Surveying Model Collapse: Empirics, Universality, and Implications

Alvan Caleb Arulandu
Harvard University
Cambridge, MA 02138
`aarulandu@college.harvard.edu`

Abstract

Given the near-exhaustion of available English internet corpora for pre-training, traditional scaling laws have ended and will remain plateaued unless synthetic data proves viable. A recent phenomenon known as model collapse has evidenced a performance degradation for certain classes of models when trained on synthetic datasets. This work surveys the extent and limitations of empirical and theoretical arguments for model collapse as well as the recent universalization of the $\pi^2/6$ argument by Dey and Donoho (2024). We reproduce relevant experimentation and proof while analyzing the implications for pre-training and model scaling, given the present deployment of large-scale generative models.

1 Introduction

On December 13, 2024, Ilya Sutskever confirmed an ongoing suspicion of top machine learning scientists for the last month. “Pre-training as we know it will unquestionably end," he declared, comparing the saturation of real training data to the “fossil fuel of AI" in his NeurIPS presentation. Literature has long understood real training data as a “tragedy of the commons," and early forecasts from Villalobos et al. (2022) indicated that all high-quality English data might be exhausted by 2028, provided scaling laws at the time continued.

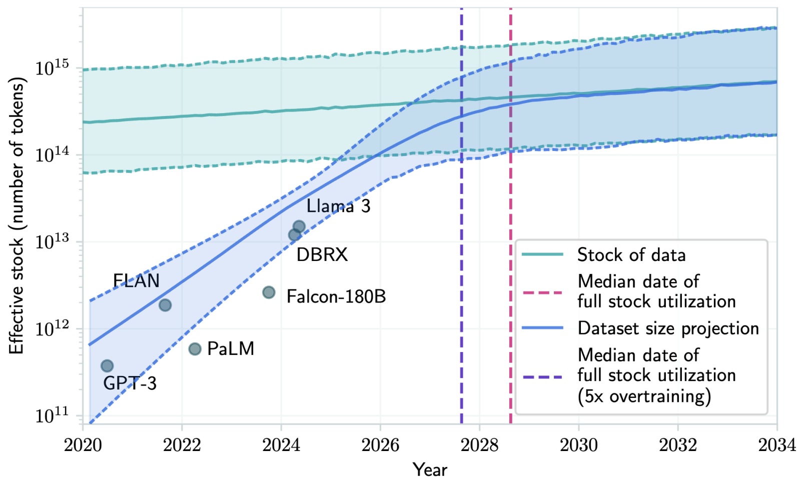


Figure 1: Projected Training Data Exhaustion (Villalobos et al., 2022)

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

While a common crawl of the internet yields $\approx 1.3 \times 10^{14}$ tokens (Villalobos et al., 2022) despite Llama 70B only pre-training on $\approx 1.5 \times 10^{13}$ tokens (Touvron et al., 2023), accounting for data pre-processing needs, we've already surpassed the initial data exhaustion projection of Figure 1. It follows that future efforts to scale pre-training involve either better use of existing data or the generation and incorporation of synthetic data into the training set. While the former is possible, beyond exploiting multi-modal pre-training data (Ding et al., 2023), expected gains from better analysis of existing corpora are marginal. Further, by the "No Free Lunch Theorem," it is reasonable to believe that the latter comes with a caveat. This caveat is model collapse.

2 Background

Broadly, synthetic data is a potential solution for remedying data sparsity and diversity concerns (H. Chen et al., 2024). A notable example in real-world decision making is self-driving; while companies, ex. Tesla, are able to record large quantities of high-quality training data for standard driving conditions, data for inclement weather and a wide variety of emergency scenarios are scarce. Training data generated from real-world-driven *world models* are hoped to fill this gap (Wang et al., 2023), and similarly, synthetic data is hoped to build more complete datasets for real-world applications.

However, data poisoning from existing generative models is a simultaneous concern. Given the unprecedented progress in large language models which are now deployed at scale (Mohamadi et al., 2023), the proportion of real data with respect to a common internet crawl is declining. Naturally, as users post generative-inspired text, media from diffusion models, and more, data poisoning worsens. While there do exist substantive efforts to distinguish between synthetic and generative data, future researchers will undeniably be faced with the harsh reality of indistinguishably poisoned data sets. However, this view seems to be at odds with the former; if synthetic data assists training, why is data poisoning even a concern to begin with?

Crucially, synthetic data does not universally improve training, but rather must be carefully used in tandem with real data to improve performance. This balancing act is heavily dependent on the model and problem space, with experimentalists observing drastic effects in certain settings. A slew of works including Alemohammad et al. (2023) initially noted declined performance in generative image models which were trained in a self-consuming loop.

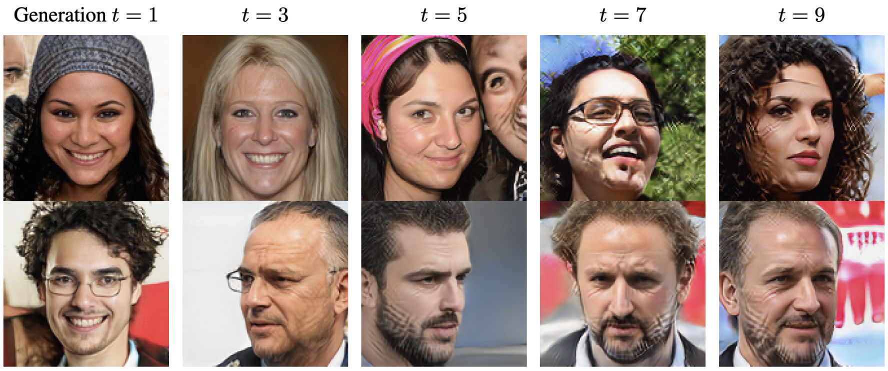


Figure 2: Self-Consuming Training of StyleGAN-2 (Alemohammad et al., 2023)

Particularly, while generation $t = 1$ is trained on the input data distribution, generation $t \geq 2$ is obtained via a training set synthesized solely from the trained model at generation $t - 1$. As pictured in Figure 2, one of the initial model collapse experiments with StyleGAN-2, as the model generation increases, cross-hatched artifacts are progressively amplified in each new generation. While such deficiencies in the generative image setting were known even prior to the diffusion era, only recently have researchers been able to make theoretical headway toward a rigorous treatment of model collapse.

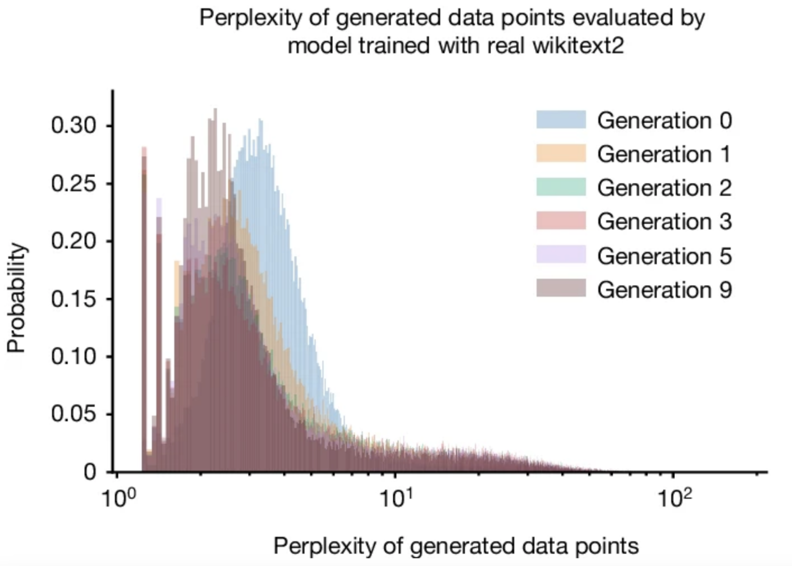


Figure 3: Self-Consuming Training, with 10% Preservation, of OPT-125m (Shumailov et al., 2024)

As in Figure 3, Shumailov et al. (2024) famously studied the language model settings, finding that OPT-125m (Zhang et al., 2022) when trained in a self-consuming fashion yields an increasing tail mass as the generation increases, meaning subsequent generations begin “misperceiving reality" based on ancestral outliers. While Shumailov et al. (2024) also study Stable Diffusion (Rombach et al., 2022) and ChatGPT (Gozalo-Brizuela and Garrido-Merchan, 2023), the authors pivotally introduce the first theoretical handhold: proven collapse in the Gaussian setting.

3 Proving Collapse

3.1 Gaussians with Sample Estimators

Suppose we want to learn some 1D Gaussian $X_0 \sim \mathcal{N}(\mu, \sigma^2)$ and consider the following iterative process via parameter point estimates. For all $i \geq 1$,

$$\mu_i = \frac{1}{T_{i-1}} \sum_j X_{i-1}^j, \quad \sigma_i^2 = \frac{1}{T_{i-1}-1} \sum_j (X_{i-1}^j - \mu_i)^2, \quad X_i^j | \mu_i, \sigma_i^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_i, \sigma_i^2) \tag{1}$$

meaning T_i samples from the estimated distribution at generation i are used to compute the parameter estimates, that is the sample mean and sample variance, for the subsequent generation. From literature, we know that X_i^j follows a variance-gamma distribution (Fischer, Gaunt, and Sarantsev, 2023) with $\mu_1 \sim \mathcal{N}(\mu, \frac{\sigma^2}{T_0})$ and $\sigma_1^2 \sim \frac{\sigma^2}{T_0 - 1} \cdot \Gamma\left(\frac{T_0 - 1}{2}, \frac{1}{2}\right)$ (Cochran, 1934). Applying this recursively,

$$X_0^j = \mu + \sigma Z_0^j \Rightarrow X_1^j = \mu + \frac{\sigma}{\sqrt{T_0}} Z_1 + \sigma \sqrt{S_1} Z_1^j$$

$$X_g^j = \mu + \frac{\sigma}{\sqrt{T_0}} Z_1 + \frac{\sigma}{\sqrt{T_1}} \sqrt{S_1} \cdot Z_2 + \cdots + \frac{\sigma}{\sqrt{T_{g-1}}} \sqrt{\prod_{j=1}^{g-1} S_j} \cdot Z_g + \sigma \sqrt{\prod_{j=1}^{g} S_j} \cdot Z_g^j$$

where $S_i \sim \frac{1}{T_{i-1}-1} \Gamma\left(\frac{T_{i-1}-1}{2}, \frac{1}{2}\right)$ and $Z_i^j \sim \mathcal{N}(0,1)$. Taking $T_j = T$ for all j such that n is large and expanding to second order in $1/T_i$, Shumailov et al. (2024) find that

$$\frac{1}{\sigma^2} \operatorname{Var}(X_g^j) = \sum_{j=0}^{g-1} \frac{1}{T_j} + 1 + O(\max_j M_j^{-2}) \Rightarrow \operatorname{Var}(X_g^j) = \sigma^2 \left(1 + \frac{g}{T}\right), \quad \mathbb{E}(X_g^j) = \mu$$

meaning that the variance diverges linearly in g. By taking the Wasserstein-2 distance between the g-th estimated distribution and the original data distribution,

$$R^g_{W_2} \triangleq \mathbb{W}_2^2(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\mu_g, \sigma_g^2)_)) = \|\mu_g - \mu\|^2 + \|\sigma_g - \sigma\|^2$$

we are able to measure the fidelity of the point estimates μ_g, σ_g^2. Doing so, Shumailov et al. (2024) find that

$$\mathbb{E}_{\mu_g,\sigma_g^2}[R^g_{W_2}] = \frac{3\sigma^2}{2}\sum_{j=0}^{g}\frac{1}{T_j} + O(\max_j M_j^{-2}) \tag{2}$$

$$\mathrm{Var}_{\mu_g,\sigma_g^2}[R^g_{W_2}] = \frac{\sigma^4}{2}\left(\sum_{j=0}^{g}\frac{3}{T_j^2} + \sum_{i\neq j}\frac{4}{T_i T_j}\right) + O(\max_j M_j^{-3})$$

meaning that the risk due to finite sampling, $\mathbb{E}_{\mu_g,\sigma_g^2}[R^g_{W_2}]$, does not diverge if and only if T_g increases superlinearly in g by the p-series test.

3.1.1 Multidimensional Gaussians

Generalizing the above approach, consider any self-consuming process on Gaussians such that

$$\mu_i = \frac{1}{T_{i-1}}\sum_j X^j_{i-1} + \epsilon_i(X^j_{i-1}), \quad \mathbb{E}_{X^j_{i-1}}(\Sigma_i) = \Sigma_{i-1}, \quad \mathbb{E}[\epsilon_i|\mu_{i-1}, \Sigma_{i-1}] = 0$$

where ϵ_{i+1} are deviations from the typical unbiased sample estimators. Shumailov et al. (2024) prove the following risk lower bound, provided $\mathrm{Cov}(\epsilon_i, \mu_i|\mu_{i-1}, \Sigma_{i-1}) = 0$:

$$\mathbb{E}[R^g_{W_2}] \geq \mathrm{Tr}\Sigma\sum_{i=0}^{g-1}\frac{1}{T_i} + \sum_{i=1}^{n+1}\mathbb{E}(\|\epsilon_i\|^2)$$

which recovers (2). Thus, for sample sizes T_g that are constant with respect to g, the Gaussian setting with point estimates yields model collapse.

3.2 Linear Model with Ridge Regression and Gaussians

While the aforementioned Gaussian example theoretically validates the intuition of compounding finite sample errors across each subsequent generation, Dohmatob, Feng, and Kempe (2024) find that the following simple regression setting is rich enough to describe a range of model collapse regimes.

$$(X_k)_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), \quad (E_k)_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_k^2), \quad \hat{w}_k = \mathrm{Fit}(X_{k-1}, \overline{Y}_{k-1}), \quad \overline{Y}_k \triangleq X_k\hat{w}_k + E_k$$

where $\hat{w}_k, w_0 \in \mathbb{R}^d$, $(X_k)_i$ denotes the i-th row of design matrix $X_k \in \mathbb{R}^{T_k \times d}$, and $\overline{Y}_k, E_k \in \mathbb{R}^{T_k}$. Here $\mathrm{Fit}(X_k, Y_k) = RX_k^\top Y_k/T_k$ is the ridge estimator, $R = (\hat{\Sigma} + \lambda I_d)^{-1}$ is the resolvent, and $\hat{\Sigma} = X_k X_k^\top/T_k$ is the sample covariance matrix.

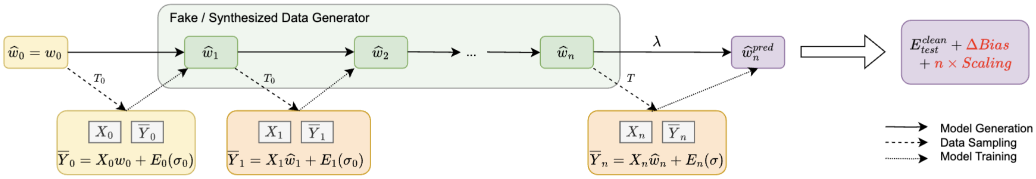

Fake / Synthesized Data Generator
$\widehat{w}_0 = w_0$
$\widehat{w}_1$
$\widehat{w}_2$
...
$\widehat{w}_n$
λ
$\widehat{w}_n^{pred}$
$E_{test}^{clean} + \Delta Bias + n \times Scaling$
T_0
T_0
T
X_0
$\overline{Y}_0$
$\overline{Y}_0 = X_0 w_0 + E_0(\sigma_0)$
X_1
$\overline{Y}_1$
$\overline{Y}_1 = X_1\widehat{w}_1 + E_1(\sigma_0)$
X_n
$\overline{Y}_n$
$\overline{Y}_n = X_n\widehat{w}_n + E_n(\sigma)$
Model Generation
Data Sampling
Model Training

Figure 4: Theoretical Framework for Regression

Pictured end-to-end in Figure 4, the self-consuming process begins with data generated from the true weight w_0 and at the k-th generation performs ridge regression on a sample of T_k points from the distribution fixed by $\hat{w}_k$ to derive $\hat{w}_{k+1}$. For simplicity, we take $T_k = T_0, \sigma_k = \sigma_0$ for all $k \leq g - 1$

and denote $T_g = T, \sigma_g = \sigma$, assuming the under-parameterized regime of $T_0 \geq d + 2$. By standard results (Hastie et al., 2022), the classical ridge regression on the true data distribution yields:

$$E_{\text{test}}^{\text{clean}} \triangleq \text{Bias} + \text{Var}, \quad \text{Bias} \triangleq \mathbb{E}\|\hat{\Sigma} R w_0 - w_0\|_{\Sigma}^2, \quad \text{Var} \triangleq \frac{\sigma^2}{T} \cdot \text{Tr}\Sigma R^2 \hat{\Sigma}$$

which corresponds to the test error of generation $k = 1$. Dohmatob, Feng, and Kempe (2024) show that after g generations of self-consuming training,

$$E_{\text{test}}(\hat{w}_g^{\text{pred}}) \simeq E_{\text{test}}^{\text{clean}} + g\sigma_0^2\rho, \quad \rho \triangleq \frac{1}{T_0 - d - 1}\mathbb{E}\text{Tr}\Sigma^{-1}\hat{\Sigma}R\Sigma\hat{\Sigma}R \tag{3}$$

Of course, in the low-dimensional limit for fixed d as $T \to \infty$, $\hat{\Sigma}_k = X_k^\top X_k / T_k \to \Sigma$ meaning that in the under-parameterized regime,

$$\rho \simeq \frac{\text{df}_2(\lambda)}{T_0 - d}, \quad \text{df}_m(\lambda) \triangleq \text{Tr}\Sigma^m(\Sigma + \lambda I_d)^{-m}$$

where $\text{df}_m(\lambda)$ is known as the m-th order degree of freedom of Σ which is $\leq d$ always. It follows that the test error after g generations is the classical test error plus an additive factor that scales linearly with g. In the more realistic RMT limit (Wei, W. Hu, and Steinhardt, 2022) where $T, d \to \infty$ such that $d^{1/C} \simeq T \simeq d^C$ and $\|\Sigma\|_{\text{op}}, \|\Sigma^{-1}\|_{\text{op}} = O(1)$, meaning $\log d, \log T$ are of the same order, Dohmatob, Feng, and Kempe (2024) derive that in the under-parameterized regime,

$$\rho = \frac{\text{Tr}\Sigma^4(\Sigma + \kappa_0 I)^{-2}(\Sigma + \kappa I)^{-2}}{T_0 - \text{df}_2(\kappa_0)} + \frac{\kappa^2 \text{Tr}\Sigma^2(\Sigma + \kappa_0 I)^{-2}(\Sigma + \kappa I)^{-2}}{T_0 - \text{df}_2(\kappa_0)} \cdot \frac{\text{df}_2(\kappa)}{T - \text{df}_2(\kappa)}$$

where $\kappa_0 \triangleq \kappa(0, T_0), \kappa = \kappa(\lambda, T)$, and $\kappa(\lambda, T)$ satisfies the fixed point equation $\kappa(\lambda, T) - \lambda = \kappa(\lambda, T) \cdot \text{df}_1(\kappa(\lambda, T))/T$. The calculation of this trace term ρ follows the standard application of RMT tools to the anisotropic ridge regression setting, (Bach, 2024).

Dohmatob, Feng, and Kempe (2024) extend their results to the adaptive ridge regression setting where $\lambda = \lambda(T) \asymp T^{-\ell}$ for some fixed $\ell \geq 0$ and also note that similar calculations apply to the kernel setting where x is replaced with a feature map induced by a kernel K. Having proven model collapse via a linear test error lower bound in the generation number g, Dohmatob, Feng, and Kempe (2024) argue that this degradation suggests that large language models will pollute the web till learning is no longer possible from un-curated online data.

4 Accumulation

While Shumailov et al. (2024) and Alemohammad et al. (2023) do consider settings in which limited data from generations prior to $k - 1$ is used to augment training the k-th generation, a major criticism of recent literature propounding model collapse is the impractical nature of self-consuming training. In solely self-consuming loops, information from the initial dataset is bound to be lost while finite sample errors are amplified across each generation, since no meaningful reference to the original data is made beyond generation 1.

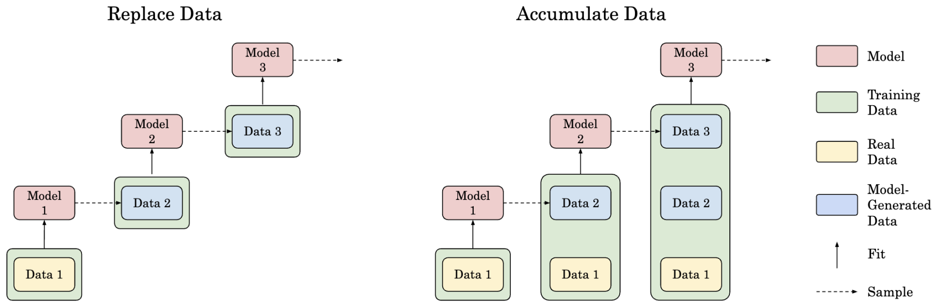


Figure 5: Settings for Studying Model Collapse (Gerstgrasser et al., 2024)

In the real world, generative outputs pollute the common crawl via accumulation, not replacement; while the real training data diminishes in proportion, as portrayed in Figure 5, it remains available for training at any generation, potentially allowing finite sample errors to be corrected to an extent in the generational limit.

4.1 Linear Model with Ridge Regression and Gaussians

Astonishingly, Gerstgrasser et al. (2024) demonstrate that accumulating data mitigates model collapse in the same setting studied by Dohmatob, Feng, and Kempe (2024). Formally, assuming X has full column rank, $T \geq d+2$ (under-parameterized), and $X^\top X$ is invertible, Gerstgrasser et al. (2024) take the ridgeless iterative fitting scheme:

$$\hat{w}_k = \tilde{X}_k^\dagger \tilde{Y}_k, \quad \hat{Y}_k = X\hat{w}_{k-1} + E_k, \quad E_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 I_T)$$

where $A^\dagger = (A^\top A)^{-1} A^\top$ is the Moore-Penrose pseudo-inverse and accumulation is applied via

$$\tilde{Y}_k^\top = [\tilde{Y}_{k-1}^\top; \hat{Y}_k^\top], \tilde{Y}_1 \triangleq \hat{Y}_1, \quad \tilde{X}_k^\top = [\tilde{X}_{k-1}^\top; \hat{X}_k^\top], \tilde{X}_1 \triangleq \hat{X}_1$$

for all $k \geq 2$. Note that this is comparable to the ridgless analog of Dohmatob, Feng, and Kempe (2024) with $T_0 = \cdots = T_g = T$ and $\sigma_0^2 = \cdots = \sigma_g^2 = \sigma^2$. In this setting, Gerstgrasser et al. (2024) find that

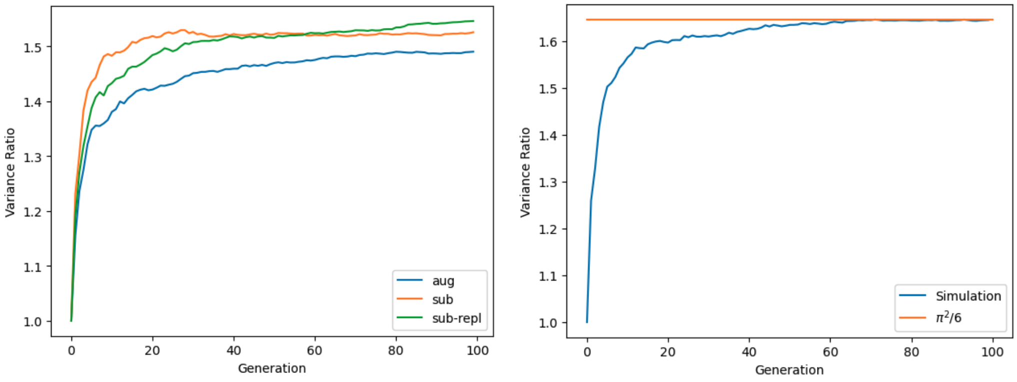
$$\hat{w}_n = w^* + X^\dagger \left(\sum_{i=1}^{n} \frac{E_i}{i} \right) \tag{4}$$

and if $\Sigma = I_d$ meaning the features are isotropic,

$$E_{\text{test}}^{\text{Replace}} = \frac{\sigma^2 d}{T-d-1} \cdot n \tag{5}$$

$$E_{\text{test}}^{\text{Accum}} = \frac{\sigma^2 d}{T-d-1} \left(\sum_{i=1}^{n} \frac{1}{i^2} \right) \leq \frac{\sigma^2 d}{T-d-1} \cdot \frac{\pi^2}{6} \tag{6}$$

Unlike the linear test error (5) the replace scheme, recovering (3) in the isotropic case, the accumulation scheme (6) yields bounded test error, by the famous Basel problem. We proceed with our own simulation of this phenomenon.



(a) Comparison of Aggregation Methods

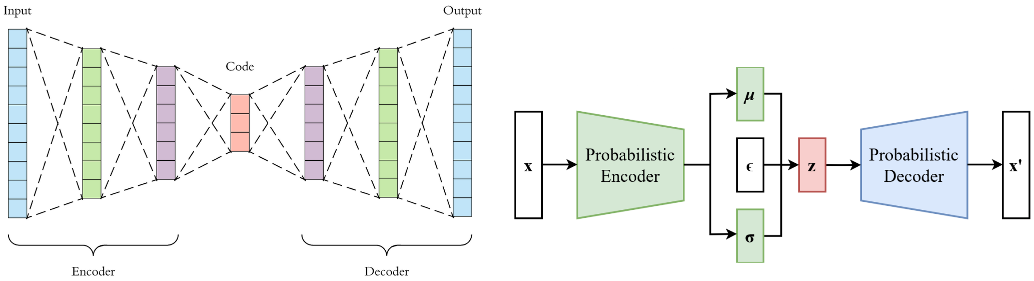
(b) Accumulation Saturates the Theoretical Bound

Figure 6: Variance Ratio (ARE) over Iterative Training Generations

In Figure 6b, we see that the ratio of empirical variances from the k-th to 1st generation across 10^4 trials limits to $\pi^2/6$, as predicted by (6). In Figure 6a, we simulate various accumulation strategies: “aug" being the classical accumulation, “sub" being the sub-sample procedure (Gerstgrasser et al., 2024) of sampling n of the $n \cdot k$ samples in $\tilde{Y}_k$ at the k-th generation, and “sub-repl" being the previous procedure with replacement. Figure 6a suggests that “sub" and “sub-repl" have convergent variance ratios with higher limiting variance compared with the accumulation procedure, which confirms a claim of Dey and Donoho (2024).

4.2 Generative Image Models

Motivated by their theoretical results for regression, Gerstgrasser et al. (2024) empirically analyze the generative image setting, particularly variational auto-encoders (VAE) on the CelebA dataset.



(a) AE

(b) VAE

Figure 7: A Diagram of a Classical and Variational Auto-Encoder

As pictured in Figure 7a, an auto-encoder (AE) is a generative model, typically composed of two sequentially coupled dense networks. These networks, called the encoder and decoder, are trained to compress input data to an intermediate vector in latent space by back-propagating a reconstruction loss of the input data with the final output data throughout the coupled network. A variational auto-encoder (Kingma, 2013) adds a probabilistic parametrization of the latent space, $\mathbb{R}^d$, such that each forward pass of the network samples a fresh $\epsilon \sim \mathcal{N}(0, I_d)$ and constructs a probabilistic latent vector $z = \mu + \sigma\epsilon$. The VAE trains via reconstruction loss and an added Kullback-Leibler divergence term, computed via the reparameterization trick (Kingma, 2013). Being arguably the simplest generative setting, Gerstgrasser et al. (2024) find that the accumulation scheme mitigates model collapse.

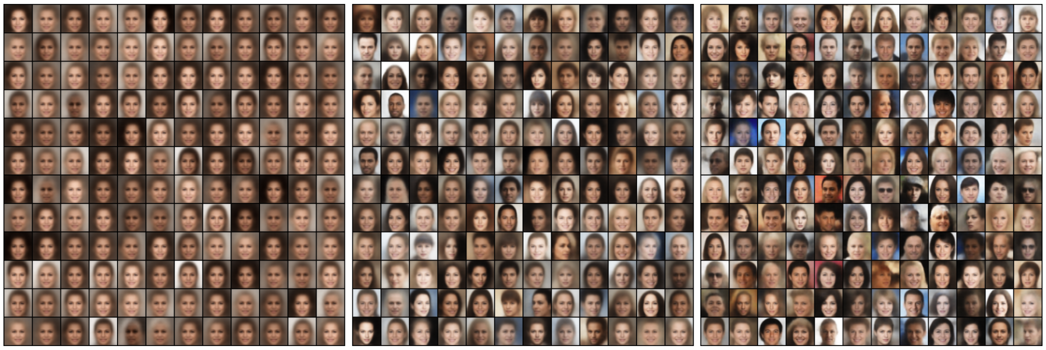
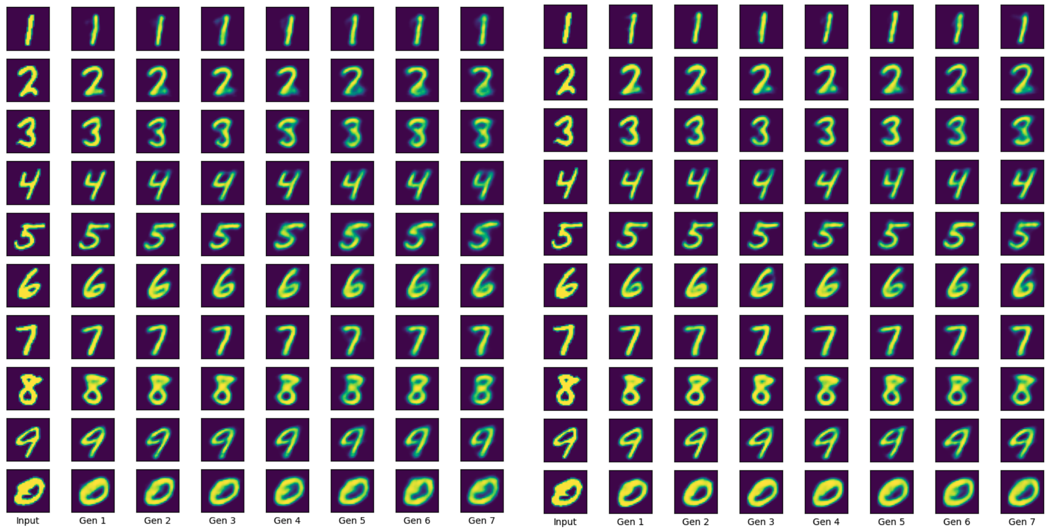


Figure 8: VAE Degradation: Replacement (Left), Accumulation (Middle), Baseline (Right)

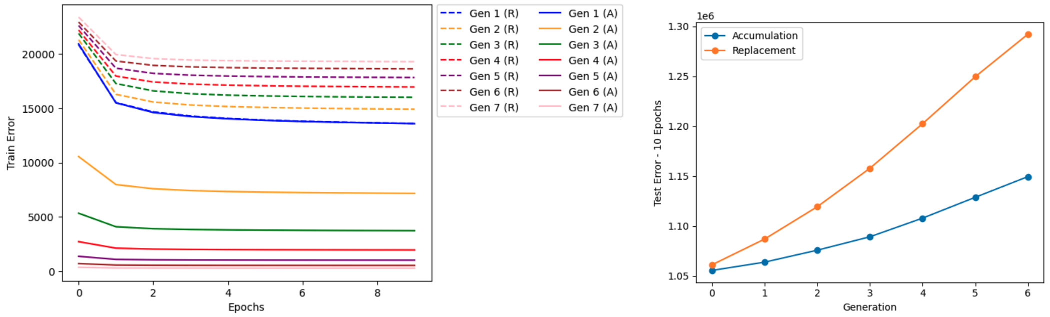
Compared to the complete mode collapse in the replacement scheme, Figure 8 shows that the accumulation strategy considerably maintains diversity, though it still suffers degradation with respect to a baseline VAE training run. Gerstgrasser et al. (2024) find that the empirical test loss under the accumulation strategy still increases approximately linearly, though at a much slower rate than that of replacement. We confirm this with our own VAE simulations on the MNIST dataset. We simulate VAE, with both data schemes, across 7 generations of iterative training with 10 training epochs per generation. Figure 9 compares the visual degradation of each handwritten digit over generation for each data scheme, and Figure 10 records the training and test loss for each scheme across each iterative generation. We also simulate an AE via a similar configuration, with Figure 11 recording the respective training and test loss.



(a) Replacement

(b) Accumulation

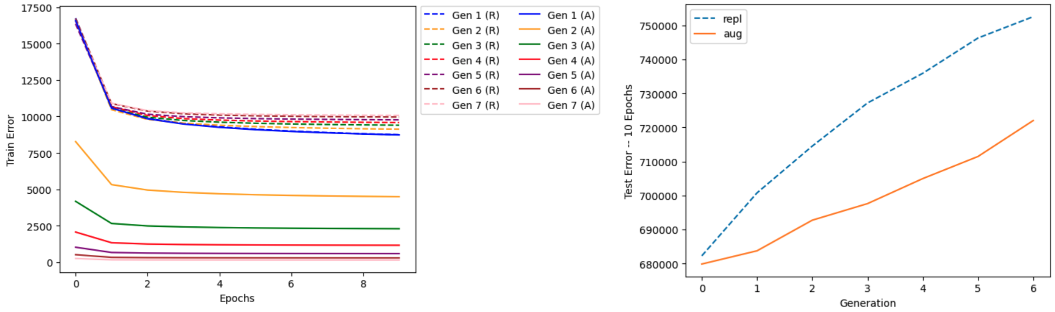
Figure 9: Visual Degradation of VAE on MNIST across Iterative Training



(a) Train Loss

(b) Test Loss

Figure 10: Training and Test Loss of VAE across Iterative Generations



(a) Train Loss

(b) Test Loss

Figure 11: Training and Test Loss of AE across Iterative Generations

Both Figures 10b and 11b confirm the hypothesis of Gerstgrasser et al. (2024) that accumulation does not prevent collapse in such settings, but merely slows it. Interestingly, we also find that in both Figures 10a and 11a, training error increases across generation in the replacement scheme while it decreases across generation in the accumulation scheme. We have yet to find an intellectually satisfying resolution to this observation.

4.3 Universality

While model collapse does not seem to hold in the generative image setting, Dey and Donoho (2024) have recently universalized the $\pi^2/6$ pathway from (6) to the logistic model via proof of the following general tool. First, consider the following class of iterative fitting algorithms.

Algorithm 1 Iterative Model Training by Synthetic Data Augmentation (Dey and Donoho, 2024)

```
Require: Positive integers $d_X, d_Y, d_\eta, d_\Theta$; parametric generative probability model $\{p(\cdot|\eta) : \eta \in \mathbb{R}^{d_\eta}\}$ defined on $\mathbb{R}^{d_Y}$; function $\eta : \mathbb{R}^{d_X} \times \mathbb{R}^{d_\Theta} \to \mathbb{R}^{d_\eta}$.
Start with a dataset $\mathcal{Z}_1 = \{(X_{1,1}, Y_{1,1}), \cdots, (X_{1,n}, Y_{1,n})\}$ where $X_{1,i} \in \mathbb{R}^{d_X}$ and $Y_{1,i} \in \mathbb{R}^{d_Y}$ for each $1 \leq i \leq n$.
for each generation $G \geq 1$ do
    Estimate $\hat{\theta}_G$ from $\mathcal{Z}_G$.
    Generate $\mathcal{X}_{G+1} = \{X_{G+1,1}, \cdots, X_{G+1,n}\}$.
    Generate new $\mathcal{Y}_{G+1} = \{Y_{G+1,1}, \cdots, Y_{G+1,n}\}$ with $Y_{G+1,i} \sim p(\cdot|\eta(X_{G+1,i}, \hat{\theta}_G))$ independently for each $1 \leq i \leq n$.
    Set $\mathcal{D}_{G+1} = \{(X_{G+1,1}, Y_{G+1,1}), \cdots, (X_{G+1,n}, Y_{G+1,n})\}$.
    Augment the existing data corpus with the newly generated data: $\mathcal{Z}_{G+1} = \mathcal{Z}_G \cup \mathcal{D}_{G+1}$.
end for
```

With respect to Algorithm 1, Gerstgrasser et al. (2024)'s original linear regression setting is simply a special case with a particular $p(\cdot|\eta)$ and an estimator $\hat{\theta}_G$ that weights all members of $\mathcal{Z}_G$ equally. Consider any iterative model training algorithm of the form in Algorithm 1 that satisfies the following assumptions.

1. Suppose that $X_{G,1}, \ldots, X_{G,n} \overset{\text{i.i.d.}}{\sim} H$ such that H is free of G and has all finite moments.
2. Suppose that for each $G \geq 1$, given feature $X_{G,i}$, which may be some transformation of the raw feature vector, the response-generating distribution $p(\cdot|\eta(X_{G,i}, \theta))$ comes from the exponential family with natural parameter $\eta(X_{G,i}, \theta) \equiv X_{G,i}\theta$ and sufficient statistic $T(\cdot)$ such that
$$p(\cdot|\eta(X_{G,i}, \theta)) \equiv \exp(\theta^\top X_{G,i}^\top T(y) - A(X_{G,i}\theta))h(y)$$
and A satisfies certain regularity conditions.
3. Suppose that for each $G \geq 1$, $\hat{\theta}_G$ is asymptotically approximately linear (AAL).

Then, define the following.

$$W_{n,T}(G) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{G,i}^\top (T(Y_{G,i}) - \nabla A(X_{G,i}\theta_0)), \quad W_{n,\Theta}(G) = \sqrt{n}(\hat{\theta}_G - \theta_0) \tag{7}$$

Consider a reference distribution $\mathbb{P}^{ref}$ such that $X_{G,i} \overset{i.i.d.}{\sim} H$ and $Y_{G,i}|X_{G,i} \sim p(\cdot|\eta(X_{G,i}, \theta_0))$ under $\mathbb{P}^{ref}$. Then, under $\mathbb{P}^{ref}$,

$$(W_{n,T}(g), W_{n,\Theta}(g))_{g=1}^G \xrightarrow{d} (W_T^{ref}(g), W_\Theta^{ref}(g))_{g=1}^G \tag{8}$$

where $(W_T^{ref}(g), W_\Theta^{ref}(g))_{g=1}^G$ are jointly mean zero Gaussians. Further,

$$(W_{n,T}(g), W_{n,\Theta}(g))_{g=1}^G \xrightarrow{d} (W_T(g), W_\Theta(g))_{g=1}^G \tag{9}$$

where $(W_T(g), W_\Theta(g))_{g=1}^G$ is a sequential Gaussian process such that

$$\begin{aligned} &W_T(g) = W_T^{ref}(g) = \mathbb{E}_0[X^\top \nabla^2 A(X\theta_0)X]W_\Theta(g-1) \\ &W_\Theta(g) \sim \mathbb{P}^{ref}(\cdot|W_T^{ref}(g) = W_T(g) \cdots W_T^{ref}(i) = W_T(i) \cdots W_T^{ref}(1) = W_T(1)) \end{aligned} \tag{10}$$

In many machine learning settings, the conditions described above are broad enough to relate a iterative model training algorithm of interest to an analogous Gaussian process, which Dey and Donoho (2024) argues is easier to analyze. With this tool, Dey and Donoho (2024) re-derive (5) and (6) while generalizing the $\pi^2/6$ bound to the logistic model with limited additional effort. The authors also support the claim that the variance ratio for the accumulation method outperforms the subsample method which outperforms the replacement method, as aforementioned in Figure 6a.

As for the generative setting, while the classical AE yields a response-generating distribution in the exponential family, a similar statement for a VAE in full generality is nontrivial given the introduction of a fresh $\epsilon \sim \mathcal{N}(0, 1)$ in the forward pass. Conditions aside, it is also unclear whether the process in (10) can be manipulated to achieve test error bounds in the generative case. Further, even if convergence is proven via this universality, by Figures 10 and 11, it may be the case that constant factors make model collapse detrimental even if its effects are bounded in the generational limit.

5 Discussion

While the proof and technical details of the above universality tool are beyond the scope of this work, Dey and Donoho (2024) published a pre-print of this technique on October 30th, 2024, which was mere days after this manuscript was underway! We hope that this recent development in the field inspires future work on the boundaries of model collapse.

We acknowledge that our empirical study is limited due to computational constraints, particularly in the generative setting. For the accumulation scheme, each generation doubles the available training data, which for Figure 9, become unsustainable beyond the 7-th generation on a single MacBook Pro in a reasonable timescale. Parties with additional computational resources are encouraged to explore concrete, experimental generative models beyond auto-encoders. While (Dohmatob, Feng, Subramonian, et al., 2024) show model collapse in the self-consuming setting for large language models under certain regimes, exploring this setting under accumulation as well as the extent of the universality claim of Dey and Donoho (2024) is a worthy pursuit in the authors eyes.

While most of the accumulation strategies discussed treat each piece of data equally, recent work on faulty verifiers for distinguishing real and synthetic data have shown promise in substantively improving test performance (Feng et al., 2024). Along these lines, synthetic watermarking (Wenger, 2024) has also demonstrated efficacy. However, watermarks are unstable and removable (Y. Hu et al., 2024), and implementing a universal watermarking scheme across company lines will likely be a logistical nightmare.

Regardless, any combination of these ideas is a relevant subject of activate research, and we look forward to future work on mitigating data pollution to ensure a welcoming environment for future models in training.

6 Acknowledgement

We would like to thank Professor Yue Lu and TF Weiyu Li for their support. While this work began as a final project for Harvard's AM 254, it has evolved into a rewarding intellectual exploration. Source code for this work can be made available upon request, and we are looking forward to continued study of this curious phenomenon.

References

[1] Sina Alemohammad et al. “Self-consuming generative models go mad”. In: *arXiv preprint arXiv:2307.01850* (2023).

[2] Francis Bach. “High-dimensional analysis of double descent for linear regression with random projections”. In: *SIAM Journal on Mathematics of Data Science* 6.1 (2024), pp. 26–50.

[3] Hao Chen et al. “On the Diversity of Synthetic Data and its Impact on Training Large Language Models”. In: *arXiv preprint arXiv:2410.15226* (2024).

[4] William G Cochran. “The distribution of quadratic forms in a normal system, with applications to the analysis of covariance”. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 30. 2. Cambridge University Press. 1934, pp. 178–191.

[5] Apratim Dey and David Donoho. “Universality of the $\pi^2/6$ Pathway in Avoiding Model Collapse”. In: *arXiv preprint arXiv:2410.22812* (2024).

[6] Ruixue Ding et al. “Mgeo: Multi-modal geographic language model pre-training”. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2023, pp. 185–194.

[7] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. “Model collapse demystified: The case of regression”. In: *arXiv preprint arXiv:2402.07712* (2024).

[8] Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, et al. “Strong model collapse”. In: *arXiv preprint arXiv:2410.04840* (2024).

[9] Yunzhen Feng et al. “Beyond Model Collapse: Scaling Up with Synthesized Data Requires Reinforcement”. In: *arXiv preprint arXiv:2406.07515* (2024).

[10] Adrian Fischer, Robert E Gaunt, and Andrey Sarantsev. “The variance-gamma distribution: A review”. In: *arXiv preprint arXiv:2303.05615* (2023).

[11] Matthias Gerstgrasser et al. “Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data”. In: *arXiv preprint arXiv:2404.01413* (2024).

[12] Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. “ChatGPT is not all you need. A State of the Art Review of large Generative AI models”. In: *arXiv preprint arXiv:2301.04655* (2023).

[13] Trevor Hastie et al. “Surprises in high-dimensional ridgeless least squares interpolation”. In: *Annals of statistics* 50.2 (2022), p. 949.

[14] Yuepeng Hu et al. “Stable Signature is Unstable: Removing Image Watermark from Diffusion Models”. In: *arXiv preprint arXiv:2405.07145* (2024).

[15] Diederik P Kingma. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[16] Salman Mohamadi et al. “ChatGPT in the age of generative AI and large language models: a concise survey”. In: *arXiv preprint arXiv:2307.04251* (2023).

[17] Robin Rombach et al. “High-resolution image synthesis with latent diffusion models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[18] Ilia Shumailov et al. “AI models collapse when trained on recursively generated data”. In: *Nature* 631.8022 (2024), pp. 755–759.

[19] Hugo Touvron et al. “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288* (2023).

[20] Pablo Villalobos et al. “Will we run out of data? an analysis of the limits of scaling datasets in machine learning”. In: *arXiv preprint arXiv:2211.04325* 1 (2022).

[21] Xiaofeng Wang et al. “Drivedreamer: Towards real-world-driven world models for autonomous driving”. In: *arXiv preprint arXiv:2309.09777* (2023).

[22] Alexander Wei, Wei Hu, and Jacob Steinhardt. “More than a toy: Random matrix models predict how real-world neural representations generalize”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 23549–23588.

[23] Emily Wenger. *AI produces gibberish when trained on too much AI-generated data*. 2024.

[24] Susan Zhang et al. “Opt: Open pre-trained transformer language models”. In: *arXiv preprint arXiv:2205.01068* (2022).