Fast Solvers for Discrete Diffusion Models: *Theory and Applications of High-Order Algorithms*

Yinuo Ren[*1] Haoxuan Chen[*1] Yuchen Zhu[*2] Wei Guo[*2]
Yongxin Chen[2] Grant M. Rotskoff[1] Molei Tao[2] Lexing Ying[1]

[*] Equal Contribution [1]Stanford University [2]Georgia Institute of Technology

Discrete Diffusion Model: an Introduction

- **Task:** Sample from a target distribution $p_0(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{X}$, $\mathbb{X}$ is a discrete set
- **Forward Process:** $\frac{\mathrm{d}\boldsymbol{p}_t}{\mathrm{d}t} = \boldsymbol{Q}_t \boldsymbol{p}_t$, where (i) $\forall x$, $Q_t(x,x) = -\sum_{y\neq x} Q_t(y,x)$; (ii) $\forall x \neq y$, $Q_t(x,y) \geq 0$
- **Backward Process** ($\overleftarrow{*}_t = *_{T-t}$):

$$\frac{\mathrm{d}\overleftarrow{\boldsymbol{p}}_s}{\mathrm{d}s} = \overleftarrow{\boldsymbol{Q}}_s \overleftarrow{\boldsymbol{p}}_s, \quad \text{where } \overleftarrow{Q}_s(y,x) = \begin{cases} \frac{\overleftarrow{p}_s(y)}{\overleftarrow{p}_s(x)} \overleftarrow{Q}_s(x,y), & \forall x \neq y \in \mathbb{X} \\ -\sum_{y'\neq x} \overleftarrow{Q}_s(y',x), & \forall x = y \in \mathbb{X} \end{cases}$$

- **Score Function:** $\widehat{\boldsymbol{s}}_t^\theta(x) \approx (s_t(x,y))_{y\in\mathbb{X}} := \frac{\boldsymbol{p}_t}{p_t(x)}$ parametrized by some neural network (NN) and trained by the following loss:

$$\min_\theta \int_0^T \psi_t \mathbb{E}_{x_t\sim p_t}\left[\sum_{y\neq x}\left(-\log\frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)} - 1 + \frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)}\right) s_t(x,y) Q_t(x,y)\right]\mathrm{d}t$$

Inference Schemes: Exact and Approximate

- **Exact Methods:** Uniformization [3], First-Hitting Sampler (FHS) [7], *etc.*, which may cause **redundant number of function evaluations (NFEs)**

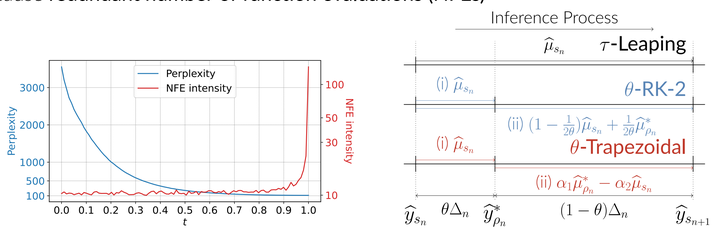


Figure 1. **Left:** An illustrative application of the uniformization algorithm to discrete diffusion models for text generation. Perplexity convergence occurs well before the NFE experiences unbounded growth. **Right:** Comparison between τ-leaping and the proposed second-order schemes (θ-RK-2 and θ-Trapezoidal).

- **Approximate Methods:** τ-Leaping scheme [1] (Euler-Maruyama)

$$\widehat{y}_{s_{n+1}} = \widehat{y}_{s_n} + \sum_{\nu\in\mathbb{D}} \nu \mathcal{P}\left(\widehat{\mu}_{s_n}(\nu)\Delta_n\right).$$

Above $\mathbb{D}$ is the set of all possible jumps from $\widehat{y}_{s_n}$, $\Delta_n = s_{n+1} - s_n$ is the step size, $\mathcal{P}(\lambda)$ denotes the Poisson distribution with parameter λ, and

$$\mu_\rho(\nu) := s_\rho(y_{\rho^-}, y_{\rho^-}+\nu)\overleftarrow{Q}^0_\rho(y_{\rho^-}, y_{\rho^-}+\nu),$$
$$\widehat{\mu}_\rho(\nu) := \widehat{s}^\theta_\rho(\widehat{y}_{\rho^-}, \widehat{y}_{\rho^-}+\nu)\overleftarrow{Q}^0_\rho(\widehat{y}_{\rho^-}, \widehat{y}_{\rho^-}+\nu)$$

are the true and estimated intensities for any time ρ, where $-$ denotes the left limit, $A^0 = A - \operatorname{diag} A$ for any matrix A. [6] shows that the error bound of τ-leaping is **first-order** w.r.t the step size κ:

$$D_{\mathrm{KL}}(p_0\|\widehat{q}_T) \lesssim \underbrace{\exp(-T)}_{\text{truncation error}} + \underbrace{\epsilon}_{\text{score estimation error}} + \underbrace{\kappa T}_{\text{numerical error}}$$

Question: *How to design faster approximate inference algorithms yielding better performance and error bound with the same NFE?*

Methodology: High-Order Inference Algorithm

Let the time discretization scheme $(s_i)_{i\in[0:N]}$ and θ-*section points* $(\rho_n)_{n\in[0:N]}$ be

$$0 = s_0 < s_1 < \cdots < s_N = T - \delta,\ \rho_n = (1-\theta)s_n + \theta s_{n+1}$$

Take $\alpha_1 = \frac{1}{2\theta(1-\theta)}$ and $\alpha_2 = \frac{(1-\theta)^2+\theta^2}{2\theta(1-\theta)}$ with $\alpha_1 - \alpha_2 = 1$. Then during the n-th step, we perform the following updates (**Figure 1, Right**):

- **Motivation: Runge-Kutta-2 methods** $(0 < \theta < 1)$ **for ODE** $\mathrm{d}x_t = f_t(x_t)\mathrm{d}t$

$$\widehat{x}^*_{t+\theta\Delta} = \widehat{x}_t + f_t(\widehat{x}_t)\theta\Delta, \quad \widehat{x}_{t+\Delta} = \widehat{x}_t + \left[(1-\tfrac{1}{2\theta})f_t(\widehat{x}_t) + \tfrac{1}{2\theta}f_{t+\theta\Delta}(\widehat{x}^*_{t+\theta\Delta})\right]\Delta.$$

- **θ-RK-2 Method**: Interpolation between s_n and ρ_n

$$\widehat{y}^*_{\rho_n} \leftarrow \widehat{y}_{s_n} + \sum_{\nu\in\mathbb{D}} \nu\mathcal{P}\left(\widehat{\mu}_{s_n}(\nu)\theta\Delta_n\right),$$
$$\widehat{y}_{s_{n+1}} \leftarrow \widehat{y}_{s_n} + \sum_{\nu\in\mathbb{D}} \nu\mathcal{P}\left(\mathbf{1}_{\widehat{\mu}_{s_n}>0}\left[(1-\tfrac{1}{2\theta})\widehat{\mu}_{s_n} + \tfrac{1}{2\theta}\widehat{\mu}^*_{\rho_n}\right]_+(\nu)\Delta_n\right).$$

- **θ-Trapezoidal Method**: Extrapolation beyond s_n and ρ_n

$$\widehat{y}^*_{\rho_n} \leftarrow \widehat{y}_{s_n} + \sum_{\nu\in\mathbb{D}} \nu\mathcal{P}\left(\widehat{\mu}_{s_n}(\nu)\theta\Delta_n\right),$$
$$\widehat{y}_{s_{n+1}} \leftarrow \widehat{y}^*_{\rho_n} + \sum_{\nu\in\mathbb{D}} \nu\mathcal{P}\left(\left(\alpha_1\widehat{\mu}^*_{\rho_n} - \alpha_2\widehat{\mu}_{s_n}\right)_+(\nu)(1-\theta)\Delta_n\right).$$

Theoretical Analysis (Informal Bound)

- **θ-RK-2 Method**: Suppose $\theta \in (0, \frac{1}{2}]$ and $(1-\frac{1}{2\theta})\widehat{\mu}_{\lfloor s\rfloor} + \frac{1}{2\theta}\widehat{\mu}^*_{\rho_s} \geq 0$, then

$$D_{\mathrm{KL}}\left(p_0\|\widehat{q}^{\mathrm{RK}}_T\right) \lesssim \exp(-T) + (\epsilon_{\mathrm{I}} + \epsilon_{\mathrm{II}})T + \boxed{\kappa^2 T}$$

- **θ-Trapezoidal Method**: Suppose $\theta \in (0, 1]$ and $\alpha_1\widehat{\mu}^*_{\rho_s} - \alpha_2\widehat{\mu}_{\lfloor s\rfloor} \geq 0$, then

$$D_{\mathrm{KL}}\left(p_0\|\widehat{q}^{\mathrm{trap}}_T\right) \lesssim \exp(-T) + (\epsilon_{\mathrm{I}} + \epsilon_{\mathrm{II}})T + \boxed{\kappa^2 T}$$

Assumptions

- **Exponential convergence of the forward process:** $D_{\mathrm{KL}}(p_T\|p_\infty) \lesssim \exp(-T)$.
- **Regularity of intensity:** Both the true intensity μ_s and the estimated intensity $\widehat{\mu}_s$ are in $\mathcal{C}^2$ and bounded for any $s \in [0, T-\delta]$.
- **Score estimation error**: The following error bounds hold for any grid point or θ-section point $s \in \cup_{n=0}^{N-1}\{s_n, \rho_n\}$:

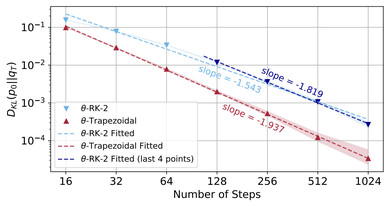
$$\mathbb{E}\left[\sum_{\nu\in\mathbb{D}}\left(\mu_s(\nu)\left(\log\frac{\mu_s(\nu)}{\widehat{\mu}_s(\nu)} - 1\right) + \widehat{\mu}_s(\nu)\right)\right] \leq \epsilon_{\mathrm{I}},$$
$$\mathbb{E}\left[\sum_{\nu\in\mathbb{D}}|\mu_s(\nu) - \widehat{\mu}_s(\nu)|\right] \leq \epsilon_{\mathrm{II}}.$$

Remarks

- **Proof Technique:** Change of measure [6] between stochastic integral with respect to *Poisson random measures*, error decomposition, and the usage of *Dynkin's Formula*;
- **Second-Order Numerical Error Guarantee:** $\mathcal{O}(\kappa^2 T)$ for both θ-RK-2 and θ-Trapezoidal methods vs. $\mathcal{O}(\kappa T)$ for τ-leaping method;
- **Range of θ:** $\theta \in (0, 1/2]$ for RK-2 method and $\theta \in (0, 1]$ for trapezoidal method, which is caused by application of *Jensen's Inequality*.

Experiments

- **Toy Model:** synthetic discrete distribution over 15 states
- **Text Generation:** RADD (GPT-2 Level) [5] on OpenWebText
- **Image Generation:** MaskGIT [2] on ImageNet-256
- **Math Reasoning:** LLaDA [4] (Instruct 8B) on GSM8K



Methods (Text)	NFE = 128	NFE = 1024
Euler	$\lesssim$ 86.28	$\lesssim$ 44.69
Tweedie τ-leap.	$\lesssim$ 85.74	$\lesssim$ 44.26
τ-leaping	$\lesssim$ 52.37	$\lesssim$ 28.80
θ-trapezoidal	$\lesssim$ **49.05**	$\lesssim$ **27.55**

Methods (Math)	NFE = 64	NFE = 128
Semi-AR (Conf.)	33.6%	32.0%
Semi-AR (Rand.)	33.8%	34.3%
θ-trapezoidal	**35.1%**	**38.4%**

Figure 2. **Toy Model, Text Generation and Math Reasoning**. **Left:** Empirical KL divergence between the true and generated distribution of the toy model (15 states) vs. NFE. **Upper Right:** Perplexity (on GPT-2 large) of generated text vs. NFE, **Bottom Right**: Response accuracy vs. NFE.

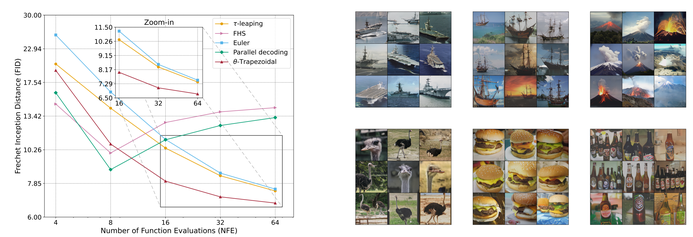


Figure 3. **Image Generation**, **Left:** FID of images generated by different sampling algorithms versus NFEs. Lower values are better. **Right:** Visualization of samples from ImageNet generated by θ-Trapezoidal.

References

[1] Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

[2] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

[3] Hongrui Chen and Lexing Ying. Convergence analysis of discrete diffusion model: Exact implementation through uniformization. *arXiv preprint arXiv:2402.08095*, 2024.

[4] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

[5] Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*, 2024.

[6] Yinuo Ren, Haoxuan Chen, Grant M Rotskoff, and Lexing Ying. How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. *arXiv preprint arXiv:2410.03601*, 2024.

[7] Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*, 2024.

Scan for Full Paper!

Scan for Code Repo!

{yinuoren, haoxuanc, rotskoff, lexing}@stanford.edu {yzhu738, wei.guo, yongchen, mtao}@gatech.edu

The 39-th Annual Conference on Neural Information Processing Systems (NeurIPS 2025), San Diego, CA