

# Error Propagation and Model Collapse in Diffusion Models

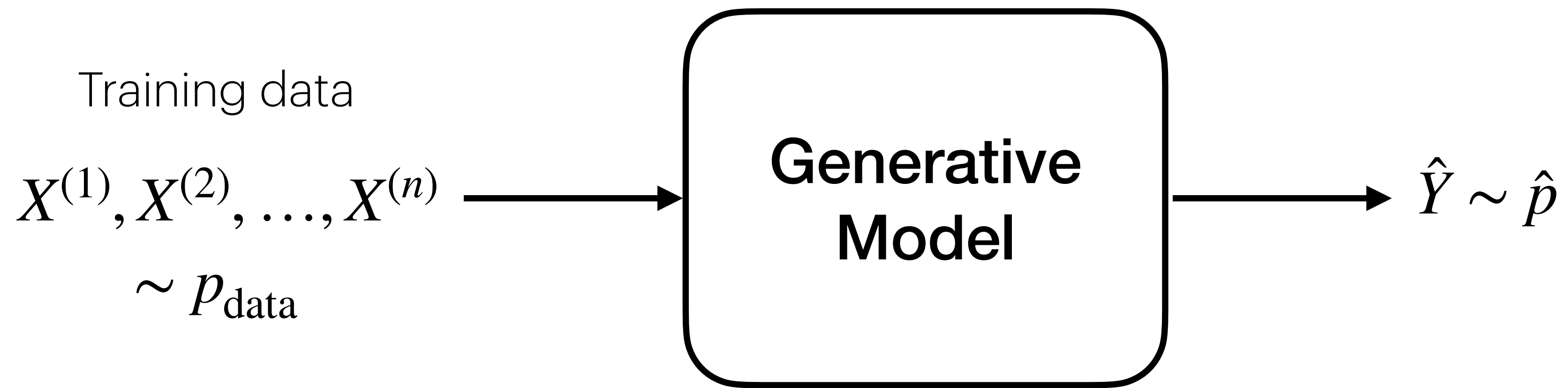
Ramji Venkataramanan



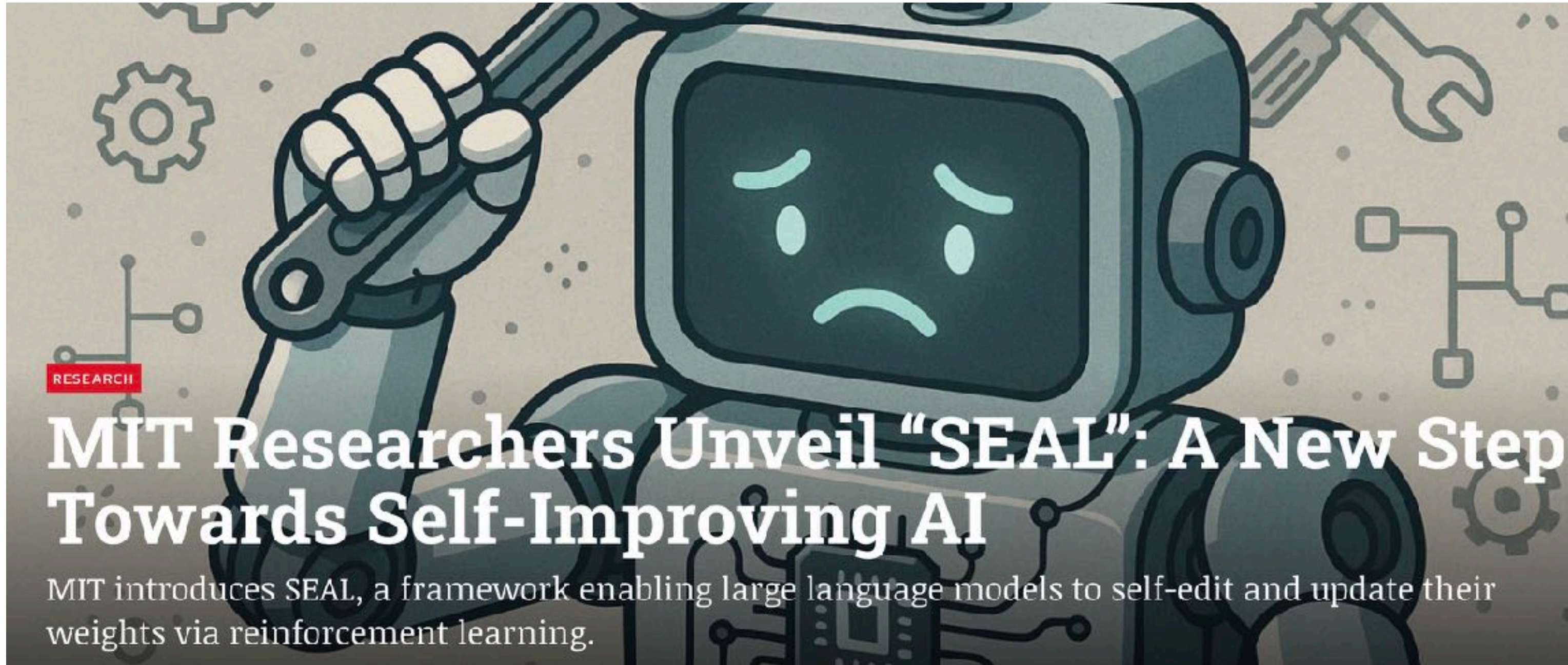
Nail Kheilfa



Richard Turner



Want  $\hat{p} \approx P_{\text{data}}$



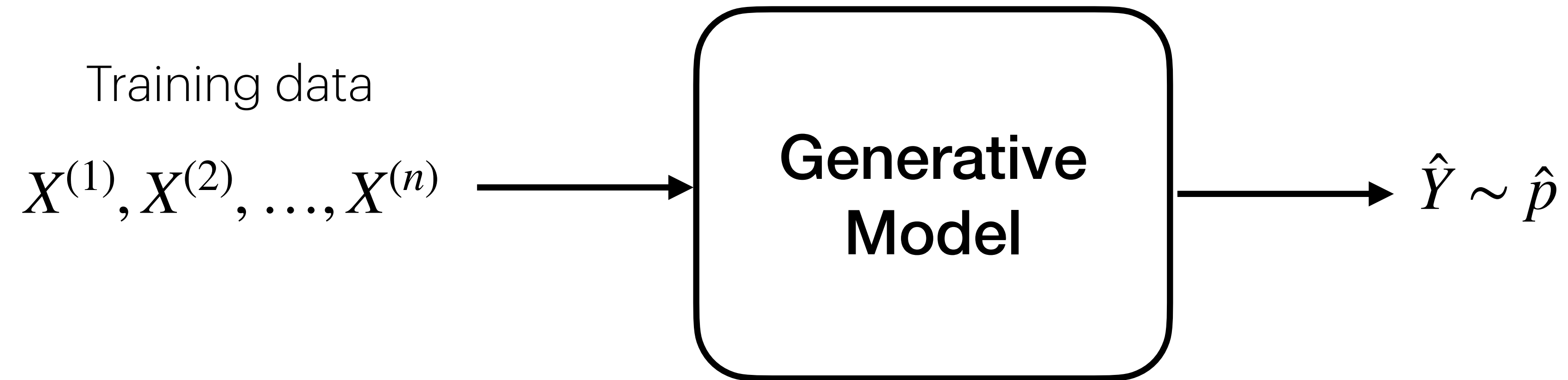
## MIT Researchers Unveil “SEAL”: A New Step Towards Self-Improving AI

MIT introduces SEAL, a framework enabling large language models to self-edit and update their weights via reinforcement learning.

## OpenAI says new coding model helped build itself

OpenAI says its own artificial intelligence systems are accelerating the pace of AI development.

# Recursive Self-Training



Training uses **combination** of fresh data  $\sim p_{\text{data}}$  and synthetic data  $\sim \hat{p}_{\text{prev}}$

Want  $\hat{p} \approx p_{\text{data}}$

# Self-Consuming Generative Models Go MAD

---

Sina Alemohammad,<sup>\*†</sup> Josue Casco-Rodriguez,<sup>\*†</sup> Lorenzo Luzi,<sup>†</sup> Ahmed Imtiaz Humayun,<sup>†</sup>  
Hossein Babaei,<sup>†</sup> Daniel LeJeune,<sup>‡</sup> Ali Siahkoohi,<sup>§</sup> Richard G. Baraniuk<sup>†</sup>

<sup>†</sup>Department of Electrical and Computer Engineering, Rice University

<sup>‡</sup>Department of Statistics, Stanford University

<sup>§</sup>Department of Computational Applied Mathematics and Operations Research, Rice University

---

## AI models collapse when trained on recursively generated data


<https://doi.org/10.1038/s41586-024-07566-y>

Received: 20 October 2023

Accepted: 14 May 2024

Published online: 24 July 2024

Open access

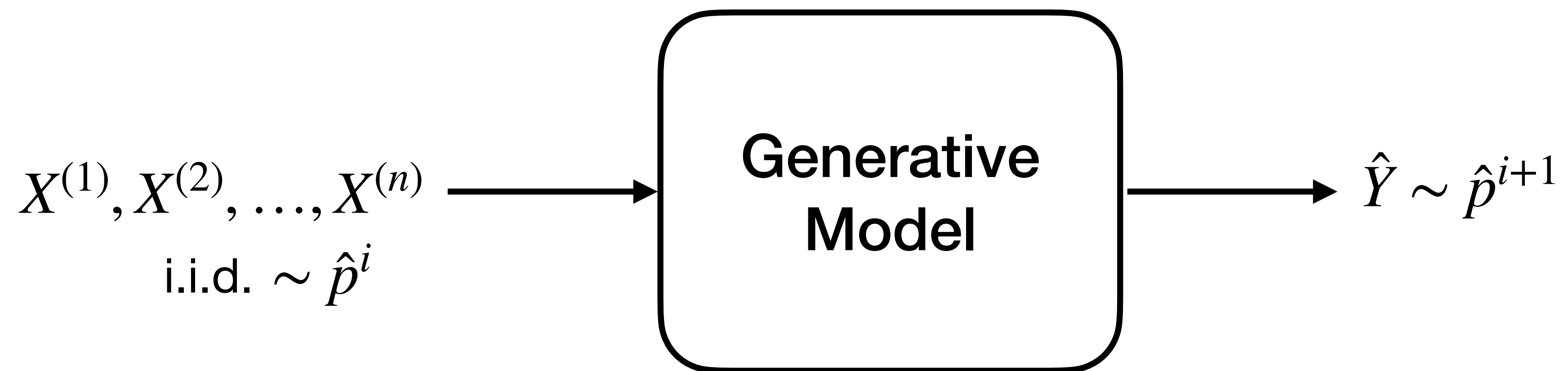
 Check for updates

Ilia Shumailov<sup>1,8✉</sup>, Zakhar Shumaylov<sup>2,8✉</sup>, Yiren Zhao<sup>3</sup>, Nicolas Papernot<sup>4,5</sup>, Ross Anderson<sup>6,7,9</sup>  
& Yarin Gal<sup>1✉</sup>

---

Stable diffusion revolutionized image creation from descriptive text. GPT-2 (ref. 1), GPT-3(.5) (ref. 2) and GPT-4 (ref. 3) demonstrated high performance across a variety of language tasks. ChatGPT introduced such language models to the public. It is now clear that generative artificial intelligence (AI) such as large language models (LLMs) is here to stay and will substantially change the ecosystem of online text and images. Here we consider what may happen to GPT- $\{n\}$  once LLMs contribute much of the text found online. We find that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear. We refer to this effect as ‘model collapse’ and show

# 1-D Gaussian with self-training



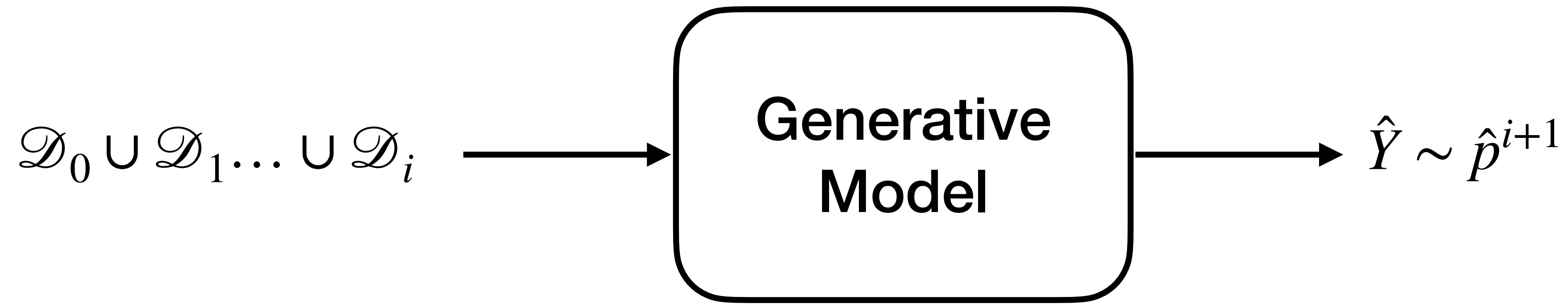
Target  $p_{\text{data}} = \mathcal{N}(\mu, \sigma^2)$

After generation  $i$  compute  $\hat{\mu}_i, \hat{\sigma}_i^2$  and generate samples from  $\hat{p}^i = \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i^2)$

Can show  $\text{Var}(\hat{Y}) = \sigma^2 \left( 1 + \frac{i}{n} \right) \rightarrow \infty$  with growing  $i$  [Shumailov et al. '24]

# Preventing Model Collapse

## Data Accumulation



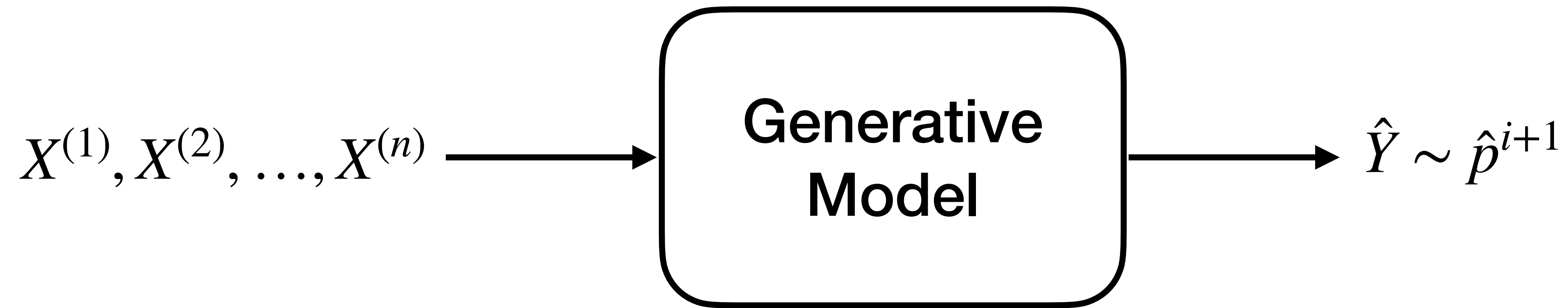
Training data for generation  $i$  :  $\mathcal{D}_0 \cup \mathcal{D}_1 \dots \cup \mathcal{D}_i$

$$\mathcal{D}_0 \sim P_{\text{data}}, \mathcal{D}_1 \sim \hat{p}^1, \dots, \mathcal{D}_i \sim \hat{p}^i$$

- Linear and Generalized linear models: [Gerstgragsser et al. '24], [Dey-Donoho '24]
- Discrete distribution estimation: [Kanabar-Gastpar '25]
- MLE for parametric distributions: [Barzilai-Shamir '25]

# Preventing Model Collapse

## Data Augmentation

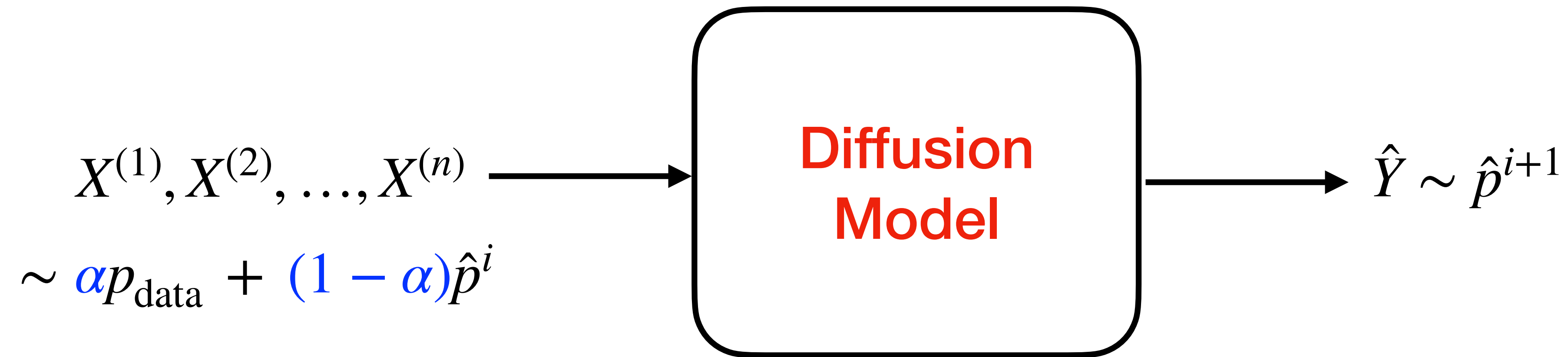


Training data for generation  $i + 1$  :

Fraction  $\alpha$  of samples  $\sim p_{\text{data}}$ , separately fraction  $(1 - \alpha)$  from  $\hat{p}^i$

- Gaussian estimation, linear regression: [He et al '25]
- Overparametrized linear models: [Garg et. al '25]

# This talk



Training data for generation  $i$  : **Mixture** of  $p_{\text{data}}$  and  $\hat{p}^i$

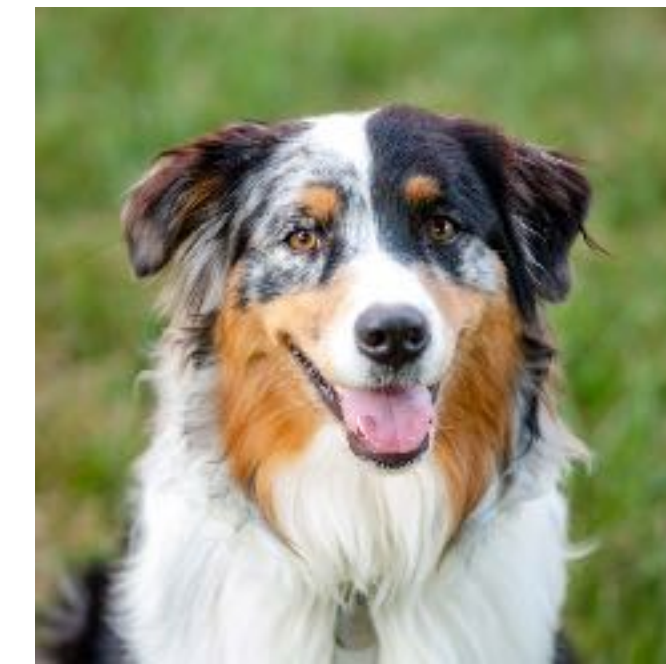
**Goal:** Quantify how divergence between  $\hat{p}^i$  and  $p_{\text{data}}$  evolves

# Diffusion Models

$X^{(1)}, X^{(2)}, \dots, X^{(n)}$

**Diffusion  
Model**

$\hat{Y} \sim \hat{p}$

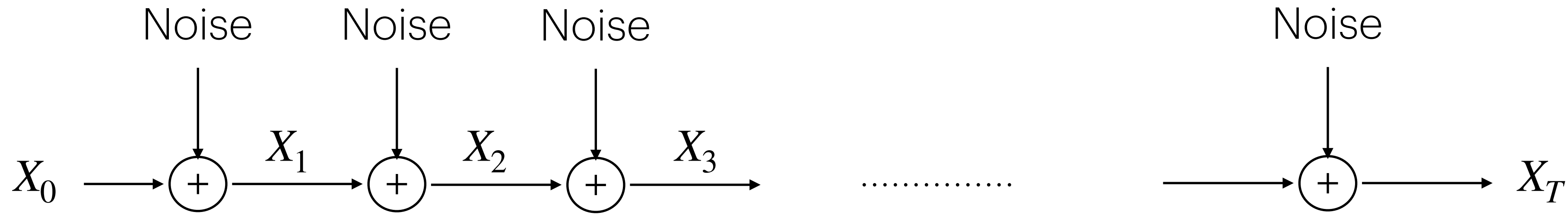


State of the art for text-to-image, video generation, protein modeling, ...

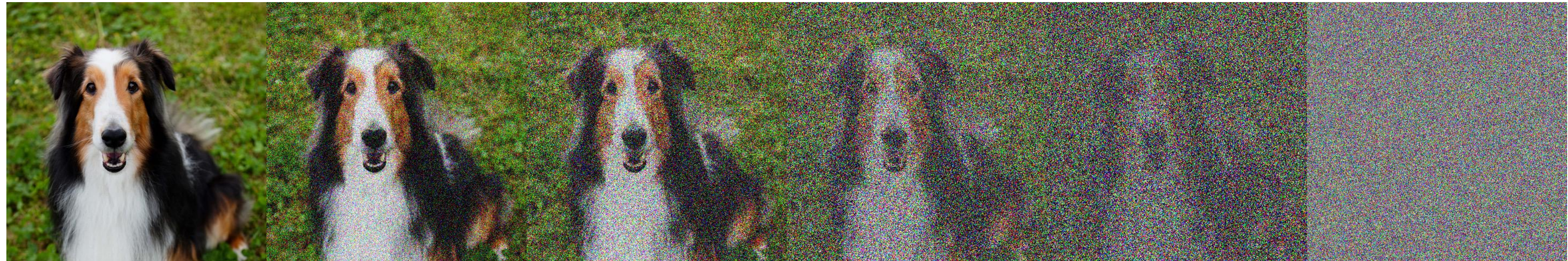
[Sohl-Dickstein et al '15], [Ho et al. '20, Song et al., '21]



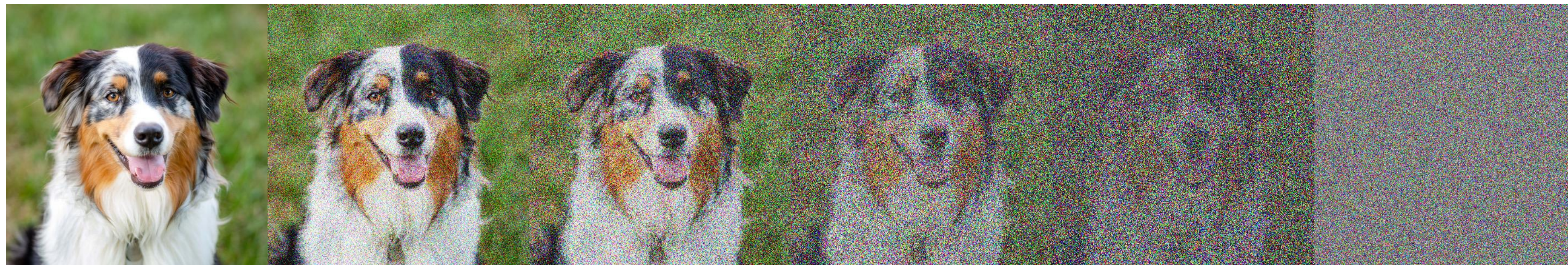
“Create a cartoon image of a bear punting in Cambridge”



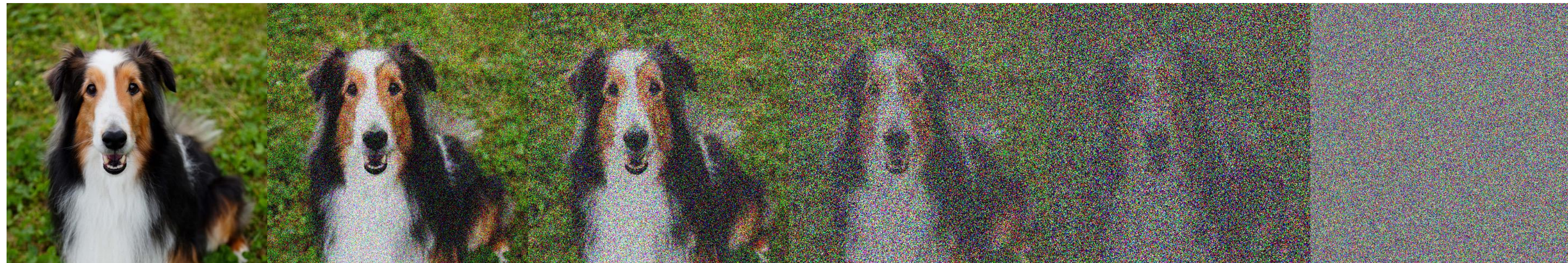
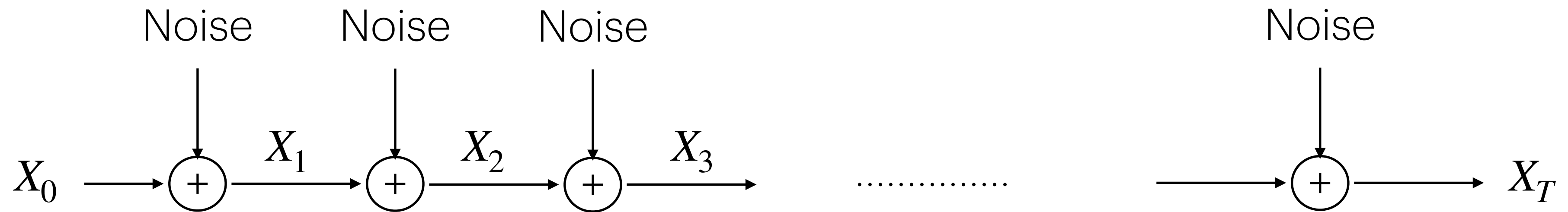
**Forward  
noising**



**Backward  
denoising**



# Forward Noising Process

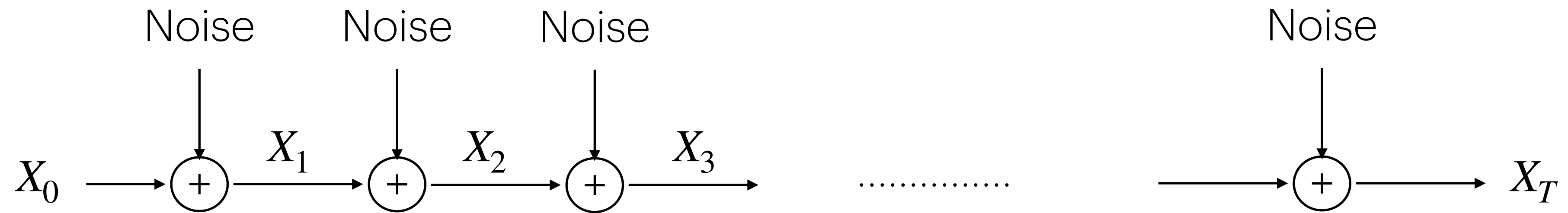


$p_{\text{data}}$  : distribution on  $\mathbb{R}^d$ . Starting from  $X_0 \sim p_{\text{data}}$ ,

$$dX_t = -\frac{1}{2}X_t dt + dB_t, \quad t \in [0, T]$$

Discretization: 
$$X_{k+1} = X_k - \frac{1}{2}X_k \Delta_k + G_k \sqrt{\Delta_k}, \quad G_k \sim \mathcal{N}(0, I)$$

# Forward Noising SDE



**Forward  
noising**



$$dX_t = -\frac{1}{2}X_t dt + dB_t, \quad t \in [0, T], \quad X_0 \sim p_{\text{data}}$$

$$X_t \mid X_0 \sim \mathcal{N}(X_0 e^{-\frac{t}{2}}, (1 - e^{-t})I)$$

# Backward Denoising

Forward SDE:  $dX_t = -\frac{1}{2}X_t dt + dB_t$ ,  $t \in [0, T]$ ,  $X_0 \sim p_{\text{data}}$

With  $q_t = \text{Law}(X_t)$ , the **score function** is  $s(x, t) = \nabla_x \log q_t(x)$

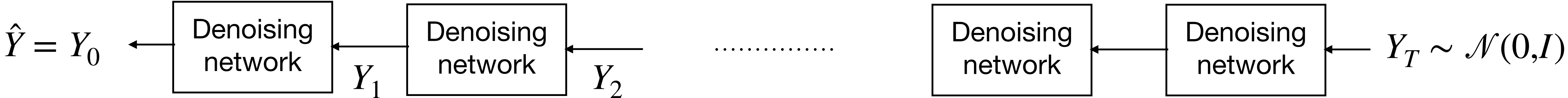
Reverse SDE:  $dY_s = \left[ -\frac{1}{2}Y_s - s(Y_s, s) \right] ds + d\bar{B}_s$ ,  $Y_T \sim q_T$

Then  $\text{Law}(Y_s) = \text{Law}(X_s) = q_s$  for  $s \in [0, T]$  [Anderson '82]

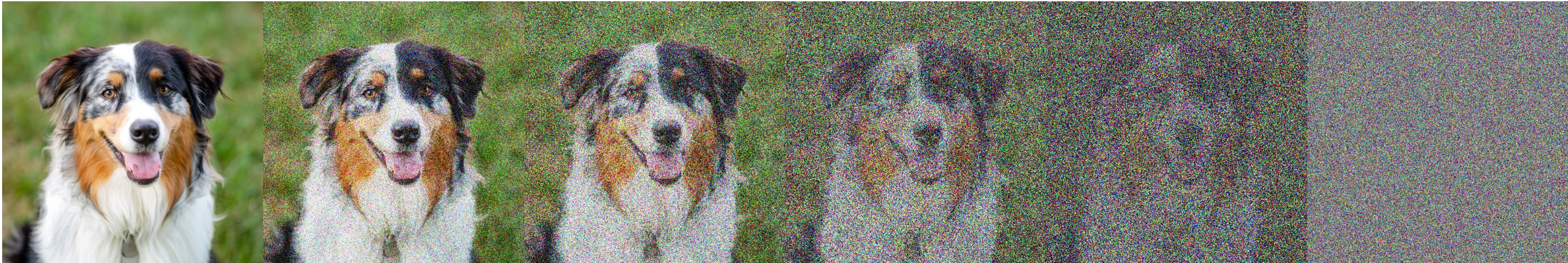
In particular,  $\text{Law}(Y_0) = p_{\text{data}}$  and  $q_T \approx \mathcal{N}(0, I) \Rightarrow$

Sampling from  $p_{\text{data}}$  is equivalent to learning the score function  $s(x, t)$

# Sampling via Learned Reverse SDE



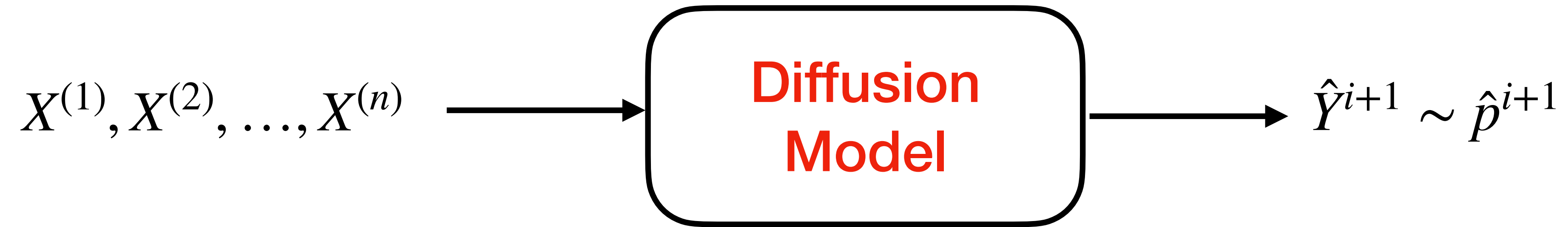
**Backward  
denoising**



Learned Reverse SDE: 
$$d\hat{Y}_s = \left[ -\frac{1}{2}\hat{Y}_s - s_\theta(\hat{Y}_s, s) \right] ds + d\bar{B}_s, \quad \hat{Y}_T \sim \mathcal{N}(0, I)$$

Score error: 
$$e(x, t) = s(x, t) - s_\theta(x, t)$$

# Recursive Training Model



Training data for generation  $i$  : **Mixture** of  $p_{\text{data}}$  and  $\hat{p}^i$

$$\sim_{\text{i.i.d}} q_i = \alpha p_{\text{data}} + (1 - \alpha) \hat{p}^i$$

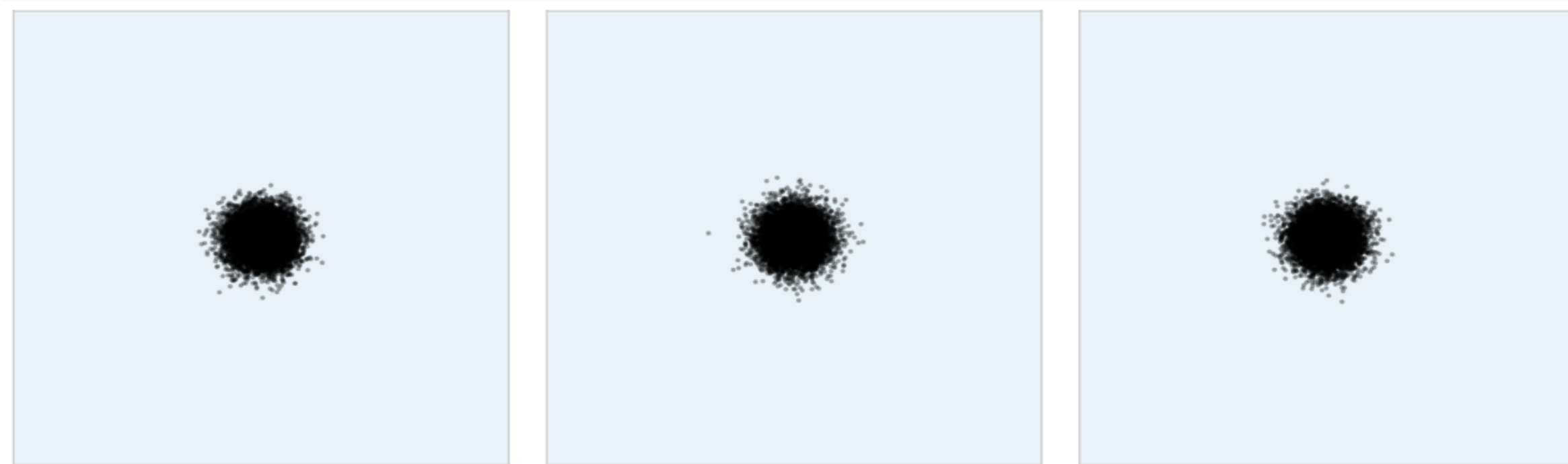
Track the recursion using two  $\chi^2$ -divergences:

- Intra-generation divergence  $I_i := \chi^2(\hat{p}^{i+1} \| q_i)$
- Accumulated divergence  $D_i := \chi^2(\hat{p}^i \| p_{\text{data}})$

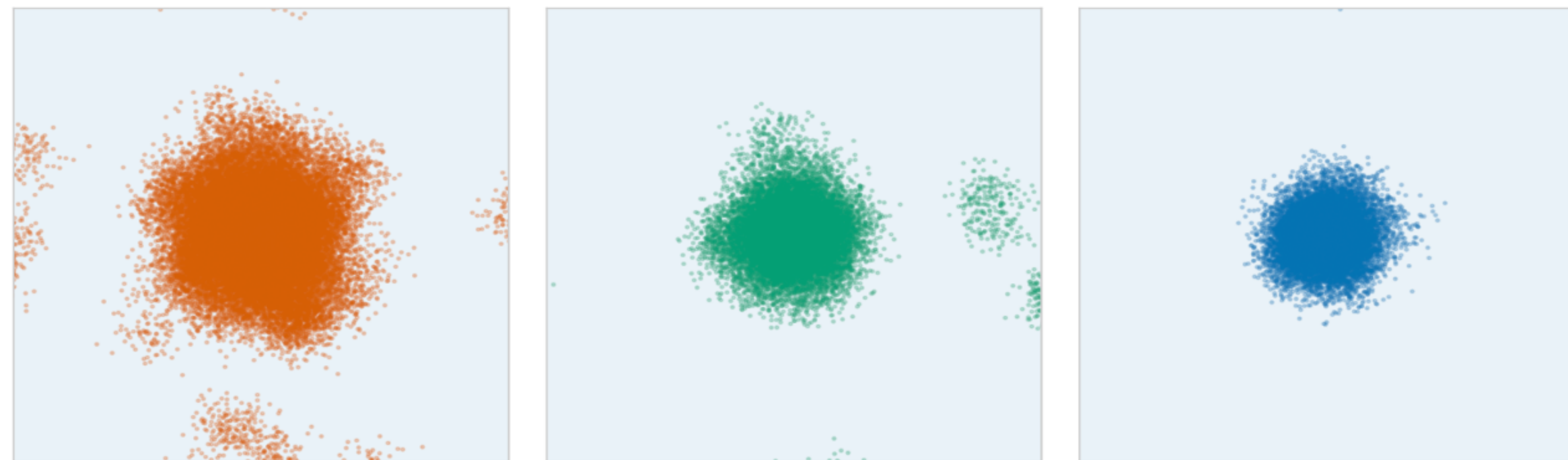
$$\chi^2(\hat{p}^{i+1} \| q_i) = \mathbb{E}_{q_i} \left[ \left( \frac{\hat{p}^{i+1}}{q_i} - 1 \right)^2 \right]$$

$p_{\text{data}}$  : Mixture of 5 isotropic Gaussians in  $\mathbb{R}^{10}$

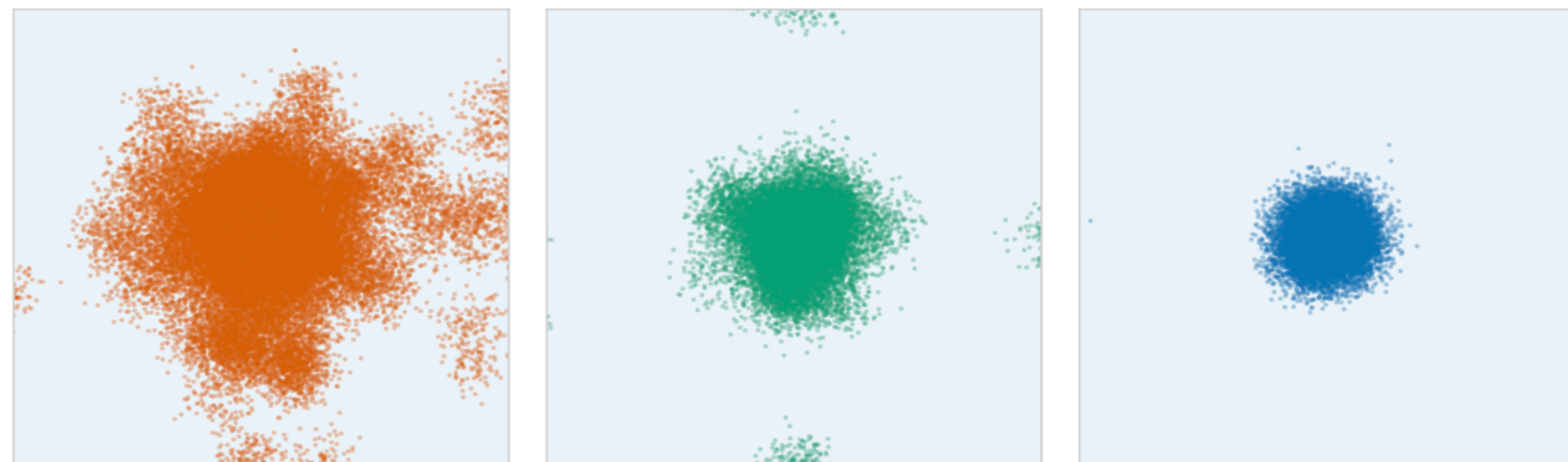
Generation 0



Generation 10



Generation 20



$\alpha = 0.1$

$\alpha = 0.5$

$\alpha = 0.9$

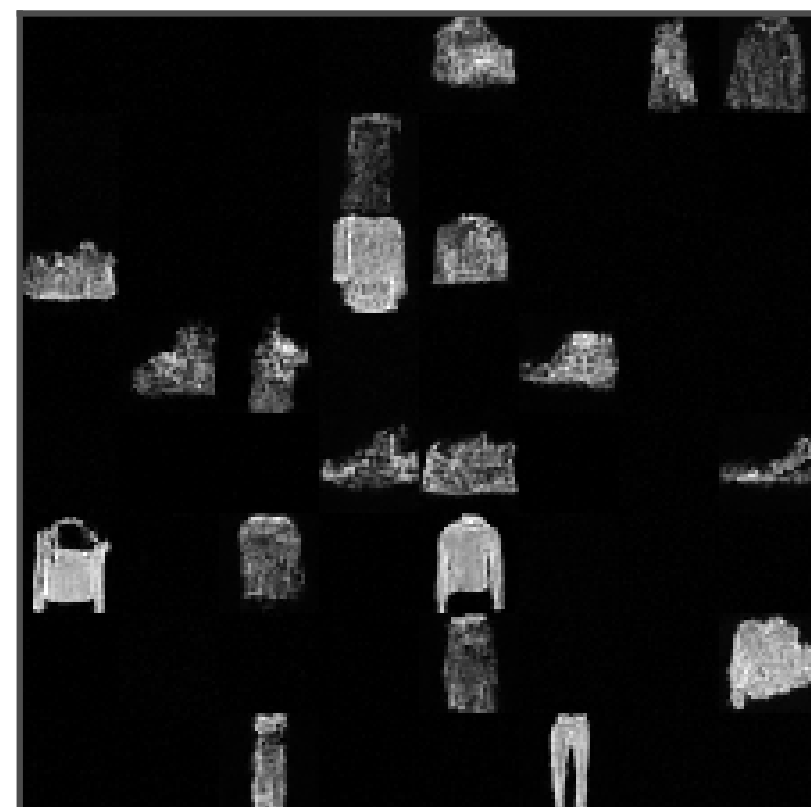
Gen 0

Gen 10

Gen 15

Gen 25

$\alpha = 0.1$



$\alpha = 0.5$



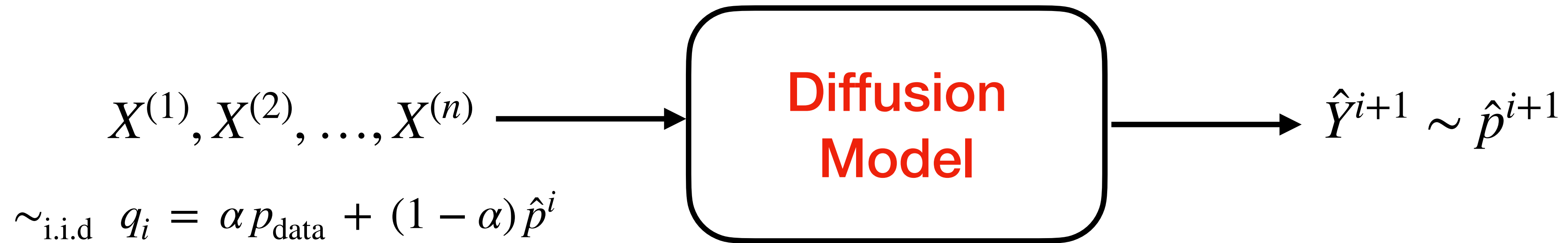
$\alpha = 0.9$



$P_{\text{data}}$

Fashion-MNIST

# Path Laws



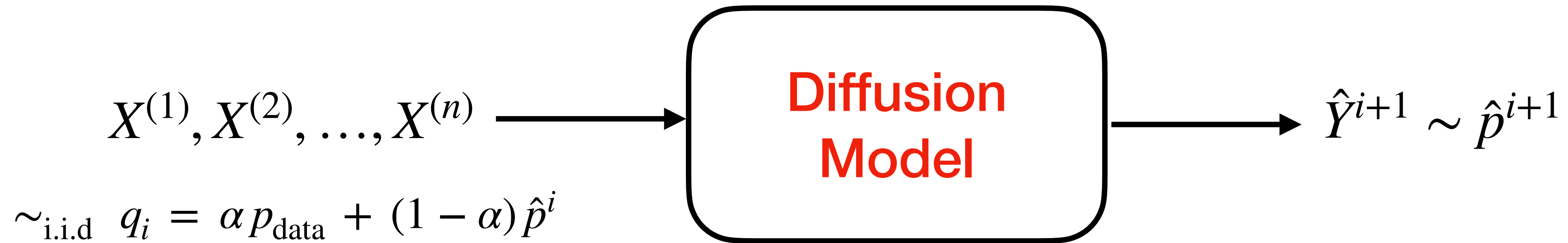
Ideal Reverse SDE:  $dY_s^i = \left[ -\frac{1}{2}Y_s^i - \mathbf{s}(Y_s^i, s) \right] ds + d\bar{B}_s, \quad Y_T^i \sim \mathcal{N}(0, I)$

Learned Reverse SDE:  $d\hat{Y}_s^i = \left[ -\frac{1}{2}\hat{Y}_s^i - \mathbf{s}_\theta(\hat{Y}_s^i, s) \right] ds + d\bar{B}_s, \quad \hat{Y}_T^i \sim \mathcal{N}(0, I)$

Start at  $s = T$  and stop at  $s = t_0 > 0$

- Path Laws  $\mathbb{P}_i = \text{Law} \left( (Y_s^i)_{s \in [t_0, T]} \right), \quad \hat{\mathbb{P}}_i = \text{Law} \left( (\hat{Y}_s^i)_{s \in [t_0, T]} \right)$
- Time  $t_0$  marginals  $Y_{t_0}^i \sim q_i, \quad \hat{Y}_{t_0}^i \sim \hat{p}^{i+1}$

# Score Error Energy



Ideal Reverse SDE:  $dY_s^i = \left[ -\frac{1}{2}Y_s^i - \mathbf{s}_i(Y_s^i, s) \right] ds + d\bar{B}_s, \quad Y_T^i \sim \mathcal{N}(0, I) \quad (\mathbb{P}^i)$

Learned Reverse SDE:  $d\hat{Y}_s^i = \left[ -\frac{1}{2}\hat{Y}_s^i - \mathbf{s}_\theta(\hat{Y}_s^i, s) \right] ds + d\bar{B}_s, \quad \hat{Y}_T^i \sim \mathcal{N}(0, I) \quad (\hat{\mathbb{P}}^i)$

Score error:  $e_i(x, t) = \mathbf{s}_i(x, t) - \mathbf{s}_\theta(x, t)$

$$\varepsilon_i^2 := \mathbb{E}_{\mathbb{P}^i} \left[ \int_{t_0}^T \|e_i(Y_s^i, s)\|_2^2 ds \right], \quad \hat{\varepsilon}_i^2 := \mathbb{E}_{\hat{\mathbb{P}}^i} \left[ \int_{t_0}^T \|e_i(\hat{Y}_s^i, s)\|_2^2 ds \right]$$

Score error:  $e_i(x, t) = s(x, t) - s_\theta(x, t)$

$$\varepsilon_i^2 := \mathbb{E}_{\mathbb{P}_i} \left[ \int_{t_0}^T \|e_i(Y_s^i, s)\|_2^2 ds \right]$$

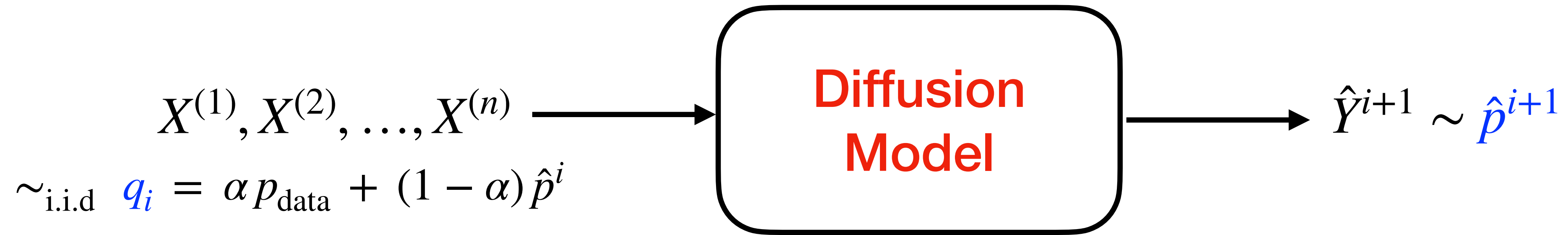
Under suitable assumptions, minimax score estimation error:

$$\varepsilon_i^2 \gtrsim \frac{\text{polylog}(n_i)}{n_i} \left( \frac{1}{t_0} \right)^{d/2}$$

Small  $\varepsilon_i^2$  requires number of training samples  $\sim (1/t_0)^{d/2}$

[Zhang et al '24], [Okon et al '23], [Dou et al '24], [Wibisono et al '24]...

# Intra-generation upper bound



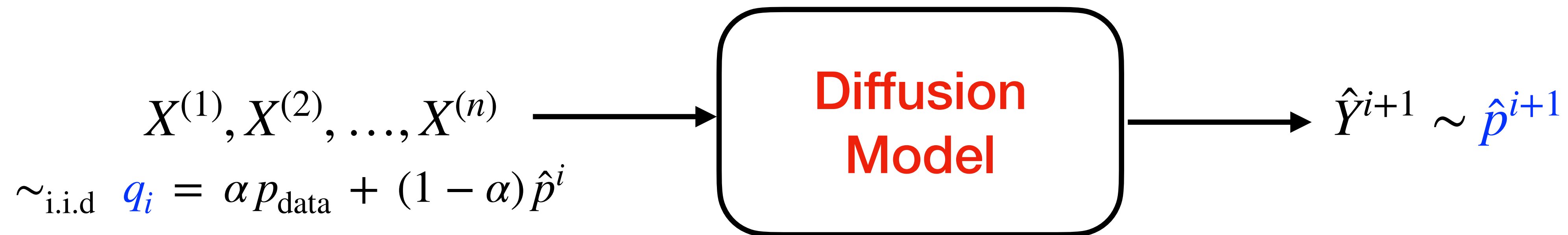
Learned Reverse SDE:  $d\hat{Y}_s = \left[ -\frac{1}{2}\hat{Y}_s - s(\hat{Y}_s, s) + e_i(\hat{Y}_s, s) \right] ds + d\bar{B}_s, \quad \hat{Y}_T \sim \mathcal{N}(0, I) \quad (\hat{\mathbb{P}}^i)$

**Proposition:** Under mild assumptions,

$$\text{KL}(\hat{p}^{i+1} \| q_i) \leq \text{KL}(\hat{\mathbb{P}}_i \| \mathbb{P}_i) = \frac{1}{2} \hat{\varepsilon}_i^2$$

where  $\hat{\varepsilon}_i^2 := \mathbb{E}_{\hat{\mathbb{P}}_i} \left[ \int_{t_0}^T \|e_i(\hat{Y}_s^i, s)\|_2^2 ds \right]$

# Intra-generation upper bound

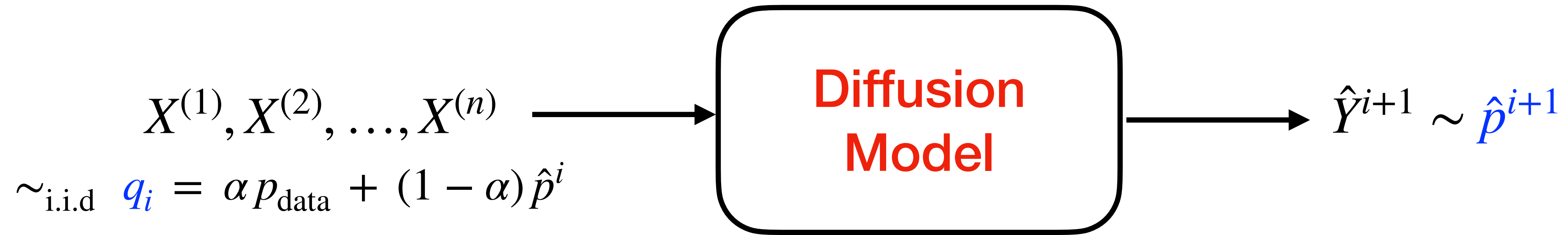


**Proposition:**  $\text{KL}(\hat{p}^{i+1} \| q_i) \leq \text{KL}(\hat{\mathbb{P}}_i \| \mathbb{P}_i) = \frac{1}{2} \hat{\varepsilon}_i^2$

where  $\hat{\varepsilon}_i^2 := \mathbb{E}_{\hat{\mathbb{P}}_i} \left[ \int_{t_0}^T \|e_i(\hat{Y}_s^i, s)\|_2^2 ds \right]$

- Standard application of Girsanov's theorem gives  $\frac{d\hat{\mathbb{P}}_i}{d\mathbb{P}_i}$
- Inequality via data processing

# Girsanov path density

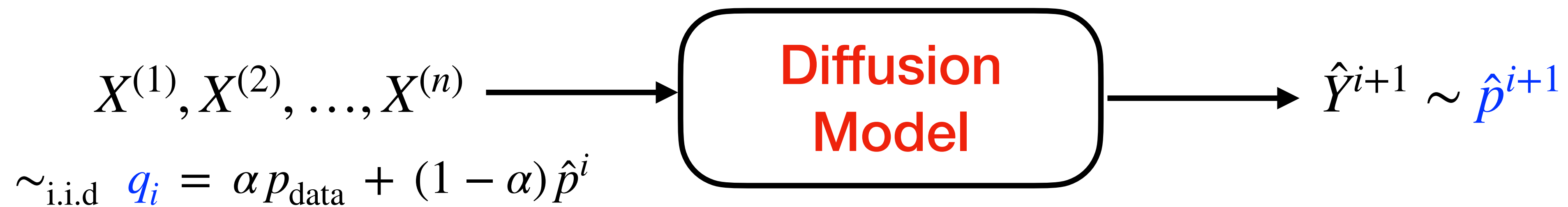


Ideal Reverse SDE:  $dY_s^i = \left[ -\frac{1}{2}Y_s^i - \mathbf{s}(Y_s^i, s) \right] ds + d\bar{B}_s, \quad Y_T^i \sim \mathcal{N}(0, I) \quad (\mathbb{P}^i)$

Girsanov's theorem:  $\frac{d\hat{\mathbb{P}}_i}{d\mathbb{P}_i} = \exp(Z_i)$  where  $Z_i = M_i - \frac{1}{2}\langle M_i \rangle$ ,

$$M_i = - \int_{t_0}^T e_i(Y_s^i, s) \cdot d\bar{B}_s \quad \langle M_i \rangle = \int_{t_0}^T \|e_i(Y_s^i, s)\|^2 ds$$

# Ratio of Marginals



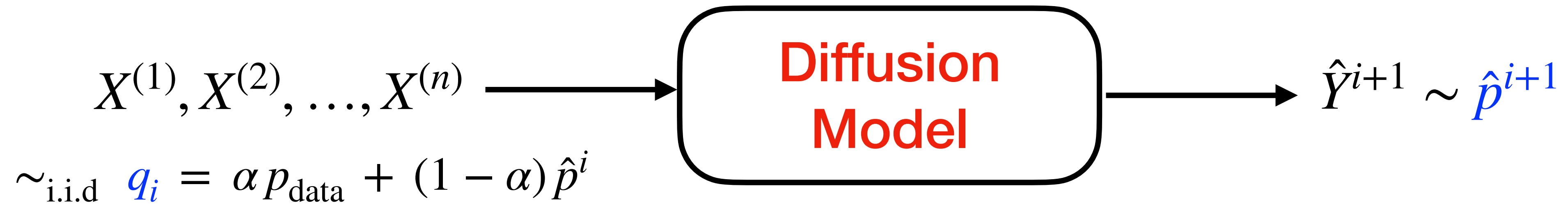
Ideal Reverse SDE:  $dY_s^i = \left[ -\frac{1}{2}Y_s^i - s_i(Y_s^i, s) \right] ds + d\bar{B}_s, \quad Y_T^i \sim \mathcal{N}(0, I) \quad (\mathbb{P}^i)$

Girsanov's theorem:  $\frac{d\hat{\mathbb{P}}_i}{d\mathbb{P}_i} = \exp(Z_i)$  where  $Z_i = M_i - \frac{1}{2}\langle M_i \rangle$ ,

$$M_i = - \int_{t_0}^T e_i(Y_s^i, s) \cdot d\bar{B}_s \quad \langle M_i \rangle = \int_{t_0}^T \|e_i(Y_s^i, s)\|^2 ds$$

Marginal density ratio  $\frac{\hat{p}^{i+1}(y)}{q_i(y)} = \mathbb{E}_{\mathbb{P}_i} \left[ \exp(Z_i) \mid Y_{t_0}^i = y \right]$  (Projection of path density ratio)

# Intra-generation Lower Bound



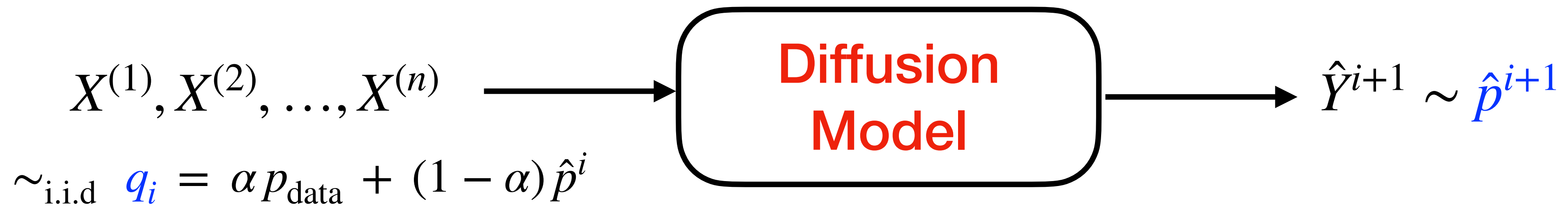
## Proposition

Recall  $\varepsilon_i^2 := \mathbb{E}_{\mathbb{P}_i} \left[ \int_{t_0}^T \|e_i(Y_s^i, s)\|_2^2 ds \right]$ . Under mild assumptions, if  $\varepsilon_i^2 < 1$ ,

$$\chi^2(\hat{p}^{i+1} \| q_i) \geq \eta_i \varepsilon_i^2 - C \varepsilon_i^4$$

$\eta_i$  captures the **observability** of the score error at  $t_0$

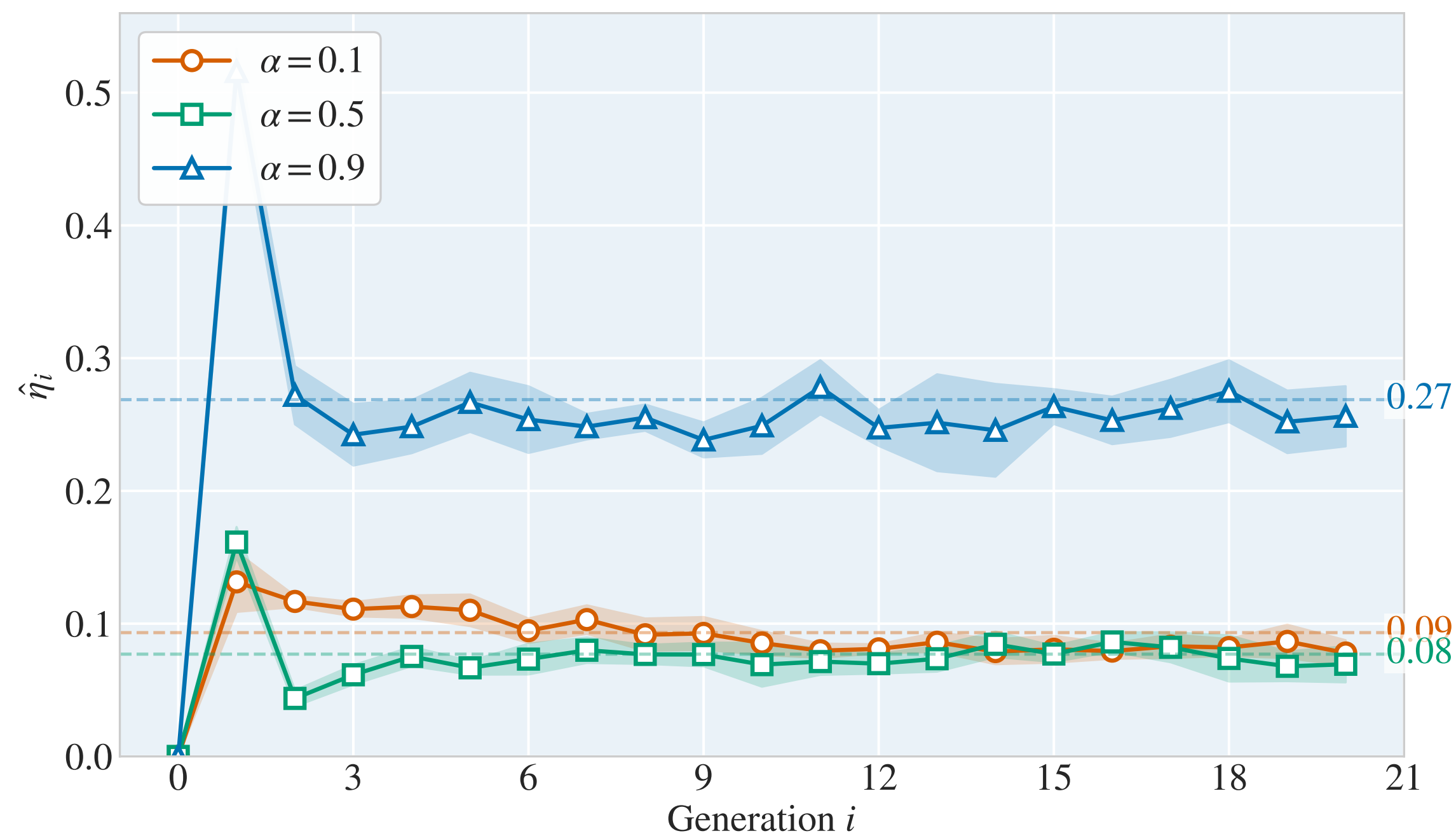
# Observability of Errors



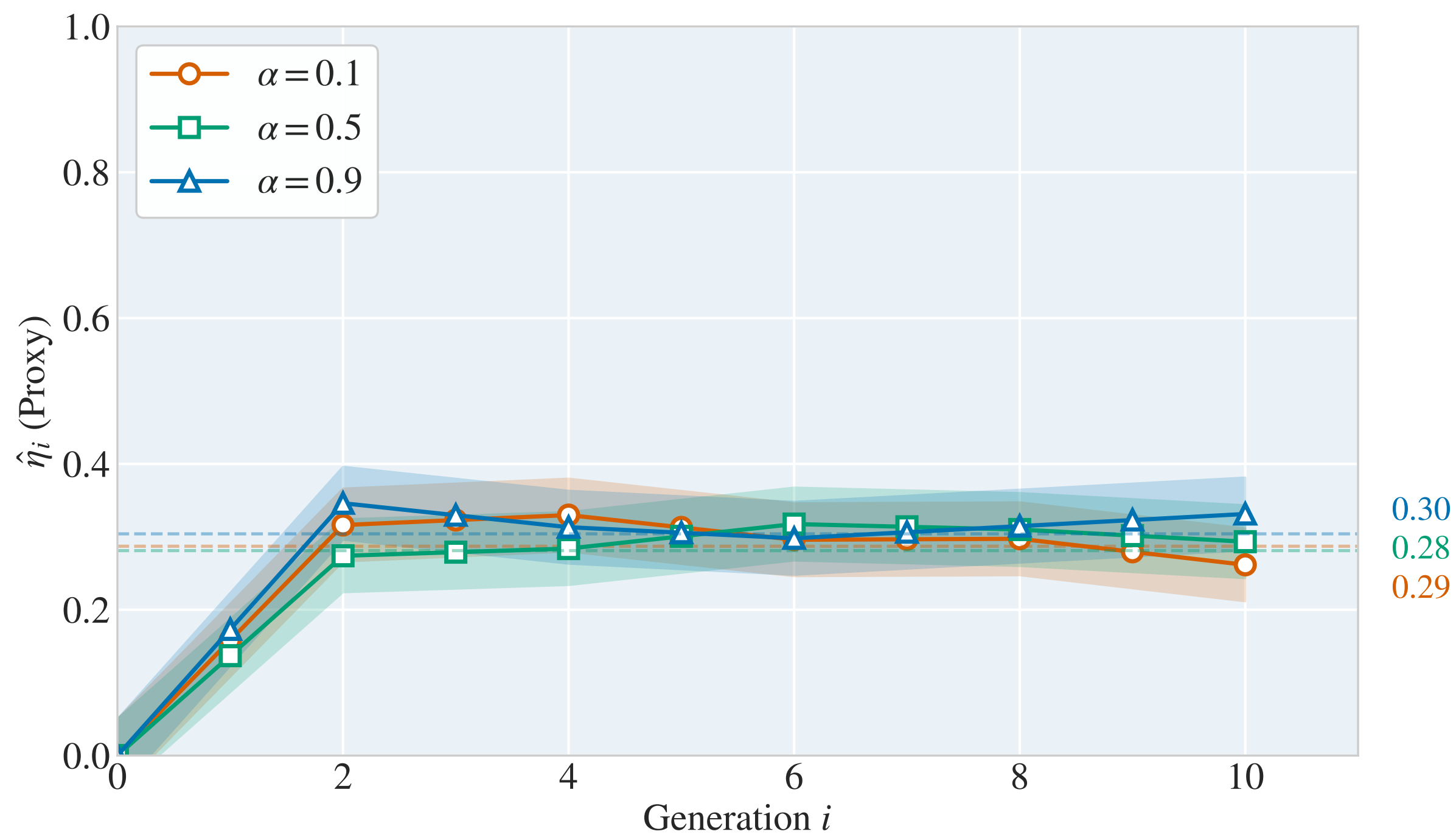
Density ratio:  $\frac{\hat{p}^{i+1}(y)}{q_i(y)} = \mathbb{E} \left[ \exp(Z_i) \mid Y_{t_0}^i = y \right]$  where  $Z_i = M_i - \frac{1}{2} \langle M_i \rangle$ ,

$$M_i = - \int_{t_0}^T e_i(Y_s^i, s) \cdot d\bar{B}_s \quad \langle M_i \rangle = \int_{t_0}^T \|e_i(Y_s^i, s)\|^2 ds, \quad \text{Var}_{\mathbb{P}_i}(M_i) = \mathbb{E}_{\mathbb{P}_i}[\langle M_i \rangle] = \varepsilon_i^2$$

Observability coefficient  $\eta_i := \frac{\text{Var}_{\mathbb{P}_i}(\mathbb{E}[M^i \mid Y_{t_0}^i])}{\text{Var}_{\mathbb{P}_i}(M_i)} = \frac{\text{Var}_{\mathbb{P}_i}(\mathbb{E}[M_i \mid Y_{t_0}^i])}{\varepsilon_i^2} \in [0, 1]$

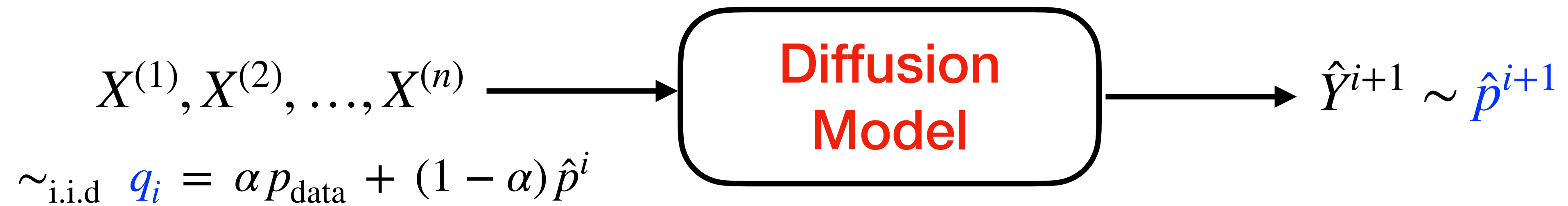


$P_{\text{data}}$  : 10-dimensional Gaussian mixture



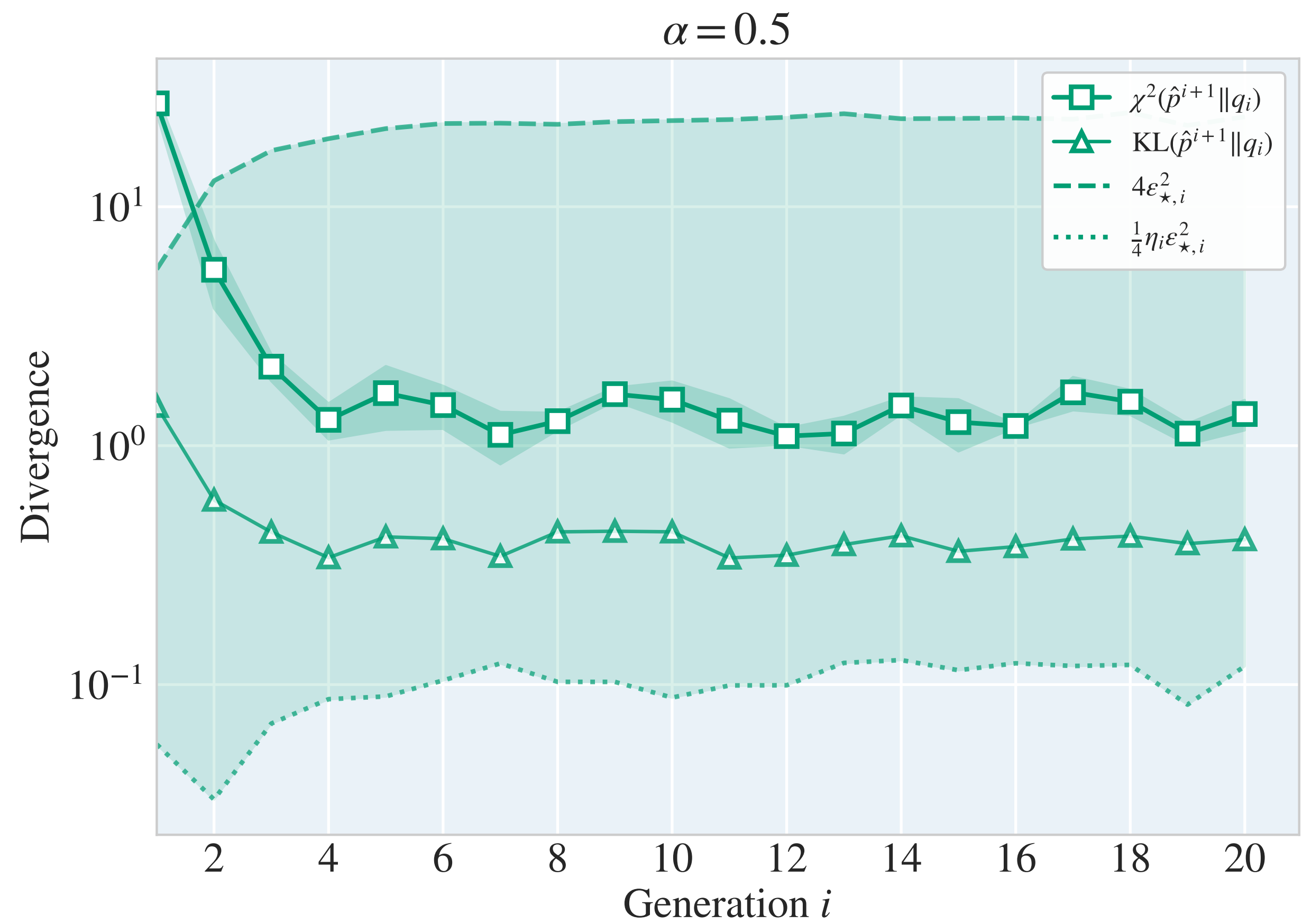
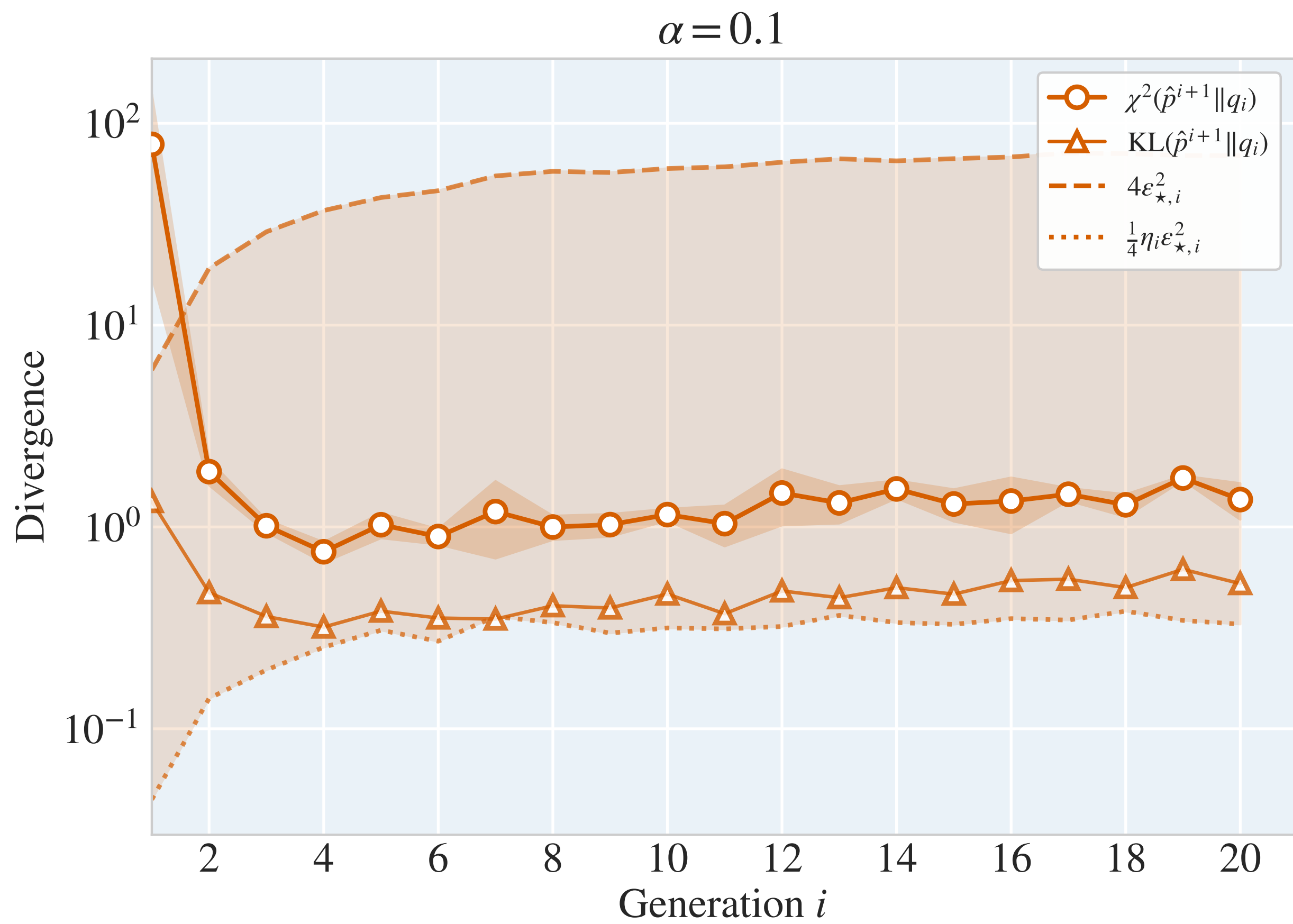
$P_{\text{data}}$  : CIFAR-10

# Two-sided Intra-generation Bound



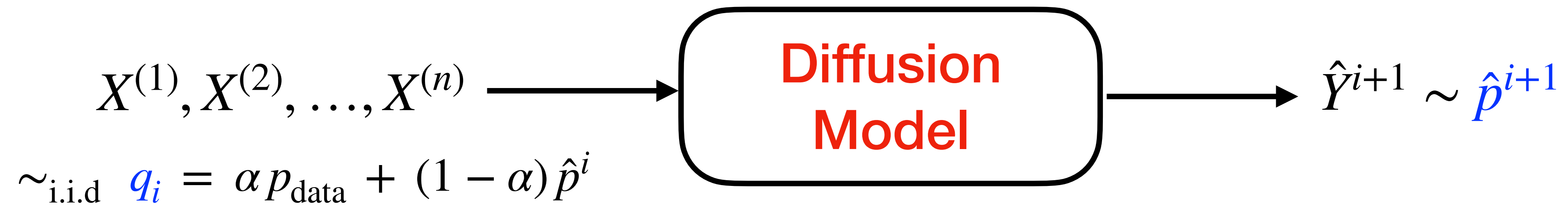
**Theorem:** Under mild assumptions, if  $\varepsilon_i^2 < \min \left\{ 1, \frac{\eta_i}{8C} \right\}$ , then

$$\eta_i \varepsilon_i^2 - C \varepsilon_i^4 \leq \chi^2(\hat{p}^{i+1} \| q_i) \leq 4 \varepsilon_i^2 + c \varepsilon_i^4$$



$p_{\text{data}}$  : 10-dimensional mixture of 5 isotropic Gaussians

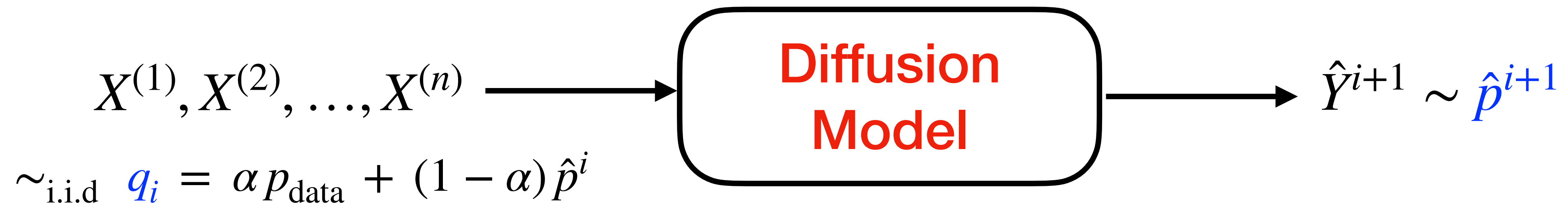
# Error Accumulation



Want to track  $D_i := \chi^2(\hat{p}^i \| p_{\text{data}})$

- Fresh data contracts accumulated divergence:  $\chi^2(q^i \| p_{\text{data}}) = (1 - \alpha)^2 \chi^2(\hat{p}^i \| p_{\text{data}})$
- Score error  $\varepsilon_i^2$  in each round  $i$  increases accumulated divergence

# Error Accumulation



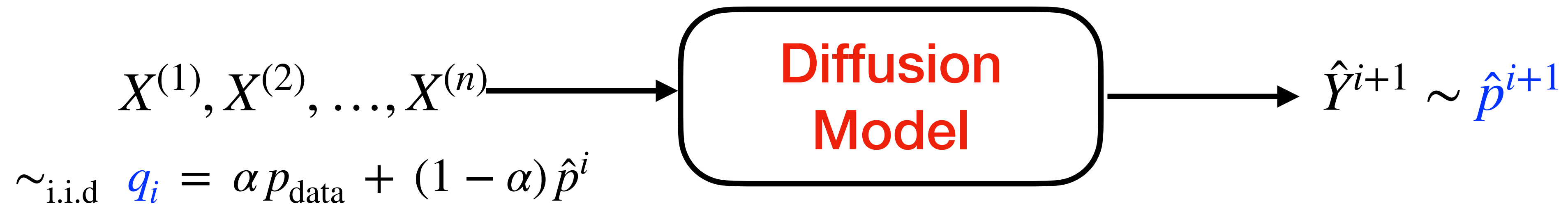
**Proposition** (Non-summable score errors).

For generations  $i \geq i_0$ , let  $\eta_i \geq \underline{\eta}$  and  $\varepsilon_i^2 < \min \left\{ 1, \frac{\eta_i}{8C} \right\}$ .

1. If  $\sum_{i \geq i_0} \varepsilon_i^2 = \infty$  then  $\sum_{i \geq 0} D_i = \infty$ .

2. If  $\varepsilon_i^2 \geq \underline{\varepsilon}^2$  and  $\eta_i \geq \underline{\eta}$  for sufficiently large  $i$ , then  $\limsup_{i \rightarrow \infty} D_i \geq C_{\alpha} \underline{\eta} \underline{\varepsilon}^2$

# Error Accumulation



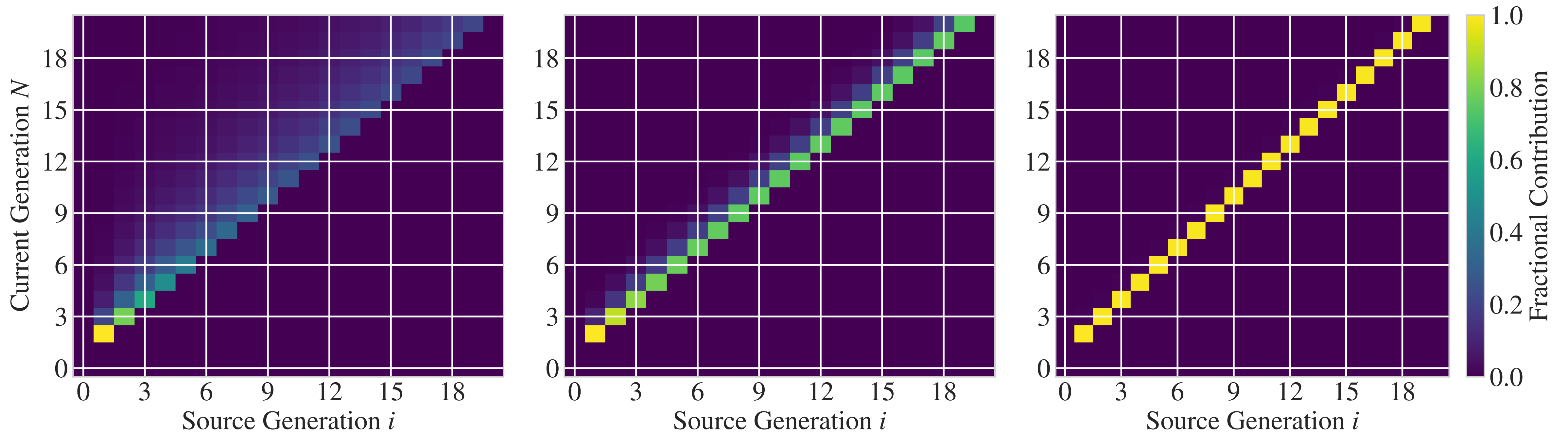
**Theorem (Summable score errors).** Assume that for generations  $i \geq i_0$ , we have  $\eta_i \geq \underline{\eta}$  and  $\varepsilon_i^2 < \min\{1, \eta_i / 8C\}$ . Also assume  $\sum_i \varepsilon_i^2 < \infty$ . Then for each  $N \geq i_0$ ,

$$D_{N+1} + C_{\text{bias}} \asymp \sum_{i=i_0}^N (1 - \alpha)^{2(N-i)} \varepsilon_i^2 + (1 - \alpha)^{2(N+1-i_0)} D_{i_0}$$

Accumulated divergence: **geometrically-discounted sum of score errors**

Theorem requires additional assumption on tail moment of  $\frac{\hat{p}^i}{p_{\text{data}}}$

Contribution of  $\varepsilon_i^2$  to  $D_N = \chi^2(\hat{p}^N \parallel p_{\text{data}})$  for  $i \leq N$



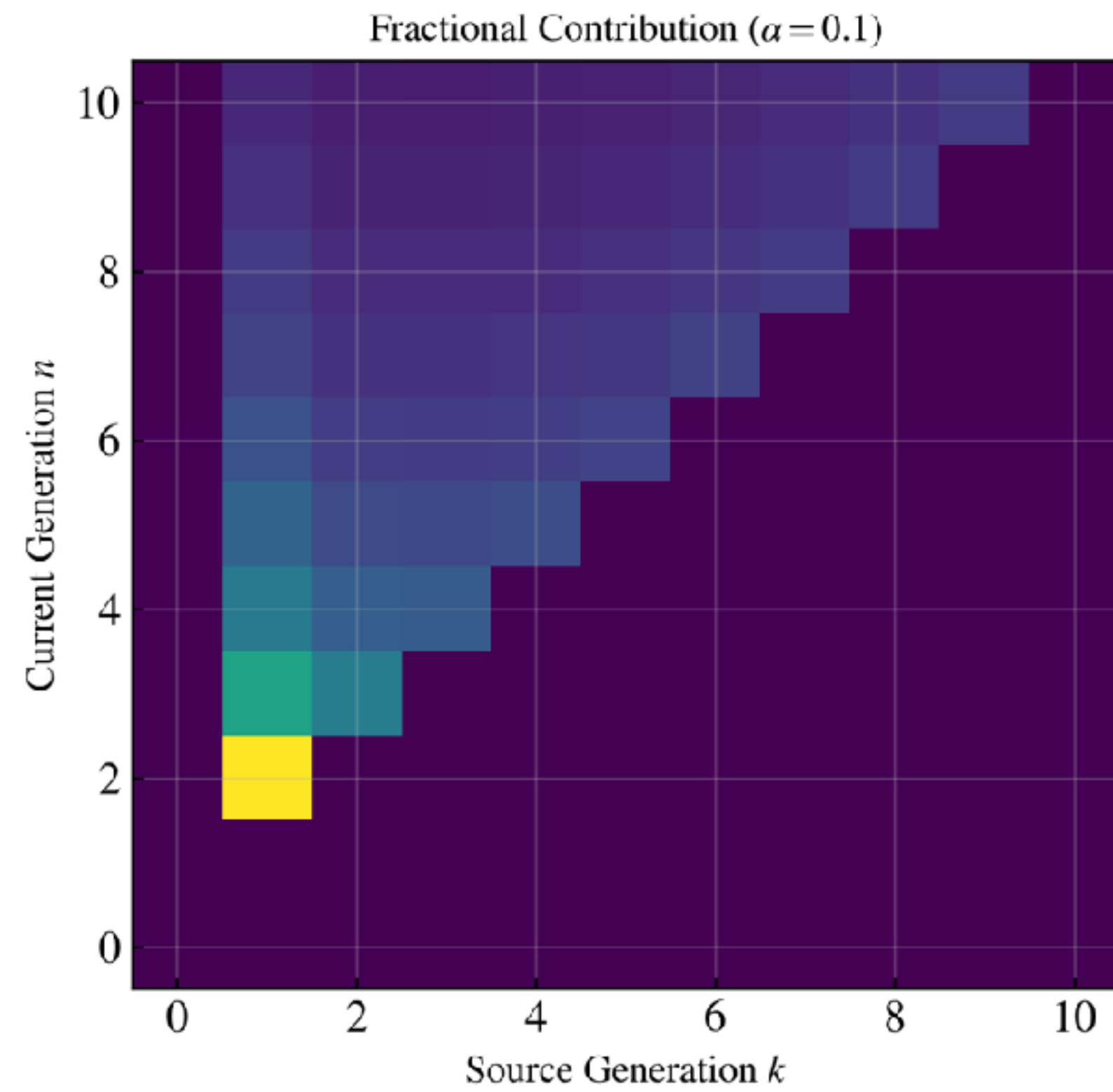
$\alpha = 0.1$

$\alpha = 0.5$

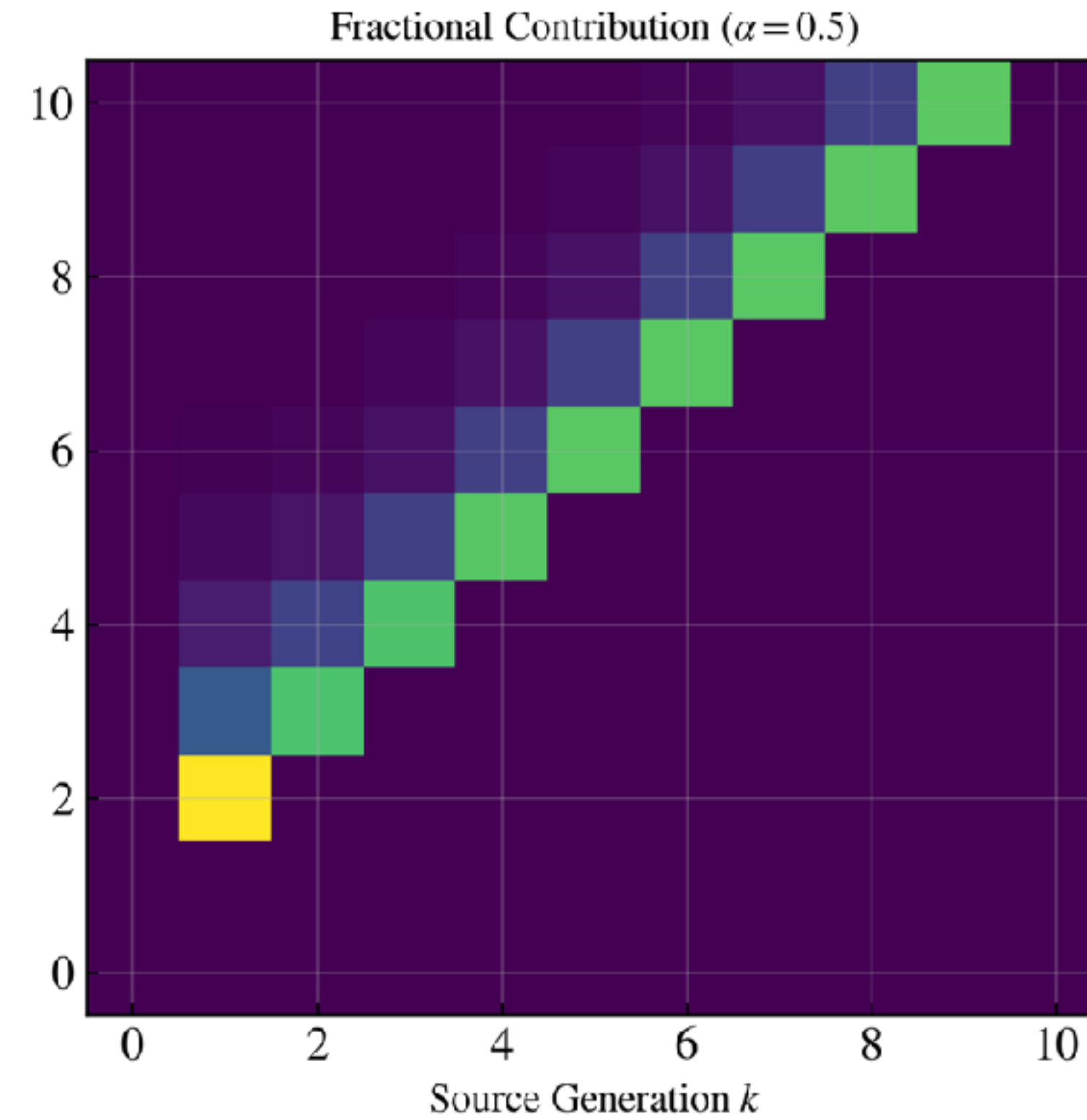
$\alpha = 0.9$

10-dimensional mixture of 5 isotropic Gaussians

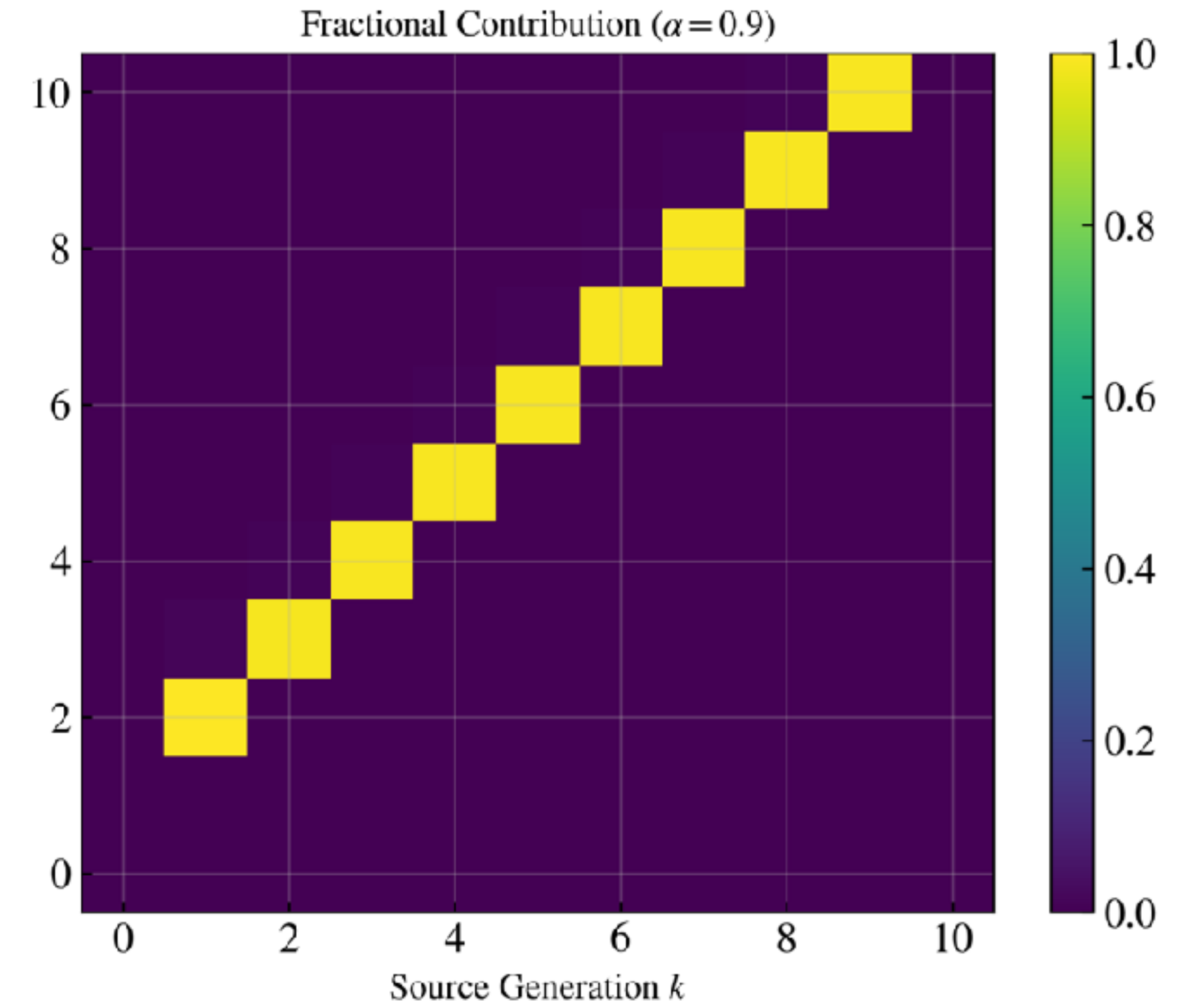
Contribution of  $\varepsilon_i^2$  to  $D_N = \chi^2(\hat{p}^N \parallel p_{\text{data}})$  for  $i \leq N$



$\alpha = 0.1$



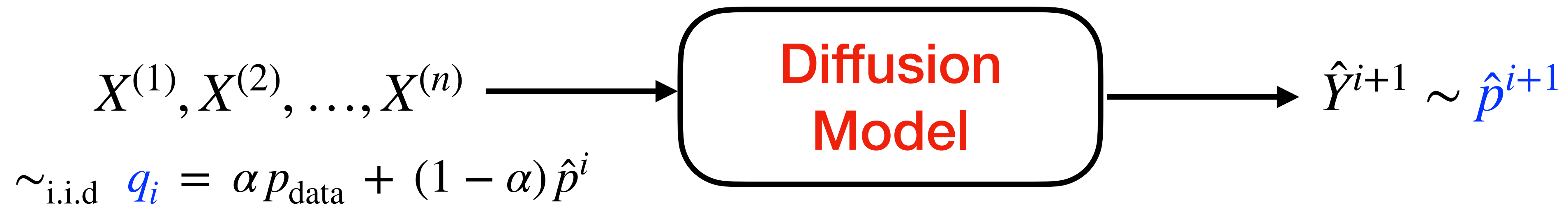
$\alpha = 0.5$



$\alpha = 0.9$

$p_{\text{data}}$  : Fashion-MNIST

# Limiting distribution as $i \rightarrow \infty$



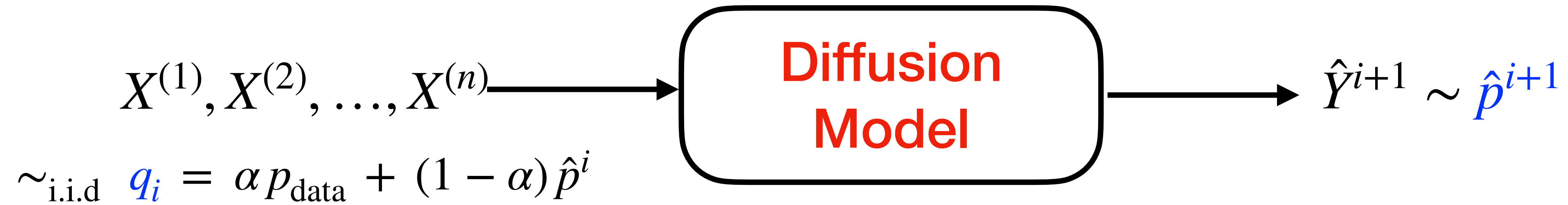
Gaussian smoothing:  $\mathcal{U}_t(p) = \text{Law}(e^{-t/2}X + \sqrt{1 - e^{-t/2}}Z), \quad X \sim p \perp Z \sim \mathcal{N}(0, I)$

Without score errors or discretization errors:  $\hat{p}^\infty = \alpha \sum_{k=0}^{\infty} (1 - \alpha)^k \mathcal{U}_{(k+1)t_0}(p_{\text{data}})$

Geometric convergence of  $\hat{p}^i$  to  $\hat{p}^\infty$  in Wasserstein-2

With score/discretization errors of magnitude  $\delta$ , concentrates on  $\mathcal{W}_2$  ball around  $\hat{p}^\infty$

# Summary



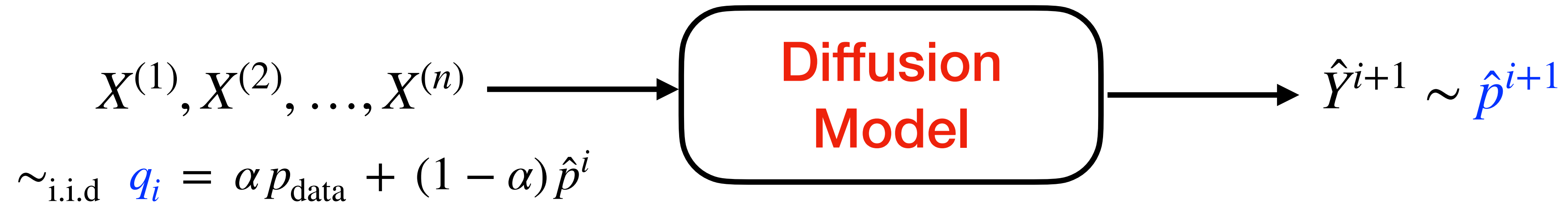
Two effects: Error **accumulation** (imperfect score learning) and Error **mitigation** (fresh data)

**Intra-generational** divergence bounds for small score errors:

- $$\underbrace{\chi^2(\hat{p}^{i+1} \| q_i)}_{\text{Intra-Generational Divergence}} \gtrsim \underbrace{\eta_i}_{\text{Observability of Errors}} \cdot \underbrace{\varepsilon_i^2}_{\text{Score Error Energy}}$$

- For small  $\varepsilon_i^2$ , we have  $\chi^2(\hat{p}^{i+1} \| q_i) \asymp \varepsilon_i^2$

# Summary

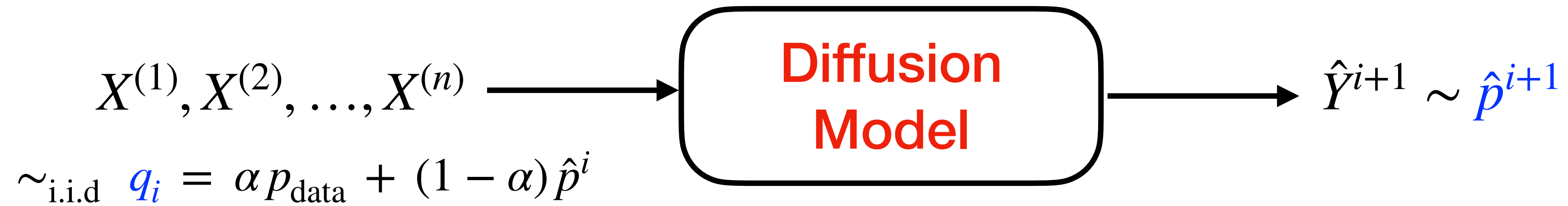


Two effects: Error **accumulation** (imperfect score learning) and Error **mitigation** (fresh data)

**Accumulated** divergence for small score errors:

- If  $\sum_{i \geq i_0} \varepsilon_i^2 < \infty$ , then  $D_{N+1} + C_{\text{bias}} \asymp \sum_{i=i_0}^N (1 - \alpha)^{2(N-i)} \varepsilon_i^2$

# Open Questions



- Divergence bounds when score errors may be large
- Moment assumptions on density ratio
- Effect of time-discretization

<https://arxiv.org/abs/2602.16601>