Theory on Score-Mismatched Diffusion Models and Zero-Shot Conditional Samplers

Yuchen Liang[†], Peizhong Ju[‡], Yingbin Liang[†], Ness Shroff[†]

[†]The Ohio State University [‡]University of Kentucky

Presented at ICLR 2025

Overview

Background on Diffusion Models

- Problem: Generate samples from a complex distribution (e.g., images)
- Denoising Diffusion Probabilistic Models (DDPMs) (Ho *et al.*, 2020)

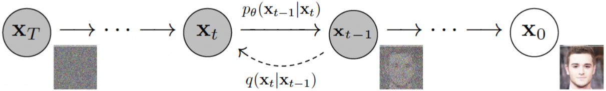


Figure 2: The directed graphical model considered in this work.

1. Forward process Q: $x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} w_t$, for **training**
   - Train for posterior-mean (with score-matching)

$$\widehat{\mu}_t(x_t) \approx \mathbb{E}_{Q_{t-1|t}}[X_{t-1}|x_t] = \frac{1}{\sqrt{\alpha_t}} x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}} \nabla \log q_t(x_t)$$

2. Reverse process $\widehat{P}$: $x_{t-1} = \widehat{\mu}_t(x_t) + \sqrt{\frac{1-\alpha_t}{\alpha_t}} z_t$, for **sampling**

Score-mismatch and Zero-shot Samplers

- Three different dists: Target, Training, Sampling
  - Ideally, matched-score: Target $=$ Training ($= Q_0$)
  - In practice, (usually) Target $\neq$ Training
  - E.g., distributed sensors, privacy concerns

Score-mismatch and Zero-shot Samplers

- Three different dists: Target, Training, Sampling
  - Ideally, matched-score: Target = Training ($= Q_0$)
  - In practice, (usually) Target $\neq$ Training
  - E.g., distributed sensors, privacy concerns
- Interesting example: Conditional Samplers
  - Target: $Q_{0|y}$ (with $y =$ conditioned img) (e.g., image SR)
- One common solution: Classifier Guidance (& CFG)

$$\nabla_x \log q_{t|y}(x_t|y) = \nabla_x \log q_t(x_t) + \nabla_x \log q_{y|t}(y|x_t)$$

  - Pro: No mismatch!
  - Con: $\nabla_x \log q_{y|t}(y|x_t)$ requires extra training...
- Alternative: **Zero-shot** samplers, no extra training, yet mismatched

Main Questions

1. How does **score mismatch** affect theoretical performance?
2. What does it imply for **zero-shot** conditional samplers?

Overview

General Mismatch: Main Result

- Target: $\tilde{Q}_t$; Sampling: P_t ($\widehat{P}_t$ if approx); Training: Q_t
- (Score-)mismatch: $\Delta_t(x_t) := \frac{\sqrt{\alpha_t}}{1-\alpha_t}(\mathbb{E}_{\tilde{Q}_{t-1|t}} - \mathbb{E}_{P_{t-1|t}})[X_{t-1}|x_t]$
- List of Assumptions:
  1. Finite-var targets
  2. Posterior-mean est (for zero-shot: unconditional score est)
  3. Regular target derivatives
  4. Bounded mismatch (in moments)
- Theorem 1: Under [A1]–[A4],

$$\mathrm{KL}(\tilde{Q}_0 \| \widehat{P}_0) \lesssim \mathcal{W}_{\text{oracle}} + \mathcal{W}_{\text{vanish}} + \mathcal{W}_{\text{bias}}$$

  - $\mathcal{W}_{\text{oracle}} = \tilde{O}(T^{-1})$ is error with no mismatch, i.e., $\Delta_t \equiv 0$;
  - $\mathcal{W}_{\text{vanish}} = \tilde{O}(T^{-1})$ captures vanishing effect of mismatch;
  - $\mathcal{W}_{\text{bias}} = O(1)$ captures the **asymptotic bias**
  - This bias is a weighted-sum of Fisher divergence, usually linear in d

Overview

Zero-shot Sampler: Problem Setup

- **Linear** conditional model: $y = Hx_0 + n$, with $n \sim \mathcal{N}(0, \sigma_y^2 I_p)$
- Zero-shot sampler: $\mu_{t,y}(x_t) = \frac{1}{\sqrt{\alpha_t}}(x_t + (1-\alpha_t)g_{t,y}(x_t))$ with

$$g_{t,y} := (I_d - H^\dagger H)\nabla \log q_t(x_t) + f_{t,y}(x_t)$$

  such that $(I_d - H^\dagger H) f_{t,y}(x) \equiv 0$.
- Many applications admit such $f_{t,y}$: CCDF (Chung *et al.*, 2022), DDNM (Wang *et al.*, 2023), etc.

Bias-Optimal Zero-shot Sampler

Theorem 3: Fix $t \geq 1$ and y. Define the forward variance $\Sigma_{t|0,y} := \bar{\alpha}_t \sigma_y^2 H^\dagger (H^\dagger)^\intercal + (1 - \bar{\alpha}_t) I_d$. If

$$f_{t,y}^*(x_t) := \Sigma_{t|0,y}^{-1} \left(\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H x_t \right),$$

then

$$f_{t,y}^* \in \underset{f_{t,y}:(I_d - H^\dagger H) f_{t,y} \equiv 0}{\arg\min} \left\| \Delta_{t,y} \right\|^2, \quad Q_{t|y}\text{–almost surely.}$$

Bias-Optimal Zero-shot Sampler

Theorem 3: Fix $t \geq 1$ and y. Define the forward variance $\Sigma_{t|0,y} := \bar{\alpha}_t \sigma_y^2 H^\dagger (H^\dagger)^\intercal + (1 - \bar{\alpha}_t) I_d$. If

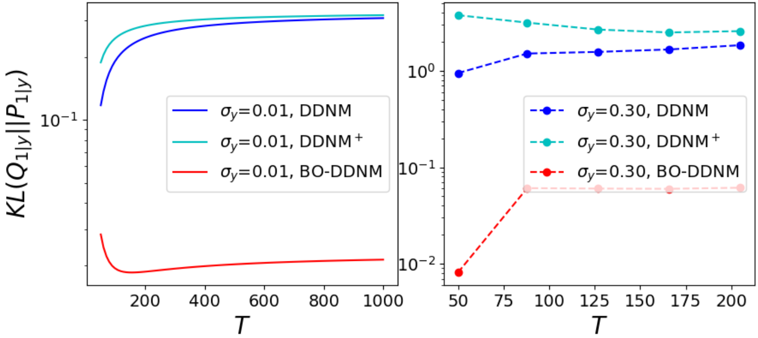
$$f_{t,y}^*(x_t) := \Sigma_{t|0,y}^{-1} \left(\sqrt{\bar{\alpha}_t} H^\dagger y - H^\dagger H x_t \right),$$

then

$$f_{t,y}^* \in \underset{f_{t,y}:(I_d - H^\dagger H) f_{t,y} \equiv 0}{\arg\min} \|\Delta_{t,y}\|^2, \quad Q_{t|y}\text{–almost surely.}$$

- We call $\mu_{t,y}$ with $f_{t,y}^*$ **Bias-Optimal (BO) DDNM** sampler
- If $\sigma_y^2 = 0$, then BO-DDNM = DDNM
- Detailed **parameter dependency** for Gaussian mixture (see paper)

Numerical Performance: BO-DDNM vs. DDNM



- Left: Gaussian; Right: Gaussian mixture

Conclusion

- First non-asymptotic analysis for score-mismatched DDPMs
- Bias-optimal zero-shot sampler for linear conditional models
- System dependencies (d and y) of BO-DDNM