ICML 2024 Debriefing

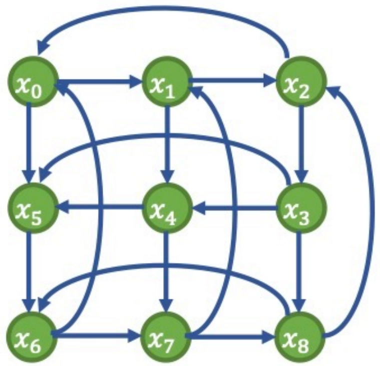
Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution

Aaron Lou, Chenlin Meng, Stefano Ermon

Presenter: Ruizhong Qiu

October 3, 2024

Prior Work: Concrete Score Matching



- We can typically define *neighbors* for discrete data
- *Concrete score* of a sample x with neighbors $\mathcal{N}(x) = \{x_{n_1}, \dots, x_{n_k}\}$:

$$\boldsymbol{c}_{p_{\text{data}}}(\mathbf{x}; \mathcal{N}) \triangleq \left[\frac{p_{\text{data}}(\mathbf{x}_{n_1}) - p_{\text{data}}(\mathbf{x})}{p_{\text{data}}(\mathbf{x})}, \dots, \frac{p_{\text{data}}(\mathbf{x}_{n_k}) - p_{\text{data}}(\mathbf{x})}{p_{\text{data}}(\mathbf{x})} \right]^T$$

- Inference with Metropolis–Hastings: accept proposed x' with probability:

$$A(\mathbf{x}'|\mathbf{x}) = \min\left(1, \frac{p_{\text{data}}(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p_{\text{data}}(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}\right)$$

- Concrete score matching (CSM):

$$\begin{aligned}\mathcal{L}_{\text{CSM}}(\theta) &= \sum_{\mathbf{x}} p_{\text{data}}(\mathbf{x}) \left\| \boldsymbol{c}_\theta(\mathbf{x}; \mathcal{N}) - \boldsymbol{c}_{p_{\text{data}}}(\mathbf{x}; \mathcal{N}) \right\|_2^2 \\ &= \sum_{\mathbf{x}} \sum_{i=1}^{|\mathcal{N}(\mathbf{x})|} p_{\text{data}}(\mathbf{x}) \left(\boldsymbol{c}_\theta(\mathbf{x}; \mathcal{N})_i^2 + 2\boldsymbol{c}_\theta(\mathbf{x}; \mathcal{N})_i \right) - \sum_{\mathbf{x}} \sum_{i=1}^{|\mathcal{N}(\mathbf{x})|} 2p_{\text{data}}(\mathbf{x}_{n_i}) \boldsymbol{c}_\theta(\mathbf{x}; \mathcal{N})_i + \text{const}\end{aligned}$$

[1] Meng et al. Concrete score matching: Generalized score matching for discrete data. *NeurIPS*, 2022.

Continuous-Time Discrete Diffusion

- Discrete diffusion via a continuous-time Markov chain (CTMC):

$$\frac{dp_t}{dt} = Q_t p_t \quad p_0 \approx p_{\text{data}} \qquad p(x_{t+\Delta t} = y | x_t = x) = \delta_{xy} + Q_t(y, x)\Delta t + O(\Delta t^2)$$

- Inference with the reverse process needs the likelihood ratio:

$$\frac{dp_{T-t}}{dt} = \overline{Q}_{T-t} p_{T-t} \quad \overline{Q}_t(y, x) = \frac{p_t(y)}{p_t(x)} Q_t(x, y)$$

$$\overline{Q}_t(x, x) = -\sum_{y \neq x} \overline{Q}_t(y, x)$$

- If we employ CSM here, the loss will be:

$$\mathcal{L}_{\text{CSM}} = \frac{1}{2} \mathbb{E}_{x \sim p_t} \left[\sum_{y \neq x} \left(s_\theta(x_t, t)_y - \frac{p_t(y)}{p_t(x)} \right)^2 \right]$$

  - CSM does not sufficiently penalize negative or zero values

[1] Campbell et al. A continuous time framework for discrete denoising models. *NeurIPS*, 2022.
[2] Sun et al. Score-based continuous-time discrete diffusion models. *ICLR*, 2023.
[3] Meng et al. Concrete score matching: Generalized score matching for discrete data. *NeurIPS*, 2022.

Score Entropy

- To generalize the cross entropy from distributions to any positive vectors
- Bregman divergence w.r.t. convex $K(a) := a(\log a - 1)$; $K'(a) = \log a$

$$D_K(r, s) = K(r) - K(s) - K'(s) \cdot (r - s) = s - r \log s + K(r)$$

- The *score entropy* (SE) is defined by the expected $D_K\left(\frac{p(y)}{p(x)}, s_\theta(x)_y\right)$

Definition 3.1. *The **score entropy** $\mathcal{L}_{\mathrm{SE}}$ for a distribution p, weights $w_{xy} \geq 0$ and a score network $s_\theta(x)_y$ is*

$$\mathbb{E}_{x\sim p}\left[\sum_{y\neq x} w_{xy}\left(s_\theta(x)_y - \frac{p(y)}{p(x)}\log s_\theta(x)_y + K\left(\frac{p(y)}{p(x)}\right)\right)\right]$$

- Satisfies some desiderata…

Properties of Score Entropy

- Consistency: SE can recover ground truth concrete scores in the limit

Proposition 3.2 (Consistency of Score Entropy). *Suppose p is fully supported and $w_{xy} > 0$. As the number of samples and model capacity approaches ∞, the optimal θ^* that minimizes Equation 5 satisfies $s_{\theta^*}(x)_y = \frac{p(y)}{p(x)}$ for all pairs x, y Furthermore, $\mathcal{L}_{\mathrm{SE}}$ will be 0 at θ^*.*

- A log-barrier to keep $s_\theta > 0$:

$$\nabla_{s_\theta(x)_y} \mathcal{L}_{\mathrm{SE}} = \frac{1}{s_\theta(x)_y} \nabla_{s_\theta(x)_y} \mathcal{L}_{\mathrm{CSM}}$$

- Can be made computationally tractable…

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

Denoising Score Entropy

- SE is intractable because $\frac{p_t(y)}{p_t(x)}$ is unknown
- Following [1], this work develops a tractable *denoising SE* (DSE)

Theorem 3.4 (Denoising Score Entropy). *Suppose p is a perturbation of a base density p_0 by a transition kernel $p(\cdot|\cdot)$, ie $p(x) = \sum_{x_0} p(x|x_0)p_0(x_0)$. The score entropy $\mathcal{L}_{\mathrm{SE}}$ is equivalent (up to a constant independent of θ) to the* ***denoising score entropy*** *$\mathcal{L}_{\mathrm{DSE}}$ is*

$$\mathop{\mathbb{E}}_{\substack{x_0 \sim p_0 \\ x \sim p(\cdot|x_0)}} \left[\sum_{y \neq x} w_{xy} \left(s_\theta(x)_y - \frac{p(y|x_0)}{p(x|x_0)} \log s_\theta(x)_y \right) \right]$$

[1] Campbell et al. A continuous time framework for discrete denoising models. *NeurIPS*, 2022.
[2] Lou et al. Discrete diffusion modeling by estimating the ratios of the data distribution. *ICML*, 2024.

Diffusion Weighted DSE

$$\frac{dp_{T-t}}{dt} = \overline{Q}_{T-t} p_{T-t} \quad \overline{Q}_t(y,x) = \frac{p_t(y)}{p_t(x)} Q_t(x,y)$$
$$\overline{Q}_t(x,x) = -\sum_{y \neq x} \overline{Q}_t(y,x) \tag{3}$$

- Parameterized reverse matrix:

Definition 3.5. *For our time dependent score network $s_\theta(\cdot, t)$, the parameterized reverse matrix is $\overline{Q}_t^\theta(y,x) = \begin{cases} s_\theta(x,t)_y Q_t(x,y) & x \neq y \\ -\sum_{z \neq x} \overline{Q}_t^\theta(z,y) & x = y \end{cases}$ found by replacing the ground truth scores in Equation 3. Our parameterized densities p_t^θ thus satisfy the following differential equation:*

$$\frac{dp^\theta_{T-t}}{dt} = \overline{Q}^\theta_{T-t} p^\theta_{T-t} \quad p^\theta_T = p_{\text{base}} \approx p_T$$

- Final objective function: *diffusion weighted DSE* (DWDSE)

Theorem 3.6 (Likelihood Training and Evaluation)**.** *For the diffusion and forward probabilities defined above,*

$$-\log p_0^\theta(x_0) \leq \mathcal{L}_{\text{DWDSE}}(x_0) + D_{KL}(p_{T|0}(\cdot|x_0) \parallel p_{\text{base}}) \tag{9}$$

*where $\mathcal{L}_{\text{DWDSE}}(x_0)$ is the **diffusion weighted denoising score entropy** for data point x_0*

$$\int_0^T \mathbb{E}_{x_t \sim p_{t|0}(\cdot|x_0)} \sum_{y \neq x_t} Q_t(x_t, y) \Bigg(s_\theta(x_t,t)_y - \frac{p_{t|0}(y|x_0)}{p_{t|0}(x_t|x_0)} \log s_\theta(x_t,t)_y + K\left(\frac{p_{t|0}(y|x_0)}{p_{t|0}(x_t|x_0)}\right) \Bigg) dt$$

Practical Implementation

- Suppose that data are sequences $\boldsymbol{x} = x^1 \cdots x^d$
- To avoid exponential-size Q_t, they perturb each token independently

$$Q_t(x^1 \dots x^i \dots x^d, x^1 \dots \widehat{x}^i \dots x^d) = Q_t^{\text{tok}}(x^i, \widehat{x}^i)$$

- Score network $s_\theta(\cdot, t) : \{1, \dots, n\}^d \to \mathbb{R}^{d \times n}$ as a seq-to-seq map:

$$(s_\theta(x^1 \dots x^i \dots x^d, t))_{i, \widehat{x}^i} \approx \frac{p_t(x^1 \dots \widehat{x}^i \dots x^d)}{p_t(x^1 \dots x^i \dots x^d)}$$

- Given noise level σ: $\mathbb{R}_+ \to \mathbb{R}_+$, let $\bar{\sigma}(t) \coloneqq \int_0^t \sigma(s)\mathrm{d}s$. Decomposition:

$$p_{t|0}^{\text{seq}}(\widehat{\mathbf{x}}|\mathbf{x}) = \prod_{i=1}^{d} p_{t|0}^{\text{tok}}(\widehat{x}^i|x^i) \qquad p_{t|0}^{\text{tok}}(\cdot|x) = x\text{-th column of } \exp\left(\bar{\sigma}(t) Q^{\text{tok}}\right)$$

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

Tweedie τ-Leaping

- Previous works use Euler τ-leaping [1] to simulate the reverse CTMC:

$$\delta_{x_t^i}(x_{t-\Delta t}^i) + \Delta t Q_t^{\text{tok}}(x_t^i, x_{t-\Delta t}^i) s_\theta(\mathbf{x}_t, t)_{i, x_{t-\Delta t}^i}$$

- This work derives the closed form of the denoiser:

Theorem 4.1 (Discrete Tweedie's Theorem). *Suppose that p_t follows the diffusion ODE $dp_t = Qp_t$. Then the true denoiser is given by*

$$p_{0|t}(x_0|x_t) = \left(\exp(-tQ)\left[\frac{p_t(i))}{p_t(x_t)}\right]_{i=1}^{N}\right)_{x_0} \exp(tQ)(x_t, x_0)$$

- However, we only know ratios of neighboring points
- They propose Tweedie τ-leaping and show it is the optimal τ-leaping

$$\left(\exp(-\sigma_t^{\Delta t} Q) s_\theta(\mathbf{x}_t, t)_i\right)_{x_{t-\Delta t}^i} \exp(\sigma_t^{\Delta t} Q)(x_t^i, x_{t-\Delta t}^i)$$

$$\text{where } \sigma_t^{\Delta t} = (\overline{\sigma}(t) - \overline{\sigma}(t - \Delta t))$$

Theorem 4.2 (Tweedie τ-leaping). *Let $p_{t-\Delta t|t}^{\text{tweedie}}(\mathbf{x}_{t-\Delta t}|\mathbf{x}_t)$ be the probability of the token update rule defined by Equation 19. Assuming s_θ is learned perfectly, this minimizes the KL divergence with the true reverse $p_{t-\Delta t|t}(\mathbf{x}_{t-\Delta t}|\mathbf{x}_t)$ for all τ-leaping strategies (i.e. token transitions are applied independently and simultaneously).*

[1] Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics 115*, 2001.
[2] Lou et al. Discrete diffusion modeling by estimating the ratios of the data distribution. *ICML*, 2024.

Main Results

- Evaluate the proposed method SEDD on zero-shot text generation
- Baselines: GPT-2 and other discrete diffusion models

$$Q^{\text{uniform}} = \begin{bmatrix} 1-N & 1 & \cdots & 1 \\ 1 & 1-N & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1-N \end{bmatrix}$$

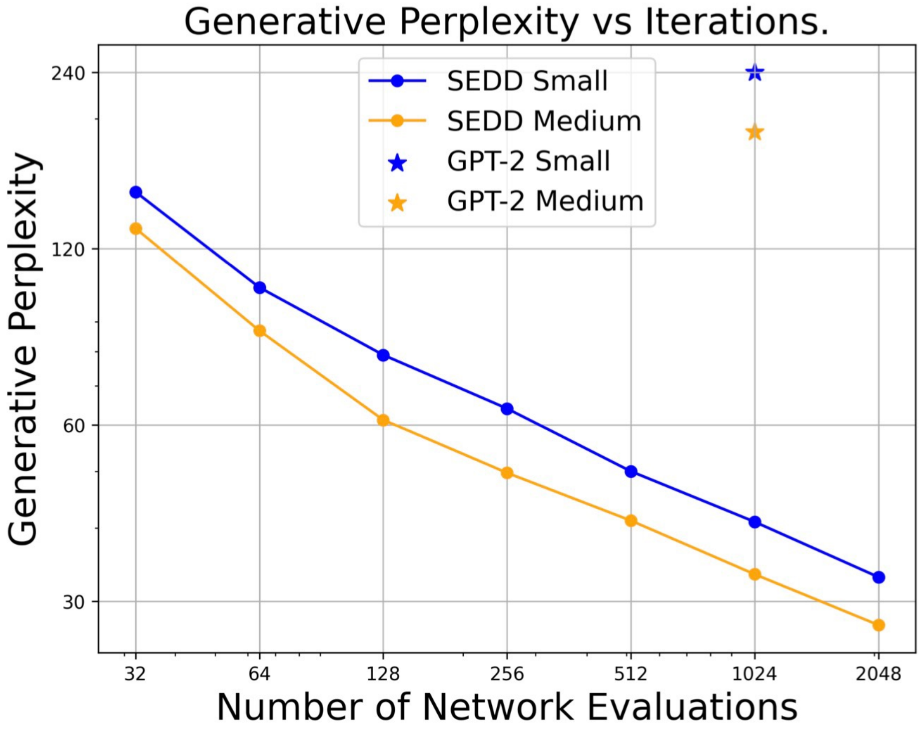
$$Q^{\text{absorb}} = \begin{bmatrix} -1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 0 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}$$

Size	Model	LAMBADA	WikiText2	PTB	WikiText103	1BW
Small	GPT-2	**45.04**	42.43	138.43	41.60	**75.20**
	SEDD Absorb	≤50.92	≤**41.84**	≤**114.24**	≤**40.62**	≤79.29
	SEDD Uniform	≤65.40	≤50.27	≤140.12	≤49.60	≤101.37
	D3PM	≤93.47	≤77.28	≤200.82	≤75.16	≤138.92
	PLAID	≤57.28	≤51.80	≤142.60	≤50.86	≤91.12
Medium	GPT-2	**35.66**	31.80	123.14	31.39	**55.72**
	SEDD Absorb	≤42.77	≤**31.04**	≤**87.12**	≤**29.98**	≤61.19
	SEDD Uniform	≤51.28	≤38.93	≤102.28	≤36.81	≤79.12

Table 1: **Zero-shot unconditional perplexity (↓) on a variety of datasets.** For a fixed size, the best perplexity is **bolded**. Our SEDD model with absorbing transition beats GPT-2 (Radford et al., 2019) on a majority of the tasks and entirely outperforms prior language diffusion models (Austin et al., 2021; Gulrajani & Hashimoto, 2023).

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

Main Results (Cont'd)



(a) Generative Perplexity (↓) vs. Sampling Iterations.

GPT-2 S	a hiring platform that "includes a fun club meeting place," says petitioner's AQQFredericks. They's the adjacent marijuana-hop. Others have allowed 3B Entertainment
GPT-2 M	misused, whether via Uber, a higher-order reality of quantified impulse or the No Mass Paralysis movement, but the most shamefully universal example is gridlock
SEDD S	As Jeff Romer recently wrote, "The economy has now reached a corner - 64% of household wealth and 80% of wealth goes to credit cards because of government austerity
SEDD M	Wyman worked as a computer science coach before going to work with the U.S. Secret Service in upstate New York in 2010. Without a license, the Secret Service will have to

(b) Generated Text (small models)

Figure 1: **Quality evaluation of unconditionally generated text.** We compare SEDD and GPT-2 by the perplexity of their analytically generated sequences. Our SEDD models consistently outperform GPT-2, interpolating between a 32× speedup and a 6-8× improvement based on the chosen step size. The generated text reflects this improved generation capability, as our samples are far more coherent. Additional samples and ablations can be found in Appendix D.3

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

Thank you!

Boat fry bear

rock tree ABCs

thing sleep

boat tennis