Information-Theoretic Discrete Diffusion

Moongyu Jeon, Sangwoo Shin, Dongjae Jeon, Albert No

Motivation: I-MMSE in Gaussian Diffusion

Gaussian Diffusion Model[1]

- Forward (noising) process: $\mathbf{Z}_\gamma = \sqrt{\gamma}\,\mathbf{X} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$
  - γ denotes the signal-to-noise ratio (SNR), serves as a reparametrized time variable.
- MSE (score-matching) loss[2]: $\mathcal{L}_{\text{MSE}} = \frac{1}{2}\int_0^\infty \mathbb{E}\left[\|\mathbf{X} - \hat{\mathbf{X}}_\theta(\mathbf{Z}_\gamma, \gamma)\|^2\right] d\gamma.$

I-MMSE relation[3]

$$\frac{d}{d\gamma} I(\mathbf{X}; \mathbf{Z}_\gamma) = \frac{1}{2}\text{mmse}(\gamma).$$

- $\text{mmse}(\gamma) = \mathbb{E}[\|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Z}_\gamma]\|^2].$

[1] Song & Ermon, Generative modeling by estimating gradients of the data distribution, NeurIPS, 2019.
[2] Kingma et al., Variational diffusion models, NeurIPS, 2021.
[3] Guo et al., Mutual information and minimum mean-square error in Gaussian channels, IEEE Trans. Inf. Theory, 2005.

Motivation: I-MMSE in Gaussian Diffusion

- **Pointwise I-MMSE relation**[4]:

$$\frac{d}{d\gamma} D_{\mathrm{KL}}\left(p(\mathbf{Z}_\gamma|\mathbf{X}_0)\,\|\,p(\mathbf{Z}_\gamma)\right) = \frac{1}{2}\,\mathrm{mmse}(\mathbf{X}_0,\gamma).$$

  - $\mathrm{mmse}(\mathbf{X}_0,\gamma) = \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Z}_\gamma]\|^2 \,|\, \mathbf{X} = \mathbf{X}_0\right]$.

- **NLL (Negative Log-Likelihood) decomposition via pointwise I-MMSE**[5]:

$$-\log p_0(\mathbf{X}_0) = \frac{1}{2}\int_0^\infty \mathrm{mmse}(\mathbf{X}_0,\gamma)\,d\gamma + \mathrm{const}.$$

[4] Venkat & Weissman, Pointwise relations between information and estimation in Gaussian noise, IEEE Trans. Inf. Theory, 2012.
[5] Kong et al., Information-theoretic diffusion, ICLR, 2023.

Discrete Diffusion Models and DSE Loss

- **Forward process.** Stochastic transitions between finite discrete states.
  - The transition probability from x to y is determined by the transition rate $Q_t(y, x)$.
- **DSE (Denoising Score Entropy) loss** [6]:

$$\mathcal{L}_{\mathrm{DSE}}^{T}(x_0) = \int_0^T \mathbb{E}_{p_{t|0}(x_t|x_0)} \left[\ell_{\mathrm{DSE}}(x_0, x_t, t, s_t^{\theta}) \right] dt, \ \text{where}$$

$$\ell_{\mathrm{DSE}}(x_0, x_t, t, s_t^{\theta}) = \sum_{y \neq x_t} Q_t(x_t, y) \left(s_t^{\theta}(x_t)_y - \frac{p_{t|0}(y|x_0)}{p_{t|0}(x_t|x_0)} \log s_t^{\theta}(x_t)_y + K\left(\frac{p_{t|0}(y|x_0)}{p_{t|0}(x_t|x_0)} \right) \right).$$

  - A variational bound of the NLL.
  - Matches the true distribution ratio (concrete score) $s_t^{\star}(x_t)_y = \frac{p_t(y)}{p_t(x_t)}$.
- **Do there exist analogous results in discrete diffusion models?**
  - **Information-theoretic identity** • **Exact NLL estimation**

[6] Lou et al., Discrete diffusion language modeling by estimating the ratios of the data distribution, ICML, 2024.

[Main Result] I-MDSE: The Discrete Counterpart of I-MMSE

- **I-MDSE relation**:

$$\frac{d}{dt} I(x_0; x_t) = -\mathrm{mdse}(t).$$

  - $\mathrm{mdse}(t) = \mathbb{E}_{p(x_0,x_t)} \left[\ell_{\mathrm{DSE}}(x_0, x_t, t, s_t^\star)\right]$

- **Pointwise I-MDSE relation**:

$$\frac{d}{dt} D_{\mathrm{KL}} \left(p_{t|0}(\cdot|x_0) \,\|\, p_t\right) = -\mathrm{mdse}(x_0, t).$$

  - $\mathrm{mdse}(x_0, t) = \mathbb{E}_{p_{t|0}(x_t|x_0)} \left[\ell_{\mathrm{DSE}}(x_0, x_t, t, s_t^\star)\right].$

[Main Result] NLL Decomposition via Pointwise I-MDSE

- **NLL as an integral of MDSE:**

$$-\log p_0(x_0) = \int_0^T \mathrm{mdse}(x_0, t)\, dt + D_{\mathrm{KL}}\left(p_{T|0}(\cdot|x_0) \| p_T\right) = \int_0^\infty \mathrm{mdse}(x_0, t)\, dt.$$

- **NLL estimation using learned s^θ:**

$$-\log p_0(x_0) \approx \int_0^T \mathbb{E}_{p_{t|0}(x_t|x_0)}\left[\ell_{\mathrm{DSE}}(x_0, x_t, t, s_t^\theta)\right] dt = \mathcal{L}_{\mathrm{DSE}}^T(x_0).$$

  - **Not as an ELBO, but as an equality-based characterization of DSE objective.**
  - **Enables exact likelihood estimation, no higher-order corrections needed.**

Masked Diffusion Model and DCE Loss

- A **particular discrete diffusion** with an absorbing transition rate matrix Q_t.
- **Time-free reparametrization** of the score network[7]:
  - From score $s_t^\theta(\mathbf{x})_{i,\hat{x}^i} \approx \dfrac{p_t(\hat{\mathbf{x}})}{p_t(\mathbf{x})}$ to conditional $c^\theta(\mathbf{x})_{i,\hat{x}^i} \approx p_0(\hat{x}^i|\mathbf{x}^{\mathrm{UM}})$.
- **DCE (Denoising Cross-Entropy) loss**:

$$\mathcal{L}_{\mathrm{DCE}}^{\Lambda}(\mathbf{x}_0) = \int_0^{\Lambda} \frac{1}{\lambda}\mathbb{E}_{p_{\lambda|0}(\mathbf{x}_\lambda|\mathbf{x}_0)}\left[\ell_{\mathrm{DCE}}(\mathbf{x}_0, \mathbf{x}_\lambda, c^\theta)\right] d\lambda,$$

$$\ell_{\mathrm{DCE}}(\mathbf{x}_0, \mathbf{x}_\lambda, c^\theta) = \sum_{i=1}^{L} \mathbb{1}[x_\lambda^i = [\mathbf{M}]] \log \frac{1}{c^\theta(\mathbf{x}_\lambda)_{i,x_0^i}}.$$

  - **Equivalent to the DSE loss** under a reparametrization of the score network.
  - A variational bound of the NLL.
  - Matches the true conditional distribution $c^\star(\mathbf{x}_\lambda)_{i,x_0^i} = p_0(x_0^i|\mathbf{x}_\lambda^{\mathrm{UM}})$.

[7] Ou et al., Your absorbing discrete diffusion secretly models the conditional distributions of clean data, ICLR, 2025.

[Main Result] I-MDCE: Specialization to Masked Diffusion

- **I-MDCE relation:**

$$\frac{d}{d\lambda} I(\mathbf{x}_0; \mathbf{x}_\lambda) = -\frac{1}{\lambda}\,\mathrm{mdce}(\lambda).$$

  - $\mathrm{mdce}(\lambda) = \mathbb{E}_{p(\mathbf{x}_0, \mathbf{x}_\lambda)}\left[\ell_{\mathrm{DCE}}(\mathbf{x}_0, \mathbf{x}_\lambda, c^\star)\right].$

- **Pointwise I-MDCE relation**:

$$\frac{d}{d\lambda} D_{\mathrm{KL}}\left(p_{\lambda|0}(\cdot|\mathbf{x}_0) \| p_\lambda\right) = -\frac{1}{\lambda}\,\mathrm{mdce}(\mathbf{x}_0, \lambda).$$

  - $\mathrm{mdce}(\mathbf{x}_0, \lambda) = \mathbb{E}_{p_{\lambda|0}(\mathbf{x}_\lambda|\mathbf{x}_0)}\left[\ell_{\mathrm{DCE}}(\mathbf{x}_0, \mathbf{x}_\lambda, c^\star)\right].$

[Main Result] NLL Decomposition via Pointwise I-MDCE

- **NLL as an integral of MDCE:**

$$-\log p_0(\mathbf{x}_0) = \int_0^\Lambda \frac{1}{\lambda}\,\mathrm{mdce}(\mathbf{x}_0, \lambda)\, d\lambda + D_{\mathrm{KL}}\left(p_{\Lambda|0}(\cdot|\mathbf{x}_0) \| p_\Lambda\right) = \int_0^1 \frac{1}{\lambda}\,\mathrm{mdce}(\mathbf{x}_0, \lambda)\, d\lambda.$$

- **NLL estimation using learned c^θ:**

$$-\log p_0(x_0) \approx \int_0^1 \frac{1}{\lambda}\mathbb{E}_{p_{\lambda|0}(\mathbf{x}_\lambda|\mathbf{x}_0)}\left[\ell_{\mathrm{DCE}}(\mathbf{x}_0, \mathbf{x}_\lambda, c^\theta)\right] d\lambda = \mathcal{L}^1_{\mathrm{DCE}}(\mathbf{x}_0).$$

  - **Not as an ELBO, but as an equality-based characterization of DCE objective.**
  - **Enables exact likelihood estimation and supports practical extensions.**

Extending I-MDCE: Time-Free Likelihood Estimation

- **Time-free NLL decomposition**

$$-\log p_0(\mathbf{x}_0) = H_L \mathbb{E}_{p(I)}\left[\sum_{i \notin I} \log \frac{1}{p_0(x_0^i|\mathbf{x}_0^I)}\right].$$

  - H_L is the L^{th} harmonic number.
  - I is the set of unmasked indices sampled from $p(I) = \frac{B(L-|I|,|I|+1)}{H_L}$.

- **Time-free NLL estimation using learned c^θ:**

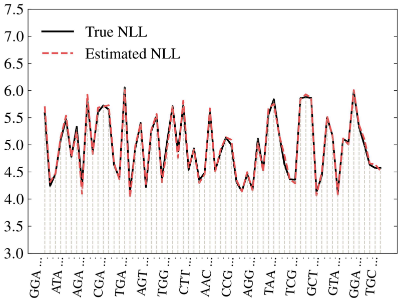
$$-\log p_0(\mathbf{x}_0) \approx H_L \mathbb{E}_{p(I)}\left[\sum_{i \notin I} \log \frac{1}{c^\theta(\tilde{\mathbf{x}}_0^I)_{i,x_0^i}}\right].$$

  - $\tilde{\mathbf{x}}_0^I$ denotes the sequence obtained from $\mathbf{x}_0$ by masking all tokens with indices $\notin I$.

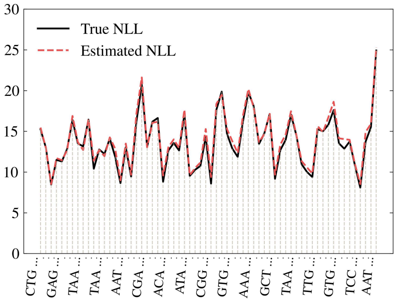
- Exhibits **significantly lower variance** than time-integral estimation.

Experiment: Accuracy of Time-Free Likelihood Estimation

- Trained the masked diffusion model (RADD) on a toy dataset.
- Demonstrated the **exact match** between the **true (black)** and **estimated (red)** NLL.



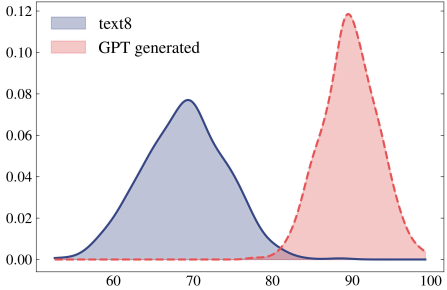
(a) Unconditional



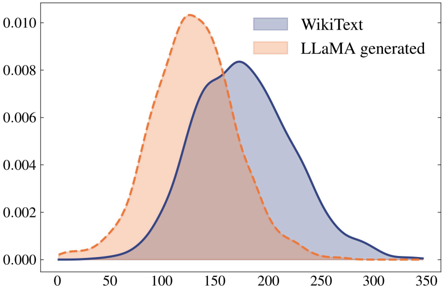
(b) Conditional

Experiment: Real-World Auditing Tasks

- (Fig. c) RADD trained on `text8`: Using conditional likelihood estimation, distinguishes OOD (GPT-generated) from in-distribution samples.
- (Fig. d) In LLaDA, LLaMA-generated responses show higher likelihood than true continuations (`WikiText`) $\to$ model influenced by LLaMA distribution.



(c) (NLL) Detecting OOD inputs



(d) (NLL) Analysis of distributions in LLaDA

Recap: Key Contributions

- Established **I-MDSE** and **I-MDCE**, discrete counterparts of the I-MMSE relation.
- Proved that **DSE and DCE losses** are **exact estimators of the NLL**, not potentially loose variational bounds.

Further results and experiments are presented in the paper.

See you at the poster session!

Exhibit Hall C, D, E

Thu, Dec 4, 11 a.m. – 2 p.m. (PST)

Paper Link ↑