, \ldots, L-1\}$, all the γ'_l coincide, and by (93) their common value satisfies

$$\phi(1/\gamma'_l) \;=\; \ln\left(\frac{2}{N_1 - d - 1}\right) + \psi_d(N_1) \;=:\; y(N_1) \;=:\; y. \tag{97}$$

Write $u_k := \phi(\xi_k)$. Point 1 fixes the two endpoints, $u_0 = \phi(1/\gamma_0) = \phi\left(\frac{N_0-d-1}{N_0}\right)$ and $u_{L-1} = \phi(1) = 0$, while point 2 is the single recursion

$$(1+\rho)\,u_{k+1} \;\le\; u_k - y, \qquad k = 0, 1, \ldots, L-2. \tag{98}$$

Because (98) only bounds u_{k+1} from above, the pointwise largest admissible sequence is obtained by taking equality throughout, which unrolls to

$$u_k \;=\; \frac{u_0}{(1+\rho)^k} - \frac{y}{\rho}\left(1-(1+\rho)^{-k}\right), \qquad k=0,1,\dots,L-1, \tag{99}$$

with the convention $\frac{(1+\rho)^m-1}{\rho} = \sum_{i=0}^{m-1}(1+\rho)^i = m$ at $\rho=0$. This sequence is non-increasing, so every u_k is non-negative — as $\phi \geq 0$ requires — if and only if $u_{L-1} \geq 0$, i.e. iff

$$\frac{u_0}{(1+\rho)^{L-1}} \;\geq\; \frac{y}{\rho}\left(1-(1+\rho)^{-(L-1)}\right) \;\Longleftrightarrow\; y \;\leq\; \frac{\rho\,\phi(1/\gamma_0)}{(1+\rho)^{L-1}-1} \;=:\; \tau. \tag{100}$$

The same inequality is precisely what permits the right endpoint $u_{L-1}=0$ of point 1: imposing it, the final constraint of (98) reads $0 \leq u_{L-2} - y$.

In fact the admissible sequence can be written down explicitly. Take

$$\xi_k \;=\; \phi_-^{-1}\!\left(\frac{(1+\rho)^{L-1-k}-1}{(1+\rho)^{L-1}-1}\,\phi\!\left(\frac{N_0-d-1}{N_0}\right)\right), \qquad k=0,1,\dots,L-1, \tag{101}$$

where at $\rho=0$ the ratio is read as its limit $\frac{L-1-k}{L-1}$. Both endpoints of point 1 are then automatic: at $k=0$ the ratio equals 1, so $\phi(\xi_0)=\phi(1/\gamma_0)$ and hence $\xi_0 = 1/\gamma_0$ — both lying in $(0,1]$, the range of ϕ_-^{-1} — which is $\beta^{(0)}=1$; and at $k=L-1$ the ratio vanishes, so $\xi_{L-1}=1$. For point 2, writing $u_k=\phi(\xi_k)$ and $D := (1+\rho)^{L-1}-1$,

$$(1+\rho)\,u_{k+1} = \frac{(1+\rho)^{L-1-k}-(1+\rho)}{D}\,u_0 = u_k - \tau \tag{102}$$

identically, and $\tau \geq y$ is exactly the hypothesis $N_1 \geq N_1^*(\rho,N_0)$, so $(1+\rho)u_{k+1} = u_k - \tau \leq u_k - y$ as required. Note that (101) does not involve N_1: the sample size enters only through the feasibility condition $\tau \geq y$, so one and the same sequence serves every admissible N_1. Equivalently, $\phi(\xi_k)$ decays geometrically at rate $(1+\rho)$ from the teacher's bottleneck $\phi(1/\gamma_0)$ at $k=0$ to 0 at $k=L-1$.

It remains to convert $y(N_1) \leq \tau$ into the stated threshold. The map $N \mapsto y(N)$ is strictly decreasing, so $y(N_1) \leq \tau$ iff $N_1 \geq y^{-1}(\tau)$. By Eq (3), $y(h_d(M)) = \phi(1/\gamma(M))$, so $y^{-1}(\tau) = h_d(M)$ for the M with $\phi(1/\gamma(M)) = \tau$; since $1/\gamma(M) \leq 1$ this is the branch $1/\gamma(M) = \phi_-^{-1}(\tau)$, i.e. $M = \frac{d+1}{1-\phi_-^{-1}(\tau)}$. As $\tau = \frac{\rho}{(1+\rho)^{L-1}-1}\,\phi\!\left(\frac{N_0-d-1}{N_0}\right)$, this is exactly $N_1^*(\rho,N_0)$ of (96). □

Two elementary facts. Claims 3 and 4 use, for $y(N) := \ln\left(\frac{2}{N-d-1}\right) + \psi_d(N)$ and $N > d+1$,

$$\frac{d+1}{2(N-1)} \;<\; y(N) \;<\; \frac{d+1}{2(N-d-1)}, \tag{103}$$

established in §E, together with the elementary sandwich

$$\frac{(x-1)^2}{2x} \;\leq\; \phi(x) \;\leq\; \frac{(x-1)^2}{x}, \qquad x \in (0,1]. \tag{104}$$

For the lower half of (104), the function $g(x) := \ln x - \frac{x}{2} + \frac{1}{2x}$ has $g'(x) = -\frac{(x-1)^2}{2x^2} \leq 0$ and $g(1)=0$, so $g \geq 0$ on $(0,1]$, i.e. $\ln x \geq \frac{x}{2} - \frac{1}{2x}$, whence $\phi(x) \geq \frac{x}{2} + \frac{1}{2x} - 1 = \frac{(x-1)^2}{2x}$; the upper half is $\ln x \leq x-1$. Throughout write $G := \frac{(1+\rho)^{L-1}-1}{\rho} = \sum_{i=0}^{L-2}(1+\rho)^i \geq 1$ and $S := G\,\frac{N_0^2}{d+1}$.

Claim 3 (upper half of the sandwich).

$$\frac{N_1^*(\rho,N_0)}{d+1} \;<\; 2\,G\left[\frac{N_0}{d+1}\right]^2. \tag{105}$$

Proof. Write $\tau := \frac{\rho}{(1+\rho)^{L-1}-1}\,\phi(1/\gamma_0)$ and $M := \frac{d+1}{1-\phi_-^{-1}(\tau)}$, so that $N_1^*(\rho,N_0) = h_d(M)$ and, exactly as in the proof of Claim 2, $y(h_d(M)) = \phi(1/\gamma(M)) = \tau$.

Step 1 (bounding h_d). By the upper half of (103) evaluated at $N = h_d(M)$, where y equals τ, we get $\tau < \frac{d+1}{2(h_d(M)-d-1)}$, i.e.

$$h_d(M) \;<\; (d+1) + \frac{d+1}{2\tau}. \tag{106}$$

Step 2 (bounding $\phi(1/\gamma_0)$ from below). By the lower half of (104) at $x = 1/\gamma_0 = 1 - \frac{d+1}{N_0} \in (0,1)$,

$$\phi(1/\gamma_0) \;\geq\; \frac{\left(\frac{d+1}{N_0}\right)^2}{2\left(1-\frac{d+1}{N_0}\right)} \;\geq\; \frac{(d+1)^2}{2N_0^2}. \tag{107}$$

Step 3 (combining). By Step 2,

$$\frac{d+1}{2\tau} = \frac{d+1}{2}\cdot\frac{G}{\phi(1/\gamma_0)} \;\leq\; \frac{d+1}{2}\cdot G\cdot\frac{2N_0^2}{(d+1)^2} \;=\; S, \tag{108}$$

so (106) gives $h_d(M) < (d+1) + S$. Finally $G \geq 1$ and $N_0 > d+1$ give $\frac{N_0^2}{d+1} > d+1$, hence $S > d+1$ and $h_d(M) < (d+1) + S < 2S$, which is the claim after dividing by $d+1$. □

Claim 4. The matching lower bound holds true:

$$h_d\left(\frac{d+1}{1-\phi_-^{-1}\left(\frac{(\kappa-1)\phi(1/\gamma_0)}{\kappa^{L-1}-1}\right)}\right) \;>\; 1 + \frac{1}{2}\left(\frac{\kappa^{L-1}-1}{\kappa-1}\right)\left(\frac{N_0\,(N_0-d-1)}{d+1}\right). \tag{109}$$

Proof. Keep τ and M as in the proof of Claim 3, so that the left-hand side is $h_d(M)$ and $y(h_d(M)) = \tau$. The argument is the mirror image of Claim 3, with each of the two ingredient inequalities reversed.

Step 1 (bounding h_d from below). By the *lower* half of the two-sided bottleneck bound, $y(N) > \frac{d+1}{2(N-1)}$ for every $N > d+1$. Evaluating at $N = h_d(M)$, where y equals τ, gives $\tau > \frac{d+1}{2(h_d(M)-1)}$, i.e.

$$h_d(M) \;>\; 1 + \frac{d+1}{2\tau}. \tag{110}$$

Step 2 (bounding $\phi(1/\gamma_0)$ from above). Since $\ln x \leq x - 1$ for all $x > 0$,

$$\phi(x) \;=\; \ln x + \frac{1}{x} - 1 \;\leq\; (x-1) + \frac{1}{x} - 1 \;=\; \frac{(x-1)^2}{x}, \qquad x \in (0,1]. \tag{111}$$

Applying this at $x = 1/\gamma_0 = 1 - \frac{d+1}{N_0}$,

$$\phi(1/\gamma_0) \;\leq\; \frac{\left(\frac{d+1}{N_0}\right)^2}{1-\frac{d+1}{N_0}} \;=\; \frac{(d+1)^2}{N_0\,(N_0-d-1)}. \tag{112}$$

Step 3 (combining). Substituting Step 2 into (110),

$$\frac{d+1}{2\tau} = \frac{d+1}{2}\cdot\frac{\kappa^{L-1}-1}{(\kappa-1)\,\phi(1/\gamma_0)} \;\geq\; \frac{d+1}{2}\cdot\frac{\kappa^{L-1}-1}{\kappa-1}\cdot\frac{N_0\,(N_0-d-1)}{(d+1)^2}, \tag{113}$$

which is the claim.

The two bounds together. With $S := \left(\frac{\kappa^{L-1}-1}{\kappa-1}\right)\frac{N_0^2}{d+1}$ as in Claim 3, and since $\frac{N_0(N_0-d-1)}{d+1} = \frac{N_0^2}{(d+1)\,\gamma_0}$, the bound just proved reads $h_d(M) > 1 + \frac{S}{2\gamma_0}$. Hence Claims 3 and 4 together give

$$\frac{S}{2\gamma_0} \;<\; h_d(M) \;<\; 2S. \tag{114}$$

The threshold is therefore pinned to within a factor $4\gamma_0$, which tends to 4 as $N_0/(d+1) \to \infty$. In particular the two bounds have the *same* dependence on κ, L, N_0 and d, so the threshold is $\Theta\Big(\Big(\frac{\kappa^{L-1}-1}{\kappa-1}\Big)\frac{N_0^2}{d+1}\Big)$: the growth rate in Claim 3 is order-correct, not merely an upper estimate. The residual factor of 2 between the two arguments is exactly the gap between the ingredient inequalities $\frac{(x-1)^2}{2x} \le \phi(x) \le \frac{(x-1)^2}{x}$ on $(0,1]$. □

Proof of the theorem. Fix $\rho \ge 0$ and suppose $N_1 = \cdots = N_{L-1} \ge N_1^*(\rho, N_0)$. Claim 2 supplies a sequence $\{\xi_k\}$ satisfying points 1–2; setting $\beta^{(k)} = \gamma_k\, \xi_k/\xi_{k-1}$, Claim 1 gives, for every $k = 0, \dots, L-2$,

$$\mathbb{E}\left[\mathsf{KL}\left(p_* \,||\, p_{\theta_{k+1}}\right) - \mathsf{KL}\left(p_* \,||\, p_{\theta_k}\right)\right] \;=\; \mathbb{E}\left[\Delta_{k:k+1}\right] \;\le\; -\frac{d\rho}{2}\,\phi(\xi_{k+1}) \;\le\; 0, \tag{115}$$

so each learner improves on its predecessor, strictly so whenever $\rho > 0$ and $\xi_{k+1} \neq 1$. The accompanying sandwich bound on $N_1^*(\rho, N_0)$ is the content of Claims 3 and 4. Finally, the two boundary conditions in point 1 are exactly the normalisations of the chain: $\xi_0 = 1/\gamma_0$ is the unscaled teacher $\beta^{(0)} = 1$, and $\xi_{L-1} = 1$ makes the last learner the minimiser $\beta^{(L-1)} = \beta_*^{(L-1)} = \gamma_{L-1}/\xi_{L-2}$ of its own gap. □

H Proof of Theorem 5

The teacher $\mathcal{E}_0$ is a $(1,1)$-scaled LDA estimator applied to N_0 labelled samples from p_{θ^*}; the student $\mathcal{E}_1$ is a $(\overline{\beta}, \underline{\beta})$-scaled LDA estimator applied to N_1 labelled samples $(\mathbf{x}, y) \sim p_{\theta_0}$ generated by the teacher, and $\Delta := \mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_1}\right) - \mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_0}\right)$ is now the gap between the two *joint* laws on $\mathbb{R}^d \times \mathcal{Y}$. The chain rule for the KL divergence separates the label marginal from the class-conditionals, and the shared covariance then collapses the K class-conditional terms into a single covariance term plus K mean terms,

$$\begin{aligned}
\mathsf{KL}\left(p_{\theta^*} \,||\, p_\theta\right) &= \underbrace{\mathsf{KL}\left(\boldsymbol{\pi}^* \,||\, \boldsymbol{\pi}\right)}_{\text{label marginal}} + \sum_{k=1}^{K} \pi_k^*\, \mathsf{KL}\left(\mathcal{N}(\mu_k^*, \Sigma_*) \,||\, \mathcal{N}(\mu_k, \Sigma)\right) \\
&= \mathsf{KL}\left(\boldsymbol{\pi}^* \,||\, \boldsymbol{\pi}\right) + \underbrace{\frac{1}{2}\left[\operatorname{tr}\!\left(\Sigma^{-1}\Sigma_*\right) - d + \ln\frac{\det\Sigma}{\det\Sigma_*}\right]}_{\text{covariance}} \\
&\quad + \underbrace{\frac{1}{2}\sum_{k=1}^{K} \pi_k^*\,(\mu_k - \mu_k^*)^\top \Sigma^{-1}(\mu_k - \mu_k^*)}_{\text{class means}}.
\end{aligned} \tag{116}$$

The covariance term is exactly the object analyzed in §2; two features distinguish the present setting from Theorem 1. First, the class means are no longer known to be zero: centering each class at $\hat{\mu}_k$ costs K Wishart degrees of freedom, so that $N\hat{S} \sim \mathcal{W}_d(\Sigma_*, N-K)$, and every occurrence of $d+1$ in the main theorem is replaced by $d+K+1$. Second, the estimated means contribute the Mahalanobis terms above, which the student inherits *twice* — once from its own finite synthetic sample and once from the teacher's displaced means. Both effects are captured exactly below. We state the result for *stratified, proportion-matched* budgets, i.e. the per-class counts $N_k^{(1)}, N_k^{(2)}$ are fixed by design and the student generates synthetic data in the teacher's class proportions,

$$\frac{N_k^{(2)}}{N_1} = \frac{N_k^{(1)}}{N_0} \qquad \text{for every } k \in [K], \tag{117}$$

so that $\boldsymbol{\pi}^{(2)} = \boldsymbol{\pi}^{(1)}$ and the label term of (116) cancels in Δ; the i.i.d. case is treated in the remarks.

The proof reuses the computation of §D wholesale. The point is that the LDA gap is *the same functional of the covariance estimates* as the zero-mean gap of Theorem 1, once the fixed target Σ_* is replaced by a random *augmented target* that absorbs the class means. Only two ingredients are genuinely new — the degrees of freedom lost to centering, and the mean of that augmented target — and both are isolated in Lemma 1; everything else is quoted from §D.

Throughout, $\mathcal{F}_1 := \sigma\big(\theta^{(1)}\big)$ denotes the σ-algebra generated by the teacher's output, and for $l \in \{1,2\}$ we write $\mu_k^{(l)}$, $\Sigma^{(l)}$ and $\hat{S}^{(l)}$ for the class means, the covariance estimate and the pooled scatter (17) of the lth learner. Define the *augmented targets*

$$M^{(l)} \;:=\; \Sigma_* \;+\; \sum_{k=1}^{K} \pi_k^* \left(\mu_k^{(l)} - \mu_k^*\right)\left(\mu_k^{(l)} - \mu_k^*\right)^\top \;\succeq\; 0, \qquad l \in \{1,2\}, \tag{118}$$

each a positive-definite perturbation of Σ_* by the class-mean errors.

Lemma 1 (Cochran decomposition and augmented targets)**.** *Let $\{(\mathbf{x}_i, y_i)\}_{i\in[N]}$ have deterministic class counts N_k and let $\mathbf{x}_i \mid y_i = k$ be i.i.d. $\mathcal{N}(\mu_k, \Sigma)$.*

(i) The class means and the pooled scatter (17) satisfy

$$\hat{\mu}_k \sim \mathcal{N}\left(\mu_k, \frac{\Sigma}{N_k}\right) \;\; \textit{independently over } k \in [K], \qquad N\hat{S} \sim \mathcal{W}_d(\Sigma, N-K)\,, \tag{119}$$

and $\{\hat{\mu}_k\}_{k\in[K]}$ is independent of $\hat{S}$. Hence if $N \geq d + K + 2$ then, with $\gamma := \big(1 - \frac{d+K+1}{N}\big)^{-1}$,

$$\mathbb{E}\Big[\hat{S}^{-1}\Big] = \gamma\,\Sigma^{-1}, \tag{120}$$

$$\mathbb{E}\Big[\ln\det\hat{S}\Big] = \ln\det\Sigma + d\ln\left(\frac{2}{N}\right) + \sum_{j=1}^{d}\psi\left(\frac{N-K+1-j}{2}\right). \tag{121}$$

(ii) The augmented targets (118) satisfy

$$\mathbb{E}\left[M^{(1)}\right] = \left(1+\rho^{(1)}\right)\Sigma_*, \qquad \mathbb{E}\left[M^{(2)}\,\middle|\,\mathcal{F}_1\right] = M^{(1)} + \rho^{(2)}\,\Sigma_0, \tag{122}$$

and $M^{(l)}$ is independent of $\hat{S}^{(l)}$ (conditionally on $\mathcal{F}_1$ when $l = 2$).

Proof. (i) Within class k, Cochran's theorem gives $\hat{\mu}_k \sim \mathcal{N}(\mu_k, \Sigma/N_k)$ independent of $\sum_{i:y_i=k}(\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^\top \sim \mathcal{W}_d(\Sigma, N_k - 1)$. The K classes are independent because the counts are deterministic, so the K within-class scatters are independent Wisharts with common scale Σ and may be added, $\sum_k (N_k - 1) = N - K$, giving $N\hat{S} \sim \mathcal{W}_d(\Sigma, N-K)$ independent of $\{\hat{\mu}_k\}$. For (120), write $W := N\hat{S}$ and use the moments of an *unnormalized* Wishart $W \sim \mathcal{W}_d(\Sigma, \nu)$ with $\nu := N - K$ degrees of freedom (Lemma 5, stated there for the normalized scatter W/ν),

$$\mathbb{E}\left[W^{-1}\right] = \frac{\Sigma^{-1}}{\nu - d - 1}, \qquad \mathbb{E}\left[\ln\det W\right] = d\ln 2 + \ln\det\Sigma + \sum_{j=1}^{d}\psi\left(\frac{\nu+1-j}{2}\right). \tag{123}$$

The first is finite precisely when $\nu > d+1$, i.e. $N \geq d+K+2$, and gives $\mathbb{E}[\hat{S}^{-1}] = N\,\mathbb{E}[W^{-1}] = \frac{N}{N-K-d-1}\Sigma^{-1} = \gamma\Sigma^{-1}$; the second gives the $\ln\det$ identity after subtracting $d\ln N$.

(ii) By (i) applied to the teacher, $\mu_k^{(1)} - \mu_k^* \sim \mathcal{N}(0, \Sigma_*/N_k^{(1)})$, so $\mathbb{E}\big[(\mu_k^{(1)} - \mu_k^*)(\mu_k^{(1)} - \mu_k^*)^\top\big] = \Sigma_*/N_k^{(1)}$ and, summing against π_k^*, $\mathbb{E}[M^{(1)}] = \Sigma_* + \Sigma_*\sum_k \pi_k^*/N_k^{(1)} = (1+\rho^{(1)})\Sigma_*$. For the student, decompose $\mu_k^{(2)} - \mu_k^* = (\mu_k^{(2)} - \mu_k^{(1)}) + (\mu_k^{(1)} - \mu_k^*)$. Given $\mathcal{F}_1$ the second summand is deterministic and, by (i) applied to the student's sample from p_{θ_0}, the first is centred with covariance $\Sigma_0/N_k^{(2)}$; the cross terms therefore vanish and $\mathbb{E}[M^{(2)} \mid \mathcal{F}_1] = M^{(1)} + \Sigma_0\sum_k \pi_k^*/N_k^{(2)}$, which is (122). The independence claims are those of (i), since $M^{(l)}$ is a function of the class means alone. □

Proof of Theorem 5. **Step 1 (the LDA gap is the gap of Theorem 1 with target $M^{(l)}$).** Under (19) both learners report the same label marginal, $\pi_k^{(2)} = N_k^{(2)}/N_1 = N_k^{(1)}/N_0 = \pi_k^{(1)}$, so the label term of (116) is common to the two divergences and cancels in Δ. In the surviving terms of (116) the $-d$ and $\ln\det\Sigma_*$ contributions cancel as well, and each mean term is the quadratic form

$\pi_k^*(\mu_k^{(l)} - \mu_k^*)^\top(\Sigma^{(l)})^{-1}(\mu_k^{(l)} - \mu_k^*) = \langle(\Sigma^{(l)})^{-1}, \pi_k^*(\mu_k^{(l)} - \mu_k^*)(\mu_k^{(l)} - \mu_k^*)^\top\rangle$. Collecting these rank-one matrices into (118) gives the exact identity

$$\Delta = \frac{1}{2}\left[\ln\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right) + \langle(\Sigma_1)^{-1}, M^{(2)}\rangle - \langle(\Sigma_0)^{-1}, M^{(1)}\rangle\right]. \tag{124}$$

This is *verbatim* the quantity analyzed in §D, with the single fixed target Σ_* there replaced by the two random targets $M^{(1)}, M^{(2)}$ here.

Step 2 (transfer of the argument of §D). The argument of §D applies to (124) without change. Explicitly, it consists of three moves, each of which survives the substitution $\Sigma_* \rightsquigarrow M^{(l)}$:

(M1) *The Loewner sandwich*, which replaces the $(\underline{\beta}, \overline{\beta})$-scaled Σ_1 by a scalar multiple of $\hat{S}^{(2)}$ at the cost of $\frac{d}{2}\ln(\overline{\beta}/\underline{\beta})$. It uses only that $\Sigma_1 \succeq \underline{\beta}\hat{S}^{(2)}$ implies $\langle(\Sigma_1)^{-1}, P\rangle \le \underline{\beta}^{-1}\langle(\hat{S}^{(2)})^{-1}, P\rangle$ for *any* $P \succeq 0$, together with monotonicity of $\det$ on the positive-definite cone. Since $M^{(2)} \succeq 0$ by construction, the same inequality applies here, and it now covers the covariance and the K mean terms simultaneously.

(M2) *The two-stage conditioning* $\mathbb{E}[\,\cdot\,] = \mathbb{E}_{\mathcal{F}_1}\big[\mathbb{E}[\,\cdot \mid \mathcal{F}_1]\big]$, which is unchanged.

(M3) *The Wishart moment evaluations* of Lemma 5 and Proposition 3, applied to $\Sigma_0 = \hat{S}^{(1)}$ and to $\hat{S}^{(2)}$. Here Lemma 1(i) supplies the only modification: the pooled scatter carries $N^{(l)} - K$ degrees of freedom while retaining the normalisation $1/N^{(l)}$, so that in every formula of §D

$$\frac{N^{(l)}}{N^{(l)} - d - 1} \rightsquigarrow \gamma^{(l)}, \qquad \psi\left(\frac{N_1 + 1 - j}{2}\right) \rightsquigarrow \psi\left(\frac{N_1 - K + 1 - j}{2}\right), \tag{125}$$

with $\ln(2/N_1)$ unchanged.

It therefore remains only to evaluate the two inner-product terms of (124), which is where the augmented targets enter. Write $\Psi := \ln\left(\frac{2}{N_1}\right) + \frac{1}{d}\sum_{j=1}^{d}\psi\left(\frac{N_1-K+1-j}{2}\right)$ for the $\ln\det$ shrinkage. By Lemma 1, $M^{(1)}$ is independent of $\hat{S}^{(1)}$, so

$$\begin{aligned}\mathbb{E}\left[\langle(\Sigma_0)^{-1}, M^{(1)}\rangle\right] &= \langle\mathbb{E}\left[\left(\hat{S}^{(1)}\right)^{-1}\right], \mathbb{E}\left[M^{(1)}\right]\rangle, \\ &= \langle\gamma^{(1)}(\Sigma_*)^{-1}, \left(1 + \rho^{(1)}\right)\Sigma_*\rangle, \\ &= d\,\gamma^{(1)}\left(1 + \rho^{(1)}\right) = dc,\end{aligned} \tag{126}$$

which is the interpretation of the constant c of (20): it is the teacher's total expected penalty, covariance and means together. Similarly, taking $\Sigma_1 = \beta\hat{S}^{(2)}$ for the moment and using conditional independence given $\mathcal{F}_1$ followed by (122) and (126),

$$\mathbb{E}\left[\langle(\Sigma_1)^{-1}, M^{(2)}\rangle\right] = \frac{\gamma^{(2)}}{\beta}\,\mathbb{E}\left[\langle(\Sigma_0)^{-1}, M^{(1)} + \rho^{(2)}\Sigma_0\rangle\right] = \frac{\gamma^{(2)}}{\beta}\left[dc + d\rho^{(2)}\right]. \tag{127}$$

Substituting (126), (127) and $\mathbb{E}\big[\ln\det\Sigma_1 - \ln\det\Sigma_0\big] = d\ln\beta + d\Psi$ into (124) gives

$$\mathbb{E}[\Delta] = \frac{d}{2}\left[\ln\beta + \Psi + \frac{\gamma^{(2)}}{\beta}\left(c + \rho^{(2)}\right) - c\right] = \Delta_0^{\mathsf{LDA}}(\beta), \tag{128}$$

which is statement 1; carrying the same substitution through (M1) as in §D gives statement 3, (24). Note that (127) is the only place where the mean terms leave a trace beyond the factor $(1 + \rho^{(1)})$ inside c: they contribute the single extra summand $\gamma^{(2)}\rho^{(2)}/\beta$.

Step 3 (statement 4: optimal scaling). Write $A := \gamma^{(2)}(c + \rho^{(2)}) > 0$ and $u := \ln\beta$, so that $\frac{2}{d}\Delta_0^{\mathsf{LDA}} = u + Ae^{-u} + (\Psi - c)$. The map $u \mapsto u + Ae^{-u}$ has second derivative $Ae^{-u} > 0$, is therefore strictly convex, and is stationary at $e^u = A$; hence $\beta^* = A$ is the unique minimizer. At $\beta = \beta^*$ we have $A/\beta^* = 1$, giving $\Delta_0^{\mathsf{LDA}}(\beta^*) = \frac{d}{2}[\ln\beta^* + \Psi + 1 - c]$. Finally $\gamma^{(2)} > 1$, $\gamma^{(1)} > 1$

and $\rho^{(1)}, \rho^{(2)} > 0$ give $c > 1$ and hence $\beta^* > 1$. For balanced budgets $N_k^{(l)} = N^{(l)}/K$ one has $\rho^{(l)} = \sum_k \pi_k^* K/N^{(l)} = K/N_l$ and $c = \gamma^{(1)}(1 + K/N_0) = (N_0 + K)/(N_0 - d - K - 1)$.

Step 4 (statement 2: the unbiased student loses). This is the one statement that does not follow from §D by substitution, because the sign of $\Delta_0^{\mathsf{LDA}}(1)$ results from a competition between the (negative) $\ln\det$ shrinkage Ψ and the (positive) trace term, and the substitution (125) moves both. Put $m := N_1 - K - d \geq 2$, so that $\gamma^{(2)} = N_1/(m-1)$ and $\gamma^{(2)} - 1 = (d + K + 1)/(m-1)$, and split $\frac{2}{d}\Delta_0^{\mathsf{LDA}}(1) = \Psi + B$ with $B := \gamma^{(2)}(c + \rho^{(2)}) - c = c\left(\gamma^{(2)} - 1\right) + \gamma^{(2)}\rho^{(2)}$. Since $c > 1$, $\gamma^{(2)} > 1$ and $\rho^{(2)} > 0$,

$$B \;>\; \gamma^{(2)} - 1 \;=\; \frac{d + K + 1}{m - 1} \;>\; \frac{d + K + 1}{m + 1}. \tag{129}$$

For Ψ, the digamma arguments are $\frac{N_1 - K + 1 - j}{2} = \frac{m+i}{2}$ with $i := d + 1 - j$ running over $[d]$, and $N_1 = m + d + K$. The lower bound $\psi(x) > \ln x - \frac{1}{2x} - \frac{1}{12x^2}$ of Theorem 6 at $x = \frac{m+i}{2}$ gives $\psi\left(\frac{m+i}{2}\right) > \ln\frac{m+i}{2} - \frac{1}{m+i} - \frac{1}{3(m+i)^2}$, whence

$$\Psi \;>\; \frac{1}{d}\sum_{i=1}^{d}\left[\ln\left(\frac{m+i}{m+d+K}\right) - \frac{1}{m+i} - \frac{1}{3(m+i)^2}\right]. \tag{130}$$

Applying $\ln(1-t) \geq -\frac{t}{1-t}$ with $t = \frac{d+K-i}{N_1}$, so that $1 - t = \frac{m+i}{N_1}$, yields $\ln\frac{m+i}{m+d+K} \geq -\frac{d+K-i}{m+i}$, and therefore, bounding each denominator below by $m + 1$,

$$\Psi \;>\; -\frac{1}{d}\sum_{i=1}^{d}\frac{d + K + 1 - i}{m + i} - \frac{1}{3d}\sum_{i=1}^{d}\frac{1}{(m+i)^2} \;\geq\; -\frac{1}{m+1}\left[\frac{d+1}{2} + K\right] - \frac{1}{3(m+1)^2}, \tag{131}$$

where we used $\sum_{i=1}^{d}(d + K + 1 - i) = d\left[\frac{d+1}{2} + K\right]$. Adding (129) and (131), the K's cancel and

$$\frac{2}{d}\Delta_0^{\mathsf{LDA}}(1) \;>\; \frac{1}{m+1}\left[d + K + 1 - \frac{d+1}{2} - K - \frac{1}{3(m+1)}\right] = \frac{1}{m+1}\left[\frac{d+1}{2} - \frac{1}{3(m+1)}\right]. \tag{132}$$

Since $m \geq 2$ we have $\frac{1}{3(m+1)} \leq \frac{1}{9} \leq \frac{3}{10}(d+1)$, so the bracket is at least $\frac{d+1}{5}$ and

$$\Delta_0^{\mathsf{LDA}}(1) \;>\; \frac{d(d+1)}{10\,(m+1)} = \frac{d(d+1)}{10\,(N_1 - d - K + 1)} \;>\; 0, \tag{133}$$

which is (23). This concludes the proof. □

We derive an improved probability guarantee for the KL gap Δ by bounding its variance and applying Chebyshev's inequality, avoiding the loose concentration-based approach of Theorem 2. First step, in using Chebyshev's inequality is computing the variance of Δ, since we are dealing with covariance matrices of a Gaussian distribution, it is natural that we encounter Wishart distributions. Wishart distributions often require computation of polygamma functions. We will state a result from Chen (2005) which provides tight bounds for polygamma functions.

Theorem 6 (Bounds of polygamma functions from Theorem 1 (Chen, 2005))**.** *Denote the Gamma function by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$, the digamma function by $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$, and the polygamma functions by $\psi^{(n)}(x) = \frac{d^n}{dx^n}\psi(x)$. For $x > 0$, and integer $n \geq 1$, we have*

$$\ln(x) - \frac{1}{2x} - \frac{1}{12x^2} < \psi(x) < \ln(x) - \frac{1}{2x} \tag{134}$$

and

$$\frac{(n-1)!}{x^n} + \frac{n!}{2x^{n+1}} < (-1)^{n+1}\psi^{(n)}(x) < \frac{(n-1)!}{x^n} + \frac{n!}{2x^{n+1}} + \frac{(n+1)!}{12x^{n+2}}. \tag{135}$$

Next, we will compute the variance of Δ which involves computing the variance of KL-divergence taking the form $\ln|\Sigma| + \langle\Sigma^{-1}, A\rangle$, where Σ is the co-variance and A is some fixed matrix. We will state and prove variance of such an object for general use.

Lemma 2 (Variance of general KL-divergence). *Let C be some fixed positive definite matrix in d-dimensions. Suppose $W \sim \mathcal{W}_d(I, N)$. If $N \geq d + 4$, then*

$$\mathsf{Var}\left(\ln(|W| + \langle W^{-1}, C\rangle)\right) = 2\frac{\left(\mathsf{tr}\,(C)^2 + (N-d-1)\,\mathsf{tr}(C^2)\right)}{(N-d)(N-d-3)(N-d-1)^2} \tag{136}$$

$$- 4\frac{\mathsf{tr}(C)}{(N-d-1)^2} + \left[\sum_{i=1}^{d} \psi^{(1)}\left(\frac{N+1-i}{2}\right)\right]. \tag{137}$$

Proof. The crucial step here is to compute the co-variance between the random variable $\ln|W|$ and $\langle W^{-1}, C\rangle$, we have

$$\mathsf{Cov}(\ln|W|, \langle W^{-1}, C\rangle) = \mathbb{E}[\langle \ln|W| \cdot W^{-1}, C\rangle] - \mathbb{E}[\ln|W|] \cdot \mathbb{E}[\langle W^{-1}, C\rangle]. \tag{138}$$

From the properties of Wishart distribution we already have that $\mathbb{E}[\ln|W|] = d\ln(2) + \sum_{i=1}^{d} \psi\left(\frac{N+1-i}{2}\right)$, and $\mathbb{E}[\langle W^{-1}, C\rangle] = \mathsf{tr}(C)/(N-d-1)$. Next we will compute the cross product term, recall that gradient of log of Wishart density is given by

$$\nabla_W \ln p(W) = \frac{N-d-1}{2} W^{-1} - \frac{1}{2} I, \tag{139}$$

therefore we can re-write the cross product term as

$$\mathbb{E}\left[\langle \ln(|W|)W^{-1}, C\rangle\right] = \langle \mathbb{E}\left[\frac{\ln(|W|)}{N-d-1}\left(2\nabla_W \ln p(W) + I\right)\right], C\rangle, \tag{140}$$

$$= \langle 2\mathbb{E}\left[\frac{\ln(|W|)\nabla_W \ln p(W)}{N-d-1}\right] + \mathbb{E}\left[\frac{\ln(|W|)}{N-d-1} I\right], C\rangle, \tag{141}$$

$$= \frac{\mathsf{tr}(C)}{N-d-1}\left(d\ln 2 + \sum_{i=1}^{d} \psi\left(\frac{N+1-i}{2}\right) - \frac{2}{N-d-1}\right), \tag{142}$$

in the last equality we have used Stein's identity. The covariance is given by

$$\mathsf{Cov}(\ln|W|, \langle W^{-1}, C\rangle) = \frac{\mathsf{tr}(C)}{(N-d-1)}\left(d\ln 2 + \sum_{i=1}^{d} \psi\left(\frac{N+1-i}{2}\right) - \frac{2}{N-d-1}\right) \tag{143}$$

$$- \frac{\mathsf{tr}(C)}{(N-d-1)}\left(d\ln 2 + \sum_{i=1}^{d} \psi\left(\frac{N+1-i}{2}\right)\right), \tag{144}$$

$$= -2\frac{\mathsf{tr}(C)}{(N-d-1)^2}. \tag{145}$$

Finally the variance is given by

$$\mathsf{Var}\left(\ln|W| + \langle W^{-1}, C\rangle\right) = \mathsf{Var}\left(\ln|W|\right) + \mathsf{Var}\left(\langle W^{-1}, C\rangle\right) + 2\mathsf{Cov}(\ln|W|, \langle W^{-1}, C\rangle). \tag{146}$$

The individual variance computation is given in Lemma 3, and Lemma 4. This completes the proof. □

Lemma 3 (Variance of Wishart log-determinant). *Suppose $W \sim \mathcal{W}_d(I, N)$. If $N > d$, then*

$$\mathsf{Var}(\ln|W|) = \sum_{i=1}^{d} \psi^{(1)}\left(\frac{N+1-i}{2}\right), \tag{147}$$

where $\psi(x)$ is the digamma function. Moreover, we have the inequality

$$2d\left(\frac{1}{N} + \frac{1}{N^2}\right) < \mathsf{Var}(\ln|W|) < 2d\left(\frac{1}{N+1-d} + \frac{1}{(N+1-d)^2} + \frac{1}{3(N+1-d)^3}\right). \tag{148}$$

Proof. Using the Bartlett decomposition of the Wishart distribution, we can express W as $W = LDL^\top$, where L is a lower-triangular matrix with unit diagonal and D is a diagonal matrix with

independent chi-squared entries. The log-determinant can be written as $\ln|W| = \ln|L| + \ln|D| + \ln|L^\top| = 2\ln|L| + \ln|D|$. Since L has unit diagonal, $\ln|L| = 0$, and thus $\ln|W| = \ln|D|$. The diagonal entries of D are $\chi^2_N, \chi^2_{N-1}, \ldots, \chi^2_{N-d+1}$, so $\ln|W| = \sum_{i=1}^d \ln\chi^2_{N+1-i}$. Since $\mathsf{Var}(\ln\chi^2_k) = \psi^{(1)}\left(\frac{k}{2}\right)$ (χ^2_k is a Gamma random variable with shape parameter $k/2$ and scale parameter 2, and variance of the log of such a Gamma random variable is $\psi^{(1)}\left(\frac{k}{2}\right)$), the variance of $\ln|W|$ is the sum of the variances of the independent terms, giving the exact formula. Finally, we use the result from Theorem 6 that $1/x + 1/2x^2 \le \psi^{(1)}(x) \le 1/x + 1/2x^2 + 1/6x^3$ for $x > 0$. This concludes the proof. □

Lemma 4. *Suppose $S \sim \mathcal{W}_d^{-1}(A, N)$ where $A \in \mathbb{S}^d$, and $N > d + 3$. Then*

$$\mathsf{Var}(\mathsf{tr}(S)) = 2\left(\frac{\mathsf{tr}(A)}{N-d-1}\right)^2 \frac{\left(1 + (N-d-1)\frac{\mathsf{tr}(A^2)}{\mathsf{tr}(A)^2}\right)}{(N-d)(N-d-3)}, \tag{149}$$

$$\mathbb{E}\left[(\mathsf{tr}(S))^2\right] = \left(\frac{\mathsf{tr}(A)}{N-d-1}\right)^2 \left[1 + \frac{2\left(1 + (N-d-1)\frac{\mathsf{tr}(A^2)}{\mathsf{tr}(A)^2}\right)}{(N-d)(N-d-3)}\right], \tag{150}$$

$$\mathbb{E}\left[\mathsf{tr}(S^2)\right] = \frac{(\mathsf{tr}\,A)^2}{(N-d-1)}\left[\frac{1 + (N-d-1)\frac{\mathsf{tr}(A^2)}{(\mathsf{tr}\,A)^2}}{(N-d)(N-d-3)}\right]. \tag{151}$$

Proof. The first claim follows from the second-order moments of the inverse-Wishart satisfy $\mathsf{Cov}(S_{ij}, S_{kl}) = \frac{2A_{ij}A_{kl} + (N-d-1)(A_{ik}A_{jl} + A_{il}A_{kj})}{(N-d)(N-d-1)^2(N-d-3)}$ (S. James Press, 2005). Since $\mathsf{tr}(S) = \sum_{i=1}^d S_{ii}$, we have

$$\mathsf{Var}(\mathsf{tr}(S)) = \sum_{i,j=1}^d \mathsf{Cov}(S_{ii}, S_{jj}) \tag{152}$$

$$= \sum_{i,j=1}^d \frac{2A_{ii}A_{jj} + (N-d-1)(A_{ij}^2 + A_{ji}^2)}{(N-d)(N-d-1)^2(N-d-3)} \tag{153}$$

$$= \frac{2\left[(\mathsf{tr}\,A)^2 + (N-d-1)\,\mathsf{tr}(A^2)\right]}{(N-d-1)^2(N-d)(N-d-3)}. \tag{154}$$

Next, the second moment $\mathbb{E}[(\mathsf{tr}(S))^2]$ can be computed as

$$\mathbb{E}[(\mathsf{tr}(S))^2] = \mathsf{Var}(\mathsf{tr}(S)) + (\mathbb{E}[\mathsf{tr}(S)])^2 \tag{155}$$

$$= \frac{2\left[(\mathsf{tr}\,A)^2 + (N-d-1)\,\mathsf{tr}(A^2)\right]}{(N-d-1)^2(N-d)(N-d-3)} + \left(\frac{\mathsf{tr}\,(A)}{N-d-1}\right)^2. \tag{156}$$

The third equality uses the fact that $\mathbb{E}[S] = \frac{A}{N-d-1}$ for an inverse-Wishart distribution. Finally, the second moment $\mathbb{E}[\mathsf{tr}(S^2)]$ can be computed as

$$\mathbb{E}[\mathsf{tr}(S^2)] = \sum_{i,j=1}^d \mathbb{E}[S_{ij}^2] = \sum_{i,j=1}^d \left(\mathsf{Var}(S_{ij}) + (\mathbb{E}[S_{ij}])^2\right) \tag{157}$$

$$= \sum_{i,j=1}^d \left(\frac{2A_{ij}^2 + (N-d-1)(A_{ii}A_{jj} + A_{ij}^2)}{(N-d)(N-d-1)^2(N-d-3)} + \left(\frac{A_{ij}}{N-d-1}\right)^2\right). \tag{158}$$

Putting the two terms over a common denominator $(N-d-1)^2(N-d)(N-d-3)$ and collecting coefficients:

$$\mathbb{E}[\mathsf{tr}(S^2)] = \frac{\sum_{ij}\left[2A_{ij}^2 + (N-d-1)(A_{ii}A_{jj} + A_{ij}^2) + (N-d)(N-d-3)A_{ij}^2\right]}{(N-d-1)^2(N-d)(N-d-3)} \tag{159}$$

$$= \frac{\mathsf{tr}(A^2)[2 + (N-d-1) + (N-d)(N-d-3)] + (N-d-1)(\mathsf{tr}\,A)^2}{(N-d-1)^2(N-d)(N-d-3)}. \tag{160}$$

Note that $2 + (N-d-1) + (N-d)(N-d-3) = (N-d-1)^2$ since $(N-d)(N-d-3) = (N-d-1)^2 - (N-d-1) - 2$. Hence

$$\mathbb{E}[\mathsf{tr}(S^2)] = \frac{(N-d-1)^2\,\mathsf{tr}(A^2) + (N-d-1)(\mathsf{tr}\,A)^2}{(N-d-1)^2(N-d)(N-d-3)} \tag{161}$$

$$= \frac{(\mathsf{tr}\,A)^2 + (N-d-1)\,\mathsf{tr}(A^2)}{(N-d-1)(N-d)(N-d-3)}. \tag{162}$$

This concludes the proof. □

Proposition 1 (Variance of Δ with scaled student)**.** *Under the setting of Theorem 1 with $\underline{\beta} = \overline{\beta} = \beta > 0$. For any $N_0 \geq d+4$ and $N_1 \geq d+4$ we have*

$$\mathsf{Var}(\Delta) = \frac{1}{4}\left[\sum_{i=1}^{d}\psi^{(1)}\left(\frac{N_1+1-i}{2}\right)\right] - \left(\frac{\beta^*}{\beta}\right)\left(\frac{N_1}{N_1-d-1}\right)\left[\frac{d}{N_1}\right] + \frac{g_1 g_2}{2}\left(\frac{\beta^*}{\beta^2}\right)\left[\frac{d}{N_1}\right]$$
$$+ \frac{g_2}{2}\left(\frac{\beta^*}{\beta}\right)^2\left[\frac{d}{N_1}\right]^2 + \frac{g_1 g_2}{d^2}\left(\frac{\beta^*}{\beta}\right)^2\left[\frac{d}{N_0}\right]\left[\frac{d}{N_1}\right]^2 + \frac{g_1}{2}\left[\frac{\beta^*}{\beta} - \frac{N_0}{N_0-d-1}\right]^2\left[\frac{d}{N_0}\right], \tag{163}$$

where $\beta^* = \frac{N_0 N_1}{(N_0-d-1)(N_1-d-1)}$, $g_1 = \frac{(N_0)(N_0-1)}{(N_0-d)(N_0-d-3)}$ $g_2 = \frac{(N_1)^2}{(N_1-d)(N_1-d-3)}$.
Furthermore, the minimum variance is achieved when

$$\beta = \frac{g_1\left[\frac{d}{N_0}\right] + \frac{g_1 g_2}{\beta^*}\left[\frac{d}{N_1}\right] + g_2\left[\frac{d}{N_1}\right]^2 + \frac{2g_1 g_2}{d^2}\left[\frac{d}{N_0}\right]\left[\frac{d}{N_1}\right]^2}{g_1\left(\frac{N_1-d-1}{N_1}\right)\left[\frac{d}{N_0}\right] + \frac{N_0-d-1}{N_0}\left[\frac{d}{N_1}\right]}. \tag{164}$$

Proof. Recall that $\Delta = \mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_1}\right) - \mathsf{KL}\left(p_{\theta^*} \,||\, p_{\theta_0}\right)$, where p_{θ_0} is the teacher's Gaussian with covariance Σ_0, and p_{θ_1} is the student's Gaussian with covariance Σ_1. Conditioned on Σ_0, only p_{θ_1} is random, so we can decompose the variance of Δ using the law of total variance:

$$\begin{aligned}
\mathsf{Var}(\Delta) &= \mathbb{E}\left[(\Delta - \mathbb{E}[\Delta])^2\right], \\
&= \mathbb{E}\left[(\Delta - \mathbb{E}[\Delta \mid \Sigma_0] + \mathbb{E}[\Delta \mid \Sigma_0] - \mathbb{E}[\Delta])^2\right], \\
&= \mathbb{E}\left[(\Delta - \mathbb{E}[\Delta \mid \Sigma_0])^2\right] + \mathbb{E}\left[(\mathbb{E}[\Delta \mid \Sigma_0] - \mathbb{E}[\Delta])^2\right], \\
&= \mathbb{E}[\mathsf{Var}(\Delta \mid \Sigma_0)] + \mathsf{Var}(\mathbb{E}[\Delta \mid \Sigma_0]).
\end{aligned} \tag{165}$$

We now analyze the two terms separately.
Student variance: $\mathbb{E}[\mathsf{Var}(\Delta \mid \Sigma_0)]$. Recall the closed-form expression for the KL divergence difference is given by

$$\Delta = \frac{1}{2}\left[\ln\frac{|\Sigma_1|}{|\Sigma_0|} + \langle(\Sigma_1)^{-1} - (\Sigma_0)^{-1}, \Sigma_*\rangle\right] \tag{166}$$

We now compute the upper bound for conditional variance given the randomness of the teacher:

$$\begin{aligned}
\mathsf{Var}(\Delta \mid \Sigma_0) &= \frac{1}{4}\mathsf{Var}\left(\ln|N_1\Sigma_1| + \langle(N_1\Sigma_1)^{-1}, N_1\Sigma_*\rangle \mid \Sigma_0\right), \\
&= \frac{1}{4}\mathsf{Var}\Big(\ln|\beta^{-1}N_1\left[\Sigma_0\right]^{-1/2}\Sigma_1\left[\Sigma_0\right]^{-1/2}| \\
&\quad + \langle\left(\beta^{-1}N_1\left[\Sigma_0\right]^{-1/2}\Sigma_1\left[\Sigma_0\right]^{-1/2}\right)^{-1}, \beta^{-1}N_1\left[\Sigma_0\right]^{-1/2}\Sigma_*\left[\Sigma_0\right]^{-1/2}\rangle \mid \Sigma_0\Big).
\end{aligned} \tag{167}$$

Observe that $\beta^{-1}N_1\left[\Sigma_0\right]^{-1/2}\Sigma_1\left[\Sigma_0\right]^{-1/2} \sim \mathcal{W}_d(I, N_1)$, making it obvious to invoke Lemma 2 giving us

$$\begin{aligned}
\mathsf{Var}(\Delta \mid \Sigma_0) = \left(\frac{N_1}{\beta}\right)^2 &\frac{\left(\left(\mathsf{tr}((\tilde{\Sigma}^{(1)})^{-1})\right)^2 + (N_1-d-1)\,\mathsf{tr}((\tilde{\Sigma}^{(1)})^{-2})\right)}{2\,(N_1-d-1)^2\,(N_1-d)(N_1-d-3)} \\
&- \left(\frac{N_1}{\beta}\right)\frac{\mathsf{tr}((\tilde{\Sigma}^{(1)})^{-1})}{(N_1-d-1)^2} + \frac{1}{4}\left[\sum_{i=1}^{d}\psi^{(1)}\left(\frac{N_1+1-i}{2}\right)\right],
\end{aligned} \tag{168}$$

where $\tilde{\Sigma}^{(1)} := \sqrt{\Sigma_0}\,(\Sigma_*)^{-1}\sqrt{\Sigma_0}$. Now we apply expectation over the randomness of Σ_0, this gives us

$$\mathbb{E}\left[\mathsf{Var}(\Delta \mid \Sigma_0)\right] = \beta^{-2}\left(\frac{N_1}{N_1-d-1}\right)^2 \frac{\mathbb{E}\left[\left(\mathsf{tr}((\tilde{\Sigma}^{(1)})^{-1})\right)^2\right] + (N_1-d-1)\,\mathbb{E}\left[\mathsf{tr}((\tilde{\Sigma}^{(1)})^{-2})\right]}{2(N_1-d)(N_1-d-3)} - \beta^{-1}\left(\frac{N_1}{N_1-d-1}\right)\frac{\mathbb{E}\left[\mathsf{tr}((\tilde{\Sigma}^{(1)})^{-1})\right]}{N_1-d-1} + \frac{1}{4}\left[\sum_{i=1}^{d}\psi^{(1)}\left(\frac{N_1+1-i}{2}\right)\right], \tag{169}$$

we now invoke Lemma 4 to obtain the expected values, that is,

$$\mathbb{E}\left[\mathsf{tr}((\tilde{\Sigma}^{(1)})^{-1})\right] = \frac{N_0}{N_0-d-1}d, \tag{170}$$

$$\mathbb{E}\left[\left(\mathsf{tr}((\tilde{\Sigma}^{(1)})^{-1})\right)^2\right] = \left(\frac{N_0}{N_0-d-1}\right)^2\left[1+\frac{2(N_0-1)}{d(N_0-d)(N_0-d-3)}\right]d^2, \tag{171}$$

$$\mathbb{E}\left[\mathsf{tr}((\tilde{\Sigma}^{(1)})^{-2})\right] = \left(\frac{N_0}{N_0-d-1}\right)^2\frac{(N_0-1)(N_0-d-1)}{(N_0-d)(N_0-d-3)}d. \tag{172}$$

Substituting the earlier values, we obtain

$$\mathbb{E}\left[\mathsf{Var}(\Delta \mid \Sigma_0)\right] = d\left(\frac{\beta^*}{\beta}\right)^2\frac{d+(2+(N_1-d-1)(N_0-d-1))\,\frac{(N_0-1)}{(N_0-d)(N_0-d-3)}}{2(N_1-d)(N_1-d-3)} - \left(\frac{\beta^*}{\beta}\right)\frac{d}{N_1-d-1} + \frac{1}{4}\left[\sum_{i=1}^{d}\psi^{(1)}\left(\frac{N_1+1-i}{2}\right)\right], \tag{173}$$

where $\beta^* = \left(\frac{N_0}{N_0-d-1}\right)\left(\left(\frac{N_1}{N_1-d-1}\right)\right)$. We can-rewrite the earlier as

$$\begin{aligned}\mathbb{E}\left[\mathsf{Var}(\Delta \mid \Sigma_0)\right] = {} & \frac{1}{4}\left[\sum_{i=1}^{d}\psi^{(1)}\left(\frac{N_1+1-i}{2}\right)\right] \\ & - \left(\frac{\beta^*}{\beta}\right)\left(\frac{N_1}{N_1-d-1}\right)\left[\frac{d}{N_1}\right] \\ & + \frac{1}{2}\left(\frac{\beta^*}{\beta}\right)^2\left(\frac{N_1(N_1-d-1)}{(N_1-d)(N_1-d-3)}\right)\left(\frac{(N_0-1)(N_0-d-1)}{(N_0-d)(N_0-d-3)}\right)\left[\frac{d}{N_1}\right] \\ & + \frac{1}{d^2}\left(\frac{\beta^*}{\beta}\right)^2\frac{(N_1)^2}{(N_1-d)(N_1-d-3)}\frac{N_0(N_0-1)}{(N_0-d)(N_0-d-3)}\left[\frac{d}{N_0}\right]\left[\frac{d}{N_1}\right]^2 \\ & + \frac{1}{2}\left(\frac{\beta^*}{\beta}\right)^2\frac{(N_1)^2}{(N_1-d)(N_1-d-3)}\left[\frac{d}{N_1}\right]^2.\end{aligned} \tag{174}$$

Teacher variance: $\mathsf{Var}(\mathbb{E}[\Delta \mid \Sigma_0])$**.** Conditional mean $\mathbb{E}[\Delta \mid \Sigma_0]$ is given by the difference of the KL divergences computed by

$$\begin{aligned}\mathbb{E}[\Delta \mid \Sigma_0] &= \frac{1}{2}\mathbb{E}\left[\ln\frac{|\Sigma_1|}{|\Sigma_0|} + \langle(\Sigma_1)^{-1}-(\Sigma_0)^{-1},\Sigma_*\rangle \mid \Sigma_0\right], \\ &= \frac{1}{2}\left[d\ln\left(\frac{2}{N_1}\right) + d\psi_d(N_1) + d\ln(\beta) + \left[\frac{N_1}{N_1-d-1}\beta^{-1}-1\right]\langle(\Sigma_0)^{-1},\Sigma_*\rangle\right], \\ &= \frac{1}{2}\left[d\ln\left(\frac{2}{N_1}\right) + d\psi_d(N_1) + d\ln(\beta)\right. \\ &\quad \left. + \left[\frac{N_1}{N_1-d-1}\beta^{-1}-1\right]\mathsf{tr}\left(\sqrt{\Sigma_*}\,(\Sigma_0)^{-1}\sqrt{\Sigma_*}\right)\right].\end{aligned} \tag{175}$$

Recall that $\sqrt{\Sigma_*}\,(N_0\Sigma_0)^{-1}\sqrt{\Sigma_*} \sim \mathcal{W}_d^{-1}(I, N_0)$, so we have

$$\begin{aligned}\mathsf{Var}(\mathbb{E}[\Delta \mid \Sigma_0]) &= \frac{1}{4}\left[\frac{N_1}{N_1-d-1}\beta^{-1}-1\right]^2 (N_0)^2\mathsf{Var}\left(\mathsf{tr}\left(\sqrt{\Sigma_*}\,(N_0\Sigma_0)^{-1}\sqrt{\Sigma_*}\right)\right), \\ &= \frac{1}{2}\left[\frac{\beta^*}{\beta}-\frac{N_0}{N_0-d-1}\right]^2\left(\frac{N_0(N_0-1)}{(N_0-d)(N_0-d-3)}\right)\left[\frac{d}{N_0}\right].\end{aligned} \tag{176}$$

The last line follows from Lemma 4 applied to the inverse-Wishart distribution of $\sqrt{\Sigma_*}\,(N_0\Sigma_0)^{-1}\sqrt{\Sigma_*}$.

Now we have the student variance and teacher variance bounds, we can combine them to get the final variance bound for Δ, giving us:

$$\begin{aligned}\mathsf{Var}(\Delta) = \frac{1}{4}&\left[\sum_{i=1}^{d}\psi^{(1)}\left(\frac{N_1+1-i}{2}\right)\right] \\ &-\left(\frac{\beta^*}{\beta}\right)\left(\frac{N_1}{N_1-d-1}\right)\left[\frac{d}{N_1}\right] \\ &+\frac{1}{2}\left(\frac{\beta^*}{\beta}\right)^2\left(\frac{N_1(N_1-d-1)}{(N_1-d)(N_1-d-3)}\right)\left(\frac{(N_0-1)(N_0-d-1)}{(N_0-d)(N_0-d-3)}\right)\left[\frac{d}{N_1}\right] \\ &+\frac{1}{2}\left(\frac{\beta^*}{\beta}\right)^2\frac{(N_1)^2}{(N_1-d)(N_1-d-3)}\left[\frac{d}{N_1}\right]^2 \\ &+\frac{1}{d^2}\left(\frac{\beta^*}{\beta}\right)^2\frac{(N_1)^2}{(N_1-d)(N_1-d-3)}\frac{N_0(N_0-1)}{(N_0-d)(N_0-d-3)}\left[\frac{d}{N_0}\right]\left[\frac{d}{N_1}\right]^2 \\ &+\frac{1}{2}\left[\frac{\beta^*}{\beta}-\frac{N_0}{N_0-d-1}\right]^2\left(\frac{N_0(N_0-1)}{(N_0-d)(N_0-d-3)}\right)\left[\frac{d}{N_0}\right].\end{aligned} \tag{177}$$

This completes the proof. □

I MISCELLANEOUS

Proposition 2. *The KL-divergence between two n-dimensional Gaussians $p_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $p_2 = \mathcal{N}(\mu_2, \Sigma_2)$ is given by:*

$$\mathsf{KL}\left(p_1 \,||\, p_2\right) = \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} - n + \langle\Sigma_2^{-1}, \Sigma_1 + (\mu_2-\mu_1)(\mu_2-\mu_1)^\top\rangle\right). \tag{178}$$

Lemma 5. *Suppose $S \sim \mathcal{W}_p(\Sigma/N, N)$ represents a random variable which follows Wishart distribution formed via empirical covariance N i.i.d. samples of p dimensional zero Gaussian with covariance Σ. The following properties hold true:*

- $\mathbb{E}[\ln(|S|)] = p\ln 2 + \ln|\Sigma| + \sum_{i\in[p]}\psi\left(\frac{N+1}{2} - \frac{i}{2}\right)$.
- *If $C \in \mathbb{R}^{q\times p}$ is a rank q matrix then $CSC^\top \sim \mathcal{W}_q(C\Sigma C^\top, N)$.*
- *If $N > p+1$, then $\mathbb{E}[S^{-1}] = \frac{N}{N-p-1}\Sigma^{-1}$.*

Proposition 3. *Let Σ is non-singular in p dimensions. Let $\mathcal{N}(0, C\hat{\Sigma}_N C^\top)$ be Gaussian where $\hat{\Sigma}_N$ is empirical covariance matrix formed from N i.i.d. samples from $\mathcal{N}(0,\Sigma)$, and $C \in \mathbb{R}^{p\times p}$ is a full row rank matrix. For some fixed $\Sigma' \in \mathbb{S}^p_{++}$, the following holds*

$$\mathsf{KL}\left(\mathcal{N}(0,\Sigma') \,||\, \mathcal{N}(0, C\hat{\Sigma}_N C^\top)\right) = \frac{1}{2}\left[\ln\left(\frac{|C\hat{\Sigma}_N C^\top|}{|\Sigma'|}\right) - p + \langle\left(C\hat{\Sigma}_N C^\top\right)^{-1}, \Sigma'\rangle\right]. \tag{179}$$

Moreover, taking expectation over all the randomness of sampling give us

$$\mathbb{E}\left[\mathsf{KL}\left(\mathcal{N}(0,\Sigma')\,||\,\mathcal{N}(0,C\hat{\Sigma}_N C^\top)\right)\right] = \frac{1}{2}\left[p\ln(2/N) + \sum_{i\in[p]}\psi\left(\frac{N+1}{2}-\frac{i}{2}\right) + \ln\left(\frac{|C\Sigma C^\top|}{|\Sigma'|}\right) - p + \frac{N}{N-p-1}\langle\left(C\Sigma C^\top\right)^{-1}, \Sigma'\rangle\right]. \tag{180}$$

Remark. The result in Eq (180) matches closely with existing analysis of § 3 from Bongiorno & Lamrani (2025).

Proof. The KL divergence between two Gaussian is given by

$$\mathsf{KL}\left(p\,||\,p'\right) = \frac{1}{2}\left[\ln\left(\frac{|NC\hat{\Sigma}_N C^\top|}{|\Sigma'|}\right) - p - p\ln(N) + N\langle\left(NC\hat{\Sigma}_N C^\top\right)^{-1}, \Sigma'\rangle\right]. \tag{181}$$

Now we apply expectation on both the sides and use the properties from Lemma 5, this gives us

$$\mathbb{E}\left[\mathsf{KL}\left(p\,||\,p'\right)\right] = \frac{1}{2}\left[p\ln(2/N) + \sum_{i\in[p]}\psi\left(\frac{N+1}{2}-\frac{i}{2}\right) + \ln\left(\frac{|C\Sigma C^\top|}{|\Sigma'|}\right) - p + \frac{N}{N-p-1}\langle\left(C\Sigma C^\top\right)^{-1}, \Sigma'\rangle\right]. \tag{182}$$

This completes the proof. □