

扩散模型 (DIFFUSION MODEL)

黄政宇

北京大学北京国际数学研究中心
北京大学国际机器学习研究中心



生成模型

➤ 应用：视频、音乐生成





生成模型

➤ 目标分布

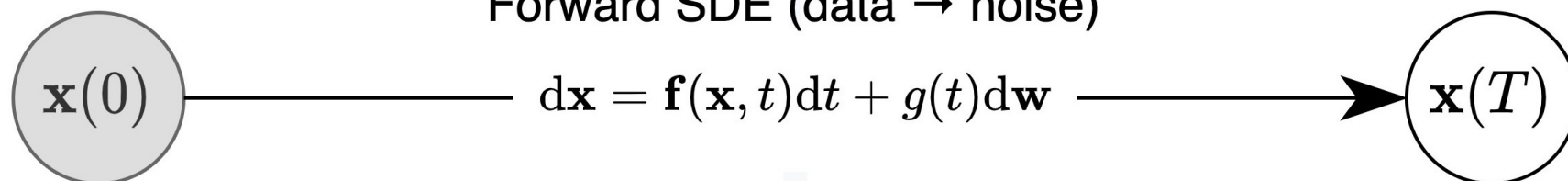
$$q_0(x) \approx \frac{1}{N} \sum_i \delta(x - x^{*i})$$

目标: 采样 $x \sim q_0(x)$

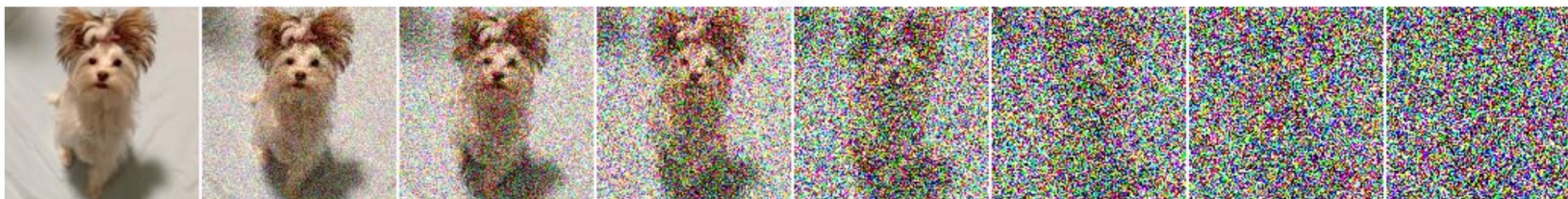


基于分数的生成模型

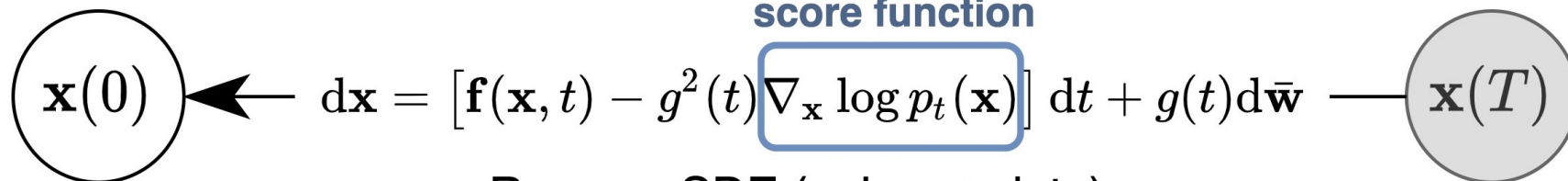
Forward SDE (data $\rightarrow$ noise)



x^{*i}



score function



Reverse SDE (noise $\rightarrow$ data)



本堂课大纲

- 扩散过程简介
 - 前向过程
 - 后向过程
 - 去噪扩散概率模型 (denoising diffusion probabilistic model, DDPM)
 - 概率流常微分方程 (probability flow ODE)
- 分数匹配 (score matching)
 - U-Net: 分数表示 (score representation)
- 理论分析
- 数值实验
 - 混合高斯模型
 - 手写数字识别 (MNIST)



扩散过程

➤ 前向过程 ($0 \rightarrow T$, 图片 $\rightarrow$ 白噪音)

随机常微分方程 (stochastic differential equation, SDE):

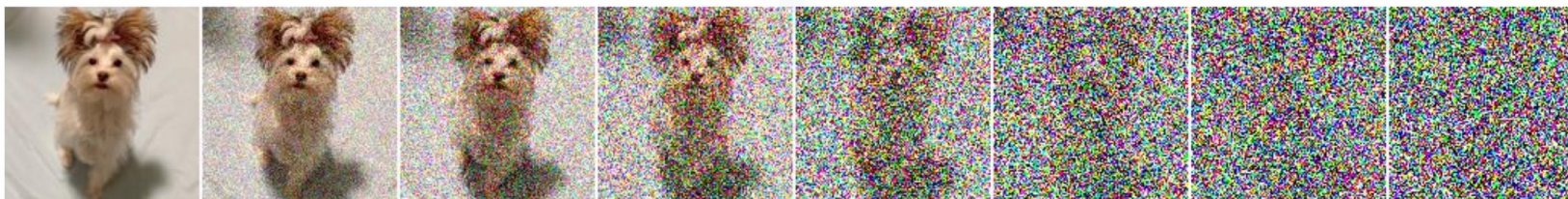
$$dX_t = f(t)X_t dt + g(t)dW_t$$

分布: $X_t = \lambda_t X_0 + \sigma_t W$, $W \sim \mathcal{N}(0, I)$

$$q_{0t}(x_t | x_0) = \mathcal{N}(x_t | \lambda_t x_0, \sigma_t^2 I)$$

其中

$$f(t) = \frac{d \log \lambda_t}{dt} \quad g(t)^2 = \frac{d\sigma_t^2}{dt} - 2 \frac{d \log \lambda_t}{dt} \sigma_t^2$$



证明思路: 考虑 $L(t, x) = \frac{x}{\lambda_t} \quad dL(t, x_t)$



扩散过程

➤ 前向过程 ($0 \rightarrow T$, 图片 $\rightarrow$ 白噪音)

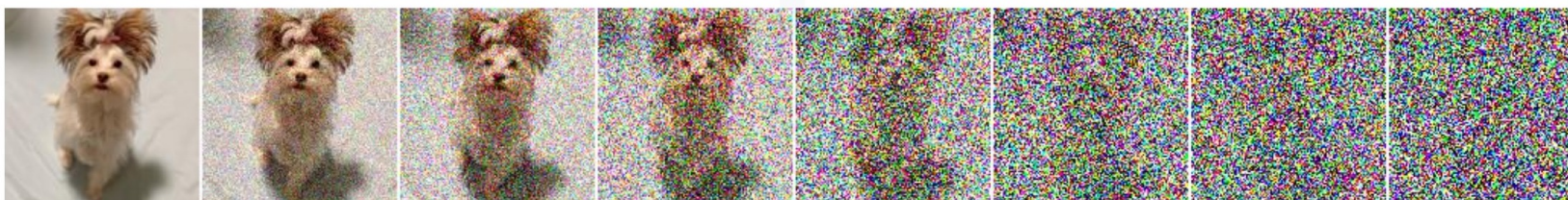
随机常微分方程:

$$dX_t = f(t)X_t dt + g(t)dW_t$$

密度函数: $X_t \sim q_t(x)$

满足Fokker Planck 方程:

$$\partial_t q_t(x) = -\nabla \cdot (f(t)xq_t(x)) + \frac{1}{2}\Delta(g(t)^2 q_t(x))$$





扩散过程

➤ 反向过程 ($T \rightarrow 0$, 白噪音 $\rightarrow$ 图片)

$$-\partial_t q_{T-t}(x) = -\nabla \cdot (f(T-t)x q_{T-t}) + \frac{1}{2} \Delta (g(T-t)^2 q_{T-t})$$

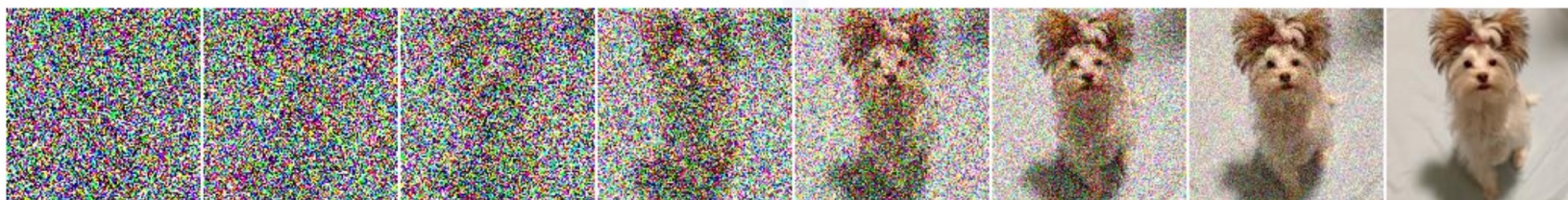
定义: $\rho_t(x) = q_{T-t}(x)$ ($0 \rightarrow T$), 满足

$$\partial_t \rho_t(x) = \nabla \cdot (f(T-t)x \rho_t) - \frac{1}{2} \Delta (g(T-t)^2 \rho_t)$$

➤ 粒子动力学 ($0 \rightarrow T$)

生成白噪音: $Y_0 \sim \rho_0(x) = q_T(x)$ ($0 \rightarrow T$)

求解: $dY_t = \hat{f}(t, Y_t)dt + \hat{g}(t)dW_t$





前向过程

➤ Ornstein-Uhlenbeck (OU) 过程 ($0 \rightarrow T$)

$$dX_t = -\frac{1}{2}\beta(t)X_tdt + \sqrt{\beta(t)}dW_t$$

$$dX_t = -X_tdt + \sqrt{2}dW_t$$

➤ 方差爆炸 (Variance exploding) SDE动力学 ($0 \rightarrow T$)

$$dX_t = \sigma(t)dW_t$$

$$dX_t = dW_t$$



OU 过程

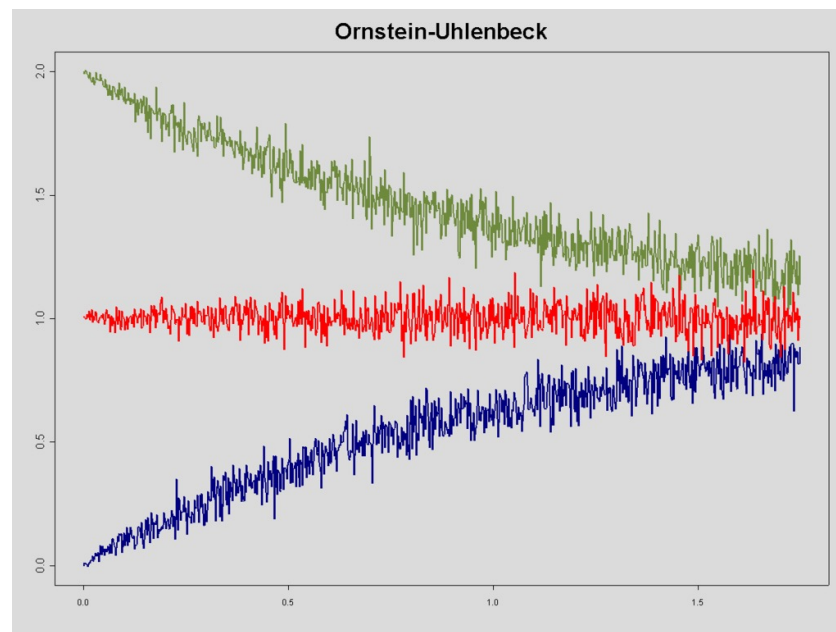
➤ 前向方程 ($0 \rightarrow T$)

$$\text{SDE: } dX_t = -X_t dt + \sqrt{2} dW_t$$

$$\lambda_t = e^{-t} \quad \sigma_t = \sqrt{1 - e^{-2t}}$$

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} W$$

$$X_t \rightarrow \mathcal{N}(0, I)$$



密度函数: $X_t \sim q_t(x)$

满足 Fokker Planck 方程:

$$\partial_t q_t(x) = \nabla \cdot ((x + \nabla \log q_t(x)) q_t(x))$$



OU过程

➤ 后向方程 ($T \rightarrow 0$)

$$-\partial_t q_{T-t}(x) = \nabla \cdot ((x + \nabla \log q_{T-t}(x)) q_{T-t}(x))$$

定义: $\rho_t(x) = q_{T-t}(x) \quad (0 \rightarrow T)$

$$\partial_t \rho_t(x) = -\nabla \cdot ((x + \nabla \log q_{T-t}(x)) \rho_t)$$

➤ 粒子动力学 ($0 \rightarrow T$)

生成白噪音: $Y_0 \sim \mathcal{N}(0, I) \approx q_T(x) \quad (0 \rightarrow T)$

- 去噪扩散概率模型 (DDPM)

$$dY_t = (Y_t + 2\nabla \log q_{T-t}(Y_t))dt + \sqrt{2}dW_t$$

- 概率流常微分方程 (probability flow ODE)

$$\frac{dY_t}{dt} = Y_t + \nabla \log q_{T-t}(Y_t)$$



基于分数的生成模型

➤ 基于分数的生成模型

步骤 1: 添加噪声

$$X_t = e^{-t}X_0 + \sqrt{1 - e^{-2t}}W \quad W \sim \mathcal{N}(0, I)$$

步骤 2: 分数匹配

$$s_\theta(t, x) \approx \nabla \log q_{T-t}(x)$$

步骤 3: 去噪生成

初始化: $\hat{Y}_0 \sim q_T(x) \approx \mathcal{N}(0, I)$

- 去噪扩散概率模型 (DDPM)

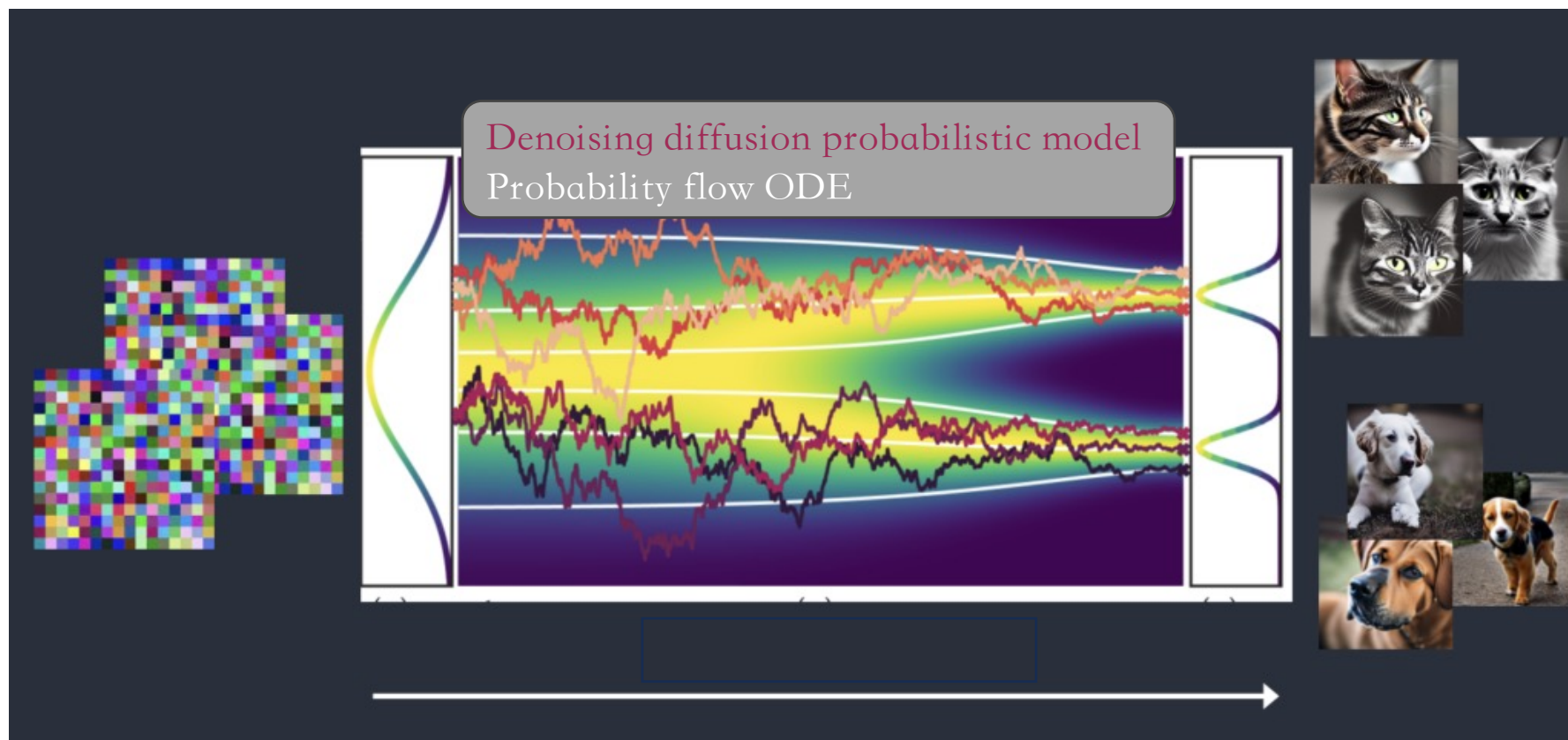
$$d\hat{Y}_t = (\hat{Y}_t + 2s_\theta(t, \hat{Y}_t))dt + \sqrt{2}dW_t$$

- 概率流常微分方程 (probability flow ODE)

$$\frac{d\hat{Y}_t}{dt} = \hat{Y}_t + s_\theta(t, \hat{Y}_t)$$



基于分数的生成模型





分数匹配 (Score Matching)

➤ 目标函数

$$\min_{\theta} \int_0^T w_{T-t} \mathbb{E}_{q_{T-t}} \| s_{\theta}(t, x) - \nabla \log q_{T-t}(x) \|^2 dt$$

数据:

$$X_t = e^{-t} X_0 + \sqrt{1 - e^{-2t}} W \sim q_t(x)$$

概率密度函数:

$$q_{0t}(x|x_0) = \mathcal{N}(x; e^{-t}x_0, (1 - e^{-2t})I)$$

$$q_t(x) = \int \mathcal{N}(x; e^{-t}x_0, (1 - e^{-2t})I) q_0(x_0) dx_0$$

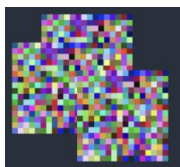
➤ 数值实现 (Monte Carlo 积分)

$$\min_{\theta} \int_0^T w_{T-t} \mathbb{E}_{q_0(x_0)} \mathbb{E}_{q_{0T-t}(x|x_0)} \| s_{\theta}(t, x) - \nabla \log q_{0T-t}(x|x_0) \|^2$$

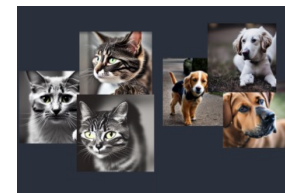


分数匹配 (Score Matching)

➤ 时间离散



$$0 = t_0 < t_1 < \cdots t_N = T - \tau$$



τ : 避免奇异 (singularity)

练习: $q_0 = \delta(x)$, 计算 $\nabla \log q_{T-t}(x)$

$$q_t(x) = \int \mathcal{N}(x; e^{-t}x_0, (1 - e^{-2t})I) q_0(x_0) dx_0$$

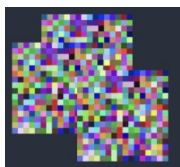
$$= \mathcal{N}(x; 0, (1 - e^{-2t})I)$$

$$s(t, x) = \nabla \log q_{T-t}(x) = -\frac{x}{1 - e^{-2(T-t)}} \approx \mathcal{O}\left(\frac{1}{T-t}\right), \text{ 当 } t \rightarrow T$$

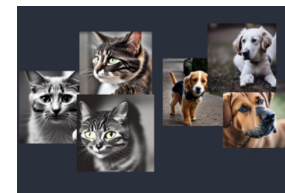


分数匹配 (Score Matching)

➤ 时间离散



$$0 = t_0 < t_1 < \cdots t_N = T - \tau$$



$$\min_{\theta} \int_0^{T-\tau} w_{T-t} \mathbb{E}_{q_0(x_0)} \mathbb{E}_{q_{0T-t}(x|x_0)} \| s_{\theta}(t, x) - \nabla \log q_{0T-t}(x|x_0) \|^2$$

- 固定 t_i 进行训练
- 随机采样 t_i 进行训练



分数匹配 (Score Matching)

➤ 估计 $s(x, t) \approx \nabla \log q_{T-t}$

$$\min_{\theta} \int_{\tau}^T w_t \mathbb{E}_{q_0(x_0)} \mathbb{E}_{q_{0t}(x|x_0)} \| s_{\theta}(x, T-t) - \nabla \log q_{0t}(x|x_0) \|^2$$

- 权重选取

$$q_t(x|x_0) = \mathcal{N}(x; e^{-t}x_0, (1 - e^{-2t})I)$$

$$\nabla \log q_t(x|x_0) = -\frac{x - e^{-t}x_0}{1 - e^{-2t}}$$

$$w_t \propto \frac{1}{\mathbb{E}_{q_t(x|x_0)} \|\nabla \log q_t(x|x_0)\|^2} \propto 1 - e^{-2t}$$

- 时间嵌入 (time embedding)

输入: $x, [\sin(2\pi\omega t); \cos(2\pi\omega t)]$

- 噪音预测模型

$$s_{\theta}(x, T-t) := -\epsilon_{\theta}(x, T-t) / \sqrt{1 - e^{-2t}}$$



分数匹配 (Score Matching)

➤ 噪音预测模型

$$\min_{\theta} \int_{\tau}^T w_t \mathbb{E}_{q_0(x_0)} \mathbb{E}_{q_{0t}(x|x_0)} \| s_{\theta}(x, T-t) - \nabla \log q_{0t}(x|x_0) \|^2$$

换元:

$$s_{\theta}(x, T-t) := -\epsilon_{\theta}(x, T-t) / \sqrt{1 - e^{-2t}}$$

目标函数:

$$\min_{\theta} \int_{\tau}^T \frac{w_t}{\sigma_t^2} \mathbb{E}_{q_0(x_0)} \mathbb{E}_{x=\lambda_t x_0 + \sigma_t W} \| \epsilon_{\theta}(x, T-t) - W \|^2$$

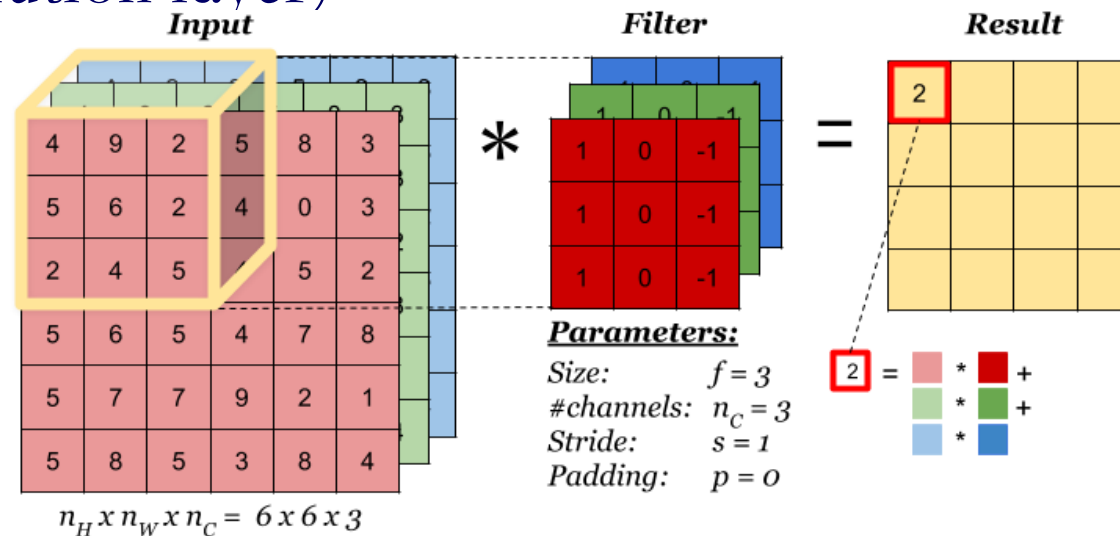
其中

$$\lambda_t = e^{-t} \quad w_t \propto \sigma_t^2 = 1 - e^{-2t}$$

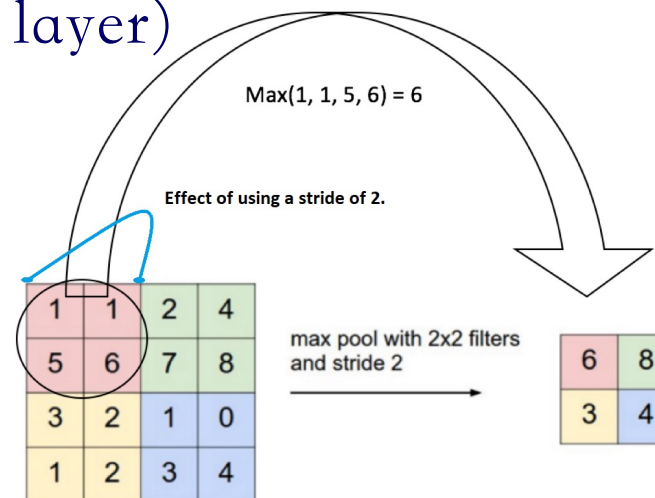


U-Net: 分数表示 $\epsilon_{\theta}(x, t)$

➤ 卷积层 (Convolution layer)



➤ 池化层 (Pooling layer)

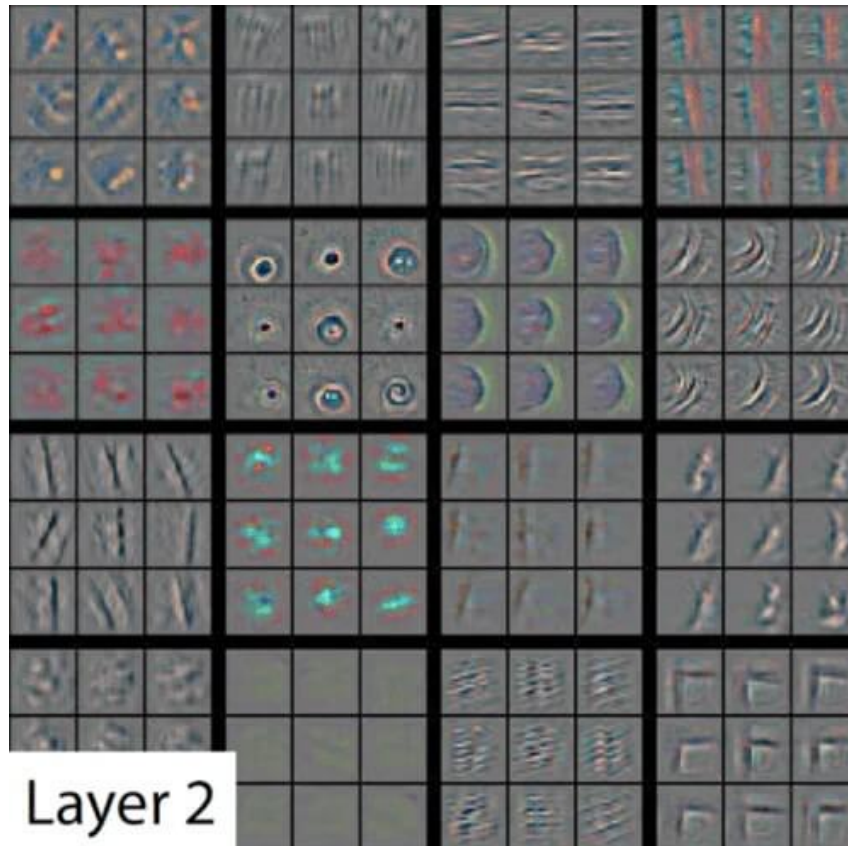


We do the pooling with 2x2 filters, so we will divide an input image on 2x2 regions, and we will use a stride of 2. Because we are using a stride of 2, these regions don't overlap.



U-Net: 分数表示 $\epsilon_{\theta}(x, t)$

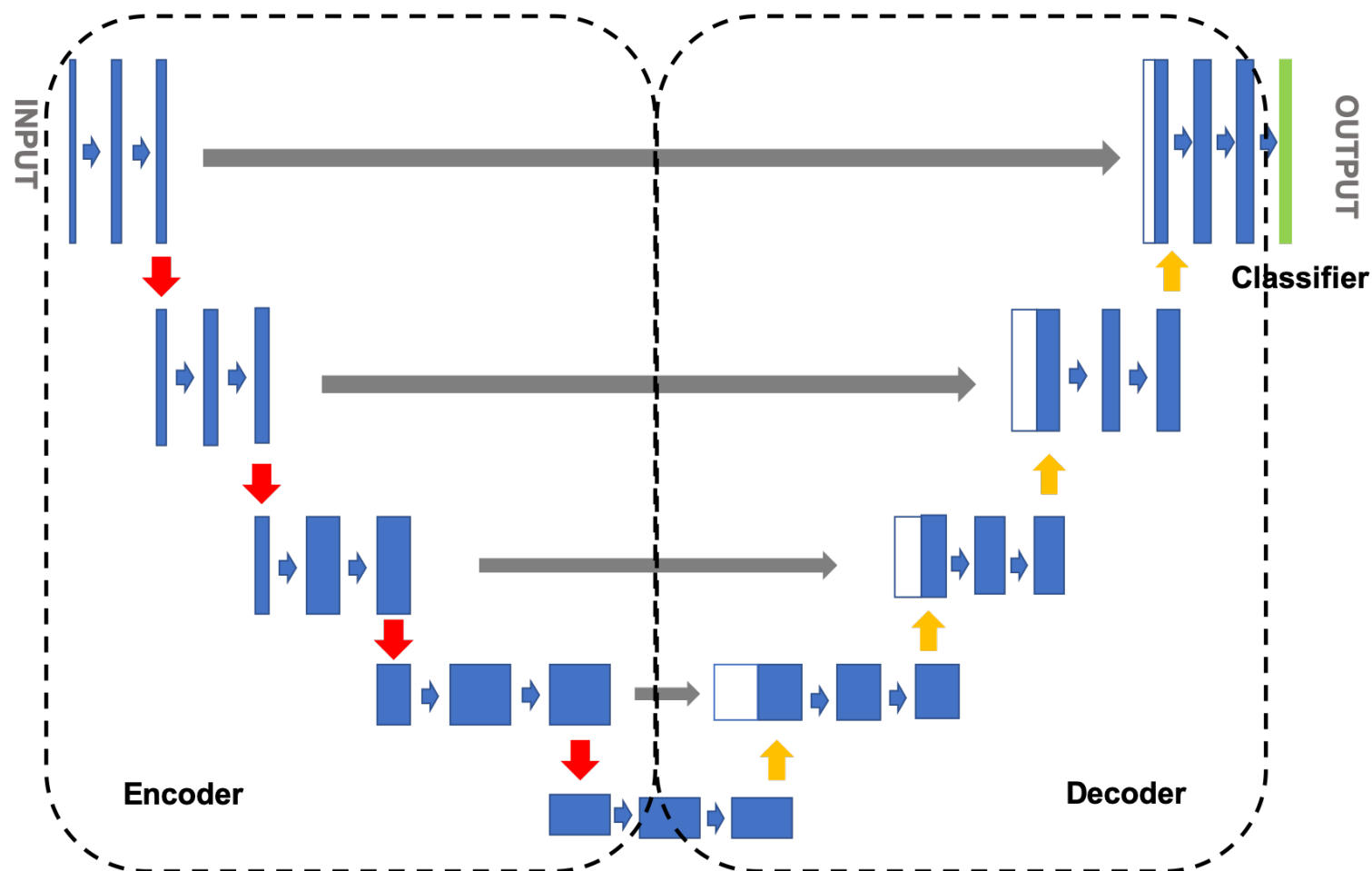
➤ 卷积层 (Convolution layer)





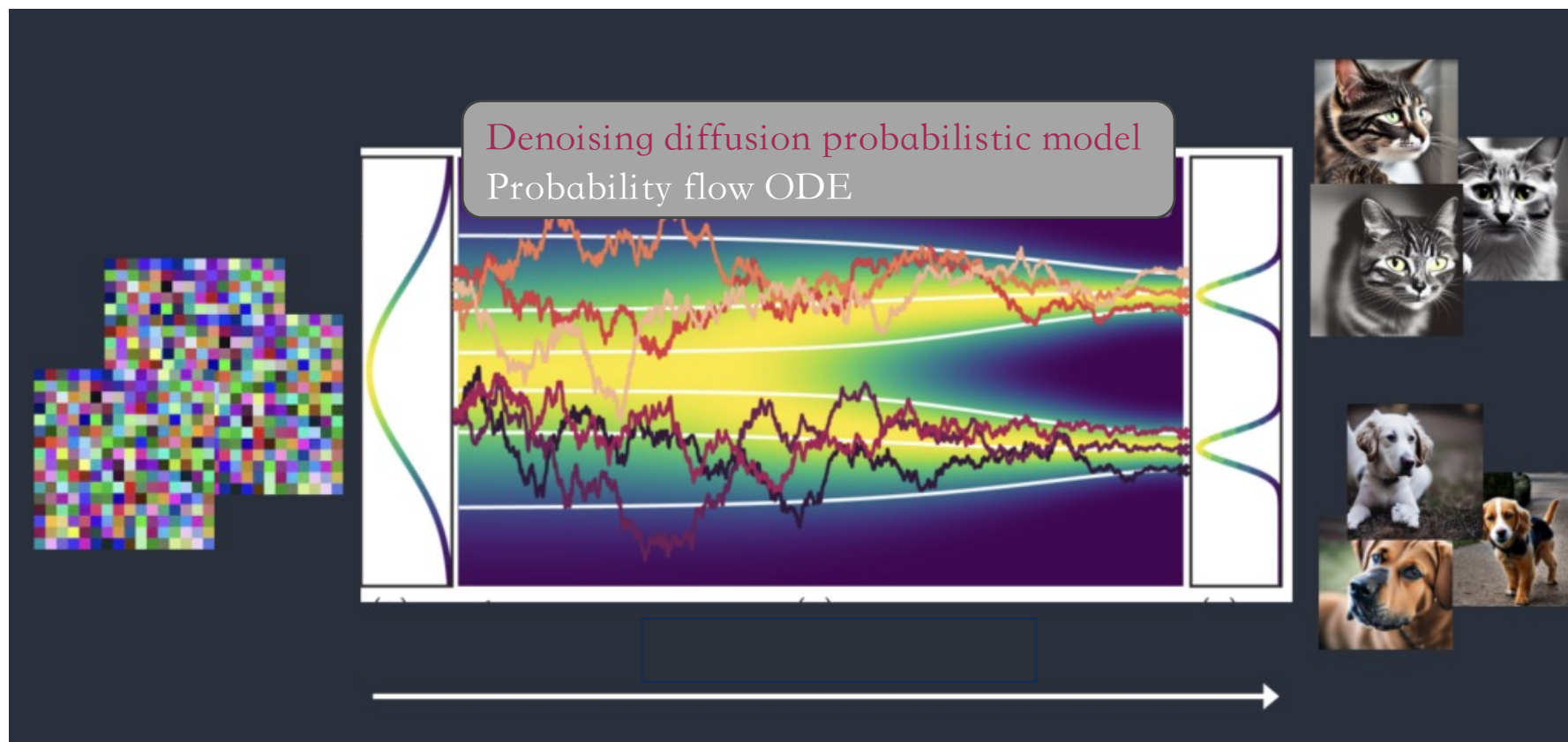
U-Net: 分数表示 $\epsilon_{\theta}(x, t)$

➤ U-Net





基于分数的生成模型





去噪扩散概率模型 (DDPM)

➤ 去噪扩散概率模型

$$d\hat{Y}_t = (\hat{Y}_t + 2s_\theta(t, \hat{Y}_t)) dt + \sqrt{2}dW_t$$

- Euler-Maruyama

$$d\hat{Y}_t = \left(\hat{Y}_{t_k} + 2s_\theta(t_k, \hat{Y}_{t_k}) \right) dt + \sqrt{2}dW_t \quad t \in [t_k, t_{k+1}]$$

- 指数积分器 (exponential integrator)

$$d\hat{Y}_t = \left(\hat{Y}_t + 2s_\theta(t_k, \hat{Y}_{t_k}) \right) dt + \sqrt{2}dW_t \quad t \in [t_k, t_{k+1}]$$

$\hat{Y}_t \sim \hat{\rho}_t$, $\hat{\rho}_{t_N}$ 和 ρ_T 之间的误差是多少?

$$\partial_t \rho_t(x) = -\nabla \cdot ((x + \nabla \log q_{T-t}(x)) \rho_t)$$

$$dY_t = (Y_t + 2\nabla \log q_{T-t}(Y_t))dt + \sqrt{2}dw_t$$



去噪扩散概率模型 (DDPM)

Girsanov 定理

假设 $y_0 \sim q_T$, 考虑在 $C([0, T]; \mathbb{R}^d)$ 上的路径测度 Q_T 和 P_T , 它们对应以下两种扩散过程:

$$Q_T: dY_t = (Y_t + 2\nabla \log q_{T-t}(Y_t))dt + \sqrt{2}dW_t$$

$$P_T: d\hat{Y}_t = (\hat{Y}_t + 2s_\theta(t_k, \hat{Y}_{t_k}))dt + \sqrt{2}dW_t \quad t \in [t_k, t_{k+1}]$$

$$\begin{aligned} \text{KL}[\rho_{t_N} \parallel \hat{\rho}_{t_N}] &\leq \text{KL}[Q_{t_N} \parallel P_{t_N}] \\ &\leq \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{Q_{t_N}} \|s_\theta(t_k, y_{t_k}) - \nabla \log q_{T-t}(y_t)\|^2 dt \end{aligned}$$



去噪扩散概率模型 (DDPM)

误差分析 (Chen et al. 2023; Benton et al. 2024)

A1: 分数匹配误差

$$\sum_{k=0}^{N-1} (t_{k+1} - t_k) \mathbb{E}_{q_{t_k}} \|s_{\theta}(t_k, x) - \nabla \log q_{T-t_k}(x)\| \leq \delta^2$$

A2: 数据假设 $\text{Cov}_{q_0}(x) = I_d$

我们有, 当 $0 = t_0 < t_1 \cdots < t_N = T - \tau$, 且 $t_{k+1} - t_k \leq \kappa \min\{1, T - t_{k+1}\}$, 那么

$$KL[\rho_T \parallel \hat{\rho}_{t_N}] \lesssim \delta^2 + d\kappa^2 N + \kappa dT + de^{-2T}$$

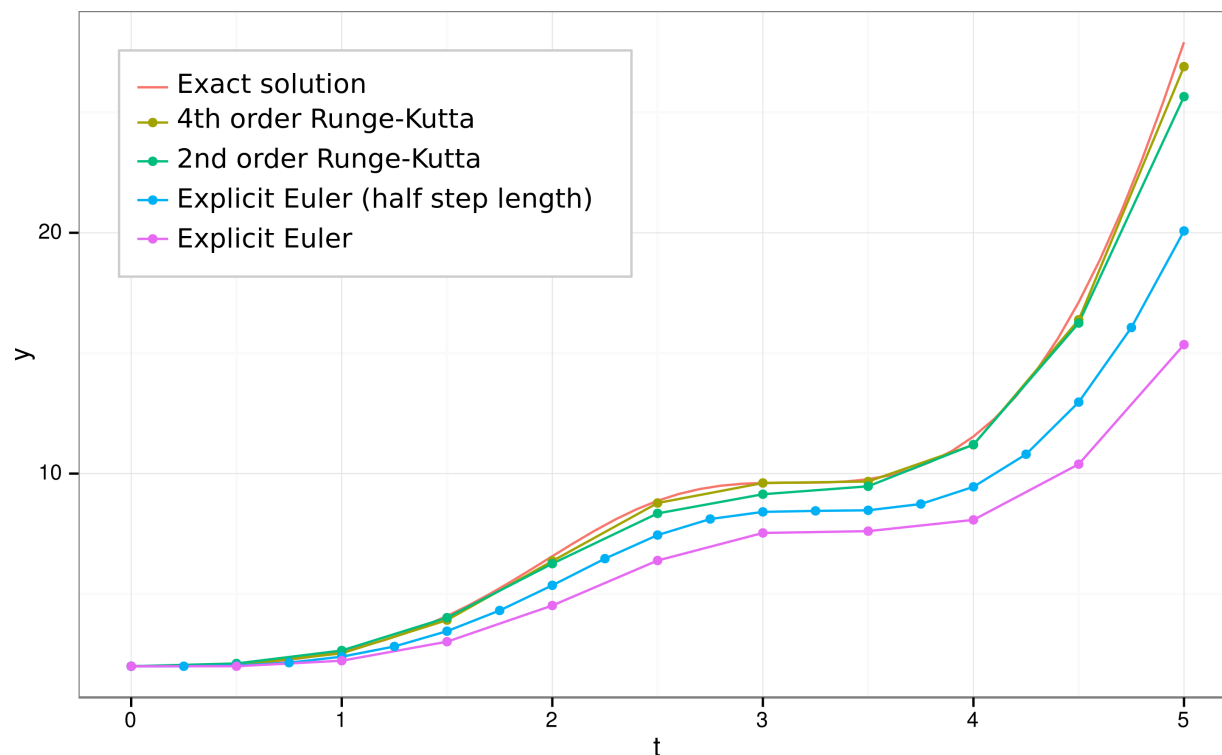


概率流常微分方程

➤ 概率流常微分方程 (Probability flow ODE)

$$\frac{d\hat{Y}_t}{dt} = \hat{Y}_t + s_\theta(t, \hat{Y}_t)$$

高阶格式!
 $O(10)$ 个时间步, 更高效!





概率流常微分方程

➤ 概率流常微分方程 (Probability flow ODE)

$$\frac{d\hat{Y}_t}{dt} = \hat{Y}_t + s_\theta(t, \hat{Y}_t)$$

- (高阶) Runge-Kutta 格式

$$d\hat{Y}_t = \left(\hat{Y}_{t_k} + s_\theta(t_k, \hat{Y}_{t_k}) \right) dt \quad t \in [t_k, t_{k+1}]$$

- (高阶) 指数积分器 (exponential integrator)

$$d\hat{Y}_t = \left(\hat{Y}_t + s_\theta(t_k, \hat{Y}_{t_k}) \right) dt \quad t \in [t_k, t_{k+1}]$$



概率流常微分方程

引理 (Huang et al. 2024)

假设 $\rho_t, \hat{\rho}_t \in C^1([0, T]; R^d) \cap L^1([0, T]; R^d)$ 是下面两个对流方程的解

$$\partial_t \rho_t(x) = \nabla \cdot (U_t(x) \rho_t(x))$$

$$\partial_t \hat{\rho}_t(x) = \nabla \cdot (\hat{U}_t(x) \hat{\rho}_t(x))$$

我们有

$$\begin{aligned} & |\text{TV}(\rho_T, \hat{\rho}_T) - \text{TV}(\rho_0, \hat{\rho}_0)| \\ & \leq \frac{1}{2} \int_0^T \int |\nabla \cdot ((U_t(x) - \hat{U}_t(x)) \rho_t(x))| dx dt \end{aligned}$$



概率流常微分方程

假设

A1. 数据有紧支撑 $\{x \in R^d: \|x\|_\infty \leq D\}$

A2. 分数匹配误差

$$\int_0^{T-\tau} \mathbb{E}_{q_{T-t}} \|s_\theta(t, x) - \nabla \log q_{T-t}(x)\|^2 dt \leq \delta^2$$

A3. (正则性) $s_\theta(t, x)$ 是 L -Lipschitz连续的。即为了使用 p -th 阶 Runge-Kutta 方法, 我们假设 $s_\theta(t, x)$ 前 $p+1$ 阶导数模长小于 L 。



概率流常微分方程

误差分析

如果 $\hat{\rho}_{t_N}$ 是从 $\mathcal{N}(0, I_d)$ 初始化的概率流常微分方程 (ODE) 的输出。使用 p 阶 Runge-Kutta 方法, 均匀步长 $h = \frac{T-\tau}{N}$, 得到:

$$\text{TV}(\rho_{t_N}, \hat{\rho}_{t_N}) \leq e^{-T} dD + dT^{\frac{3}{4}} (L + \tau^{-2} D^3)^{\frac{1}{2}} \delta^{\frac{1}{2}} + d(dh)^p (LD)^{p+1} \log \frac{T}{\tau}$$

初始化误差

得分匹配误差

离散化误差

- 我们只估计到时间 $T - \tau$ 处的误差, 而不是 q_0 。 q_0 和 $q_\tau = \rho_{t_N}$ 之间的 Wasserstein-2 距离有上界 $\mathcal{O}(\tau + d(\tau D)^2)$
- 理论误差界为 $\mathcal{O}(d\sqrt{\delta} + d(dh)^p)$



数值实验

➤ 数据分布 q_0

R^d 中 5-模态的混合高斯

➤ 得分匹配误差 ($\delta(t, x) = s(t, x) - \nabla \log q_{T-t}(x)$)

1. 常数误差: $\delta(t, x) = \delta \frac{1}{d}$

2. 线性误差: $\delta(t, x) = \delta \frac{x-m}{d}$

3. 正弦误差: $\delta(t, x) = \delta \sin x \frac{x-m}{d}$

其中, m 为高斯混合分布的均值, $\sin x$ 作用于每个分量

➤ 设置

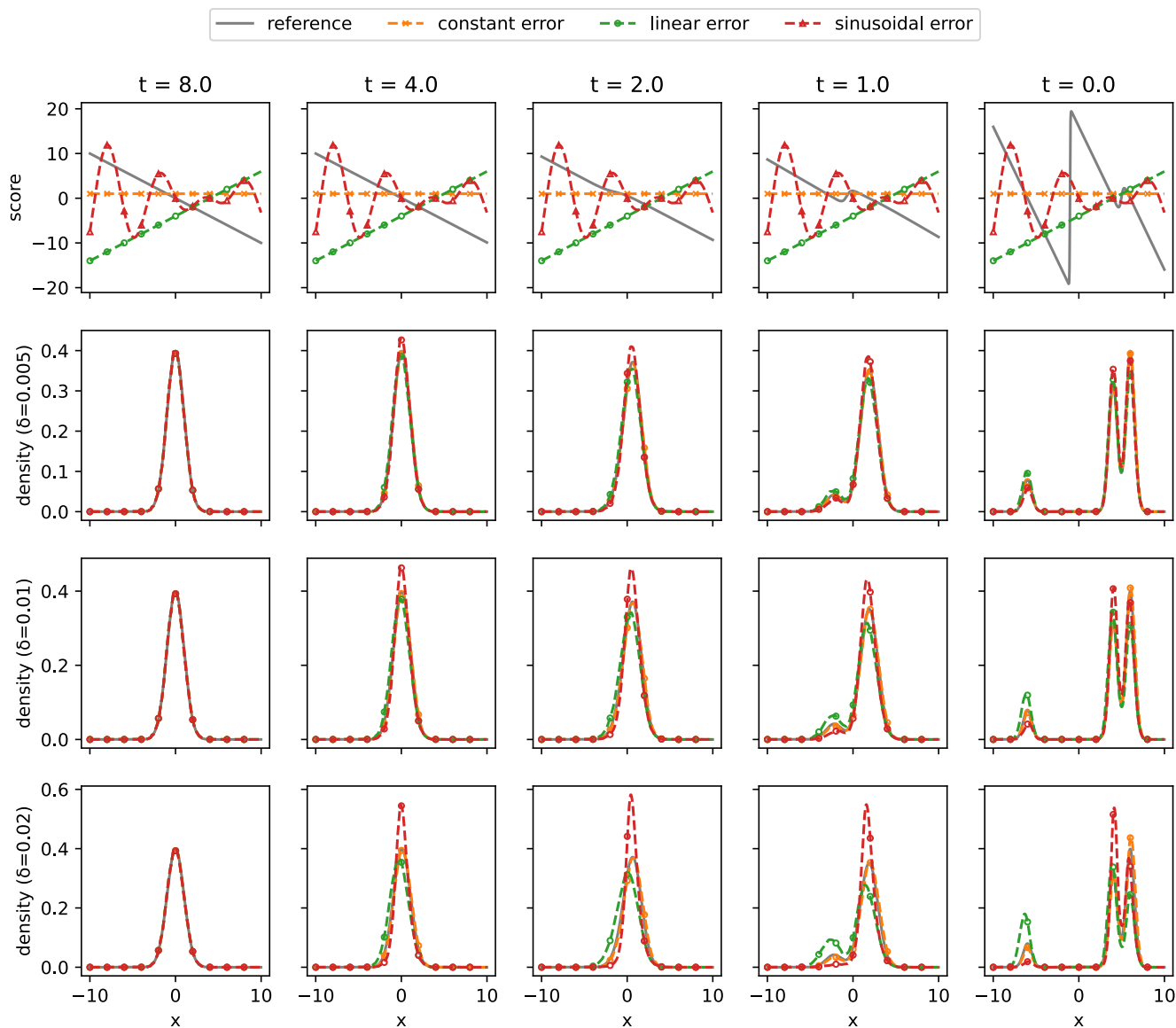
二阶 Heun 方法

时间 $T = 8$, 并从 $\mathcal{N}(0, I_d)$ 初始化 4×10^4 粒子



数值实验

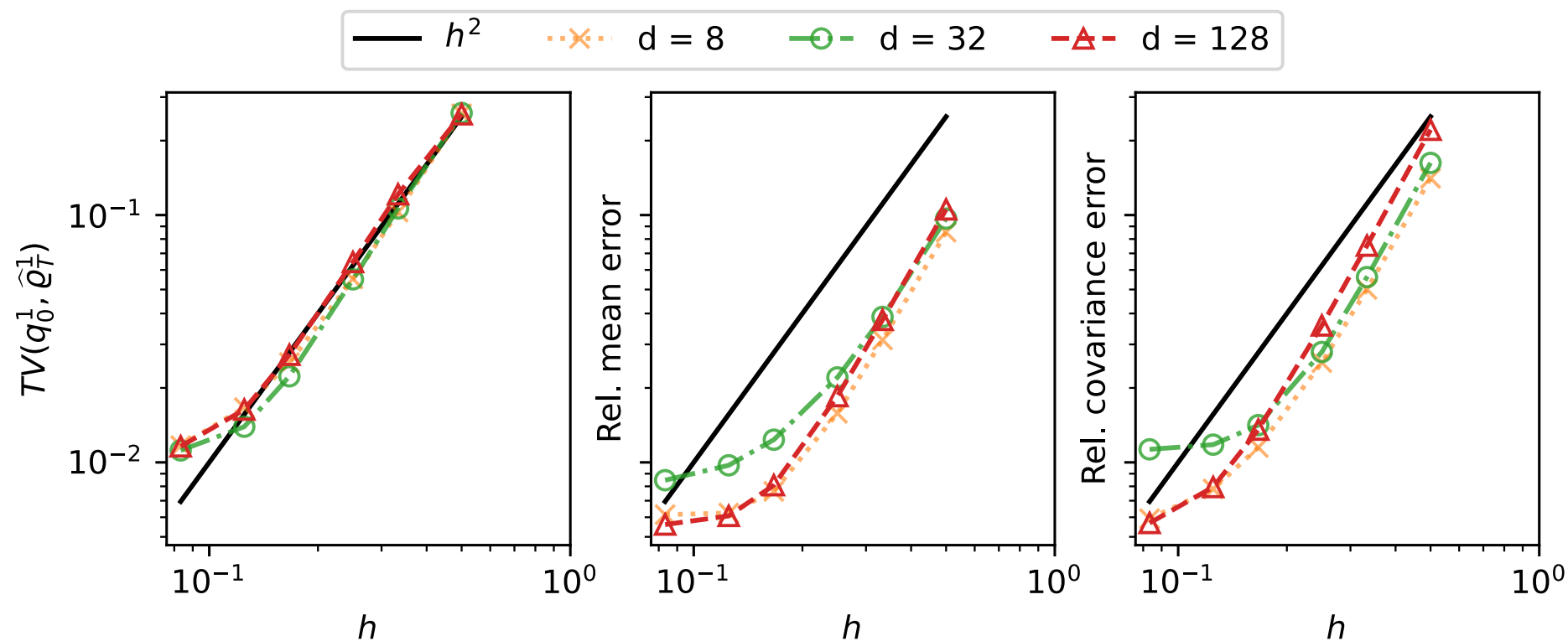
➤ 概率密度函数演化





数值实验

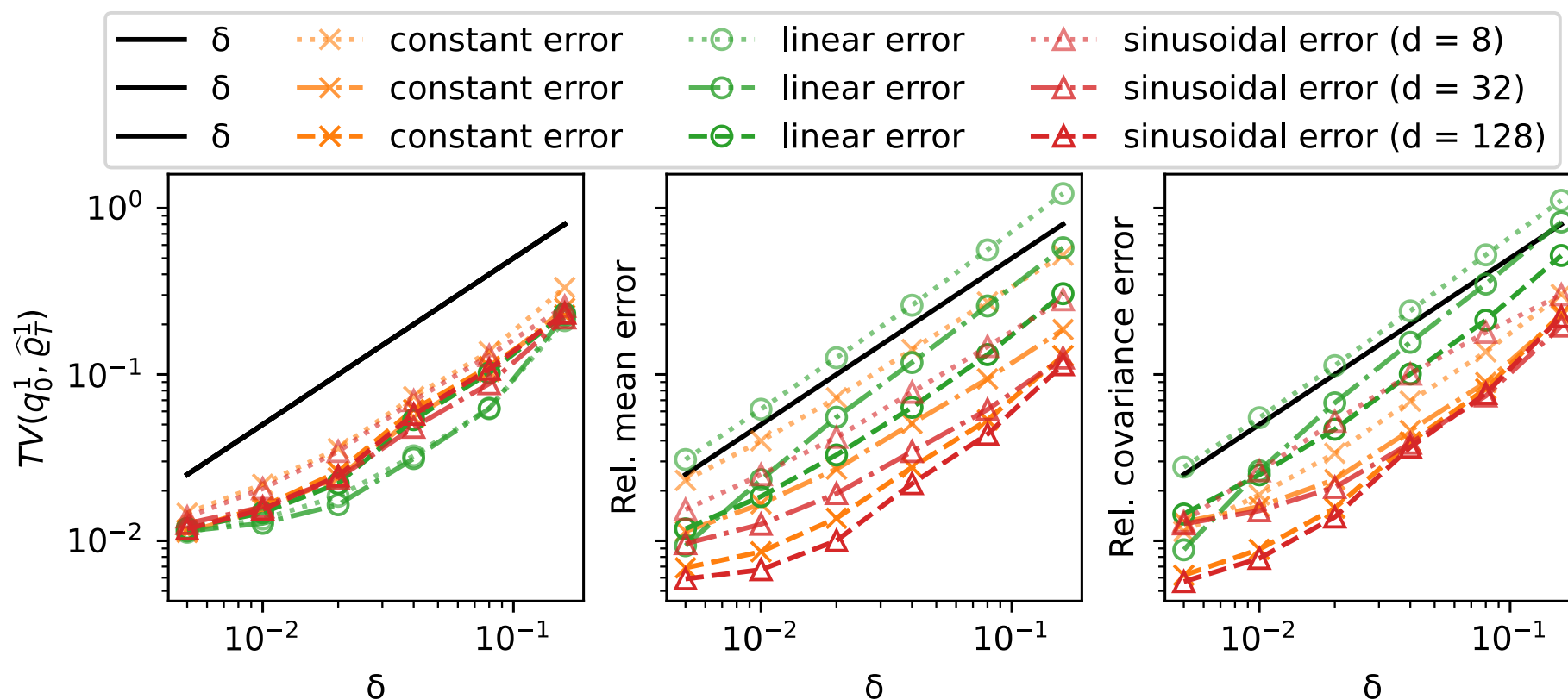
➤ 离散误差 ($\delta=0$)





数值实验

➤ 全变差($h^2 \approx \delta$)

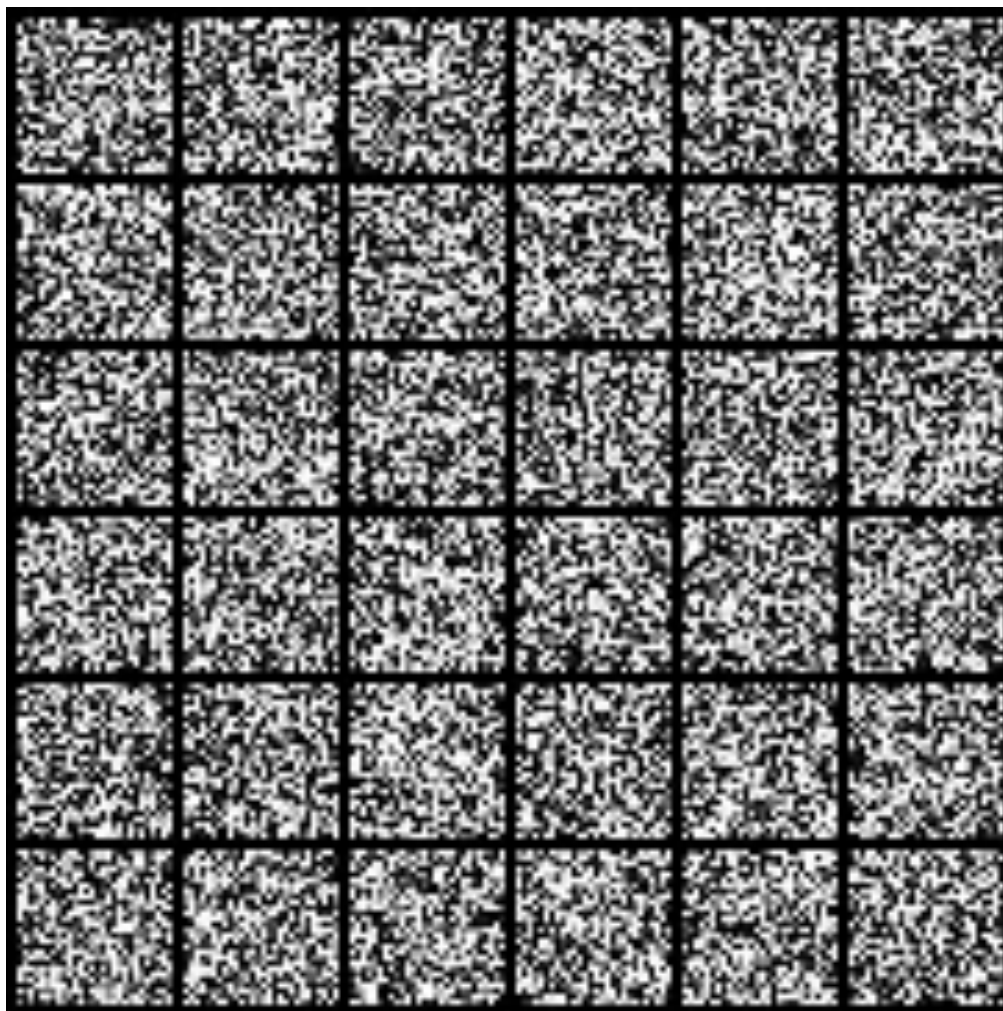


- 理论误差界 $\mathcal{O}(d\sqrt{\delta} + d(dh)^p)$.
- 数值实验误差界 $\mathcal{O}(\delta + h^p)$.



数值实验

➤ MNIST (<https://github.com/bot66/MNISTDiffusion>)





扩展阅读

➤ 方法

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." *International Conference on Learning Representations*, 2021.

Lu, Cheng, et al. "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps." *Advances in Neural Information Processing Systems* 2022.

➤ 理论

Chen, Sitan, et al. "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions." *arXiv preprint arXiv:2209.11215* (2022).

Benton, Joe, et al. "Linear convergence bounds for diffusion models via stochastic localization." *arXiv preprint arXiv:2308.03686* (2023).

Huang, Daniel Zhengyu, Jiaoyang Huang, and Zhengjiang Lin. "Convergence Analysis of Probability Flow ODE for Score-based Generative Models." *arXiv preprint arXiv:2404.09730* (2024).



随机过程

Fokker Planck 方程

$(\theta_t)_{t \geq 0} \in R^d$ 满足 $d\theta_t = b(\theta_t, t)dt + \sigma(\theta_t, t)dB_t$, 若 $\theta_t \sim \rho_t$, 那么

$$\dot{\rho}_t = -\nabla_{\theta} \cdot (b_t \rho_t) + \sum_i \sum_j \frac{\partial^2}{\partial_i \partial_j} [D_{ij} \rho_t]$$

其中 $D = \frac{1}{2} \sigma_t \sigma_t^T$

系统（参数）在时间上随机变化 $\Leftrightarrow$ 概率测度的演化