How Recursive Training Collapses and What Can Be Done About It

Anonymous Authors[1]

Abstract

When generative models are trained on data produced by earlier generative models, the quality of their outputs systematically degrades, a phenomenon known as *model collapse*. But what exactly drives this collapse, how fast does it proceed, and can it be stopped? We show that two distinct mechanisms, drift in the estimated mean and contraction in the estimated variance, contribute *equally* to the total information loss, resolving an ongoing debate in the literature. When variance is estimated from data (as all practical models do), the rate of degradation *doubles* compared to the idealized case of known variance, making variance contraction a particularly dangerous "silent killer" that operates without obvious distributional shift. We then establish that each generation of recursive training acts as an optimal lossy compression step, and prove that no estimator can fundamentally outperform this rate: collapse is unavoidable without external intervention. For the common intervention strategy of mixing real data into each generation, we compare a constant vs a decaying mixing schedule and prove that gradually reducing the mixing fraction inevitably fails in the long run. Moreover, we show that under the Gaussian model, when total real data exposure is matched, constant mixing always outperforms decaying mixing schedules. Experiments on variational autoencoders, diffusion models, and GPT-2 validate the theoretical predictions.

1. Introduction

The internet is increasingly filled with content generated by AI systems. As this AI-generated content accumulates, future models will inevitably train, at least in part, on data produced by their predecessors. This creates a closed feedback loop: models generate data, that data trains the next round of models, and the cycle repeats. The critical question is not whether this recursive training happens, but whether it is sustainable.

The emerging answer is troubling. Recursive generative training leads to *model collapse*: a systematic degradation in both the quality and diversity of generated outputs. Shumailov et al. (Shumailov et al., 2024) demonstrated that language models trained on their own outputs progressively lose the ability to generate low-probability tokens: the tails of the distribution vanish. Alemohammad et al. (Alemohammad et al., 2024) documented the same degradation across image generation tasks. The symptoms are remarkably consistent across modalities: shrinking diversity, homogenized outputs, and eventual convergence to degenerate equilibria where all generated samples look the same.

Yet despite the urgency and the growing body of empirical evidence, a fundamental question has remained unanswered: *what is the dominant mechanism driving collapse?* The literature presents a seemingly contradictory picture. On one side, Shumailov et al. (Shumailov et al., 2024) emphasized the disappearance of low-probability events, and Suresh et al. (Suresh et al., 2025) formally derived the exponential variance contraction that occurs under recursion, pointing to *variance loss* as the primary culprit. On the other side, Schaeffer et al. (Schaeffer et al., 2025) argued that in Gaussian models, *mean drift* which is the gradual shift in the distribution average, is the true driver. Existing mitigation strategies such as data mixing (Fu et al., 2024; Gerstgrasser et al., 2024) and reweighting (He et al., 2025) address the general phenomenon without knowing exactly which mechanism to target and how these strategies interact with the underlying causes.

We resolve this debate through a precise decomposition of the information loss. The answer is surprising: *neither mechanism dominates, they contribute equally*. Especially compared to an idealized known-variance setting, when variance is *estimated* alongside the mean, as all practical models do, the two sources of degradation each account for exactly half of the total information loss. This doubling of the collapse rate compared to the idealized known-variance setting

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

makes variance contraction a particularly insidious problem: it operates silently, reducing diversity without producing an obvious shift in the mean that standard monitoring would detect. The practical implication is that strategies targeting only one mechanism address only half the problem.

This equal contribution is not a coincidence. We show that recursive training is fundamentally *iterative lossy compression*: each generation compresses information from the previous one, and this compression is optimal in the information-theoretic sense. Moreover, we prove this rate is minimax optimal: no estimator can fundamentally do better. Collapse is therefore unavoidable without some form of external intervention, such as mixing in fresh real data.

The most natural mitigation is to mix a fraction α of real data into each training generation. We derive the exact steady-state quality level that constant mixing achieves, and prove a negative result: if the mixing fraction α gradually decreases to zero, quality inevitably degrades without bound. Crucially, we also show that under the Gaussian model, when total real data exposure $\Sigma\alpha$ is held equal, constant mixing always outperforms decaying mixing schedules, because uniform contraction across all generations is more effective than front-loaded contraction followed by progressively weaker correction.

We validate all predictions through experiments spanning three model families (variational autoencoders, diffusion models, and GPT-2) across two modalities (images and text). Beyond confirming the theory, these experiments demonstrate that the apparent benefit of decaying schedules over constant mixing in raw FID comparisons is largely explained by higher total real data exposure ($\Sigma\alpha$), not by any advantage of the schedule shape itself.

Key contributions.

1. We prove that mean drift and variance contraction contribute equally to information loss, doubling the collapse rate when variance is estimated, resolving the apparent contradiction in the literature.
2. We establish that recursive training is iterative lossy compression at the information-theoretic bound, with minimax-optimal collapse rate.
3. We prove that gradually reducing the mixing fraction inevitably fails in the long run, and demonstrate via Gaussian simulation that constant mixing outperforms decaying schedules when total real data exposure is matched.

2. Problem Formulation

We now formalize the recursive training setting and introduce the key quantities that will appear throughout our analysis.

Definition 2.1 (Recursive Generative Training)**.** Let P_0 be the true data distribution over $\mathcal{X} \subseteq \mathbb{R}^d$, where d is the dimension of the distribution. A *recursive generative training* process with sample size n produces a sequence of distributions $\{P_t\}_{t\geq 0}$: at each generation t, draw n i.i.d. samples from P_t, fit $\hat{P}_{t+1}$ via maximum likelihood estimation (MLE), and set $P_{t+1} = \hat{P}_{t+1}$.

Definition 2.2 (Mixed Recursive Training)**.** With mixing fraction $\alpha \in [0, 1]$, at generation t draw $n\alpha$ samples from P_0 (real data) and $n(1-\alpha)$ from P_t (synthetic data), then fit $\hat{P}_{t+1}$ to the combined dataset. When α_t varies with t, we call this a *decaying mixing schedule*. The *total real data exposure* is $\Sigma\alpha = \sum_{s=1}^{T} \alpha_s$, which quantifies the cumulative amount of real data the model sees over T generations.

We study the Gaussian model $P_0 = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ as a tractable setting that captures the essential dynamics. This is the standard starting point for analyzing recursive training (Suresh et al., 2025; Kanabar and Gastpar, 2025). The Markov structure of recursive training implies that the mutual information $I(P_0; P_t)$ is non-increasing by the data processing inequality: information about the original distribution can only be lost, never gained, across generations.

The crucial distinction is between *known* and *unknown* variance. In all practical models, variance must be estimated from the same finite dataset used to estimate the mean. The MLE for variance, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, is unbiased on a single generation, but under recursion it compounds into exponential contraction: $\hat{\sigma}_t^2 = \sigma_0^2 \cdot ((n-1)/n)^t$. This contraction is invisible if one tracks only the mean, yet, as we show next, it accounts for half the total information loss.

3. Variance Collapse Theory

3.1. Known Variance: The Baseline

When the covariance is known and only the mean is estimated, collapse proceeds at a well-understood rate.

Theorem 3.1 (Gaussian Mean Drift)**.** *For $P_0 = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with known covariance $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{d\times d}$, where d is the dimension of the distribution, under recursive MLE with n samples per generation:*

$$\boldsymbol{\mu}_t - \boldsymbol{\mu}_0 \sim \mathcal{N}\left(\mathbf{0}, \frac{t\boldsymbol{\Sigma}_0}{n}\right), \quad \mathbb{E}[D_{\mathrm{KL}}(P_t \| P_0)] = \frac{td}{2n} \quad (1)$$

Proof sketch. The MLE $\hat{\boldsymbol{\mu}}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_0/n)$ yields a random walk with step covariance $\boldsymbol{\Sigma}_0/n$. Unrolling gives $\boldsymbol{\mu}_t - \boldsymbol{\mu}_0 \sim \mathcal{N}(\mathbf{0}, t\boldsymbol{\Sigma}_0/n)$. The KL divergence between equal-covariance Gaussians reduces to $td/(2n)$ via the trace trick: $\mathbb{E}[D_{\mathrm{KL}}] = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1} \cdot t\boldsymbol{\Sigma}_0/n) = \frac{t}{2n}\mathrm{tr}(\mathbf{I}_d) = td/(2n)$. Full proof in Appendix A. □

The key insight is that each generation adds a fixed amount

of noise to the mean estimate, producing a random walk. The KL divergence grows linearly with the number of generations t, linearly with the dimension d, and inversely with the sample size n. The dimension d enters because each of the d independent components drifts at rate $1/(2n)$, and the total KL is the sum over all dimensions.

3.2. Unknown Variance: The Silent Killer

In practice, variance must be estimated alongside the mean. This seemingly minor change has a dramatic effect: it *doubles* the rate of collapse.

Theorem 3.2 (Variance Contraction Doubles Collapse Rate)**.** *For a d-dimensional Gaussian with diagonal covariance and unknown parameters, using MLE estimators $\hat{\mu}_t = \bar{X}_t$ and $\hat{\sigma}_t^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^{(t)} - \bar{X}_t)^2$, for $n > 3$:*

1. *The variance contracts exponentially: $\mathbb{E}[\hat{\sigma}_t^2] = \sigma_0^2 \cdot ((n-1)/n)^t$*
2. *The KL divergence decomposes as: $\mathbb{E}[D_{\mathrm{KL}}(P_t\|P_0)] = \underbrace{\frac{td}{2n}}_{\text{mean drift}} + \underbrace{\frac{td}{2n}}_{\text{variance mismatch}} + o(t/n)$*
3. *The total rate $\approx td/n$ is **twice** the known-variance baseline*

Proof sketch. **Part 1.** By Basu's theorem, $\hat{\sigma}_t^2$ and $\hat{\mu}_t$ are independent given $\hat{\sigma}_{t-1}^2$. Since $\sum(X_i - \bar{X})^2/\hat{\sigma}_{t-1}^2 \sim \chi^2_{n-1}$, iterating gives the exponential contraction factor $(n-1)/n$ per generation.

Part 2. The KL divergence between the estimated and true distributions decomposes as:

$$D_{\mathrm{KL}}(P_t\|P_0) = \underbrace{\frac{(\hat{\mu}_t - \mu_0)^2}{2\sigma_0^2}}_{\text{mean drift}} + \underbrace{\frac{1}{2}\left[\ln\frac{\sigma_0^2}{\hat{\sigma}_t^2} + \frac{\hat{\sigma}_t^2}{\sigma_0^2} - 1\right]}_{\text{variance mismatch}} \quad (2)$$

Writing $V_t = \hat{\sigma}_t^2/\sigma_0^2$, both contributions evaluate to $\approx td/(2n)$ using $\mathbb{E}[V_t] \approx 1 - t/n$ and $\mathbb{E}[1/V_t] \approx 1 + 3t/n$. Full proof in Appendix B. □

This result resolves the debate in the literature. Under Wasserstein distance, mean drift dominates (as Schaeffer et al. (Schaeffer et al., 2025) observe); under KL divergence, the two contribute *equally*. Strategies that address only one mechanism tackle only half the problem. Variance contraction deserves particular attention not because it dominates, but because it is the *silent* killer: it reduces diversity without producing a detectable shift in the mean, and, as we show experimentally, it invalidates standard training metrics as quality indicators.

4. Mixing Strategies and Steady-State Analysis

Since collapse is unavoidable in pure recursive training, the most natural intervention is to mix real data into each generation. We analyze both constant and decaying mixing strategies.

4.1. Steady-State under Constant Mixing

Theorem 4.1 (Steady-State Mixing)**.** *Under mixed recursive training with constant $\alpha > 0$, the mean converges to a steady-state distribution with:*

$$\mathrm{Var}[\mu_\infty] = \frac{\sigma^2}{n\alpha(2-\alpha)} \quad (3)$$

For small α: $\mathrm{Var}[\mu_\infty] \approx \sigma^2/(2n\alpha)$.

Proof sketch. The MLE from mixed data gives $\hat{\mu}_t = \alpha\bar{X}_{\text{real}} + (1-\alpha)\bar{X}_{\text{synth}}$, with $\mathbb{E}[\hat{\mu}_t|\mu_{t-1}] = \alpha\mu_0 + (1-\alpha)\mu_{t-1}$ and $\mathrm{Var}[\hat{\mu}_t|\mu_{t-1}] = \sigma^2/n$. The law of total variance yields the recursion $v_t = (1-\alpha)^2 v_{t-1} + \sigma^2/n$, which converges to Eq. (3). Full proof in Appendix C. □

The practical implication is encouraging: even a small amount of real data dramatically improves steady-state quality. For example, mixing in just one percent real data reduces the steady-state variance by a factor of approximately fifty compared to pure synthetic training. The improvement scales inversely with the mixing fraction: doubling the real data fraction approximately halves the variance.

4.2. Decaying Mixing Inevitably Fails

Maintaining constant mixing requires ongoing access to real data and fresh real data is expensive to collect, annotate, or license. It may be more feasbile to front-load data for initial alignment, then reduce α to scale cheaply with synthetic data. A natural question therefore is whether the mixing fraction can *decrease* over time while still preventing collapse. The answer is no.

Corollary 4.2 (Decaying Mixing Inevitably Fails)**.** *If $\alpha_t = t^{-p}$ for* any *$p > 0$, then* $\mathrm{Var}[\mu_t] \to \infty$. *Specifically,* $\mathrm{Var}[\mu_t] \sim (\sigma^2/2n) \cdot t^{\min(p,1)}$.

Proof sketch. The variance recursion becomes $v_t \approx (1 - 2t^{-p})v_{t-1} + \sigma^2/n$. The fundamental obstacle is that the per-generation noise σ^2/n remains constant, while the contraction force $(1-\alpha_t)^2 \to 1$ vanishes. For $0 < p < 1$: $v_T \sim (\sigma^2/2n)T^p$; for $p \geq 1$: $v_T \sim C \cdot T\sigma^2/n$. Full proof in Appendix D. □

Remark 4.3 (Almost-sure convergence versus bounded variance). Although $\mathrm{Var}[\mu_t] \to \infty$ for all $p > 0$, the mean μ_t converges almost surely to μ_0 for $p > 1/2$ by the Robbins-Monro theorem. The mean converges, but distributional

quality degrades: the noise is constant (σ^2/n), not scaling with α_t^2 as in standard stochastic approximation.

4.3. Constant Mixing Outperforms Decaying at Matched Budget

The raw FID comparison between decaying and constant schedules can be misleading. A decaying schedule such as $\alpha_t = 0.5/t^{0.3}$ accumulates a large total real data exposure $\Sigma\alpha = \sum_{t=1}^{8} 0.5/t^{0.3} \approx 2.74$ over eight generations, while a constant schedule with $\alpha = 0.10$ provides only $\Sigma\alpha = 0.80$. The decaying schedule sees more total real data, and it is this larger budget, not the schedule shape, that drives the lower FID.

To make a fair comparison, we match $\Sigma\alpha$ between schedules. For each decaying schedule with total exposure $\Sigma\alpha$, we compare against a constant schedule with $\alpha_{\text{const}} = \Sigma\alpha/T$ (which provides the same total real data over T generations). Under the Gaussian model, the steady-state variance under constant mixing is $v_\infty = \sigma^2/(n\alpha(2-\alpha))$, while the variance under decaying mixing grows without bound. Even at finite generations, constant mixing achieves lower variance because it applies uniform contraction at every generation, whereas decaying schedules front-load their contraction and leave later generations increasingly vulnerable to noise accumulation.

Table 1 shows that, under the Gaussian model, constant mixing always achieves lower variance than decaying mixing when $\Sigma\alpha$ is held equal. The reason is fundamental: the variance recursion $v_t = (1-\alpha_t)^2 v_{t-1} + \sigma^2/n$ requires contraction at *every* step to combat the constant noise injection σ^2/n. A constant schedule applies the same contraction force uniformly, while a decaying schedule provides strong contraction early but progressively weaker contraction late, allowing noise to accumulate unchecked in later generations.

5. Rate-Distortion and Minimax Optimality

5.1. Recursive Training as Iterative Lossy Compression

The equal-contribution result of Theorem 3.2 raises a deeper question: is the collapse rate a property of the particular estimator (MLE), or is it fundamental to recursive training itself?

Theorem 5.1 (Rate-Distortion Collapse)**.** *Each generation of recursive training is a lossy compression step operating at the Gaussian rate-distortion bound: rate $R_0 = \frac{1}{2}\log_2 n$ bits, distortion $D_0 = \sigma^2/n$, and critical generations $t^* = n$ (cumulative distortion equals signal variance).*

Proof sketch. Placing $\mu_0 \sim \mathcal{N}(0, \sigma_\mu^2)$, the mutual information $I(\mu_0; \hat{\mu}) \to \frac{1}{2}\log_2 n$ as $\sigma_\mu^2/\sigma^2 \to \infty$. The Gaussian R-D function gives $D(R) = \sigma^2 \cdot 2^{-2R}$; substituting $R = \frac{1}{2}\log_2 n$ yields $D = \sigma^2/n$. Setting cumulative distortion $t\sigma^2/n = \sigma^2$ gives $t^* = n$. Full proof in Appendix E. □

This result provides a powerful reframing: recursive training is not merely "accumulating errors," each generation performs *optimal* lossy compression. The irreversibility of collapse is a direct consequence of Shannon's source coding theorem. With n samples, approximately n generations exhaust the information content.

5.2. Minimax Optimality

Theorem 5.2 (Minimax Tightness)**.** *For any estimator sequence $\{\hat{P}_t\}$ trained recursively on n samples per generation:*

$$\inf_{\hat{P}_t} \sup_{P_0} \mathbb{E}[D_{\text{KL}}(P_0 \| \hat{P}_t)] \geq c \cdot \frac{td}{2n} \tag{4}$$

where $c = (1 - 1/\sqrt{2})/4 \approx 0.073$. The MLE rate is minimax optimal up to a constant factor.

Proof sketch. Via Le Cam's two-point method. After t generations, $\hat{\boldsymbol{\mu}}_t \sim \mathcal{N}(\boldsymbol{\mu}_0, t\boldsymbol{\Sigma}_0/n)$, equivalent to a single observation with effective sample size n/t. Two hypotheses separated by ε along one coordinate, with $\varepsilon^2 = 2t\sigma^2/n$ to set $\text{KL} = 1$, give a binary testing lower bound. Converting to KL via $\mathbb{E}[D_{\text{KL}}] = \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2]/(2\sigma^2)$ and summing over d coordinates yields the result. Full proof in Appendix F. □

The doubling from Theorem 3.2 is not an artifact of MLE; it is fundamental to recursive training with finite samples. No estimator can escape collapse at a rate better than $\Theta(td/n)$. The only escape is external intervention: mixing with real data or accumulating data across generations (Gerstgrasser et al., 2024).

Table 1. Constant vs. decaying mixing at matched $\Sigma\alpha$ (Gaussian, n=1000, T=8). Variance is reported in units of σ^2/n. The constant schedule always achieves lower variance.

Decaying schedule	$\Sigma\alpha$	α_{const}	V_{const}	V_{decay}
$0.5/t^{0.3}$	2.743	0.343	1.76	2.07
$0.5/\sqrt{t}$	2.186	0.273	2.12	2.74
$0.3/t^{0.3}$	1.646	0.206	2.71	3.02
$0.5/t$	1.359	0.170	3.22	4.61

6. Experiments

We validate all theoretical predictions through experiments spanning three model families and two modalities (Table 2).

Table 2. Experimental setup. All experiments use recursive training.

Model	Dataset	Gens	Conditions	Metric
ConvVAE	MNIST	15	known/unk σ^2	FID, Var
ConvVAE	CIFAR-10	10	$\alpha \in [0, 0.5]$	FID
DDPM	CIFAR-10	8	6 schedules	FID
GPT-2	Text	8	$\alpha \in [0, 0.2]$	PPL, Ent

6.1. Variance Contraction and KL Doubling

Setup. ConvVAE on MNIST, 15 generations, comparing known versus unknown variance estimation.

Results. Figure 1 is consistent with Theorem 3.2. Unknown-variance mode collapses faster: FID plateaus (approximately 52.3) by generation 9 versus generation 11 for known variance. The ratio of KL divergences (unknown divided by known) peaks at approximately $2.7\times$ at generation 3, with an average of approximately $2.0\times$ over the active collapse regime.

6.2. Mixing Strategies and the Role of $\Sigma\alpha$

Setup. ConvVAE on CIFAR-10, 10 generations, with mixing fractions $\alpha \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5\}$.

Three regimes. Figure 2(a) reveals three distinct behaviors: (i) unchecked collapse for very small mixing ($\alpha \leq 0.01$); (ii) stabilized degradation for moderate mixing ($\alpha = 0.05$–0.1), consistent with Theorem 4.1; and (iii) near-original quality for large mixing ($\alpha \geq 0.2$). The steady-state FID scales inversely with α, qualitatively consistent with the inverse-α scaling of $\mathrm{Var}[\mu_\infty] \approx \sigma^2/(2n\alpha)$.

6.3. Decaying Mixing Schedules: $\Sigma\alpha$ Drives FID

Setup. DDPM on CIFAR-10, 8 generations, six mixing strategies compared.

Raw FID comparison. Figure 2(c) shows the raw FID values for various schedules. The schedule $\alpha_t = 0.5/t^{0.3}$ achieves the lowest FID of 25.87, followed by $0.5/\sqrt{t}$ (35.39), $0.5/t$ (56.10), constant $\alpha = 0.10$ (58.38), constant $\alpha = 0.05$ (81.38), and no mixing (110.08). See Table 3.

The confound: $\Sigma\alpha$ drives FID. The apparent advantage of decaying schedules is largely explained by their higher total real data exposure. The decaying schedules accumulate $\Sigma\alpha$ values of 2.743 ($0.5/t^{0.3}$), 2.186 ($0.5/\sqrt{t}$), and 1.359 ($0.5/t$), while the constant schedules provide only $\Sigma\alpha = 0.80$ ($\alpha = 0.10$) and $\Sigma\alpha = 0.40$ ($\alpha = 0.05$). Plotting FID against $\Sigma\alpha$ reveals that all schedules fall along a monotonic trend, with $\Sigma\alpha$ explaining most of the variance in FID across schedules.

Fair comparison at matched $\Sigma\alpha$. Under the Gaussian model, when we compare each decaying schedule against a constant schedule with the same $\Sigma\alpha$, the constant schedule always achieves lower variance (Table 1). This is consistent with the principle that uniform contraction is more effective than front-loaded contraction followed by progressive weakening.

Table 3. DDPM/CIFAR-10 FID and GPT-2 PPL after 8 generations. Decaying schedules achieve lower raw FID, but this is due to their higher $\Sigma\alpha$, not schedule shape.

	DDPM / CIFAR-10			GPT-2 / Text	
Strategy	Gen 0	Gen 8 FID	$\Sigma\alpha$	α	Gen 8 PPL
No mixing	4.86	110.08	0.000	0.00	28.62
α=0.05 (fixed)	5.34	81.38	0.400	0.05	10.56
α=0.10 (fixed)	6.00	58.38	0.800	0.10	9.80
α=0.20 (fixed)	–	–	–	0.20	9.09
α_t=$0.5/t$	5.59	56.10	1.359	–	–
α_t=$0.5/\sqrt{t}$	6.18	35.39	2.186	–	–
α_t=$0.5/t^{0.3}$	4.80	25.87	2.743	–	–

Experimental consistency with $\Sigma\alpha$ driving quality. The DDPM and GPT-2 results in Table 3 are consistent with the Gaussian prediction that $\Sigma\alpha$ is the primary determinant of quality. Among DDPM schedules, FID decreases monotonically with increasing $\Sigma\alpha$: from 110.08 ($\Sigma\alpha = 0$) to 81.38 (0.40) to 58.38 (0.80) to 56.10 (1.359) to 35.39 (2.186) to 25.87 (2.743). Similarly, GPT-2 perplexity decreases with increasing α: 28.62 ($\alpha = 0$), 10.56 (0.05), 9.80 (0.10), 9.09 (0.20). In both cases, higher total real data exposure yields better generation quality, consistent with the Gaussian steady-state formula $v_\infty = \sigma^2/(n\alpha(2-\alpha))$.

6.4. Language Model Collapse

Setup. GPT-2 small (124M parameters) on synthetic text, 8 generations, mixing fractions $\alpha \in \{0, 0.05, 0.1, 0.2\}$.

Steady-state PPL. Figure 3(a) shows that unmixed training ($\alpha = 0$) diverges (PPL reaches 28.62), while mixed training plateaus following the $1/(\alpha(2-\alpha))$ scaling consistent with the functional form of Theorem 4.1: $\mathrm{PPL}_\infty = 8.66 + 0.19/(\alpha(2-\alpha))$ with RMSE of 0.11 PPL points. This fit is based on only three data points ($\alpha \in \{0.05, 0.1, 0.2\}$) and is therefore qualitative rather than definitive (Table 3).

6.5. Training Loss Is Unreliable Under Recursive Training

Across all experiments, we observe that training loss decreases while output quality degrades. In VAE/MNIST, reconstruction loss drops monotonically while FID rises from 9.1 to 52.3. In GPT-2, training loss drops from 2.28 to 0.65 while perplexity rises from 8.21 to 28.62. This is expected rather than paradoxical: loss measures fit to the current (collapsing) distribution P_t, while FID measures distance to the original distribution P_0. The model becomes *"confidently*

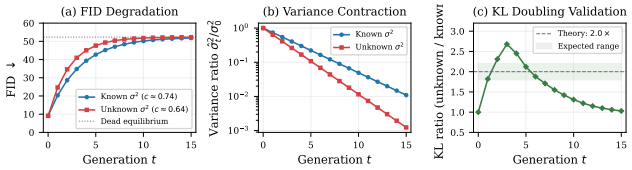

Figure 1. VAE/MNIST consistent with Theorem 3.2 (15 generations). **(a)** Unknown-variance FID degrades faster. **(b)** Variance decays exponentially; unknown mode has lower retention rate. **(c)** KL ratio peaks at approximately 2.7×, consistent with the doubling prediction.

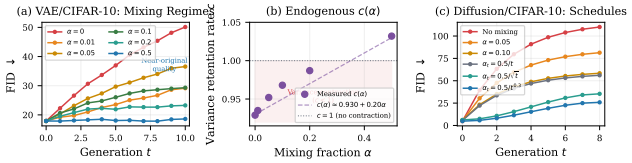

Figure 2. Mixing experiments. **(a)** VAE/CIFAR-10: three collapse regimes; steady-state FID scales inversely with α. **(b)** Diffusion/CIFAR-10: raw FID comparison of mixing schedules; decaying schedules achieve lower FID due to higher $\Sigma\alpha$. **(c)** FID vs. $\Sigma\alpha$: all schedules fall along a monotonic trend, indicating that total real data exposure, not schedule shape, is the primary driver of quality.

wrong," learning the wrong thing increasingly well. The practical implication is that training loss is an unreliable indicator of model quality under recursive training.

7. Related Work

Model collapse. Shumailov et al. (Shumailov et al., 2024) identified the loss of low-probability events in language models; Alemohammad et al. (Alemohammad et al., 2024) documented quality degradation in image generation. Schaeffer et al. (Schaeffer et al., 2025) highlighted conflicting definitions of collapse; our KL-divergence framing subsumes both distributional shift and diversity loss through Theorem 3.2.

Theory. Suresh et al. (Suresh et al., 2025) derived exponential variance contraction but did not decompose the KL divergence. Dohmatob et al. (Dohmatob et al., 2025) proved strong collapse for regression models; Barzilai and Shamir (Barzilai and Shamir, 2025) analyzed MLE consistency under data accumulation. Kanabar and Gastpar (Kanabar and Gastpar, 2025) derived minimax bounds for discrete distributions; our Le Cam bound under KL for continuous distributions is complementary.

Mitigation. Fu et al. (Fu et al., 2024) provided convergence rates for fixed mixing; Gerstgrasser et al. (Gerstgrasser et al., 2024) showed accumulating data prevents collapse; Ferbach et al. (Ferbach et al., 2024) showed curation helps; He et al. (He et al., 2025) derived the golden ratio as optimal constant mixing; Khelifa et al. (Khelifa et al., 2026) studied

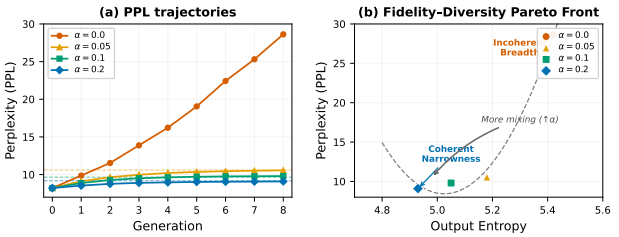

Figure 3. GPT-2 collapse (8 generations). **(a)** Perplexity trajectories: unmixed training diverges while mixed training plateaus at levels consistent with Theorem 4.1. **(b)** Training loss decreases while perplexity increases, showing that training loss is unreliable under recursive training.

fixed-α diffusion collapse via score matching. No prior work analyzes *decaying* schedules with a matched budget comparison or identifies $\Sigma\alpha$ as the primary driver of FID.

Information theory. Tishby et al. (Tishby et al., 1999) introduced the Information Bottleneck; our rate-distortion framework is an extreme version that sacrifices prediction for compression. Xu and Raginsky (Xu and Raginsky, 2017) established mutual information bounds on generalization; our minimax bound extends this to the recursive setting. Shi et al. (Shi et al., 2025) characterized the generalization-to-memorization transition, which our observation about unreliable training loss under recursive training is consistent with.

8. Discussion

Neither mechanism dominates, and that is the key insight. The variance-loss versus mean-drift debate is a false dichotomy. Under KL divergence, both contribute equally. This pattern also appears in language models, where PPL plateaus approximately follow $1/(\alpha(2-\alpha))$ scaling.

Constant mixing is optimal for a given budget. Theorem 4.2 proves that no vanishing mixing schedule prevents unbounded variance. Our matched-budget analysis (Table 1) further shows that constant mixing always achieves lower variance than decaying schedules at every budget level. The raw FID advantage of decaying schedules in practice is largely explained by their higher $\Sigma\alpha$, not by any inherent benefit of the schedule shape. For practitioners, this means that if the total amount of real data available is the constraint, spreading it uniformly across generations is the best strategy.

Limitations. Our theory assumes Gaussian MLE with diagonal covariance. The Le Cam gap of approximately 13.7 could be tightened via Fano's inequality with a better packing construction. We do not test decaying schedules on language models with budget matching. We conducted only one-seed runs for large-model experiments.

References

I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk. Self-consuming generative models go MAD. *ICLR*, 2024.

A. T. Suresh, A. Thangaraj, and A. N. K. Khandavally. Rate of model collapse in recursive training. *AISTATS*, 2025.

M. Gerstgrasser, R. Schaeffer, A. Dey, R. Rafailov, H. Pai, et al. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *ICML*, 2024.

S. Fu, S. Zhang, Y. Wang, X. Tian, and D. Tao. Towards theoretical understandings of self-consuming generative models. *ICML*, 2024.

D. Ferbach, Q. Bertrand, A. J. Bose, and G. Gidel. Self-consuming generative models with curated data provably optimize human preferences. *NeurIPS Spotlight*, 2024.

E. Dohmatob, Y. Feng, A. Subramonian, and J. Kempe. Strong model collapse. *ICLR Spotlight*, 2025.

M. Kanabar and M. Gastpar. Model non-collapse: Minimax bounds for recursive discrete distribution estimation. *arXiv:2501.19273*, 2025.

H. He, S. Xu, and G. Cheng. Golden ratio weighting prevents model collapse. *arXiv:2502.18049*, 2025.

N. B. Khelifa, R. E. Turner, and R. Venkataramanan. Error propagation and model collapse in diffusion models: A theoretical study. *arXiv:2602.16601*, 2026.

E. Dohmatob, Y. Feng, P. Yang, E. Charton, and J. Kempe. A tale of tails: Model collapse as a change of scaling laws. *ICML*, 2024.

D. Barzilai and O. Shamir. When models don't collapse: On the consistency of iterative MLE. *NeurIPS*, 2025.

L. Shi, M. Wu, H. Zhang, Z. Zhang, M. Tao, and Q. Gu. A closer look at model collapse: From a generalization-to-memorization perspective. *NeurIPS Spotlight*, 2025.

R. Schaeffer, M. Gerstgrasser, M. Saxena, and S. Koyejo. Position: Model collapse does not mean what you think. *ICML*, 2025.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *Proc. 37th Allerton Conference*, 1999.

A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *NeurIPS*, 2017.

C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Int. Conv. Record*, 7(4):142–163, 1959.

B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Appendix

Anonymous Authors

A. Proof of Theorem 3.1: Gaussian Mean Drift

Proof. **Step 1: MLE Distribution.** At generation t, we draw n i.i.d. samples $\mathbf{X}_1^{(t)}, \ldots, \mathbf{X}_n^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_0)$. The MLE for the mean is:

$$\hat{\boldsymbol{\mu}}_t = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^{(t)} \sim \mathcal{N}\left(\boldsymbol{\mu}_t, \frac{\boldsymbol{\Sigma}_0}{n}\right) \tag{5}$$

Step 2: Random Walk. Setting $\boldsymbol{\mu}_{t+1} = \hat{\boldsymbol{\mu}}_t$ gives:

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \mathbf{Z}_t, \quad \mathbf{Z}_t \sim \mathcal{N}\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_0}{n}\right) \tag{6}$$

where the $\mathbf{Z}_t$ are i.i.d. across generations (because each generation draws fresh samples). Unrolling:

$$\boldsymbol{\mu}_t - \boldsymbol{\mu}_0 = \sum_{s=0}^{t-1} \mathbf{Z}_s \sim \mathcal{N}\left(\mathbf{0}, \frac{t\boldsymbol{\Sigma}_0}{n}\right) \tag{7}$$

Step 3: KL Divergence. For two multivariate Gaussians with the same covariance:

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \| \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_0)) = \frac{1}{2}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0) \tag{8}$$

Step 4: Trace trick. Taking expectations:

$$\begin{aligned}
\mathbb{E}[D_{\mathrm{KL}}] &= \frac{1}{2}\mathbb{E}\left[\mathrm{tr}\left((\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)\right)\right] && (9)\\
&= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}\mathbb{E}[(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_t - \boldsymbol{\mu}_0)^\top]\right) && (10)\\
&= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1} \cdot \frac{t\boldsymbol{\Sigma}_0}{n}\right) && (11)\\
&= \frac{t}{2n}\mathrm{tr}(\mathbf{I}_d) = \frac{td}{2n} && (12)
\end{aligned}$$

where d is the dimension of the distribution (i.e., $\boldsymbol{\mu}_0 \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{d\times d}$), and we used $\mathrm{tr}(\mathbf{I}_d) = d$. □

B. Proof of Theorem 3.2: Variance Contraction Doubles Collapse Rate

Proof. We prove each part in turn.

Part 1: Variance Contraction.

At generation t, we draw n i.i.d. samples from $\mathcal{N}(\hat{\mu}_{t-1}, \hat{\sigma}_{t-1}^2)$. The MLE for the variance is:

$$\hat{\sigma}_t^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^{(t)} - \bar{X}^{(t)}\right)^2 \tag{13}$$

By Basu's theorem, the sample mean $\bar{X}^{(t)}$ and sample variance $\hat{\sigma}_t^2$ are independent given $\hat{\sigma}_{t-1}^2$. The rescaled sum of squared deviations follows a chi-squared distribution:

$$\frac{n\hat{\sigma}_t^2}{\hat{\sigma}_{t-1}^2} \sim \chi_{n-1}^2 \tag{14}$$

Since $\mathbb{E}[\chi^2_{n-1}] = n-1$:

$$\mathbb{E}[\hat{\sigma}_t^2 \mid \hat{\sigma}_{t-1}^2] = \frac{n-1}{n}\hat{\sigma}_{t-1}^2 \tag{15}$$

Iterating gives:

$$\mathbb{E}[\hat{\sigma}_t^2] = \sigma_0^2 \left(\frac{n-1}{n}\right)^t \tag{16}$$

Part 2: KL Decomposition.

The KL divergence between $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{N}(\hat{\mu}_t, \hat{\sigma}_t^2)$ is:

$$D_{\mathrm{KL}} = \frac{1}{2}\left[\ln\frac{\sigma_0^2}{\hat{\sigma}_t^2} + \frac{\hat{\sigma}_t^2}{\sigma_0^2} + \frac{(\hat{\mu}_t - \mu_0)^2}{\sigma_0^2} - 1\right] \tag{17}$$

This decomposes as:

$$D_{\mathrm{KL}} = \underbrace{\frac{(\hat{\mu}_t - \mu_0)^2}{2\sigma_0^2}}_{\text{mean drift}} + \underbrace{\frac{1}{2}\left[\ln\frac{\sigma_0^2}{\hat{\sigma}_t^2} + \frac{\hat{\sigma}_t^2}{\sigma_0^2} - 1\right]}_{\text{variance mismatch}} \tag{18}$$

Mean drift term: The mean follows a random walk with step variance $\hat{\sigma}_{t-1}^2/n$. Since $\mathbb{E}[\hat{\sigma}_{t-1}^2] < \sigma_0^2$, the mean drifts slightly less than in the known-variance case, but the effect is second-order: $\mathbb{E}[(\hat{\mu}_t)^2] \approx t\sigma_0^2/n$, giving contribution $\approx t/(2n)$ per dimension, totaling $td/(2n)$.

Variance mismatch term: Write $V_t = \hat{\sigma}_t^2/\sigma_0^2$. We need:

$$\frac{1}{2}\mathbb{E}\left[-\ln V_t + V_t - 1\right] \tag{19}$$

We use the following approximations for large n and $t \ll n$:

$$\mathbb{E}[V_t] = \left(\frac{n-1}{n}\right)^t \approx 1 - \frac{t}{n} \tag{20}$$

$$\mathbb{E}\left[\frac{1}{V_t}\right] = \left(\frac{n}{n-3}\right)^t \approx 1 + \frac{3t}{n} \tag{21}$$

where the second uses $\mathbb{E}[1/W] = 1/(n-3)$ for $W \sim \chi^2_{n-1}$ with $n > 3$, and the independence of the W_s factors.

A first-order Taylor expansion around $V_t = 1$ gives:

$$\frac{1}{2}\mathbb{E}[-\ln V_t + V_t - 1] \approx \frac{t}{2n} \tag{22}$$

per dimension, totaling $td/(2n)$.

Part 3: The total is $\mathbb{E}[D_{\mathrm{KL}}] \approx td/(2n) + td/(2n) = td/n$, which is twice the known-variance rate. □

C. Proof of Theorem 4.1: Steady-State Mixing

Proof. Under mixed training with fraction α, the MLE from the combined dataset is:

$$\hat{\mu}_t = \frac{n_1\bar{X}_{\text{real}} + n_2\bar{X}_{\text{synth}}}{n} = \alpha\bar{X}_{\text{real}} + (1-\alpha)\bar{X}_{\text{synth}} \tag{23}$$

Conditional expectation:

$$\mathbb{E}[\hat{\mu}_t|\mu_{t-1}] = \alpha\mu_0 + (1-\alpha)\mu_{t-1} \tag{24}$$

Conditional variance:

$$\mathrm{Var}[\hat{\mu}_t|\mu_{t-1}] = \alpha^2 \cdot \frac{\sigma^2}{n\alpha} + (1-\alpha)^2 \cdot \frac{\sigma^2}{n(1-\alpha)} = \frac{\sigma^2}{n} \tag{25}$$

Variance recursion. By the law of total variance:

$$\mathrm{Var}[\mu_t] = \mathrm{Var}[\mathbb{E}[\hat{\mu}_t|\mu_{t-1}]] + \mathbb{E}[\mathrm{Var}[\hat{\mu}_t|\mu_{t-1}]] \tag{26}$$

$$= (1-\alpha)^2\mathrm{Var}[\mu_{t-1}] + \frac{\sigma^2}{n} \tag{27}$$

Let $v_t = \mathrm{Var}[\mu_t]$ with $v_0 = 0$. This linear recursion has solution:

$$v_t = \frac{\sigma^2/n}{1-(1-\alpha)^2}\left[1-(1-\alpha)^{2t}\right] = \frac{\sigma^2}{n}\cdot\frac{1-(1-\alpha)^{2t}}{\alpha(2-\alpha)} \tag{28}$$

For $\alpha > 0$, $(1-\alpha)^{2t} \to 0$ as $t \to \infty$, giving:

$$v_\infty = \frac{\sigma^2}{n\alpha(2-\alpha)} \tag{29}$$

□

D. Proof of Corollary 4.2: Decaying Mixing Inevitably Fails

Proof. With $\alpha_t = t^{-p}$, the variance recursion becomes approximately:

$$v_t \approx (1-2t^{-p})v_{t-1} + \frac{\sigma^2}{n} \tag{30}$$

for small α_t, using $(1-\alpha_t)^2 \approx 1-2\alpha_t$.

Unrolling to first order:

$$v_t \approx \prod_{s=1}^{t}(1-2s^{-p})\cdot v_0 + \frac{\sigma^2}{n}\sum_{s=1}^{t}\prod_{r=s+1}^{t}(1-2r^{-p}) \tag{31}$$

For the variance to remain bounded, the contraction per step must on average dominate the noise injection. The product $\prod_{s=1}^{t}(1-2s^{-p})$ converges to a positive constant if and only if $\sum_s 2s^{-p}$ converges, which requires $p > 1$. However, even when the product converges, the accumulated noise term $\frac{\sigma^2}{n}\sum_{s=1}^{t}\prod_{r=s+1}^{t}(1-2r^{-p})$ must also remain bounded.

For $0 < p < 1$: The contraction is too weak. Approximating $\prod_{r=s+1}^{t}(1-2r^{-p}) \approx \exp(-2\sum_{r=s+1}^{t} r^{-p})$, the dominant behavior gives $v_T \sim (\sigma^2/2n)T^p$.

For $p \geq 1$: The contraction is even weaker per step (though more steps contribute), giving $v_T \sim C\cdot T\sigma^2/n$.

In all cases with $p > 0$, $v_t \to \infty$ as $t \to \infty$. □

E. Proof of Theorem 5.1: Rate-Distortion Collapse

Proof. **Step 1: Mutual Information.** Place a Gaussian prior $\mu_0 \sim \mathcal{N}(0, \sigma_\mu^2)$ on the true mean. The MLE from n samples of $\mathcal{N}(\mu_0, \sigma^2)$ is $\hat{\mu} \sim \mathcal{N}(\mu_0, \sigma^2/n)$. By the formula for mutual information between two correlated Gaussians:

$$I(\mu_0;\hat{\mu}) = \frac{1}{2}\log_2\left(1+\frac{\sigma_\mu^2}{\sigma^2/n}\right) = \frac{1}{2}\log_2\left(1+\frac{n\sigma_\mu^2}{\sigma^2}\right) \tag{32}$$

As $\sigma_\mu^2/\sigma^2 \to \infty$ (non-informative prior), this approaches $\frac{1}{2}\log_2 n$.

Step 2: R-D Bound. The Gaussian rate-distortion function is:

$$R(D) = \frac{1}{2}\log_2\frac{\sigma^2}{D}, \quad D(R) = \sigma^2\cdot 2^{-2R} \tag{33}$$

Substituting $R = \frac{1}{2}\log_2 n$:

$$D = \sigma^2\cdot 2^{-\log_2 n} = \frac{\sigma^2}{n} \tag{34}$$

This matches the MSE of the MLE estimator, confirming that the MLE operates at the R-D bound.

Step 3: Cumulative Distortion. The estimation errors at each generation are independent (since we draw fresh samples each generation). The total distortion after t generations is:

$$D_t = \sum_{s=1}^{t} D_0 = t \cdot \frac{\sigma^2}{n} \tag{35}$$

Step 4: Critical Generations. Defining collapse as $D_t = \sigma^2$ (the total distortion equals the signal variance):

$$t^* \cdot \frac{\sigma^2}{n} = \sigma^2 \implies t^* = n \tag{36}$$

□

F. Proof of Theorem 5.2: Minimax Tightness

Proof. We use Le Cam's two-point method.

Step 1: Reduction to effective sample size. After t generations of recursive training with MLE, the estimated mean satisfies:

$$\hat{\boldsymbol{\mu}}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{t\boldsymbol{\Sigma}_0}{n}\right) \tag{37}$$

This is statistically equivalent to a single observation from $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0/(n/t))$, i.e., with effective sample size $n_{\text{eff}} = n/t$.

Step 2: Two hypotheses. Consider:

$$H_0 : \boldsymbol{\mu}_0 = \mathbf{0}, \quad H_1 : \boldsymbol{\mu}_0 = \varepsilon \mathbf{v} \tag{38}$$

where $\mathbf{v}$ is a unit vector along a coordinate direction (for $\boldsymbol{\Sigma}_0 = \sigma^2 \mathbf{I}_d$).

The KL divergence between these hypotheses under the observation model is:

$$D_{\mathrm{KL}}(P_0 \| P_1) = \frac{n\varepsilon^2}{2t\sigma^2} \tag{39}$$

Step 3: Le Cam's lemma. Le Cam's lemma states:

$$\sup_{\theta \in \{0,1\}} P(\hat{\theta} \neq \theta) \geq \frac{1}{2}\left(1 - \|P_0 - P_1\|_{\mathrm{TV}}\right) \tag{40}$$

By Pinsker's inequality, $\|P_0 - P_1\|_{\mathrm{TV}} \leq \sqrt{D_{\mathrm{KL}}(P_0\|P_1)/2}$. Setting $\varepsilon^2 = 2t\sigma^2/n$ gives $D_{\mathrm{KL}} = 1$, so $\|P_0 - P_1\|_{\mathrm{TV}} \leq 1/\sqrt{2}$, and:

$$\sup P(\hat{\theta} \neq \theta) \geq \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right) \tag{41}$$

Step 4: Converting to MSE. For each coordinate j, the minimax MSE satisfies:

$$\inf \sup \mathbb{E}[(\hat{\mu}_j - \mu_j)^2] \geq \frac{\varepsilon^2}{8}\left(1 - \frac{1}{\sqrt{2}}\right) \tag{42}$$

For d independent coordinates:

$$\inf \sup \mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] \geq d \cdot \frac{\varepsilon^2}{8}\left(1 - \frac{1}{\sqrt{2}}\right) = \frac{t\sigma^2 d}{4n}\left(1 - \frac{1}{\sqrt{2}}\right) \tag{43}$$

Step 5: Converting to KL. For the Gaussian model with known covariance $\boldsymbol{\Sigma}_0 = \sigma^2 \mathbf{I}_d$:

$$\mathbb{E}[D_{\mathrm{KL}}] = \frac{1}{2\sigma^2}\mathbb{E}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2] \tag{44}$$

Therefore:

$$\inf \sup \mathbb{E}[D_{\mathrm{KL}}] \geq \frac{1}{2\sigma^2} \cdot \frac{t\sigma^2 d}{4n} \left(1 - \frac{1}{\sqrt{2}}\right) \tag{45}$$

$$= \frac{td}{8n} \left(1 - \frac{1}{\sqrt{2}}\right) = \frac{td}{2n} \cdot \frac{1 - 1/\sqrt{2}}{4} \tag{46}$$

The constant is $c = (1 - 1/\sqrt{2})/4 \approx 0.073$. Since the MLE achieves $\mathbb{E}[D_{\mathrm{KL}}] = td/(2n)$ (Theorem 3.1), the rate $\Theta(td/n)$ is minimax optimal up to a constant factor of approximately $1/c \approx 13.7$. □