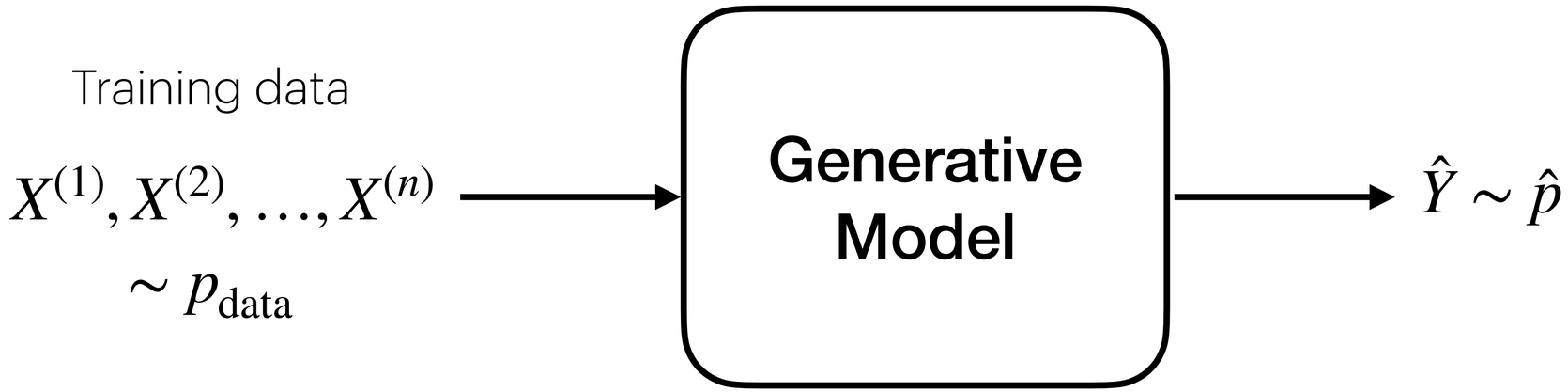
Error Propagation and Model Collapse in Diffusion Models

Ramji Venkataramanan
University of Cambridge

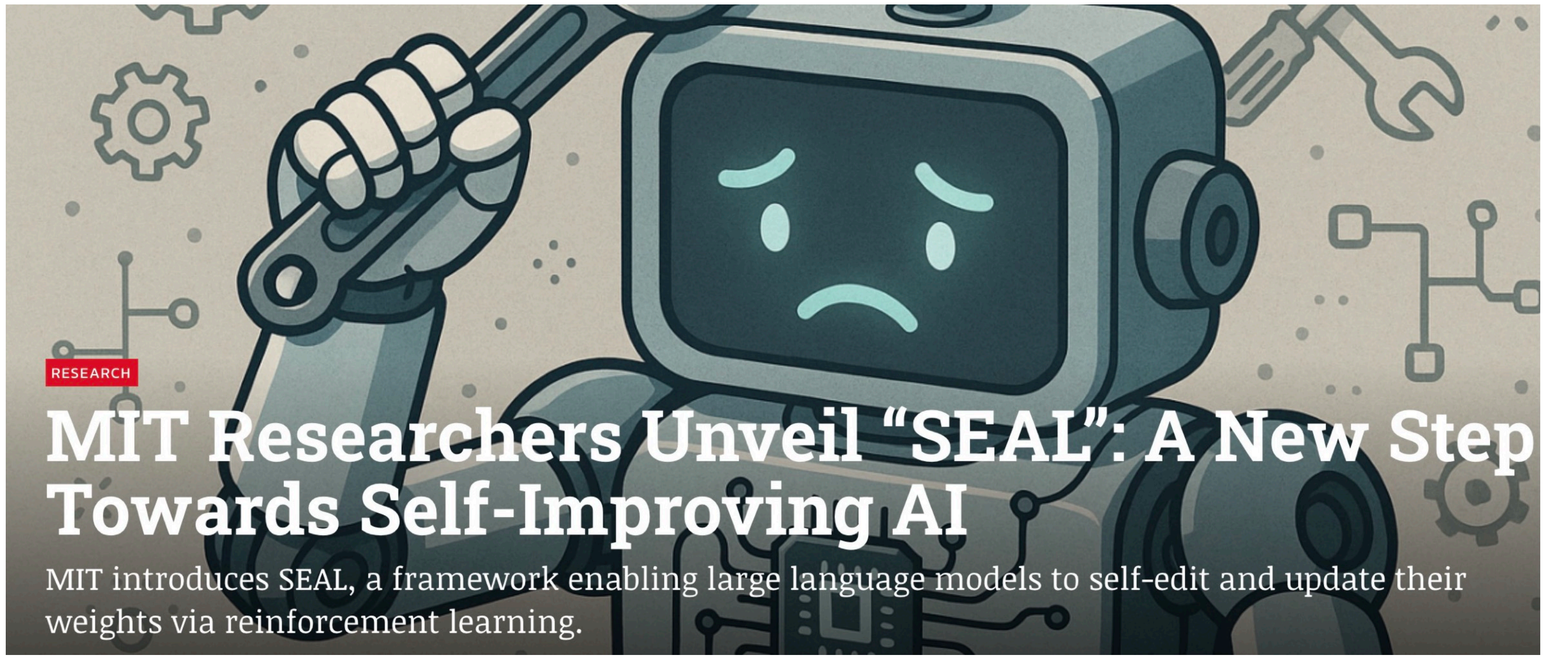
LITH Workshop 2026

Naïl Kheilfa

Richard Turner



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



RESEARCH

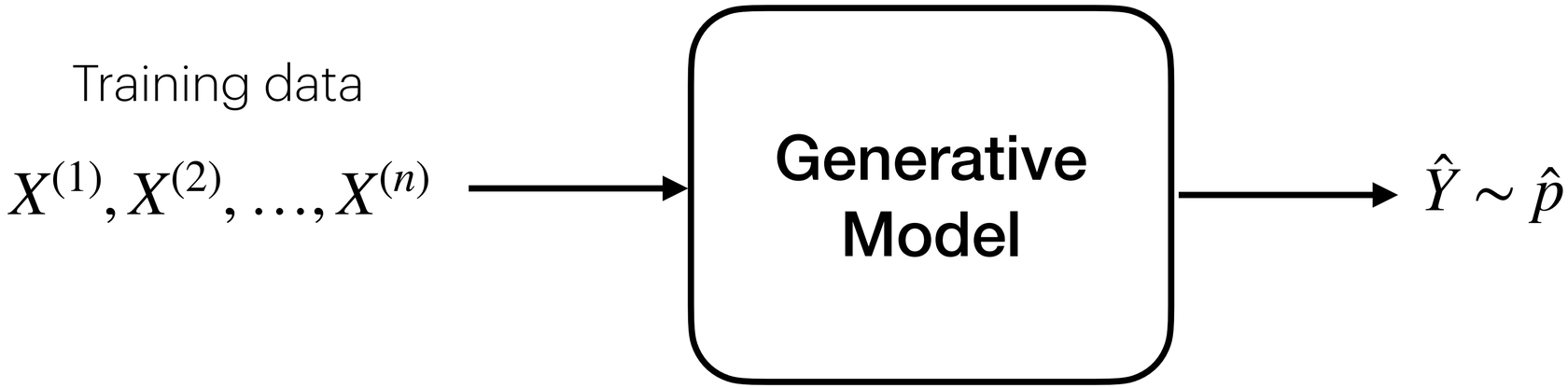
MIT Researchers Unveil “SEAL”: A New Step Towards Self-Improving AI

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,[*†] Josue Casco-Rodriguez,[*†] Lorenzo Luzi,[†] Ahmed Imtiaz Humayun,[†]
Hossein Babaei,[†] Daniel LeJeune,[‡] Ali Siahkoohi,[§] Richard G. Baraniuk[†]
[†] Department of Electrical and Computer Engineering, Rice University
[‡] Department of Statistics, Stanford University
[§] 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

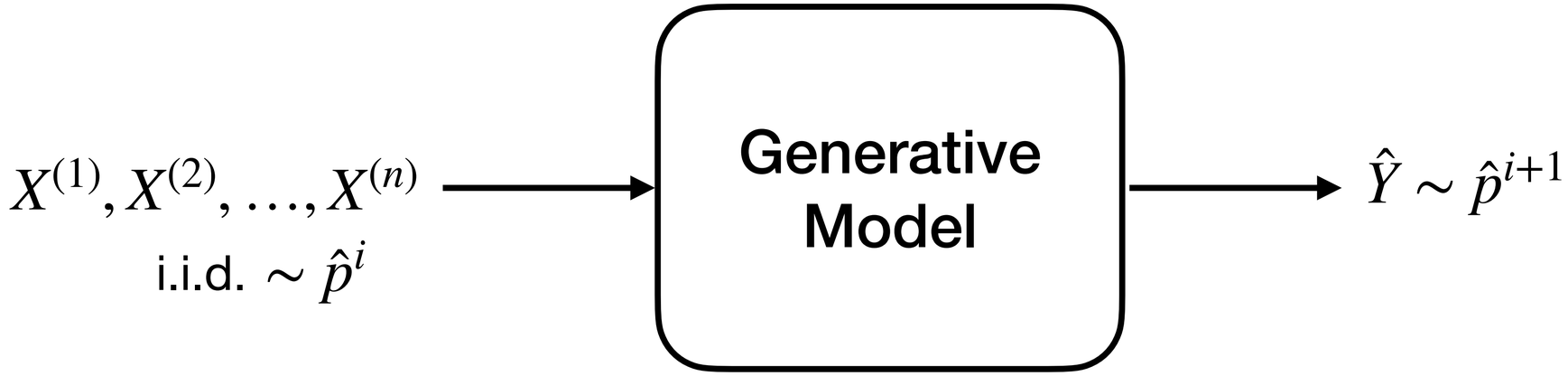
Open access

Check for updates

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

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



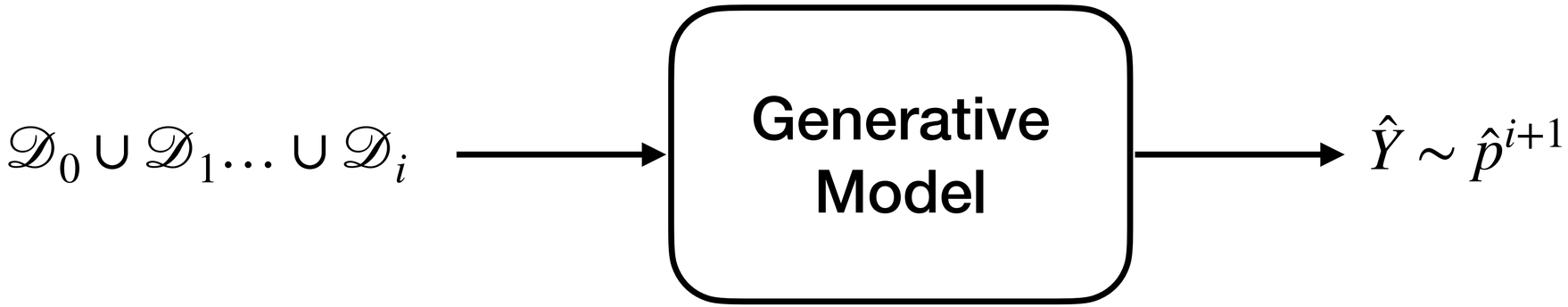
Target $p_{\text{data}} = \mathcal{N}(\mu, \sigma^2)$

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



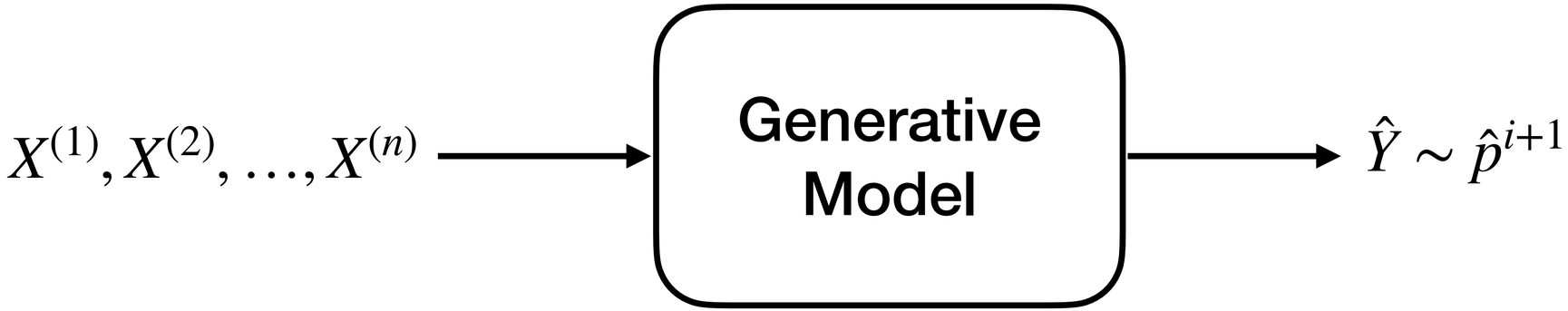
Training data for generation i : $\mathscr{D}_0 \cup \mathscr{D}_1 \ldots \cup \mathscr{D}_i$

$\mathscr{D}_0 \sim p_{\text{data}}$, $\mathscr{D}_1 \sim \hat{p}^1$, …, $\mathscr{D}_i \sim \hat{p}^i$

- 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

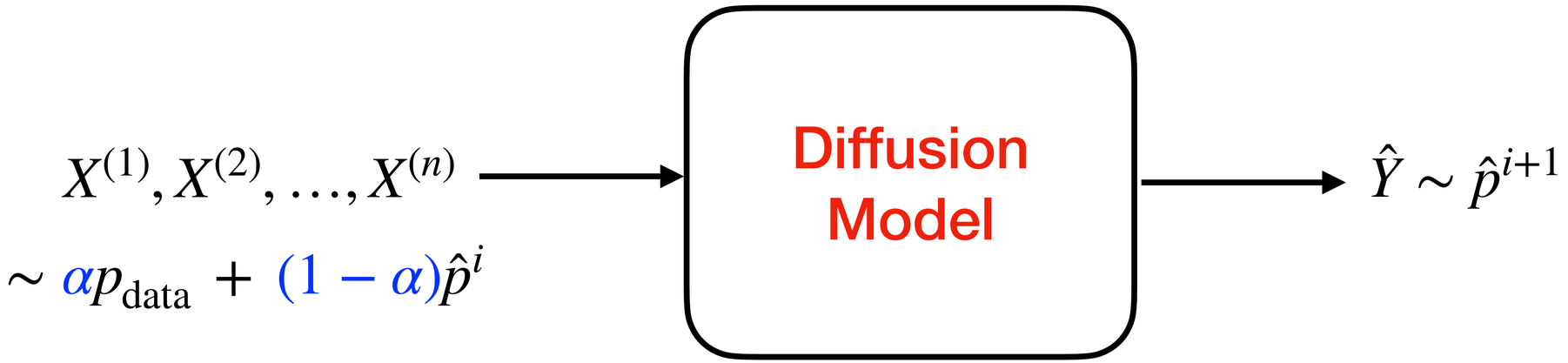


Training data for generation $i+1$:

Fraction α 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

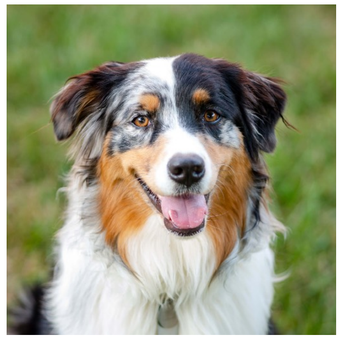


Training data for generation i : Mixture of p_{data} and $\hat{p}^i$

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

Diffusion Models

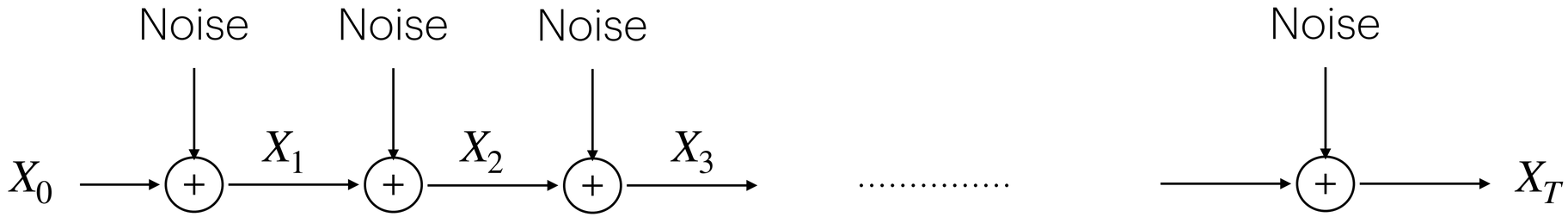
$X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ → **Diffusion Model** → $\hat{Y} \sim \hat{p}$



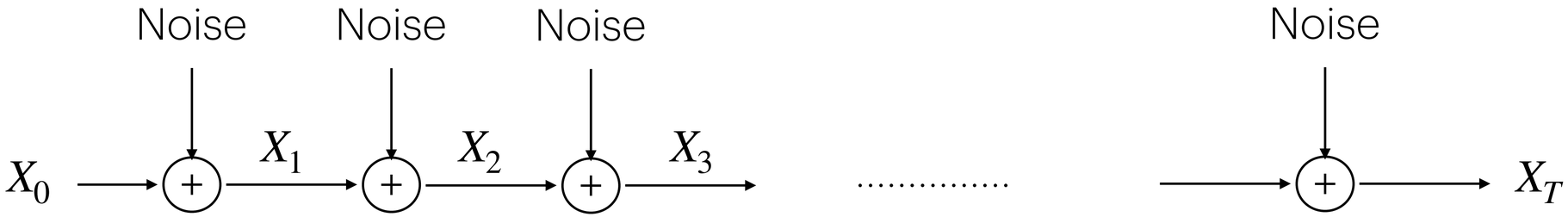
State of the art for text-to-image, video generation, protein modeling, …

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

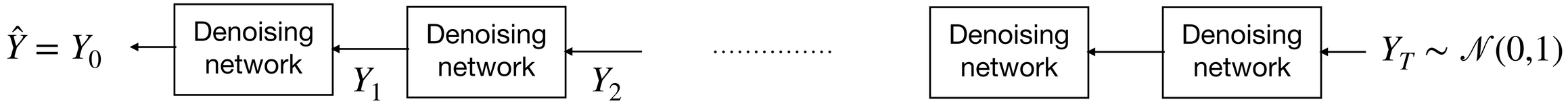
“Create a cartoon image of a panda skiing in Switzerland”



Forward noising

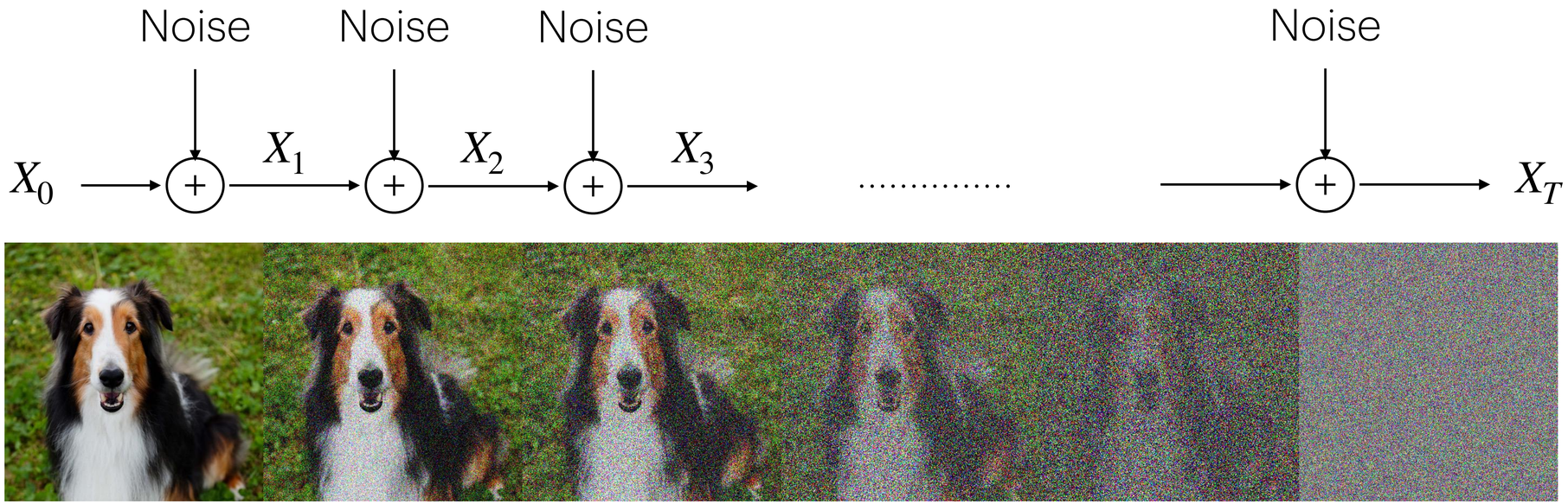


Forward noising



Backward denoising

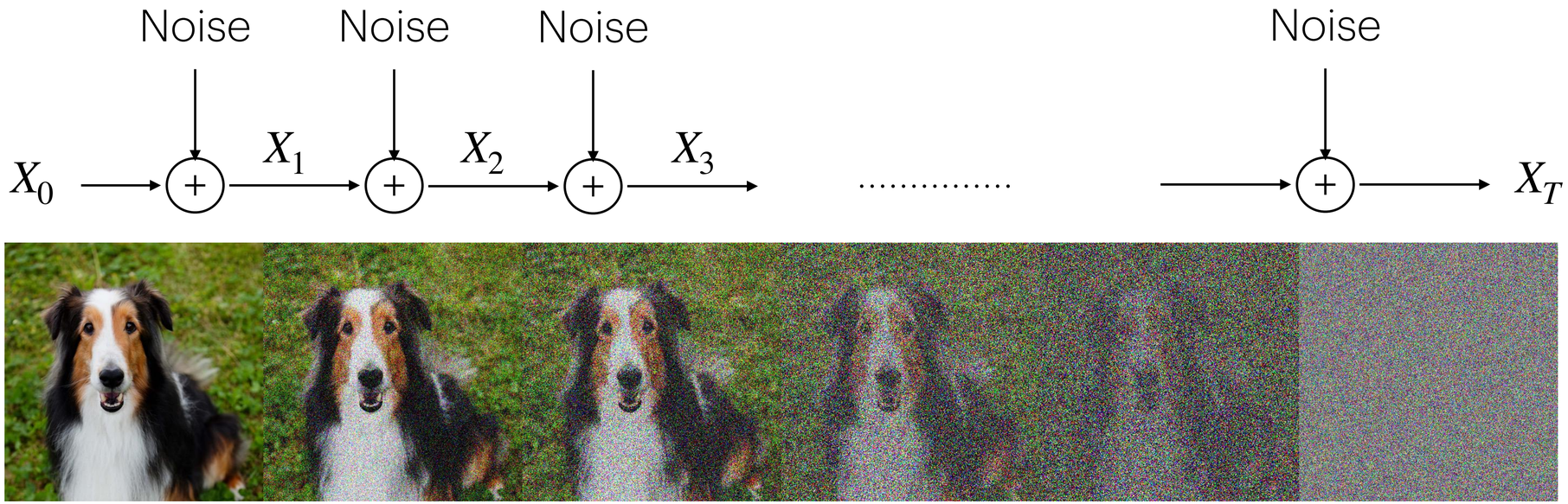
Forward Noising Process



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

$$\mathrm{d}X_t = -\frac{1}{2}X_t\,\mathrm{d}t + \mathrm{d}B_t, \qquad t \in [0,T]$$

Forward Noising Process



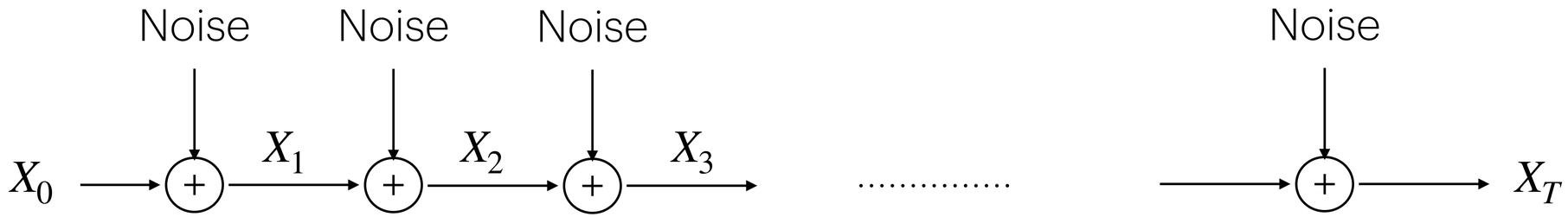
p_{data} : distribution on $\mathbb{R}^d$. Starting from $X_0 \sim p_{\text{data}}$,

$$\mathrm{d}X_t = -\frac{1}{2}X_t\,\mathrm{d}t + \mathrm{d}B_t, \qquad t \in [0,T]$$

Discretization:

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

Forward Noising SDE



Forward noising

$$\mathrm{d}X_t = -\frac{1}{2}X_t\,\mathrm{d}t + \mathrm{d}B_t, \qquad t \in [0,T], \quad X_0 \sim p_{\text{data}}$$

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

Backward Denoising

Forward SDE: $\mathrm{d}X_t = -\frac{1}{2}X_t\,\mathrm{d}t + \mathrm{d}B_t, \qquad t \in [0,T], \quad X_0 \sim p_{\text{data}}$

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

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

Backward Denoising

Forward SDE: $\mathrm{d}X_t = -\frac{1}{2}X_t\,\mathrm{d}t + \mathrm{d}B_t, \qquad t \in [0,T], \quad X_0 \sim p_{\text{data}}$

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

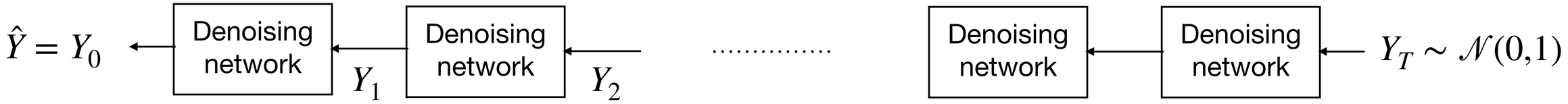
Reverse SDE: $\mathrm{d}Y_s = \left[-\frac{1}{2}Y_s - \mathsf{s}(Y_s, s)\right]\mathrm{d}s + \mathrm{d}\bar{B}_s, \qquad Y_T \sim q_T$

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 $q_T \approx \mathcal{N}(0,1)$ $\Rightarrow$

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

Sampling via Learned Reverse SDE

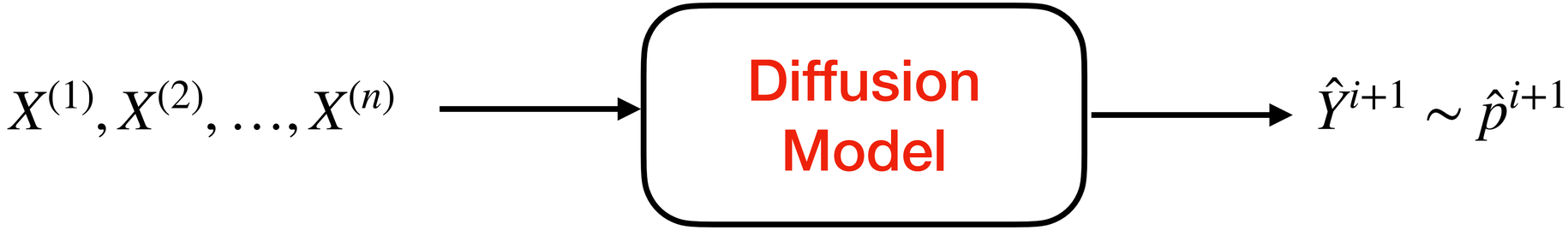


Backward denoising

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

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

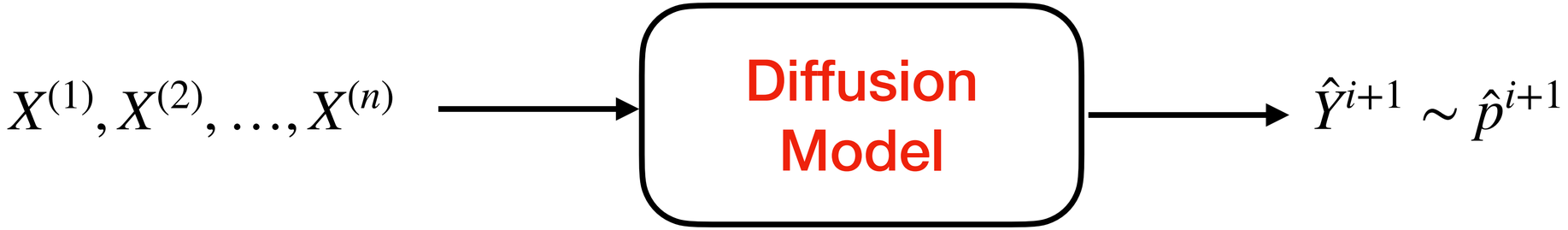
Recursive Training Model



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

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

Recursive Training Model



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

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

Track the recursion using two χ^2-divergences:

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

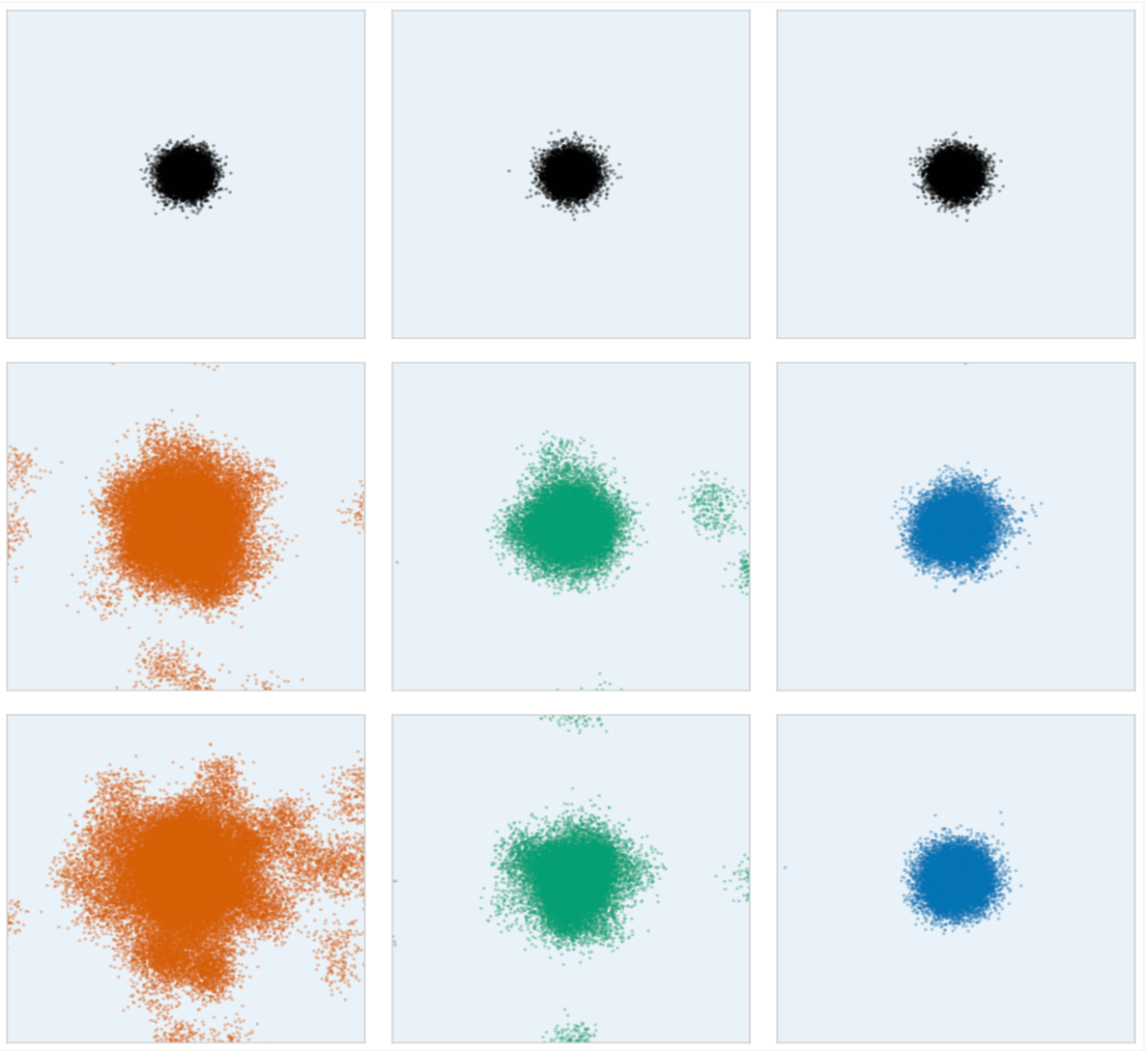
$$\chi^2(\hat{p}^{i+1} \| q_i) = \mathbb{E}_{q_i}\left[\left(\frac{\hat{p}^{i+1}}{q_i} - 1\right)^2\right]$$

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

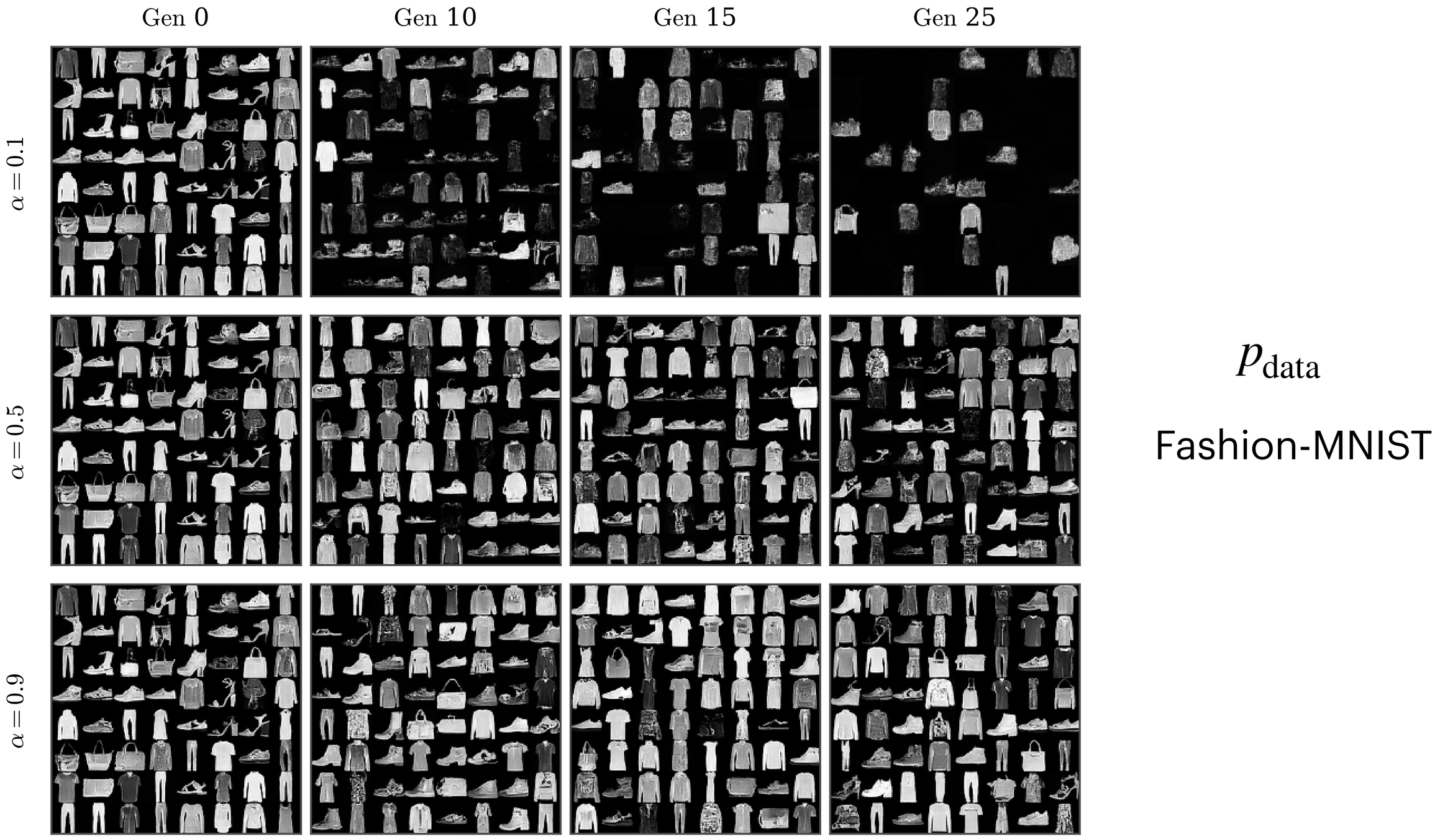
Generation 0

Generation 10

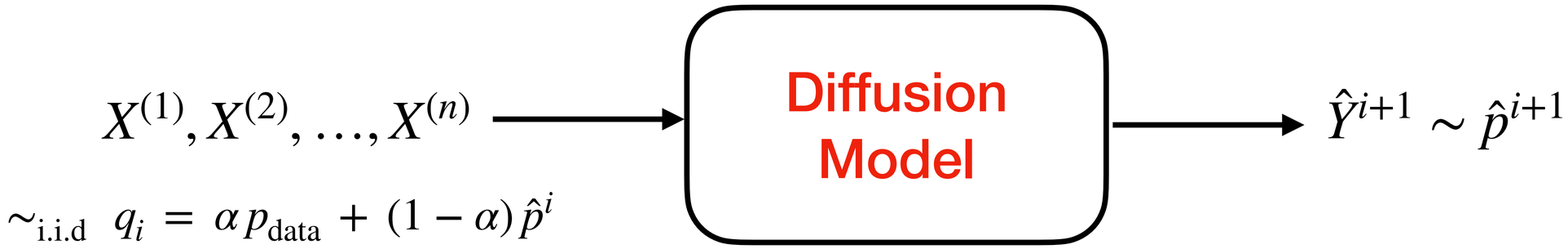
Generation 20



$\alpha = 0.1$ $\alpha = 0.5$ $\alpha = 0.9$


Gen 0
Gen 10
Gen 15
Gen 25
$\alpha = 0.1$
$\alpha = 0.5$
$\alpha = 0.9$
p_{data}
Fashion-MNIST

Path Laws



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

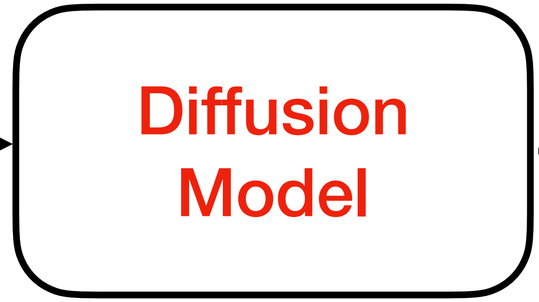
Learned Reverse SDE: $\mathrm{d}\hat{Y}_s^i = \left[-\frac{1}{2}\hat{Y}_s^i - \mathrm{s}_\theta(\hat{Y}_s^i, s)\right]\mathrm{d}s + \mathrm{d}\bar{B}_s\,, \qquad \hat{Y}_T^i \sim \mathcal{N}(0,1)$

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

- **Path Laws** $\mathbb{P}_i = \mathrm{Law}\left((Y_s^i)_{s\in[t_0,T]}\right), \quad \hat{\mathbb{P}}_i = \mathrm{Law}\left((\hat{Y}_s^i)_{s\in[t_0,T]}\right)$
- **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

$$X^{(1)}, X^{(2)}, \ldots, X^{(n)} \sim_{\text{i.i.d}} q_i = \alpha\, p_{\text{data}} + (1-\alpha)\,\hat{p}^i$$



$$\hat{Y}^{i+1} \sim \hat{p}^{i+1}$$

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

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

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

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

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

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

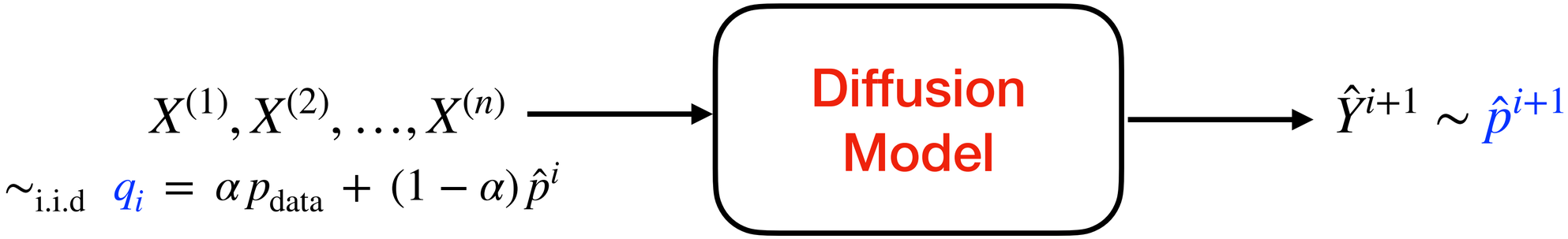
Under suitable assumptions, minimax score estimation error:

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

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

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

Intra-generation upper bound



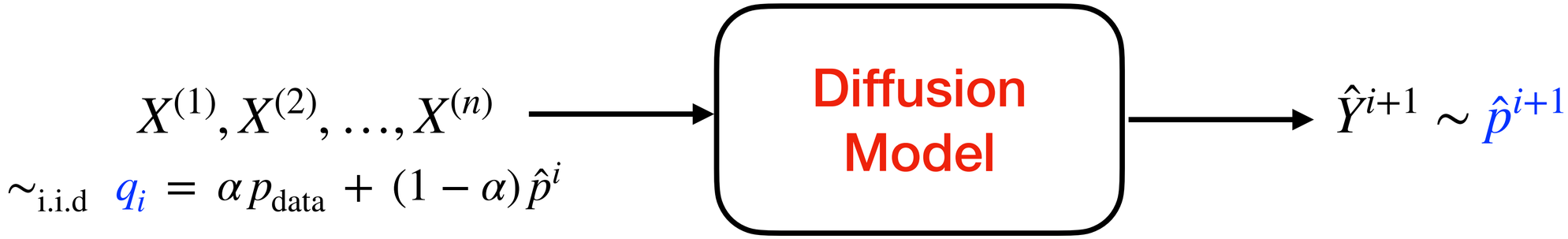
Learned Reverse SDE: $\mathrm{d}\hat{Y}_s = \left[-\frac{1}{2}\hat{Y}_s - \mathrm{s}(\hat{Y}_s, s) + e_i(\hat{Y}_s, s)\right]\mathrm{d}s + \mathrm{d}\bar{B}_s, \quad \hat{Y}_T \sim \mathcal{N}(0,1) \quad (\hat{\mathbb{P}}^i)$

Proposition: Under mild assumptions,

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

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

Intra-generation upper bound



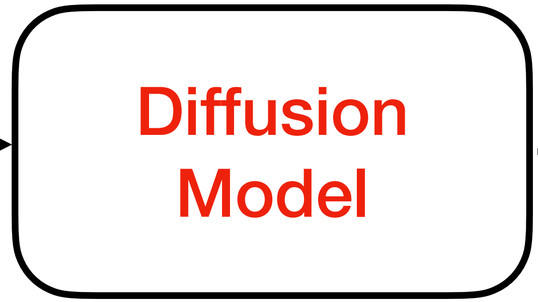
Proposition: $\mathrm{KL}(\hat{p}^{\,i+1} \| q_i) \le \mathrm{KL}(\hat{\mathbb{P}}_i \| \mathbb{P}_i) = \frac{1}{2}\,\hat{\varepsilon}_i^2$

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

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

Girsanov path density

$$X^{(1)}, X^{(2)}, \ldots, X^{(n)} \sim_{\text{i.i.d}} q_i = \alpha\, p_{\text{data}} + (1-\alpha)\,\hat{p}^i$$



$$\hat{Y}^{i+1} \sim \hat{p}^{i+1}$$

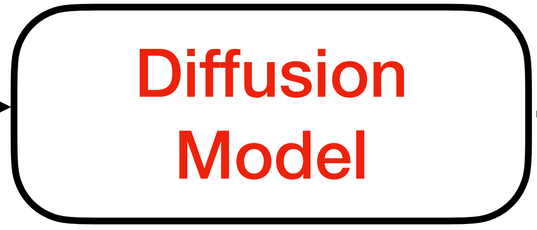
Ideal Reverse SDE: $\mathrm{d}Y_s^i = \left[-\frac{1}{2}Y_s^i - \mathbf{s}(Y_s^i, s)\right]\mathrm{d}s + \mathrm{d}\bar{B}_s\,, \qquad Y_T^i \sim \mathcal{N}(0,1) \quad (\mathbb{P}^i)$

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

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

Ratio of Marginals

$$X^{(1)}, X^{(2)}, \ldots, X^{(n)} \sim_{\text{i.i.d}} q_i = \alpha\, p_{\text{data}} + (1-\alpha)\,\hat{p}^i$$



$$\hat{Y}^{i+1} \sim \hat{p}^{i+1}$$

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

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

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

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

Intra-generation Lower Bound

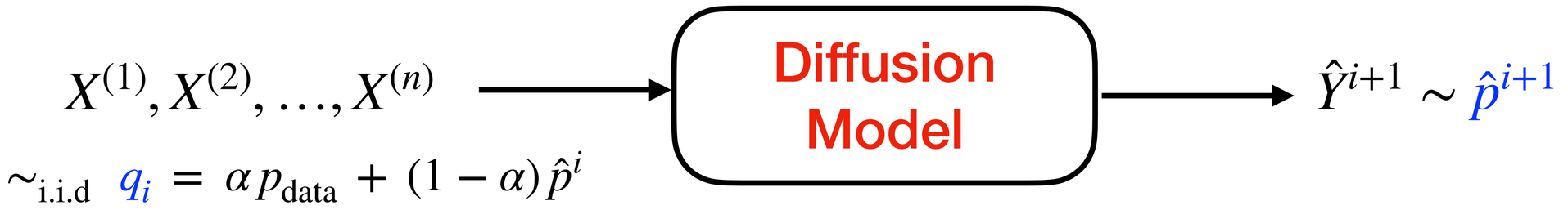
Proposition

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

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

η_i captures the observability of the score error at t_0

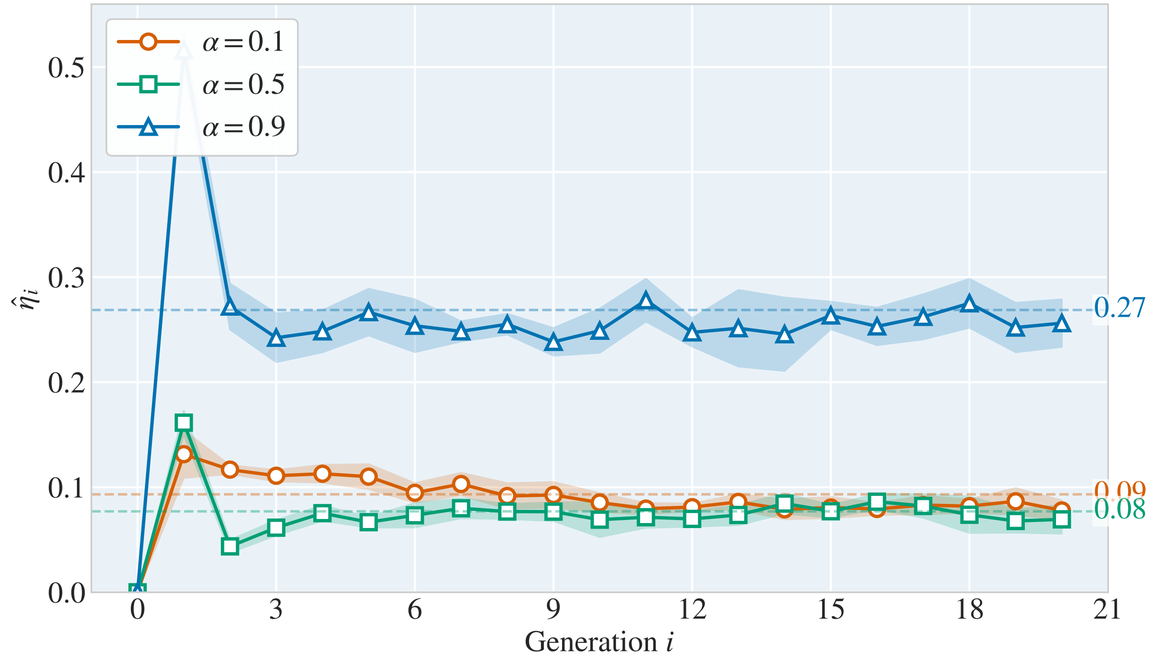
Observability of Errors



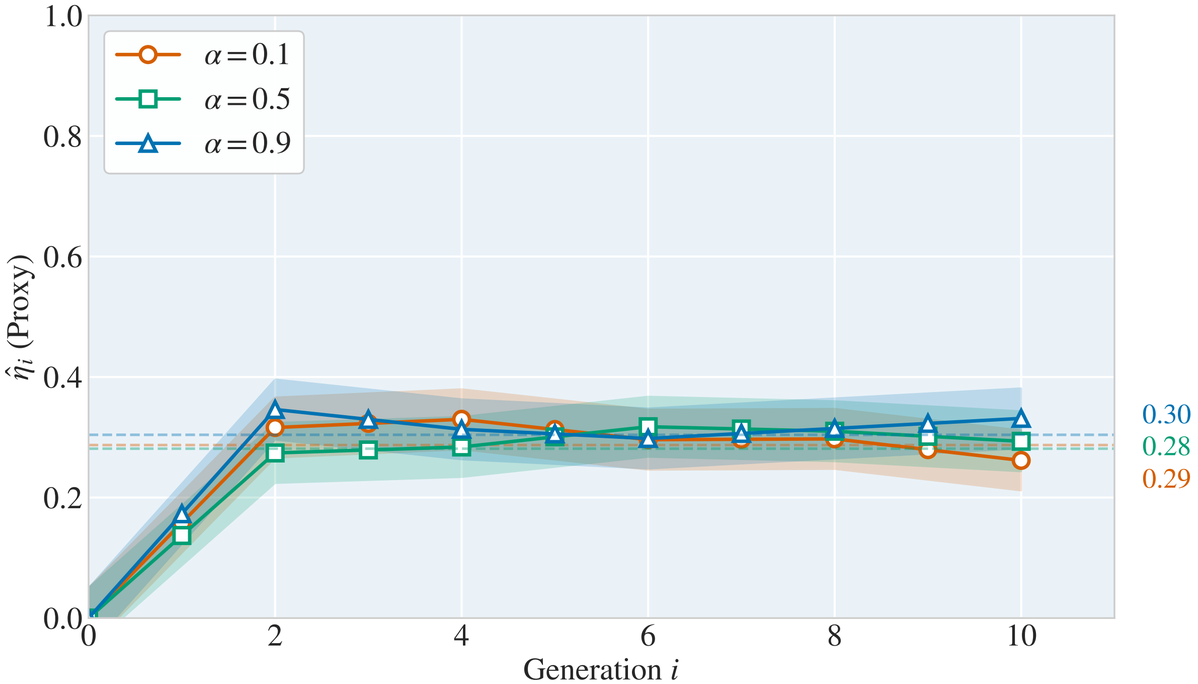
Density ratio: $\dfrac{\hat{p}^{i+1}(y)}{q_i(y)} = \mathbb{E}\left[\exp(Z_i) \mid Y^i_{t_0} = y\right]$ where $Z_i = M_i - \frac{1}{2}\langle M_i \rangle$,

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

Observability coefficient $\eta_i := \dfrac{\mathrm{Var}_{\mathbb{P}_i}\left(\mathbb{E}[M^i \mid Y^i_{t_0}]\right)}{\mathrm{Var}_{\mathbb{P}_i}(M_i)} = \dfrac{\mathrm{Var}_{\mathbb{P}_i}\left(\mathbb{E}[M_i \mid Y^i_{t_0}]\right)}{\varepsilon_i^2} \in [0,1]$

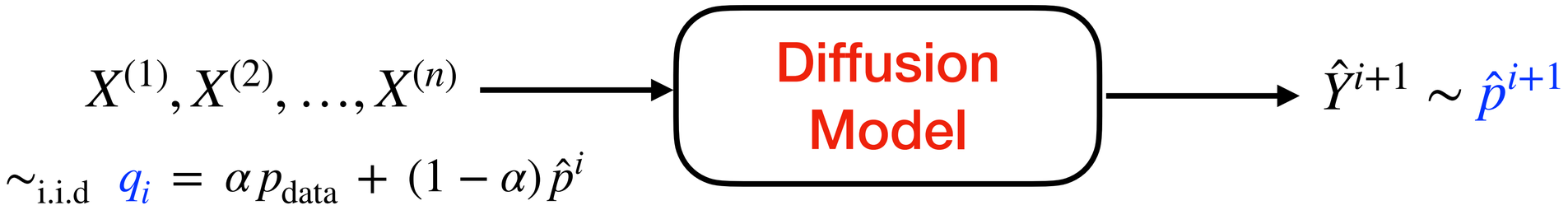


p_{data} : 10-dimensional Gaussian mixture



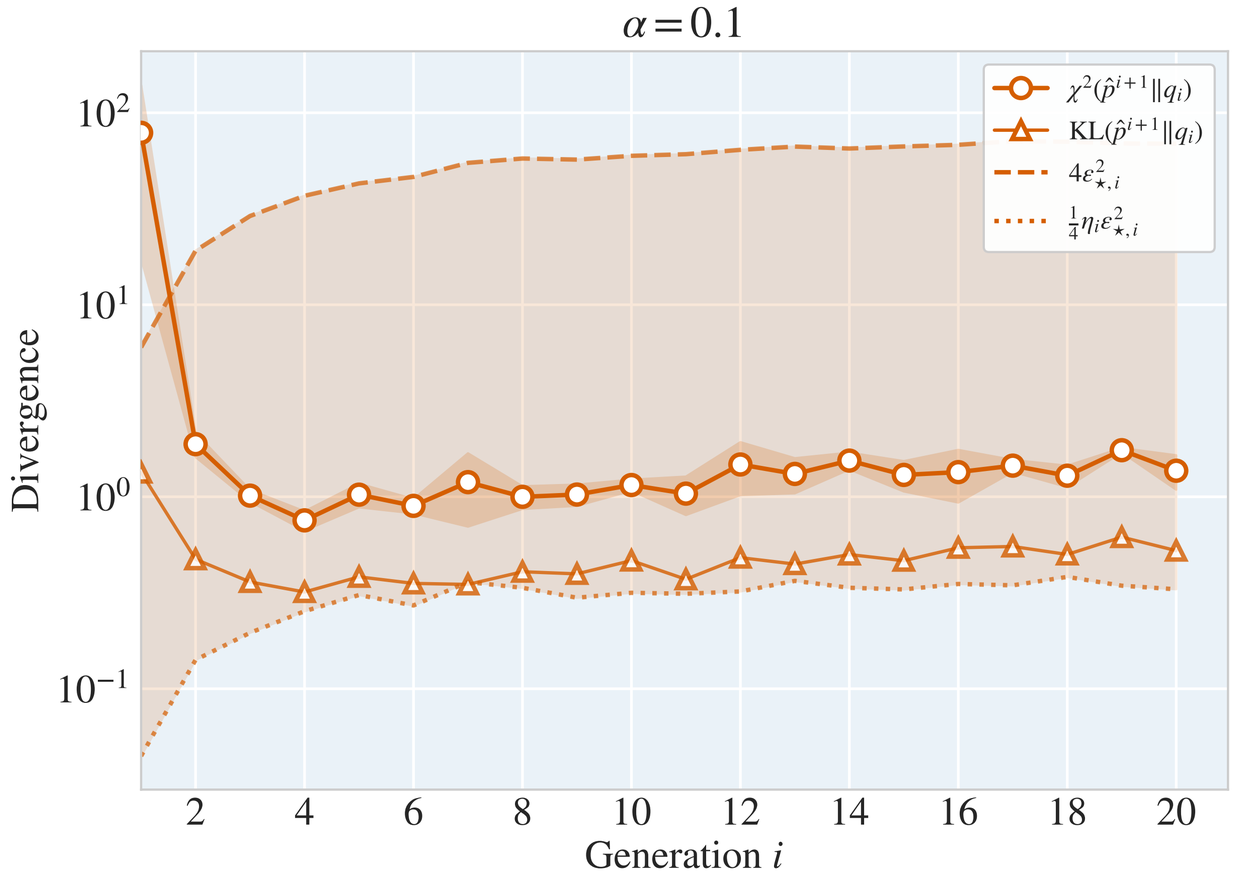
p_{data} : CIFAR-10

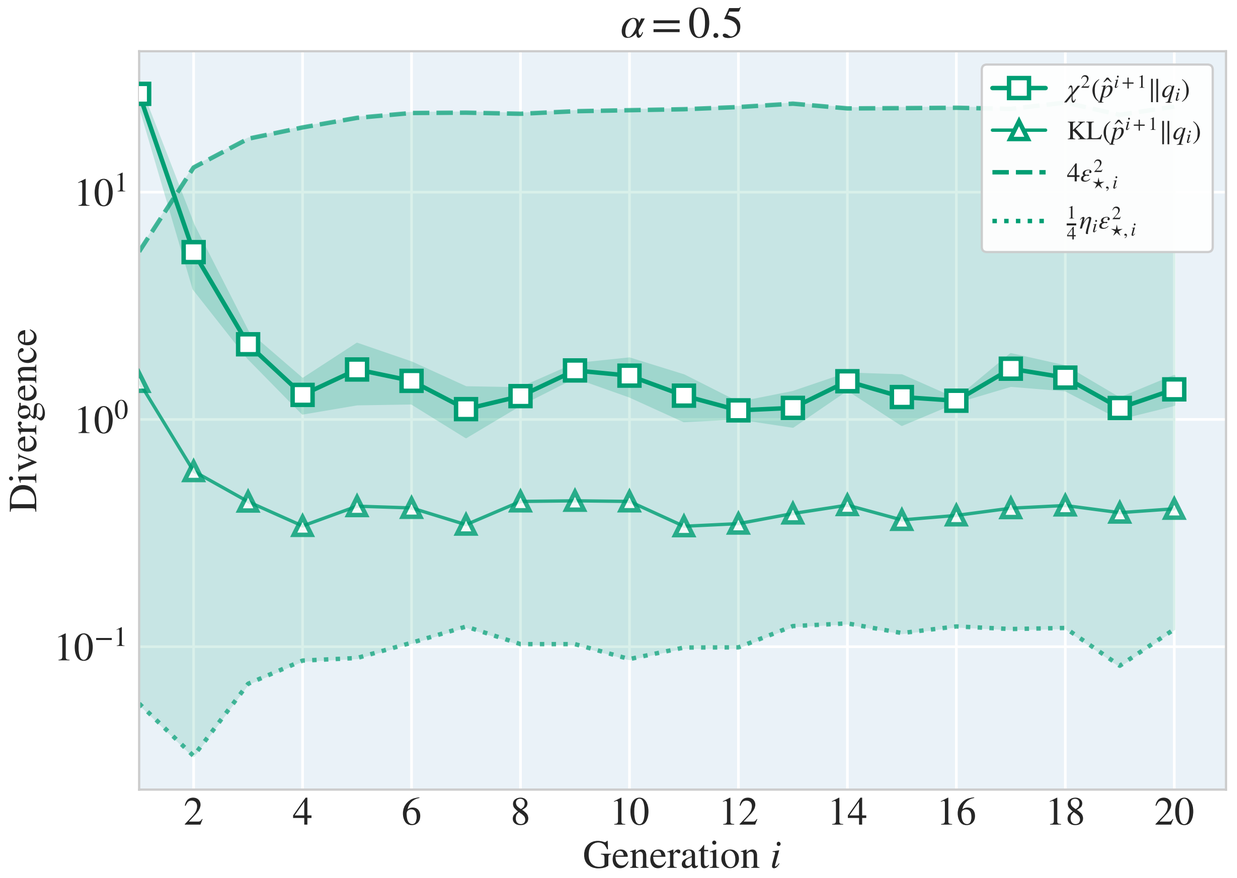
Two-sided Intra-generation Bound



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

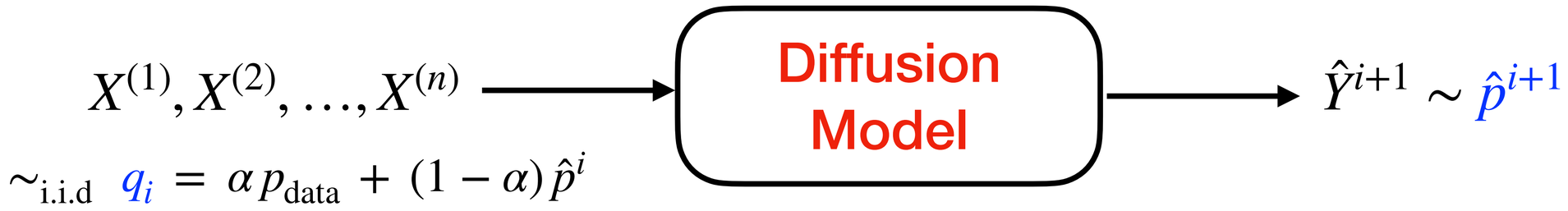
$$\eta_i\, \varepsilon_i^2 \; - C\varepsilon_i^4 \;\leq\; \chi^2(\hat{p}^{i+1} \| q_i) \;\leq\; 4\varepsilon_i^2 \;+\; c\varepsilon_i^4$$





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

Error Accumulation



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

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

Error Accumulation

Proposition (Non-summable score errors).

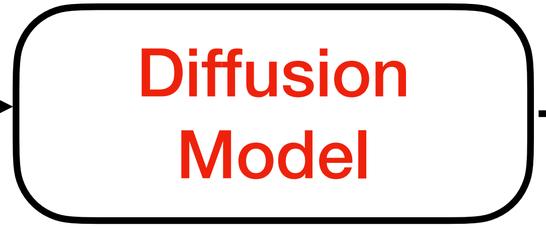
For generations $i \geq i_0$, let $\eta_i \geq \underline{\eta}$ and $\varepsilon_i^2 < \min\left\{1, \frac{\eta_i}{8C}\right\}$.

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

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

Error Accumulation

$$X^{(1)}, X^{(2)}, \ldots, X^{(n)} \sim_{\text{i.i.d}} q_i = \alpha\, p_{\text{data}} + (1-\alpha)\,\hat{p}^i$$



$$\hat{Y}^{i+1} \sim \hat{p}^{i+1}$$

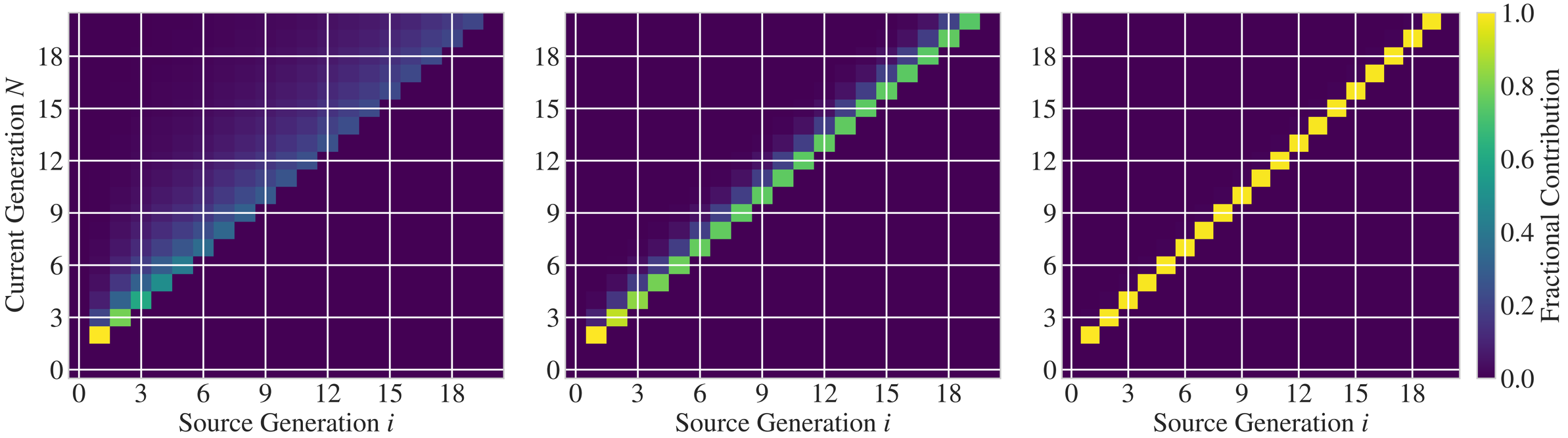
Theorem (Summable score errors). Assume that for generations $i \geq i_0$, we have $\eta_i \geq \underline{\eta}$ and $\varepsilon_i^2 < \min\{1, \eta_i / 8C\}$. Also assume $\sum_i \varepsilon_i^2 < \infty$. Then for each $N \geq i_0$,

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

Accumulated divergence: geometrically-discounted sum of score errors

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

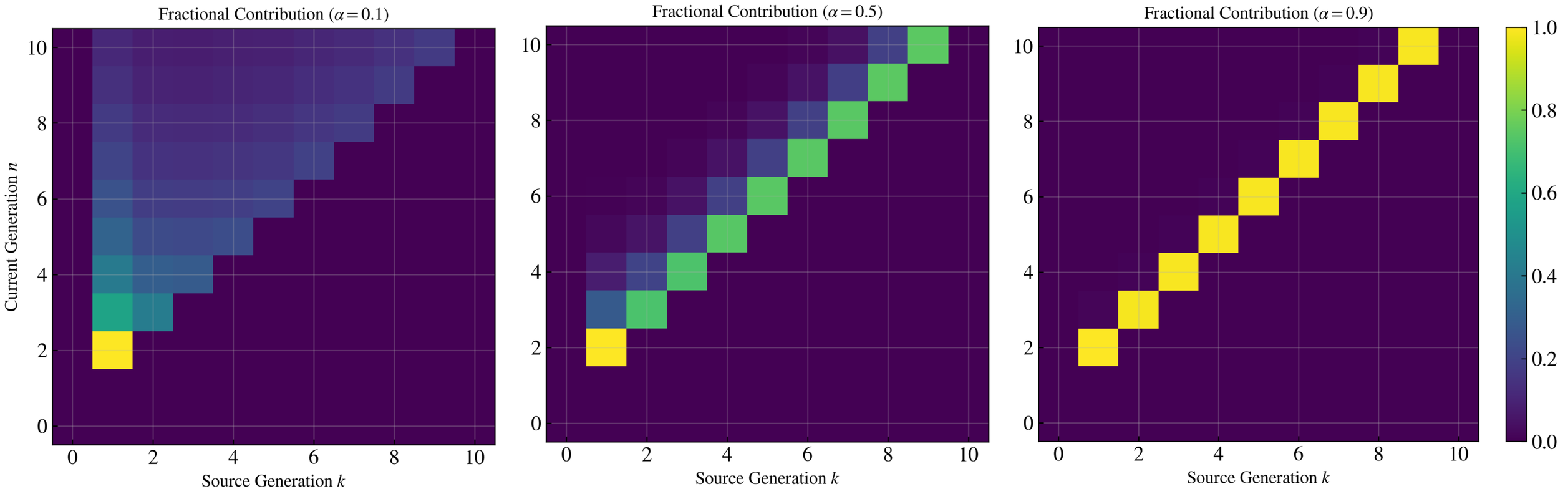
Contribution of ε_i^2 to $D_N = \chi^2(\hat{p}^N \,\|\, p_{\text{data}})$ for $i \leq N$



$\alpha = 0.1$ $\alpha = 0.5$ $\alpha = 0.9$

10-dimensional mixture of 5 isotropic Gaussians

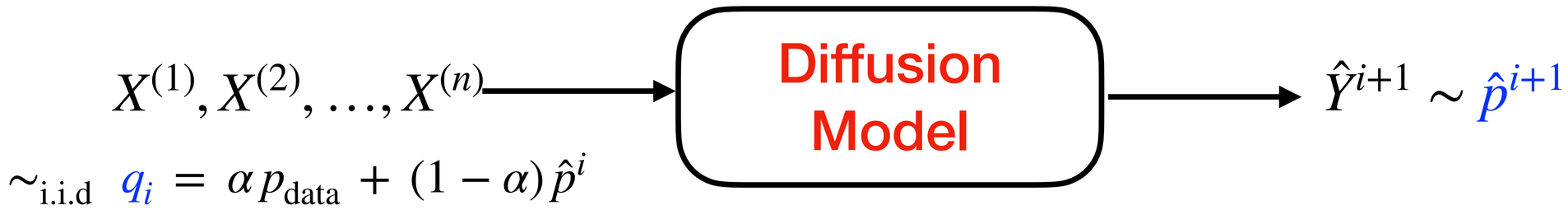
Contribution of ε_i^2 to $D_N = \chi^2(\hat{p}^N \,\|\, p_{\text{data}})$ for $i \le N$



$\alpha = 0.1$ $\alpha = 0.5$ $\alpha = 0.9$

p_{data} : Fashion-MNIST

Summary



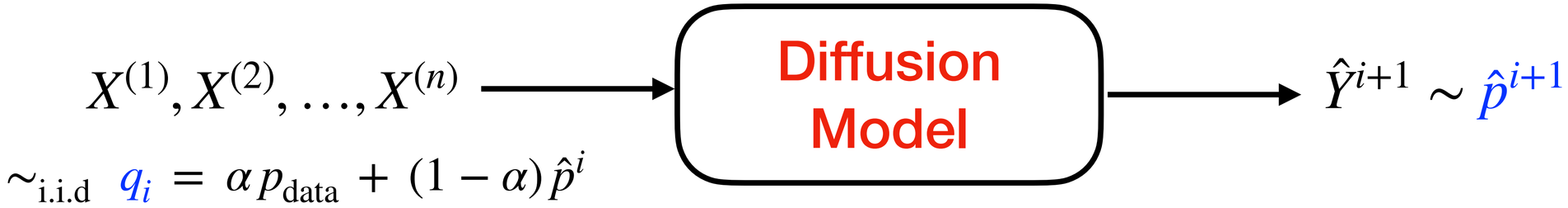
Two effects: Error accumulation (imperfect score learning) and Error mitigation (fresh data)

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

Summary

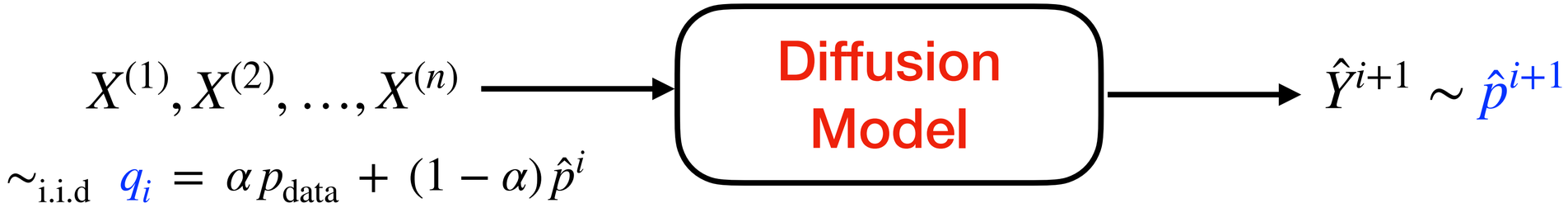


Two effects: Error accumulation (imperfect score learning) and Error mitigation (fresh data)

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
- What is the limiting distribution?

https://arxiv.org/abs/2602.16601