MODEL COLLAPSE IS AVOIDABLE EVEN AS REAL DATA VANISHES: A SHARP CRITERION FOR ITERATIVE GAUSSIAN ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large generative models are commonly trained on data collected from the open web, which may contain data produced by previous generations of the same model. Previous work has shown that recursively training on model-generated data can cause a progressive degradation in model performance, a phenomenon known as *model collapse*. Yet, currently, generative models continue to improve dramatically over time. Existing theoretical research suggests that this is enabled by the availability of new real data at each iteration, kept separate from model-generated data. However, exact separation is not realistic, and uncurated data is still commonly used. We bridge this gap by considering a less restrictive and more realistic setup in which real and generated data are not handled differently during training. Specifically, we present a comprehensive theoretical study of iterative estimation of the mean and covariance of a multivariate Gaussian data distribution. At each iteration, samples are drawn from a mixture of the real data distribution and the estimated distribution from the previous iteration, but are used indistinguishably. We provide a single equivalent condition for consistency of the mean and covariance estimators in Euclidean norms, and convergence to zero of the expected squared 2-Wasserstein distance between the estimated and true distributions. Using this result, we then show that consistency holds, i.e., model collapse is avoided, for any fixed real data fraction above zero, and even for a fraction that vanishes across iterations, provided it does not vanish too quickly. However, more real data (or data curation) remains important at practical finite times: we show that the convergence rate of the mean estimator deteriorates as the real data fraction approaches zero. We support our results with an empirical investigation of the theoretical setup, and complement our theoretical work with experiments on iterative training of a diffusion model on CIFAR-10 that correlate with our theory.

1 INTRODUCTION

A growing body of empirical and theoretical work has raised the concern that machine learning generative models trained iteratively on data produced by earlier generations of themselves — rather than on real, independently sourced data — can drift away from the true underlying distribution over successive generations, a phenomenon often called *model collapse* (Shumailov et al., 2024; Alemohammad et al., 2024). As more of the data on which models are trained is itself generated by other models, understanding the conditions under which this recursive process is destructive to the model and how we can deal with that has become an important theoretical question (Bertrand et al., 2024; Gerstgrasser et al., 2024; Kazdan et al., 2025; Seddik et al., 2024; Dohmatob et al., 2025; 2024; Barzilai & Shamir, 2025; He et al., 2026; Garg et al., 2026).

There are several approaches to iterative retraining, two of which are commonly referred to as *Accumulate* and *Replace*. In the *Accumulate* method, we start with a set of real data, and at each iteration, we augment the training set with newly generated data while retaining all data from previous iterations. In contrast, in the *Replace* method, the training set at each iteration consists only of the newly generated data from the previous iteration; all data from earlier iterations are discarded. The Accumulate method is generally considered a more realistic setting for iterative retraining and has been shown to exhibit degradation that remains bounded across iterations (Gerstgrasser et al., 2024;

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.

Kazdan et al., 2025; Seddik et al., 2024). In contrast, the Replace method can lead to unbounded degradation, with the learned distribution progressively diverging from the original data distribution (Shumailov et al., 2024; Kazdan et al., 2025).

Existing literature tries to theoretically explain the dramatic improvement of current top-tier generative models. This is done by considering the existence of independent real data in each iteration. However, it has only been investigated under the assumption that the real data can be perfectly separated from the generated data (Alemohammad et al., 2024; He et al., 2026; Garg et al., 2026). In this case, an optimal weighting scheme for the separate real and generated data used for training has been studied (He et al., 2026). However, exact separation is not realistic (Sadasivan et al., 2025; Li et al., 2025), and uncurated data is still commonly used. In this paper, to the best of our knowledge, we are the first to bridge this gap by considering a less restrictive and more realistic setup in which real and generated data are not handled differently during training.

We study *Accumulate*-based iterative estimation of the mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ of a Gaussian data distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a standard setting in theoretical works on model collapse (Shumailov et al., 2024; Alemohammad et al., 2024; Kazdan et al., 2025; He et al., 2026). At each round t, a batch of new samples is drawn from a *mixture* of (i) the current fitted Gaussian model and (ii) the true distribution, and all data accumulated so far is used to refit the mean and covariance by standard empirical estimators, $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$, but *without distinguishing* between real and generated samples. This setting strips away the complications of neural networks while retaining the essential recursive feedback loop and admits rigorous analysis.

Our guiding question is: under what conditions does this process *not* collapse. In other words, when do the iterated estimators $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ converge to the true $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and when do they instead settle on a biased, degenerate limit? We answer this question completely for both estimators. For the mean estimator, we also analyze the rate of convergence. Here, the *mixing fraction* $\alpha_t \in [0, 1]$, the expected fraction of data in the new batch drawn from the current model, plays a decisive role, and we provide an explicit, checkable condition on how it may vary across rounds while still guaranteeing consistency. That separates collapse from non-collapse. Our main contributions include:

- We prove that the mean estimator $\boldsymbol{\mu}^{(t)}$ is unbiased, while the covariance estimator $\boldsymbol{\Sigma}^{(t)}$ is biased, with the exact bias given by $-\operatorname{Cov}(\boldsymbol{\mu}^{(t)})$.
- We provide a closed-form equivalent condition for the consistency of the estimated mean in L^2, the consistency of the covariance estimator in expected Frobenius norm, and the convergence to zero of the expected squared 2-Wasserstein distance between the estimated and true distributions.
- We use our result to show that several scenarios that naturally arise when scraping the web for data are sufficient to guarantee convergence, thereby preventing model collapse. First, convergence holds for an arbitrarily small but fixed positive fraction of real data $(1 - \alpha)$ at every iteration. Second, convergence still holds when the fraction of real data vanishes across iterations, provided that it does not vanish too quickly.
- For a fixed synthetic fraction $0 \leq \alpha < 1$ (i.e., when not all data is generated), we derive the exact rate of convergence (up to constant factors) of the mean estimator to the true mean. Specifically, we prove that if less than half of each new batch is generated, then the rate of convergence matches the optimal rate. However, when more than half of each new batch is generated, the rate of convergence becomes arbitrarily slow as the fraction of synthetic data increases. This result motivates the ongoing research on the detection and curation of generated samples: it shows that for practical finite t in the high-synthetic regime, synthetic data detection can substantially improve estimation accuracy.
- We present an empirical investigation of the theoretical model that supports our theoretical results and provides further insight into the rate of convergence.
- We complement our theoretical work with an empirical study of how different fixed real-data ratios affect the FID of a diffusion model iteratively trained on CIFAR-10, showing results that correlate with our theory.

2 Related work

Existing works have shown that under the *Replace* workflow, iterative retraining degrades without bound (Shumailov et al., 2024; Dohmatob et al., 2024; Kazdan et al., 2025), whereas under the

Accumulate workflow the degradation remains bounded (Gerstgrasser et al., 2024; Kazdan et al., 2025; Seddik et al., 2024).

Several works show that collapse can be avoided, but under restrictive conditions. Bertrand et al. (2024) prove the stability of iterative retraining on a mixture of real and synthetic data, but only locally, when the initial model is sufficiently close to the data distribution and the proportion of real data is sufficiently large. Barzilai & Shamir (2025) provide non-asymptotic guarantees under data accumulation; however, for a fixed sample size n, their bound holds only up to a finite number of iterations t that depends on n, and thus does not cover the regime of fixed n with $t \to \infty$. Moreover, beyond an initial (or fixed) real dataset, these works do not incorporate new real data at subsequent iterations.

In practice, data scraped from the web at each iteration contains new real data alongside generated data. Several works study this setting (Alemohammad et al., 2024; He et al., 2026; Garg et al., 2026), but either discard earlier data or assume that real and generated samples can be told apart, e.g., in order to weight them separately. The latter is unrealistic, since reliably detecting generated content remains an open problem (Sadasivan et al., 2025; Li et al., 2025), although detection, when possible, can substantially mitigate collapse (Drayson et al., 2025). To the best of our knowledge, no prior work analyzes iterative retraining with accumulated data in which fresh real samples are not distinguished from generated ones.

3 THEORY

In this section, we formalize the iterative retraining process, define the corresponding mean and covariance estimators, and state our main theoretical results, which characterize when the estimators converge to the true parameters and the rate at which this convergence occurs. Finally, we validate our results through experiments and provide additional insights into the rate of decay.

3.1 PROBLEM SETUP: DATA GENERATING PROCESS AND ESTIMATORS

In the theoretical model, we consider the accumulate method, where the process is initialized with a dataset of real samples. Similar to a vast body of theoretical work on model collapse (Shumailov et al., 2024; Alemohammad et al., 2024; Kazdan et al., 2025; Dohmatob et al., 2024; He et al., 2026; Dohmatob et al., 2025), we assume that the real data distribution is a multivariate Gaussian distribution. At each iteration, the training dataset is augmented with both generated and real data. Since the model does not know which samples are real and which are generated, it treats all samples identically and estimates the underlying distribution using the entire accumulated dataset. Formally, at each iteration, each newly added sample is drawn from the real distribution with some probability and from the current estimated distribution with the complementary probability.

Data-generating process. Let $\boldsymbol{\mu} \in \mathbb{R}^d$, and let $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ be symmetric positive-definite. The process unfolds in discrete iterations $t = 1, 2, 3, \ldots$

Iteration 1 (real data). Observe $n_r \geq 1$ i.i.d. samples $\{\boldsymbol{x}_j^{(0)}\}_{j=1}^{n_r} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and form the empirical mean and covariance

$$\boldsymbol{\mu}^{(1)} = \frac{1}{n_r}\sum_{j=1}^{n_r} \boldsymbol{x}_j^{(0)}, \qquad \boldsymbol{\Sigma}^{(1)} = \frac{1}{n_r}\sum_{j=1}^{n_r} \big(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}^{(1)}\big)\big(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}^{(1)}\big)^\top. \tag{1}$$

Iteration $t \geq 2$ ***(mixed data).*** Given $\boldsymbol{\mu}^{(t-1)}$ and $\boldsymbol{\Sigma}^{(t-1)}$, draw $n_b(t-1) \geq 1$ i.i.d. samples from the mixture

$$\boldsymbol{x}_i^{(t-1)} \sim \alpha_{t-1}\mathcal{N}\big(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)}\big) + (1-\alpha_{t-1})\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \alpha_{t-1} \in [0,1], \tag{2}$$

where $1 \leq i \leq n_b(t-1)$. The parameter $\alpha_{t-1} \in [0,1]$ controls the fraction of the new batch that is drawn from the current model. Specifically, $\alpha_{t-1} = 1$ means the new batch is entirely synthetic, while $\alpha_{t-1} = 0$ means it is entirely real.

Denote the total number of samples at iteration t by N_t and the inverse ratio of the growth by ρ_t:

$$N_t := n_r + \sum_{s=1}^{t-1} n_b(s), \qquad \rho_t := \frac{N_t}{N_{t+1}} \qquad (t \geq 1). \tag{3}$$

Note that $\rho_t \in (0, 1)$ for every $t \geq 1$.

Estimators. At every round, the model is refit using the pooled empirical mean and covariance of all samples accumulated so far. For every $t \geq 1$, we write $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{N_t}$ for the N_t accumulated points, consisting of the original real batch $\{\boldsymbol{x}_j^{(0)}\}_{j=1}^{n_r}$ and the accumulated batches $\{\boldsymbol{x}_i^{(s)}\}_{i=1}^{n_b(s)}$, $1 \leq s \leq t-1$. Define the pooled empirical estimators:

$$\boldsymbol{\mu}^{(t)} \coloneqq \frac{1}{N_t} \sum_{k=1}^{N_t} \boldsymbol{y}_k, \qquad \boldsymbol{\Sigma}^{(t)} \coloneqq \frac{1}{N_t} \sum_{k=1}^{N_t} \big(\boldsymbol{y}_k - \boldsymbol{\mu}^{(t)}\big)\big(\boldsymbol{y}_k - \boldsymbol{\mu}^{(t)}\big)^\top, \tag{4}$$

which for $t = 1$ coincides with (1). These iterative estimators are standard and widely used in the literature (Shumailov et al., 2024; Kazdan et al., 2025). Conveniently, they admit online recursions in terms of the previous estimates and the newly drawn batch. Finally, we write

$$\mathcal{D}_t \coloneqq \mathcal{N}\big(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\big), \qquad \mathcal{D}_{\text{real}} \coloneqq \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{5}$$

for the estimated distribution at iteration t and the true data distribution, respectively. We measure the discrepancy between them by the 2-Wasserstein distance, $W_2(\cdot, \cdot)$, defined in Appendix A.4.4, which has also been used by Shumailov et al. (2024). For $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$ themselves, we measure the discrepancy from the true values by the Euclidean norm, $\|\cdot\|_2$, and the Frobenius norm, $\|\cdot\|_F$.

Finally, we state our standing assumption on how new samples are generated.

Assumption 1 (Conditional independence)**.** *For every $t \geq 2$, **conditionally** on $(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)})$, the new samples $\{\boldsymbol{x}_i^{(t-1)}\}_{i=1}^{n_b(t-1)}$ are drawn i.i.d. from the mixture* (2) *and are independent of all previously drawn samples $\{\boldsymbol{x}_i^{(s)}\}$, $0 \leq s \leq t-2$.*

In words, the past influences a new batch only through the current fitted model. This is not a restriction on the setting, but a formalization of the data-generating process described above.

3.2 Theoretical results

In this section, we present our key theoretical results. In Section 3.2.1, we establish that the mean estimator is unbiased and derive an explicit formula for the bias of the covariance estimator. In Section 3.2.2, we provide an equivalent (if and only if) condition for the convergence of the estimated parameters to the truth, and hence for the convergence of the estimated distribution to the true distribution. We then show that the model converges when the fraction of real data is fixed and arbitrarily small, as well as when the fraction of real data tends to zero sufficiently slowly. In Section 3.2.3, we analyze the convergence rate of the mean estimator.

The proofs for all the claims are deferred to Appendix A.

3.2.1 Bias of estimators and covariance of the mean estimator

Here, we derive the bias of each estimator and the covariance of the mean estimator. First, let us denote by $\boldsymbol{C}_t$ the covariance of the mean estimator

$$\boldsymbol{C}_t \coloneqq \mathrm{Cov}(\boldsymbol{\mu}^{(t)}) = \mathbb{E}\left[\big(\boldsymbol{\mu}^{(t)} - \mathbb{E}[\boldsymbol{\mu}^{(t)}]\big)\big(\boldsymbol{\mu}^{(t)} - \mathbb{E}[\boldsymbol{\mu}^{(t)}]\big)^\top\right], \tag{6}$$

and by $\boldsymbol{\delta}^{(t)}$ the deviation of the mean estimator from the truth:

$$\boldsymbol{\delta}^{(t)} \coloneqq \boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}. \tag{7}$$

We now establish the unbiasedness of the mean estimator and the bias of the covariance estimator.

Theorem 3.1 (Bias of estimators)**.** *Under the setup of Section 3.1, for all $t \geq 1$*

(a) The mean estimator is unbiased:

$$\mathbb{E}\big[\boldsymbol{\mu}^{(t)}\big] = \boldsymbol{\mu}. \tag{8}$$

(b) The bias of the covariance estimator is:

$$\mathbb{E}\big[\boldsymbol{\Sigma}^{(t)}\big] = \boldsymbol{\Sigma} - \boldsymbol{C}_t. \tag{9}$$

Proof idea. Part (a) follows by induction using the recursion formula obtained for $\boldsymbol{\mu}^{(t)}$ (Eq. (29) in the appendix). For part (b), we write $\boldsymbol{\Sigma}^{(t)}$ as the empirical second moment around the *true* mean $\boldsymbol{\mu}$ minus $\boldsymbol{\delta}^{(t)}\boldsymbol{\delta}^{(t)\top}$ (Lemma A.2), and then use induction.

Remark. The covariance estimator is *negatively* biased by exactly the covariance of the mean estimator $\boldsymbol{C}_t \succeq 0$ ('$\succeq 0$' should be understood in the positive semi-definite sense). In particular, since $\boldsymbol{C}_1 = \boldsymbol{\Sigma}/n_r$, we have $\mathbb{E}[\boldsymbol{\Sigma}^{(1)}] = \frac{n_r-1}{n_r}\boldsymbol{\Sigma}$, recovering the familiar downward bias of the maximum likelihood covariance estimator under Gaussian distribution. We now show that this bias never exceeds twice its value at $t = 1$, regardless of how much generated data is used.

Lemma 3.2 (Bound on $\boldsymbol{C}_t$)**.** *Under the setup of Section 3.1, for all $t \geq 1$:*

$$\mathbf{0} \preceq \boldsymbol{C}_t \preceq \frac{2}{n_r}\boldsymbol{\Sigma}.$$

Remark. By Theorem 3.1(b), the upper bound in Lemma 3.2 is equivalent to $\mathbb{E}\big[\boldsymbol{\Sigma}^{(t)}\big] \succeq \big(1-\frac{2}{n_r}\big)\boldsymbol{\Sigma}$: the downward bias of the covariance estimator is at most twice its value at $t = 1$. Note that for $n_r = 1$ the bound is weaker than $\boldsymbol{C}_t \preceq \boldsymbol{\Sigma}$, which also holds for every $t \geq 1$: it follows from Theorem 3.1(b) and $\boldsymbol{\Sigma}^{(t)} \succeq \mathbf{0}$ (by construction).

Since $\boldsymbol{\mu}^{(t)}$ is unbiased (Theorem 3.1(a)), from (6) we have that

$$\boldsymbol{C}_t = \mathbb{E}\big[(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu})^\top\big] = \mathbb{E}\big[\boldsymbol{\delta}^{(t)}\boldsymbol{\delta}^{(t)\top}\big]. \tag{10}$$

Using $\|\boldsymbol{v}\|_2^2 = \mathrm{Tr}(\boldsymbol{v}\boldsymbol{v}^\top)$, the linearity of the trace and expectation, and (10),

$$\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] = \mathbb{E}\big[\|\boldsymbol{\delta}^{(t)}\|_2^2\big] = \mathbb{E}\big[\mathrm{Tr}\big(\boldsymbol{\delta}^{(t)}\boldsymbol{\delta}^{(t)\top}\big)\big] = \mathrm{Tr}\Big(\mathbb{E}\big[\boldsymbol{\delta}^{(t)}\boldsymbol{\delta}^{(t)\top}\big]\Big) = \mathrm{Tr}(\boldsymbol{C}_t). \tag{11}$$

Thus, $\mathrm{Tr}(\boldsymbol{C}_t)$ is the mean-squared error of $\boldsymbol{\mu}^{(t)}$ around $\boldsymbol{\mu}$.

3.2.2 Convergence to the true distribution

Crucially, we analyze in closed form when this iterative estimation process converges to the true distribution. Below, we establish that the condition for convergence of the estimated parameters to the truth collapses to a single elementary series in $\{\alpha_t\}_{t\geq 1}$, $\{n_b(t)\}_{t\geq 1}$, and n_r. Recall from (3) that $\rho_t \coloneqq N_t/N_{t+1}$.

Theorem 3.3 (Sharp elementary criterion for consistency)**.** *Under the setup of Section 3.1, the following are equivalent:*

(a) The following series diverges:

$$\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \sum_{t=1}^{\infty}(1-\alpha_t)\frac{n_b(t)}{N_{t+1}} = \infty. \tag{12}$$

(b) The mean estimator $\boldsymbol{\mu}^{(t)}$ converges to $\boldsymbol{\mu}$ in L^2, i.e., in mean square:

$$\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \xrightarrow{t\to\infty} 0. \tag{13}$$

(c) The covariance estimator converges to $\boldsymbol{\Sigma}$ in expected Frobenius norm:

$$\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big] \xrightarrow{t\to\infty} 0. \tag{14}$$

(d) The expected squared 2-Wasserstein distance between the estimated distribution and the true distribution tends to zero:

$$\mathbb{E}\big[W_2^2(\mathcal{D}_t, \mathcal{D}_{\mathrm{real}})\big] \xrightarrow{t\to\infty} 0. \tag{15}$$

Proof idea. Since $\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] = \mathrm{Tr}(\boldsymbol{C}_t)$, we derive a recursion for $\boldsymbol{C}_t$ (Theorem A.11) and use it to show that $\boldsymbol{C}_t \to \mathbf{0}$ if and only if (12) holds. For the covariance, a similar recursion for the empirical second moment around the true mean gives sufficiency, while necessity follows from the

bias $-\boldsymbol{C}_t$ (Theorem 3.1). For W_2, the Gaussian closed form splits $W_2^2(\mathcal{D}_t, \mathcal{D}_{\text{real}})$ into $\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2$ plus a nonnegative term bounded by $\sqrt{d}\,\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F$. Taking expectations yields sufficiency from the mean and covariance results, and necessity from the mean result alone.

Remark (interpretation). The factor $1 - \alpha_t$ is the fraction of the batch added at round t that is drawn from the true distribution, while $1 - \rho_t = n_b(t)/N_{t+1}$ is the weight this batch carries in the pool. Their product is therefore the share of *genuinely fresh real information* injected at round t, and the model escapes collapse precisely when these shares are non-summable. The convergences of the mean, the covariance and the fitted distribution hold or fail together, and, by Markov's inequality, each implies convergence in probability.

Discussion (implications). In the remainder of this section, we characterize when the series $\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t)$ diverges, which, using Theorem 3.3, allows us to determine when the iterative process converges to the true parameters.

We start by providing a sufficient condition for the mixing fraction for convergence.

Lemma 3.4 (Sufficient condition on the mixing fraction)**.** *If* $\sup_{t\geq 1}\alpha_t < 1$*, then*

$$\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty \tag{16}$$

In particular, for a fixed mixing fraction $\alpha_t \equiv \alpha \in [0, 1)$, the condition of Lemma 3.4 holds. Thus, any fixed positive fraction $1-\alpha$ of real data in each batch suffices for consistency, even when the real samples added at each iteration are vastly outnumbered by the generated ones. This is encouraging in practice: scraping the web naturally yields some new real data alongside generated data, and any fixed fraction of it suffices for convergence.

The next case we consider is an asymptotically vanishing fraction of real data. We show that the model can still be consistent, provided that the synthetic fraction tends to 1 slowly enough. Specifically, take $n_b(t) \equiv n_b \geq 1$, so that

$$N_{t+1} = n_r + n_b t \implies 1 - \rho_t = \frac{n_b}{N_{t+1}} \asymp \frac{1}{t}.$$

Here, $f(t) \asymp g(t)$ means that for all sufficiently large t, the ratio $f(t)/g(t)$ is bounded above and below by positive constants (independent of t). The criterion (12) for convergence of the estimated parameters to the true parameters then reads $\sum_{t=1}^{\infty}\frac{1-\alpha_t}{t} = \infty$. For mixing fractions $\{\alpha_t\}_{t=1}^{\infty}$ tending to 1 from below, this gives

$$\alpha_t = 1 - \tfrac{1}{t},\ \ \alpha_t = 1 - \tfrac{1}{\sqrt{t}} \implies \sum_{t=1}^{\infty}\frac{1-\alpha_t}{t} < \infty,$$

$$\alpha_t = 1 - \tfrac{1}{\ln(t+2)},\ \ \alpha_t = 1 - \tfrac{1}{\ln\ln(t+16)} \implies \sum_{t=1}^{\infty}\frac{1-\alpha_t}{t} = \infty.$$

Thus, consistency holds in the last two cases, despite the fraction of real data tending to zero as $t \to \infty$. Note that all four sequences satisfy $\sup_{t\geq 1}\alpha_t = 1$, so Lemma 3.4 covers none of them.

Finally, we consider the case $\alpha_t \equiv 1$, in which no real samples are added at any iteration. Then, $\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = 0$, so the condition from Theorem 3.3 is not satisfied. Thus, the estimated parameters do not converge to the true parameters, consistent with existing results in the literature (Dohmatob et al., 2025). In a purely replacement-based scheme, the expected squared 2-Wasserstein distance W_2 can diverge (Shumailov et al., 2024). In contrast, for our accumulation-based scheme, we next show that the estimators neither converge to the true parameters nor diverge. This is consistent with prior work showing that accumulating data keeps the error bounded (Gerstgrasser et al., 2024; Kazdan et al., 2025). Here, we provide explicit two-sided bounds in terms of n_r. This highlights a fundamental difference between the two schemes.

Proposition 3.5 (Bounded error at $\alpha = 1$)**.** *Under the setup of Section 3.1, if* $\alpha_t \equiv 1$*, then*

$$\frac{\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r} \leq \mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \leq \frac{2\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r}$$

$$\frac{\|\boldsymbol{\Sigma}\|_F}{n_r} \leq \mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big] \leq \frac{2\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r} + \frac{1}{\sqrt{n_r}}\sqrt{A_{\boldsymbol{\Sigma}} + \frac{3}{n_r}\exp\left(\frac{3}{n_r}\right)B_{\boldsymbol{\Sigma}}}$$

for every $t \geq 1$, where $A_{\boldsymbol{\Sigma}} = 5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 3\|\boldsymbol{\Sigma}\|_F^2$ and $B_{\boldsymbol{\Sigma}} = 5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 4\|\boldsymbol{\Sigma}\|_F^2$.

Proof idea. For the mean, we use the recursion for $\boldsymbol{C}_t$ (Theorem A.11) to show that $\mathrm{Tr}(\boldsymbol{C}_t)$ is non-decreasing in t, starting from $\mathrm{Tr}(\boldsymbol{C}_1) = \mathrm{Tr}(\boldsymbol{\Sigma})/n_r$, and bounded. For the covariance, the lower bound follows from the bias $-\boldsymbol{C}_t$ (Theorem 3.1), and the upper bound combines the bound on $\mathrm{Tr}(\boldsymbol{C}_t)$ with an upper bound on the error of the empirical second moment around the true mean.

3.2.3 Rate of convergence

We have shown that, for every fixed $\alpha < 1$, the estimators are consistent. This result concerns the limit of infinitely many iterations. In practical training scenarios, however, only a finite number of iterations is performed. We, therefore, ask: how quickly do the estimated parameters converge to the true parameters? Below, we establish the rate of convergence for the mean estimator. In Section 3.3, we show similar behavior for the covariance estimator.

The next result gives the exact convergence rate and reveals a sharp transition at $\alpha = \frac{1}{2}$.

Theorem 3.6 (Rate of decay)**.** *Assume the setup of Section 3.1, $\alpha_t \equiv \alpha \in [0, 1]$ and $n_b(t) \equiv n_b$. Then*

$$\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \asymp \begin{cases} N_t^{-1}, & \alpha < \frac{1}{2}, \\ N_t^{-1} \ln N_t, & \alpha = \frac{1}{2}, \\ N_t^{-2(1-\alpha)}, & \alpha > \frac{1}{2}, \end{cases} \tag{17}$$

where $N_t = n_r + (t-1)n_b \asymp t$. The implied constants depend only on α, n_b, n_r and $\mathrm{Tr}(\boldsymbol{\Sigma})$.

Proof idea. We derive the asymptotic rate of each term in the closed form of $\boldsymbol{C}_t$ (Eq. (39) in the appendix) by comparing sums with integrals.

Remark (interpretation). For $\alpha < \frac{1}{2}$, the estimator attains the parametric rate N_t^{-1} of the mean of N_t i.i.d. real samples: synthetic data inflates the constant but not the exponent. At $\alpha = \frac{1}{2}$ a phase transition occurs, and for $\alpha > \frac{1}{2}$ the exponent $2(1-\alpha)$ decreases linearly to 0 as $\alpha \to 1$. Consistency thus survives for every $\alpha < 1$, but becomes arbitrarily slow: halving the error requires multiplying N_t by $2^{1/[2(1-\alpha)]}$, a factor of 32 at $\alpha = 0.9$ and of 10^{15} at $\alpha = 0.99$. Equivalently, the pooled estimator is only as accurate as the mean of order $N_t^{2(1-\alpha)}$ i.i.d. real samples, a vanishing fraction of N_t. At $\alpha = 1$, the rate reads $N_t^{-2(1-\alpha)} = 1$: the error does not vanish (Theorem 3.3), but it does not diverge either (Proposition 3.5).

Remark (the value of identifying real samples). The gain from perfect detection of synthetic samples, i.e., from keeping only the real samples, stays bounded for $\alpha < \frac{1}{2}$, but grows like $N_t^{2\alpha-1}$ for $\alpha > \frac{1}{2}$, i.e., faster as the fraction of generated samples increases. For large but finite N_t, it is largest at some α strictly between 0 and 1, i.e., when generated data dominates but real data is still being added in meaningful amounts (see Appendix B for more discussion).

3.3 Empirical verification

In this section, we simulate the theoretical setup in order to further support our theoretical results and also gain more insights into the model, e.g., the rate of convergence of the covariance estimator.

Setup. We fix $d = 4$, $\boldsymbol{\mu} = (-1, -\frac{1}{3}, \frac{1}{3}, 1)^\top$, and $[\boldsymbol{\Sigma}]_{ij} = s_i s_j 2^{-|i-j|}$ with $s = (1, \frac{4}{3}, \frac{5}{3}, 2)^\top$, so that $\mathrm{Tr}(\boldsymbol{\Sigma}) = 9.56$, $\|\boldsymbol{\Sigma}\|_F = 6.18$, and $\boldsymbol{\Sigma}$ has a condition number 9.58. Additionally, we take $n_b(t) \equiv n_b = 10$ and $n_r = 20$. Two families of mixing fractions are used: a grid of 15 constant values $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 1\}$, spanning the transition of Theorem 3.6 and both endpoints of Theorem 3.3, and the four sequences $\alpha_t = 1 - \frac{1}{t}$, $1 - \frac{1}{\sqrt{t}}$, $1 - \frac{1}{\ln(t+2)}$, $1 - \frac{1}{\ln\ln(t+16)}$, all with $\sup_{t \geq 1} \alpha_t = 1$. Each new sample is drawn independently from the mixture (2), and in each iteration we update $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ according to the iterative estimation procedure. Each regime is run for $T = 10^5$ iterations ($N_T \approx 10^6$). Unless stated otherwise, expectations are estimated by averages over $M = 20{,}000$ independent runs, each drawing its own n_r initial real samples. At this M, the Monte-Carlo standard error of the averaged mean-square error is about 1% of its value, below the resolution of the figures (see also Appendix C).

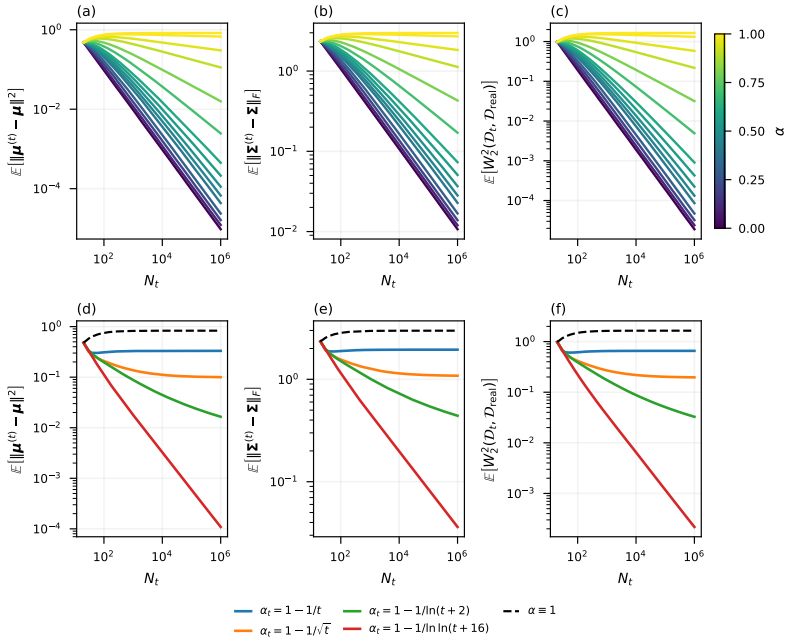

Figure 1: Convergence under constant and time-varying synthetic fractions. Top: constant α, from 0 (dark) to 1 (light, see color bar): (a) $\mathbb{E}[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2]$, (b) $\mathbb{E}[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F]$, and (c) $\mathbb{E}[W_2^2(\mathcal{D}_t, \mathcal{D}_{\text{real}})]$. Bottom: (d)–(f) the same quantities for $\alpha_t = 1 - \frac{1}{t}$, $1 - \frac{1}{\sqrt{t}}$, $1 - \frac{1}{\ln(t+2)}$ and $1 - \frac{1}{\ln\ln(t+16)}$, with $\alpha \equiv 1$ (dashed) for reference.

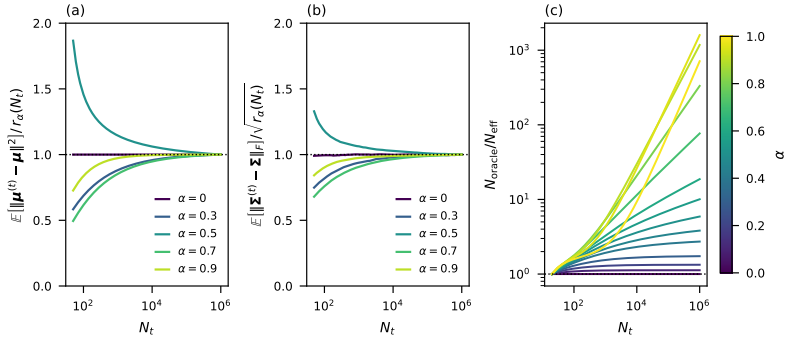

Figure 2: Rates of convergence and the effect of synthetic-sample detection. (a) $\mathbb{E}[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2]/r_\alpha(N_t)$ and (b) $\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big]/\sqrt{r_\alpha(N_t)}$, where r_α is the rate of Theorem 3.6, for $\alpha \in \{0, 0.3, 0.5, 0.7, 0.9\}$. Each curve is scaled to equal 1 at the largest N_t. (c) The gain from *perfect* detection of synthetic samples, $N_{\text{oracle}}/N_{\text{eff}}$, for constant $\alpha \in [0, 1)$ (see color bar).

Convergence at fixed α. Every curve in Figure 1 is a Monte-Carlo average over the $M = 20{,}000$ runs. The exception is W_2^2, computed from its closed form for Gaussians ((94) and (102)): since each evaluation requires a matrix square root, we average it over $10{,}000$ of the runs and evaluate it at 55 logarithmically spaced iterations rather than at every iteration. The top panels (a)–(c) illustrate Theorem 3.3 for constant α: for every $0 \le \alpha < 1$, the errors of the mean (a), the covariance (b) and the fitted distribution (c) decrease with N_t, whereas at $\alpha \equiv 1$, when no real data is added after the first iteration, they level off at a positive value (Proposition 3.5). Thus, when the real-data fraction is fixed, consistency depends only on whether real data is present at all. Its proportion governs the rate instead (Theorem 3.6): the decay slows as α grows, and becomes very slow as $\alpha \to 1$. At $\alpha = 0.99$ the mean-square error decays only like $N_t^{-0.02}$, too slowly to approach zero within the simulated horizon. For large α, the mean-square error also first *increases*, as predicted by an additional result (Proposition A.27) that we deferred to Appendix A.9 due to space limitations: above $\alpha^*(n_r, n_b) \approx 0.65$ it rises before it decreases. By the recursion of Theorem A.11, it rises by 25% at $\alpha = 0.9$ and by 62% at $\alpha = 0.99$, and by 74% at $\alpha = 1$, where it never decreases.

The sharpness of the criterion. Panels (d)–(f) in Figure 1 show the four sequences mentioned above. All four have $\sup_{t\geq 1}\alpha_t = 1$ and a vanishing fraction of real data, yet they split as predicted by (12): for $\alpha_t = 1-1/t$ and $1-1/\sqrt{t}$, whose series converges, the errors approach strictly positive limits, as at $\alpha \equiv 1$, while for $1-1/\ln(t+2)$ and $1-1/\ln\ln(t+16)$, whose series diverges, they keep decreasing. Thus, what matters is not whether the real-data fraction vanishes but how fast.

Rate of convergence. Figure 2(a) divides $\mathbb{E}[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2]$ by the rate $r_\alpha(N_t)$ stated in (17). Here, the error is computed exactly by $\mathrm{Tr}(\boldsymbol{C}_t)$ (Equation (11)) using the recursion of Theorem A.11. Appendix C (Figure 3) confirms that these exact values agree with the simulation. Each curve in Figure 2(a) is normalized by its own endpoint, so that flattening at 1 confirms the predicted rate. All five values $\alpha \in \{0, 0.3, 0.5, 0.7, 0.9\}$ level off, on both sides of the transition at $\alpha = \frac{1}{2}$. At $\alpha = 0$ the curve is flat for every N_t, since then $\boldsymbol{C}_t = \boldsymbol{\Sigma}/N_t$ exactly. Figure 2(b) repeats this for the covariance, dividing $\mathbb{E}[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F]$, estimated from the $M = 20{,}000$ Monte-Carlo runs, by $\sqrt{r_\alpha(N_t)}$, the square root because the Frobenius error is a distance while the mean-square error in (a) is a squared distance. The same five curves level off, so empirically the covariance converges at the rate of the mean, including the transition at $\alpha = \frac{1}{2}$, although Theorem 3.6 covers only the mean.

The value of identifying real samples. Figure 2(c) shows the gain $G_\alpha(N_t) = N_{\mathrm{oracle}}(N_t)/N_{\mathrm{eff}}(N_t)$ from perfect detection of synthetic samples. Here, $N_{\mathrm{eff}} = \mathrm{Tr}(\boldsymbol{\Sigma})/\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big]$ is the number of real i.i.d. samples that would give the same mean-square error, and $N_{\mathrm{oracle}} = (1-\alpha)N_t + \alpha n_r$ is the expected number of real samples in the pool (discussed more in Appendix B). N_{eff} is computed exactly using the recursion of Theorem A.11. The gain saturates for $\alpha < \frac{1}{2}$ and grows like $N_t^{2\alpha-1}$ for $\alpha > \frac{1}{2}$, reaching about 10^3 at $N_t = 10^6$ for $\alpha \in [0.9, 0.95]$. At $\alpha = 0.99$ it is smaller, since the oracle also runs out of real data as $\alpha \to 1$.

4 Experiments with Diffusion Models

Our paper focuses on theoretical contributions. Yet, to demonstrate that insights generalize beyond the theoretical setup, in Appendix D we present results of iterative training of a DDPM-based diffusion model (Ho et al., 2020), on the CIFAR-10 dataset (Krizhevsky et al., 2009). We examine the effect of different fixed values of α, with no distinction between real and generated data, on the Fréchet Inception Distance (FID) quality metric. Aligning with our theory: 1) for every $\alpha < 1$, the FID improves upon its initial value at every iteration, although real and generated images are pooled indistinguishably; 2) for $\alpha = 1$, the FID never improves but remains bounded; 3) the FID is ordered by α at every iteration, and its improvement slows as α grows; 4) for $\alpha = 0.25 < \frac{1}{2}$, the curve follows that of $\alpha = 0$ up to a roughly constant offset, i.e., synthetic data inflates the constant but not the rate; and 5) discarding the generated images would improve the final FID far more for $\alpha = 0.75 > \frac{1}{2}$ than for $\alpha = 0.25 < \frac{1}{2}$. An apparent gap between the theoretical setup and the diffusion experiments is that, for a large $\alpha < 1$, the FID stops improving at some t. This may result from the simple training strategy used (which could be improved) and the relatively simple diffusion model (which could be scaled up).

5 Conclusion

Large generative models are commonly trained on data scraped from the web, an increasing share of which is generated by earlier models. Prior work shows that iterative training degrades performance: 1) without bound when earlier data is discarded, and 2) in a bounded manner when data accumulates. The continued improvement of generative models has been explained by the availability of fresh real data, but only under the unrealistic assumption that it can be perfectly separated from generated data. In this paper, we studied the more realistic setting in which fresh real and generated samples are pooled indistinguishably. We showed, for a tractable Gaussian data model, that the estimated distribution converges to the true one if and only if the share of fresh real information (quantified exactly) is not summable. Our results reveal that consistency holds for any fixed positive fraction of real data, and even for a fraction that vanishes slowly enough. Yet, the convergence rate of the mean estimator deteriorates as generated samples become more prevalent: it matches the rate of purely real data as long as generated samples are a minority of the newly collected data, and becomes arbitrarily slow once they are a majority. This motivates ongoing research on detecting and curating generated samples, and indicates when such efforts are worthwhile: as more data is collected, the gain from perfect detection is bounded when generated samples are a minority of the newly collected data, but grows without bound with the amount of collected data otherwise.

AI USE STATEMENT

We used generative AI tools to polish the writing, for retrieval and discovery of related work, and to assist research execution, including writing code and checking derivations. All AI-assisted work was reviewed by the authors: code was verified and tested for correctness, and all cited works, configurations, code, results, and proofs were validated by the authors. We take full responsibility for the content of this work, including text, claims, and 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 G. Baraniuk. Self-Consuming Generative Models Go MAD. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ShjMHfmPs0.

Daniel Barzilai and Ohad Shamir. When Models Don't Collapse: On the Consistency of Iterative MLE. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Quentin Bertrand, Avishek 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.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model Collapse Demystified: The Case of Regression. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong Model Collapse. In *The Thirteenth International Conference on Learning Representations*, 2025.

D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

George Drayson, Emine Yilmaz, and Vasileios Lampos. Machine-generated text detection prevents language model collapse. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29657–29673, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1506. URL https://aclanthology.org/2025.emnlp-main.1506/.

Anvit Garg, Sohom Bhattacharya, and Pragya Sur. Preventing Model Collapse Under Overparametrization: Optimal Mixing Ratios for Interpolation Learning and Ridge Regression. In *The Fourteenth International Conference on Learning Representations*, 2026.

Matthias Gelbrich. On a Formula for the L2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990. doi: 10.1002/mana.19901470121.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz.

Hengzhi He, Shirong Xu, and Guang Cheng. Recursive Learning Without Collapse: A Weighting-Based Stabilization Framework. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2026. doi: 10.1093/jrsssb/qkag099. arXiv:2502.18049.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29469–29494, 2025.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking AI-generated image detection in challenging scenarios. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20379–20389. IEEE, 2025.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11410–11420, 2022.

Robert T. Powers and Erling Størmer. Free states of the canonical anticommutation relations. *Communications in Mathematical Physics*, 16(1):1–33, 1970. doi: 10.1007/BF01645492.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-Generated Text be Reliably Detected? Stress Testing AI Text Detectors Under Various Attacks. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Leonardo V. Santoro and Victor M. Panaretos. Large sample theory for Bures–Wasserstein barycentres. *The Annals of Applied Probability*, 35(5):3215–3241, 2025.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. How Bad is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=t3z6U1V09o.

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, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *The Ninth International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Bohao Zou. Denoising Diffusion Probabilistic Model (DDPM) Implementation. https://github.com/zoubohao/DenoisingDiffusionProbabilityModel-ddpm-. GitHub repository.

A PROOFS

Section A.1 defines three scalar sequences, which are used widely in the proofs.

A.1 SCALAR SEQUENCES

For $t \geq 1$, we define three scalar sequences:

$$\lambda_t \coloneqq \frac{N_t + n_b(t)\alpha_t}{N_{t+1}}, \qquad \beta_t \coloneqq \frac{n_b(t)}{N_{t+1}^2}, \qquad \gamma_t \coloneqq \lambda_t^2 - \alpha_t^2 \beta_t. \tag{18}$$

Additionally, let us define

$$\Gamma_{s,t} \coloneqq \begin{cases} \prod_{k=s}^{t} \gamma_k & s \leq t \\ 1 & s > t \end{cases}. \tag{19}$$

Section A.2 collects the technical lemmas used throughout.

A.2 TECHNICAL LEMMAS

We first formalize the conditioning used throughout the proofs.

Definition 1 (Natural filtration)**.** *For every $t \geq 1$, let*

$$\mathcal{F}_t \coloneqq \sigma\Big(\{\boldsymbol{x}_j^{(0)}\}_{j=1}^{n_r},\ \{\boldsymbol{x}_i^{(s)}\}_{1 \leq i \leq n_b(s),\ 1 \leq s \leq t-1}\Big) \tag{20}$$

denote the σ-algebra generated by all samples drawn before the batch $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^{n_b(t)}$. For $t = 1$, the second family is empty, so $\mathcal{F}_1$ is generated by the original real batch alone.

Since each iteration only adds samples, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ is a filtration, and $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ is $\mathcal{F}_t$-measurable by (4). Hence, Assumption 1 is equivalent to the following: for every $t \geq 1$, conditionally on $\mathcal{F}_t$, the samples $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^{n_b(t)}$ are i.i.d. from the mixture (2) with parameters $(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$. This is the form used throughout the proofs.

Next, we record the conditional first and second moments of a newly drawn sample. These are used in the proof of Theorems 3.1 and A.11.

Lemma A.1 (Conditional moments of the mixture samples)**.** *Under Assumption 1, conditional on $\mathcal{F}_t$, each newly drawn sample $\boldsymbol{x}_i^{(t)}$ $(t \geq 1)$ satisfies, using* (7) *$(\boldsymbol{\delta}^{(t)} = \boldsymbol{\mu}^{(t)} - \boldsymbol{\mu})$:*

$$\mathbb{E}\big[\boldsymbol{x}_i^{(t)} \mid \mathcal{F}_t\big] = \alpha_t \boldsymbol{\mu}^{(t)} + (1-\alpha_t)\boldsymbol{\mu} = \boldsymbol{\mu} + \alpha_t \boldsymbol{\delta}^{(t)}, \tag{21}$$

$$\mathrm{Cov}\big(\boldsymbol{x}_i^{(t)} \mid \mathcal{F}_t\big) = \alpha_t \boldsymbol{\Sigma}^{(t)} + (1-\alpha_t)\boldsymbol{\Sigma} + \alpha_t(1-\alpha_t)\,\boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top, \tag{22}$$

$$\mathbb{E}\big[(\boldsymbol{x}_i^{(t)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(t)} - \boldsymbol{\mu})^\top \mid \mathcal{F}_t\big] = \alpha_t \boldsymbol{\Sigma}^{(t)} + (1-\alpha_t)\boldsymbol{\Sigma} + \alpha_t\,\boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top. \tag{23}$$

Proof. Given $\mathcal{F}_t$, the sample $\boldsymbol{x}_i^{(t)}$ is drawn from the two-component mixture (2) with weights α_t on $\mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ and $1-\alpha_t$ on $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Equation (21)*.* By linearity, the mixture mean is $\alpha_t \boldsymbol{\mu}^{(t)} + (1-\alpha_t)\boldsymbol{\mu} = \boldsymbol{\mu} + \alpha_t \boldsymbol{\delta}^{(t)}$.

Equation (22)*.* Let $Z \in \{1, 2\}$ be the component indicator with $\mathbb{P}(Z = 1 \mid \mathcal{F}_t) = \alpha_t$, so that $\boldsymbol{x}_i^{(t)} \mid \{Z = 1\} \sim \mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ and $\boldsymbol{x}_i^{(t)} \mid \{Z = 2\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The law of total covariance gives

$$\mathrm{Cov}\big(\boldsymbol{x}_i^{(t)} \mid \mathcal{F}_t\big) = \underbrace{\mathbb{E}\big[\mathrm{Cov}(\boldsymbol{x}_i^{(t)} \mid Z, \mathcal{F}_t) \mid \mathcal{F}_t\big]}_{A} + \underbrace{\mathrm{Cov}\big(\mathbb{E}[\boldsymbol{x}_i^{(t)} \mid Z, \mathcal{F}_t] \mid \mathcal{F}_t\big)}_{B}.$$

The first summand A reads:

$$\mathrm{Cov}(\boldsymbol{x}_i^{(t)} \mid Z, \mathcal{F}_t) = \begin{cases} \boldsymbol{\Sigma}^{(t)} & Z = 1 \\ \boldsymbol{\Sigma} & Z = 2 \end{cases} \Rightarrow A = \alpha_t \boldsymbol{\Sigma}^{(t)} + (1-\alpha_t)\boldsymbol{\Sigma}. \tag{24}$$

We now compute the second summand B. We have

$$\mathcal{P}(Z) \coloneqq \mathbb{E}[\boldsymbol{x}_i^{(t)} \mid Z, \mathcal{F}_t] = \begin{cases} \boldsymbol{\mu}^{(t)} & Z = 1 \\ \boldsymbol{\mu} & Z = 2 \end{cases} \Rightarrow B = \mathbb{E}\big[(\mathcal{P}(Z) - \mu_{\mathcal{P}})\,(\mathcal{P}(Z) - \mu_{\mathcal{P}})^\top \mid \mathcal{F}_t\big]$$

where

$$\mu_{\mathcal{P}} = \mathbb{E}[\mathcal{P}(Z) \mid \mathcal{F}_t] = \alpha_t \boldsymbol{\mu}^{(t)} + (1 - \alpha_t)\boldsymbol{\mu}.$$

That is, $B = \mathbb{E}\left[\mathcal{P}_c(Z) \mid \mathcal{F}_t\right]$, where

$$(\mathcal{P}(Z) - \mu_{\mathcal{P}}) = \begin{cases} (1-\alpha_t)(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}) & Z = 1 \\ -\alpha_t(\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}) & Z = 2 \end{cases} = \begin{cases} (1-\alpha_t)\boldsymbol{\delta}^{(t)} & Z = 1 \\ -\alpha_t \boldsymbol{\delta}^{(t)} & Z = 2 \end{cases}$$

$$\mathcal{P}_c(Z) = (\mathcal{P}(Z) - \mu_{\mathcal{P}})\,(\mathcal{P}(Z) - \mu_{\mathcal{P}})^\top = \begin{cases} (1-\alpha_t)^2 \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top & Z = 1 \\ \alpha_t^2 \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top & Z = 2 \end{cases}$$

Finally, the second part reads

$$B = \alpha_t\big((1-\alpha_t)\boldsymbol{\delta}^{(t)}\big)\big((1-\alpha_t)\boldsymbol{\delta}^{(t)}\big)^\top + (1-\alpha_t)\big(\alpha_t\boldsymbol{\delta}^{(t)}\big)\big(\alpha_t\boldsymbol{\delta}^{(t)}\big)^\top = \alpha_t(1-\alpha_t)\,\boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top,$$

using $\alpha_t(1-\alpha_t)^2 + (1-\alpha_t)\alpha_t^2 = \alpha_t(1-\alpha_t)$. Adding the two terms yields (22).

Equation (23). For any random vector $\boldsymbol{x}$ and fixed $\boldsymbol{\mu}$, $\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathrm{Cov}(\boldsymbol{x}) + (\mathbb{E}[\boldsymbol{x}] - \boldsymbol{\mu})(\mathbb{E}[\boldsymbol{x}] - \boldsymbol{\mu})^\top$. Applying this conditionally on $\mathcal{F}_t$ with $\mathbb{E}[\boldsymbol{x}_i^{(t)} \mid \mathcal{F}_t] - \boldsymbol{\mu} = \alpha_t\boldsymbol{\delta}^{(t)}$ and substituting (22),

$$\alpha_t\boldsymbol{\Sigma}^{(t)} + (1-\alpha_t)\boldsymbol{\Sigma} + \alpha_t(1-\alpha_t)\boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top + \alpha_t^2\boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top = \alpha_t\boldsymbol{\Sigma}^{(t)} + (1-\alpha_t)\boldsymbol{\Sigma} + \alpha_t\boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top,$$

since $\alpha_t(1-\alpha_t) + \alpha_t^2 = \alpha_t$. □

Next, we derive a more useful representation of $\boldsymbol{\Sigma}^{(t)}$ that is centered at $\boldsymbol{\mu}$. It is used in the proof of Theorems 3.1 and A.22.

Lemma A.2 (Pooled scatter identity)**.** *Fix $t \geq 1$. The estimators $\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}$ defined in* (4) *satisfy, using* (7) *($\boldsymbol{\delta}^{(t)} = \boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}$):*

$$\boldsymbol{\Sigma}^{(t)} = \frac{1}{N_t}\left(\sum_{j=1}^{n_r}(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top + \sum_{s=1}^{t-1}\sum_{i=1}^{n_b(s)}(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})^\top\right) - \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top. \tag{25}$$

Proof. By (4), we have

$$\sum_{k=1}^{N_t}\boldsymbol{y}_k = N_t\boldsymbol{\mu}^{(t)}.$$

Thus,

$$\sum_{k=1}^{N_t}(\boldsymbol{y}_k - \boldsymbol{\mu}) = N_t\boldsymbol{\mu}^{(t)} - N_t\boldsymbol{\mu} = N_t\boldsymbol{\delta}^{(t)}.$$

Expanding the defining sum in (4), we get

$$\begin{aligned}
N_t \boldsymbol{\Sigma}^{(t)} &= \sum_{k=1}^{N_t} \left(\boldsymbol{y}_k - \boldsymbol{\mu}^{(t)}\right)\left(\boldsymbol{y}_k - \boldsymbol{\mu}^{(t)}\right)^\top \\
&= \sum_{k=1}^{N_t} \left((\boldsymbol{y}_k - \boldsymbol{\mu}) - \boldsymbol{\delta}^{(t)}\right)\left((\boldsymbol{y}_k - \boldsymbol{\mu}) - \boldsymbol{\delta}^{(t)}\right)^\top \\
&= \sum_{k=1}^{N_t} (\boldsymbol{y}_k - \boldsymbol{\mu})(\boldsymbol{y}_k - \boldsymbol{\mu})^\top - \sum_{k=1}^{N_t} (\boldsymbol{y}_k - \boldsymbol{\mu})(\boldsymbol{\delta}^{(t)})^\top - \boldsymbol{\delta}^{(t)} \sum_{k=1}^{N_t} (\boldsymbol{y}_k - \boldsymbol{\mu})^\top + N_t \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top \\
&= \sum_{k=1}^{N_t} (\boldsymbol{y}_k - \boldsymbol{\mu})(\boldsymbol{y}_k - \boldsymbol{\mu})^\top - N_t \cdot \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top - N_t \cdot \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top + N_t \cdot \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top \\
&= \sum_{k=1}^{N_t} (\boldsymbol{y}_k - \boldsymbol{\mu})(\boldsymbol{y}_k - \boldsymbol{\mu})^\top - N_t \cdot \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top.
\end{aligned}$$

Dividing both sides by N_t, we get the required identity:

$$\boldsymbol{\Sigma}^{(t)} = \frac{1}{N_t}\left(\sum_{j=1}^{n_r} (\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top + \sum_{s=1}^{t-1} \sum_{i=1}^{n_b(s)} (\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})^\top\right) - \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top.$$

□

Next, we record an elementary consequence of the Cauchy-Schwarz inequality, used in the proofs of Lemmas A.4, A.23, and A.19.

Lemma A.3 (ℓ_1–ℓ_2 comparison)**.** *Let $n \geq 1$ and $a_1, \ldots, a_n \in \mathbb{R}$. Then*

$$\sum_{i=1}^{n} |a_i| \leq \sqrt{n}\left(\sum_{i=1}^{n} a_i^2\right)^{1/2}, \qquad \textit{and consequently} \qquad \left(\sum_{i=1}^{n} a_i\right)^2 \leq n \sum_{i=1}^{n} a_i^2. \tag{26}$$

Proof. Write $\boldsymbol{a} \coloneqq (|a_1|, \ldots, |a_n|)^\top$ and $\mathbf{1}_n \coloneqq (1, \ldots, 1)^\top$. The Cauchy–Schwarz inequality gives

$$\sum_{i=1}^{n} |a_i| = \langle \boldsymbol{a}, \mathbf{1}_n \rangle \leq \|\boldsymbol{a}\|_2 \, \|\mathbf{1}_n\|_2 = \sqrt{n}\left(\sum_{i=1}^{n} a_i^2\right)^{1/2},$$

which is the first inequality. For the second, the triangle inequality gives $\left|\sum_{i=1}^{n} a_i\right| \leq \sum_{i=1}^{n} |a_i|$. Both sides are nonnegative, so squaring and applying the first inequality yields

$$\left(\sum_{i=1}^{n} a_i\right)^2 \leq \left(\sum_{i=1}^{n} |a_i|\right)^2 \leq n \sum_{i=1}^{n} a_i^2.$$

□

We now record an elementary fact about positive-semi-definite matrices. It is used in the proofs of Lemma A.10, Theorems A.12 and A.22, and Lemma A.16.

Lemma A.4 (Trace–Frobenius comparison)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be symmetric with $\boldsymbol{A} \succeq 0$. Then:*

(a) $\|\boldsymbol{A}\|_F \leq \operatorname{Tr}(\boldsymbol{A}) \leq \sqrt{d}\,\|\boldsymbol{A}\|_F$. In particular, $\|\boldsymbol{A}\|_F^2 \leq \operatorname{Tr}(\boldsymbol{A})^2$.

(b) If $\{\boldsymbol{A}_t\}_{t \geq 1} \subset \mathbb{R}^{d \times d}$ are symmetric and positive-semi-definite, then

$$\|\boldsymbol{A}_t\|_F \xrightarrow{t \to \infty} 0 \quad \Longleftrightarrow \quad \operatorname{Tr}(\boldsymbol{A}_t) \xrightarrow{t \to \infty} 0. \tag{27}$$

Proof. (a) By the spectral theorem, $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$ with $\boldsymbol{Q}$ orthogonal and $\boldsymbol{D} = \mathrm{diag}(\nu_1, \dots, \nu_d)$, where $\nu_1, \dots, \nu_d \geq 0$ are the eigenvalues of $\boldsymbol{A}$ (real by symmetry, nonnegative since $\boldsymbol{A} \succeq 0$). As the trace and the Frobenius norm are invariant under orthogonal conjugation,

$$\mathrm{Tr}(\boldsymbol{A}) = \sum_{i=1}^{d} \nu_i, \qquad \|\boldsymbol{A}\|_F^2 = \sum_{i=1}^{d} \nu_i^2. \tag{28}$$

Lower bound. Expanding the square and discarding the cross terms, which are nonnegative because every $\nu_i \geq 0$,

$$\mathrm{Tr}(\boldsymbol{A})^2 = \left(\sum_{i=1}^{d} \nu_i\right)^2 = \sum_{i=1}^{d} \nu_i^2 + \sum_{i \neq j} \nu_i \nu_j \;\geq\; \sum_{i=1}^{d} \nu_i^2 = \|\boldsymbol{A}\|_F^2.$$

Both $\mathrm{Tr}(\boldsymbol{A})$ and $\|\boldsymbol{A}\|_F$ are nonnegative, so taking square roots gives $\|\boldsymbol{A}\|_F \leq \mathrm{Tr}(\boldsymbol{A})$, together with the squared form $\|\boldsymbol{A}\|_F^2 \leq \mathrm{Tr}(\boldsymbol{A})^2$.

Upper bound. Since every $\nu_i \geq 0$, Lemma A.3 gives

$$\mathrm{Tr}(\boldsymbol{A}) = \sum_{i=1}^{d} \nu_i \leq \sqrt{d} \left(\sum_{i=1}^{d} \nu_i^2\right)^{1/2} = \sqrt{d}\, \|\boldsymbol{A}\|_F.$$

(b) Apply (a) to each $\boldsymbol{A}_t$:

$$\|\boldsymbol{A}_t\|_F \leq \mathrm{Tr}(\boldsymbol{A}_t) \leq \sqrt{d}\, \|\boldsymbol{A}_t\|_F, \qquad t \geq 1.$$

The constants 1 and $\sqrt{d}$ do not depend on t, so the two sequences are squeezed to 0 together, which is (27). □

Next, we prove that the sequence β_t, defined in (18), is summable. It is used in the proof of Theorems A.13, 3.6, and A.22.

Lemma A.5 (Summability of β_s)**.** *The weights β_s from* (18) *satisfy*

$$\sum_{s=1}^{\infty} \beta_s \leq \frac{1}{n_r}.$$

Proof. Since $N_{s+1} = N_s + n_b(s) \geq N_s$ from (3),

$$\beta_s = \frac{n_b(s)}{N_{s+1}^2} = \frac{N_{s+1} - N_s}{N_{s+1}^2} \leq \frac{N_{s+1} - N_s}{N_s N_{s+1}} = \frac{1}{N_s} - \frac{1}{N_{s+1}},$$

where we used the definition of β_s from (18). Summing telescopes:

$$\sum_{s=1}^{t-1} \beta_s \leq \frac{1}{N_1} - \frac{1}{N_t} < \frac{1}{N_1} = \frac{1}{n_r}.$$

Taking the limit $t \to \infty$, we get $\sum_{s=1}^{\infty} \beta_s \leq 1/n_r$. □

Next, we prove a lemma that is used in the proof of Theorem A.13 and Lemma A.25.

Lemma A.6 (Product-decay criterion)**.** *Let $(r_s)_{s \geq 1} \subset (0, 1)$. If $\prod_{s=1}^{t} r_s \to 0$ as $t \to \infty$, then $\sum_{s=1}^{\infty} (1 - r_s) = \infty$.*

Proof. We prove the contrapositive: if $\sum_{s=1}^{\infty} (1 - r_s) < \infty$, then $\prod_{s=1}^{t} r_s$ has a positive limit (as $1 - r_s > 0$, the series either converges or diverges to $+\infty$, so this is indeed the negation of the conclusion).

Assume $\sum_{s=1}^{\infty} (1 - r_s) < \infty$. Then, $1 - r_s \to 0$, so there is $s_0 \geq 1$ with $r_s \geq \frac{1}{2}$ for $s \geq s_0$.

In addition, for $x \in [\frac{1}{2}, 1)$, the function $g(x) := 2(1-x) + \ln x$ is decreasing: $g'(x) = 1/x - 2 \le 0$, and $g(1) = 0$. Thus, $g \ge 0$ on $[\frac{1}{2}, 1)$, i.e., $-\ln x \le 2(1-x)$.

Finally, for $s \ge s_0$, we have $r_s \ge \frac{1}{2}$, so $\sum_{s \ge s_0}(-\ln r_s) \le 2\sum_{s \ge s_0}(1 - r_s) < \infty$. That is, $\sum_{s \ge 1}(-\ln r_s) < \infty$ (the finitely many remaining terms are finite as $r_s > 0$) and $\prod_{s=1}^{t} r_s = \exp\big(-\sum_{s=1}^{t}(-\ln r_s)\big)$ converges to a strictly positive limit, as claimed. □

We now provide more useful expressions concerning the estimators $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$.

Theorem A.7 (Estimators formulas)**.** *Fix $t \ge 2$ and consider the setup in Section 3.1. Then:*

(a) The mean obeys the recursion

$$\boldsymbol{\mu}^{(t)} = \rho_{t-1}\boldsymbol{\mu}^{(t-1)} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)} \boldsymbol{x}_i^{(t-1)}. \tag{29}$$

(b) Splitting the pool into the initial real batch and the subsequent batches, the covariance reads

$$\boldsymbol{\Sigma}^{(t)} = \frac{1}{N_t}\left(\sum_{j=1}^{n_r}\big(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}^{(t)}\big)\big(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}^{(t)}\big)^\top + \sum_{s=1}^{t-1}\sum_{i=1}^{n_b(s)}\big(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu}^{(t)}\big)\big(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu}^{(t)}\big)^\top\right). \tag{30}$$

Remark. Equation (29) expresses the updated mean as a weighted average of the previous estimate and the new batch mean. Equation (30) is the pooled sample covariance centered at the estimated mean $\boldsymbol{\mu}^{(t)}$.

Proof. **Step 1 (part (a): mean recursion** (29))**.** Split the most recent batch $s = t-1$ off from the definition (4) of $\boldsymbol{\mu}^{(t)}$:

$$\boldsymbol{\mu}^{(t)} = \frac{1}{N_t}\left(\sum_{j=1}^{n_r}\boldsymbol{x}_j^{(0)} + \sum_{s=1}^{t-2}\sum_{i=1}^{n_b(s)}\boldsymbol{x}_i^{(s)}\right) + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)}\boldsymbol{x}_i^{(t-1)}.$$

The parenthesized sum is exactly the pool of the first $N_{t-1} = n_r + \sum_{s=1}^{t-2} n_b(s)$ points, so applying the definition (4) at iteration $t-1$—which for $t = 2$ is precisely (1)—gives $\sum_{j=1}^{n_r}\boldsymbol{x}_j^{(0)} + \sum_{s=1}^{t-2}\sum_{i=1}^{n_b(s)}\boldsymbol{x}_i^{(s)} = N_{t-1}\,\boldsymbol{\mu}^{(t-1)}$. With $\rho_{t-1} = N_{t-1}/N_t$ from (3),

$$\boldsymbol{\mu}^{(t)} = \frac{N_{t-1}}{N_t}\,\boldsymbol{\mu}^{(t-1)} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)}\boldsymbol{x}_i^{(t-1)} = \rho_{t-1}\,\boldsymbol{\mu}^{(t-1)} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)}\boldsymbol{x}_i^{(t-1)},$$

which is (29).

Step 2 (part (b)). Relabeling the N_t points in (4) by provenance splits the pooled scatter into the initial real batch and the subsequent batches:

$$\boldsymbol{\Sigma}^{(t)} = \frac{1}{N_t}\left(\sum_{j=1}^{n_r}\big(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}^{(t)}\big)\big(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}^{(t)}\big)^\top + \sum_{s=1}^{t-1}\sum_{i=1}^{n_b(s)}\big(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu}^{(t)}\big)\big(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu}^{(t)}\big)^\top\right),$$

which is exactly (30). This completes the proof. □

Finally, we prove that the estimated covariance $\boldsymbol{\Sigma}^{(t)}$ is PSD.

Lemma A.8 (The covariance estimator is PSD)**.** *Fix $t \ge 1$. Consider the setup in Section 3.1. For every realization of the data, the sample covariance matrix $\boldsymbol{\Sigma}^{(t)}$ is positive-semi-definite.*

Proof. Fix $t \geq 1$. By (30) — with the convention that the double sum $\sum_{s=1}^{t-1}$ is empty when $t = 1$, so that $\boldsymbol{\Sigma}^{(1)}$ reduces to (1) — the estimator is a nonnegatively weighted sum of outer products of the centered points $\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu}^{(t)}$. Each such outer product is symmetric, hence so is $\boldsymbol{\Sigma}^{(t)}$. Moreover, for every $\boldsymbol{v} \in \mathbb{R}^d$,

$$\boldsymbol{v}^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{v} = \frac{1}{N_t} \left(\sum_{j=1}^{n_r} \big(\boldsymbol{v}^\top (\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}^{(t)})\big)^2 + \sum_{s=1}^{t-1} \sum_{i=1}^{n_b(s)} \big(\boldsymbol{v}^\top (\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu}^{(t)})\big)^2 \right) \geq 0,$$

since it is a sum of squares scaled by $1/N_t > 0$. Hence $\boldsymbol{\Sigma}^{(t)} \succeq 0$. □

Remark (singular covariance). If $\boldsymbol{\Sigma}^{(t)}$ is singular, $\mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ in (2) denotes the corresponding degenerate Gaussian, i.e., the law of $\boldsymbol{\mu}^{(t)} + (\boldsymbol{\Sigma}^{(t)})^{1/2} \boldsymbol{z}$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, where $(\boldsymbol{\Sigma}^{(t)})^{1/2}$ is the unique symmetric PSD square root, which exists by Lemma A.8.

In the following subsections, we present the proofs of Theorem 3.1, Lemma 3.2, Theorem 3.3, Lemma 3.4, Proposition 3.5, Theorem 3.6 and Proposition A.27.

A.3 Proof of Theorem 3.1

Proof of part (a): $\boldsymbol{\mu}^{(t)}$ is unbiased. We prove by induction on $t \geq 1$.

Base case ($t = 1$). By (1), $\boldsymbol{\mu}^{(1)} = \frac{1}{n_r} \sum_{j=1}^{n_r} \boldsymbol{x}_j^{(0)}$ is the sample mean of n_r i.i.d. draws from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so

$$\mathbb{E}\big[\boldsymbol{\mu}^{(1)}\big] = \frac{1}{n_r} \sum_{j=1}^{n_r} \mathbb{E}\big[\boldsymbol{x}_j^{(0)}\big] = \boldsymbol{\mu}.$$

Inductive step. Fix $t \geq 2$ and assume $\mathbb{E}\big[\boldsymbol{\mu}^{(t-1)}\big] = \boldsymbol{\mu}$. Taking expectations in (29) and using the linearity of expectation, we get

$$\mathbb{E}\big[\boldsymbol{\mu}^{(t)}\big] = \rho_{t-1}\, \mathbb{E}\big[\boldsymbol{\mu}^{(t-1)}\big] + \frac{1}{N_t} \sum_{i=1}^{n_b(t-1)} \mathbb{E}\big[\boldsymbol{x}_i^{(t-1)}\big]. \tag{31}$$

To evaluate $\mathbb{E}\big[\boldsymbol{x}_i^{(t-1)}\big]$, we condition on $\mathcal{F}_{t-1}$ (recall (20)). By Assumption 1, given $\mathcal{F}_{t-1}$, the samples $\boldsymbol{x}_i^{(t-1)}$ are i.i.d. from the mixture (2), whose mean is $\alpha_{t-1}\boldsymbol{\mu}^{(t-1)} + (1 - \alpha_{t-1})\boldsymbol{\mu}$. Hence, by the tower property and the induction hypothesis,

$$\begin{aligned}\mathbb{E}\big[\boldsymbol{x}_i^{(t-1)}\big] &= \mathbb{E}\Big[\mathbb{E}\big[\boldsymbol{x}_i^{(t-1)} \mid \mathcal{F}_{t-1}\big]\Big] = \mathbb{E}\big[\alpha_{t-1}\boldsymbol{\mu}^{(t-1)} + (1 - \alpha_{t-1})\boldsymbol{\mu}\big] \\ &= \alpha_{t-1}\, \mathbb{E}\big[\boldsymbol{\mu}^{(t-1)}\big] + (1 - \alpha_{t-1})\boldsymbol{\mu} = \alpha_{t-1}\boldsymbol{\mu} + (1 - \alpha_{t-1})\boldsymbol{\mu} = \boldsymbol{\mu}.\end{aligned}$$

Substituting $\mathbb{E}\big[\boldsymbol{\mu}^{(t-1)}\big] = \boldsymbol{\mu}$ and $\mathbb{E}\big[\boldsymbol{x}_i^{(t-1)}\big] = \boldsymbol{\mu}$ into (31),

$$\mathbb{E}\big[\boldsymbol{\mu}^{(t)}\big] = \rho_{t-1}\boldsymbol{\mu} + \frac{n_b(t-1)}{N_t}\boldsymbol{\mu} = \frac{N_{t-1} + n_b(t-1)}{N_t}\,\boldsymbol{\mu} = \boldsymbol{\mu},$$

where we used $\rho_{t-1} = N_{t-1}/N_t$ together with $N_{t-1} + n_b(t-1) = N_t$, which follows directly from (3). □

Proof of part (b): bias of $\boldsymbol{\Sigma}^{(t)}$. Step 1: the pooled scatter identity. From (10), we have

$$\mathbb{E}\big[\boldsymbol{\delta}^{(s)}(\boldsymbol{\delta}^{(s)})^\top\big] = \mathbb{E}\big[(\boldsymbol{\mu}^{(s)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(s)} - \boldsymbol{\mu})^\top\big] = \boldsymbol{C}_s, \qquad s \geq 1. \tag{32}$$

Multiplying by N_t and taking expectations in (25), we get

$$\begin{aligned}N_t\, \mathbb{E}\big[\boldsymbol{\Sigma}^{(t)}\big] &= \sum_{j=1}^{n_r} \mathbb{E}\big[(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top\big] + \sum_{s=1}^{t-1} \sum_{i=1}^{n_b(s)} \mathbb{E}\big[(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})^\top\big] - N_t \boldsymbol{C}_t \\ &= n_r \boldsymbol{\Sigma} + \sum_{s=1}^{t-1} \sum_{i=1}^{n_b(s)} \mathbb{E}\big[(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})^\top\big] - N_t \boldsymbol{C}_t,\end{aligned} \tag{33}$$

where we used that $\boldsymbol{x}_j^{(0)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are i.i.d., so each contributes $\mathbb{E}[(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top] = \boldsymbol{\Sigma}$.

Step 2: induction on t. We show $\mathbb{E}[\boldsymbol{\Sigma}^{(t)}] = \boldsymbol{\Sigma} - \boldsymbol{C}_t$.

Base case $(t = 1)$. We have $N_1 = n_r$, and thus (33) gives

$$n_r \mathbb{E}[\boldsymbol{\Sigma}^{(1)}] = n_r \boldsymbol{\Sigma} - n_r \boldsymbol{C}_1 \Rightarrow \mathbb{E}[\boldsymbol{\Sigma}^{(1)}] = \boldsymbol{\Sigma} - \boldsymbol{C}_1.$$

Inductive step. Assume $\mathbb{E}[\boldsymbol{\Sigma}^{(s)}] = \boldsymbol{\Sigma} - \boldsymbol{C}_s$ for all $1 \le s \le t-1$. For a newly drawn sample $\boldsymbol{x}_i^{(s)}$ with $1 \le s \le t-1$, the tower property, identity (23) from Lemma A.1, the induction hypothesis, and $\mathbb{E}[\boldsymbol{\delta}^{(s)}(\boldsymbol{\delta}^{(s)})^\top] = \boldsymbol{C}_s$ (from (32)) give

$$\begin{aligned}
\mathbb{E}\big[(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})^\top\big] &= \mathbb{E}\left[\mathbb{E}\big[(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})^\top \mid \mathcal{F}_s\big]\right] \\
&= \mathbb{E}\left[\alpha_s \boldsymbol{\Sigma}^{(s)} + (1-\alpha_s)\boldsymbol{\Sigma} + \alpha_s\, \boldsymbol{\delta}^{(s)}(\boldsymbol{\delta}^{(s)})^\top\right] \\
&= \alpha_s\, \mathbb{E}[\boldsymbol{\Sigma}^{(s)}] + (1-\alpha_s)\boldsymbol{\Sigma} + \alpha_s \boldsymbol{C}_s \\
&= \alpha_s(\boldsymbol{\Sigma} - \boldsymbol{C}_s) + (1-\alpha_s)\boldsymbol{\Sigma} + \alpha_s \boldsymbol{C}_s = \boldsymbol{\Sigma}.
\end{aligned}$$

Substituting into (33) and using $n_r + \sum_{s=1}^{t-1} n_b(s) = N_t$,

$$N_t\, \mathbb{E}\big[\boldsymbol{\Sigma}^{(t)}\big] = n_r \boldsymbol{\Sigma} + \sum_{s=1}^{t-1} n_b(s)\, \boldsymbol{\Sigma} - N_t \boldsymbol{C}_t = N_t \boldsymbol{\Sigma} - N_t \boldsymbol{C}_t,$$

so $\mathbb{E}[\boldsymbol{\Sigma}^{(t)}] = \boldsymbol{\Sigma} - \boldsymbol{C}_t$. □

A.4 Steps towards proving Theorem 3.3

In this subsection, we present the steps towards proving Theorem 3.3.

A.4.1 Range of scalar sequences

First, we characterize the ranges of the three sequences defined in (18) and ρ_t, defined in (3). These ranges will be used extensively in the subsequent results.

Lemma A.9 (Range of λ_t, γ_t, ρ_t)**.** $0 < \lambda_t \le 1,\ 0 < \gamma_t < 1$ *and* $0 < \rho_t < 1$ *for all* $t \ge 1$.

Remark. For $1 \le s \le t$, (19), and Lemma A.9 imply $0 < \Gamma_{s,t} < 1$.

Proof. Fix $t \ge 1$. From (3) we have $N_{t+1} = N_t + n_b(t)$. Remember that $N_t \ge n_r \ge 1$ and $n_b(t) \ge 1$.

Range of λ_t. By (18) and $\alpha_t \in [0, 1]$, we have

$$\lambda_t = \frac{N_t + n_b(t)\alpha_t}{N_{t+1}} = \frac{N_t + n_b(t)\alpha_t}{N_t + n_b(t)} \Rightarrow 0 < \lambda_t \le 1$$

Range of ρ_t. By (3) and $n_b(t) \ge 1$, we have

$$\rho_t = \frac{N_t}{N_{t+1}} = \frac{N_t}{N_t + n_b(t)} \Rightarrow 0 < \rho_t < 1$$

Positivity $\gamma_t > 0$. By (18) we have

$$\beta_t = \frac{n_b(t)}{N_{t+1}^2} = \frac{n_b(t)}{(N_t + n_b(t))^2} \tag{34}$$

Substituting (34) into the definition $\gamma_t = \lambda_t^2 - \alpha_t^2\beta_t$ from (18),

$$\begin{aligned}\gamma_t &= \left(\frac{N_t + n_b(t)\alpha_t}{N_t + n_b(t)}\right)^2 - \frac{\alpha_t^2 n_b(t)}{\left(N_t + n_b(t)\right)^2} \\ &= \frac{\left(N_t + n_b(t)\alpha_t\right)^2 - \alpha_t^2 n_b(t)}{\left(N_t + n_b(t)\right)^2} \\ &= \frac{N_t^2 + 2N_t n_b(t)\alpha_t + \alpha_t^2 n_b(t)\left(n_b(t) - 1\right)}{\left(N_t + n_b(t)\right)^2} \\ &\geq \frac{N_t^2}{\left(N_t + n_b(t)\right)^2} > 0,\end{aligned} \tag{35}$$

where we used $2N_t n_b(t)\alpha_t \geq 0$ and $\alpha_t^2 n_b(t)\left(n_b(t) - 1\right) \geq 0$ (as $n_b(t) \geq 1$ and $\alpha_t \in [0, 1]$). Finally, $\gamma_t > 0$.

Upper bound $\gamma_t < 1$. From (35),

$$\begin{aligned}1 - \gamma_t &= \frac{\left(N_t + n_b(t)\right)^2 - \left(N_t + n_b(t)\alpha_t\right)^2 + \alpha_t^2 n_b(t)}{\left(N_t + n_b(t)\right)^2} \\ &= \frac{2N_t n_b(t)\left(1 - \alpha_t\right) + n_b^2(t)\left(1 - \alpha_t^2\right) + \alpha_t^2 n_b(t)}{\left(N_t + n_b(t)\right)^2} \\ &= \frac{(1 - \alpha_t) n_b(t) \cdot \left(2N_t + n_b(t)\left(1 + \alpha_t\right)\right) + \alpha_t^2 n_b(t)}{\left(N_t + n_b(t)\right)^2} \\ &= \frac{n_b(t)}{\left(N_t + n_b(t)\right)^2} \cdot \left(\left(1 - \alpha_t\right) \cdot \left(2N_t + n_b(t)\left(1 + \alpha_t\right)\right) + \alpha_t^2\right)\end{aligned}$$

That is,

$$1 - \gamma_t = \beta_t \cdot \left(\left(1 - \alpha_t\right) \cdot \left(2N_t + n_b(t)\left(1 + \alpha_t\right)\right) + \alpha_t^2\right) \tag{36}$$

If $0 \leq \alpha_t < 1$, the expression above is strictly positive. Otherwise, $\alpha_t = 1$, and we have

$$1 - \gamma_t = \frac{n_b(t)}{\left(N_t + n_b(t)\right)^2} > 0$$

In either case, $1 - \gamma_t > 0 \Rightarrow \gamma_t < 1$. □

A.4.2 Consistency of $\boldsymbol{\mu}^{(t)}$

We start by characterizing when $\mathbb{E}\left[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\right] \xrightarrow{t\to\infty} 0$. First, we provide an equivalent condition in terms of the covariance of $\boldsymbol{\mu}^{(t)}$.

Lemma A.10 (L^2 consistency equivalence)**.** *Under Assumption 1, we have*

$$\boldsymbol{C}_t \to \mathbf{0} \Leftrightarrow \mathbb{E}\left[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\right] \xrightarrow{t\to\infty} 0. \tag{37}$$

Remark. We note $\boldsymbol{C}_t \to \mathbf{0}$ with the Frobenius norm. This is valid because $\boldsymbol{C}_t \in \mathbb{R}^{d\times d}$ is in a finite dimensional space, which means that all norms are equivalent. Additionally, from Theorem 3.1(b), Equation (37) is equivalent to the asymptotic unbiasedness of the covariance estimator $\boldsymbol{\Sigma}^{(t)}$.

Proof. By (11), $\mathbb{E}\left[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\right] \xrightarrow{t\to\infty} 0$ if and only if $\mathrm{Tr}(\boldsymbol{C}_t) \to 0$. By (10), the matrix $\boldsymbol{C}_t$ is symmetric, and it is positive-semi-definite as a covariance matrix. Hence, Lemma A.4(b) gives $\mathrm{Tr}(\boldsymbol{C}_t) \to 0 \iff \boldsymbol{C}_t \to \mathbf{0}$. Combining the two equivalences gives the goal equivalence. □

Next, let us provide a closed-form formula for the covariance of $\boldsymbol{\mu}^{(t)}$.

Theorem A.11 (Covariance of $\boldsymbol{\mu}^{(t)}$)**.** *Under Assumption 1, with $\lambda_t, \beta_t, \gamma_t$ as in* (18)*, and $\Gamma_{s,t}$ as in* (19)*:*

(a) ***Recursion.***

$$\boldsymbol{C}_t = \gamma_{t-1}\boldsymbol{C}_{t-1} + \beta_{t-1}\boldsymbol{\Sigma}, \quad t \geq 2, \qquad \boldsymbol{C}_1 = \frac{\boldsymbol{\Sigma}}{n_r}. \tag{38}$$

(b) ***Closed form.***

$$\boldsymbol{C}_t = \Gamma_{1,t-1}\,\boldsymbol{C}_1 + \boldsymbol{\Sigma}\sum_{s=1}^{t-1}\beta_s\,\Gamma_{s+1,t-1}. \tag{39}$$

(c) ***Special case*** $(\alpha_t \equiv 0)$***.*** *The covariance is $\boldsymbol{C}_t = \boldsymbol{\Sigma}/N_t$. This is the covariance of the sample mean of N_t i.i.d. real samples (indeed, no generated data is used).*

(d) ***Special case*** $(\alpha_t \equiv 1)$***.*** *The recursion becomes $\boldsymbol{C}_t = (1-\beta_{t-1})\boldsymbol{C}_{t-1} + \beta_{t-1}\boldsymbol{\Sigma}$. In this case, all new samples are synthetic.*

(e) ***Scalar form.*** *The covariance is a scalar multiple of the true covariance, $\boldsymbol{C}_t = c_t\boldsymbol{\Sigma}$, where*

$$c_1 = \frac{1}{n_r}, \qquad c_t = \gamma_{t-1}c_{t-1} + \beta_{t-1} \quad (t \geq 2). \tag{40}$$

Remark. In the recursion (38), $\gamma_{t-1} \in (0,1)$ (Lemma A.9) acts as a damping factor on the covariance from the previous iteration, while the added terms $\beta_{t-1}\boldsymbol{\Sigma}$ are summable (Lemma A.5). Together, these yield the bound $\boldsymbol{C}_t \preceq \frac{2}{n_r}\boldsymbol{\Sigma}$ of Lemma 3.2.

Proof of part (a): recursion. Apply the law of total covariance to $\boldsymbol{\mu}^{(t)}$, conditioning on $\mathcal{F}_{t-1}$:

$$\boldsymbol{C}_t = \mathrm{Cov}\big(\mathbb{E}[\boldsymbol{\mu}^{(t)} \mid \mathcal{F}_{t-1}]\big) + \mathbb{E}\big[\mathrm{Cov}(\boldsymbol{\mu}^{(t)} \mid \mathcal{F}_{t-1})\big]. \tag{41}$$

Term 1 (covariance of the conditional mean). By (29) and (21) from Lemma A.1,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}^{(t)} \mid \mathcal{F}_{t-1}] &= \mathbb{E}\left[\rho_{t-1}\boldsymbol{\mu}^{(t-1)} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)}\boldsymbol{x}_i^{(t-1)} \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= \rho_{t-1}\boldsymbol{\mu}^{(t-1)} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)}\mathbb{E}[\boldsymbol{x}_i^{(t-1)} \mid \mathcal{F}_{t-1}] \\
&= \rho_{t-1}\boldsymbol{\mu}^{(t-1)} + \frac{n_b(t-1)}{N_t}\cdot\Big(\alpha_{t-1}\boldsymbol{\mu}^{(t-1)} + (1-\alpha_{t-1})\boldsymbol{\mu}\Big) \\
&= \frac{N_{t-1} + n_b(t-1)\alpha_{t-1}}{N_t}\boldsymbol{\mu}^{(t-1)} + \frac{(1-\alpha_{t-1})n_b(t-1)}{N_t}\boldsymbol{\mu} \\
&= \lambda_{t-1}\boldsymbol{\mu}^{(t-1)} + (1-\lambda_{t-1})\boldsymbol{\mu}
\end{aligned} \tag{42}$$

where we used the definition of λ_t in (18), and

$$1 - \lambda_t = 1 - \frac{N_t + n_b(t)\alpha_t}{N_{t+1}} = \frac{N_{t+1} - N_t - n_b(t)\alpha_t}{N_{t+1}} = \frac{(1-\alpha_t)\,n_b(t)}{N_{t+1}}.$$

As λ_{t-1} is deterministic and $(1-\lambda_{t-1})\boldsymbol{\mu}$ is constant, from (42) we have

$$\mathrm{Cov}\big(\mathbb{E}[\boldsymbol{\mu}^{(t)} \mid \mathcal{F}_{t-1}]\big) = \lambda_{t-1}^2\,\mathrm{Cov}(\boldsymbol{\mu}^{(t-1)}) = \lambda_{t-1}^2\,\boldsymbol{C}_{t-1}. \tag{43}$$

Term 2 (expectation of the conditional covariance). By (29), the conditional covariance reads:

$$\begin{aligned}
\mathrm{Cov}\big(\boldsymbol{\mu}^{(t)} \mid \mathcal{F}_{t-1}\big) &= \mathrm{Cov}\left(\rho_{t-1}\boldsymbol{\mu}^{(t-1)} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)}\boldsymbol{x}_i^{(t-1)} \,\middle|\, \mathcal{F}_{t-1}\right) \\
&= \frac{n_b(t-1)}{N_t^2}\,\mathrm{Cov}\big(\boldsymbol{x}_1^{(t-1)} \mid \mathcal{F}_{t-1}\big) \\
&= \beta_{t-1}\,\mathrm{Cov}\big(\boldsymbol{x}_1^{(t-1)} \mid \mathcal{F}_{t-1}\big)
\end{aligned}$$

where we used that given $\mathcal{F}_{t-1}$, the term $\rho_{t-1}\boldsymbol{\mu}^{(t-1)}$ is deterministic, and the $n_b(t-1)$ new samples are i.i.d. (Assumption 1), so the cross-covariances vanish. We also used the definition of β_t in (18). Using (22) from Lemma A.1 with $t-1$, the conditional covariance is:

$$\mathrm{Cov}\big(\boldsymbol{\mu}^{(t)} \mid \mathcal{F}_{t-1}\big) = \beta_{t-1}\left(\alpha_{t-1}\boldsymbol{\Sigma}^{(t-1)} + (1-\alpha_{t-1})\boldsymbol{\Sigma} + \alpha_{t-1}(1-\alpha_{t-1})\boldsymbol{\delta}^{(t-1)}(\boldsymbol{\delta}^{(t-1)})^\top\right). \tag{44}$$

Taking the outer expectation and using $\mathbb{E}[\boldsymbol{\Sigma}^{(t-1)}] = \boldsymbol{\Sigma} - \boldsymbol{C}_{t-1}$ (Theorem 3.1(b)) together with $\mathbb{E}[\boldsymbol{\delta}^{(t-1)}(\boldsymbol{\delta}^{(t-1)})^\top] = \boldsymbol{C}_{t-1}$ ($\boldsymbol{\delta}^{(t-1)} = \boldsymbol{\mu}^{(t-1)} - \boldsymbol{\mu}$ and (10)), from (44) we have

$$\begin{aligned}
\mathbb{E}\big[\mathrm{Cov}(\boldsymbol{\mu}^{(t)} \mid \mathcal{F}_{t-1})\big] &= \beta_{t-1}\Big(\alpha_{t-1}(\boldsymbol{\Sigma} - \boldsymbol{C}_{t-1}) + (1-\alpha_{t-1})\boldsymbol{\Sigma} + \alpha_{t-1}(1-\alpha_{t-1})\boldsymbol{C}_{t-1}\Big) \\
&= \beta_{t-1}\Big(\alpha_{t-1}\boldsymbol{\Sigma} - \alpha_{t-1}\boldsymbol{C}_{t-1} + (1-\alpha_{t-1})\boldsymbol{\Sigma} + \alpha_{t-1}\boldsymbol{C}_{t-1} - \alpha_{t-1}^2\boldsymbol{C}_{t-1}\Big) \\
&= \beta_{t-1}\big(\boldsymbol{\Sigma} - \alpha_{t-1}^2\boldsymbol{C}_{t-1}\big).
\end{aligned} \tag{45}$$

Combining. From (41), (43) and (45), the covariance $\boldsymbol{C}_t$ reads:

$$\begin{aligned}
\boldsymbol{C}_t &= \lambda_{t-1}^2\boldsymbol{C}_{t-1} + \beta_{t-1}(\boldsymbol{\Sigma} - \alpha_{t-1}^2\boldsymbol{C}_{t-1}) \\
&= (\lambda_{t-1}^2 - \alpha_{t-1}^2\beta_{t-1})\boldsymbol{C}_{t-1} + \beta_{t-1}\boldsymbol{\Sigma} \\
&= \gamma_{t-1}\boldsymbol{C}_{t-1} + \beta_{t-1}\boldsymbol{\Sigma}
\end{aligned}$$

where we used the definition of γ_t in (18). The initial condition is

$$\boldsymbol{C}_1 = \mathrm{Cov}\left(\frac{1}{n_r}\sum_{j=1}^{n_r}\boldsymbol{x}_j^{(0)}\right) = \frac{1}{n_r^2}\cdot n_r\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma}}{n_r}. \tag{46}$$

□

Proof of part (b): closed form. We prove (39) via induction on $t \geq 1$.

- For $t = 1$, the right-hand side of (39) is $\Gamma_{1,0}\boldsymbol{C}_1 = \boldsymbol{C}_1$ (empty sum; $\Gamma_{1,0} = 1$ by (19)).
- Assuming $\boldsymbol{C}_{t-1} = \Gamma_{1,t-2}\boldsymbol{C}_1 + \boldsymbol{\Sigma}\sum_{s=1}^{t-2}\beta_s\Gamma_{s+1,t-2}$ and substituting into the recursion (38),

$$\begin{aligned}
\boldsymbol{C}_t &= \gamma_{t-1}\Big(\Gamma_{1,t-2}\boldsymbol{C}_1 + \boldsymbol{\Sigma}\sum_{s=1}^{t-2}\beta_s\Gamma_{s+1,t-2}\Big) + \beta_{t-1}\boldsymbol{\Sigma} \\
&= \underbrace{\gamma_{t-1}\Gamma_{1,t-2}}_{\Gamma_{1,t-1}}\boldsymbol{C}_1 + \boldsymbol{\Sigma}\sum_{s=1}^{t-2}\beta_s\underbrace{\gamma_{t-1}\Gamma_{s+1,t-2}}_{\Gamma_{s+1,t-1}} + \beta_{t-1}\boldsymbol{\Sigma}\underbrace{\Gamma_{t,t-1}}_{=1} \\
&= \Gamma_{1,t-1}\boldsymbol{C}_1 + \boldsymbol{\Sigma}\sum_{s=1}^{t-1}\beta_s\Gamma_{s+1,t-1},
\end{aligned}$$

using $\gamma_{t-1}\Gamma_{s,t-2} = \Gamma_{s,t-1}$ and the empty product $\Gamma_{t,t-1} = 1$, so that $\beta_{t-1}\boldsymbol{\Sigma}$ becomes the $s = t-1$ term of the sum. We used the definition of $\Gamma_{s,t}$ in (19) throughout.

□

Proof of part (c): $\alpha_t \equiv 0$. Assume $\alpha_t \equiv 0$. From (18) we have:

$$\gamma_t = \lambda_t^2 = \frac{N_t^2}{N_{t+1}^2}, \qquad \beta_t = \frac{n_b(t)}{N_{t+1}^2}. \tag{47}$$

We prove that $\boldsymbol{C}_t = \boldsymbol{\Sigma}/N_t$ by induction on $t \geq 1$:

- For $t = 1$, using $N_1 = n_r$, from (46) we have

$$\boldsymbol{C}_1 = \frac{\boldsymbol{\Sigma}}{n_r} = \frac{\boldsymbol{\Sigma}}{N_1}$$

- Assuming $\boldsymbol{C}_{t-1} = \boldsymbol{\Sigma}/N_{t-1}$, the recursion (38) gives

$$\begin{aligned}\boldsymbol{C}_t &= \gamma_{t-1}\boldsymbol{C}_{t-1} + \beta_{t-1}\boldsymbol{\Sigma} \\ &= \frac{N_{t-1}^2}{N_t^2} \cdot \frac{\boldsymbol{\Sigma}}{N_{t-1}} + \frac{n_b(t-1)}{N_t^2}\boldsymbol{\Sigma} \\ &= \frac{N_{t-1} + n_b(t-1)}{N_t^2}\boldsymbol{\Sigma} \\ &= \frac{\boldsymbol{\Sigma}}{N_t}.\end{aligned}$$

  where we used the formulas for γ_t, β_t from (47), and $N_{t-1} + n_b(t-1) = N_t$ from (3).

□

Proof of part (d): $\alpha_t \equiv 1$. Assume $\alpha_t \equiv 1$. From $N_t + n_b(t) = N_{t+1}$ (follows from (3)) we have

$$\lambda_t = \frac{N_t + n_b(t)\alpha_t}{N_{t+1}} = \frac{N_t + n_b(t)}{N_{t+1}} = 1$$

Hence, $\gamma_t = \lambda_t^2 - \alpha_t^2\beta_t = 1 - \beta_t$. The recursion (38) then reads $\boldsymbol{C}_t = (1-\beta_{t-1})\boldsymbol{C}_{t-1} + \beta_{t-1}\boldsymbol{\Sigma}$. □

Proof of part (e): scalar form. We prove by induction on $t \geq 1$. For $t = 1$, $\boldsymbol{C}_1 = \boldsymbol{\Sigma}/n_r = c_1\boldsymbol{\Sigma}$ by (46). Assuming $\boldsymbol{C}_{t-1} = c_{t-1}\boldsymbol{\Sigma}$, the recursion (38) gives

$$\boldsymbol{C}_t = \gamma_{t-1}c_{t-1}\boldsymbol{\Sigma} + \beta_{t-1}\boldsymbol{\Sigma} = c_t\boldsymbol{\Sigma}.$$

□

We can now prove Lemma 3.2, stated in the main text.

Proof of Lemma 3.2. The lower bound $\boldsymbol{C}_t \succeq \mathbf{0}$ holds since $\boldsymbol{C}_t$ is a covariance matrix (6).

For the upper bound, by Theorem A.11(e) we have $\boldsymbol{C}_t = c_t\boldsymbol{\Sigma}$, with c_t given by (40). First, $c_t \geq 0$ for every $t \geq 1$ by induction, since $c_1 = 1/n_r > 0$, and $\gamma_t > 0$ and $\beta_t > 0$ for every $t \geq 1$ (Lemma A.9 and (18)). Hence, using $\gamma_{t-1} < 1$ (Lemma A.9),

$$c_t = \gamma_{t-1}c_{t-1} + \beta_{t-1} \leq c_{t-1} + \beta_{t-1}, \qquad t \geq 2.$$

Unrolling this inequality and applying Lemma A.5,

$$c_t \leq c_1 + \sum_{s=1}^{t-1}\beta_s \leq \frac{1}{n_r} + \frac{1}{n_r} = \frac{2}{n_r}, \qquad t \geq 1,$$

where for $t = 1$ the sum is empty. Since $\boldsymbol{\Sigma} \succ 0$, we conclude

$$\boldsymbol{C}_t = c_t\boldsymbol{\Sigma} \preceq \frac{2}{n_r}\boldsymbol{\Sigma}.$$

□

We need to characterize when the estimated mean converges to the truth. By Lemma A.10, this is equivalent to characterizing when $\boldsymbol{C}_t \xrightarrow{t\to\infty} \mathbf{0}$. The following theorem addresses this.

Theorem A.12 (General necessary and sufficient condition)**.** *Under Assumption 1, with $\lambda_t, \beta_t, \gamma_t$ as in* (18)*, $\Gamma_{s,t}$ as in* (19)*, $\boldsymbol{C}_t \to 0$ as $t \to \infty$ if and only if*

$$\Gamma_{1,t-1} \to 0 \qquad \textit{and} \qquad S_t \coloneqq \sum_{s=1}^{t-1} \beta_s \, \Gamma_{s+1,t-1} \to 0. \tag{48}$$

Remark. From (37), it follows that Theorem A.12 provides a necessary and sufficient condition for $\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \xrightarrow{t\to\infty} 0$.

Proof. By the closed form of Theorem A.11(b), (39), we have

$$\boldsymbol{C}_t = \Gamma_{1,t-1}\boldsymbol{C}_1 + S_t \, \boldsymbol{\Sigma}, \qquad S_t = \sum_{s=1}^{t-1} \beta_s \, \Gamma_{s+1,t-1}. \tag{49}$$

Sufficiency. Suppose $\Gamma_{1,t-1} \to 0$ and $S_t \to 0$. By the triangle inequality and $\|c\boldsymbol{A}\|_F = |c| \, \|\boldsymbol{A}\|_F$ for a scalar c,

$$0 \le \|\boldsymbol{C}_t\|_F \le \Gamma_{1,t-1}\|\boldsymbol{C}_1\|_F + S_t\|\boldsymbol{\Sigma}\|_F \longrightarrow 0,$$

since $\|\boldsymbol{C}_1\|_F$ and $\|\boldsymbol{\Sigma}\|_F$ are finite constants. We also used that $\Gamma_{1,t-1} \ge 0, S_t \ge 0$. Hence $\boldsymbol{C}_t \to \boldsymbol{0}$.

Necessity. Suppose $\boldsymbol{C}_t \to \boldsymbol{0}$. By (10), the matrix $\boldsymbol{C}_t$ is symmetric, and it is positive-semi-definite as a covariance matrix. Hence, Lemma A.4(b) gives $\mathrm{Tr}(\boldsymbol{C}_t) \to 0$. From (49), we have

$$\mathrm{Tr}(\boldsymbol{C}_t) = \Gamma_{1,t-1} \, \mathrm{Tr}(\boldsymbol{C}_1) + S_t \, \mathrm{Tr}(\boldsymbol{\Sigma}).$$

Both summands are nonnegative: $\Gamma_{1,t-1} \ge 0$ and $S_t \ge 0$ by (19) together with $\gamma_t \in (0,1)$ and $\beta_t > 0$ (Lemma A.9), while $\mathrm{Tr}(\boldsymbol{C}_1) = \mathrm{Tr}(\boldsymbol{\Sigma})/n_r > 0$ and $\mathrm{Tr}(\boldsymbol{\Sigma}) > 0$ since $\boldsymbol{\Sigma} \succ 0$. Discarding one summand at a time and dividing by these two positive constants gives

$$0 \le \Gamma_{1,t-1} \le \frac{\mathrm{Tr}(\boldsymbol{C}_t)}{\mathrm{Tr}(\boldsymbol{C}_1)} \longrightarrow 0, \qquad 0 \le S_t \le \frac{\mathrm{Tr}(\boldsymbol{C}_t)}{\mathrm{Tr}(\boldsymbol{\Sigma})} \longrightarrow 0.$$

Hence, $\Gamma_{1,t-1} \to 0$ and $S_t \to 0$. □

Next, we show that the condition for $\boldsymbol{C}_t \xrightarrow{t\to\infty} \boldsymbol{0}$ collapses to a single elementary series in $\{\alpha_t\}_{t\ge1}$, $\{n_b(t)\}_{t\ge1}$, and n_r.

Theorem A.13 (Sharp elementary criterion for consistency)**.** *Under Assumption 1, with ρ_t as in* (3)*,*

$$\boldsymbol{C}_t \to \boldsymbol{0} \quad \Longleftrightarrow \quad \sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \sum_{t=1}^{\infty}(1-\alpha_t)\,\frac{n_b(t)}{N_{t+1}} = \infty, \tag{50}$$

and equivalently for $\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \xrightarrow{t\to\infty} 0$. In particular, $S_t \to 0$ whenever $\Gamma_{1,t-1} \to 0$, so the second condition in (48) *is redundant.*

Proof. It is convenient to route the argument through two intermediate conditions. Write

$$\text{(a) } \boldsymbol{C}_t \to \boldsymbol{0}, \qquad \text{(b) } \Gamma_{1,t-1} \to 0, \qquad \text{(c) } \sum_{t=1}^{\infty}(1-\gamma_t) = \infty, \qquad \text{(d) } \sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty,$$

so that the assertion of the theorem is (a)⇔(d). Theorem A.12 already gives (a)⇒(b); we prove (b)⇒(a), (b)⇔(c) and (c)⇔(d), which together yield (a)⇔(d) as well as the equivalences among (b), (c) and (d). The final restatement in terms of $\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \xrightarrow{t\to\infty} 0$ is Lemma A.10.

Step 1: (b)⇒(a). We first prove that $\Gamma_{1,t-1} \to 0$ implies $S_t \to 0$.

Assume that $\Gamma_{1,t-1} \to 0$, and fix $\varepsilon > 0$. By Lemma A.5 choose s_0 with $\sum_{s>s_0} \beta_s < \varepsilon$, and for $t-1 > s_0$ split

$$S_t = \sum_{s=1}^{t-1} \beta_s \Gamma_{s+1,t-1} = \underbrace{\sum_{s=1}^{s_0} \beta_s \Gamma_{s+1,t-1}}_{A_t} + \underbrace{\sum_{s=s_0+1}^{t-1} \beta_s \Gamma_{s+1,t-1}}_{B_t}.$$

By Lemma A.9 every $\gamma_k \in (0,1)$, so each $\Gamma_{s+1,t-1} = \prod_{k=s+1}^{t-1} \gamma_k \le 1$ and $B_t \le \sum_{s>s_0} \beta_s < \varepsilon$. In addition, for $s \le s_0$, from the definition of $\Gamma_{s,t}$ in (19):

$$\Gamma_{s+1,t-1} = \Gamma_{s_0+1,t-1} \cdot \Gamma_{s+1,s_0} \le \Gamma_{s_0+1,t-1}$$

Thus,

$$0 \le A_t \le \Gamma_{s_0+1,t-1} \cdot \sum_{s=1}^{s_0} \beta_s = \frac{\sum_{s=1}^{s_0} \beta_s}{\Gamma_{1,s_0}} \cdot \Gamma_{1,t-1} \to 0$$

That is, $A_t \to 0$. Therefore, $\limsup_t S_t \le \varepsilon$, and as $\varepsilon > 0$ was arbitrary and $S_t \ge 0$ we get $S_t \to 0$. Both conditions of (48) now hold, so Theorem A.12 gives $\boldsymbol{C}_t \to \mathbf{0}$. This also proves the final assertion of the theorem.

Step 2: (b)⇔(c). Assume (c). Since $\gamma_t \in (0,1)$, the bound $\ln u \le u - 1$ gives

$$\Gamma_{1,t-1} = \prod_{s=1}^{t-1} \gamma_s = \exp\left(\sum_{s=1}^{t-1} \ln \gamma_s\right) \le \exp\left(-\sum_{s=1}^{t-1} (1-\gamma_s)\right) \xrightarrow[t\to\infty]{} 0,$$

which, together with $\Gamma_{1,t-1} \ge 0$, is (b). Conversely, (b) together with $\gamma_t \in (0,1)$ is precisely the hypothesis of Lemma A.6, whose conclusion is (c).

Step 3: (c)⇔(d). Remember that from (3) and (18), we have $N_{t+1} = N_t + n_b(t)$ and $\beta_t = \frac{n_b(t)}{N_{t+1}^2}$. Using the identity (36), we have

$$1 - \gamma_t = \beta_t \big[(1-\alpha_t) K_t + \alpha_t^2\big], \qquad K_t \coloneqq 2N_t + n_b(t)(1+\alpha_t). \tag{51}$$

Because $0 \le \alpha_t^2 \le 1$, Lemma A.5 yields

$$\sum_{t=1}^{\infty} \beta_t \alpha_t^2 \le \sum_{t=1}^{\infty} \beta_t \le \frac{1}{n_r} < \infty.$$

Thus, the second term in (51) is summable and can never cause divergence. Hence

$$\sum_{t=1}^{\infty} (1-\gamma_t) = \infty \iff \sum_{t=1}^{\infty} \beta_t (1-\alpha_t) K_t = \infty. \tag{52}$$

We bound K_t on both sides by a multiple of N_{t+1}. Since $N_t \ge 0$ and $\alpha_t \ge 0$,

$$K_t = 2N_t + n_b(t)(1+\alpha_t) \ge N_t + n_b(t) = N_{t+1},$$

and since $\alpha_t \le 1$,

$$K_t \le 2N_t + 2n_b(t) = 2N_{t+1}.$$

That is,

$$(1-\alpha_t)\beta_t N_{t+1} \le \beta_t (1-\alpha_t) K_t \le 2(1-\alpha_t)\beta_t N_{t+1}.$$

By (3), we have $\beta_t N_{t+1} = \frac{n_b(t)}{N_{t+1}} = 1 - \frac{N_t}{N_{t+1}} = 1 - \rho_t$. Thus:

$$(1-\alpha_t)(1-\rho_t) \le \beta_t (1-\alpha_t) K_t \le 2(1-\alpha_t)(1-\rho_t).$$

The two series therefore diverge together, which with (52) gives (c)⇔(d). □

A.4.3 CONSISTENCY OF $\boldsymbol{\Sigma}^{(t)}$

Next, we characterize when $\mathbb{E}\Big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\Big] \xrightarrow{t\to\infty} 0$. We start with eight technical lemmas. First, we denote the Gaussian fourth-moment constant

$$C_{\boldsymbol{\Sigma}} \coloneqq \mathrm{Tr}(\boldsymbol{\Sigma})^2 + 2\|\boldsymbol{\Sigma}\|_F^2. \tag{53}$$

It equals $\mathbb{E}\|\boldsymbol{g}\|_2^4$ for $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$: by Isserlis' theorem $\mathbb{E}[g_i^2 g_j^2] = \boldsymbol{\Sigma}_{ii}\boldsymbol{\Sigma}_{jj} + 2\boldsymbol{\Sigma}_{ij}^2$, so $\mathbb{E}\|\boldsymbol{g}\|_2^4 = \sum_{i,j}(\boldsymbol{\Sigma}_{ii}\boldsymbol{\Sigma}_{jj} + 2\boldsymbol{\Sigma}_{ij}^2) = \mathrm{Tr}(\boldsymbol{\Sigma})^2 + 2\|\boldsymbol{\Sigma}\|_F^2$.

First, we isolate the dependence of the covariance estimator on the mean error.

Lemma A.14 (Real-centered second-moment statistic)**.** *Define*

$$\hat{\boldsymbol{S}}_t \coloneqq \frac{1}{N_t}\left(\sum_{j=1}^{n_r}(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top + \sum_{s=1}^{t-1}\sum_{i=1}^{n_b(s)}(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})^\top\right). \tag{54}$$

Then, using (7) ($\boldsymbol{\delta}^{(t)} = \boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}$)*:*

(i) $\hat{\boldsymbol{S}}_t \succeq 0$ *almost surely;*

(ii) $\boldsymbol{\Sigma}^{(t)} = \hat{\boldsymbol{S}}_t - \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top$*. Thus,* $\mathrm{Tr}(\boldsymbol{\Sigma}^{(t)}) = \mathrm{Tr}(\hat{\boldsymbol{S}}_t) - \|\boldsymbol{\delta}^{(t)}\|_2^2$*;*

(iii) for $t \geq 2$*,* $\hat{\boldsymbol{S}}_t = \rho_{t-1}\hat{\boldsymbol{S}}_{t-1} + \dfrac{1}{N_t}\displaystyle\sum_{i=1}^{n_b(t-1)}(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})^\top$*;*

(iv) for $t \geq 2$*,* $\mathbb{E}[\hat{\boldsymbol{S}}_t \mid \mathcal{F}_{t-1}] = \lambda_{t-1}\hat{\boldsymbol{S}}_{t-1} + (1-\lambda_{t-1})\boldsymbol{\Sigma}$*, and consequently* $\mathbb{E}[\hat{\boldsymbol{S}}_t] = \boldsymbol{\Sigma}$ *for all* $t \geq 1$*.*

Proof. (i) $\hat{\boldsymbol{S}}_t$ is a non-negatively weighted sum of outer products. Hence, it is symmetric PSD.

(ii) Using the pooled-scatter identity (25) and recalling (7) ($\boldsymbol{\delta}^{(t)} = \boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}$) gives $\boldsymbol{\Sigma}^{(t)} = \hat{\boldsymbol{S}}_t - \boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top$; the trace identity follows from $\mathrm{Tr}(\boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top) = \|\boldsymbol{\delta}^{(t)}\|_2^2$.

(iii) Fix $t \geq 2$. From (54), we have

$$\begin{aligned}
\hat{\boldsymbol{S}}_t &= \frac{1}{N_t}\left(\sum_{j=1}^{n_r}(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top + \sum_{s=1}^{t-1}\sum_{i=1}^{n_b(s)}(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})^\top\right) \\
&= \frac{1}{N_t}\left(\sum_{j=1}^{n_r}(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top + \sum_{s=1}^{t-2}\sum_{i=1}^{n_b(s)}(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(s)} - \boldsymbol{\mu})^\top + \sum_{i=1}^{n_b(t-1)}(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})^\top\right) \\
&= \frac{N_{t-1}}{N_t}\hat{\boldsymbol{S}}_{t-1} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)}(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})^\top \\
&= \rho_{t-1}\hat{\boldsymbol{S}}_{t-1} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)}(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})^\top
\end{aligned}$$

We used the definition of ρ_t from (3).

(iv) From (iii), we have

$$\begin{aligned}
\mathbb{E}\big[\hat{\boldsymbol{S}}_t \mid \mathcal{F}_{t-1}\big] &= \rho_{t-1}\mathbb{E}\big[\hat{\boldsymbol{S}}_{t-1} \mid \mathcal{F}_{t-1}\big] + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)} \mathbb{E}\big[(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})^\top \mid \mathcal{F}_{t-1}\big] \\
&= \rho_{t-1}\hat{\boldsymbol{S}}_{t-1} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)} \big(\alpha_{t-1}\boldsymbol{\Sigma}^{(t-1)} + (1-\alpha_{t-1})\boldsymbol{\Sigma} + \alpha_{t-1}\boldsymbol{\delta}^{(t-1)}(\boldsymbol{\delta}^{(t-1)})^\top\big) \\
&= \frac{N_{t-1}}{N_t}\hat{\boldsymbol{S}}_{t-1} + \frac{1}{N_t}\sum_{i=1}^{n_b(t-1)} \big(\alpha_{t-1}\hat{\boldsymbol{S}}_{t-1} + (1-\alpha_{t-1})\boldsymbol{\Sigma}\big) \\
&= \frac{N_{t-1} + \alpha_{t-1}n_b(t-1)}{N_t}\hat{\boldsymbol{S}}_{t-1} + \frac{(1-\alpha_{t-1})n_b(t-1)}{N_t}\boldsymbol{\Sigma} \\
&= \lambda_{t-1}\hat{\boldsymbol{S}}_{t-1} + (1-\lambda_{t-1})\boldsymbol{\Sigma}.
\end{aligned}$$

We used

- $\hat{\boldsymbol{S}}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable.
- $\rho_{t-1} = N_{t-1}/N_t$ from (3).
- From (23) (Lemma A.1), we have

  $$\mathbb{E}\big[(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})^\top \mid \mathcal{F}_{t-1}\big] = \alpha_{t-1}\boldsymbol{\Sigma}^{(t-1)} + (1-\alpha_{t-1})\boldsymbol{\Sigma} + \alpha_{t-1}\boldsymbol{\delta}^{(t-1)}(\boldsymbol{\delta}^{(t-1)})^\top$$

- $\boldsymbol{\Sigma}^{(t-1)} + \boldsymbol{\delta}^{(t-1)}(\boldsymbol{\delta}^{(t-1)})^\top = \hat{\boldsymbol{S}}_{t-1}$ from (ii) at index $t-1$.
- The definition of λ_t from (18), and

  $$1 - \lambda_t = 1 - \frac{N_t + n_b(t)\alpha_t}{N_{t+1}} = \frac{N_{t+1} - N_t - n_b(t)\alpha_t}{N_{t+1}} = \frac{(1-\alpha_t)\, n_b(t)}{N_{t+1}}.$$

That is, we proved that

$$\mathbb{E}\big[\hat{\boldsymbol{S}}_t \mid \mathcal{F}_{t-1}\big] = \lambda_{t-1}\hat{\boldsymbol{S}}_{t-1} + (1-\lambda_{t-1})\boldsymbol{\Sigma}. \tag{55}$$

Finally, we prove unbiasedness by induction on $t \geq 1$. At $t = 1$, we have

$$\hat{\boldsymbol{S}}_1 = \frac{1}{n_r}\sum_{j=1}^{n_r}(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top \Rightarrow \mathbb{E}[\hat{\boldsymbol{S}}_1] = \frac{1}{n_r}\sum_{j=1}^{n_r}\mathbb{E}\big[(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top\big] = \boldsymbol{\Sigma}$$

Assume that $\mathbb{E}[\hat{\boldsymbol{S}}_{t-1}] = \boldsymbol{\Sigma}$. From (55) and the tower property, we have

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{S}}_t] &= \mathbb{E}\big[\mathbb{E}[\hat{\boldsymbol{S}}_t \mid \mathcal{F}_{t-1}]\big] \\
&= \mathbb{E}[\lambda_{t-1}\hat{\boldsymbol{S}}_{t-1} + (1-\lambda_{t-1})\boldsymbol{\Sigma}] \\
&= \lambda_{t-1}\mathbb{E}[\hat{\boldsymbol{S}}_{t-1}] + (1-\lambda_{t-1})\boldsymbol{\Sigma} \\
&= \boldsymbol{\Sigma}.
\end{aligned}$$

□

We now state a technical lemma that will be helpful later on.

Lemma A.15 (Rayleigh bound)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be symmetric with $\boldsymbol{A} \succeq 0$. Then for every $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \;\leq\; \mathrm{Tr}(\boldsymbol{A})\,\|\boldsymbol{x}\|_2^2.$$

Proof. By the spectral theorem, $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$ with $\boldsymbol{Q}$ orthogonal and $\boldsymbol{D} = \operatorname{diag}(\nu_1, \dots, \nu_d)$, where $\nu_1, \dots, \nu_d$ are the eigenvalues of $\boldsymbol{A}$: real because $\boldsymbol{A}$ is symmetric, and nonnegative because $\boldsymbol{A} \succeq 0$, since $\nu_i = \boldsymbol{q}_i^\top \boldsymbol{A}\boldsymbol{q}_i \geq 0$ for the corresponding unit eigenvector $\boldsymbol{q}_i$.

Fix $\boldsymbol{x} \in \mathbb{R}^d$ and set $\boldsymbol{y} \coloneqq \boldsymbol{Q}^\top \boldsymbol{x}$. Since $\boldsymbol{Q}$ is orthogonal, $\|\boldsymbol{y}\|_2 = \|\boldsymbol{x}\|_2$, and

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top \boldsymbol{x} = \boldsymbol{y}^\top \boldsymbol{D}\boldsymbol{y} = \sum_{i=1}^{d} \nu_i \, y_i^2.$$

Every $y_i^2 \geq 0$, so bounding each ν_i by $\nu_{\max} \coloneqq \max_{1 \leq j \leq d} \nu_j$ gives

$$\sum_{i=1}^{d} \nu_i \, y_i^2 \;\leq\; \nu_{\max} \sum_{i=1}^{d} y_i^2 \;=\; \nu_{\max} \, \|\boldsymbol{y}\|_2^2 \;=\; \nu_{\max} \, \|\boldsymbol{x}\|_2^2.$$

Finally, all eigenvalues are nonnegative, so discarding the remaining $d - 1$ of them from the trace yields

$$\nu_{\max} \;\leq\; \sum_{i=1}^{d} \nu_i \;=\; \operatorname{Tr}(\boldsymbol{A}),$$

and combining the last two displays gives the claim. □

Lemma A.16 (Conditional fourth-moment bound)**.** *For every $t \geq 1$ and each newly drawn sample $\boldsymbol{x}_i^{(t)}$,*

$$\mathbb{E}\big[\|\boldsymbol{x}_i^{(t)} - \boldsymbol{\mu}\|_2^4 \mid \mathcal{F}_t\big] \leq 3\alpha_t \, \operatorname{Tr}(\hat{\boldsymbol{S}}_t)^2 + C_{\boldsymbol{\Sigma}}, \tag{56}$$

while for the real batch $\mathbb{E}\big[\|\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}\|_2^4\big] = C_{\boldsymbol{\Sigma}}$.

Proof. Given $\mathcal{F}_t$, the sample $\boldsymbol{x}_i^{(t)}$ is drawn from the mixture (2), so from the total expectation theorem

$$\mathbb{E}\big[\|\boldsymbol{x}_i^{(t)} - \boldsymbol{\mu}\|_2^4 \mid \mathcal{F}_t\big] = \alpha_t \, \mathbb{E}\big[\|\boldsymbol{Y}^{(t)} - \boldsymbol{\mu}\|_2^4 \mid \mathcal{F}_t\big] + (1 - \alpha_t) \, C_{\boldsymbol{\Sigma}}, \tag{57}$$

where $\boldsymbol{Y}^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$ given $\mathcal{F}_t$, and the real component contributes $\mathbb{E}\big[\|\boldsymbol{W} - \boldsymbol{\mu}\|_2^4\big] = C_{\boldsymbol{\Sigma}}$ for $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (by (53)). It remains to bound the synthetic term.

From (7), we have $\boldsymbol{\delta}^{(t)} = \boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}$ is $\mathcal{F}_t$-measurable. Thus, write $\boldsymbol{Y}^{(t)} - \boldsymbol{\mu} = \boldsymbol{y}^{(t)} + \boldsymbol{\delta}^{(t)}$ with $\boldsymbol{y}^{(t)} \coloneqq \boldsymbol{Y}^{(t)} - \boldsymbol{\mu}^{(t)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{(t)})$ given $\mathcal{F}_t$. Since $\boldsymbol{y}^{(t)}$ is conditionally symmetric about 0, the odd terms in $\boldsymbol{y}^{(t)}$ vanish, and

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{Y}^{(t)} - \boldsymbol{\mu}\|_2^4 \mid \mathcal{F}_t] &= \mathbb{E}[\|\boldsymbol{y}^{(t)} + \boldsymbol{\delta}^{(t)}\|_2^4 \mid \mathcal{F}_t] \\
&= \mathbb{E}[\|\boldsymbol{y}^{(t)}\|_2^4 \mid \mathcal{F}_t] + 4\,\mathbb{E}\left[\left((\boldsymbol{y}^{(t)})^\top \boldsymbol{\delta}^{(t)}\right)^2 \;\middle|\; \mathcal{F}_t\right] + 2\|\boldsymbol{\delta}^{(t)}\|_2^2 \, \mathbb{E}[\|\boldsymbol{y}^{(t)}\|_2^2 \mid \mathcal{F}_t] + \|\boldsymbol{\delta}^{(t)}\|_2^4 \\
&= \mathbb{E}[\|\boldsymbol{y}^{(t)}\|_2^4 \mid \mathcal{F}_t] + 4\,\mathbb{E}\left[\left((\boldsymbol{y}^{(t)})^\top \boldsymbol{\delta}^{(t)}\right)^2 \;\middle|\; \mathcal{F}_t\right] + 2p \cdot \mathbb{E}[\|\boldsymbol{y}^{(t)}\|_2^2 \mid \mathcal{F}_t] + p^2,
\end{aligned} \tag{58}$$

where we denote $u \coloneqq \operatorname{Tr}(\hat{\boldsymbol{S}}_t)$ and $p \coloneqq \|\boldsymbol{\delta}^{(t)}\|_2^2$. In order to further calculate and bound the above expression, we will use $\boldsymbol{y}^{(t)} \mid \mathcal{F}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{(t)})$, and use $\operatorname{Tr}(\boldsymbol{\Sigma}^{(t)}) = \operatorname{Tr}(\hat{\boldsymbol{S}}_t) - \|\boldsymbol{\delta}^{(t)}\|_2^2 = u - p$ (Lemma A.14(ii)).

First, conditionally $\boldsymbol{\Sigma}^{(t)}$ is fixed, so by (53) applied to $\boldsymbol{\Sigma}^{(t)}$,

$$\mathbb{E}[\|\boldsymbol{y}^{(t)}\|_2^4 \mid \mathcal{F}_t] = \operatorname{Tr}(\boldsymbol{\Sigma}^{(t)})^2 + 2\|\boldsymbol{\Sigma}^{(t)}\|_F^2 \leq 3 \operatorname{Tr}(\boldsymbol{\Sigma}^{(t)})^2 = 3(u - p)^2,$$

the inequality using $\|\boldsymbol{A}\|_F^2 \leq \operatorname{Tr}(\boldsymbol{A})^2$ for symmetric positive-semi-definite $\boldsymbol{A}$ (Lemma A.4(a), applied to $\boldsymbol{\Sigma}^{(t)} \succeq 0$ by Lemma A.8).

Second, $\mathbb{E}[\|\boldsymbol{y}^{(t)}\|_2^2 \mid \mathcal{F}_t] = \operatorname{Tr}(\boldsymbol{\Sigma}^{(t)}) = u - p$.

Finally,

$$\begin{aligned}
\mathbb{E}\left[\left((\boldsymbol{y}^{(t)})^\top \boldsymbol{\delta}^{(t)}\right)^2 \;\middle|\; \mathcal{F}_t\right] &= \mathbb{E}\left[\left(\boldsymbol{\delta}^{(t)}\right)^\top \boldsymbol{y}^{(t)} \left(\boldsymbol{y}^{(t)}\right)^\top \boldsymbol{\delta}^{(t)} \;\middle|\; \mathcal{F}_t\right] \\
&= (\boldsymbol{\delta}^{(t)})^\top \mathbb{E}\left[\boldsymbol{y}^{(t)} \left(\boldsymbol{y}^{(t)}\right)^\top \;\middle|\; \mathcal{F}_t\right] \boldsymbol{\delta}^{(t)} \\
&= \left(\boldsymbol{\delta}^{(t)}\right)^\top \boldsymbol{\Sigma}^{(t)} \boldsymbol{\delta}^{(t)} \\
&= \left(\boldsymbol{\delta}^{(t)}\right)^\top \hat{\boldsymbol{S}}_t \boldsymbol{\delta}^{(t)} - \|\boldsymbol{\delta}^{(t)}\|_2^4 \\
&\le \operatorname{Tr}(\hat{\boldsymbol{S}}_t)\|\boldsymbol{\delta}^{(t)}\|_2^2 - \|\boldsymbol{\delta}^{(t)}\|_2^4 \\
&= up - p^2 \\
&= p(u-p).
\end{aligned}$$

We used $\boldsymbol{\Sigma}^{(t)} = \hat{\boldsymbol{S}}_t - \boldsymbol{\delta}^{(t)} \left(\boldsymbol{\delta}^{(t)}\right)^\top$ (Lemma A.14(ii)), and Lemma A.15 with $\hat{\boldsymbol{S}}_t \succeq 0$ by Lemma A.14(i). Collecting terms in (58),

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{Y}^{(t)} - \boldsymbol{\mu}\|_2^4 \mid \mathcal{F}_t] &= \mathbb{E}[\|\boldsymbol{y}^{(t)}\|_2^4 \mid \mathcal{F}_t] + 4\,\mathbb{E}\left[\left((\boldsymbol{y}^{(t)})^\top \boldsymbol{\delta}^{(t)}\right)^2 \;\middle|\; \mathcal{F}_t\right] + 2p \cdot \mathbb{E}[\|\boldsymbol{y}^{(t)}\|_2^2 \mid \mathcal{F}_t] + p^2 \\
&\le 3(u-p)^2 + 4p(u-p) + 2p(u-p) + p^2 \\
&= 3(u-p)^2 + 6p(u-p) + p^2 \\
&= 3u^2 - 3p^2 + p^2 \\
&= 3u^2 - 2p^2 \\
&\le 3u^2 \\
&= 3\operatorname{Tr}(\hat{\boldsymbol{S}}_t)^2.
\end{aligned}$$

We used $-2p^2 \le 0$, and

$$3(u-p)^2 + 6p(u-p) = 3(u-p)\,(u-p+2p) = 3(u-p)(u+p) = 3u^2 - 3p^2.$$

Substituting into (57) and bounding $1 - \alpha_t \le 1$ gives (56). □

Lemma A.17 (Recursion unrolling)**.** *Let $\{a_t\}_{t\ge1}$, $\{r_t\}_{t\ge1}$, $\{q_t\}_{t\ge1}$ be nonnegative sequences with $a_t \le r_{t-1}a_{t-1} + q_{t-1}$ for all $t \ge 2$. Then for every $t \ge 2$,*

$$a_t \le \left(\prod_{k=1}^{t-1} r_k\right) a_1 + \sum_{s=1}^{t-1} \left(\prod_{k=s+1}^{t-1} r_k\right) q_s,$$

with the empty product equal to 1.

Proof. We prove by induction on t. The base case $t = 2$ is the hypothesis $a_2 \le r_1 a_1 + q_1$. Assuming the bound at $t-1$, we get

$$\begin{aligned}
a_t \le r_{t-1}a_{t-1} + q_{t-1} &\le r_{t-1}\left[\left(\prod_{k=1}^{t-2} r_k\right) a_1 + \sum_{s=1}^{t-2} \left(\prod_{k=s+1}^{t-2} r_k\right) q_s\right] + q_{t-1} \\
&= \left(\prod_{k=1}^{t-1} r_k\right) a_1 + \sum_{s=1}^{t-2} \left(\prod_{k=s+1}^{t-1} r_k\right) q_s + q_{t-1} \\
&= \left(\prod_{k=1}^{t-1} r_k\right) a_1 + \sum_{s=1}^{t-1} \left(\prod_{k=s+1}^{t-1} r_k\right) q_s,
\end{aligned}$$

where q_{t-1} is the $s = t-1$ term (empty product $= 1$). □

Lemma A.18 (Uniform bound on the amplification products)**.** *Let $r_t \coloneqq \lambda_t^2 + 3\alpha_t\beta_t$ for $t \geq 1$, with λ_t, β_t as in* (18)*. Then $r_t \geq 0$ for every $t \geq 1$, and for all integers $1 \leq a \leq b < \infty$,*

$$\prod_{t=a}^{b} r_t \leq \exp\left(\frac{3}{n_r}\right) < \infty. \tag{59}$$

In particular the bound is uniform in a and b, and holds for every $\{\alpha_t\} \subset [0,1]$ *and* $\{n_b(t) \geq 1\}$*, with no further hypothesis.*

Proof. Fix $t \geq 1$. By Lemma A.9 we have $\lambda_t \in (0,1]$, hence $\lambda_t^2 \leq 1$; and $\alpha_t \in [0,1]$ gives $3\alpha_t\beta_t \leq 3\beta_t$. Adding,

$$0 \leq r_t \leq 1 + 3\beta_t, \tag{60}$$

the lower bound because both summands defining r_t are nonnegative. Fix integers $1 \leq a \leq b$. From (60), we have

$$\begin{aligned}
\prod_{t=a}^{b} r_t \leq \prod_{t=a}^{b} (1+3\beta_t) &\leq \prod_{t=1}^{\infty} (1+3\beta_t) \\
&= \exp\left(\sum_{t=1}^{\infty} \ln(1+3\beta_t)\right) \\
&\leq \exp\left(3\sum_{t=1}^{\infty} \beta_t\right) \\
&\leq \exp\left(\frac{3}{n_r}\right)
\end{aligned}$$

Where we used $1 + 3\beta_t \geq 1$, $\ln(1+x) \leq x$ for $x \geq 0$, and $\sum_{t=1}^{\infty} \beta_t \leq 1/n_r$ from Lemma A.5. □

Lemma A.19 (Uniform second-moment bound on $\hat{\boldsymbol{S}}_t$)**.** *Under Assumption 1,*

$$B^* \coloneqq \sup_{t \geq 1} \mathbb{E}\big[\mathrm{Tr}(\hat{\boldsymbol{S}}_t)^2\big] < \infty.$$

Proof. Write $u_t \coloneqq \mathrm{Tr}(\hat{\boldsymbol{S}}_t)$, and denote

$$v_t \coloneqq \mathrm{Var}(u_t) = \mathbb{E}\big[\mathrm{Tr}(\hat{\boldsymbol{S}}_t)^2\big] - \mathrm{Tr}(\boldsymbol{\Sigma})^2. \tag{61}$$

We used $\mathbb{E}[u_t] = \mathbb{E}\big[\mathrm{Tr}(\hat{\boldsymbol{S}}_t)\big] = \mathrm{Tr}(\mathbb{E}[\hat{\boldsymbol{S}}_t]) = \mathrm{Tr}(\boldsymbol{\Sigma})$ (Lemma A.14(iv)). We establish a recursion on u_t. From Lemma A.14(iii), we have for every $t \geq 2$

$$\begin{aligned}
u_t &= \rho_{t-1}u_{t-1} + \frac{1}{N_t} \sum_{i=1}^{n_b(t-1)} \mathrm{Tr}\big((\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})(\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu})^\top\big) \\
&= \rho_{t-1}u_{t-1} + \frac{1}{N_t} \sum_{i=1}^{n_b(t-1)} \|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu}\|_2^2
\end{aligned} \tag{62}$$

and $u_1 = \frac{1}{n_r}\sum_{j=1}^{n_r} \|\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}\|_2^2$. From (61), we need to prove that $\sup_{t\geq 1} v_t < \infty$. We will use the law of total variance, which presupposes $u_t \in L^2$. We verify that by induction on $t \geq 1$.

Square-integrability at every finite t. For $t = 1$, the samples $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_r}$ are i.i.d. from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and thus

$$v_1 = \frac{1}{n_r}\mathrm{Var}(\|\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu}\|_2^2) \leq \frac{1}{n_r}\mathbb{E}\big[\|\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu}\|_2^4\big] = \frac{1}{n_r}C_{\boldsymbol{\Sigma}} < \infty, \tag{63}$$

where we used $\mathrm{Var}(X) \le \mathbb{E}[X^2]$. From (61), we have $\mathbb{E}[u_1^2] = v_1 + \mathrm{Tr}(\mathbf{\Sigma})^2 < \infty$, thus $u_1 \in L^2$. Assume $u_{t-1} \in L^2$, then

$$\begin{aligned}\mathbb{E}[u_t^2] &\le 2\left(\rho_{t-1}^2 \mathbb{E}[u_{t-1}^2] + \frac{n_b(t-1)}{N_t^2} \sum_{i=1}^{n_b(t-1)} \mathbb{E}\big[\|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu}\|_2^4\big]\right) \\ &\le 2\left(\rho_{t-1}^2 \mathbb{E}\big[u_{t-1}^2\big] + \frac{n_b(t-1)}{N_t^2} \sum_{i=1}^{n_b(t-1)} (3\alpha_{t-1}\mathbb{E}\big[u_{t-1}^2\big] + C_{\mathbf{\Sigma}})\right) \\ &< \infty,\end{aligned}$$

where the first inequality applies Lemma A.3 twice: with $n = 2$ to the two summands of (62), i.e., $(a+b)^2 \le 2(a^2+b^2)$, and with $n = n_b(t-1)$ to the sum $\sum_{i=1}^{n_b(t-1)} \|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu}\|_2^2$, which gives

$$\left(\sum_{i=1}^{n_b(t-1)} \|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu}\|_2^2\right)^2 \le n_b(t-1) \sum_{i=1}^{n_b(t-1)} \|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu}\|_2^4.$$

The second inequality uses the following bound: from Lemma A.16 and the tower property, we have

$$\mathbb{E}\big[\|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu}\|_2^4\big] = \mathbb{E}\big[\mathbb{E}\big[\|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu}\|_2^4 \mid \mathcal{F}_{t-1}\big]\big] \le \mathbb{E}\big[3\alpha_{t-1}\,\mathrm{Tr}(\hat{\boldsymbol{S}}_{t-1})^2 + C_{\mathbf{\Sigma}}\big] = 3\alpha_{t-1}\mathbb{E}\big[u_{t-1}^2\big] + C_{\mathbf{\Sigma}}.$$

The law of total variance. Fix $t \ge 2$. The law of total variance dictates that

$$v_t = \mathrm{Var}\big(\mathbb{E}\big[u_t \mid \mathcal{F}_{t-1}\big]\big) + \mathbb{E}\big[\mathrm{Var}(u_t \mid \mathcal{F}_{t-1})\big]. \tag{64}$$

Taking the trace of Lemma A.14(iv), $\mathbb{E}[u_t \mid \mathcal{F}_{t-1}] = \lambda_{t-1}u_{t-1} + (1-\lambda_{t-1})\,\mathrm{Tr}(\mathbf{\Sigma})$. Thus, the first summand reads

$$\mathrm{Var}\big(\mathbb{E}[u_t \mid \mathcal{F}_{t-1}]\big) = \lambda_{t-1}^2\, v_{t-1}. \tag{65}$$

For the second summand, we use (62):

$$\begin{aligned}\mathrm{Var}(u_t \mid \mathcal{F}_{t-1}) &= \frac{n_b(t-1)}{N_t^2}\,\mathrm{Var}\big(\|\boldsymbol{x}_1^{(t-1)} - \boldsymbol{\mu}\|_2^2 \mid \mathcal{F}_{t-1}\big) \le \beta_{t-1}\,\mathbb{E}\big[\|\boldsymbol{x}_1^{(t-1)} - \boldsymbol{\mu}\|_2^4 \mid \mathcal{F}_{t-1}\big] \\ &\le \beta_{t-1}\big(3\alpha_{t-1}u_{t-1}^2 + C_{\mathbf{\Sigma}}\big).\end{aligned} \tag{66}$$

We used that u_{t-1} is $\mathcal{F}_{t-1}$-measurable, and conditionally on $\mathcal{F}_{t-1}$ the samples $\{\boldsymbol{x}_i^{(t-1)}\}_{i=1}^{n_b(t-1)}$ are i.i.d. (Assumption 1), Var $\le$ second moment together with β_t definition from (18), and Lemma A.16. From (64), (65) and (66), we have

$$\begin{aligned}v_t &\le \lambda_{t-1}^2 v_{t-1} + \mathbb{E}\big[3\beta_{t-1}\alpha_{t-1}u_{t-1}^2 + \beta_{t-1}C_{\mathbf{\Sigma}}\big] \\ &= \big(\lambda_{t-1}^2 + 3\alpha_{t-1}\beta_{t-1}\big)\, v_{t-1} + \beta_{t-1}(3\alpha_{t-1}\,\mathrm{Tr}(\mathbf{\Sigma})^2 + C_{\mathbf{\Sigma}}),\end{aligned} \tag{67}$$

where we used $\mathbb{E}[u_{t-1}^2] = v_{t-1} + \mathrm{Tr}(\mathbf{\Sigma})^2$ from (61). From (67), we have:

$$v_t \le r_{t-1}\, v_{t-1} + q_{t-1}, \qquad \begin{cases} r_{t-1} \coloneqq \lambda_{t-1}^2 + 3\alpha_{t-1}\beta_{t-1}, \\ q_{t-1} \coloneqq \beta_{t-1}\big(3\alpha_{t-1}\,\mathrm{Tr}(\mathbf{\Sigma})^2 + C_{\mathbf{\Sigma}}\big). \end{cases} \tag{68}$$

Uniform bound. Applying Lemma A.17 to (68) and using Lemma A.18, we have

$$\begin{aligned}v_t &\le \left(\prod_{k=1}^{t-1} r_k\right) v_1 + \sum_{s=1}^{t-1}\left(\prod_{k=s+1}^{t-1} r_k\right) q_s \\ &\le \exp\left(\frac{3}{n_r}\right) v_1 + \exp\left(\frac{3}{n_r}\right) \sum_{s=1}^{t-1} q_s \\ &\le \exp\left(\frac{3}{n_r}\right)(v_1 + D_0)\end{aligned} \tag{69}$$

uniformly in t. Using (67), we defined

$$\begin{aligned} D_0 \coloneqq \sum_{s=1}^{\infty} q_s &\leq \left(3\,\mathrm{Tr}(\mathbf{\Sigma})^2 + C_{\mathbf{\Sigma}}\right) \sum_{t=1}^{\infty} \beta_t \\ &\leq \frac{1}{n_r}\left(C_{\mathbf{\Sigma}} + 3\,\mathrm{Tr}(\mathbf{\Sigma})^2\right). \end{aligned} \tag{70}$$

The first inequality follows from $\alpha_{t-1} \leq 1$, and the second inequality follows from Lemma A.5. Finally, from (61), (69) and (70), we have

$$\sup_{t\geq 1} \mathbb{E}\big[\mathrm{Tr}(\hat{\boldsymbol{S}}_t)^2\big] = \sup_{t\geq 1}\big(v_t + \mathrm{Tr}(\mathbf{\Sigma})^2\big) \leq \exp\left(\frac{3}{n_r}\right)(v_1 + D_0) + \mathrm{Tr}(\mathbf{\Sigma})^2 < \infty. \tag{71}$$

□

Lemma A.20 (Divergence of $1-\lambda_t^2$)**.** *If $\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty$, then $\sum_{t=1}^{\infty}(1-\lambda_t^2) = \infty$.*

Proof. From the definition of λ_t (18) and the definition of ρ_t in (3), we have

$$1 - \lambda_t = 1 - \frac{N_t + n_b(t)\alpha_t}{N_{t+1}} = \frac{N_{t+1} - N_t - n_b(t)\alpha_t}{N_{t+1}} = \frac{n_b(t)(1-\alpha_t)}{N_{t+1}} = (1-\rho_t)(1-\alpha_t),$$

where we used $N_{t+1} = N_t + n_b(t)$ and $\rho_t = N_t/N_{t+1} \Rightarrow 1 - \rho_t = n_b(t)/N_{t+1}$. As $\lambda_t \in (0,1]$ (Lemma A.9), $1-\lambda_t^2 = (1-\lambda_t)(1+\lambda_t) \geq 1-\lambda_t$, thus

$$\sum_{t=1}^{\infty}(1-\lambda_t^2) \geq \sum_{t=1}^{\infty}(1-\lambda_t) = \sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) \;=\; \infty$$

by hypothesis. □

Lemma A.21 (A useful upper bound)**.** *Let $\boldsymbol{X}$ be a random matrix with $\mathbb{E}[\|\boldsymbol{X}\|_F^2] < \infty$, and $\mathcal{G}$ be a sub-σ-algebra. Define $\boldsymbol{Z} \coloneqq \boldsymbol{X} - \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]$. Then,*

$$\mathbb{E}[\|\boldsymbol{Z}\|_F^2 \mid \mathcal{G}] \leq \mathbb{E}[\|\boldsymbol{X}\|_F^2 \mid \mathcal{G}].$$

Proof. Indeed, we have

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{Z}\|_F^2 \mid \mathcal{G}] &= \mathbb{E}[\mathrm{Tr}(\boldsymbol{Z}^\top \boldsymbol{Z}) \mid \mathcal{G}] \\ &= \mathrm{Tr}\Big(\mathbb{E}[\boldsymbol{Z}^\top \boldsymbol{Z} \mid \mathcal{G}]\Big) \\ &= \mathrm{Tr}\Big(\mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}])^\top (\boldsymbol{X} - \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]) \mid \mathcal{G}]\Big) \\ &= \mathrm{Tr}\Big(\mathbb{E}\Big[\boldsymbol{X}^\top \boldsymbol{X} \;\Big|\; \mathcal{G}\Big] - \mathbb{E}\Big[\boldsymbol{X}^\top \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}] \;\Big|\; \mathcal{G}\Big] - \mathbb{E}\big[\mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]^\top \boldsymbol{X} \mid \mathcal{G}\big] + \mathbb{E}\big[\mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]^\top \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}] \mid \mathcal{G}\big]\Big) \\ &= \mathbb{E}[\mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{X}) \mid \mathcal{G}] - \mathrm{Tr}\big(\mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]^\top \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}] + \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]^\top \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}] - \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]^\top \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]\big) \\ &= \mathbb{E}[\mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{X}) \mid \mathcal{G}] - \mathrm{Tr}\big(\mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]^\top \mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]\big) \\ &= \mathbb{E}[\|\boldsymbol{X}\|_F^2 \mid \mathcal{G}] - \|\mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]\|_F^2 \\ &\leq \mathbb{E}[\|\boldsymbol{X}\|_F^2 \mid \mathcal{G}]. \end{aligned}$$

We used $\|\boldsymbol{A}\|_F^2 = \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{A})$, $\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}) \mid \mathcal{G}] = \mathrm{Tr}(\mathbb{E}[\boldsymbol{A} \mid \mathcal{G}])$ and that $\mathbb{E}[\boldsymbol{X} \mid \mathcal{G}]$ is $\mathcal{G}$-measurable. □

Finally, we characterize when $\mathbf{\Sigma}^{(t)}$ is consistent (in expected Frobenius norm).

Theorem A.22 (Frobenius norm consistency)**.** *Under Assumption 1,*

$$\mathbb{E}\Big[\|\mathbf{\Sigma}^{(t)} - \mathbf{\Sigma}\|_F\Big] \xrightarrow{t\to\infty} 0 \quad\Longleftrightarrow\quad \sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty.$$

Proof. Let $P_t \coloneqq \mathbb{E}\big[\|\hat{\boldsymbol{S}}_t - \boldsymbol{\Sigma}\|_F^2\big]$.

P_t is finite. Indeed,

$$P_t \le 2\Big(\mathbb{E}\Big[\|\hat{\boldsymbol{S}}_t\|_F^2\Big] + \|\boldsymbol{\Sigma}\|_F^2\Big) \le 2\Big(\mathbb{E}\Big[\mathrm{Tr}(\hat{\boldsymbol{S}}_t)^2\Big] + \|\boldsymbol{\Sigma}\|_F^2\Big) < \infty.$$

We used $\|\boldsymbol{A}-\boldsymbol{B}\|_F^2 \le 2\big(\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{B}\|_F^2\big)$ and the monotonicity of the expectation. Then, since $\hat{\boldsymbol{S}}_t$ is symmetric positive-semi-definite (Lemma A.14(i)), we have $\|\hat{\boldsymbol{S}}_t\|_F^2 \le \mathrm{Tr}(\hat{\boldsymbol{S}}_t)^2$ (Lemma A.4(a)). The final bound inequality follows from Lemma A.19, together with $\|\boldsymbol{\Sigma}\|_F < \infty$.

Steps 1–3 establish sufficiency, i.e. that $\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty$ implies $\mathbb{E}[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F] \to 0$. Step 4 establishes necessity.

Step 1 (A recursion for P_t). For $t \ge 2$, define

$$\tilde{\boldsymbol{V}}_t \coloneqq \hat{\boldsymbol{S}}_t - \boldsymbol{\Sigma}, \quad \tilde{\boldsymbol{\xi}}_t \coloneqq \hat{\boldsymbol{S}}_t - \mathbb{E}\big[\hat{\boldsymbol{S}}_t \mid \mathcal{F}_{t-1}\big]. \tag{72}$$

A recursion for $\tilde{\boldsymbol{V}}_t$. From Lemma A.14(iv), we have

$$\tilde{\boldsymbol{V}}_t - \tilde{\boldsymbol{\xi}}_t = \mathbb{E}\big[\hat{\boldsymbol{S}}_t \mid \mathcal{F}_{t-1}\big] - \boldsymbol{\Sigma} = \lambda_{t-1}\Big(\hat{\boldsymbol{S}}_{t-1} - \boldsymbol{\Sigma}\Big) = \lambda_{t-1}\tilde{\boldsymbol{V}}_{t-1}.$$

Thus,

$$\tilde{\boldsymbol{V}}_t = \lambda_{t-1}\tilde{\boldsymbol{V}}_{t-1} + \tilde{\boldsymbol{\xi}}_t. \tag{73}$$

Conditional expectation of $\tilde{\boldsymbol{\xi}}_t$. From (72), the following identity holds:

$$\mathbb{E}[\tilde{\boldsymbol{\xi}}_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[\hat{\boldsymbol{S}}_t \mid \mathcal{F}_{t-1}] - \mathbb{E}[\hat{\boldsymbol{S}}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{0}. \tag{74}$$

Final computation. From (72), we have $P_t = \mathbb{E}\big[\|\tilde{\boldsymbol{V}}_t\|_F^2\big]$. Thus,

$$\begin{aligned}
P_t &= \mathbb{E}\big[\langle \tilde{\boldsymbol{V}}_t, \tilde{\boldsymbol{V}}_t\rangle_F\big] \\
&= \mathbb{E}\big[\langle \lambda_{t-1}\tilde{\boldsymbol{V}}_{t-1} + \tilde{\boldsymbol{\xi}}_t, \lambda_{t-1}\tilde{\boldsymbol{V}}_{t-1} + \tilde{\boldsymbol{\xi}}_t\rangle_F\big] \\
&= \lambda_{t-1}^2\mathbb{E}\big[\langle \tilde{\boldsymbol{V}}_{t-1}, \tilde{\boldsymbol{V}}_{t-1}\rangle_F\big] + 2\lambda_{t-1}\mathbb{E}\big[\langle \tilde{\boldsymbol{V}}_{t-1}, \tilde{\boldsymbol{\xi}}_t\rangle_F\big] + \mathbb{E}\big[\langle \tilde{\boldsymbol{\xi}}_t, \tilde{\boldsymbol{\xi}}_t\rangle_F\big] \\
&= \lambda_{t-1}^2\mathbb{E}\big[\|\tilde{\boldsymbol{V}}_{t-1}\|_F^2\big] + \mathbb{E}\big[\|\tilde{\boldsymbol{\xi}}_t\|_F^2\big] \\
&= \lambda_{t-1}^2 P_{t-1} + \mathbb{E}\big[\|\tilde{\boldsymbol{\xi}}_t\|_F^2\big].
\end{aligned} \tag{75}$$

We defined the inner product $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_F \coloneqq \mathrm{Tr}(\boldsymbol{A}^\top\boldsymbol{B}) = \sum_{i=1}^d\sum_{j=1}^d A_{ij}B_{ij}$, and used $\mathbb{E}\big[\langle \tilde{\boldsymbol{V}}_{t-1}, \tilde{\boldsymbol{\xi}}_t\rangle_F\big] = 0$. Indeed, $\tilde{\boldsymbol{V}}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable, and from (74) we have:

$$\begin{aligned}
\mathbb{E}\big[\langle \tilde{\boldsymbol{V}}_{t-1}, \tilde{\boldsymbol{\xi}}_t\rangle_F\big] &= \mathbb{E}\big[\mathbb{E}\big[\mathrm{Tr}\big(\tilde{\boldsymbol{V}}_{t-1}^\top\tilde{\boldsymbol{\xi}}_t\big) \mid \mathcal{F}_{t-1}\big]\big] \\
&= \mathbb{E}\big[\mathrm{Tr}\big(\mathbb{E}\big[\tilde{\boldsymbol{V}}_{t-1}^\top\tilde{\boldsymbol{\xi}}_t \mid \mathcal{F}_{t-1}\big]\big)\big] \\
&= \mathbb{E}\big[\mathrm{Tr}\big(\tilde{\boldsymbol{V}}_{t-1}^\top\mathbb{E}\big[\tilde{\boldsymbol{\xi}}_t \mid \mathcal{F}_{t-1}\big]\big)\big] \\
&= \mathbb{E}\big[\mathrm{Tr}\big(\tilde{\boldsymbol{V}}_{t-1}^\top\boldsymbol{0}\big)\big] \\
&= 0.
\end{aligned}$$

Finally, from (75), we have

$$P_t = \lambda_{t-1}^2\, P_{t-1} + \mathbb{E}\big[\|\tilde{\boldsymbol{\xi}}_t\|_F^2\big]. \tag{76}$$

Bounding $\mathbb{E}\big[\|\tilde{\boldsymbol{\xi}}_t\|_F^2\big]$. For every $s \ge 1$ and $1 \le i \le n_b(s)$, define

$$\boldsymbol{z}_i^{(s)} \coloneqq \boldsymbol{x}_i^{(s)} - \boldsymbol{\mu}. \tag{77}$$

Fix $t \geq 2$. For every $1 \leq s \leq t-2$, the random vector $\boldsymbol{z}_i^{(s)}$ is $\mathcal{F}_{t-1}$-measurable, and so is $\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu}$ for $1 \leq j \leq n_r$. Hence their outer products equal their conditional expectations given $\mathcal{F}_{t-1}$, e.g.,

$$\mathbb{E}\Big[\boldsymbol{z}_i^{(s)} \left(\boldsymbol{z}_i^{(s)}\right)^\top \Big| \mathcal{F}_{t-1}\Big] = \boldsymbol{z}_i^{(s)} \left(\boldsymbol{z}_i^{(s)}\right)^\top.$$

From (54) and (72), these terms cancel, and we get

$$\tilde{\boldsymbol{\xi}}_t = \frac{1}{N_t} \sum_{i=1}^{n_b(t-1)} \Big(\boldsymbol{z}_i^{(t-1)}\big(\boldsymbol{z}_i^{(t-1)}\big)^\top - \mathbb{E}\Big[\boldsymbol{z}_i^{(t-1)}\big(\boldsymbol{z}_i^{(t-1)}\big)^\top \Big| \mathcal{F}_{t-1}\Big]\Big). \tag{78}$$

Let us define

$$\boldsymbol{Z}_i^{(t-1)} \coloneqq \boldsymbol{z}_i^{(t-1)}\big(\boldsymbol{z}_i^{(t-1)}\big)^\top - \mathbb{E}\Big[\boldsymbol{z}_i^{(t-1)}\big(\boldsymbol{z}_i^{(t-1)}\big)^\top \Big| \mathcal{F}_{t-1}\Big]. \tag{79}$$

Finally, from (78) and (79), we have

$$\tilde{\boldsymbol{\xi}}_t = \frac{1}{N_t} \sum_{i=1}^{n_b(t-1)} \boldsymbol{Z}_i^{(t-1)}. \tag{80}$$

Conditionally on $\mathcal{F}_{t-1}$, the samples $\boldsymbol{x}_i^{(t-1)}$ are i.i.d. Thus, from (77) and (79), the random matrices $\boldsymbol{Z}_i^{(t-1)}$ are also i.i.d. as a deterministic transformation of $\boldsymbol{x}_i^{(t-1)}$. That is,

$$\begin{aligned}
\mathbb{E}[\|\tilde{\boldsymbol{\xi}}_t\|_F^2 \mid \mathcal{F}_{t-1}] &= \frac{n_b(t-1)}{N_t^2}\, \mathbb{E}[\|\boldsymbol{Z}_i^{(t-1)}\|_F^2 \mid \mathcal{F}_{t-1}] \\
&\leq \beta_{t-1}\, \mathbb{E}[\|\boldsymbol{z}_i^{(t-1)}\|_2^4 \mid \mathcal{F}_{t-1}] \\
&= \beta_{t-1}\, \mathbb{E}[\|\boldsymbol{x}_i^{(t-1)} - \boldsymbol{\mu}\|_2^4 \mid \mathcal{F}_{t-1}] \\
&\leq \beta_{t-1}\big(3\alpha_{t-1}\, \mathrm{Tr}(\hat{\boldsymbol{S}}_{t-1})^2 + C_{\boldsymbol{\Sigma}}\big).
\end{aligned}$$

We used the definition of β_t from (18) and Lemma A.21 with $\mathcal{G} = \mathcal{F}_{t-1}$ and $\boldsymbol{X} = \boldsymbol{z}_i^{(t-1)}\big(\boldsymbol{z}_i^{(t-1)}\big)^\top$, for which

$$\|\boldsymbol{X}\|_F^2 = \Big\|\boldsymbol{z}_i^{(t-1)}\big(\boldsymbol{z}_i^{(t-1)}\big)^\top\Big\|_F^2 = \Big\|\boldsymbol{z}_i^{(t-1)}\Big\|_2^4,$$

so that its hypothesis $\mathbb{E}[\|\boldsymbol{X}\|_F^2] < \infty$ holds: taking expectations in Lemma A.16 (using the tower property) and using Lemma A.19, $\mathbb{E}\big[\|\boldsymbol{z}_i^{(t-1)}\|_2^4\big] \leq 3\alpha_{t-1}\, \mathbb{E}\big[\mathrm{Tr}(\hat{\boldsymbol{S}}_{t-1})^2\big] + C_{\boldsymbol{\Sigma}} < \infty$. We also used (77) and Lemma A.16 for the conditional fourth moment. This gives

$$\mathbb{E}[\|\tilde{\boldsymbol{\xi}}_t\|_F^2 \mid \mathcal{F}_{t-1}] \leq \beta_{t-1}\big(3\alpha_{t-1}\, \mathrm{Tr}(\hat{\boldsymbol{S}}_{t-1})^2 + C_{\boldsymbol{\Sigma}}\big). \tag{81}$$

Using the tower property in (81), we have

$$\begin{aligned}
\mathbb{E}[\|\tilde{\boldsymbol{\xi}}_t\|_F^2] &= \mathbb{E}\Big[\mathbb{E}\Big[\|\tilde{\boldsymbol{\xi}}_t\|_F^2 \mid \mathcal{F}_{t-1}\Big]\Big] \\
&\leq \mathbb{E}\Big[\beta_{t-1}\Big(3\alpha_{t-1}\, \mathrm{Tr}(\hat{\boldsymbol{S}}_{t-1})^2 + C_{\boldsymbol{\Sigma}}\Big)\Big] \\
&= \beta_{t-1}\Big(3\alpha_{t-1}\mathbb{E}\Big[\mathrm{Tr}(\hat{\boldsymbol{S}}_{t-1})^2\Big] + C_{\boldsymbol{\Sigma}}\Big) \\
&\leq (3B^* + C_{\boldsymbol{\Sigma}})\, \beta_{t-1}.
\end{aligned} \tag{82}$$

We used $\alpha_{t-1} \leq 1$ and the uniform bound $\mathbb{E}\big[\mathrm{Tr}(\hat{\boldsymbol{S}}_{t-1})^2\big] \leq B^*$ from Lemma A.19. Finally, from (82) we have:

$$\mathbb{E}[\|\tilde{\boldsymbol{\xi}}_t\|_F^2] \leq D\, \beta_{t-1}, \qquad D \coloneqq 3B^* + C_{\boldsymbol{\Sigma}} < \infty. \tag{83}$$

Step 2 ($P_t \to 0$). From (76) and (83), we have $P_t \leq \lambda_{t-1}^2 P_{t-1} + D\beta_{t-1}$. Unrolling by Lemma A.17 with $r_{t-1} = \lambda_{t-1}^2 \leq 1$ (Lemma A.9) and $q_{t-1} = D\beta_{t-1}$,

$$P_t \leq \underbrace{\left(\prod_{k=1}^{t-1} \lambda_k^2\right)}_{\text{(I)}} P_1 + \underbrace{D \sum_{s=1}^{t-1} \left(\prod_{k=s+1}^{t-1} \lambda_k^2\right) \beta_s}_{\text{(II)}}. \tag{84}$$

Term (I). By the hypothesis $\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t)=\infty$ and Lemma A.20,

$$\prod_{k=1}^{t-1}\lambda_k^2=\exp\left(\sum_{k=1}^{t-1}\ln\lambda_k^2\right)\le\exp\left(-\sum_{k=1}^{t-1}(1-\lambda_k^2)\right)\xrightarrow[t\to\infty]{}0. \tag{85}$$

We used $\ln(x)\le x-1$ with $x=\lambda_k^2\ge 0$ (indeed, $0\le\lambda_k\le 1$ from Lemma A.9)

Term (II). Fix $\varepsilon>0$. By Lemma A.5, choose $s_0\ge 2$ with

$$\sum_{s=s_0}^{\infty}\beta_s<\frac{\varepsilon}{2}, \tag{86}$$

and from (85), choose $s_1\ge 1$ such that for every $t\ge s_1$,

$$\prod_{k=s_0}^{t-1}\lambda_k^2\le\frac{1}{\sum_{s=1}^{s_0-1}\beta_s}\frac{\varepsilon}{2}, \tag{87}$$

where we used $\prod_{k=s_0}^{t-1}\lambda_k^2=\frac{\prod_{k=1}^{t-1}\lambda_k^2}{\prod_{k=1}^{s_0-1}\lambda_k^2}\xrightarrow[t\to\infty]{}0$.

From (84), (86) and (87), for all $t\ge\max(s_0+2,s_1)$:

$$\begin{aligned}
\mathrm{II}&=\sum_{s=1}^{s_0-1}\left(\prod_{k=s+1}^{t-1}\lambda_k^2\right)\beta_s+\sum_{s=s_0}^{t-1}\left(\prod_{k=s+1}^{t-1}\lambda_k^2\right)\beta_s\\
&\le\prod_{k=s_0}^{t-1}\lambda_k^2\cdot\sum_{s=1}^{s_0-1}\beta_s+\sum_{s=s_0}^{t-1}\beta_s\\
&\le\prod_{k=s_0}^{t-1}\lambda_k^2\cdot\sum_{s=1}^{s_0-1}\beta_s+\sum_{s=s_0}^{\infty}\beta_s\\
&\le\frac{\varepsilon}{2}+\frac{\varepsilon}{2}\\
&=\varepsilon,
\end{aligned} \tag{88}$$

where we used $\lambda_k^2\le 1$ for all $k\ge 1$ (Lemma A.9). Finally, (88) holds for every $\varepsilon>0$, and thus $(\mathrm{II})\to 0$. Together with $(\mathrm{I})\to 0$ from (85), we have $P_t\to 0$ (indeed, from (84), $P_t\le(\mathrm{I})\cdot P_1+D\cdot(\mathrm{II})\to 0$). That is,

$$P_t\xrightarrow{t\to\infty}0. \tag{89}$$

Step 3 (from $\hat{\boldsymbol{S}}_t$ to $\boldsymbol{\Sigma}^{(t)}$). The triangle inequality gives

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)}-\boldsymbol{\Sigma}\|_F\big]&\le\mathbb{E}\big[\|\hat{\boldsymbol{S}}_t-\boldsymbol{\Sigma}\|_F\big]+\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)}-\hat{\boldsymbol{S}}_t\|_F\big]\\
&=\mathbb{E}\big[\|\hat{\boldsymbol{S}}_t-\boldsymbol{\Sigma}\|_F\big]+\mathbb{E}\big[\|\boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^\top\|_F\big]\\
&\le\sqrt{\mathbb{E}\big[\|\hat{\boldsymbol{S}}_t-\boldsymbol{\Sigma}\|_F^2\big]}+\mathbb{E}\big[\|\boldsymbol{\delta}^{(t)}\|_2^2\big]\\
&=\sqrt{P_t}+\mathrm{Tr}(\boldsymbol{C}_t).
\end{aligned} \tag{90}$$

We used Lemma A.14(ii) and the identity $\|\boldsymbol{x}\boldsymbol{x}^\top\|_F=\|\boldsymbol{x}\|_2^2$. Additionally, $(\cdot)^2$ is convex, and Jensen's inequality with $X=\|\hat{\boldsymbol{S}}_t-\boldsymbol{\Sigma}\|_F$ gives

$$\mathbb{E}[X]^2\le\mathbb{E}[X^2]\quad\Rightarrow\quad\mathbb{E}[X]\le\sqrt{\mathbb{E}[X^2]},$$

where $X\ge 0$ a.s. Then, we used $P_t=\mathbb{E}\big[\|\hat{\boldsymbol{S}}_t-\boldsymbol{\Sigma}\|_F^2\big]$ and $\mathbb{E}\big[\|\boldsymbol{\delta}^{(t)}\|_2^2\big]=\mathrm{Tr}(\boldsymbol{C}_t)$, which follows from (11). That is, from (90), we have shown that

$$\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)}-\boldsymbol{\Sigma}\|_F\big]\le\sqrt{P_t}+\mathrm{Tr}(\boldsymbol{C}_t). \tag{91}$$

By (89), we have $P_t \to 0$. Additionally, from our hypothesis $\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty$ and Theorem A.13, we have $\boldsymbol{C}_t \to \mathbf{0}$. Hence, $\mathrm{Tr}(\boldsymbol{C}_t) \to 0$ (Lemma A.10). Thus, both terms of (91) vanish, so $\mathbb{E}[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F] \to 0$.

Step 4 (necessity). Conversely, suppose

$$\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) < \infty. \tag{92}$$

Since $\|\cdot\|_F$ is convex, Jensen's inequality and Theorem 3.1(b) give

$$\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big] \geq \big\|\mathbb{E}[\boldsymbol{\Sigma}^{(t)}] - \boldsymbol{\Sigma}\big\|_F = \|-\boldsymbol{C}_t\|_F = \|\boldsymbol{C}_t\|_F, \tag{93}$$

where we used Theorem 3.1(b). By (92) and Theorem A.13, $\boldsymbol{C}_t \not\to \mathbf{0}$. Hence, $\|\boldsymbol{C}_t\|_F \not\to 0$ (Remark after Lemma A.10). From (93), we have $\mathbb{E}[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F] \not\to 0$. This establishes the stated equivalence. □

A.4.4 Convergence of the fitted distribution to the true distribution

We now characterize when the fitted distribution $\mathcal{D}_t$ converges to the true distribution $\mathcal{D}_{\text{real}}$ (see (5)) in the 2-Wasserstein distance W_2, a standard metric between probability distributions (Villani, 2009).

Definition 2 (2-Wasserstein distance)**.** *For probability distributions P, Q on $\mathbb{R}^d$ with finite second moments,*

$$W_2^2(P,Q) \coloneqq \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\pi}\big[\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\big],$$

where $\Pi(P,Q)$ is the set of couplings of P and Q, i.e., joint distributions with marginals P and Q.

First, we prove a bound on a metric between two symmetric positive-semi-definite matrices. It is used in the proof of Theorem A.24.

Lemma A.23 (Bures–Frobenius bound)**.** *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$ be symmetric with $\boldsymbol{A} \succeq 0$ and $\boldsymbol{B} \succeq 0$. Define the Bures metric $\mathcal{B}$ as*

$$\mathcal{B}^2(\boldsymbol{A},\boldsymbol{B}) = \mathrm{Tr}(\boldsymbol{A}) + \mathrm{Tr}(\boldsymbol{B}) - 2\,\mathrm{Tr}\big((\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2})^{1/2}\big). \tag{94}$$

Then

$$\mathcal{B}^2(\boldsymbol{A},\boldsymbol{B}) \;\leq\; \big\|\boldsymbol{A}^{1/2} - \boldsymbol{B}^{1/2}\big\|_F^2 \;\leq\; \|\boldsymbol{A}-\boldsymbol{B}\|_* \;\leq\; \sqrt{d}\,\|\boldsymbol{A}-\boldsymbol{B}\|_F, \tag{95}$$

where $\|\cdot\|_*$ denotes the nuclear (trace) norm, i.e. the sum of the singular values.

Proof. First inequality. Expanding the Frobenius norm, we get

$$\begin{aligned}\big\|\boldsymbol{A}^{1/2} - \boldsymbol{B}^{1/2}\big\|_F^2 &= \mathrm{Tr}\big((\boldsymbol{A}^{1/2} - \boldsymbol{B}^{1/2})(\boldsymbol{A}^{1/2} - \boldsymbol{B}^{1/2})^\top\big)\\ &= \mathrm{Tr}\big(\boldsymbol{A} - \boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2} - \boldsymbol{B}^{1/2}\boldsymbol{A}^{1/2} + \boldsymbol{B}\big)\\ &= \mathrm{Tr}(\boldsymbol{A}) + \mathrm{Tr}(\boldsymbol{B}) - 2\,\mathrm{Tr}(\boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2})\end{aligned} \tag{96}$$

Indeed, for any $\boldsymbol{C} \succeq 0$, its symmetric square root $\boldsymbol{C}^{1/2}$ satisfies $\boldsymbol{C}^{1/2}\boldsymbol{C}^{1/2} = \boldsymbol{C}$. Finally, we have $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{Tr}(\boldsymbol{B}\boldsymbol{A})$. Using the definition of $\mathcal{B}^2$ in (94) and (96), the claim is equivalent to

$$\mathrm{Tr}(\boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2}) \leq \mathrm{Tr}\big((\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2})^{1/2}\big). \tag{97}$$

Set $\boldsymbol{M} \coloneqq \boldsymbol{B}^{1/2}\boldsymbol{A}^{1/2}$. Then

$$\boldsymbol{M}^\top\boldsymbol{M} = \boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2}\boldsymbol{B}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}. \tag{98}$$

Since the singular values of $\boldsymbol{M}$ are the eigenvalues of the positive-semi-definite matrix $(\boldsymbol{M}^\top\boldsymbol{M})^{1/2}$, the nuclear norm satisfies, by (98),

$$\|\boldsymbol{M}\|_* = \mathrm{Tr}\big((\boldsymbol{M}^\top\boldsymbol{M})^{1/2}\big) = \mathrm{Tr}\big((\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2})^{1/2}\big). \tag{99}$$

By cyclicity,

$$\mathrm{Tr}(\boldsymbol{M}) = \mathrm{Tr}(\boldsymbol{A}^{1/2}\boldsymbol{B}^{1/2}). \tag{100}$$

Using (97), (99), (100), the claim is $\mathrm{Tr}(\boldsymbol{M}) \le \|\boldsymbol{M}\|_*$, which follows immediately from von Neumann's trace inequality (Horn & Johnson, 2012).

Second inequality. This is the Powers–Størmer inequality (Powers & Størmer, 1970); see also Santoro & Panaretos (2025, Lemma 1).

Third inequality. Let $s_1, \dots, s_d \ge 0$ be the singular values of $\boldsymbol{A} - \boldsymbol{B}$, so that $\|\boldsymbol{A} - \boldsymbol{B}\|_* = \sum_{i=1}^d s_i$ and $\|\boldsymbol{A} - \boldsymbol{B}\|_F = \left(\sum_{i=1}^d s_i^2\right)^{1/2}$. Lemma A.3 gives

$$\|\boldsymbol{A} - \boldsymbol{B}\|_* = \sum_{i=1}^d s_i \le \sqrt{d}\left(\sum_{i=1}^d s_i^2\right)^{1/2} = \sqrt{d}\,\|\boldsymbol{A} - \boldsymbol{B}\|_F.$$

□

We now characterize when the fitted distribution converges to the true distribution in the 2-Wasserstein distance.

Theorem A.24 (W_2 consistency)**.** *Under Assumption 1,*

$$\mathbb{E}\big[W_2^2(\mathcal{D}_t, \mathcal{D}_{\mathrm{real}})\big] \xrightarrow{t\to\infty} 0 \quad\Longleftrightarrow\quad \sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty.$$

Proof. Recall Definition 2. For Gaussians, it admits the closed form (Dowson & Landau, 1982; Gelbrich, 1990)

$$W_2^2\big(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\big) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathcal{B}^2(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2), \tag{101}$$

where $\mathcal{B}$ is the Bures metric (94). Applying (101) with $\boldsymbol{\mu}_1 = \boldsymbol{\mu}^{(t)}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}^{(t)}$, $\boldsymbol{\mu}_2 = \boldsymbol{\mu}$, $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, and writing $\mathcal{B}_t := \mathcal{B}(\boldsymbol{\Sigma}^{(t)}, \boldsymbol{\Sigma})$, we obtain from (101) the pointwise identity

$$W_2^2(\mathcal{D}_t, \mathcal{D}_{\mathrm{real}}) = \|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2 + \mathcal{B}_t^2, \tag{102}$$

in which both summands are nonnegative. Taking expectations in (102) and using $\mathbb{E}\big[\|\boldsymbol{\delta}^{(t)}\|_2^2\big] = \mathrm{Tr}(\boldsymbol{C}_t)$ from (11),

$$\mathbb{E}\big[W_2^2(\mathcal{D}_t, \mathcal{D}_{\mathrm{real}})\big] = \mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] + \mathbb{E}\big[\mathcal{B}_t^2\big]. \tag{103}$$

Sufficiency. Suppose $\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty$. By Lemma A.8 we have $\boldsymbol{\Sigma}^{(t)} \succeq 0$, and $\boldsymbol{\Sigma} \succ 0$ by assumption, so Lemma A.23 applies and gives the pointwise bound

$$\mathcal{B}_t^2 \;\le\; \sqrt{d}\,\big\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\big\|_F.$$

Substituting into (103),

$$\mathbb{E}\big[W_2^2(\mathcal{D}_t, \mathcal{D}_{\mathrm{real}})\big] \le \mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] + \sqrt{d}\,\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big]. \tag{104}$$

By Theorem A.13, the hypothesis gives $\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \to 0$. By Theorem A.22, the hypothesis also gives $\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big] \to 0$. Both terms on the right-hand side of (104) therefore vanish, so $\mathbb{E}\big[W_2^2(\mathcal{D}_t, \mathcal{D}_{\mathrm{real}})\big] \to 0$.

Necessity. Conversely, suppose $\mathbb{E}\big[W_2^2(\mathcal{D}_t, \mathcal{D}_{\mathrm{real}})\big] \to 0$. Since $\mathbb{E}\big[\mathcal{B}_t^2\big] \ge 0$, identity (103) gives

$$0 \le \mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \le \mathbb{E}\big[W_2^2(\mathcal{D}_t, \mathcal{D}_{\mathrm{real}})\big] \xrightarrow{t\to\infty} 0.$$

Thus, $\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \to 0$. Using Theorem A.13 and Lemma A.10, we have $\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty$. This establishes the stated equivalence. □

A.5 Proof of Theorem 3.3

Theorem 3.3 characterizes in closed form when this iterative estimation process converges to the true distribution. We now assemble it from Lemma A.10 and Theorems A.13, A.22, and A.24.

Proof. It suffices to show that (a) is equivalent to each of (b), (c), and (d). First, (a)⇔(b) follows by combining Theorem A.13 with Lemma A.10. Second, (a)⇔(c) is exactly Theorem A.22. Finally, (a)⇔(d) is precisely Theorem A.24. □

A.6 Proof of Lemma 3.4

Before proving Lemma 3.4, we prove an important technical lemma.

Lemma A.25 (Diverging series)**.** *Let $0 < a_s \to \infty$ be strictly increasing; then*

$$\sum_{s=1}^{\infty} \left(1 - \frac{a_s}{a_{s+1}}\right) = \infty$$

Proof. Take $\{a_s\}_{s=1}^{\infty}$ strictly increasing such that $0 < a_s \to \infty$. We now use Lemma A.6 with $r_s = \frac{a_s}{a_{s+1}} \in (0,1)$ (where we used $a_s > 0$ and $a_s < a_{s+1}$). The condition of the lemma holds:

$$\prod_{s=1}^{t} r_s = \prod_{s=1}^{t} \frac{a_s}{a_{s+1}} = \frac{a_1}{a_{t+1}} \xrightarrow[t\to\infty]{} 0,$$

where we used $a_t \to \infty$. This proves

$$\sum_{s=1}^{\infty} \left(1 - \frac{a_s}{a_{s+1}}\right) = \sum_{s=1}^{\infty} (1 - r_s) = \infty$$

□

Proof of Lemma 3.4. Let $\bar{\alpha} := \sup_{t\geq 1} \alpha_t \in [0,1)$. We prove that $\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) = \infty$. Indeed, note that

$$\sum_{t=1}^{\infty}(1-\alpha_t)(1-\rho_t) \geq (1-\bar{\alpha})\sum_{t=1}^{\infty}(1-\rho_t) = (1-\bar{\alpha})\sum_{t=1}^{\infty}\left(1 - \frac{N_t}{N_{t+1}}\right) = \infty.$$

We used $1 - \bar{\alpha} > 0$ and

$$\sum_{t=1}^{\infty}\left(1 - \frac{N_t}{N_{t+1}}\right) = \infty.$$

This directly follows from Lemma A.25 with $a_t = N_t$. Indeed, $N_t = N_{t-1} + n_b(t-1) > N_{t-1}$ is strictly increasing, and $N_t = n_r + \sum_{s=1}^{t-1} n_b(s) \geq t - 1 \to \infty \Rightarrow N_t \to \infty$. □

A.7 Proof of Proposition 3.5

Proof. Assume $\alpha_t \equiv 1$ throughout. We start by bounding $\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big]$.

Step 1 (lower bound). By Theorem A.11(d), the recursion becomes

$$\boldsymbol{C}_t = (1-\beta_{t-1})\boldsymbol{C}_{t-1} + \beta_{t-1}\boldsymbol{\Sigma}, \tag{105}$$

and hence

$$\boldsymbol{C}_t - \boldsymbol{C}_{t-1} = \beta_{t-1}\big(\boldsymbol{\Sigma} - \boldsymbol{C}_{t-1}\big), \qquad t \geq 2.$$

By Theorem 3.1(b) and Lemma A.8, $\boldsymbol{C}_{t-1} = \boldsymbol{\Sigma} - \mathbb{E}\big[\boldsymbol{\Sigma}^{(t-1)}\big] \preceq \boldsymbol{\Sigma}$, and $\beta_{t-1} > 0$ by (18). Hence, $\boldsymbol{C}_t - \boldsymbol{C}_{t-1} \succeq 0$, i.e., $\boldsymbol{C}_t \succeq \boldsymbol{C}_{t-1}$ for every $t \geq 2$. By induction,

$$\boldsymbol{C}_t \succeq \boldsymbol{C}_1 \tag{106}$$

for all $t \geq 1$. Since the trace is monotone with respect to the positive semidefinite order,

$$\operatorname{Tr}(\boldsymbol{C}_t) \geq \operatorname{Tr}(\boldsymbol{C}_1) = \frac{\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r},$$

where we used (46). Therefore, by (11),

$$\frac{\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r} \leq \mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big]. \tag{107}$$

Step 2 (upper bound). Applying trace in (105), we get

$$\begin{aligned}\operatorname{Tr}(\boldsymbol{C}_t) &= (1-\beta_{t-1})\operatorname{Tr}(\boldsymbol{C}_{t-1}) + \beta_{t-1}\operatorname{Tr}(\boldsymbol{\Sigma}) \\ &\leq \operatorname{Tr}(\boldsymbol{C}_{t-1}) + \beta_{t-1}\operatorname{Tr}(\boldsymbol{\Sigma}),\end{aligned}$$

where we used $\beta_{t-1} > 0$ by (18). By induction,

$$\begin{aligned}\operatorname{Tr}(\boldsymbol{C}_t) &\leq \operatorname{Tr}(\boldsymbol{C}_1) + \left(\sum_{s=2}^{t} \beta_{s-1}\right)\operatorname{Tr}(\boldsymbol{\Sigma}) \\ &\leq \frac{\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r} + \left(\sum_{s=1}^{\infty} \beta_s\right)\operatorname{Tr}(\boldsymbol{\Sigma}) \\ &\leq \frac{2\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r},\end{aligned} \tag{108}$$

where we used (46), $\beta_t > 0$ from (18) and Lemma A.5. Therefore, by (11),

$$\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \leq \frac{2\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r}. \tag{109}$$

Combining (107) and (109), we get

$$\frac{\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r} \leq \mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] \leq \frac{2\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r}. \tag{110}$$

Finally, we bound $\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big]$.

Step 3 (lower bound). Using (93) and (106)

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big] &\geq \|\boldsymbol{C}_t\|_F \\ &\geq \|\boldsymbol{C}_1\|_F \\ &\geq \frac{\|\boldsymbol{\Sigma}\|_F}{n_r}\end{aligned} \tag{111}$$

where we used the monotonicity of the Frobenius norm with respect to the positive semidefinite order and (46).

Step 4 (upper bound). Note that (76), (83), (84) and (90) in the proof of Theorem A.22 do not use the hypothesis $\sum_{s=1}^{\infty}(1-\alpha_s)(1-\rho_s) = \infty$. Hence, they hold for $\alpha_t \equiv 1$. Combining (90) and (108), we get

$$\begin{aligned}\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big] &\leq \sqrt{P_t} + \operatorname{Tr}(\boldsymbol{C}_t) \\ &\leq \sqrt{P_t} + \frac{2\operatorname{Tr}(\boldsymbol{\Sigma})}{n_r}.\end{aligned} \tag{112}$$

Using (18) and $\alpha_t \equiv 1$, we have for every $t \geq 1$,

$$\lambda_t = \frac{N_t + n_b(t)\alpha_t}{N_{t+1}} = \frac{N_t + n_b(t)}{N_{t+1}} = \frac{N_{t+1}}{N_{t+1}} = 1.$$

Thus, $\lambda_t \equiv 1$. Thus, (84) reads

$$\begin{aligned} P_t &\leq \left(\prod_{k=1}^{t-1} \lambda_k^2\right) P_1 + D \sum_{s=1}^{t-1} \left(\prod_{k=s+1}^{t-1} \lambda_k^2\right) \beta_s \\ &= P_1 + D \sum_{s=1}^{t-1} \beta_s \\ &\leq P_1 + D \sum_{s=1}^{\infty} \beta_s \\ &\leq P_1 + \frac{D}{n_r}, \end{aligned} \tag{113}$$

where we used $\beta_t > 0$ from (18) and Lemma A.5. It is left to bound P_1 and D.

Bounding D. From (83), we have

$$D = 3B^* + C_{\mathbf{\Sigma}}, \tag{114}$$

where $C_{\mathbf{\Sigma}}$ is defined in (53), and $B^* = \sup_{t \geq 1} \mathbb{E}\big[\mathrm{Tr}(\hat{\boldsymbol{S}}_t)^2\big]$ is finite from Lemma A.19.
We now bound B^*. Using (71), we have

$$B^* \leq \exp\left(\frac{3}{n_r}\right)(v_1 + D_0) + \mathrm{Tr}(\mathbf{\Sigma})^2. \tag{115}$$

Using (63) and (70),

$$v_1 \leq \frac{1}{n_r} C_{\mathbf{\Sigma}}, \tag{116}$$

and

$$D_0 \leq \frac{1}{n_r}\left(C_{\mathbf{\Sigma}} + 3\,\mathrm{Tr}(\mathbf{\Sigma})^2\right). \tag{117}$$

Combining (114), (115), (116), and (117), we get

$$\begin{aligned} D &\leq 3\left(\exp\left(\frac{3}{n_r}\right)(v_1 + D_0) + \mathrm{Tr}(\mathbf{\Sigma})^2\right) + C_{\mathbf{\Sigma}} \\ &\leq 3\left(\exp\left(\frac{3}{n_r}\right)\left(\frac{1}{n_r} C_{\mathbf{\Sigma}} + \frac{1}{n_r}\left(C_{\mathbf{\Sigma}} + 3\,\mathrm{Tr}(\mathbf{\Sigma})^2\right)\right) + \mathrm{Tr}(\mathbf{\Sigma})^2\right) + C_{\mathbf{\Sigma}} \\ &= \frac{3}{n_r}\exp\left(\frac{3}{n_r}\right)\left(2C_{\mathbf{\Sigma}} + 3\,\mathrm{Tr}(\mathbf{\Sigma})^2\right) + 3\,\mathrm{Tr}(\mathbf{\Sigma})^2 + C_{\mathbf{\Sigma}}. \end{aligned} \tag{118}$$

Bounding P_1. For a random matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, we write $\mathrm{Cov}(\boldsymbol{X}) \coloneqq \mathrm{Cov}(\mathrm{vec}(\boldsymbol{X})) \in \mathbb{R}^{d^2\times d^2}$. As for random vectors, Cov is additive over independent summands. We have

$$\begin{aligned}
P_1 &= \mathbb{E}\big[\|\hat{\boldsymbol{S}}_1 - \boldsymbol{\Sigma}\|_F^2\big] \\
&= \mathbb{E}\big[\|\hat{\boldsymbol{S}}_1 - \mathbb{E}[\hat{\boldsymbol{S}}_1]\|_F^2\big] \\
&= \mathrm{Tr}\Big(\mathrm{Cov}(\hat{\boldsymbol{S}}_1)\Big) \\
&= \mathrm{Tr}\left(\mathrm{Cov}\left(\frac{1}{n_r}\sum_{j=1}^{n_r}(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top\right)\right) \\
&= \mathrm{Tr}\left(\frac{1}{n_r^2}\sum_{j=1}^{n_r}\mathrm{Cov}\Big((\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_j^{(0)} - \boldsymbol{\mu})^\top\Big)\right) \\
&= \frac{1}{n_r}\mathrm{Tr}\Big(\mathrm{Cov}\Big((\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu})^\top\Big)\Big) \\
&= \frac{1}{n_r}\Big(\mathbb{E}\big[\|(\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu})^\top\|_F^2\big] - \|\mathbb{E}[(\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu})^\top]\|_F^2\Big) \\
&= \frac{1}{n_r}\Big(\mathbb{E}\big[\|\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu}\|_2^4\big] - \|\boldsymbol{\Sigma}\|_F^2\Big) \\
&= \frac{1}{n_r}\left(C_{\boldsymbol{\Sigma}} - \|\boldsymbol{\Sigma}\|_F^2\right).
\end{aligned} \tag{119}$$

First, we used the definitions of P_t and $\hat{\boldsymbol{S}}_t$ from Theorem A.22 and (54), respectively. Then, we used Lemma A.14(iv), the additivity of the covariance for the i.i.d. samples $\{\boldsymbol{x}_j^{(0)}\}_{j=1}^{n_r}$ and the following property

$$\mathrm{Tr}(\mathrm{Cov}(\boldsymbol{X})) = \mathbb{E}[\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\|_F^2] = \mathbb{E}[\|\boldsymbol{X}\|_F^2] - \|\mathbb{E}[\boldsymbol{X}]\|_F^2.$$

Finally, we used the identities $\mathbb{E}[(\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu})(\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu})^\top] = \boldsymbol{\Sigma}$, $\|\boldsymbol{x}\boldsymbol{x}^\top\|_F^2 = \|\boldsymbol{x}\|_2^4$, and $C_{\boldsymbol{\Sigma}} = \mathbb{E}\big[\|\boldsymbol{x}_1^{(0)} - \boldsymbol{\mu}\|_2^4\big]$.

Combining (113), (118) and (119), we get

$$\begin{aligned}
P_t &\le \frac{1}{n_r}\left(C_{\boldsymbol{\Sigma}} - \|\boldsymbol{\Sigma}\|_F^2\right) + \frac{3}{n_r^2}\exp\left(\frac{3}{n_r}\right)\left(2C_{\boldsymbol{\Sigma}} + 3\,\mathrm{Tr}(\boldsymbol{\Sigma})^2\right) + \frac{1}{n_r}\left(3\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + C_{\boldsymbol{\Sigma}}\right) \\
&= \frac{3}{n_r^2}\exp\left(\frac{3}{n_r}\right)\left(2C_{\boldsymbol{\Sigma}} + 3\,\mathrm{Tr}(\boldsymbol{\Sigma})^2\right) + \frac{1}{n_r}\left(2C_{\boldsymbol{\Sigma}} + 3\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 - \|\boldsymbol{\Sigma}\|_F^2\right) \\
&= \frac{3}{n_r^2}\exp\left(\frac{3}{n_r}\right)\left(5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 4\|\boldsymbol{\Sigma}\|_F^2\right) + \frac{1}{n_r}\left(5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 3\|\boldsymbol{\Sigma}\|_F^2\right),
\end{aligned} \tag{120}$$

where we used (53).

Then, combining (112) and (120), we get

$$\begin{aligned}
\mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big] &\le \sqrt{\frac{3}{n_r^2}\exp\left(\frac{3}{n_r}\right)(5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 4\|\boldsymbol{\Sigma}\|_F^2) + \frac{1}{n_r}(5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 3\|\boldsymbol{\Sigma}\|_F^2)} + \frac{2\,\mathrm{Tr}(\boldsymbol{\Sigma})}{n_r} \\
&= \frac{2\,\mathrm{Tr}(\boldsymbol{\Sigma})}{n_r} + \frac{1}{\sqrt{n_r}}\sqrt{(5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 3\|\boldsymbol{\Sigma}\|_F^2) + \frac{3}{n_r}\exp\left(\frac{3}{n_r}\right)(5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 4\|\boldsymbol{\Sigma}\|_F^2)}.
\end{aligned} \tag{121}$$

Finally, defining $A_{\boldsymbol{\Sigma}} = 5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 3\|\boldsymbol{\Sigma}\|_F^2$ and $B_{\boldsymbol{\Sigma}} = 5\,\mathrm{Tr}(\boldsymbol{\Sigma})^2 + 4\|\boldsymbol{\Sigma}\|_F^2$, and combining (111) and (121), we get

$$\frac{\|\boldsymbol{\Sigma}\|_F}{n_r} \le \mathbb{E}\big[\|\boldsymbol{\Sigma}^{(t)} - \boldsymbol{\Sigma}\|_F\big] \le \frac{2\,\mathrm{Tr}(\boldsymbol{\Sigma})}{n_r} + \frac{1}{\sqrt{n_r}}\sqrt{A_{\boldsymbol{\Sigma}} + \frac{3}{n_r}\exp\left(\frac{3}{n_r}\right)B_{\boldsymbol{\Sigma}}}. \tag{122}$$

From (110) and (122), the proposition follows.

□

A.8 Proof of Theorem 3.6

We first state a lemma that will be used in the following proof to compare sums and integrals.

Lemma A.26 (Sum and integral equivalence)**.** *Let $a \in \mathbb{Z}$, let $f : [a, \infty) \to (0, \infty)$ be non-increasing, and fix $h \geq 0$. Then for every integer $b > a$:*

(a) $\displaystyle\int_a^b f(x)\,dx \leq \sum_{t=a}^{b-1} f(t) \leq f(a) + \int_a^b f(x)\,dx$*; in particular* $\sum_{t=a}^{b-1} f(t) = \int_a^b f(x)\,dx + O(1)$.

(b) $\displaystyle\sum_{t=a}^{b-1} f(t) \asymp \int_a^b f(x)\,dx.$

(c) $\displaystyle\int_{a+h}^{b+h} f(x)\,dx \asymp \int_a^b f(x)\,dx.$

All implied constants are independent of b and depend only on f and h. If moreover $\int_a^\infty f(x)\,dx = \infty$, then the ratio of the two sides in (b) and (c) tends to 1 as $b \to \infty$, so $\asymp$ may be replaced by $\sim$.

Proof. Since f is monotone on $[a, \infty)$, it is Riemann integrable on every compact sub-interval, so all integrals below are well defined and finite. Note also that $f > 0$ everywhere, so every quantity appearing in a denominator is nonzero.

Throughout, we use the following elementary comparison: for every real $y \geq a$, monotonicity gives $f(y) \geq f(x) \geq f(y+1)$ for all $x \in [y, y+1]$, and hence

$$f(y) \;\geq\; \int_y^{y+1} f(x)\,dx \;\geq\; f(y+1). \tag{123}$$

(a) Fix an integer $b > a$. Applying the first inequality in (123) at $y = t$ and summing over $t = a, \ldots, b-1$,

$$\sum_{t=a}^{b-1} f(t) \;\geq\; \sum_{t=a}^{b-1} \int_t^{t+1} f(x)\,dx \;=\; \int_a^b f(x)\,dx,$$

which is the lower bound. For the upper bound, apply the second inequality in (123) at $y = t-1$, i.e. $f(t) \leq \int_{t-1}^t f(x)\,dx$, and sum over $t = a+1, \ldots, b-1$ (an empty sum if $b = a+1$):

$$\sum_{t=a}^{b-1} f(t) \;=\; f(a) + \sum_{t=a+1}^{b-1} f(t) \;\leq\; f(a) + \sum_{t=a+1}^{b-1} \int_{t-1}^{t} f(x)\,dx = f(a) + \int_a^{b-1} f(x)\,dx \;\leq\; f(a) + \int_a^b f(x)\,dx,$$

the last step because $f > 0$. Consequently, writing $D(b) := \sum_{t=a}^{b-1} f(t) - \int_a^b f(x)\,dx$, we have shown

$$0 \;\leq\; D(b) \;\leq\; f(a) \qquad \text{for every integer } b > a, \tag{124}$$

which is precisely the assertion $\sum_{t=a}^{b-1} f(t) = \int_a^b f(x)\,dx + O(1)$, with implied constant $f(a)$ independent of b.

(b) Fix an integer $b > a$; then $b \geq a+1$, so by (123) applied at $y = a$,

$$\int_a^b f(x)\,dx \;\geq\; \int_a^{a+1} f(x)\,dx \;\geq\; f(a+1) \;>\; 0 \Rightarrow \frac{1}{f(a+1)} \int_a^b f(x)\,dx \geq 1. \tag{125}$$

Combining (124) and (125),

$$\sum_{t=a}^{b-1} f(t) \;\leq\; f(a) + \int_a^b f(x)\,dx \;\leq\; \frac{f(a)}{f(a+1)} \int_a^b f(x)\,dx + \int_a^b f(x)\,dx,$$

so that, together with the lower bound from (a),

$$\int_a^b f(x)\,dx \;\le\; \sum_{t=a}^{b-1} f(t) \;\le\; \Big(1+\tfrac{f(a)}{f(a+1)}\Big)\int_a^b f(x)\,dx.$$

Both constants are independent of b, proving (b).

(c) Fix $h \ge 0$ and an integer $b > a$. The change of variables $x = y + h$ gives $\int_{a+h}^{b+h} f(x)\,dx = \int_a^b f(y+h)\,dy$, and since $f(y+h) \le f(y)$ for $h \ge 0$,

$$\int_{a+h}^{b+h} f(x)\,dx \;\le\; \int_a^b f(x)\,dx. \tag{126}$$

For the reverse bound, additivity of the integral yields

$$\int_a^b f(x)\,dx - \int_{a+h}^{b+h} f(x)\,dx \;=\; \int_a^{a+h} f(x)\,dx - \int_b^{b+h} f(x)\,dx \;\le\; \int_a^{a+h} f(x)\,dx \;\le\; hf(a), \tag{127}$$

using $f > 0$ and then $f \le f(a)$ on $[a, a+h]$, because f is non-increasing. Since $b \ge a+1$, the interval $[a+h, b+h]$ contains $[a+h, a+h+1]$, so by (123) at $y = a+h$,

$$\int_{a+h}^{b+h} f(x)\,dx \;\ge\; \int_{a+h}^{a+h+1} f(x)\,dx \;\ge\; f(a+h+1) \;>\; 0 \Rightarrow \frac{1}{f(a+h+1)}\int_{a+h}^{b+h} f(x)\,dx \ge 1.$$

Feeding this into (127),

$$\int_a^b f(x)\,dx \;\le\; \int_{a+h}^{b+h} f(x)\,dx + hf(a) \;\le\; \Big(1+\tfrac{hf(a)}{f(a+h+1)}\Big)\int_{a+h}^{b+h} f(x)\,dx,$$

which together with (126) proves (c) with constants independent of b.

The case $\int_a^\infty f(x)\,dx = \infty$. Here $\int_a^b f(x)\,dx \to \infty$ as $b \to \infty$, so we may divide by it for all large b. By (124),

$$\frac{\sum_{t=a}^{b-1} f(t)}{\int_a^b f(x)\,dx} \;=\; 1 + \frac{D(b)}{\int_a^b f(x)\,dx} \;\in\; \left[1,\; 1+\frac{f(a)}{\int_a^b f(x)\,dx}\right] \;\xrightarrow[b\to\infty]{}\; 1,$$

giving $\sum_{t=a}^{b-1} f(t) \sim \int_a^b f(x)\,dx$. Similarly, both $\int_a^{a+h} f$ and $\int_b^{b+h} f$ lie in $[0, hf(a)]$, so by (127) the difference $\int_a^b f - \int_{a+h}^{b+h} f$ lies in $[-hf(a), hf(a)]$; dividing by $\int_a^b f(x)\,dx \to \infty$ gives

$$\left|1 - \frac{\int_{a+h}^{b+h} f(x)\,dx}{\int_a^b f(x)\,dx}\right| \le \frac{hf(a)}{\int_a^b f(x)\,dx} \xrightarrow{b\to\infty} 0 \Rightarrow \frac{\int_{a+h}^{b+h} f(x)\,dx}{\int_a^b f(x)\,dx} \xrightarrow{b\to\infty} 1.$$

□

Proof of Theorem 3.6. Throughout $\alpha_t \equiv \alpha \in [0,1]$ and $n_b(t) \equiv n_b$ are fixed, so that $N_t = n_r + (t-1)n_b$ and $\kappa = 2(1-\alpha) \in [0,2]$ (note that $\kappa = 0$ when $\alpha = 1$). Since $n_b = N_{t+1} - N_t$ and $N_t \ge n_r$, we have

$$\frac{N_{t+1}}{N_t} = \frac{N_t + n_b}{N_t} = 1 + \frac{n_b}{N_t} \le 1 + \frac{n_b}{n_r},$$

so $N_t \asymp N_{t+1} \asymp tn_b$; all implied constants below depend only on α, n_b, n_r and $\mathrm{Tr}(\mathbf{\Sigma})$.

We use that for strictly positive sequences, an $\asymp$ relation valid for all large indices extends to all of them by enlarging the constants over the remaining finite set. Additionally, by (11), note that $\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] = \mathrm{Tr}(\boldsymbol{C}_t)$. Thus, giving an asymptotic analysis for $\mathrm{Tr}(\boldsymbol{C}_t)$ suffices.

Step 1: $\sum_{t=1}^{m}(1-\gamma_t) = \kappa \ln N_{m+1} + O(1)$. By (51) with $\alpha_t \equiv \alpha$,

$$\begin{aligned}
1-\gamma_t &= \frac{n_b\big[(1-\alpha)\big(2N_t + n_b(1+\alpha)\big) + \alpha^2\big]}{N_{t+1}^2} \\
&= \kappa\, n_b\, \frac{N_t}{N_{t+1}^2} + \frac{n_b\big[(1-\alpha)n_b(1+\alpha) + \alpha^2\big]}{N_{t+1}^2} \\
&= \kappa n_b \frac{N_{t+1} - n_b}{N_{t+1}^2} + \frac{n_b[n_b(1-\alpha^2) + \alpha^2]}{N_{t+1}^2} \\
&= \frac{\kappa n_b}{N_{t+1}} + \big(n_b(1-\alpha^2) + \alpha^2 - \kappa n_b\big) \cdot \frac{n_b}{N_{t+1}^2} \\
&= \frac{\kappa n_b}{N_{t+1}} + \big(\alpha^2 - n_b(1-\alpha)^2\big)\,\beta_t,
\end{aligned}$$

where we used

$n_b(1-\alpha^2)+\alpha^2-\kappa n_b = n_b\big(1-\alpha^2-2(1-\alpha)\big)+\alpha^2 = n_b\big(-\alpha^2+2\alpha-1\big)+\alpha^2 = \alpha^2 - n_b(1-\alpha)^2$.

Thus,

$$\sum_{t=1}^{m}(1-\gamma_t) = \kappa n_b \cdot \sum_{t=1}^{m}\frac{1}{N_{t+1}} + \big(\alpha^2 - n_b(1-\alpha)^2\big)\sum_{t=1}^{m}\beta_t = \kappa \ln N_{m+1} + O(1). \tag{128}$$

We used Lemma A.5 ($\sum_{t=1}^{\infty}\beta_t < \infty$) and

$$\sum_{t=1}^{m}\frac{1}{N_{t+1}} = \frac{1}{n_b}\ln N_{m+1} + O(1).$$

It follows from Lemma A.26(a) with $a=1, b=m+1$, and the function $f(x) = 1/\left(n_r + n_b x\right)$. It is positive and non-increasing, thus:

$$\begin{aligned}
\sum_{t=1}^{m}\frac{1}{N_{t+1}} = \sum_{t=1}^{m}\frac{1}{n_r + n_b t} &= \sum_{t=1}^{m} f(t) \\
&= \int_1^{m+1} f(x)\,dx + O(1) \\
&= \int_1^{m+1} \frac{1}{n_r + n_b x}\,dx + O(1) \\
&= \left(\frac{1}{n_b}\ln(n_r + n_b x)\right)\Big|_1^{m+1} + O(1) \\
&= \frac{1}{n_b}\left(\ln\left(n_r + n_b(m+1)\right) - \ln(n_r + n_b)\right) + O(1) \\
&= \frac{1}{n_b}\cdot\ln(N_{m+2}) + O(1) \\
&= \frac{1}{n_b}\ln\left(N_{m+1}\right) + O(1).
\end{aligned}$$

We used $N_{m+2} \asymp N_{m+1} \Rightarrow \ln(N_{m+2}) = \ln(N_{m+1}) + O(1)$. This holds uniformly in m: $1 \le N_{m+2}/N_{m+1} \le 1 + n_b/n_r$.

Step 2: $\Gamma_{1,m} \asymp N_{m+1}^{-\kappa}$, uniformly in $m \ge 1$. From (51) and $\alpha_t \in [0,1]$, we have

$$\begin{aligned}
0 \le 1-\gamma_t &= \frac{n_b}{N_{t+1}^2}\cdot\left[(1-\alpha_t)\cdot(2N_t + n_b(1+\alpha_t)) + \alpha_t^2\right] \\
&\le \frac{n_b}{N_{t+1}^2}\cdot(2N_t + 2n_b + 1) \\
&= \frac{n_b}{N_{t+1}^2}\cdot(2N_{t+1} + 1) \\
&\le \frac{3n_b}{N_{t+1}} \xrightarrow{t\to\infty} 0
\end{aligned}$$

That is, $1-\gamma_t \to 0$, and there is t_1 with $1-\gamma_t \le \frac{1}{2}$ for $t \ge t_1$. That is,

$$t \ge t_1 \Rightarrow \gamma_t \ge \frac{1}{2}. \tag{129}$$

Finally, from (19), we have

$$\Gamma_{1,m} = \prod_{t=1}^{m} \gamma_t = \exp\left(\sum_{t=1}^{m} \ln \gamma_t\right) = \exp\left(-\sum_{t=1}^{m} -\ln \gamma_t\right) = \exp\left(-B_m\right), \tag{130}$$

where we defined

$$B_m \coloneqq \sum_{t=1}^{m} -\ln \gamma_t. \tag{131}$$

Before we analyze B_m, we record two elementary bounds on the logarithm. First, $-\ln u \ge 0$ for every $u \in (0,1]$, by monotonicity of $\ln$. Second,

$$(1-u) \le -\ln u \le (1-u) + (1-u)^2, \qquad u \in \left[\tfrac{1}{2}, 1\right]. \tag{132}$$

Indeed, writing $x \coloneqq 1-u \in [0, \frac{1}{2}]$, the series $-\ln u = -\ln(1-x) = \sum_{k=1}^{\infty} x^k/k$ converges, and

$$0 \le -\ln u - (1-u) = \sum_{k=2}^{\infty} \frac{x^k}{k} \le \frac{1}{2}\sum_{k=2}^{\infty} x^k = \frac{x^2}{2(1-x)} = \frac{(1-u)^2}{2u} \le (1-u)^2,$$

the first inequality because every term is nonnegative, and the last because $u \ge \frac{1}{2}$. Rearranging gives (132).

Fix $m \ge t_1$. From (131), an upper bound on B_m reads:

$$\begin{aligned}
B_m &= \underbrace{\sum_{t=1}^{t_1-1} -\ln \gamma_t}_{O(1)} + \sum_{t=t_1}^{m} -\ln \gamma_t \\
&\le O(1) + \sum_{t=t_1}^{m} (1-\gamma_t) + \sum_{t=t_1}^{m} (1-\gamma_t)^2 \\
&\le O(1) + \sum_{t=1}^{m} (1-\gamma_t) + \sum_{t=1}^{m} \frac{9n_b^2}{N_{t+1}^2} \\
&= O(1) + \kappa \ln N_{m+1} + 9n_b \sum_{t=1}^{m} \beta_t \\
&= O(1) + \kappa \ln N_{m+1}.
\end{aligned}$$

We used (129) and (132). Then, we used (128), $(1-\gamma_t) \le 3n_b/N_{t+1}$, the definition of β_t from (18), and the summability of β_t from Lemma A.5. That is,

$$B_m \le \kappa \ln N_{m+1} + O(1) \tag{133}$$

Additionally, from (131), a lower bound on B_m is:

$$\begin{aligned}
B_m &\ge \sum_{t=t_1}^{m} -\ln \gamma_t \\
&\ge \sum_{t=t_1}^{m} (1-\gamma_t) \\
&= \sum_{t=1}^{m} (1-\gamma_t) + O(1) \\
&= \kappa \ln N_{m+1} + O(1).
\end{aligned} \tag{134}$$

In the first step, we discarded the terms $t < t_1$, which is legitimate since $-\ln \gamma_t \geq 0$ for every t, by $\gamma_t \in (0,1)$ (Lemma A.9). In the second, we applied the left-hand inequality of (132), valid because $\gamma_t \in [\frac{1}{2}, 1)$ for $t \geq t_1$ by (129). In the third, we re-added the finitely many terms $t < t_1$, each lying in $[0,1]$, at a cost of $O(1)$. The last step is (128).

That is, from (133) and (134),

$$B_m = \kappa \ln N_{m+1} + O(1). \tag{135}$$

Finally, from (130) and (135), we have

$$\begin{aligned} \Gamma_{1,m} &= \exp\left(-\kappa \ln N_{m+1} + O(1)\right) \\ &= \exp\left(\ln(N_{m+1}^{-\kappa}) + O(1)\right) \\ &\asymp N_{m+1}^{-\kappa}. \end{aligned} \tag{136}$$

This was derived for $m \geq t_1$. For each of the finitely many $m < t_1$, the ratio $\Gamma_{1,m}/N_{m+1}^{-\kappa}$ is a strictly positive real number, so over this finite range, it is bounded between its minimum and its maximum, both positive. Enlarging the implied constants accordingly extends the equivalence to every $m \geq 1$.

That is, $\Gamma_{1,m} \asymp N_{m+1}^{-\kappa}$, where the implied constants are uniform in $m \geq 1$.

Step 3: $\Gamma_{s+1,t-1} \asymp \left(N_{s+1}/N_t\right)^{\kappa}$. Indeed, from Step 2 and (19), we have

$$\Gamma_{s+1,t-1} = \frac{\Gamma_{1,t-1}}{\Gamma_{1,s}} \asymp \frac{N_t^{-\kappa}}{N_{s+1}^{-\kappa}} = \left(\frac{N_{s+1}}{N_t}\right)^{\kappa}$$

The implied constants are uniform in s and t.

Step 4: evaluation of S_t. By Step 3 and $\beta_s = n_b/N_{s+1}^2$ from (18), we have

$$\begin{aligned} S_t = \sum_{s=1}^{t-1} \beta_s \Gamma_{s+1,t-1} &\asymp \sum_{s=1}^{t-1} \frac{n_b}{N_{s+1}^2} \cdot \frac{N_{s+1}^{\kappa}}{N_t^{\kappa}} \\ &= n_b N_t^{-\kappa} \sum_{s=1}^{t-1} N_{s+1}^{\kappa-2} \\ &= n_b N_t^{-\kappa} \sum_{s=1}^{t-1} (n_r + n_b s)^{\kappa-2}, \end{aligned} \tag{137}$$

where the constants of Step 3 are independent of t and therefore factor out of the sum term by term. Let us now use Lemma A.26(b) with $a = 1$ and $f(s) = (n_r + n_b s)^{\kappa-2}$. Indeed, $\kappa - 2 \leq 0$, and thus $f(s)$ is positive and non-increasing.

$$\begin{aligned} \sum_{s=1}^{t-1} (n_r + n_b s)^{\kappa-2} &\asymp \int_1^t (n_r + n_b x)^{\kappa-2}\, dx \\ &\overset{u=n_r+n_b x}{=} \frac{1}{n_b} \int_{n_r+n_b}^{n_r+n_b t} u^p\, du \\ &= \frac{1}{n_b} \int_{n_r+n_b}^{N_t+n_b} u^p\, du \\ &\asymp \frac{1}{n_b} \int_{n_r}^{N_t} u^p\, du, \end{aligned} \tag{138}$$

where we defined $p \coloneqq \kappa - 2 = -2\alpha \in [-2, 0]$, and used Lemma A.26(c) (using the decreasing and positive function $u \mapsto u^p$). Let us now split into two cases. From (138), we have

- If $p = -1$ (equivalent to $\kappa = 1$), we have

$$\sum_{s=1}^{t-1} (n_r + n_b s)^{\kappa-2} \asymp \frac{1}{n_b} \ln(u)\Big|_{n_r}^{N_t} = \frac{1}{n_b} \ln\left(\frac{N_t}{n_r}\right) \tag{139}$$

- If $p \neq -1$ (equivalent to $\kappa \neq 1$, i.e., $0 \leq \kappa < 1$ or $1 < \kappa \leq 2$), we have

$$\sum_{s=1}^{t-1} (n_r + n_b s)^{\kappa-2} \asymp \frac{1}{n_b} \frac{u^{p+1}}{p+1}\Big|_{n_r}^{N_t} = \frac{1}{n_b} \frac{N_t^{\kappa-1} - n_r^{\kappa-1}}{\kappa - 1} \asymp \begin{cases} \frac{n_r^{\kappa-1}}{n_b(1-\kappa)} & 0 \leq \kappa < 1 \\ \frac{N_t^{\kappa-1}}{n_b(\kappa-1)} & 1 < \kappa \leq 2 \end{cases} \tag{140}$$

From (139) and (140), we have

$$\sum_{s=1}^{t-1} (n_r + n_b s)^{\kappa-2} \asymp \frac{1}{n_b} \cdot \begin{cases} 1 & 0 \leq \kappa < 1 \\ \ln(N_t) & \kappa = 1 \\ N_t^{\kappa-1} & 1 < \kappa \leq 2 \end{cases} \tag{141}$$

Let us now substitute (141) into (137). We get

$$S_t \asymp \begin{cases} N_t^{-\kappa} & 0 \leq \kappa < 1 \\ N_t^{-1} \ln(N_t) & \kappa = 1 \\ N_t^{-1} & 1 < \kappa \leq 2 \end{cases} \tag{142}$$

Step 5: combination. Taking traces in the closed form (39) and using $\boldsymbol{C}_1 = \boldsymbol{\Sigma}/n_r$,

$$\mathrm{Tr}(\boldsymbol{C}_t) = \Gamma_{1,t-1} \frac{\mathrm{Tr}(\boldsymbol{\Sigma})}{n_r} + \mathrm{Tr}(\boldsymbol{\Sigma})\, S_t \asymp N_t^{-\kappa} + S_t, \tag{143}$$

where the two summands are non-negative, $\mathrm{Tr}(\boldsymbol{\Sigma}) > 0$ and $\mathrm{Tr}(\boldsymbol{C}_1) = \mathrm{Tr}(\boldsymbol{\Sigma})/n_r > 0$ are constants, and $\Gamma_{1,t-1} \asymp N_t^{-\kappa}$ by Step 2 (Equation (136)); a sum of two non-negative terms is $\asymp$ the sum of any two quantities where each $\asymp$ to the corresponding term.

Let us now split into three cases, using (142) for S_t asymptotics, and (143) for the final $\mathrm{Tr}(\boldsymbol{C}_t)$ asymptotics. We also use $\kappa = 2(1-\alpha)$.

- If $0 \leq \kappa < 1$ (equivalent to $\frac{1}{2} < \alpha \leq 1$), we have $S_t \asymp N_t^{-\kappa}$. Both terms in (143) have the same asymptotics, and we obtain $\mathrm{Tr}(\boldsymbol{C}_t) \asymp N_t^{-\kappa} = N_t^{-2(1-\alpha)}$.
- If $\kappa = 1$ (equivalent to $\alpha = \frac{1}{2}$), we have $S_t \asymp N_t^{-1} \ln(N_t)$. The second term in (143) dominates, and we obtain $\mathrm{Tr}(\boldsymbol{C}_t) \asymp N_t^{-1} \ln(N_t)$.
- If $1 < \kappa \leq 2$ (equivalent to $0 \leq \alpha < \frac{1}{2}$), we have $S_t \asymp N_t^{-1}$. Together with $N_t^{-\kappa} = o(N_t^{-1})$, the second term in (143) dominates, and we get $\mathrm{Tr}(\boldsymbol{C}_t) \asymp N_t^{-1}$.

Finally, we get

$$\mathrm{Tr}(\boldsymbol{C}_t) \asymp \begin{cases} N_t^{-1} & 0 \leq \alpha < \frac{1}{2} \\ N_t^{-1} \ln(N_t) & \alpha = \frac{1}{2} \\ N_t^{-2(1-\alpha)} & \frac{1}{2} < \alpha \leq 1 \end{cases}$$

Which is exactly (17). Note, that for $\alpha = 1$ we get $\mathrm{Tr}(\boldsymbol{C}_t) \asymp 1$, which means that $\mathrm{Tr}(\boldsymbol{C}_t)$ stays bounded as $t \to \infty$. □

Remark. When $\alpha = 0$, (17) gives $\mathrm{Tr}(\boldsymbol{C}_t) \asymp N_t^{-1}$, in agreement with the exact identity $\boldsymbol{C}_t = \boldsymbol{\Sigma}/N_t$ of Theorem A.11(c).

A.9 Additional result: the error of the mean estimator increases initially for large α

In Figure 1(a), the mean-square error of $\boldsymbol{\mu}^{(t)}$ first increases before it decreases when α is large. The following proposition characterizes exactly when this happens, shows that the error then has a single peak, and bounds how large the increase can be.

Proposition A.27 (Initial growth of the error)**.** *Under Assumption 1, let* $\alpha_t \equiv \alpha \in [0,1]$ *and* $n_b(t) \equiv n_b$*. Let*

$$\alpha^*(n_r, n_b) := \begin{cases} \dfrac{\sqrt{n_r^2 + (n_b-1)(n_r+n_b)} - n_r}{n_b - 1}, & n_b \geq 2, \\ \dfrac{n_r+1}{2n_r}, & n_b = 1, \end{cases} \tag{144}$$

which is the unique root in $[0,1]$ *of* $g(\alpha) := (n_b-1)\alpha^2 + 2n_r\alpha - (n_r+n_b)$ *and satisfies* $\frac{1}{2} < \alpha^* \leq 1$*. Then:*

(a) ***Threshold.*** *The error increases at the first iteration,* $c_2 > c_1$*, if and only if* $\alpha > \alpha^*$*. For* $\alpha \leq \alpha^*$*,* c_t *is non-increasing in* t*.*

(b) ***Single peak.*** *There is* $t^* \in \{1, 2, \ldots\} \cup \{\infty\}$ *such that* c_t *is strictly increasing for* $t \leq t^*$ *and non-increasing for* $t \geq t^*$*. If* $\alpha < 1$*, then* $t^* < \infty$ *and* $N_{t^*} \leq \max\{n_r,\ \frac{n_r}{2(1-\alpha)} + n_b\}$*.*

(c) ***Bounded growth.*** *For every* t*,* $c_t \leq c_1 + \sum_{s \geq 1} \beta_s \leq 2c_1$*.*

Remark (interpretation). When $\alpha > \alpha^*$, the error first *rises* above its initial value $\mathrm{Tr}(\mathbf{\Sigma})/n_r$, then decreases for $\alpha < 1$, and never comes back down at $\alpha = 1$. Three observations follow. First, since $\alpha^* > \frac{1}{2}$, the initial rise occurs only where the rate of Theorem 3.6 is already degraded. Second, α^* depends only on n_r and n_b, and not on $\mathbf{\Sigma}$ or d, as c_t is independent of both. As $n_b/n_r \to \infty$, i.e., when each batch is much larger than the initial pool of real samples, we have $\alpha^* \to 1$: the real samples in the batch, $(1-\alpha)n_b$ of them in expectation, outnumber the initial pool unless α is close to 1, so the error rises only if almost the entire batch is generated. As $n_b/n_r \to 0$, i.e., when each batch is much smaller than the initial pool, we have $\alpha^* \to \frac{1}{2}$: the initial estimate is already accurate, so a mostly generated batch adds more error than it removes. Third, the rise is modest: by (c), the error at most doubles. For $n_r = 20$ and $n_b = 10$, $\alpha^* \approx 0.654$, and at $\alpha = 1$ the error rises to 1.74 times its initial value.

Proof. Throughout, $\alpha_t \equiv \alpha$ and $n_b(t) \equiv n_b$, so that $N_t = n_r + (t-1)n_b$ is increasing in t, and β_t, γ_t are given by (18).

Reduction to a scalar sequence. Taking traces in Theorem A.11(e) and using (11), we have $\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] = c_t \,\mathrm{Tr}(\mathbf{\Sigma})$. It therefore suffices to study the scalar sequence c_t defined in (40):

$$c_1 = \frac{1}{n_r}, \qquad c_{t+1} = \gamma_t c_t + \beta_t \quad (t \geq 1)$$

Step 1: important definitions and identities. We need to analyze the difference sequences:

$$\begin{aligned} c_{t+1} - c_t &= \gamma_t c_t + \beta_t - c_t \\ &= \beta_t - (1-\gamma_t)c_t \\ &= (1-\gamma_t)\left(\frac{\beta_t}{1-\gamma_t} - c_t\right) \\ &= (1-\gamma_t)(\theta_t - c_t). \end{aligned} \tag{145}$$

We defined

$$\theta_t := \frac{\beta_t}{1-\gamma_t} = \frac{1}{(1-\alpha)K_t + \alpha^2}, \qquad K_t = 2N_t + n_b(1+\alpha), \tag{146}$$

where we used (36). Using (145), we have

$$\begin{aligned} c_{t+1} - \theta_t &= (c_{t+1} - c_t) + (c_t - \theta_t) \\ &= -(1-\gamma_t)(c_t - \theta_t) + (c_t - \theta_t) \\ &= \gamma_t(c_t - \theta_t) \end{aligned} \tag{147}$$

Step 2: once the error stops increasing, it never increases again. Fix $t \geq 1$, and suppose that $c_{t+1} \leq c_t$. Then, from (145) and $\gamma_t < 1$ (Lemma A.9), we have $\theta_t - c_t \leq 0 \Rightarrow c_t \geq \theta_t$. Thus, from (147) and $\gamma_t > 0$ (Lemma A.9), we have

$$c_{t+1} - \theta_t \geq 0 \Rightarrow c_{t+1} \geq \theta_t. \tag{148}$$

From (146), θ_t is non-increasing in t, thus $\theta_t \geq \theta_{t+1}$, and from (148), we have $c_{t+1} \geq \theta_{t+1}$. Using (145), we get $c_{t+2} - c_{t+1} \leq 0 \Rightarrow c_{t+2} \leq c_{t+1}$. Hence, by induction $c_{s+1} \leq c_s$ for all $s \geq t$.

(a) Threshold. Using (145) and $\gamma_t < 1$ (Lemma A.9) at $t = 1$, $c_2 > c_1$ if and only if $c_1 < \theta_1$. With $c_1 = 1/n_r$, $N_1 = n_r$ and (146), this reads

$$\begin{aligned}\frac{1}{n_r} < \frac{1}{(1-\alpha)\left(2n_r + n_b(1+\alpha)\right) + \alpha^2} &\Leftrightarrow (1-\alpha)\left(2n_r + n_b(1+\alpha)\right) + \alpha^2 < n_r \\ &\Leftrightarrow \alpha^2 + n_b(1-\alpha^2) + 2n_r(1-\alpha) - n_r < 0 \\ &\Leftrightarrow (1-n_b)\alpha^2 - 2n_r\alpha + (n_r + n_b) < 0\end{aligned}$$

Define

$$g(\alpha) = (n_b - 1)\alpha^2 + 2n_r\alpha - (n_r + n_b). \tag{149}$$

Thus, $c_2 > c_1$ if and only if $g(\alpha) > 0$. On $[0, 1]$, g is strictly increasing, as

$$g'(\alpha) = 2(n_b - 1)\alpha + 2n_r > 0,$$

and

$$g(0) = -(n_r + n_b) < 0, \qquad g\left(\tfrac{1}{2}\right) = -\tfrac{3n_b+1}{4} < 0, \qquad g(1) = n_r - 1 \geq 0.$$

Hence g has a unique root $\alpha^* \in (\frac{1}{2}, 1]$, which solving the quadratic $g(\alpha) = 0$ shows to be given by (144):

$$\alpha^*(n_r, n_b) \coloneqq \begin{cases} \dfrac{\sqrt{n_r^2 + (n_b - 1)(n_r + n_b)} - n_r}{n_b - 1}, & n_b \geq 2, \\ \dfrac{n_r + 1}{2n_r}, & n_b = 1, \end{cases}$$

And $g(\alpha) > 0$ if and only if $\alpha > \alpha^*$. Finally, if $\alpha \leq \alpha^*$, then $c_2 \leq c_1$, and by Step 2 with $t = 1$, c_t is non-increasing for all t.

(b) Single peak. Let $t^* \coloneqq \min\{t \geq 1 :\ c_{t+1} \leq c_t\}$, with $t^* = \infty$ if no such t exists. By definition, $c_{t+1} > c_t$ for all $t < t^*$, so c_t is strictly increasing for $t \leq t^*$. If $t^* < \infty$, Step 2 with $t = t^*$ shows that c_t is non-increasing for $t \geq t^*$.

It remains to show that $t^* < \infty$ when $\alpha < 1$, and to bound N_{t^*}. Since $K_t \geq 2N_t$ and $\alpha^2 \geq 0$, (146) gives

$$\theta_t \leq \frac{1}{2(1-\alpha)N_t}.$$

Since $N_t \to \infty$, there is a first index t_1 with $N_{t_1} \geq n_r/\big(2(1-\alpha)\big)$, and then $\theta_{t_1} \leq 1/n_r = c_1$. We claim that $t^* \leq t_1$. Assume, by contradiction, that $t^* > t_1$. Then c_t is increasing up to t_1, so $c_{t_1} \geq c_1 \geq \theta_{t_1}$. By (145), $c_{t_1+1} - c_{t_1} = (1 - \gamma_{t_1})(\theta_{t_1} - c_{t_1}) \leq 0$, so t_1 belongs to the set defining t^*. Hence $t^* \leq t_1$, which is a contradiction. Therefore, $t^* \leq t_1$.

Since N_t is increasing, $N_{t^*} \leq N_{t_1}$. If $t_1 = 1$, then $N_{t_1} = n_r$. Otherwise, by the minimality of t_1, $N_{t_1-1} < n_r/\big(2(1-\alpha)\big)$, and hence $N_{t_1} = N_{t_1-1} + n_b < n_r/\big(2(1-\alpha)\big) + n_b$. In both cases,

$$N_{t^*} \leq \max\Big\{n_r,\ \frac{n_r}{2(1-\alpha)} + n_b\Big\}.$$

(c) Bounded growth. Since $\gamma_t < 1$ (Lemma A.9),

$$c_{t+1} = \gamma_t c_t + \beta_t \leq c_t + \beta_t.$$

Summing over $t \geq 1$ and using Lemma A.5, we get

$$c_t \leq c_1 + \sum_{s \geq 1} \beta_s \leq \frac{1}{n_r} + \frac{1}{n_r} = 2c_1.$$

□

B THE VALUE OF IDENTIFYING REAL SAMPLES

We quantify the gain from perfect detection of synthetic samples under the setting of Theorem 3.6, i.e., $\alpha_t \equiv \alpha$ and $n_b(t) \equiv n_b$. The pool contains, in expectation, $N_{\text{oracle}} \coloneqq (1-\alpha)N_t + \alpha n_r$ real samples, which is approximately the sample size available to an oracle that identifies and keeps only them. On the other hand, the sample mean of N i.i.d. real samples has an error of $\operatorname{Tr}(\boldsymbol{\Sigma})/N$, so the pooled estimator has the effective sample size $N_{\text{eff}} \coloneqq \operatorname{Tr}(\boldsymbol{\Sigma})/\mathbb{E}\big[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2\big] = \operatorname{Tr}(\boldsymbol{\Sigma})/\operatorname{Tr}(\boldsymbol{C}_t)$ (by (11)), i.e., the number of i.i.d. real samples that attain the same error. The gain from perfect detection is the ratio

$$G_\alpha(N_t) \coloneqq \frac{N_{\text{oracle}}}{N_{\text{eff}}} = \big((1-\alpha)N_t + \alpha n_r\big)\,\frac{\operatorname{Tr}(\boldsymbol{C}_t)}{\operatorname{Tr}(\boldsymbol{\Sigma})}. \tag{150}$$

As $N_t \to \infty$, Theorem 3.6 gives

$$G_\alpha(N_t) \asymp \begin{cases} 1, & 0 \le \alpha < \frac{1}{2}, \\ \ln N_t, & \alpha = \frac{1}{2}, \\ N_t^{2\alpha-1}, & \frac{1}{2} < \alpha < 1. \end{cases} \tag{151}$$

In words, when generated samples are a minority of each new batch, perfect detection improves the effective sample size by at most a constant factor, no matter how much data is collected. When they are a majority, the improvement grows without bound as more data is collected. Moreover, the gain grows with α in the long run: for any $0 \le \alpha < \alpha' < 1$ with $\alpha' > \frac{1}{2}$, we have $G_{\alpha'}(N_t)/G_\alpha(N_t) \to \infty$ as $N_t \to \infty$. Thus, asymptotically, the gain from detection is largest exactly where the pooled estimator converges most slowly.

At a fixed, finite N_t, however, the gain does not increase all the way to $\alpha = 1$. At the two extremes,

$$G_0(N_t) = 1, \qquad 1 \le G_1(N_t) = n_r\,\frac{\operatorname{Tr}(\boldsymbol{C}_t)}{\operatorname{Tr}(\boldsymbol{\Sigma})} \le 2, \tag{152}$$

where the bounds follow from Proposition 3.5 and Lemma 3.2. At $\alpha = 0$ there is nothing to detect, and at $\alpha = 1$ the oracle retains only the n_r initial real samples. For fixed t, $G_\alpha(N_t)$ is continuous in α, and by (151), for any fixed $\alpha \in (\frac{1}{2}, 1)$ it exceeds 2 once N_t is large enough. Hence, for large but finite N_t, the gain is maximized at some α strictly between 0 and 1. In practice, detection is therefore worth most when generated data dominates, but real data is still being added in meaningful amounts. The potential for such detection to improve performance is supported empirically by prior work on identifying and curating model-generated data (Drayson et al., 2025).

C ADDITIONAL DETAILS ON EMPIRICAL VERIFICATION AND DISCUSSION

Under the same setup as Section 3.3, we present a numerical verification of the theoretical recursion of $\operatorname{Tr}(\boldsymbol{C}_t) = \mathbb{E}[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2]$ (Equation (11) and Theorem A.11).

Exact mean-square error. The recursion of Theorem A.11 gives the mean-square error of $\boldsymbol{\mu}^{(t)}$ exactly, without sampling; Figure 2(a) and (b) use these exact values. Figure 3 compares them with the Monte-Carlo averages of Section 3.3: in all 19 regimes and across the whole range of N_t, the two differ by at most 3.4 standard errors, i.e., only by Monte-Carlo noise.

Finally, we verify numerically that $\boldsymbol{\mu}^{(t)}$ is unbiased and that the bias of $\boldsymbol{\Sigma}^{(t)}$ is exactly $-\boldsymbol{C}_t$ (Theorem 3.1). Both checks use $T = 400$ iterations, so that N_t ranges up to about 4000.

Unbiasedness of the mean. Theorem 3.1(a) states that $\mathbb{E}[\boldsymbol{\mu}^{(t)}] = \boldsymbol{\mu}$ for every $t \ge 1$. Figure 4 illustrates this with $M = 50{,}000$ independent runs of $T = 400$ iterations. In all four regimes, the average of each coordinate lies on its true value at every N_t. The shaded bands show that this is not because single runs are accurate: they scatter widely, and at $\alpha = 0.9$ and $\alpha = 1$ the band barely narrows, since the estimator converges slowly or not at all. At $\alpha = 1$, in particular, $\boldsymbol{\mu}^{(t)}$ is unbiased without being consistent, in line with Proposition 3.5. To quantify the agreement beyond what the figure can resolve, we compare the Monte-Carlo average with the true value in units of its standard error: over all 19 regimes, all 400 iterations and all four coordinates, the difference never exceeds 2.81 standard errors, as expected from Monte-Carlo noise alone.

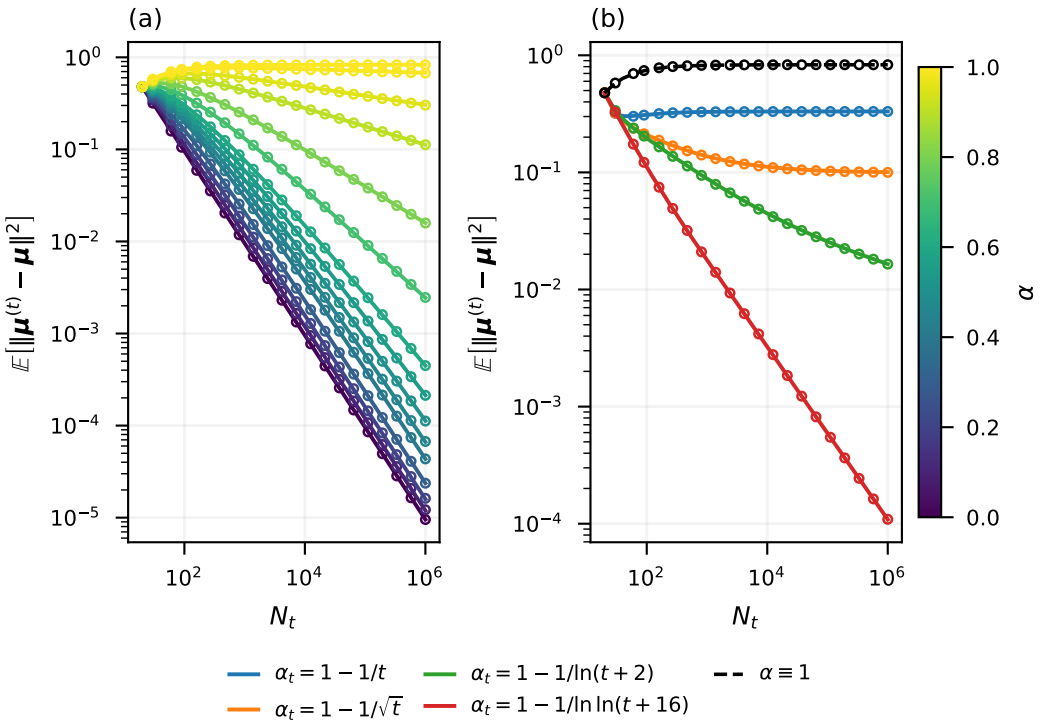

Figure 3: Exact $\mathbb{E}[\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}\|_2^2]$ against simulation. (a) Constant α (color bar) and (b) the four time-varying sequences of Figure 1, with $\alpha \equiv 1$ (dashed). Lines are the recursion of Theorem A.11, and circles are Monte-Carlo averages.

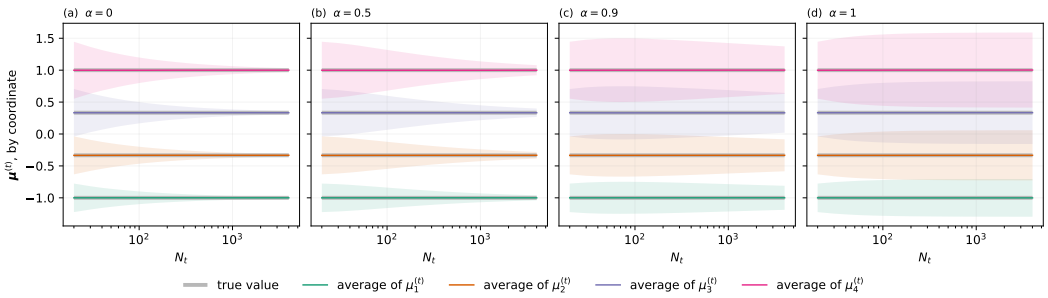

Figure 4: Unbiasedness of $\boldsymbol{\mu}^{(t)}$ for four constant values of α. For each coordinate i, the colored line is the Monte-Carlo average of $\mu_i^{(t)}$, the wide gray line underneath is the true value μ_i, and the shaded band is ± 1 standard deviation of a single run's $\mu_i^{(t)}$.

Bias of the covariance. Unlike the mean, $\boldsymbol{\Sigma}^{(t)}$ is biased: by Theorem 3.1(b), $\mathbb{E}[\boldsymbol{\Sigma}^{(t)}] - \boldsymbol{\Sigma} = -\boldsymbol{C}_t$, so the covariance is underestimated by exactly the covariance of the mean estimator. Figure 5 checks this for the trace. In all four regimes, the measured bias lies on the predicted $\mathrm{Tr}(\boldsymbol{C}_t)$ over the whole range: it decays like N_t^{-1} at $\alpha = 0$, more slowly at $\alpha = 0.5$ and $\alpha = 0.9$, and at $\alpha = 1$ it grows and then levels off at a positive value, so the covariance estimator remains biased forever (Proposition 3.5). Across all 19 regimes and all 400 iterations, the measured bias differs from the prediction by at most 3.19 standard errors, as expected from Monte-Carlo noise alone.

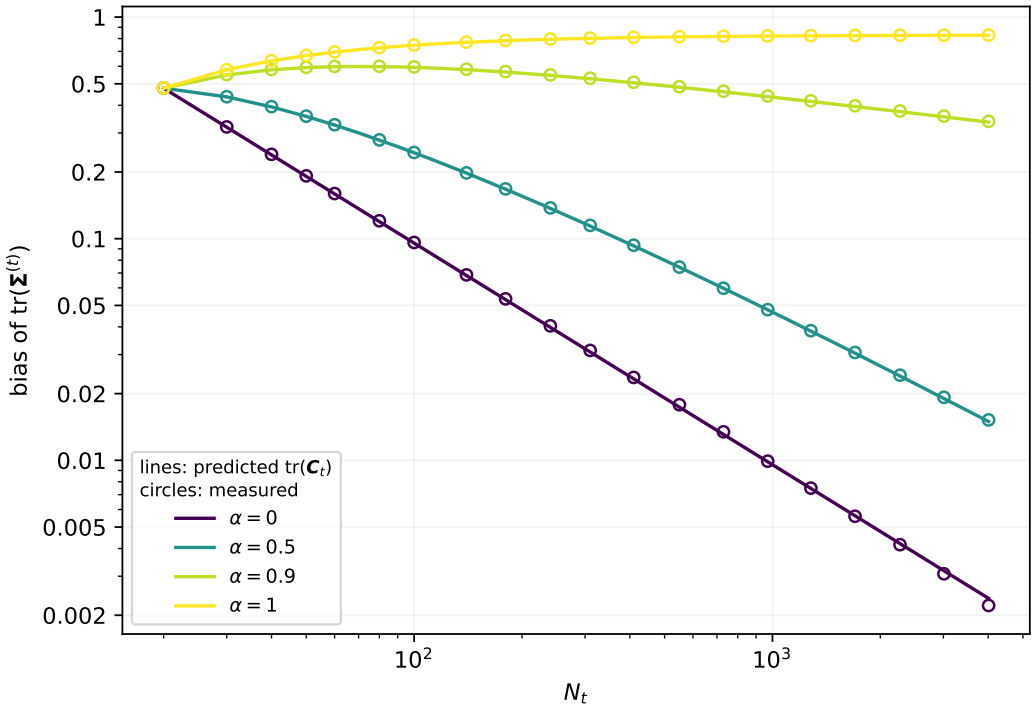

Figure 5: The bias of $\mathbf{\Sigma}^{(t)}$ for four constant values of α. Lines are the predicted downward bias of the trace, $\mathrm{Tr}(\boldsymbol{C}_t)$, computed from the recursion of Theorem A.11; circles are the measured $\mathrm{Tr}(\mathbf{\Sigma}) - \mathrm{Tr}(\widehat{\mathbb{E}}[\mathbf{\Sigma}^{(t)}])$, averaged over 2×10^6 Monte-Carlo runs.

D Experiments with Diffusion Models

While our paper focuses on theoretical contributions, in this section we empirically examine the extent to which the behavior observed in a practical deep learning setting is consistent with our theoretical results.

Data and protocol. We consider CIFAR-10 (Krizhevsky et al., 2009) and a diffusion model (Sohl-Dickstein et al., 2015; Song et al., 2021b) based on the DDPM framework (Ho et al., 2020), using the U-Net implementation of Zou (`https://github.com/zoubohao/DenoisingDiffusionProbabilityModel-ddpm-`), and train unconditional models. We split the 50,000 CIFAR-10 training images into a class-balanced validation set of 5,000 images and 45,000 images available for training; the 10,000 test images are used only for evaluation. We set $n_r = n_b = 5{,}000$ and $t_{\max} = 9$, so that the pool at iteration t contains $N_t = n_r + (t-1)n_b$ images and the final pool, $N_9 = 45{,}000$, uses all available training images. We consider four fixed mixing fractions, $\alpha \in \{0, 0.25, 0.75, 1\}$. At $t = 1$ we train a DDPM on n_r class-balanced real images; this model is shared by all four values of α. At each iteration $t \geq 2$, we accumulate the previous pool with αn_b samples generated by the model of iteration $t-1$ and $(1-\alpha)n_b$ previously unused, class-balanced real images, and the training treats real and generated images identically: minibatches are drawn uniformly from the pool, with no labels, reweighting or filtering.

Unlike the theoretical setup, where each sample is drawn from the mixture independently, here the numbers of generated and real images in each batch are fixed. This keeps the consumption of real images exact and class-balanced, and differs negligibly from independent draws, whose count of generated images fluctuates by less than 1% of the batch (one standard deviation) for $n_b = 5{,}000$.

Training. All models share one recipe: 1,000 diffusion steps with a linear noise schedule from 10^{-4} to 0.02, the ϵ-prediction objective, Adam with learning rate 2×10^{-4} after a 5,000-step linear warm-up, batch size 128, gradient clipping at 1, dropout 0.1, and an exponential moving average

(EMA) of the weights with decay 0.9999. All sampling uses the EMA weights and DDIM (Song et al., 2021a) with 200 steps. We fix the *number of optimization steps* rather than the number of epochs, so that pool size is not confounded with the amount of computation. In each iteration t the training is from scratch on the whole pool, except for $\alpha = 0.75$ and $t \geq 4$, where we use warm-start from the previous iteration's model with learning rate of lower order.

Checkpoint selection. Since diffusion models overfit on datasets of this size, each model should be trained neither too little nor too long, so that pools of different sizes are compared fairly. We save a checkpoint every $50{,}000$ steps and select the one minimizing the FID between $10{,}000$ generated samples and the validation set; we report the FID of the same samples against the test set, which is never used for selection. The default budget is $200{,}000$ steps; iterations in which the validation FID was still decreasing at the end of the budget were retrained with a larger budget (up to $800{,}000$ steps), which was needed only for $\alpha \in \{0, 0.25\}$. Validation and test FIDs selected the same checkpoint in every iteration, and over all checkpoints they differed by a nearly constant offset of 2.44 ± 0.07.

FID values. We compute FID with clean-fid (Parmar et al., 2022) from $10{,}000$ DDIM samples against the held-out test set. Published DDPM results on CIFAR-10 (e.g., 3.17 in Ho et al. (2020)) use $50{,}000$ samples, $1{,}000$-step sampling and training-set statistics, which yields substantially lower values: for our best model ($\alpha = 0$, $t = 9$, trained on $45{,}000$ images), $50{,}000$ samples against the training set give 5.1, compared with 9.7 on the test set. Our FIDs are therefore comparable across the curves of Figure 6, but not to published numbers.

Reporting. We report the test FID of the selected checkpoint as a function of N_t for each α, for a single training seed. The horizontal line is the $\alpha = 0$ model at $t = 9$, i.e., a DDPM trained on all $45{,}000$ training images with the same protocol. From duplicate runs, differences below ≈ 0.5 FID between consecutive iterations are within run-to-run variation. We expect the curves for $\alpha < 1$ to approach this reference value, whereas for $\alpha = 1$, where no new real data are introduced, we expect the FID to approach a finite limiting value.

All experiments were run on a server with eight NVIDIA RTX 6000 Ada GPUs, training one model per GPU. Training costs about 3.9 GPU-hours per $100{,}000$ steps including checkpoint evaluation, e.g., about 8, 15 and 31 hours for budgets of $200{,}000$, $400{,}000$ and $800{,}000$ steps; the runs shown in Figure 6 total roughly 400 GPU-hours.

Comparison with the theory. The central prediction of our theory holds: although real and generated images are pooled and treated identically, with no detection, labeling or reweighting, every $\alpha < 1$ improves on the initial model, whereas $\alpha = 1$ degrades. At every iteration $t \geq 2$, the FID for $\alpha \in \{0, 0.25, 0.75\}$ is below its initial value of 33.8, while for $\alpha = 1$ it never falls below it. Thus, a fixed fraction of fresh real data suffices to avoid collapse even when it cannot be told apart from generated data. This comparison is meaningful beyond the Gaussian setting because FID is the squared 2-Wasserstein distance between Gaussians fitted to the Inception features of generated and test images, i.e., the quantity of Theorem 3.3 evaluated in a feature space. The finer predictions carry over as well. First, the curves are ordered by α at every iteration $t \geq 2$: a larger fraction of generated data always yields a worse model. Second, the improvement slows as α grows, in line with Theorem 3.6: from $N_1 = 5{,}000$ to $N_3 = 15{,}000$, the FID decreases by 19.9 for $\alpha = 0$, 16.7 for $\alpha = 0.25$ and only 6.8 for $\alpha = 0.75$, the one value above $\frac{1}{2}$, whose predicted rate $N_t^{-1/2}$ is slower than the rate N_t^{-1} of $\alpha < \frac{1}{2}$. For $\alpha = 0.25$, the curve follows that of $\alpha = 0$ with a roughly constant offset of 3–4 FID, consistent with synthetic data inflating the constant but not the exponent. Third, at $\alpha = 1$ the FID rises from 33.8 to 39.7, with increments that shrink over the last iterations, matching the bounded, non-decreasing error of Propositions 3.5 and A.27.

The value of identifying real samples (Section 3.2.3) is also visible. Discarding all generated images would leave, at $t = 9$, $35{,}000$ real images for $\alpha = 0.25$ and $15{,}000$ for $\alpha = 0.75$, i.e., the $\alpha = 0$ model at $N_t = 35{,}000$ (FID 10.3) and at $N_t = 15{,}000$ (FID 13.8). Compared with FIDs of 13.5 and above 27, respectively, perfect detection would gain about 3 FID for $\alpha = 0.25 < \frac{1}{2}$, but more than 13 for $\alpha = 0.75 > \frac{1}{2}$, as the theory predicts.

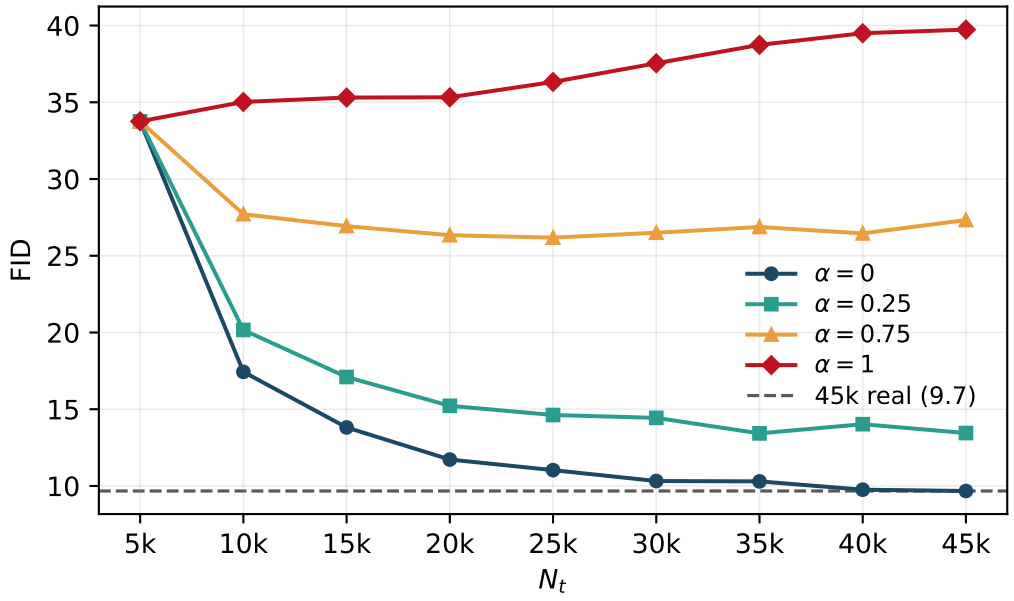

Figure 6: Test FID of DDPM on CIFAR-10 as a function of the pool size N_t, for mixing fractions $\alpha \in \{0, 0.25, 0.75, 1\}$ (single seed). At each iteration, the checkpoint with the lowest validation FID is selected and its test FID is reported. The dashed line is the $\alpha = 0$ model at $t = 9$, trained on all $45{,}000$ training images (FID 9.7 on the test set, and 5.1 with $50{,}000$ samples against the training set; see text). Differences below ≈ 0.5 FID between consecutive iterations are within run-to-run variation.