Model Collapse

Presented by Xuyang Chen & Xianglong Hou

- **AI Models Collapse When Trained on Recursively Generated Data**

  Shumailov et al. (2023)

- **Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data**

  Gerstgrasser et al. (2024)

- **Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World**

  Kazdan et al. (2024)

- **Universality of the $\pi^2/6$ Pathway in Avoiding Model Collapse**

  Dey et al. (2024)

AI Models Collapse When Trained on Recursively Generated Data

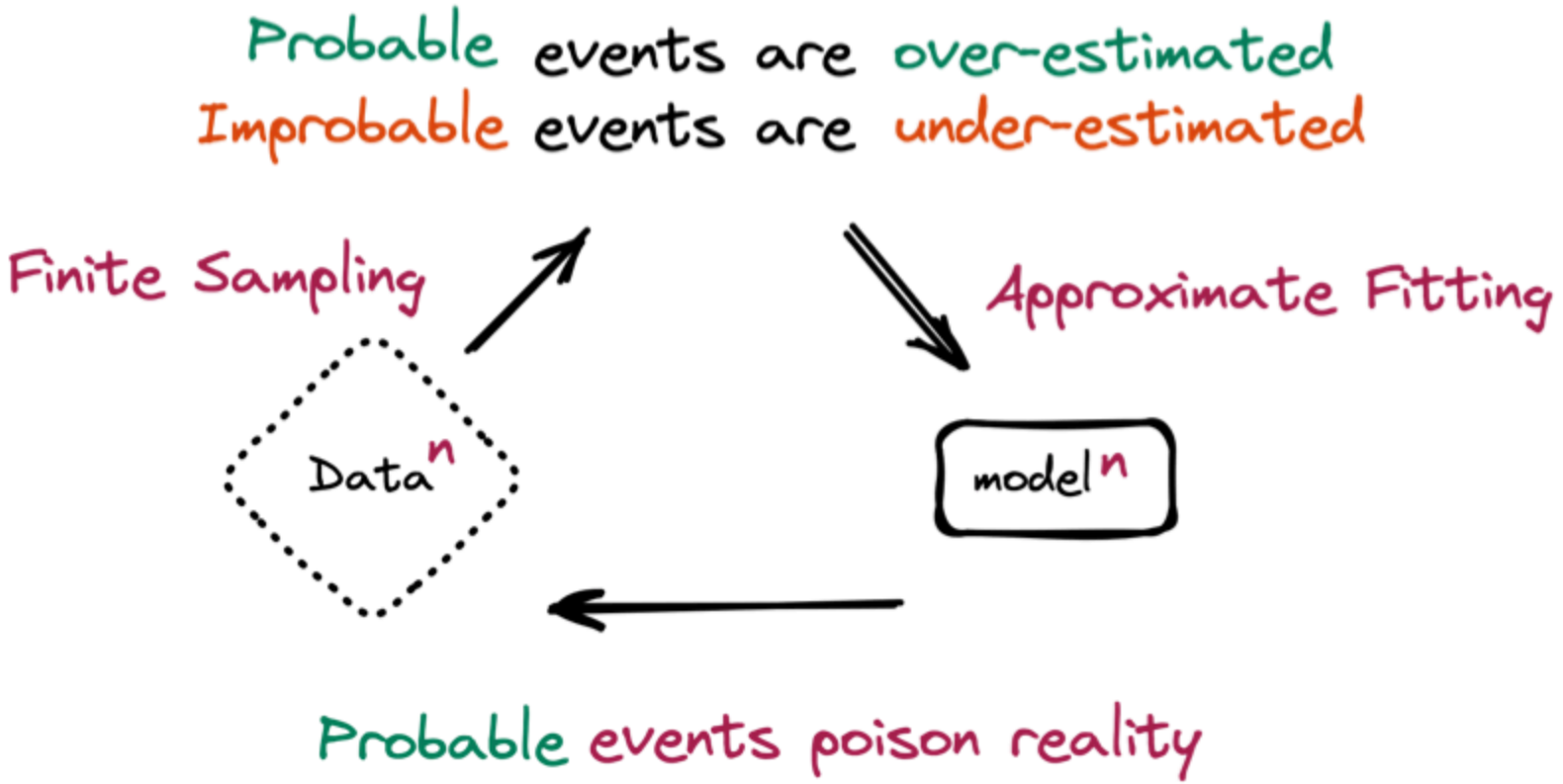
Shumailov et al. (2023)

Background

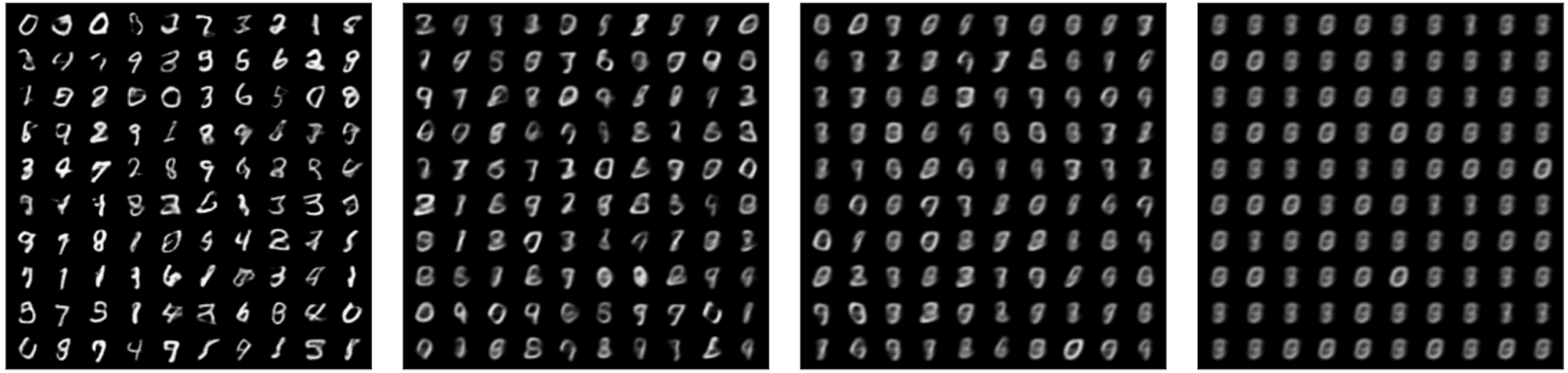
- Generative models are trained on vast amounts of diverse data from the internet
- A large amount of AI generated content has been mixed into human generated data.
- GPT-2, GPT-3, GPT-4, $\cdots$, GPT-n ?

What is Model Collapse?

- A degenerative process in which models trained on their own synthetic outputs.
- Overestimate probable events and underestimate imporbable events.



Example: Variational Autoencoders



(a) Original model (b) Generation 5 (c) Generation 10 (d) Generation 20

Figure 9: Random latent reconstructions from VAEs. No training data comes from the original distribution. Over the generations, different modes of the original distribution get entangled and generated data starts looking unimodal.

Example: Gaussian Mixture Models

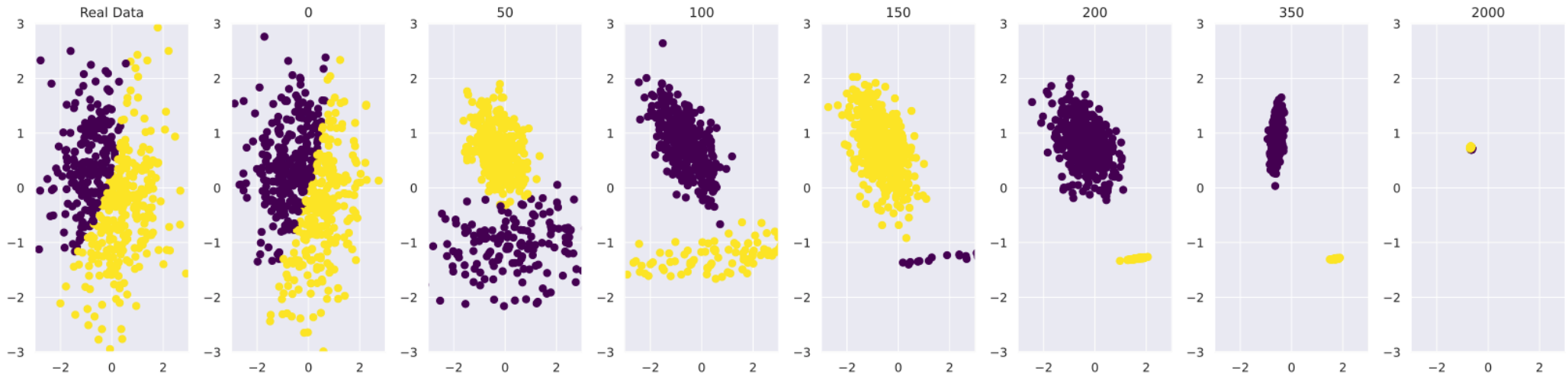
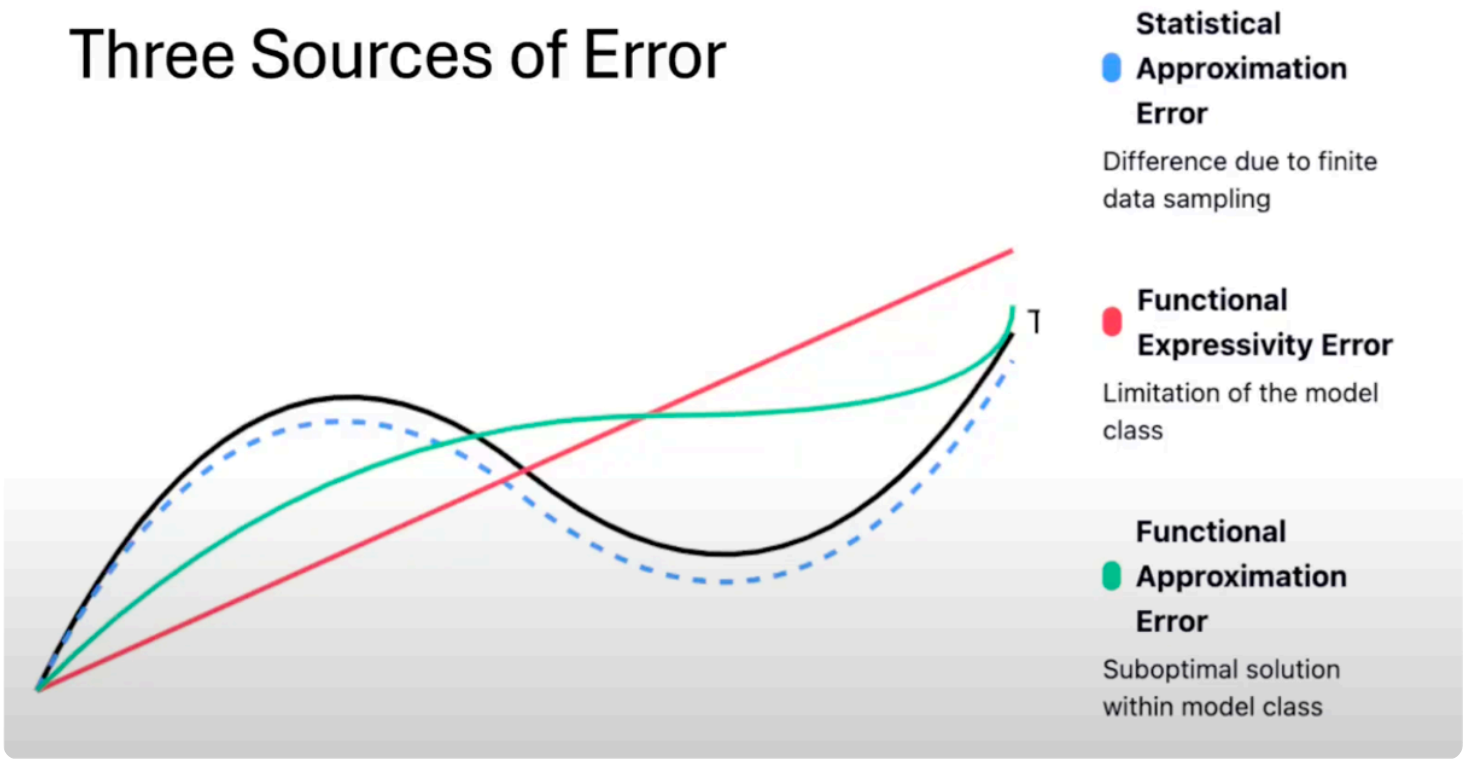


Figure 7: An examples of GMM fitting data at iterations $\{0, 50, 100, 150, 200, 350, 2000\}$. At first the model fits data very well as is shown on the left; yet even at generation 50 the perception of the underlying distribution completely changes. At generation 2000 it converges to a state with very little variance. GMM is sampled a thousand times.

Analysis: 3 types of error

- Statistical approximation error
- Functional expressivity error
- Functional approximation error



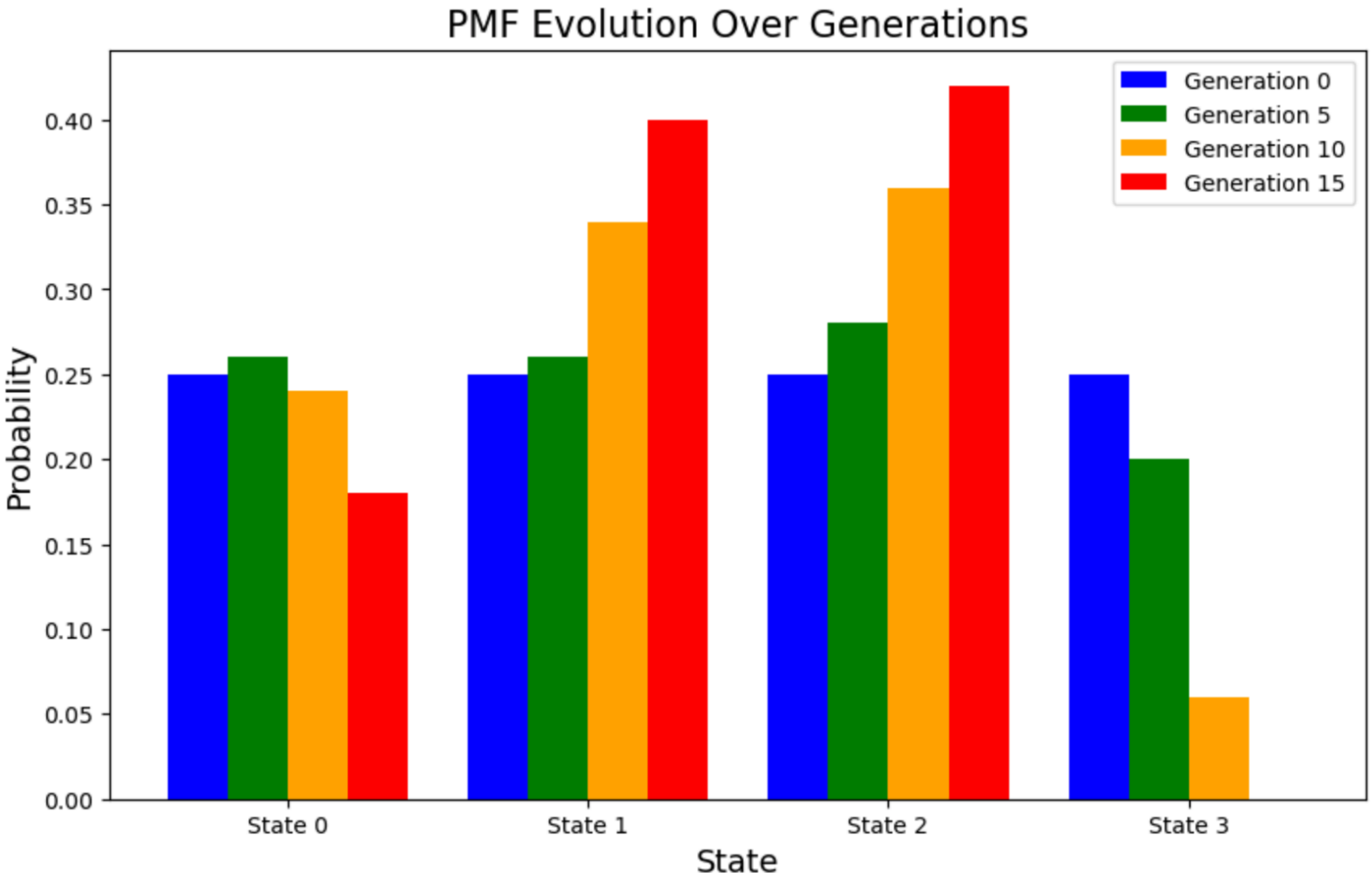
Theoretical intuition 1: Discrete distributions with exact approximation

- **Setup**:
  Assume $\mathbf{X_i}$ is the dataset generated by model $\mathcal{F}$ at generation step i, and the approximation approximation is perfect—i.e., $\mathcal{F}(p) = p$.
- **Recursive Process**:

$$\mathbf{X}_i \quad \xrightarrow{\mathcal{F}} \quad p_{i+1} \quad \xrightarrow{\text{sampling}} \quad \mathbf{X}_{i+1}$$

- **Convergence to Absorbing States**:
  The Markov chain will converge to the absorbing states ($P(s \to s) = 1$). In our problem, these absorbing states correspond to delta functions, meaning the distribution collapses to a single point.

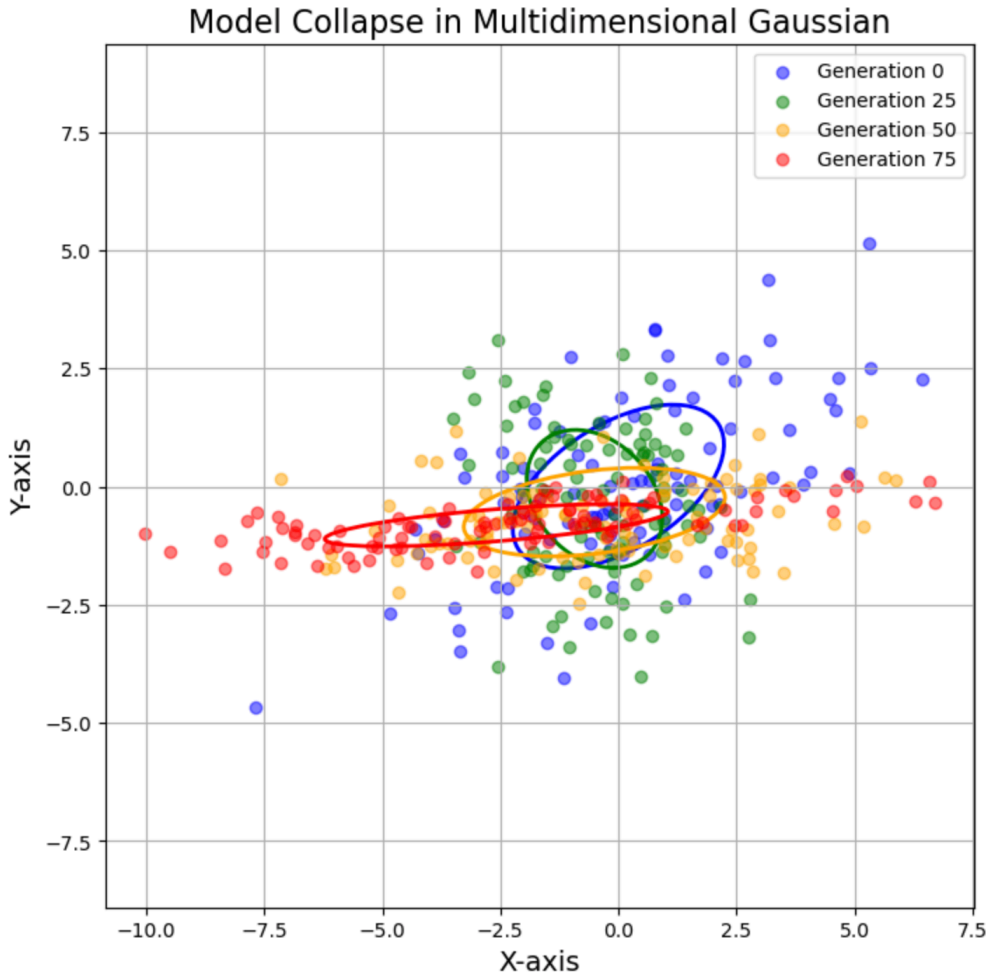
Theoretical intuition 1: Discrete distributions with exact approximation



Theoretical intuition 2: Multidimensional Gaussian

- **Setup**:
  At Each generation step n, we fits a Gaussian $\mathcal{N}(\mu_n, \Sigma_n)$ using the unbiased sample mean and covariance of $\mathbf{X^{n-1}}$. Then $\mathbf{X^n}$ is sampled from $\mathcal{N}(\mu_n, \Sigma_n)$.
- **Results**:
  - $$\mathbb{E}[\mathbb{W}_2^2(\mathcal{N}(\mu_n, \Sigma_n), \mathcal{D}_0)] \to \infty$$
  - $$\Sigma_n \overset{\text{a.s.}}{\to} 0 \text{ as } n \to \infty.$$

Theoretical intuition 2: Multidimensional Gaussian



Model collapse in language models: Setup

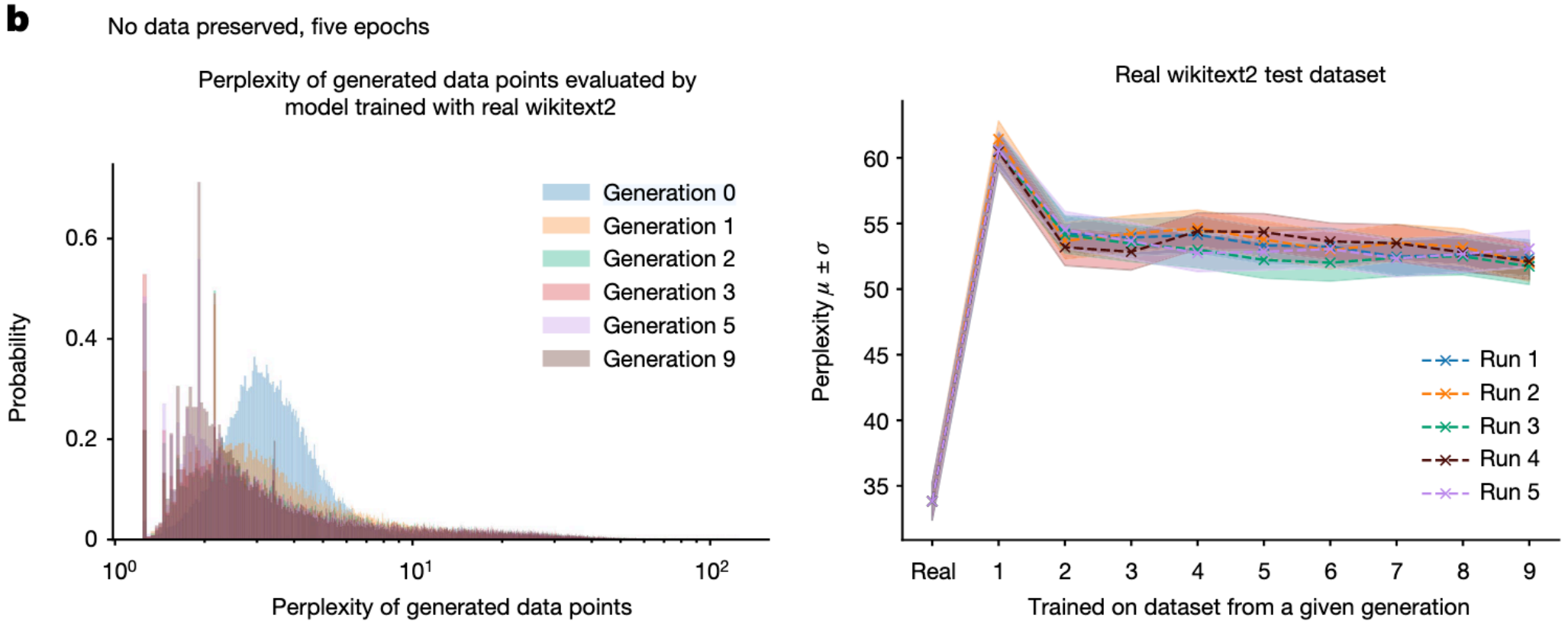
- **Base model**: OPT-125m,
- **Dataset**: wikitext2 (64 tokens $\rightarrow$ 64 tokens)
- **Training**: sequentially fine-tuned on synthetic data generated by a previous generation

  No Original Data vs. 10% Original Data Preserved
- **Metrics**: perplexity(PP): how "confused" the model is

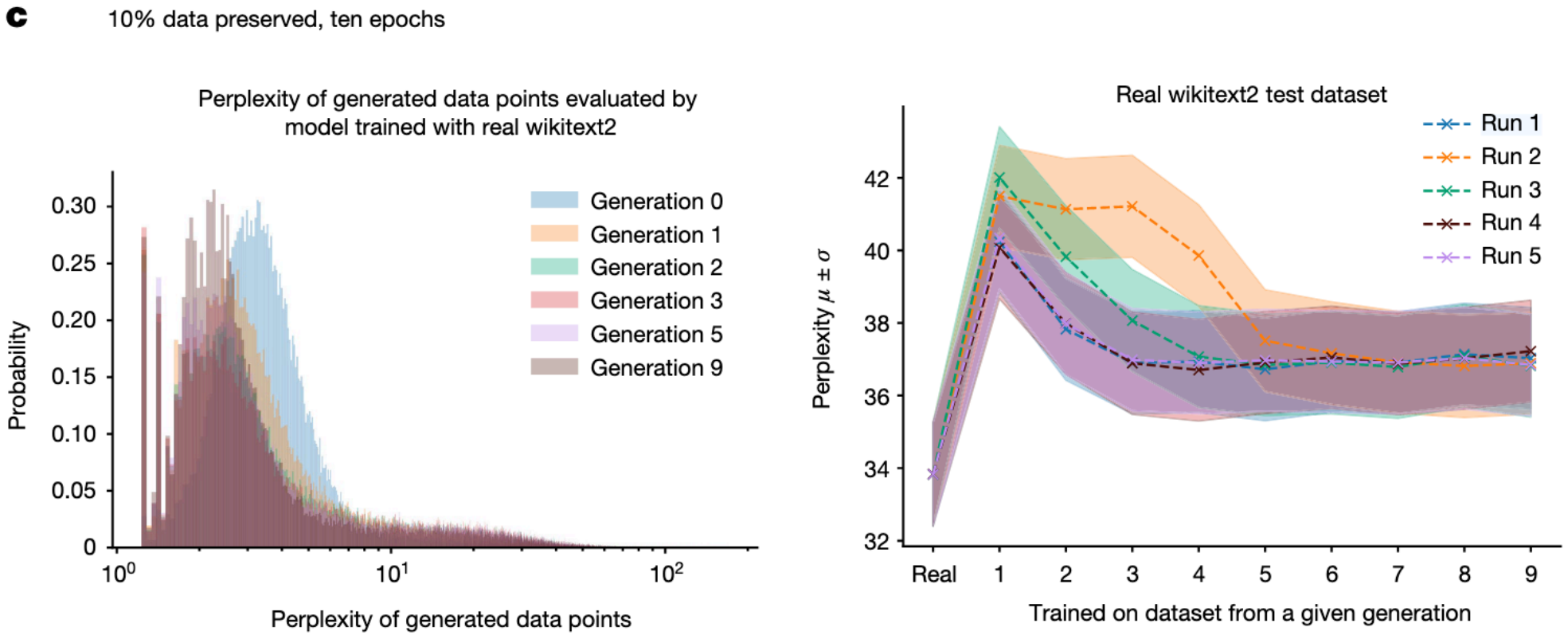
$$L = -\frac{1}{N}\sum_{t=1}^{N}\log p(y_t \mid x, y_{<t})$$

$$\mathrm{PP} = \exp(L)$$

Model collapse in language models: Results (No data preserved)



Model collapse in language models: Results (10% data preserved)



Model collapse in language models: Conclusion

- Model tends to produce samples that the original model trained with real data is more likely to produce
- Later generations start producing samples that would never be produced by the original model

Model collapse in language models: Example

Example 1. Example of text outputs of an OPT-125m model affected by model collapse—models degrade over generations, for which each new generation is trained on data produced by the previous generation.

- **Input:** some started before 1360 — was typically accomplished by a master mason and a small team of itinerant masons, supplemented by local parish labourers, according to Poyntz Wright. But other authors reject this model, suggesting instead that leading architects designed the parish church towers based on early examples of Perpendicular.
- **Outputs:**
- **Gen 0:** Revival architecture such as St. John's Cathedral in London. The earliest surviving example of Perpendicular Revival architecture is found in the 18th @-@ century Church of Our Lady of Guernsey, which dates from the late 19th century. There are two types of perpendicular churches : those.
- **Gen 1:** architecture such as St. Peter's Basilica in Rome or St. Peter's Basilica in Buenos Aires. There is no evidence that any of these buildings were built during the reign of Pope Innocent III, but it is possible that they may have been built during the reign of his successor, Pope Innocent.

- **Gen 5:** ism, which had been translated into more than 100 languages including English, French, German, Italian, Spanish, Portuguese, Dutch, Swedish, Norwegian, Polish, Hungarian, Slovak, Lithuanian, Estonian, Finnish, Romanian, Bulgarian, Turkish, Croatian, Serbian, Ukrainian, Russian, Kazakh, Kyrgyz.
- **Gen 9:** architecture. In addition to being home to some of the world's largest populations of black @-@ tailed jackrabbits, white @-@ tailed jackrabbits, blue @-@ tailed jackrabbits, red @-@ tailed jackrabbits, yellow @-.

Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data

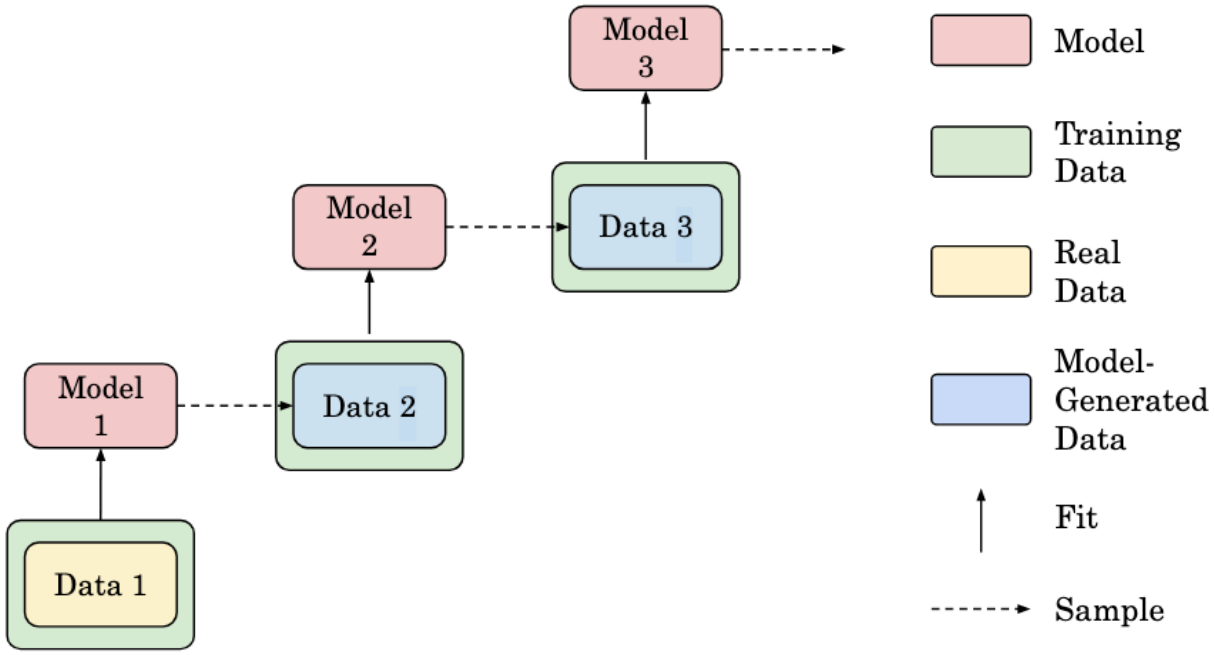
Gerstgrasser et al. (2024)

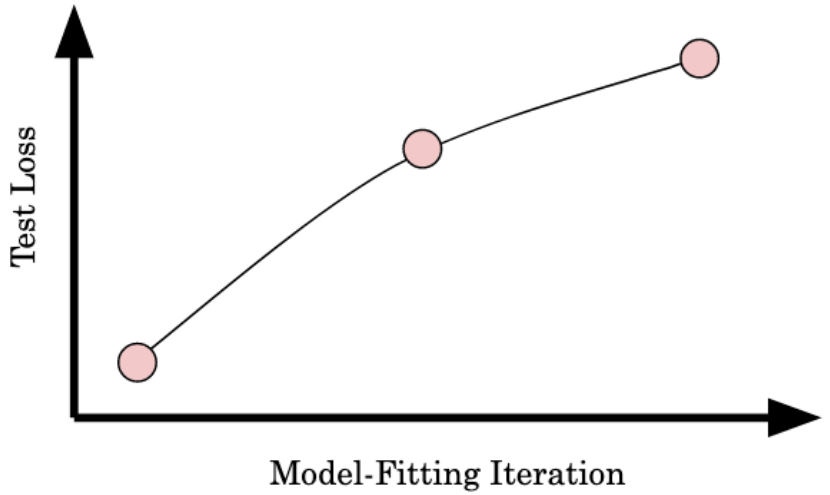
Avoiding model collapse

- Test error has a finite upper bound over such iterations.

Two Settings to Study Model Collapse

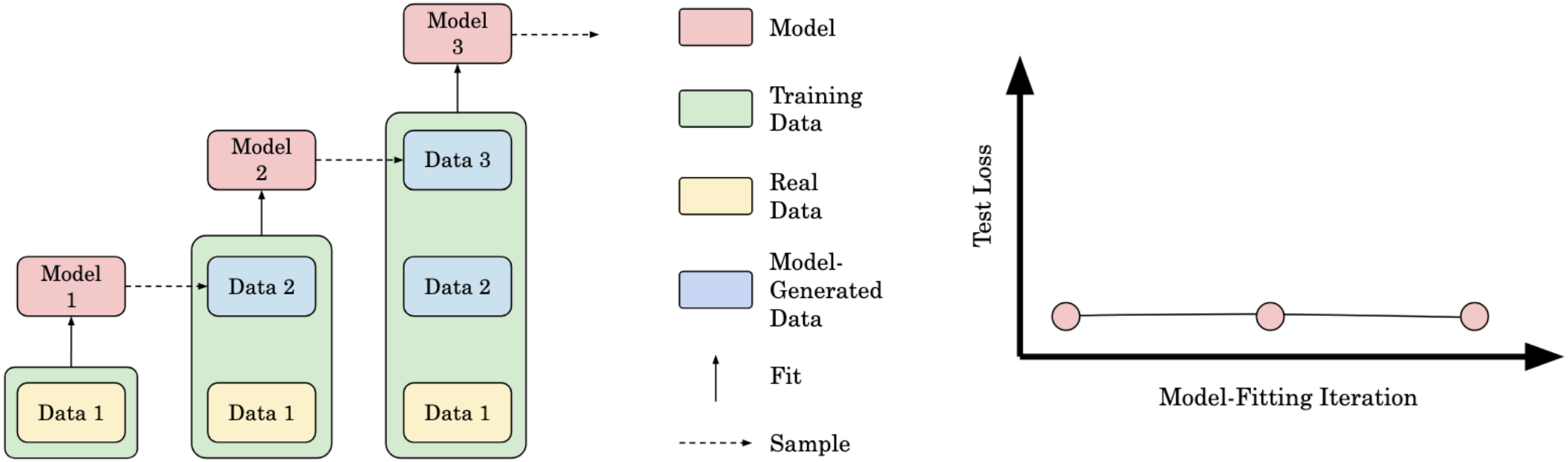
- Replace data (model's generated data replaces previous data)





Two Settings to Study Model Collapse

- Accumulate data



Accumulating Data Avoids Model Collapse in Deep Generative Models

- Transformer-Based Causal Language Modeling
- Diffusion Models on Molecular Conformation Data
- Variational Autoencoders on Image Data

Transformer-Based Causal Language Modeling

- Models: GPT-2 (9M), Llama-2 (12M, 42M, 126M)
- Dataset: TinyStories (470M token GPT-3.5/4-generated dataset of short stories at a kindergarten reading level)
- Loss: Cross Entropy

Transformer-Based Causal Language Modeling

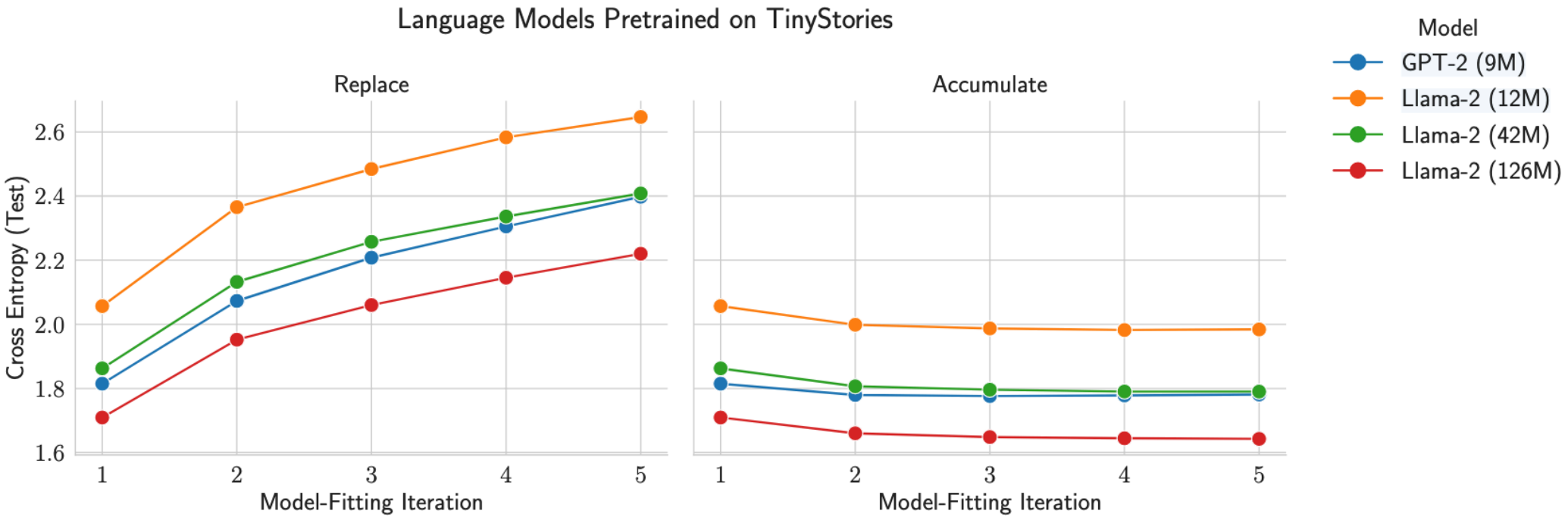


Figure 2: **Data Accumulation Avoids Model Collapse in Language Modeling.** Sequences of causal transformer-based language models are pretrained on TinyStories (Eldan & Li, 2023). Cross-entropy validation loss increases when replacing data (left), but not when accumulating data (right). Synthetic data was sampled with temperature = 1.0.

Transformer-Based Causal Language Modeling

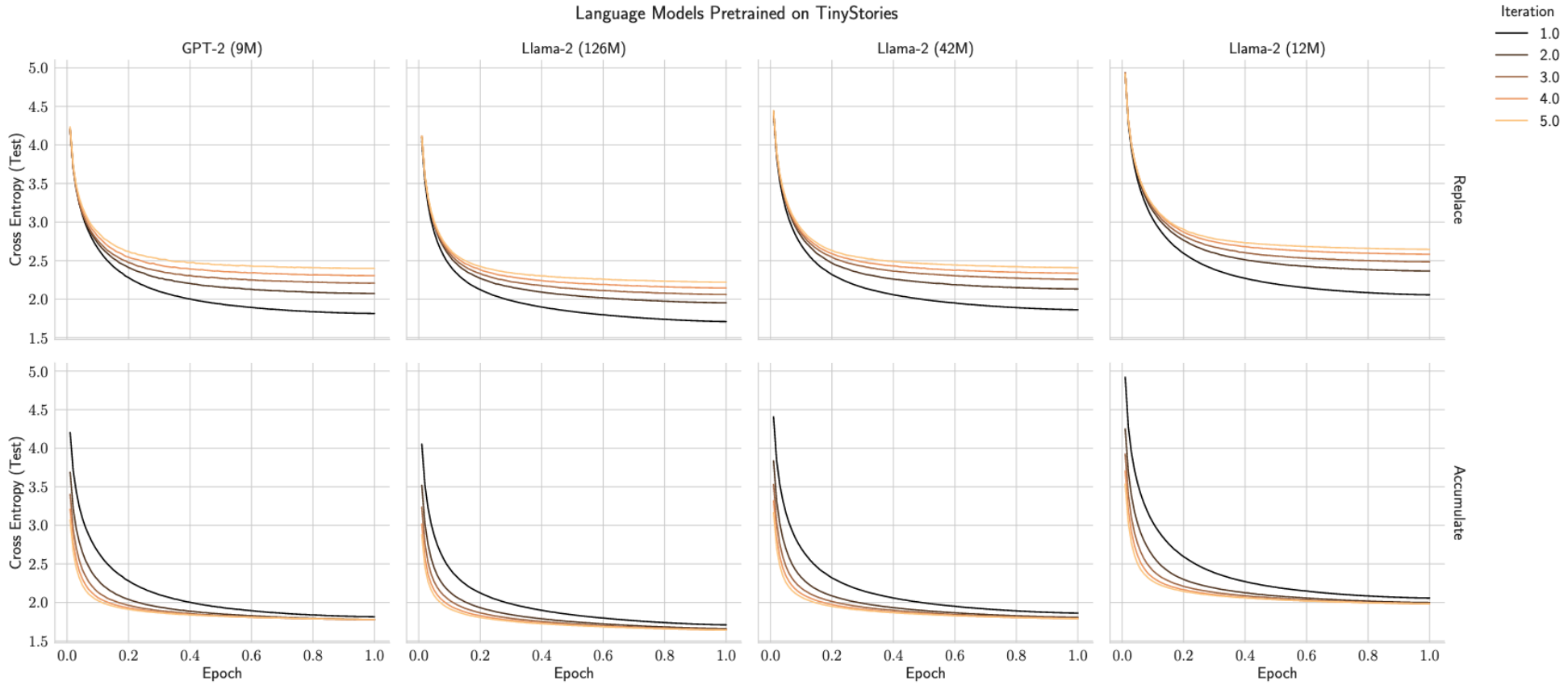


Figure 3: **Data Accumulation Avoids Model Collapse in Language Modeling.** Learning curves for individual model-fitting iterations when repeatedly *replacing* data (top), and when *accumulating* data (bottom). Note: Epochs correspond to more gradient steps for accumulate than replace because the number of training data grows for accumulate.

Diffusion Models on Molecular Conformation Data

- Model: GeoDiff, a geometric diffusion model for molecular conformation generation
- Dataset: GEOM-Drugs (containing molecular structures found in drugs)
- Loss: a weighted variational lower bound to the conditional likelihood

Diffusion Models on Molecular Conformation Data

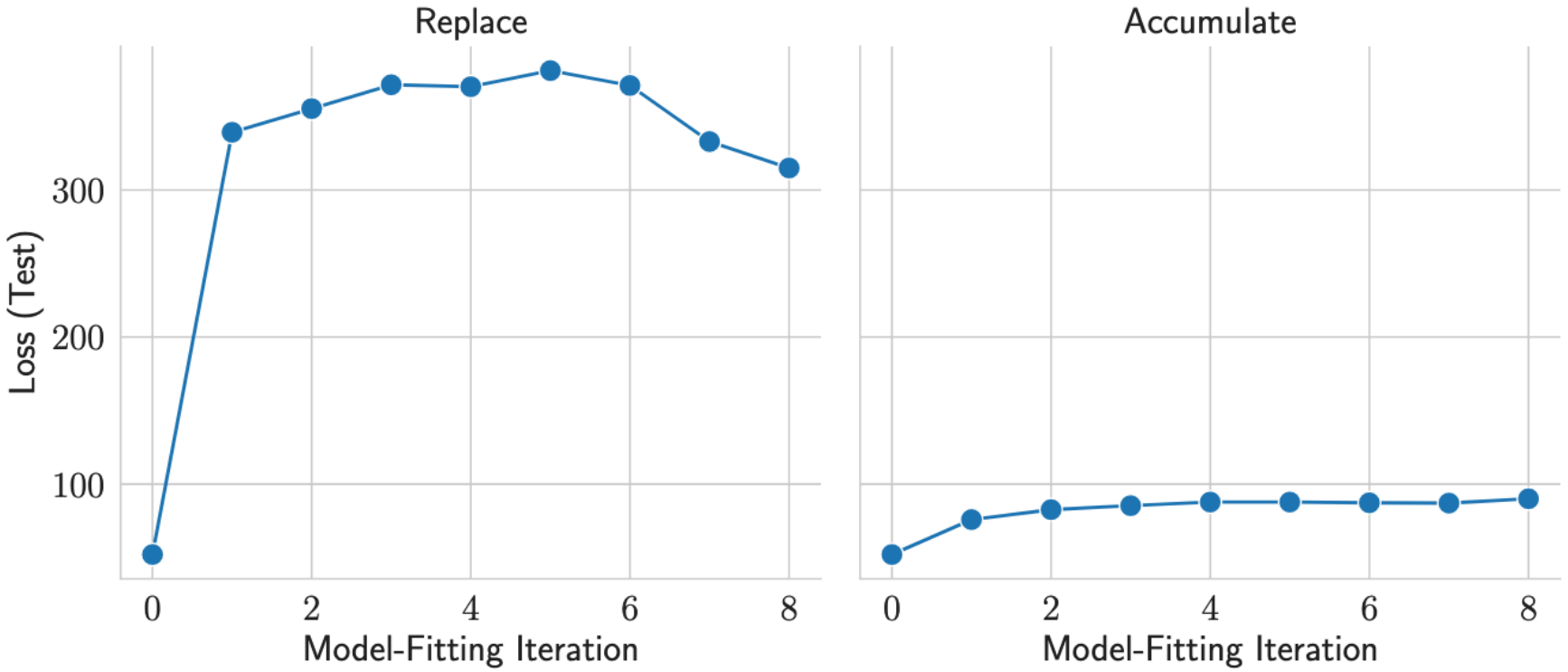


Figure 4: **Data Accumulation Avoids Model Collapse in Geometric Diffusion Modeling.** GeoDiff, a diffusion-based molecular conformation generation model, is trained on a subset of Drugs data containing molecular structures found in drugs. Test loss degrades when replacing data (left) but not when accumulating data (right).

Variational Autoencoders on Image Data

- Model: Variational Autoencoder (VAE)
- Dataset: CelebA (200K images of human faces)
- Loss: reconstruction error + KL divergence between the encoder's output Gaussian and the isotropic Gaussian prior.

Variational Autoencoders on Image Data

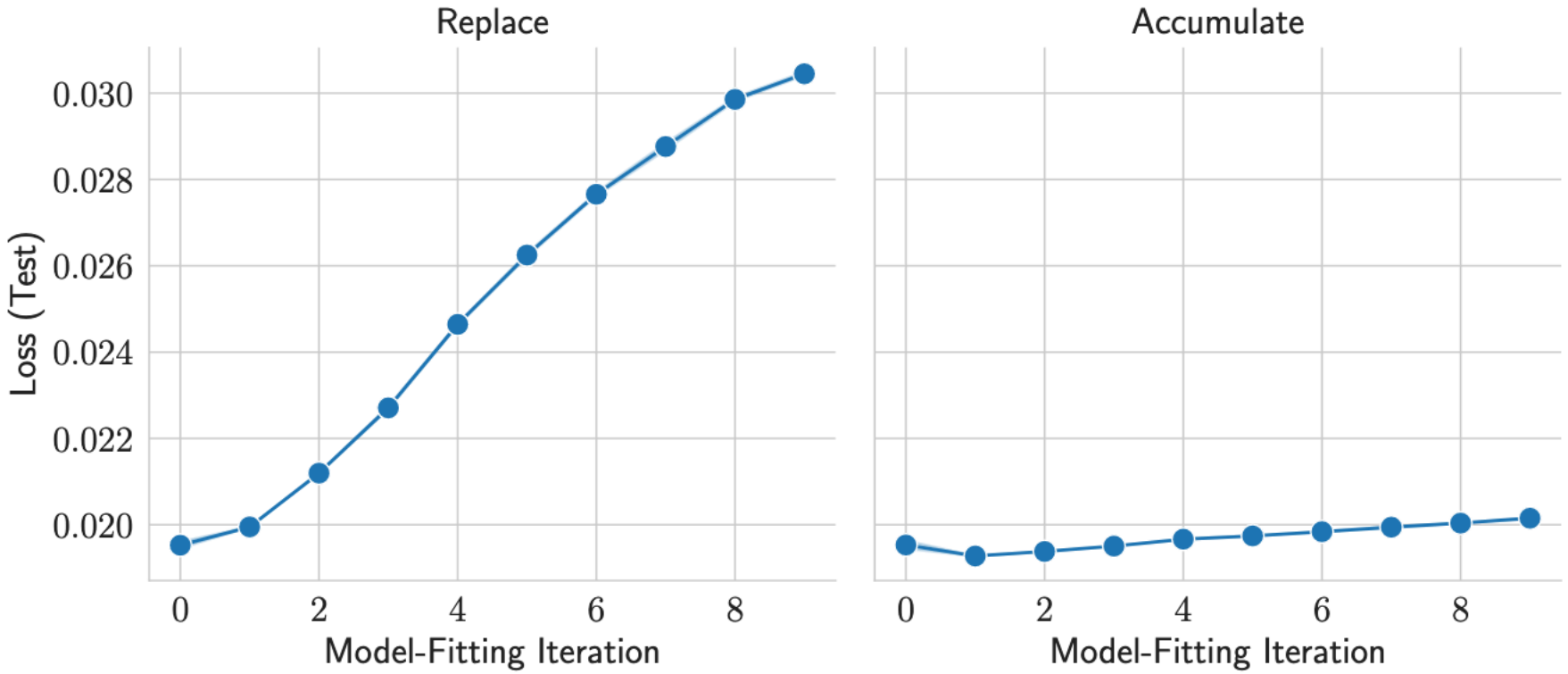


Figure 5: **Data Accumulation Avoids Model Collapse in Variational Autoencoders for Image Generation.** Sequences of variational autoencoders (VAEs) are trained on CelebA. Test loss degrades when replacing data (left) but not when accumulating data (right).

Variational Autoencoders on Image Data

Figure 6: **Sampled Images from** Left: Replacing data with data generated by the previous iteration's newly trained VAE yields lower quality and eventually leads to complete mode collapse. Middle: Accumulating data with data generated by the previous iteration's newly trained VAE preserves the quality and diversity of generated data across iterations. Right: Baseline samples after 100 training epochs on the dataset.

Accumulating Data Avoids Model Collapse in Linear Models

- Original data distribution $\mathbb{P}_{\Sigma,w^*,\sigma^2}$ over $\mathbb{R}^d \times \mathbb{R}$:
  - Input: $x \sim N(0, \Sigma) \in \mathbb{R}^d$
  - Noise: $\epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$
  - Label: $y = x^T w^* + \epsilon$
- Design matrix $X = (x_1, \cdots, x_T)^T \in \mathbb{R}^{T \times d}$.
- Also form Y, E
- X is full column rank ($T \geq d$)
- $E_{\text{test}}(w) \overset{\text{def}}{=} \mathbb{E}\left[(x_{\text{test}}^\top w - y_{\text{test}})^2\right] - \sigma^2 = \mathbb{E}\left[\|w - w^*\|_\Sigma^2\right].$

Synthetic Data Generation Process

- $P_{\Sigma,w^*,\sigma^2} \to P_{\Sigma,\hat{w}_1,\sigma^2} \to \cdots \to P_{\Sigma,\hat{w}_n,\sigma^2}$, $n \in \mathbb{N}$ is the number of iterations.
- For $n = 1$:
  - Accumulate Covariates/Features: $\tilde{X}_1 \overset{\text{def}}{=} X$
  - Accumulate Targets: $\tilde{Y}_1 \overset{\text{def}}{=} \hat{Y}_1 \overset{\text{def}}{=} Xw^* + E_1$, where $E_1 \overset{\text{def}}{=} E \sim \mathcal{N}(0, \sigma^2 I_T)$
  - Fit linear model: $\hat{w}_1 = \tilde{X}_1^+ \tilde{Y}_1$
  - Sample synthetic data for the next iteration: $\hat{Y}_2 \overset{\text{def}}{=} X\hat{w}_1 + E_2$, where $E_2 \sim \mathcal{N}(0, \sigma^2 I_T)$

Synthetic Data Generation Process

- For $n \geq 2$:
  - Accumulate data: $\tilde{X}_n^\top = [\tilde{X}_{n-1}^\top; X^\top] \in \mathbb{R}^{d \times nT}$,$\tilde{Y}_n^\top = [\tilde{Y}_{n-1}^\top; \hat{Y}_n^\top] \in \mathbb{R}^{1 \times nT}$.
  - Fit linear model: $\hat{w}_n \stackrel{\text{def}}{=} \tilde{X}_n^+ \tilde{Y}_n$
  - Sample synthetic data for the next iteration: $\hat{Y}_{n+1} \stackrel{\text{def}}{=} X\hat{w}_n + E_{n+1}$, where $E_{n+1} \sim \mathcal{N}(0, \sigma^2 I_T)$

Simulation

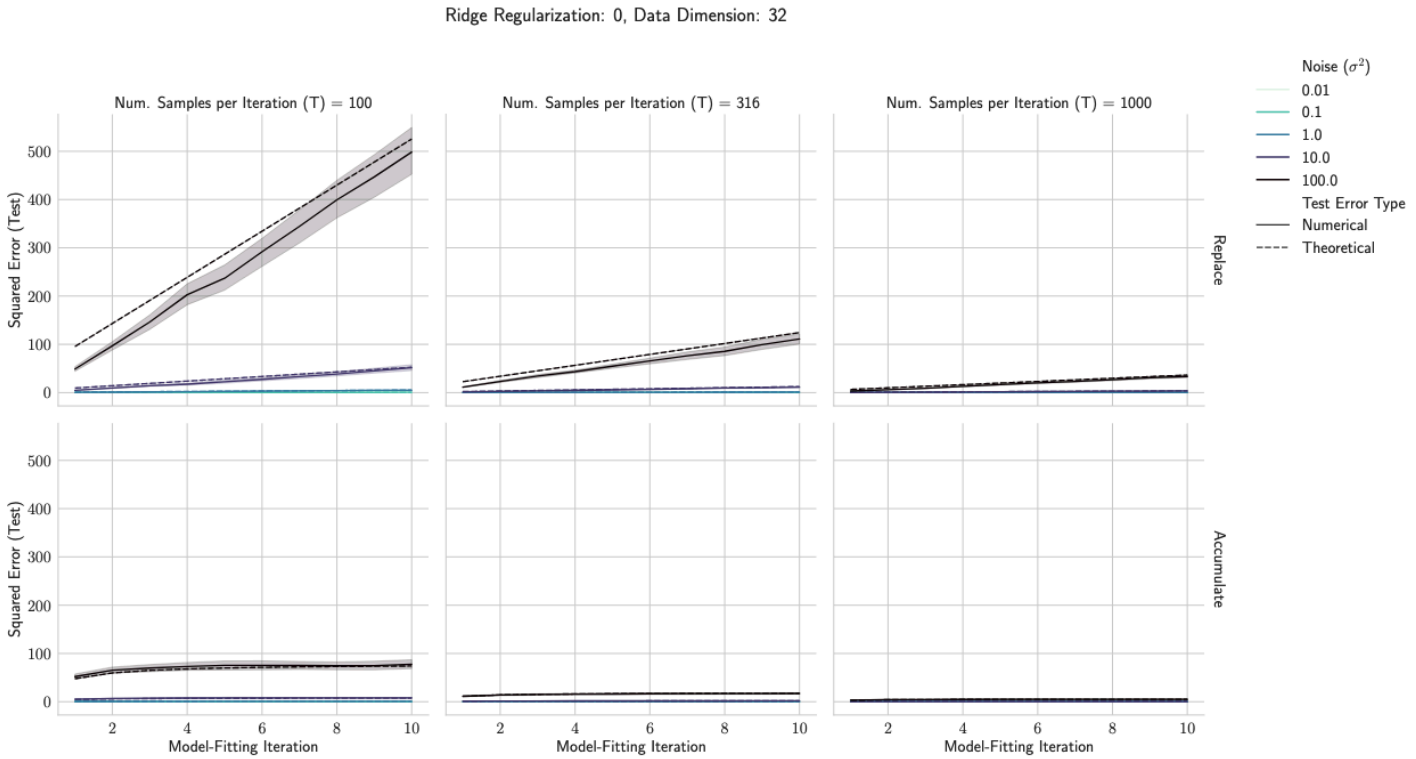


Figure 7: **Accumulating Data Avoids Model Collapse in Linear Regression.** We consider sequences of linear models recurrently fit to generated targets by previous iterations of models. Top: If each linear model is fit to the generated targets of *only* the preceding linear model, i.e., data are replaced, then the test error grows linearly with the number of iterations n. Bottom: If each linear model is instead fit to the generate targets of *all* the preceding linear models, i.e., data accumulate, then the test error has a finite upper bound independent of the number of iterations. This suggests that data accumulation might be a robust solution for mitigating model collapse. For log test error and higher iterations, see Appendix Fig. 16.

Theorem

1. In the data accumulation setting, $\forall n \geq 1, \hat{w}_n = w^* + (X^\top X)^{-1} X^\top \left(\sum_{i=1}^{n} \frac{E_i}{i} \right)$.
2. With $T \geq d + 2$ and isotropic features ($\Sigma \overset{\text{def}}{=} I_d$),
   $E_{\text{test}}^{\text{Accum}}(\hat{w}_n) = \frac{\sigma^2 d}{T-d-1} \left(\sum_{i=1}^{n} \frac{1}{i^2} \right) \leq \frac{\sigma^2 d}{T-d-1} \times \frac{\pi^2}{6}$.
3. $E_{\text{test}}^{\text{Replace}}(\hat{w}_n) = \frac{\sigma^2 d}{T-d-1} \times n$.

- Halfway: replace the previous dataset with a pure synthetic dataset of size iT at the i-th iteration.

4. $E_{\text{test}}^{\text{halfway}}(\hat{w}_n) \leq \frac{\sigma^2 d}{T-d-1} \times \log n$.

Proof

1. Mathematical Induction
2. $\mathbb{E}_X\left[\operatorname{tr}\left((X^\top X)^{-1}\right)\right] = \frac{d}{T-d-1}$ by Dohmatob et al. , then

$$
\begin{aligned}
E_{\text{test}}(\hat{w}_n) &= \mathbb{E}\left[\left(\sum_{i=1}^{n}\frac{E_i}{i}\right)^\top X(X^\top X)^{-2}X^\top\left(\sum_{i=1}^{n}\frac{E_i}{i}\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\frac{\sigma^2}{i^2}\operatorname{tr}\left(X(X^\top X)^{-2}X^\top\right)\right] = \sum_{i=1}^{n}\frac{\sigma^2}{i^2}\mathbb{E}\left[\operatorname{tr}\left((X^\top X)^{-1}\right)\right] \\
&= \frac{\sigma^2 d}{T-d-1}\sum_{i=1}^{n}\frac{\sigma^2}{i^2} \\
&< \frac{\sigma^2 d}{T-d-1}\times\frac{\pi^2}{6}.
\end{aligned}
$$

Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World

Kazdan et al. (2024)

- Different generative modeling settings
- Fixed compute budget
- Cardinality and proportion of real data

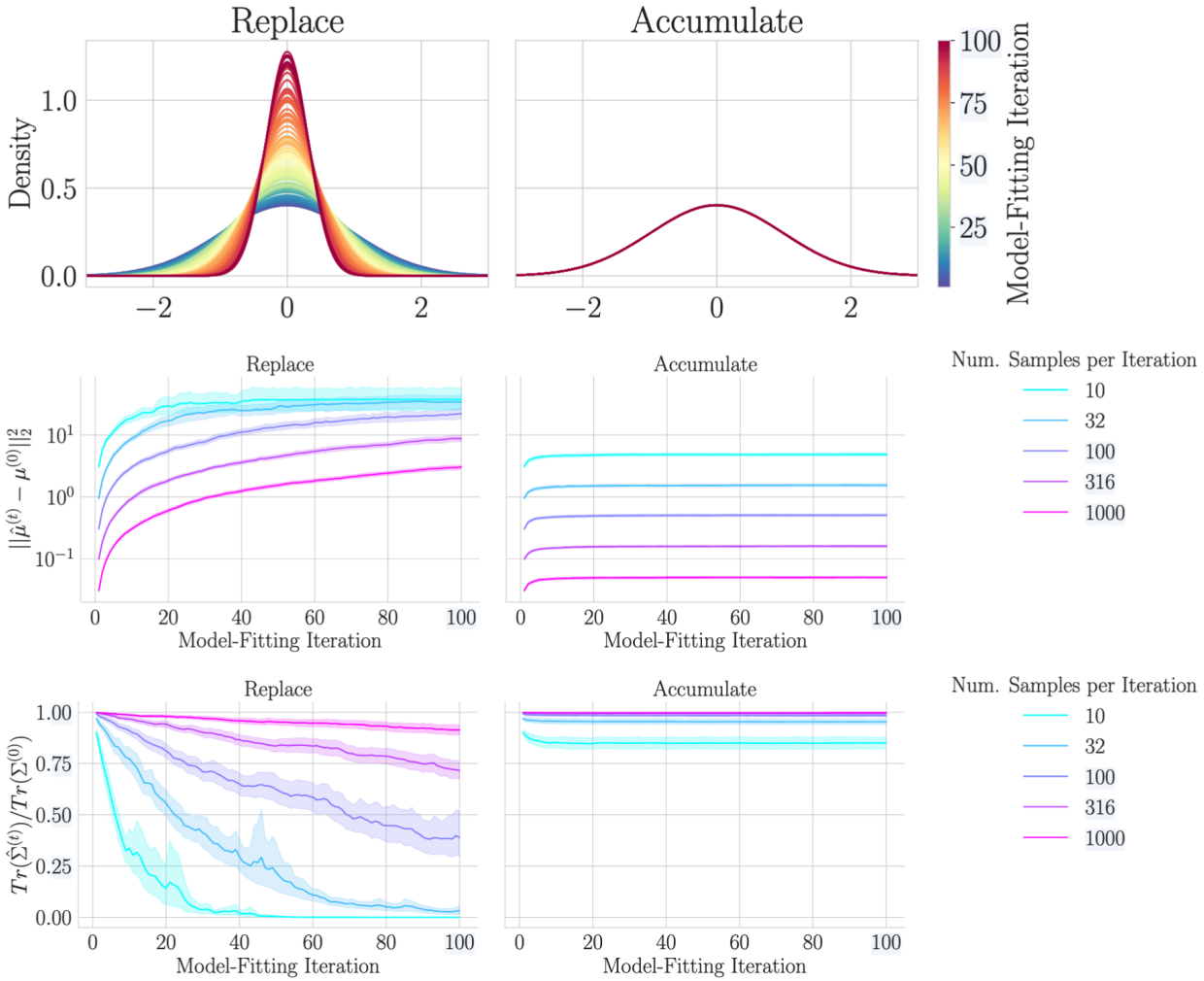
Testing Two Model Collapse Claims in Three Generative Modeling Settings

- Multivariate Gaussian Modeling
- Kernel Density Estimation
- Supervised Finetuning of Language Models

Multivariate Gaussian Modeling

- Real data: $\mathbf{X}_1^{(0)}, \ldots, \mathbf{X}_n^{(0)} \sim_{\text{i.i.d.}} \mathcal{N}(\mu^{(0)}, \Sigma^{(0)})$.
- Replace
  - $\hat{\mu}_{\text{Replace}}^{(t+1)} = \dfrac{1}{n} \displaystyle\sum_{j=1}^{n} X_j^{(t)}$.
  - $\hat{\Sigma}_{\text{Replace}}^{(t+1)} = \dfrac{1}{n-1} \displaystyle\sum_{j=1}^{n} \left(X_j^{(t)} - \hat{\mu}_{\text{Replace}}^{(t+1)} \right) \left(X_j^{(t)} - \hat{\mu}_{\text{Replace}}^{(t+1)} \right)^T$.
- Accumulate
  - $\hat{\mu}_{\text{Accumulate}}^{(t+1)} = \dfrac{1}{n(t+1)} \displaystyle\sum_{i=0}^{t} \sum_{j=1}^{n} X_j^{(i)}$.
  - $\hat{\Sigma}_{\text{Accumulate}}^{(t+1)} = \dfrac{1}{n(t+1)-1} \displaystyle\sum_{i=0}^{t} \sum_{j=1}^{n} \left(X_j^{(i)} - \hat{\mu}_{\text{Accumulate}}^{(t+1)} \right) \left(X_j^{(i)} - \hat{\mu}_{\text{Accumulate}}^{(t+1)} \right)^T$.
- Synthetic data: $X_1^{(t)}, \ldots, X_n^{(t)} \,\Big|\, \hat{\mu}^{(t)}, \hat{\Sigma}^{(t)} \sim_{\text{i.i.d.}} \mathcal{N}(\hat{\mu}^{(t)}, \hat{\Sigma}^{(t)})$.

Multivariate Gaussian Modeling



- Replace
  - Fit means drift away.
  - Fit covariances collapse.
- Accumulate
  - Fit means stabilize.
  - Fit covariances stabilize.

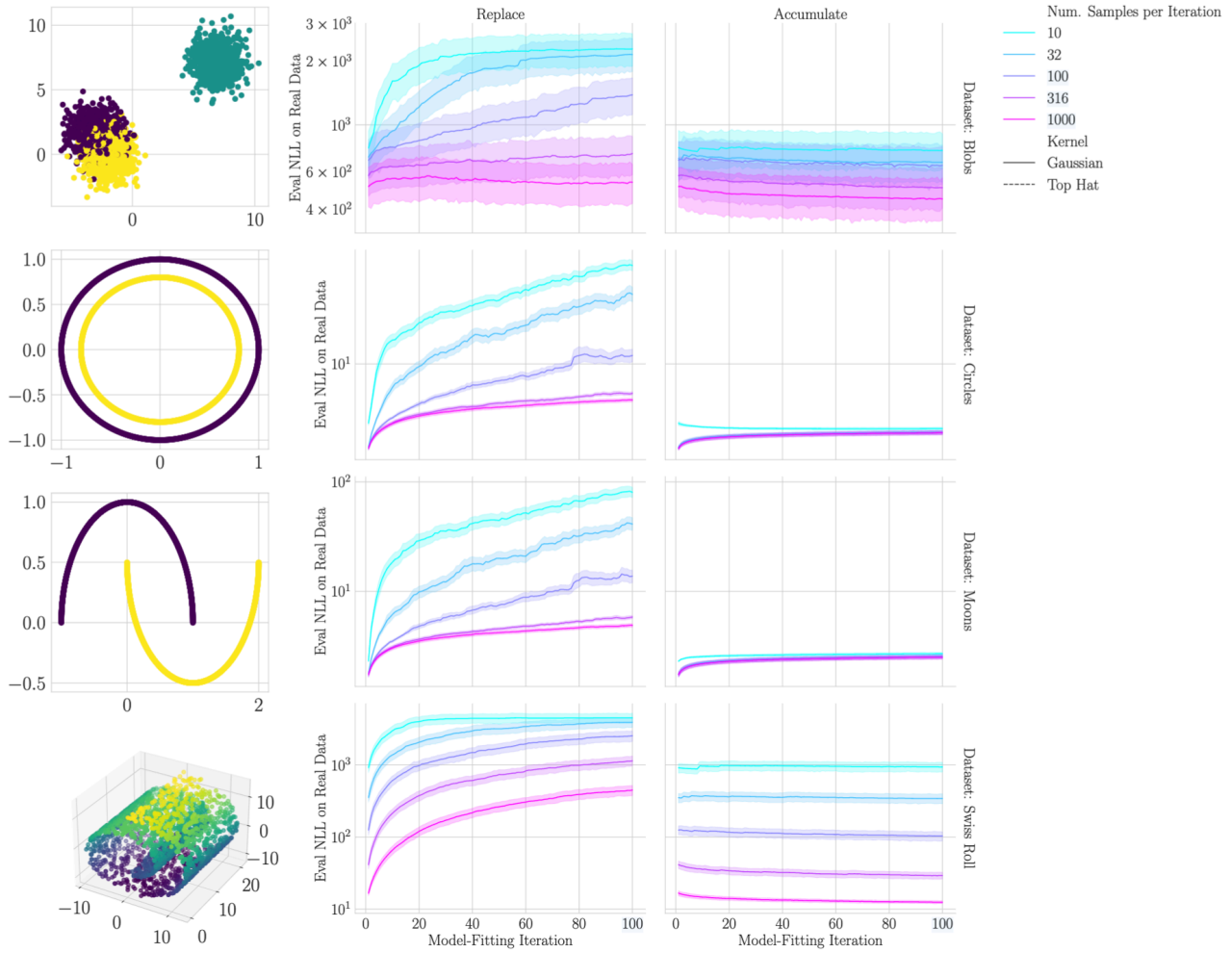
Multivariate Gaussian Modeling

- Replace
  - $\hat{\Sigma}_{\text{Replace}}^{(t+1)} \xrightarrow{\text{a.s.}} 0.$
  - $\mathbb{E}\left[W_2^2\left(\mathcal{N}(\hat{\mu}_{\text{Replace}}^{(t+1)}, \hat{\Sigma}_{\text{Replace}}^{(t+1)}), \mathcal{N}(\mu^{(0)}, \Sigma^{(0)})\right)\right] \to \infty \quad \text{as } t \to \infty.$
- Accumulate
  - For a univariate Gaussian
  - $\mathbb{E}(\sigma_t^2) = \sigma_0^2 \cdot \prod_{k=1}^{t}\left(1 - \frac{1}{k^2 n}\right) \xrightarrow{t\to\infty} \sigma_0^2 \cdot \left(\frac{\sin(\pi/\sqrt{n})}{\pi/\sqrt{n}}\right).$
  - $\mathbb{E}[(\mu_t - \mu_0)^2] = \sigma_0^2 \cdot \left(1 - \prod_{k=1}^{t}\left(1 - \frac{1}{k^2 n}\right)\right) \xrightarrow{t\to\infty} \sigma_0^2 \cdot \left(1 - \frac{\sin(\pi/\sqrt{n})}{\pi/\sqrt{n}}\right).$

Kernel Density Estimation

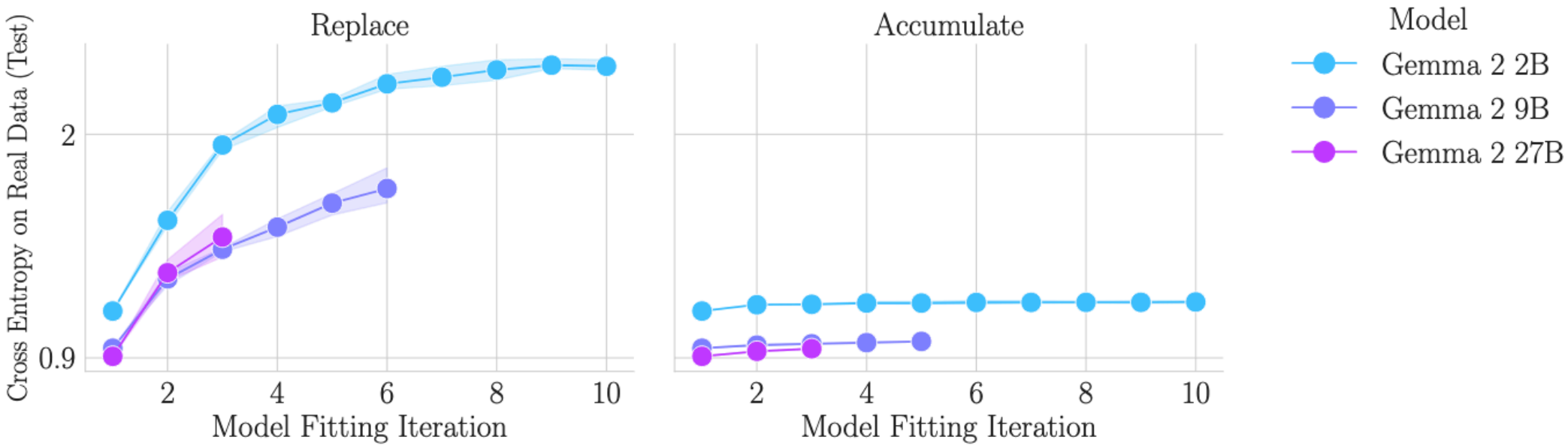
- Real data: $X_1^{(0)}, \ldots, X_n^{(0)} \sim_{\text{i.i.d.}} p^{(0)}$.
- K: kernel function, h: bandwidth parameter
- Replace: $\hat{p}^{(t+1)}_{\text{Replace}}(x) := \dfrac{1}{nh} \displaystyle\sum_{j=1}^{n} K\left(\dfrac{x - X_j^{(t)}}{h}\right)$.
- Accumulate: $\hat{p}^{(t+1)}_{\text{Accumulate}}(x) := \dfrac{1}{nh(t+1)} \displaystyle\sum_{i=0}^{t} \sum_{j=1}^{n} K\left(\dfrac{x - X_j^{(i)}}{h}\right)$.
- Synthetic data: $X_1^{(t)}, \ldots, X_n^{(t)} \sim_{\text{i.i.d.}} p^{(t)}$.

Kernel Density Estimation

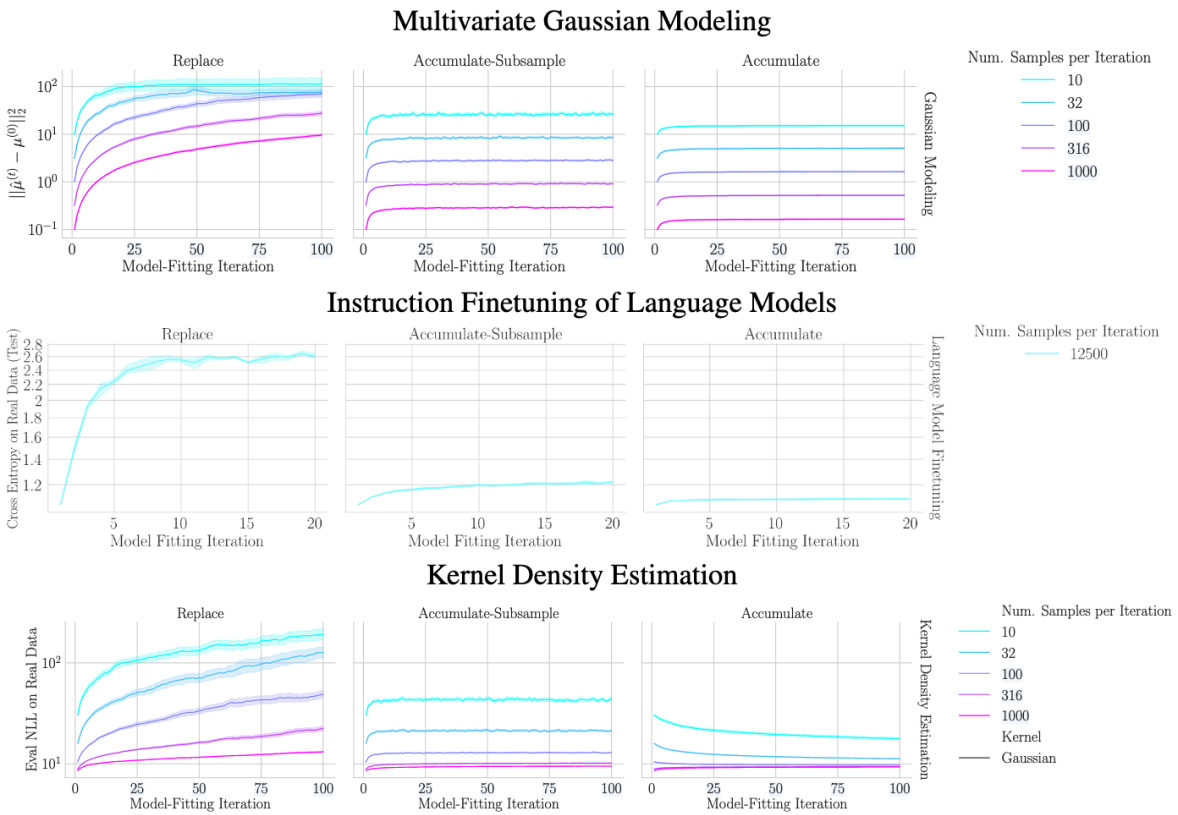


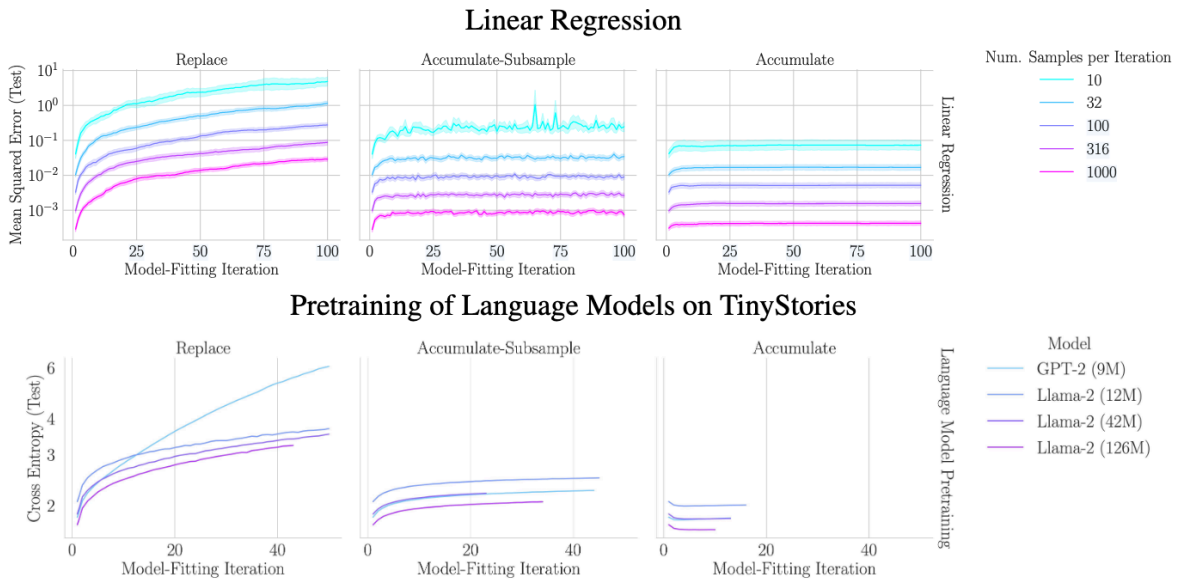
- Data: Blobs, Circles, Moons and Swiss Roll.
- Replace: NLL increases
- Accumulate: NLL is stable
  - lower loss on real test data than training on real data alone.
- Test NLL diverges!
  - Optimal bandwidth for the number of data at each iteration.

Supervised Finetuning of Language Models

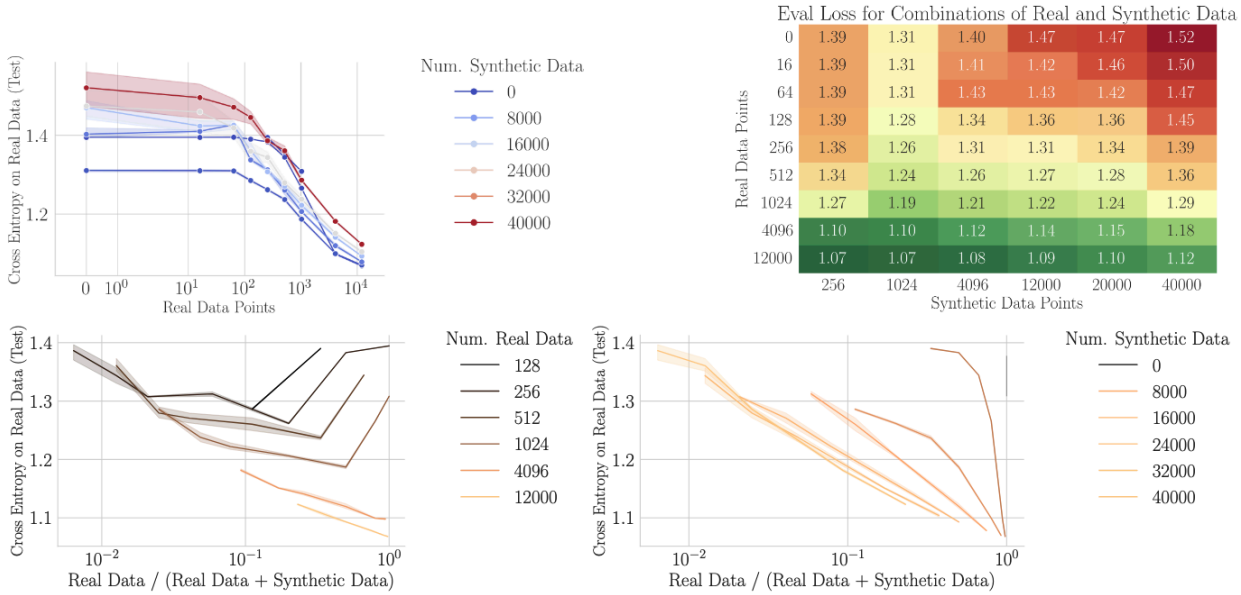


Fixed Compute Budget





Cardinality of Real Data vs Proportion of Real Data In Mitigating Model Collapse



- SFT Gemma 2 2B on HelpSteer2 dataset.
- Optimal fraction of synthetic data.
- Removing low-quality data > collecting more high-quality data.

Universality of the $\pi^2/6$ Pathway in Avoiding Model Collapse

Dey et al. (2024)

Workflow

Workflow 1 Iterative Model Training by Synthetic Data Augmentation

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 $\eta : \mathbb{R}^{d_X} \times \mathbb{R}^{d_\Theta} \to \mathbb{R}^{d_\eta}$.

1: 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 $1 \leq i \leq n$.

2: **for** each generation $G \geq 1$ **do**

3: Estimate $\hat{\theta}_G$ from $\mathcal{Z}_G$.

4: Generate $\mathcal{X}_{G+1} = \{X_{G+1,1}, \cdots, X_{G+1,n}\}$.

5: 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 $1 \leq i \leq n$.

6: Set $\mathcal{D}_{G+1} = \{(X_{G+1,1}, Y_{G+1,1}), \cdots, (X_{G+1,n}, Y_{G+1,n})\}$.

7: **Augment** the existing data corpus with the newly generated data: $\mathcal{Z}_{G+1} = \mathcal{Z}_G \cup \mathcal{D}_{G+1}$.

8: **end for**

Notations

- θ_0: true parameter
- X is some transformation of raw feature
- $Z = (X, Y)$, loss function L
- Given data $Z_1, \cdots, Z_n$, weighted M-estimator:

$$\hat{\theta}_n \in \arg\min_{\theta} \sum_{i=1}^{n} \omega_{n,i} L(Z_i; \theta).$$

How to compare the efficiency of estimators?

- "this estimator can achieve the same results as this other estimator, using only a fraction of f as much data".
- $\hat{\theta}$ achieves mean-squared error $MSE(\hat{\theta}, n)$ on a sample of size n.
- The asymptotic relative efficiency (ARE) of $\hat{\theta}_2$ relative to $\hat{\theta}_1$ is $100f\%$, if:

$$\frac{MSE(\hat{\theta}_2, n_2)}{MSE(\hat{\theta}_1, n_1)} \to 1, \quad n_2 \to \infty, \quad n_1 \sim f \cdot n_2 \to \infty.$$

Asymptotic Relative Efficiency (ARE)

- $\sqrt{n_i}(\hat{\theta}_i - \theta_0) \xrightarrow{d} N(0, V_i)$ $(i = 1, 2)$.
- $MSE(\hat{\theta}_i, n_i) \sim V_i/n_i$
- $ARE(\hat{\theta}_2, \hat{\theta}_1) = V_1/V_2$.
- When $V_1 > V_2$, $\hat{\theta}_2$ is more efficient.

Asymptotic Relative Efficiency (ARE)

- $\sqrt{n_i}(\hat{\theta}_i - \theta_0) \xrightarrow{d} N(0, V_i)\ (i = 1, 2).$
- $MSE(\hat{\theta}_i, n_i) \sim V_i / n_i$
- $ARE(\hat{\theta}_2, \hat{\theta}_1) = V_1 / V_2.$
- When $V_1 > V_2, \hat{\theta}_2$ is more efficient.
- $\forall G \geq 1, ARE(\hat{\theta}_G^{augment}, \hat{\theta}_1) \geq 6/\pi^2 > 60\%.$
- $ARE(\hat{\theta}_G^{discard}, \hat{\theta}_1) \overset{G \to \infty}{\to} 0.$

Assumptions

At each generation $G \geq 1$,

1. Features $\{X_{G,i}\}_{i=1}^{n} \overset{i.i.d.}{\sim} H$, a known distribution with finite moments of all orders.

Assumptions

At each generation $G \geq 1$,

1. Features $\{X_{G,i}\}_{i=1}^{n} \overset{i.i.d.}{\sim} H$, a known distribution with finite moments of all orders.
2. Response y is sampled from exponential family model (includes normal, binomial, poisson etc. distributions):

$$p(y \mid \eta(X, \theta)) = \exp\left(\eta(X, \theta)^{\top} T(y) - A(\eta(X, \theta))\right) h(y).$$

   - η: natural parameter, T: sufficient statistics, A: log-partition function.
   - $\eta(X, \theta) = X\theta \in \mathbb{R}^{d_\eta}$.
   - $\exists r > 0$ and g (has finite moments of all orders w.r.t. H), such that $\forall \tilde{\theta} \in B(\theta_0, r)$, $\forall X \in \mathbb{R}^{d_X}$,

$$\left| \frac{\partial^3 A(\eta)}{\partial \eta_i \eta_j \eta_k} \Big|_{\eta = X\tilde{\theta}} \right| \leq g(X), \quad \text{for all } 1 \leq i, j, k \leq d_\eta.$$

Assumptions

3. M-estimator $\hat{\theta}_G$ is asymptotically approximately linear (AAL)

$$\sqrt{nG}(\hat{\theta}_G - \theta_0) = \frac{1}{\sqrt{nG}} \sum_{g=1}^{G} \sum_{i=1}^{n} \omega_{G,g,i} \psi(Z_i; \theta) + o_{\mathbb{P}_0}(1).$$

- $\omega_{G,g,i}$: weight (possibly random) associated with $Z_{g,i}$ but is independent of $Z_{g,i}$.
- $\psi(z; \theta) = A_{n,G}(\theta)^{-1} \nabla_\theta L(z; \theta)$ for a loss function $L(\cdot; \cdot)$,

$$A_{n,G}(\theta) = \frac{1}{nG} \sum_{g=1}^{G} \sum_{i=1}^{n} \omega_{G,g,i} \nabla_\theta^2 L(Z_{g,i}; \theta).$$

- $\mathbb{E}_0[\nabla_\theta L(Z; \theta_0)] = 0$.

Theorem 5.1

- $W_{n,\Theta}(G) = \sqrt{n}(\hat{\theta}_G - \theta_0)$.
- $W_{n,T}(G) = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} X_{G,i}^{\top}\big(T(Y_{G,i}) - \nabla A(X_{G,i}\theta_0)\big)$.

Theorem 5.1

- $W_{n,\Theta}(G) = \sqrt{n}(\hat{\theta}_G - \theta_0)$.
- $W_{n,T}(G) = \dfrac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{G,i}^\top \big(T(Y_{G,i}) - \nabla A(X_{G,i}\theta_0)\big)$.
- Log likelihood ratio of $\mathbb{P}$ and $\mathbb{P}^{\text{ref}}$.
- $\mathbb{P}$: distribution of actual sequential data generating mechanism in workflow.
- $\mathbb{P}^{\text{ref}}$: reference distribution. All data are generated from the true distribution. $\{X_{G,i}\}_{G,i} \overset{i.i.d.}{\sim} H$ and $Y_{G,i}|X_{G,i} \sim \exp(\theta_0^\top X_{G,i}^\top T(Y_{G,i}) - A(X_{G,i}\theta_0))h(Y_{G,i})$.

Theorem 5.1

(a) Under $\mathbb{P}^{\text{ref}}$, $(W_{n,T}(g), W_{n,\Theta}(g))_{g=1}^{G} \xrightarrow{d} (W_T^{\text{ref}}(g), W_\Theta^{\text{ref}}(g))_{g=1}^{G}$.

- $(W_T^{\text{ref}}(g), W_\Theta^{\text{ref}}(g))_{g=1}^{G}$ is a collection of jointly Gaussian mean zero variables.

Theorem 5.1

(a) Under $\mathbb{P}^{\text{ref}}$, $(W_{n,T}(g), W_{n,\Theta}(g))_{g=1}^{G} \xrightarrow{d} (W_T^{\text{ref}}(g), W_\Theta^{\text{ref}}(g))_{g=1}^{G}$.

- $(W_T^{\text{ref}}(g), W_\Theta^{\text{ref}}(g))_{g=1}^{G}$ is a collection of jointly Gaussian mean zero variables.

(b) Under $\mathbb{P}$, $(W_{n,T}(g), W_{n,\Theta}(g))_{g=1}^{G} \xrightarrow{d} (W_T(g), W_\Theta(g))_{g=1}^{G}$.

- $(W_T(g), W_\Theta(g))_{g=1}^{G}$ is a sequential Gaussian process.
- $W_T(g) = W_T^{\text{ref}}(g) + \mathbb{E}_0[X^\top \nabla^2 A(X\theta_0) X] W_\Theta(g-1)$.
  - $W_T^{\text{ref}}(g)$ independent of $\{W_\Theta(i), W_T(i)\}_{1 \le i < g}$.
- Given $\{W_T(g), W_\Theta(g-1), \ldots, W_T(1)\}$, $W_\Theta(g)$ generate from conditional normal distribution
  $W_\Theta(g) \sim \mathbb{P}^{\text{ref}}(\cdot \mid W_T^{\text{ref}}(g) = W_T(g), \{W_\Theta^{\text{ref}}(i) = W_\Theta(i), W_T^{\text{ref}}(i) = W_T(i)\}_{1 \le i < g})$.

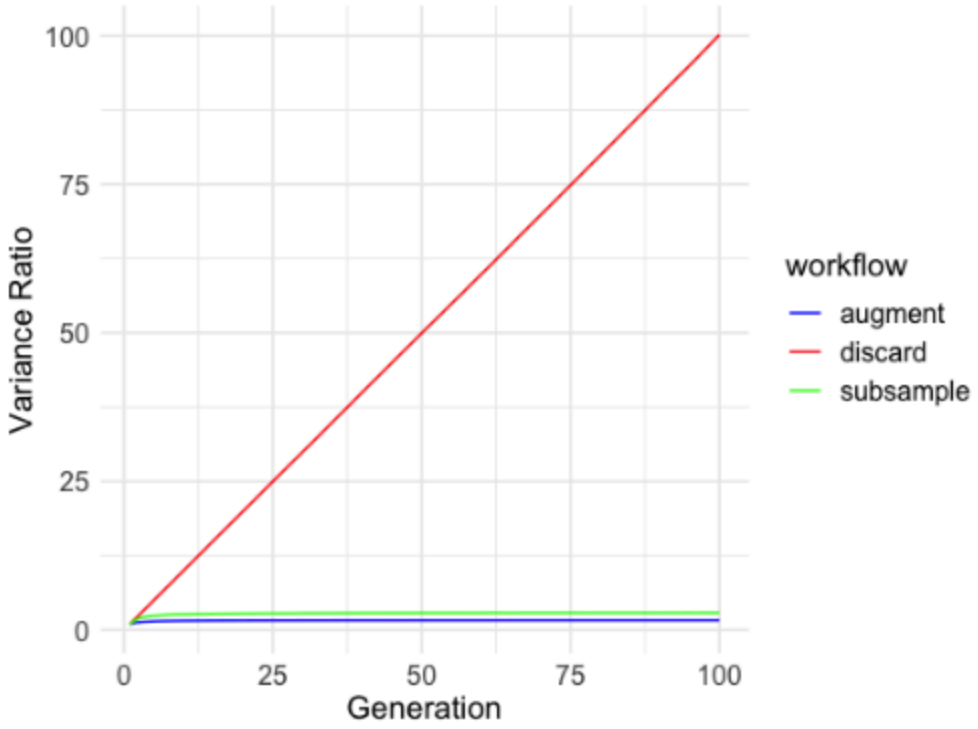
Implication on estimation

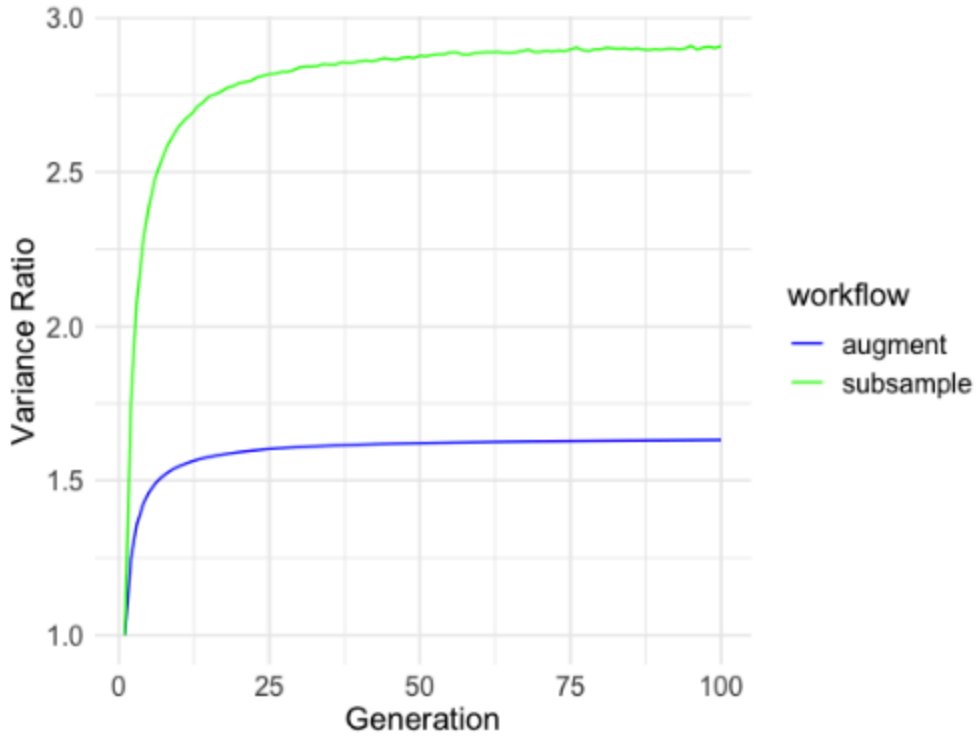
- Discard workflow: $w_{G,g,i} = \mathbf{1}_{g=G}$
  - $Var(W_\Theta(G)) = GVar(W_\Theta(1))$.
  - $ARE(\hat{\theta}_G^{dis}, \hat{\theta}_1) = 1/G \overset{G\to\infty}{\to} 0$.
- Augment workflow: $w_{G,g,i} = 1$
  - $Var(W_\Theta(G)) = (\sum_{g=1}^{G} 1/g^2) Var(W_\Theta(1))$.
  - $ARE(\hat{\theta}_G^{aug}, \hat{\theta}_1) = (\sum_{g=1}^{G} 1/g^2)^{-1} \geq 6/\pi^2 > 60\%$.

Implication on prediction

- Fit distribution $p(\cdot|X\hat{\theta})$ to the conditional distribution $Y \mid X$.
- Cross entropy: $CE(\hat{\theta}) = -\mathbb{E}_0(\log p(Y|X\hat{\theta})) = D_{KL}(\hat{\theta}||\theta_0) + H(p(Y|X\theta_0))$
  - $D_{KL}(\hat{\theta}||\theta_0) := D_{KL}(p(Y|X\hat{\theta})||p(Y|X\theta_0))$.
- Discard workflow: $\mathbb{E}D_{KL}(\hat{\theta}_G^{dis}||\theta_0)/\mathbb{E}D_{KL}(\hat{\theta}_1||\theta_0) \overset{n\to\infty}{\approx} G$.
- Augment workflow: $\mathbb{E}D_{KL}(\hat{\theta}_G^{aug}||\theta_0)/\mathbb{E}D_{KL}(\hat{\theta}_1||\theta_0) \overset{n\to\infty}{\approx} (\sum_{g=1}^{G} 1/g^2) \leq \pi^2/6$.

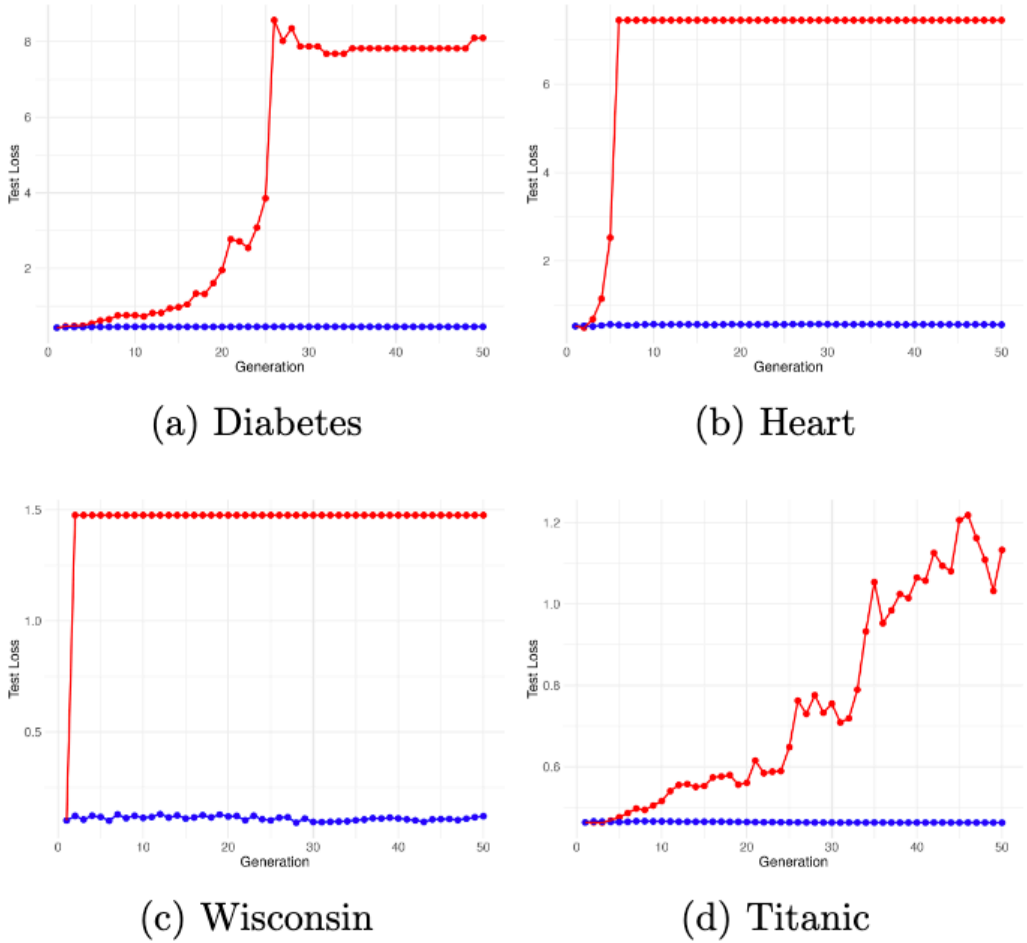
Simulation





- Ratio of variances of limit Gaussian variables.
- Discard workflow: explodes linearly.
- Augment workflow: concentrates around $\pi^2/6$.

Real-world data



- Classification test losses.
- Discard curves are significantly higher than the augment curves in all the cases.

Proof Sketch of Theorem

1. Show $\mathbb{P}$ ($\mathbb{P}_{G,nG}$) is contiguous to $\mathbb{P}^{\text{ref}}$ ($\mathbb{P}^{\text{ref}}_{G,nG}$)

Theorem 14.3.1 *Given P_n and Q_n, consider the likelihood ratio L_n defined in (14.36). Let G_n denote the distribution of L_n under P_n. Suppose G_n converges weakly to a distribution G. If G has mean 1, then Q_n is contiguous to P_n.*

2. Get the limiting distribution of $\mathbb{P}$ from the limiting distribution of $\mathbb{P}^{\text{ref}}$

Theorem 8.1. *[Theorem 14.3.3 from Lehmann et al. (1986)]* **Le Cam's Third Lemma.** *Suppose $\mathbb{Q}_N \triangleleft \mathbb{P}_N$ and define $L_N(z_1, \ldots, z_N) = (d\mathbb{Q}_N/d\mathbb{P}_N)(z_1, \cdots, z_N)$ to be the likelihood ratio. Suppose for a sequence of random variables $R_N \equiv R_N(Z_1, \cdots, Z_N)$, for every bounded continuous g, we know that*

$$\mathbb{E}_{\mathbb{P}_N}[g(R_N, L_N)] \overset{N\to\infty}{\longrightarrow} \int g(r, \ell) dF(r, \ell), \tag{32}$$

for some specific distribution function F. Then,

$$\mathbb{E}_{\mathbb{Q}_N}[g(R_N, L_N)] \overset{N\to\infty}{\longrightarrow} \int g(r, \ell) \cdot \ell \cdot dF(r, \ell). \tag{33}$$

1.

Define

- $S_{n,G} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{g,i}^{\top} (T(Y_{g,i}) - \nabla A(X\theta_0)), \sqrt{n}(\hat{\theta}_g - \theta_0) \right)_{g=1}^{G},$
- $W_T(G) = \text{asymp.lim.} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{G,i}^{\top} (T(Z_{G,i}) - \nabla A(\theta_0)) \right] \sim \mathcal{N}(0, \mathcal{V}_T)$
  - $\mathcal{V}_T = \mathbb{E}_0(X^{\top} \nabla^2 A(X\theta_0) X)$
- $W_\Theta(G) = \text{asymp.lim.} \left[\sqrt{n}(\hat{\theta}_G - \theta_0) \right].$
- $S_{n,G} \xrightarrow{d} S_G = (W_T(g), W_\Theta(g))_{g=1}^{G}.$

2.

- Denote $\hat{\theta}_0 = \theta_0$, by Taylor expansion to $f(\theta) = A(X\theta)$,

$$\log L_{G,nG}(Z_1, \ldots, Z_{nG}) = \sum_{g=1}^{G} \left[\sum_{i=1}^{n} \left((\hat{\theta}_{g-1} - \theta_0)^\top X_{g,i}^\top (T(Y_{g,i}) - \nabla A(X_{g,i}\theta_0)) \right.\right.$$
$$\left.\left. -\frac{1}{2}(\hat{\theta}_{g-1} - \theta_0)^\top X_{g,i}^\top \nabla^2 A(X_{g,i}\theta_0) X_{g,i} (\hat{\theta}_{g-1} - \theta_0) \right) \right] + o_p(1).$$

- $\log L_{G,nG}(Z_1, \ldots, Z_{nG})$ is continuous w.r.t. $S_{n,G}$, thus $\log L_{G,nG}(S_{n,G}) \xrightarrow{d} \log L_{G,nG}(S_G)$.
- $\frac{1}{n} \sum_{i=1}^{n} X_{g,i}^\top \nabla^2 A(X_{g,i}\theta_0) X_{g,i} \xrightarrow{\mathbb{P}^{\text{ref}}} \mathbb{E}(X^\top \nabla^2 A(X\theta_0) X)$.
- By Slutsky's lemma, $\log L_{G,nG}(S_{n,G}) \xrightarrow{d}$
$$\sum_{g=1}^{G} \left[W_\Theta(g-1)^\top W_T(g) - \frac{1}{2} W_\Theta(g-1)^\top \mathbb{E}(X^\top \nabla^2 A(X\theta_0) X) \cdot W_\Theta(g-1) \right].$$

2.

- $L_{G,nG}(Z_1, \ldots, Z_{nG}) \xrightarrow{d} L_{G,\infty}(S_G) := \exp(\cdots)$.
- Letting $\mathcal{F}_g = \sigma(S_g)$ (the σ-algebra generated by the random variables up to generation g),

$$\mathbb{E}[L_{G,\infty}(S_G)] = \mathbb{E}\left[\mathbb{E}\left(\exp\left(W_\Theta(G-1)^\top W_T(G) - \frac{1}{2} W_\Theta(G-1)^\top \cdot \mathcal{V}_T \cdot W_\Theta(G-1)\right) \mid \mathcal{F}_{G-1}\right) \times L_{G-1,\infty}(S_{G-1})\right].$$

- Because $W_T(G)$ is independent of $\mathcal{F}_{G-1}$,

  $$\mathbb{E}[\exp(W_\Theta(G-1)^\top W_T(G) \mid \mathcal{F}_{G-1}] = \exp\left(\frac{1}{2} W_\Theta(G-1)^\top \cdot \mathcal{V}_T \cdot W_\Theta(G-1)\right).$$

- Thus $\mathbb{E}[L_{G,\infty}(S_G)] = \cdots = \mathbb{E}[L_{1,\infty}(S_1)] = 1$. (no distribution shift in the first generation)

3.

- $\mathbb{P}^{\text{ref}}(W_T^{\text{ref}}(g), W_\Theta^{\text{ref}}(g), g \leq G) = \left(\prod_{g=1}^{G} \mathbb{P}^{\text{ref}}(W_T^{\text{ref}}(g)) \right) \times \prod_{G'=1}^{G} \mathbb{P}^{\text{ref}}(W_\Theta^{\text{ref}}(G') \mid W_T^{\text{ref}}(G'), \{W_T^{\text{ref}}(g), W_\Theta^{\text{ref}}(g)\}_{g<G'})$.
- $\mathbb{P}(\boldsymbol{s}_G) = \mathbb{P}^{\text{ref}}(\boldsymbol{s}_G) \times L_{G,\infty}(\boldsymbol{s}_G)$

$$= \prod_{G'=1}^{G} \mathbb{P}^{\text{ref}}(W_\Theta(G') \mid W_T(G'), S_{G'-1}) \times \left(\prod_{g=1}^{G} \mathbb{P}^{\text{ref}}(W_T(g)) \right) \times L_{G,\infty}(\boldsymbol{s}_G)$$

$$= \prod_{G'=1}^{G} \mathbb{P}(W_T(G') \mid W_\Theta(G'-1)) \times \mathbb{P}^{\text{ref}}(W_\Theta(G') \mid W_T(G'), S_{G'-1}).$$

- $\mathbb{P}(W_T(G') \mid W_\Theta(G'-1)) = \frac{1}{\sqrt{\det(2\pi \mathcal{V}_T)}} \exp\left(-\frac{1}{2} (W_T(g) - \mathcal{V}_T W_\Theta(g-1))^\top \mathcal{V}_T^{-1} (W_T(g) - \mathcal{V}_T W_\Theta(g-1)) \right)$.

Proof Sketch of Lemma

1. Law of total variance
2. Conditional Gaussian
3. Block matrix inverse