The Diffusion Duality

ICML 2025

Subham S. Sahoo[1], Justin Deschenaux[2], Aaron Gokaslan[1], Guanghan Wang[1], Justin Chiu[1], Volodymyr Kuleshov[1]

[1]Cornell Tech, [2]EPFL Lausanne

2025. 09. 12

Seungwoo Yoo

KAIST

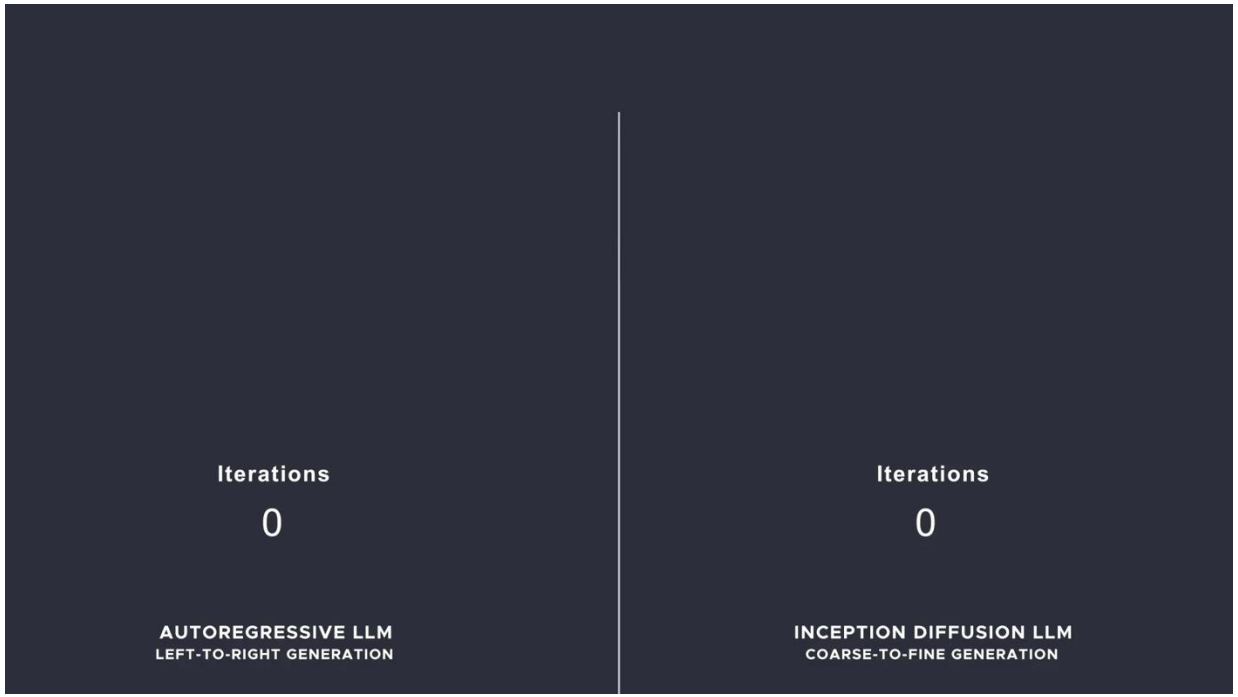
Motivation

Diffusion models have become the de facto standard for generative modeling in continuous domains, spanning images, videos, etc.

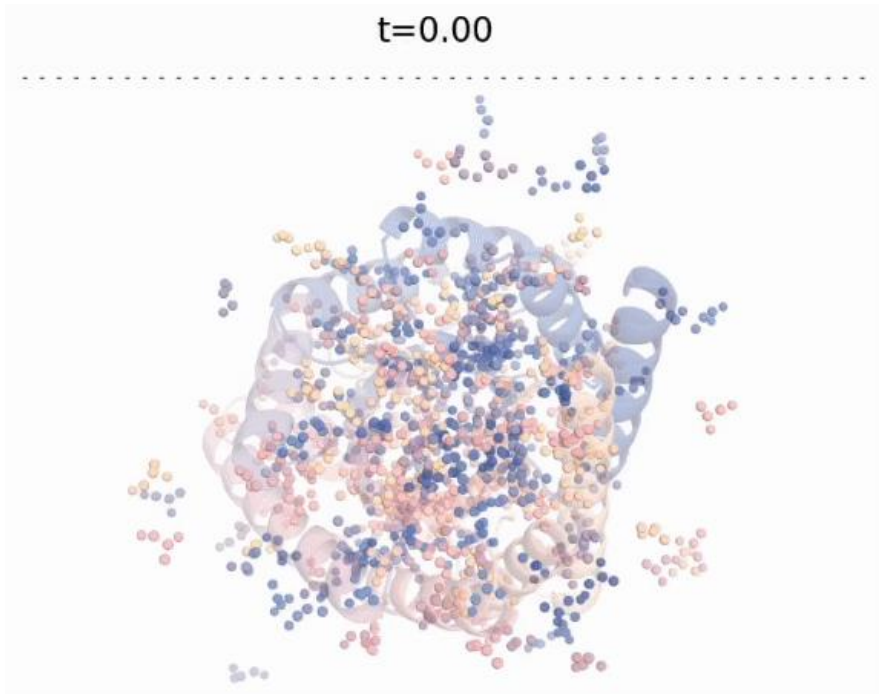
Stable Diffusion 2, Stability AI

Motivation

In recent years, efforts have been made to adapt the success of diffusion models to discrete data, which is essential for domains such as **text, graphs, and molecules**.



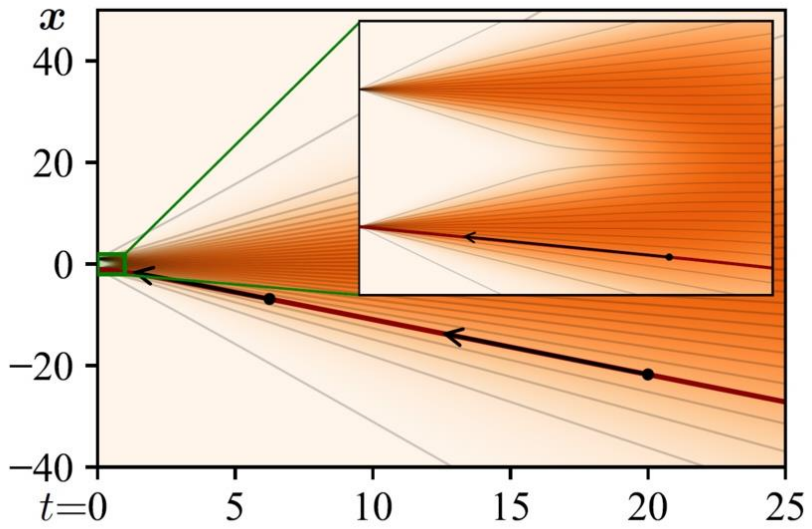
Mercury, Inception Labs



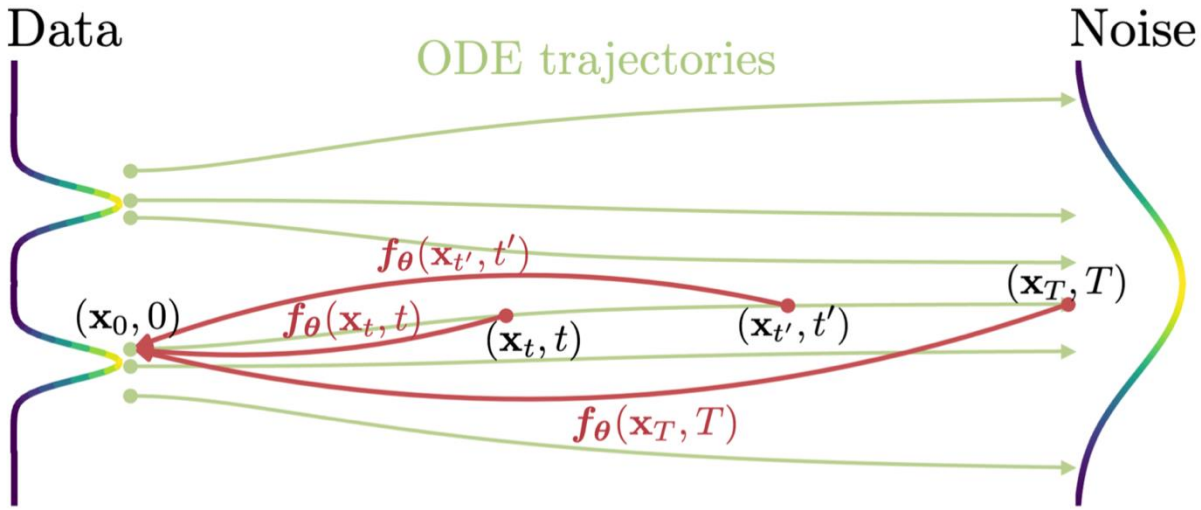
MultiFlow, Campbell *et al.*, ICML 2024

Motivation

In contrast to their continuous counterparts, many widely used techniques—such as distillation for faster sampling—remain underexplored in discrete diffusion.



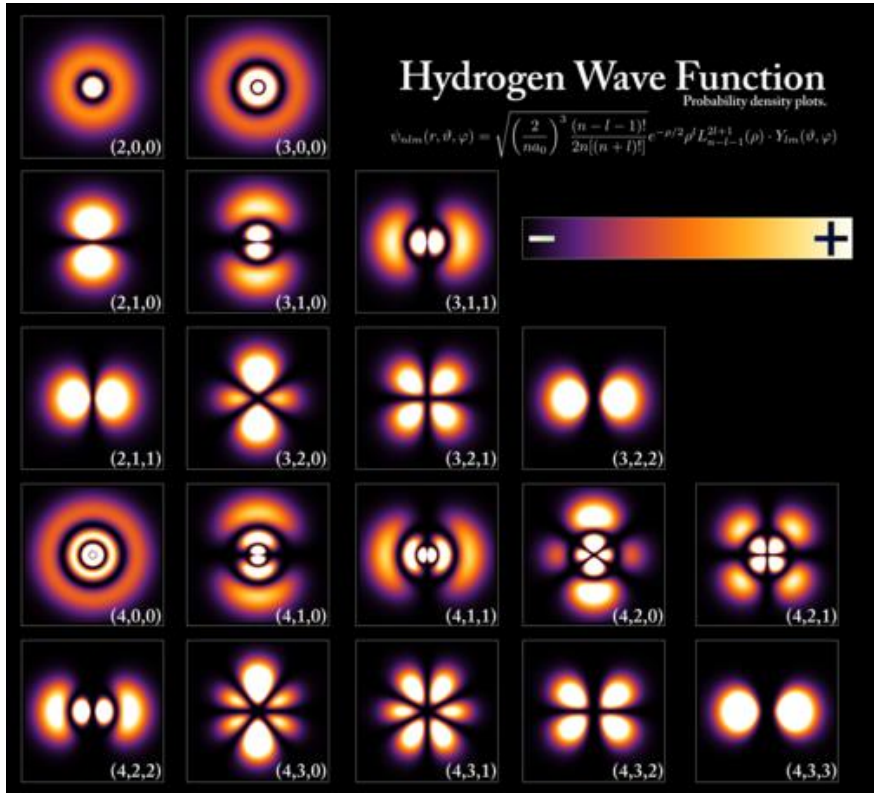
EDM, Karras *et al.*, NeurIPS 2022



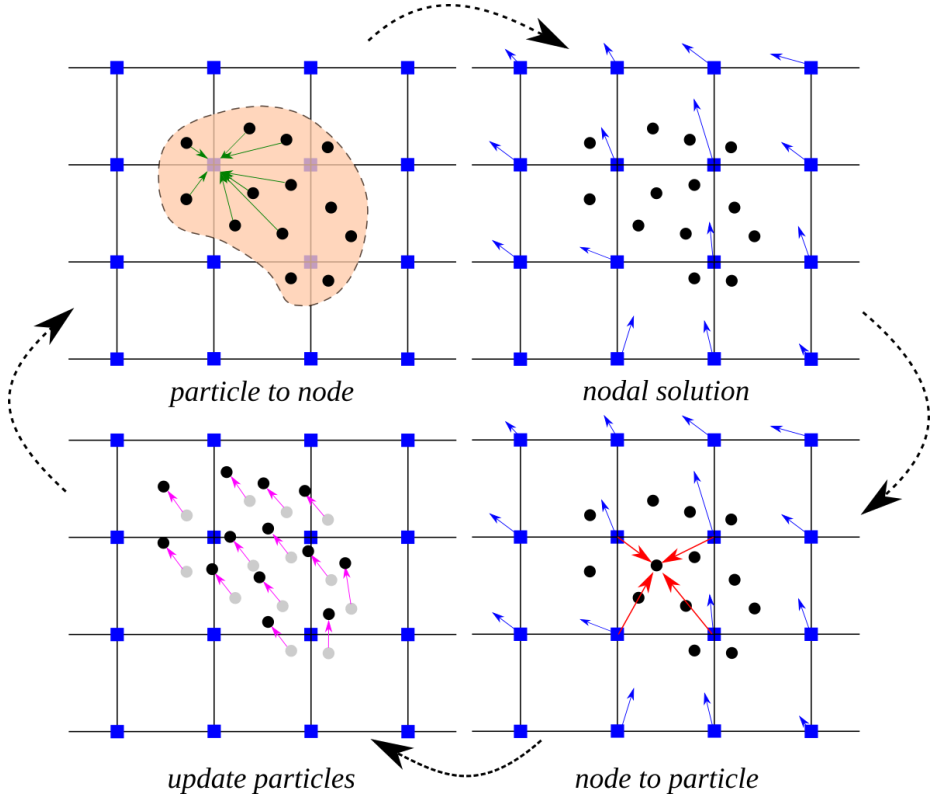
Consistency Models, Song *et al.*, ICML 2023

Motivation

In science and mathematics, surprising dualities have often provided deeper insights into complex phenomena, leading to breakthroughs in addressing challenging problems.



Orbitals as Solutions of Schrödinger's equation

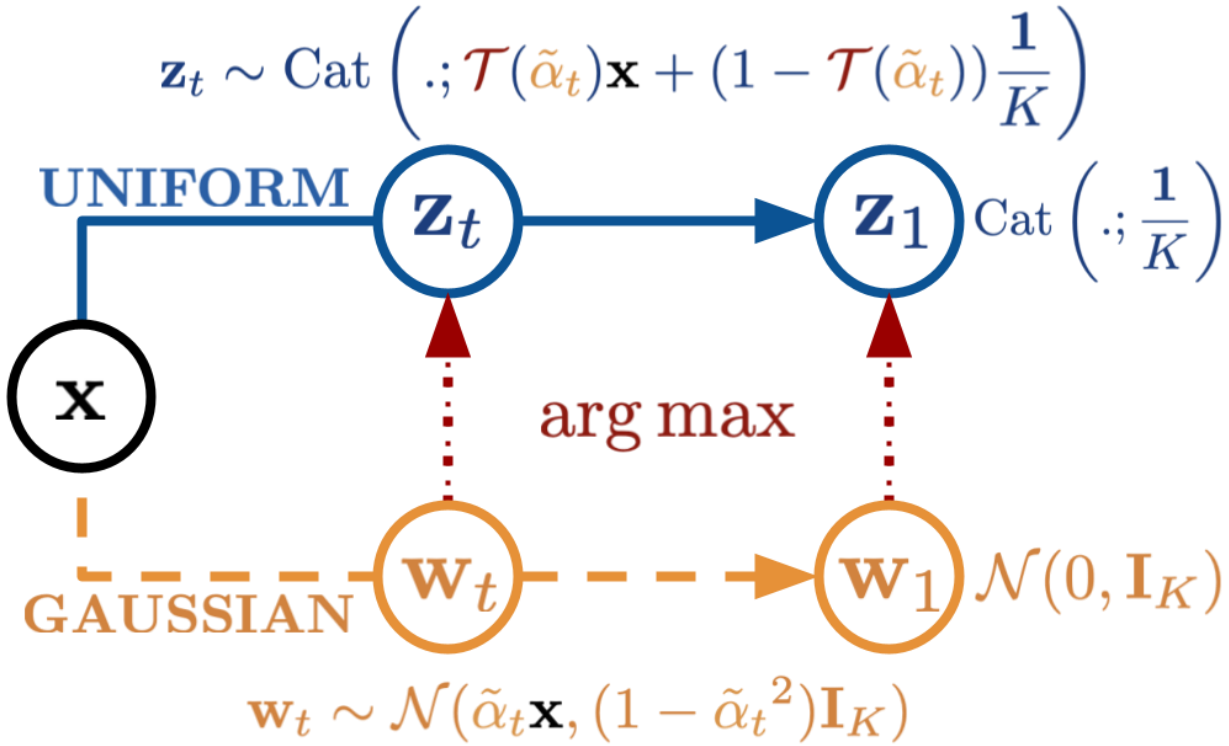


Visual Schematic of Material Point Method (MPM)

Does a fundamental duality exist between discrete and continuous diffusion?

Key Findings

This paper demonstrates that **Gaussian diffusion in the Euclidean space has a discrete counterpart in the probability simplex**, whose stationary distribution is uniform.



A visual schematic of the Diffusion Duality.

Key Findings

Leveraging this duality, the authors demonstrate that:

1. The training of uniform-state diffusion models (USDMs) can be stabilized through **curriculum learning**, leading to improved performance;

Key Findings

Leveraging this duality, the authors demonstrate that:

1. The training of uniform-state diffusion models (USDMs) can be stabilized through **curriculum learning**, leading to improved performance;
2. Discrete **Consistency Distillation** becomes feasible by emulating PF-ODE in the continuous domain, a formulation that is non-trivial to define for discrete diffusion.

Background: Discrete Diffusion Models

Let

- $\mathbf{x} \in \{0, 1\}^{\mathrm{K}}$: An instance of a scalar random variable that can take K values;
- $\mathcal{V}$: A set of one-hot vectors $\mathbf{x}$ called “dictionary”;
- $\mathrm{Cat}(\cdot; \boldsymbol{\pi})$: A categorical distribution over K classes with a probability mass function $\boldsymbol{\pi}$;
- $\mathbf{1} = \{1\}^{\mathrm{K}}$: A vector of ones;
- $\langle \mathbf{a}, \mathbf{b} \rangle$: Dot product between two vectors $\mathbf{a}$ and $\mathbf{b}$;
- $\mathbf{a} \odot \mathbf{b}$: Hadamard (Element-wise) product between two vectors $\mathbf{a}$ and $\mathbf{b}$;
- $[\mathbf{x}^l]_{l=1}^{L} \in \mathcal{V}^L$: A sequence of length L.

Background: Discrete Diffusion Models

Consider a clean token $\mathbf{x} \in \mathcal{V}$ drawn from the data distribution q_{data}.

The forward process smoothly transforms the token toward a prior distribution $\text{Cat}(\cdot; \boldsymbol{\pi})$:

$$q_t(\cdot\,|\mathbf{x}; \alpha_t) = \text{Cat}(\cdot; \alpha_t \mathbf{x} + (1 - \alpha_t)\boldsymbol{\pi}),$$

where $\alpha_t \in [0, 1]$ is a strictly decreasing function with $\alpha_0 \approx 1$ and $\alpha_1 \approx 0$.

Background: Discrete Diffusion Models

The evolution of marginal q_t is described by a linear ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t = Q_t q_t,$$

with $Q_t \in \mathbb{R}^{\mathrm{K}\times\mathrm{K}}$ denoting the state transition matrix.

Background: Discrete Diffusion Models

The matrix Q_t varies depending on the stationary distribution $\mathrm{Cat}(\cdot; \boldsymbol{\pi})$, toward which the diffusion process converges. Two common choices are:

1. Uniform: $\boldsymbol{\pi} = 1/\mathrm{K}$
2. Absorbing (Mask): $\boldsymbol{\pi} = \mathbf{m}$

where $\mathbf{m} \in \mathcal{V}$ is a special mask token appended to the dictionary $\mathcal{V}$.

Background: Discrete Diffusion Models

When $\boldsymbol{\pi} = \mathbf{1}/\mathrm{K}$, the model is called Uniform-State Diffusion Model (USDM), and its state transition matrix Q_t is given by:

$$Q_t = \frac{\alpha_t'}{\mathrm{K}\alpha_t}[\mathbf{1}\mathbf{1}^{\mathrm{T}} - \mathrm{K}\mathbf{I}],$$

where α_t' is the time derivative of α_t.

Background: Discrete Diffusion Models

Its time-reversal, which is the core of generative modeling, is given as:

$$q_{s|t}(\cdot\,|\mathbf{z}_t, \mathbf{x}) = \mathrm{Cat}(\cdot;\ \frac{\mathrm{K}\,\alpha_t \mathbf{z}_t \odot \mathbf{x} + (\alpha_{t|s} - \alpha_t)\mathbf{z}_t}{\mathrm{K}\alpha_t \langle \mathbf{z}_t, \mathbf{x}\rangle + 1 - \alpha_t} + \frac{(\alpha_s - \alpha_t)\mathbf{x} + (1 - \alpha_{t|s})(1 - \alpha_s)\mathbf{1}/\mathrm{K}}{\mathrm{K}\alpha_t \langle \mathbf{z}_t, \mathbf{x}\rangle + 1 - \alpha_t})$$

where $\alpha_{t|\mathrm{s}} = \frac{\alpha_t}{\alpha_s}$.

Since $\mathbf{x}$ is unavailable during inference, it is predicted by a neural network

$$\mathbf{x}_\theta \colon \mathcal{V} \times [0,1] \to \Delta^{\mathrm{K}},$$

yielding an approximate reverse posterior

$$p^{\theta}_{s|t}(\cdot\,|\mathbf{z}_t) = q_{s|t}(\cdot\,|\mathbf{z}_t, \mathbf{x} = \mathbf{x}_\theta(\mathbf{z}_t, t)).$$

Background: Discrete Diffusion Models

The parameters θ are updated by optimizing the **negative evidence lower bound (NELBO)**:

$$\mathrm{NELBO}(q, p_\theta; \mathbf{x}) = \mathbb{E}_{t\sim\mathcal{U}[0,1], q_t(\mathbf{z}_t|\mathbf{x};\alpha_t)} f(\mathbf{z}_t, \mathbf{x}_\theta(\mathbf{z}_t, t), \alpha_t; \mathbf{x}),$$

where f is defined as

$$f(\mathbf{z}_t, \mathbf{x}_\theta(\mathbf{z}_t, t), \alpha_t; \mathbf{x}) = -\frac{\alpha_t'}{\mathrm{K}\alpha_t}\left[\frac{\mathrm{K}}{\bar{\mathbf{x}}_i} - \frac{\mathrm{K}}{(\bar{\mathbf{x}}_\theta)_i} - \sum_j \frac{\bar{\boldsymbol{x}}_j}{\bar{\mathbf{x}}_i}\log\frac{(\bar{\mathbf{x}}_\theta)_i \cdot \bar{\mathbf{x}}_j}{(\bar{\mathbf{x}}_\theta)_j \cdot \bar{\mathbf{x}}_i}\right],$$

with $\bar{\mathbf{x}} = \mathrm{K}\,\alpha_t\mathbf{x} + (1-\alpha_t)\mathbf{1}$ and $\bar{\mathbf{x}}_\theta = \mathrm{K}\,\alpha_t\mathbf{x}_\theta(\boldsymbol{z}_t, t) + (1-\alpha_t)\mathbf{1}$.

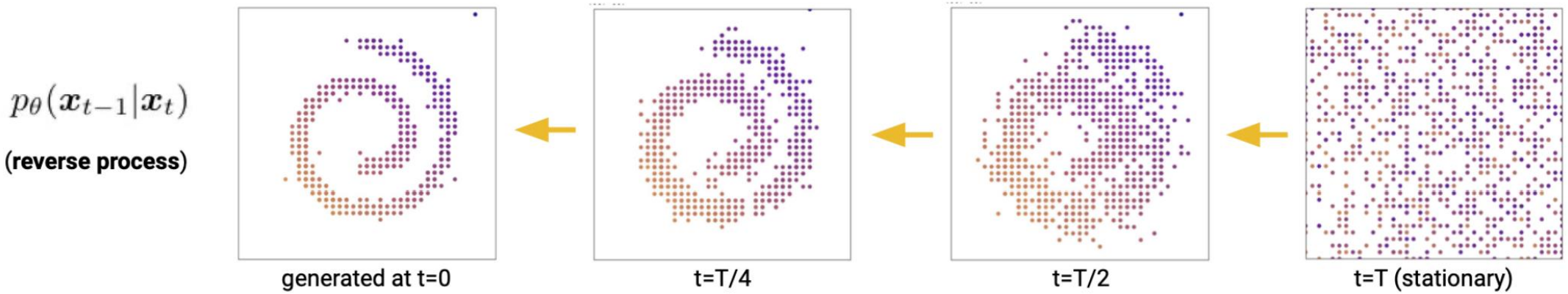
Background: Discrete Diffusion Models

After training the model, new samples are generated by first sampling from the prior

$$\mathbf{z}_{t=1} \sim \frac{\mathbf{1}}{\mathrm{K}}$$

and iteratively performing the ancestral sampling

$$\mathbf{z}_s \sim p^{\theta}_{s|t}(\cdot\,|\mathbf{z}_t).$$



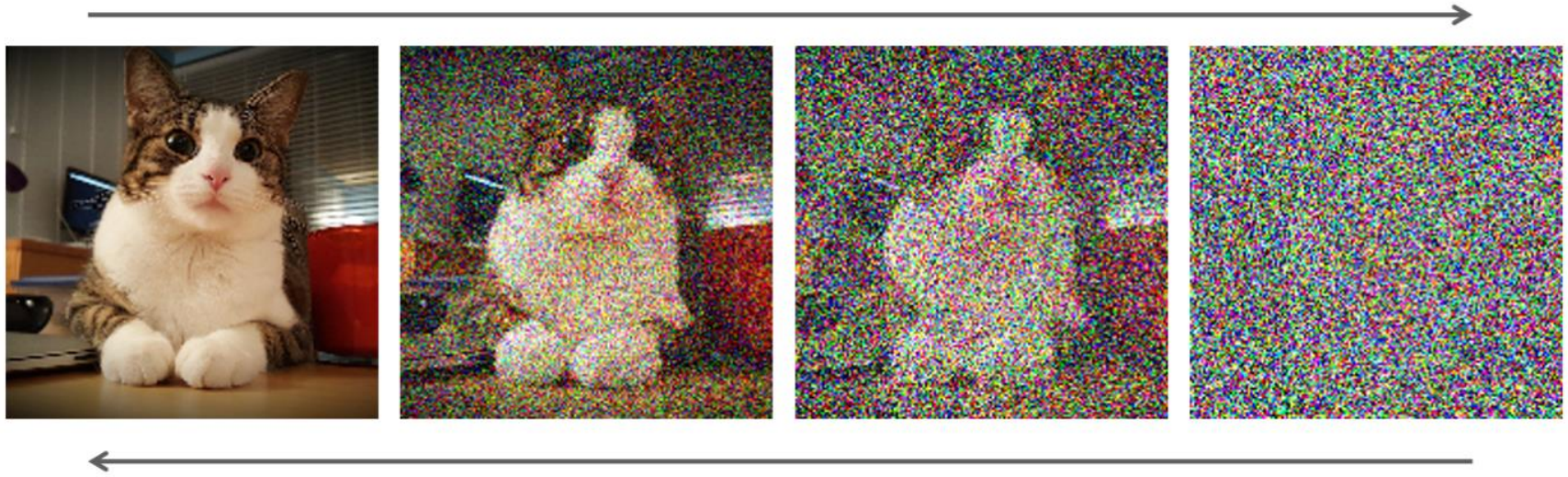
D3PM, Austin *et al.*, NeurIPS 2021

Background: Gaussian Diffusion Models

In continuous domains, Gaussian diffusion maps a data distribution q_{data} to an easy-to-sample prior distribution, typically a unit Gaussian $\mathcal{N}(0, \mathbf{I}_{\text{K}})$, along marginals

$$\tilde{q}_t(\cdot\,|\mathbf{x}; \tilde{\alpha}_t) = \mathcal{N}(\tilde{\alpha}_t \mathbf{x}, (1 - \tilde{\alpha}_t^2)\mathbf{I}_{\text{K}}),$$

where the diffusion parameter $\tilde{\alpha}_t \in [0,1]$ is a monotonically decreasing function.



Background: Gaussian Diffusion Models

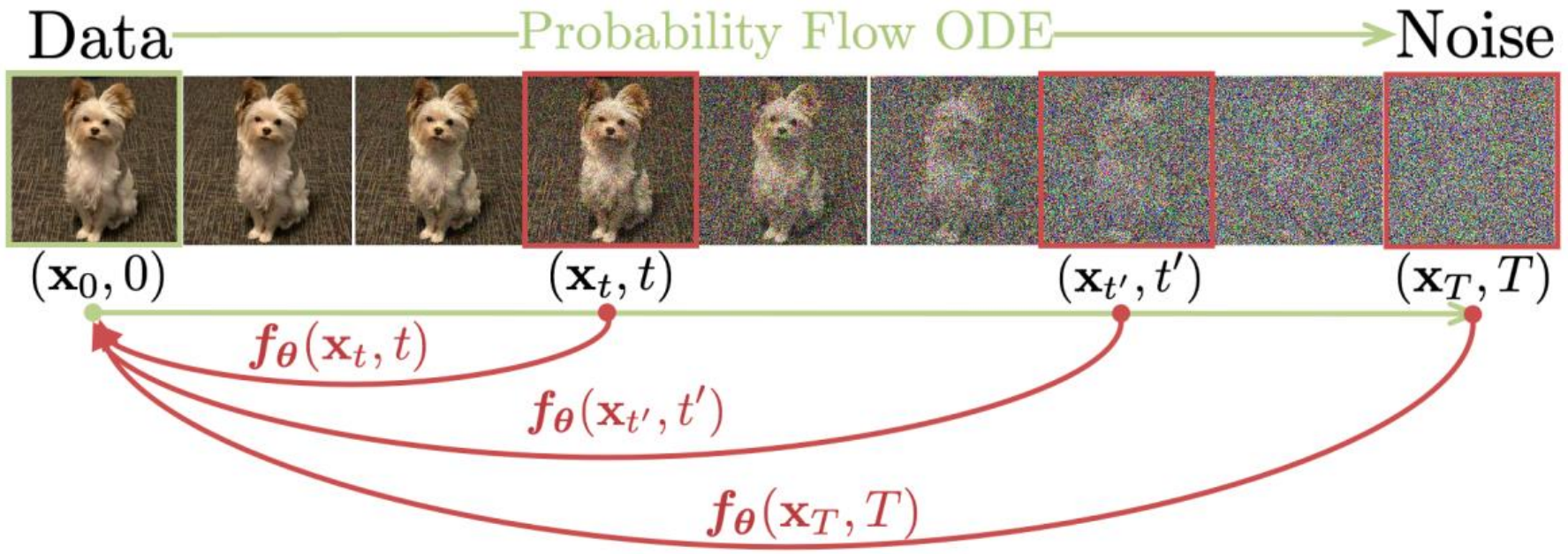
When $\tilde{\alpha}_{t=0} = 1$ and $\tilde{\alpha}_{t=1} = 0$, the NELBO for learning the process is given by:

$$\mathrm{NELO}(\tilde{q}, p_\theta, \mathbf{x}) = -\mathbb{E}_{t\sim\mathcal{U}[0,1],\tilde{q}_t(\mathbf{w}_t|\mathbf{x};\ \tilde{\alpha}_t)} v'(t)\|\mathbf{x} - \mathbf{x}_\theta(\mathbf{w}_t, t)\|_2^2$$

where $v'(t)$ is the time derivative of the signal-to-noise ratio $v(t) = \tilde{\alpha}_t^2/(1-\tilde{\alpha}_t^2)$.

Background: Consistency Distillation

Consistency Distillation [Song *et al.*, 2023] is a technique for distilling existing diffusion models for few-step generation.



Consistency Models, Song *et al.*, ICML 2023

Background: Consistency Distillation

For distillation, there must exist a deterministic PF-ODE corresponding to p_t:

$$d\mathbf{x}_t = \left[\mu(\mathbf{x}_t, t) - \frac{1}{2}\sigma(t)^2 \nabla \log p_t(\mathbf{x}_t)\right] dt.$$

With it, a student model is optimized by

1. Perturbing a data sample $\mathbf{x}$ via forward process $\mathbf{w}_t \sim \tilde{q}_t(\cdot\,|\mathbf{x})$;
2. Solving one PF-ODE step using the teacher $\mathbf{x}_{\theta^-}$, obtaining $\mathbf{w}_s$ at $s < t$;
3. Minimize the gap between clean sample estimates from the teacher and student:

$$\mathcal{L}(\theta, \theta^-) = \lambda(t) d\big(\mathbf{x}_\theta(\mathbf{w}_t, t), \mathbf{x}_{\theta^-}(\mathbf{w}_s, s)\big)$$

4. Repeat the above step until convergence.

The Diffusion Duality

Our main goal is to **bridge discrete-state and continuous-state diffusion**, enabling the transfer of techniques from the latter to improve the former.

The Diffusion Duality

Our main goal is to **bridge discrete-state and continuous-state diffusion**, enabling the transfer of techniques from the latter to improve the former.

Interestingly, a Gaussian latent $\mathbf{w}_t$ can be mapped to a discrete one-hot vector $\mathbf{z}_t$ by

$$\boldsymbol{z}_t = \text{argmax}\, \mathbf{w}_t$$

The Diffusion Duality

Our main goal is to **bridge discrete-state and continuous-state diffusion**, enabling the transfer of techniques from the latter to improve the former.

Interestingly, a Gaussian latent $\mathbf{w}_t$ can be mapped to a discrete one-hot vector $\mathbf{z}_t$ by

$$\boldsymbol{z}_t = \text{argmax}\, \mathbf{w}_t$$

However, we must show that the marginal distribution q_t of such $\boldsymbol{z}_t$'s satisfies the ODE:

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t = Q_t q_t$$

The Diffusion Duality

As in continuous domains, we can *diffuse* the discrete token $\mathbf{x} \in \mathcal{V}$ by directly applying the forward process $\mathbf{w}_t \sim \tilde{q}_t(\cdot\,|\mathbf{x};\ \tilde{\alpha}_t)$.

We define the operation $\mathrm{argmax}: \mathbb{R}^K \to \mathcal{V}$ that maps a continuous vector $\mathbf{w}_t \in \mathbb{R}^K$ to the one-hot vector corresponding to the index of its largest entry

$$\mathrm{argmax}(\mathbf{w}_t) = \mathrm{argmax}_{\mathbf{z}\in\mathcal{V}} \mathbf{z}^T \mathbf{w}_t.$$

The Diffusion Duality

After doing a bit of math, we can show that

$$\mathbf{z}_t \sim P_t(\cdot\,|\mathbf{x}; \mathcal{T}(\tilde{\alpha}_t)) = \mathrm{Cat}(\cdot; \mathcal{T}(\tilde{\alpha}_t)\mathbf{x} + \left(1 - \mathcal{T}(\tilde{\alpha}_t)\right)\frac{\mathbf{1}}{\mathrm{K}}),$$

where the function $\mathcal{T}: [0,1] \times [0,1]$ is the **Diffusion Transformation Operator**:

$$\mathcal{T}(\tilde{\alpha}_t) = \frac{\mathrm{K}}{\mathrm{K}-1}\left[\int_{-\infty}^{\infty} \phi\left(z - \frac{\tilde{\alpha}_t}{\sqrt{1-\tilde{\alpha}_t^2}}\right)\Phi^{\mathrm{K}-1}(\mathrm{z})\mathrm{dz} - \frac{1}{\mathrm{K}}\right],$$

where

- $\phi(\mathrm{z}) = \exp(-\frac{\mathrm{z}^2}{2})/\sqrt{2\pi}$: PDF of the standard Normal distribution;
- $\Phi(\mathrm{z}) = \int_{-\infty}^{\mathrm{z}} \exp\left(-\frac{t^2}{2}\right) dz/\sqrt{2\pi}$: CDF of the standard Normal distribution.

The Diffusion Duality

Furthermore, it is proven that the discrete marginal P_t satisfies the linear ODE:

$$\frac{d}{dt}P_t = -\frac{\mathcal{T}'(\tilde{\alpha}_t)}{\mathrm{K}\mathcal{T}(\tilde{\alpha}_t)}[\mathbf{1}\mathbf{1}^{\mathrm{T}} - \mathrm{K}\mathbf{I}]P_t,$$

where $\mathcal{T}'$ is the time derivative of $\mathcal{T}$.

If we define the matrix multiplied to P_t on the RHS as Q_t, then **the equation is identical to the linear ODE of the uniform-state discrete diffusion processes.**

The Diffusion Duality

So far, we have shown that

"The argmax operation transforms Gaussian diffusion into uniform-state discrete diffusion, with the diffusion parameters $\tilde{\alpha}_t$ (Gaussian) and $\mathcal{T}(\tilde{\alpha}_t)$ (Discrete) related by

$$\mathcal{T}(\tilde{\alpha}_t) = \frac{\mathrm{K}}{\mathrm{K}-1}\left[\int_{-\infty}^{\infty} \phi\left(z - \frac{\tilde{\alpha}_t}{\sqrt{1-\tilde{\alpha}_t^2}}\right) \Phi^{\mathrm{K}-1}(\mathrm{z})\mathrm{dz} - \frac{1}{\mathrm{K}}\right]"$$

The Diffusion Duality

More directly, between two marginals q_t and $\tilde{q}_t$, the following equation holds:

$$q_t(\mathbf{z}_t|\mathbf{x}; \mathcal{T}(\tilde{\alpha}_t)) = [\text{argmax}]_\star \tilde{q}_t(\mathbf{w}_t|\mathbf{x}; \tilde{\alpha}_t)$$

Where the operator $\star$ is the **pushforward** of the K-dimensional Gaussian density $\tilde{q}_t$ under the map argmax, which yields a categorical distribution q_t with K classes.

The Diffusion Duality

Despite bridged by the argmax operators, q_t and $\tilde{q}_t$ are the marginals of two independent Markov processes, which **induce different variational bounds** on the log-likelihood.

Specifically, it can be shown that

$$\log p_\theta(\mathbf{x}) \geq \mathrm{ELBO}\,(q, p_\theta; \mathbf{x}) \geq \mathrm{ELBO}(\tilde{q}, p_\theta; \mathbf{x}),$$

with the equality holds when $p_\theta(\mathbf{x})$ is the optimal denoiser. This implies that **the training objective of discrete diffusion provides a tighter bound on the log-likelihood** and is directly used for training.

The Diffusion Duality

To model sequences $\mathbf{x}^{1:\mathrm{L}} \sim q_{\mathrm{data}}$, the authors follow prior works and impose **token-wise independence assumption** to factorize both forward and reverse processes:

$$q_t\left(\mathbf{z}_t^{1:\mathrm{L}} \middle| \mathbf{x}^{1:\mathrm{L}}; \alpha_t\right) = \Pi_{l\in[\mathrm{L}]} q_t(\mathbf{z}_t^l | \mathbf{x}^l; \alpha_t)$$

$$p_\theta\left(\mathbf{z}_s^{1:\mathrm{L}} \middle| \mathbf{z}_t^{1:\mathrm{L}}\right) = \Pi_{l\in[\mathrm{L}]} q_{s|t}(\mathbf{z}_s^l | \mathbf{z}_t^{1:\mathrm{L}}, \mathbf{x}_\theta^l(\mathbf{z}_t^{1:\mathrm{L}}, t))$$

where $\mathbf{x}_\theta: \mathcal{V}^{\mathrm{L}} \times [0,1] \to \Delta^{\mathrm{L}}$ is the (learned) denoising model.

The Diffusion Duality

With the trained model, the authors use a **Greedy-Tail Sampler**, a slightly modified version of ancestral sampler that trade-offs sample quality with the entropy.

In particular, at the last denoising step, the algorithm takes the argmax during decoding:

$$\tilde{\mathbf{x}} = \text{argmax}\left(\text{p}_{0|\delta}^{\theta}(\cdot)\right),$$

instead of drawing a sample from the categorical distribution $\text{Cat}(\cdot; \text{p}_{0|\delta}^{\theta}(\cdot))$.

Applications

The authors explore two applications enabled by the found duality:

1. **A curriculum learning strategy** that reduces training variance for faster training;
2. **A (consistency) distillation algorithm** exploiting PF-ODE path on the continuous side.

Faster Training using Curriculum Learning

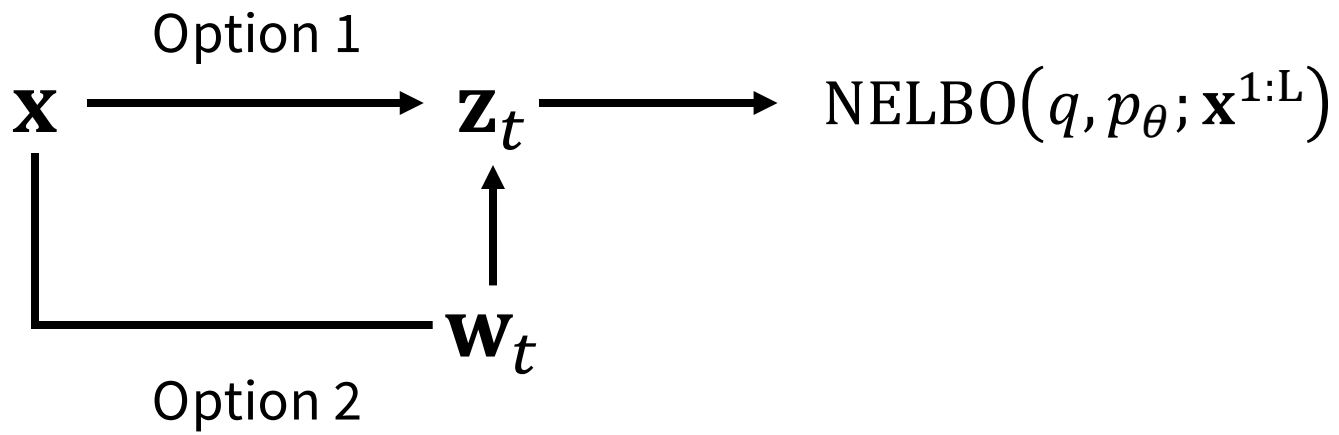
From the duality, it can be shown that the discrete diffusion NELBO for sequences

$$\mathrm{NELBO}\left(q, p_\theta; \mathbf{x}^{1:\mathrm{L}}\right) = \mathbb{E}_{t\sim\mathcal{U}[0,1],q_t} \sum_{l\in[\mathrm{L}]} f_{\mathrm{Duo}}\left(\mathbf{z}_t^l, \mathbf{x}_\theta^l\left(\mathbf{z}_t^{1:\mathrm{L}}, t\right), \alpha_t; \mathbf{x}^l\right),$$

is equivalent to

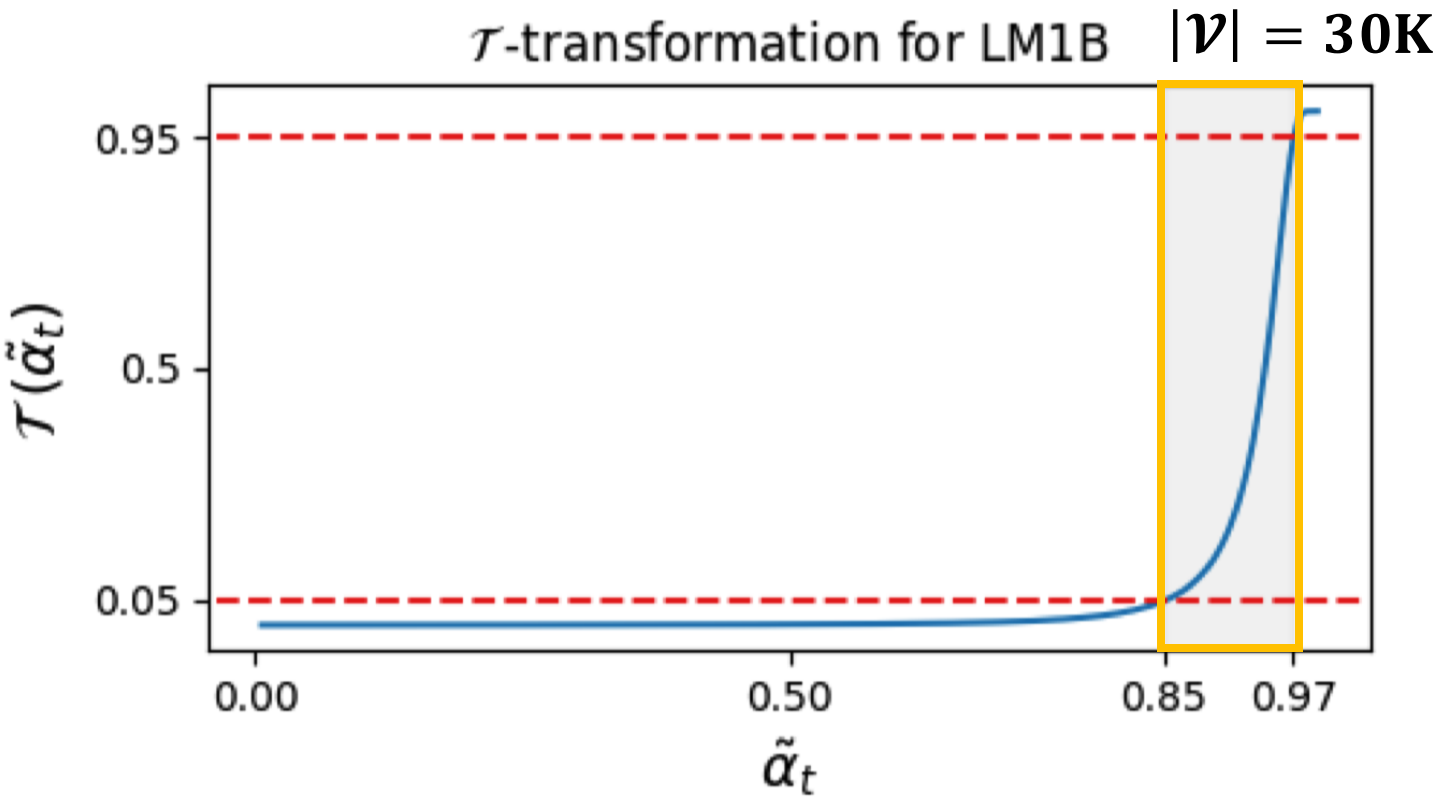
$$\mathrm{NELBO}\left(q, p_\theta; \mathbf{x}^{1:\mathrm{L}}\right)$$

$$= \mathbb{E}_{t\sim\mathcal{U}[0,1],\tilde{q}_t} \sum_{l\in[\mathrm{L}]} f_{\mathrm{Duo}}\left(\mathbf{z}_t^l = \mathrm{argmax}\left(\mathbf{w}_t^l\right), \mathbf{x}_\theta\left(\left[\mathrm{argmax}\left(\mathbf{w}_t^{l'}\right)\right]_{l'=1}^{\mathrm{L}}, t\right), \alpha_t = \mathcal{T}(\tilde{\alpha}_t); \mathbf{x}^l\right).$$



Faster Training using Curriculum Learning

Notably, when K is large, the operator $\mathcal{T}$ maps $\tilde{\alpha}_t$ for t in a small sub-interval $[a, 1]_{0 \le a \le 1}$ to the almost entire range $[0,1]$ of $\alpha_t = \mathcal{T}(\tilde{\alpha}_t)$.

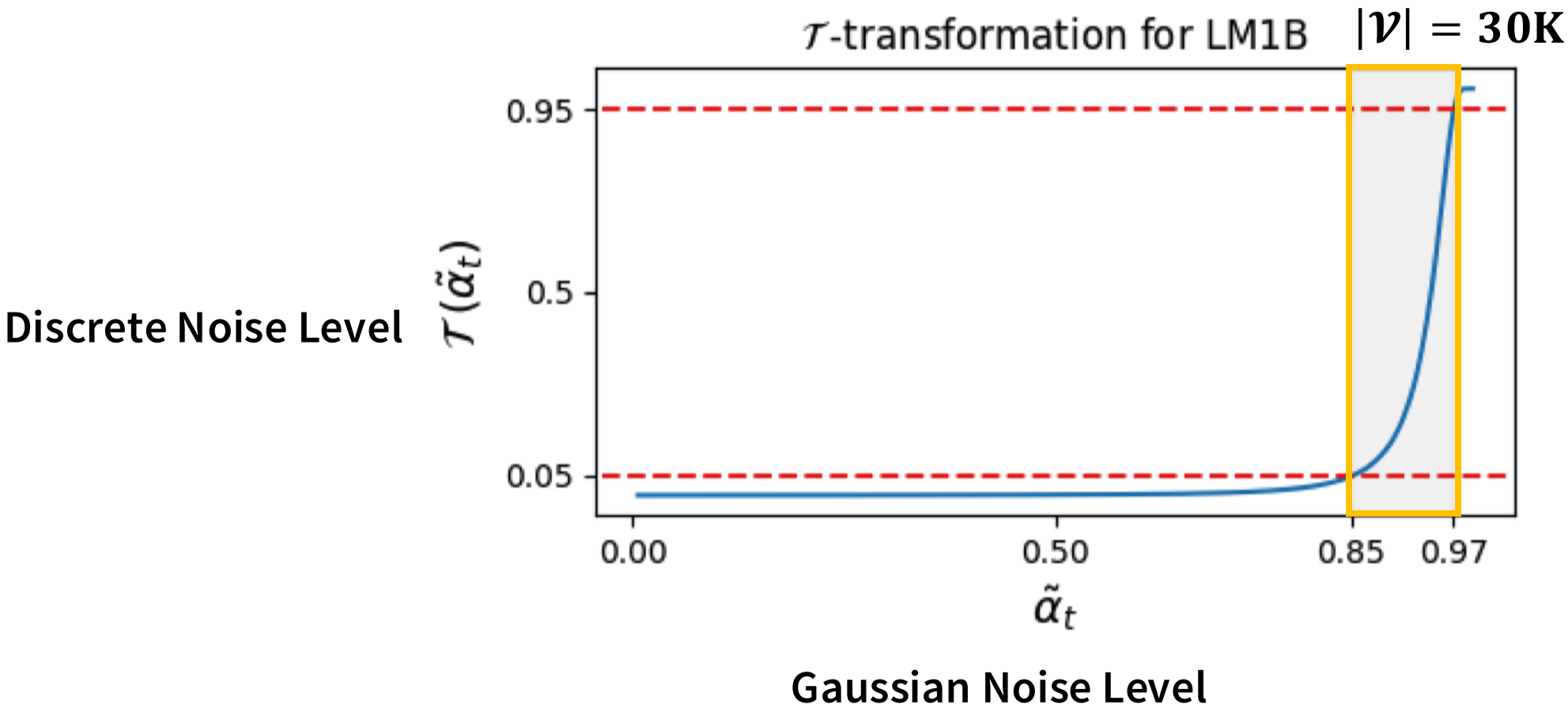


Discrete Noise Level

Gaussian Noise Level

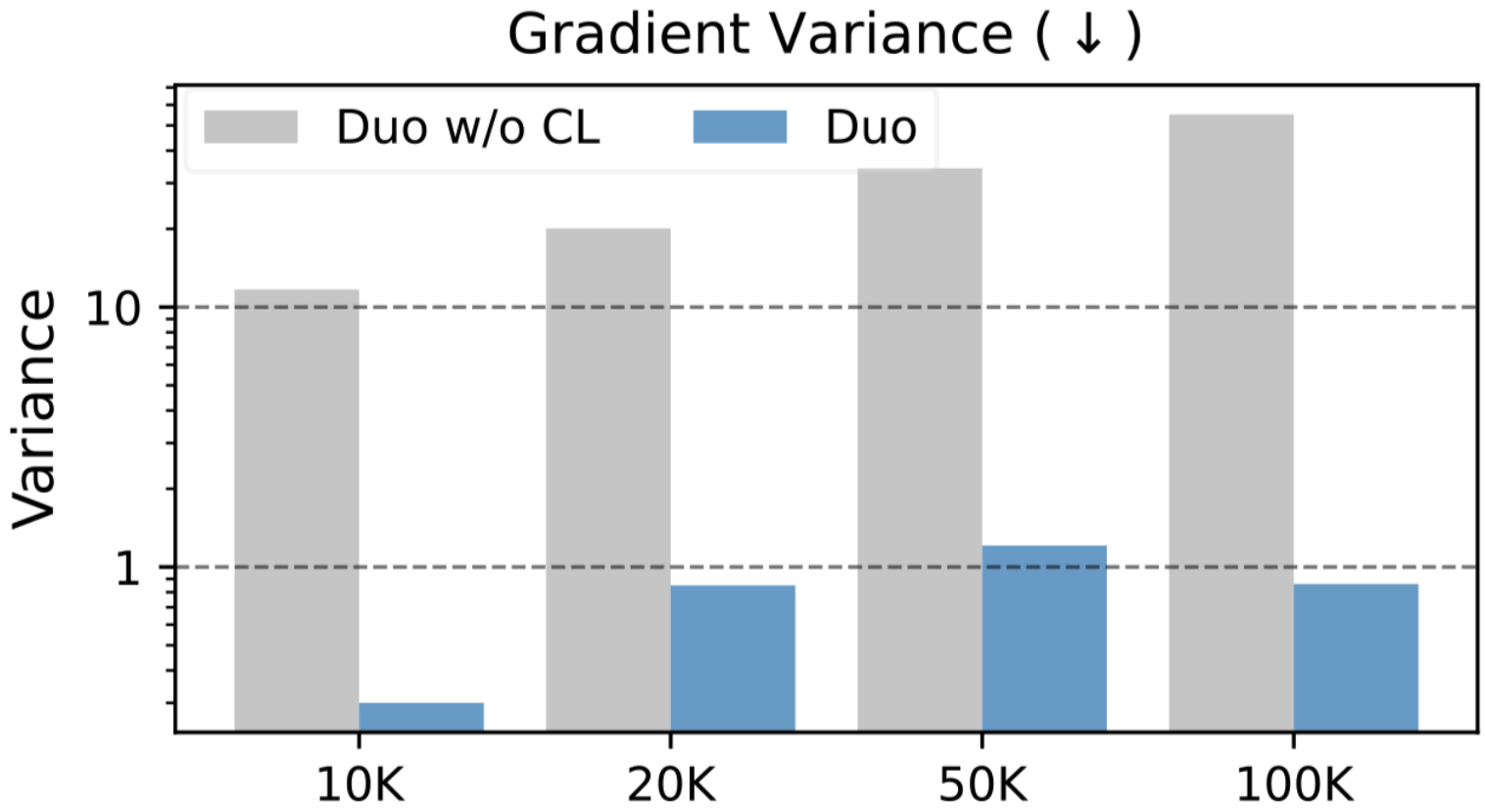
Faster Training using Curriculum Learning

This implies that the argmax operator is **highly sensitive to slight logit noise**, yielding a large noise scale in the corresponding discrete uniform diffusion process.



Faster Training using Curriculum Learning

This sensitivity induces high variance in the training loss and gradients, leading to unstable optimization and slow convergence.



Comparison of the summed gradient variance of the top 100 weights.

Faster Training using Curriculum Learning

To address this, the authors propose a curriculum learning method that **anneals the temperature parameter** τ of $\mathrm{softmax}$, an approximation of argmax during training:

$$\mathrm{argmax}(\mathbf{w}_t^l) = \lim_{\tau \to 0^+} \mathrm{softmax}(\mathbf{w}_t^l / \tau).$$

By altering τ over training, the $\mathrm{softmax}$ converges to argmax, and the authors claim that this reduce training variance by easing recovery of the clean sequence from perturbed one.

Faster Training using Curriculum Learning

For this, the denoising model $\mathbf{x}_\theta : \Delta^{\mathrm{L}} \cup \mathcal{V}^{\mathrm{L}} \times [0,1] \to \Delta^{\mathrm{L}}$ is redesigned to take both probability vectors and one-hot vectors as inputs. The model is trained by minimizing:

$$\mathrm{NELBO}\left(q, p_\theta; \mathbf{x}^{1:\mathrm{L}}\right)$$

$$= \mathbb{E}_{t\sim\mathcal{U}[\beta,\gamma],\tilde{q}_t} \sum_{l\in[\mathrm{L}]} f_{\mathrm{Duo}}\left(\mathbf{z}_t^l = \mathrm{argmax}\left(\mathbf{w}_t^l\right), \mathbf{x}_\theta\left(\left[\mathrm{softmax}\left(\mathbf{w}_t^{l'}/\tau\right)\right]_{l'=1}^{\mathrm{L}}, t\right), \alpha_t = \mathcal{T}(\tilde{\alpha}_t); \mathbf{x}^l\right).$$

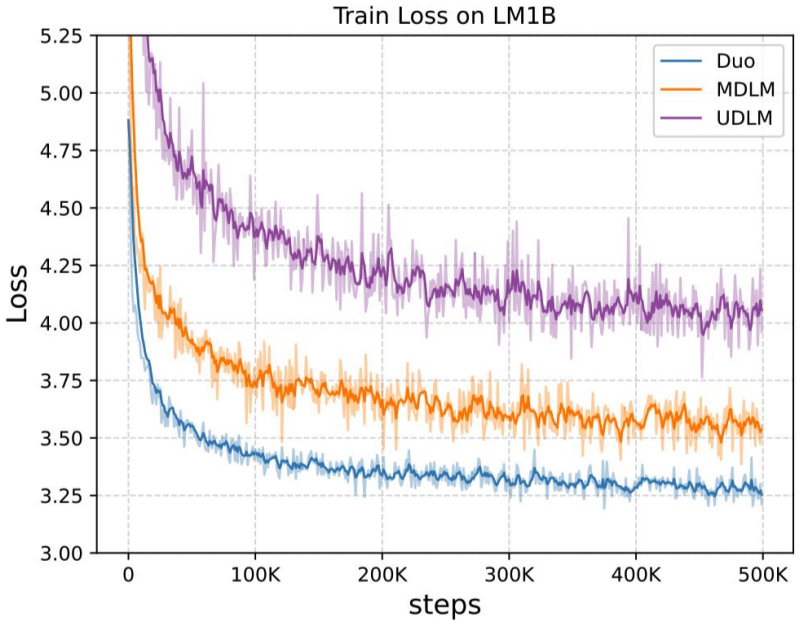
NOTE: This is not a valid NELBO except in the limiting case $\lim_{\tau\to 0^+} \mathbf{softmax}(\cdot/\tau)$ with $\beta = 0$ and $\gamma = 1$.

Faster Training using Curriculum Learning

For this, the denoising model $\mathbf{x}_\theta : \Delta^{\mathrm{L}} \cup \mathcal{V}^{\mathrm{L}} \times [0,1] \to \Delta^{\mathrm{L}}$ is redesigned to take both probability vectors and one-hot vectors as inputs. The model is trained by minimizing:

$$\mathrm{NELBO}\left(q, p_\theta; \mathbf{x}^{1:\mathrm{L}}\right)$$

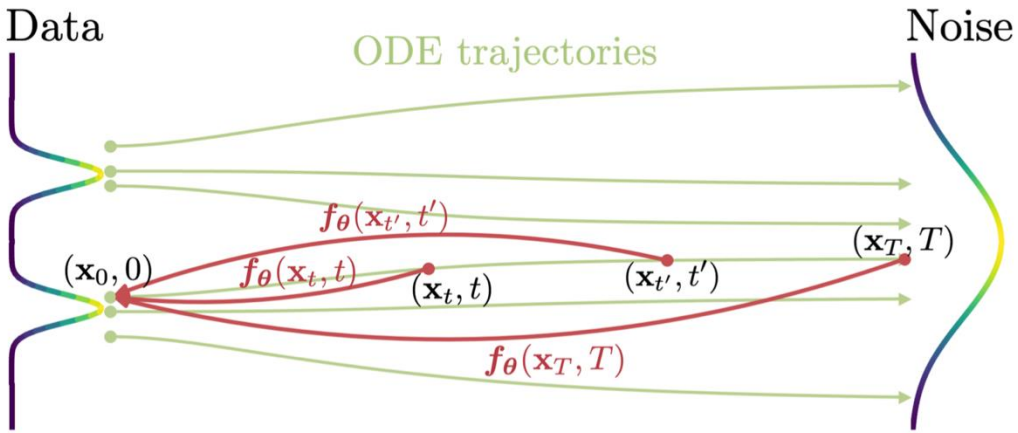
$$= \mathbb{E}_{t\sim\mathcal{U}[\beta,\gamma],\tilde{q}_t} \sum_{l\in[\mathrm{L}]} f_{\mathrm{Duo}}\left(\mathbf{z}_t^l = \mathrm{argmax}\left(\mathbf{w}_t^l\right), \mathbf{x}_\theta\left(\left[\mathrm{softmax}\left(\mathbf{w}_t^{l'}/\tau\right)\right]_{l'=1}^{\mathrm{L}}, t\right), \alpha_t = \mathcal{T}(\tilde{\alpha}_t); \mathbf{x}^l\right).$$



Better Convergence than UDLM!

Discrete Consistency Distillation (DCD)

On the other hand, continuous Gaussian diffusion enables the simulation of PF-ODE paths, which are non-trivial to define directly in discrete diffusion.



Consistency Models, Song *et al.*, ICML 2023

Discrete diffusion has PF-ODE path

↓ **Categorical Sampling!**

Continuous diffusion has PF-ODE path

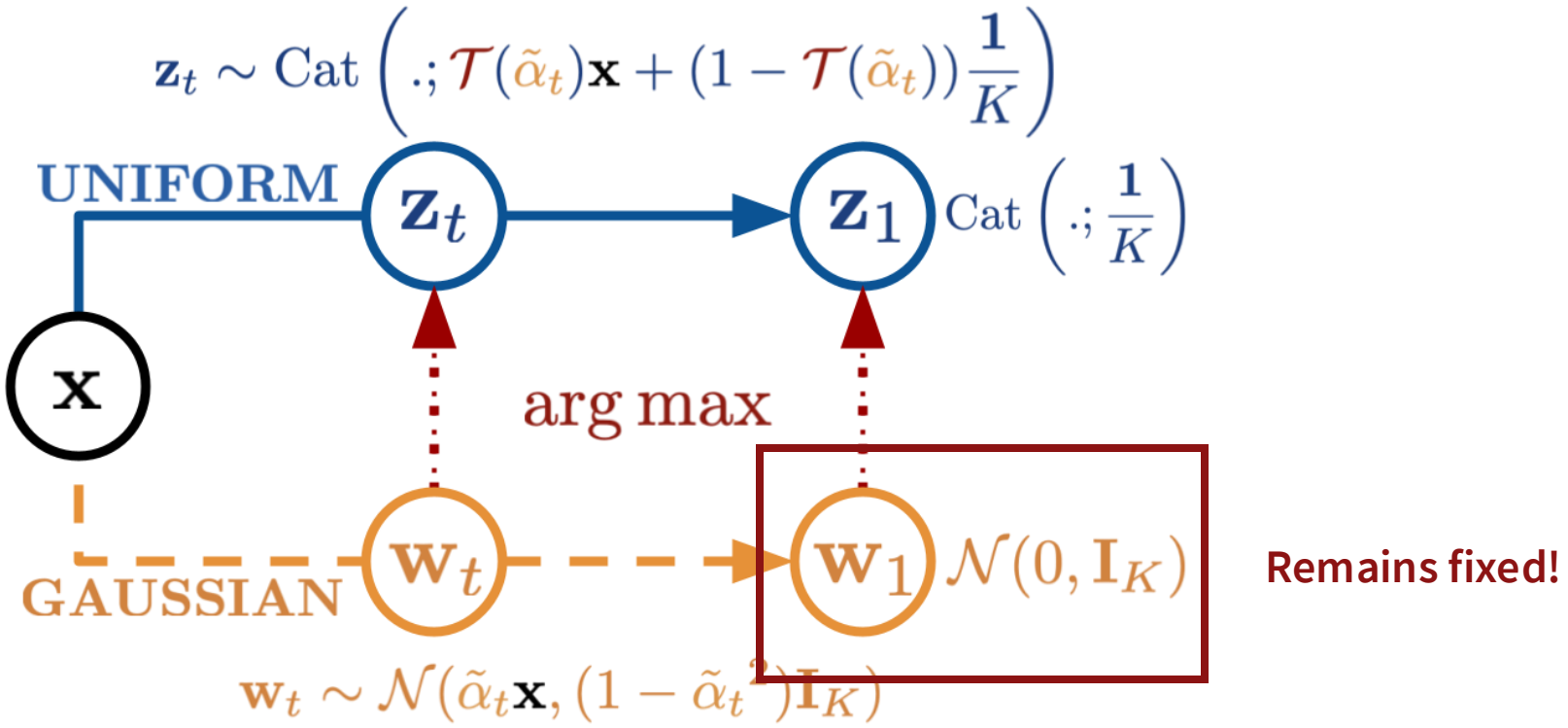
↓ **Categorical Sampling!**

Continuous diffusion has SDE path

Discrete Consistency Distillation (DCD)

To enable consistency distillation for discrete diffusion models, the authors take a detour:

Build a deterministic trajectory in Gaussian space and **map it to the discrete space**.



Remains fixed!

Discrete Consistency Distillation (DCD)

For a clean data $\mathbf{x}^{1:\mathrm{L}} \sim q_{\mathrm{data}}$ and Gaussian noise $\boldsymbol{\epsilon}^{1:\mathrm{L}} = \{\boldsymbol{\epsilon}^l \sim \mathcal{N}(0, \mathbf{I}_{\mathrm{K}}) | \forall l \in [\mathrm{L}]\}$, define

$$\mathcal{P}_{\mathrm{ODE}}\left(\mathbf{x}^{1:\mathrm{L}}, \boldsymbol{\epsilon}^{1:\mathrm{L}}\right) = \{\left[\tilde{\alpha}_t \mathbf{x}^l + \sqrt{1-\tilde{\alpha}_t^2}\boldsymbol{\epsilon}^l\right]_{l=1}^{\mathrm{L}}\}_{t\in[0,1]}$$

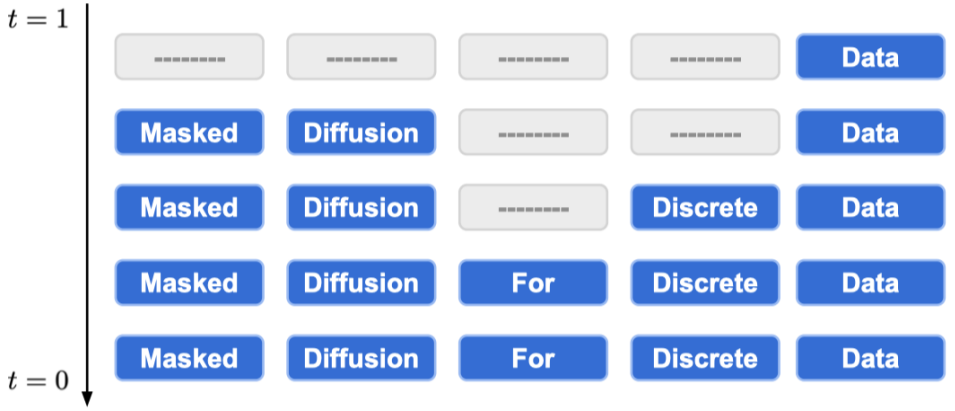
This trajectory is then projected to the discrete space via the argmax operator

$$\mathcal{P}_{\mathrm{DDT}}\left(\mathbf{x}^{1:\mathrm{L}}, \boldsymbol{\epsilon}^{1:\mathrm{L}}\right) = \{\left[\mathrm{argmax}(\tilde{\alpha}_t \mathbf{x}^l + \sqrt{1-\tilde{\alpha}_t^2}\boldsymbol{\epsilon}^l)\right]_{l=1}^{\mathrm{L}}\}_{t\in[0,1]},$$

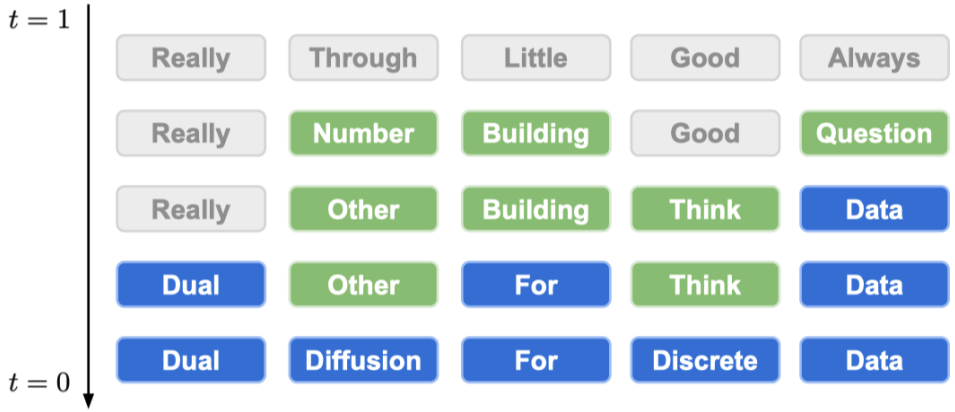
and serves as **a proxy for the PF-ODE in the discrete space.**

Discrete Consistency Distillation (DCD)

(a) Autoregressive Model



(b) Masked Diffusion

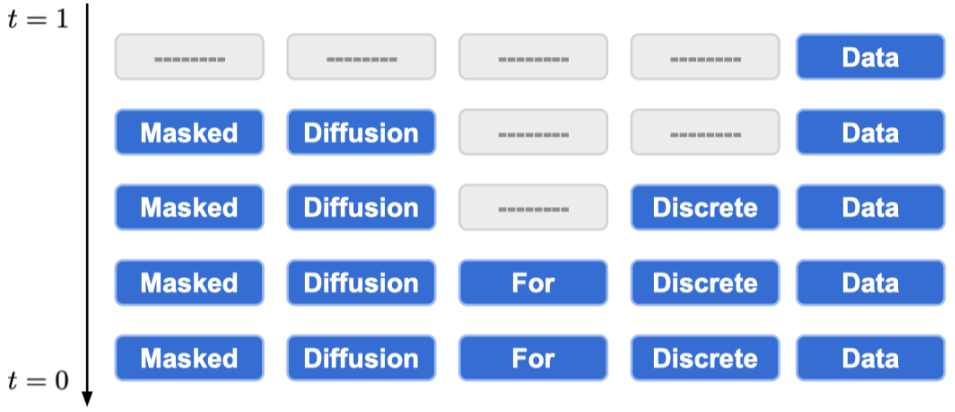


(c) Uniform-state Diffusion

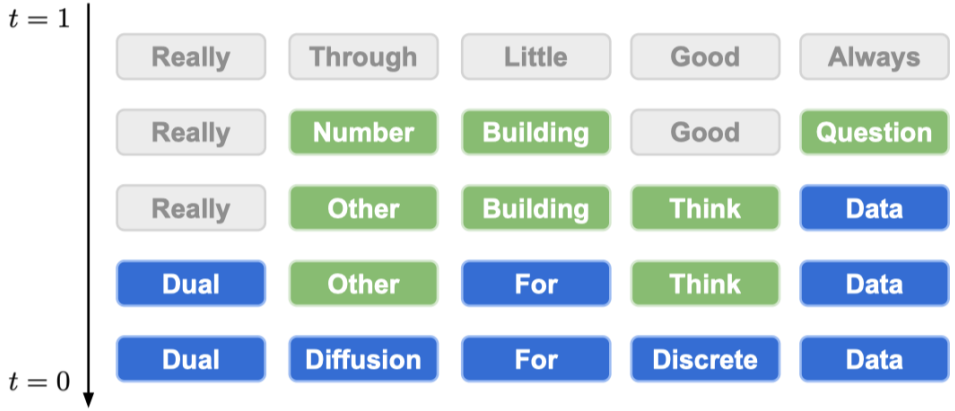
(d) $\mathcal{P}_{\text{DDT}}$

Discrete Consistency Distillation (DCD)

(a) Autoregressive Model



(b) Masked Diffusion



(c) Uniform-state Diffusion

(d) $\mathcal{P}_{\text{DDT}}$

Interpolates between 2 possibilities

Discrete Consistency Distillation (DCD)

We distill a teacher model $\mathbf{x}_{\theta^-}$ to a student model $\mathbf{x}_\theta$ that generates samples of similar quality while spending less steps. For this, we create a set of samples

$$\left(\mathbf{z}_s^{1:\mathrm{L}}, \mathbf{z}_t^{1:\mathrm{L}}\right) \sim \{(\mathbf{z}_{j-\delta}^{1:\mathrm{L}}, \mathbf{z}_j^{1:\mathrm{L}}) | \mathbf{z}_{\{\cdot\}}^{1:\mathrm{L}} \in \mathcal{P}_{\mathrm{DDT}}\left(\mathbf{x}^{1:\mathrm{L}}, \boldsymbol{\epsilon}^{1:\mathrm{L}}\right), j \in [\delta, 1]\}$$

for a given step size $\delta \in [0,1]$.

Discrete Consistency Distillation (DCD)

We distill a teacher model $\mathbf{x}_{\theta^-}$ to a student model $\mathbf{x}_\theta$ that generates samples of similar quality while spending less steps. For this, we create a set of samples

$$\left(\mathbf{z}_s^{1:\mathrm{L}}, \mathbf{z}_t^{1:\mathrm{L}}\right) \sim \{(\mathbf{z}_{j-\delta}^{1:\mathrm{L}}, \mathbf{z}_j^{1:\mathrm{L}}) | \mathbf{z}_{\{\cdot\}}^{1:\mathrm{L}} \in \mathcal{P}_{\mathrm{DDT}}\left(\mathbf{x}^{1:\mathrm{L}}, \boldsymbol{\epsilon}^{1:\mathrm{L}}\right), j \in [\delta, 1]\}$$

for a given step size $\delta \in [0,1]$.

With the pairs of noisy ($\mathbf{z}_t^{1:\mathrm{L}}$) and less noisy ($\mathbf{z}_s^{1:\mathrm{L}}$) samples, we train the student by optimizing

$$\mathcal{L}_{\mathrm{DCD}}(\theta; \theta^-) = D_{\mathrm{KL}}\left(\mathbf{x}_\theta\left(\mathbf{z}_t^{1:\mathrm{L}}, t\right), \mathbf{x}_{\theta^-}\left(\mathbf{z}_s^{1:\mathrm{L}}, s\right)\right).$$

Discrete Consistency Distillation (DCD)

Algorithm 1 Discrete Consistency Distillation (DCD)

Input: Dataset $\mathcal{D}$, learning rate η, number of distillation rounds N, number of training iterations per round M, ema μ, weights of the denoising model $\boldsymbol{\theta}$, weights of the EMA model $\boldsymbol{\theta}_{\text{ema}}$, discretization step δ.

for $i = 1$ **to** N **do**

  $\boldsymbol{\theta}^{-} \leftarrow \text{stopgrad}(\boldsymbol{\theta})$

  for $j = 1$ **to** M **do**

    Sample $\mathbf{x}^{1:L} \sim \mathcal{D}$, $t \sim \mathcal{U}[0,1]$, and $\boldsymbol{\epsilon}^{\ell} \sim \mathcal{N}(0, I_K)$.

    $s \leftarrow \max(t - \delta, 0)$

    $\mathbf{z}_s^{1:L} \leftarrow \left[\arg\max\left(\tilde{\alpha}_s \mathbf{x}^{\ell} + \sqrt{1 - \tilde{\alpha}_s{}^2}\boldsymbol{\epsilon}^{\ell}\right)\right]_{\ell=1}^{L}$

    $\mathbf{z}_t^{1:L} \leftarrow \left[\arg\max\left(\tilde{\alpha}_t \mathbf{x}^{\ell} + \sqrt{1 - \tilde{\alpha}_t{}^2}\boldsymbol{\epsilon}^{\ell}\right)\right]_{\ell=1}^{L}$

    $\mathcal{L}_{\text{DCD}}(\boldsymbol{\theta}; \boldsymbol{\theta}^{-}) \leftarrow \mathrm{D}_{\text{KL}}(\mathbf{x}_{\theta}(\mathbf{z}_t^{1:L}, t), \mathbf{x}_{\boldsymbol{\theta}^{-}}(\mathbf{z}_s^{1:L}, s))$

    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{DCD}}(\boldsymbol{\theta}; \boldsymbol{\theta}^{-})$

    $\boldsymbol{\theta}_{\text{ema}} \leftarrow \text{stopgrad}(\mu \boldsymbol{\theta}_{\text{ema}} + (1 - \mu)\boldsymbol{\theta})$

  end for

  $\delta \leftarrow 2 \cdot \delta$

end for

return $\boldsymbol{\theta}_{\text{ema}}$

Experiments

Setup

The authors compare the proposed method, coined **Duo**, on standard language modeling benchmarks:

- LM1B [Chelba *et al*., 2014]
- OpenWebText (OWT) [Gokaslan *et al*., 2019]

All methods, including **Duo**, share the DiT architecture [Peebles & Xie, 2023] containing 170M parameters for fair comparisons.

Improved Training

Summary

- The curriculum learning **accelerates training by 2x** and sets a new SotA for USDMs;
- **Duo** performs similarly to absorbing state models, surpassing ARMs on 3/7 zero-shot PPL benchmarks.

Baselines (U: Uniform State Models / A: Absorbing State Models)

1. (U) SEDD Uniform [Lou *et al.*, ICML 2024]
2. (U) UDLM [Schiff *et al.*, ICLR 2025]
3. (U) PLAID [Gulrajani & Hashimoto, arXiv 2023]
4. (A) MDLM [Sahoo *et al.*, NeurIPS 2024];
5. (A) SEDD Absorb [Lou *et al.*, ICML 2024];
6. (A) D3PM Absorb [Austin *et al.*, NeurIPS 2021];
7. AR.

Improved Training

With a properly chosen τ, the curriculum learning facilitates convergence and even improves the performance of the converged model compared to existing USDMs.

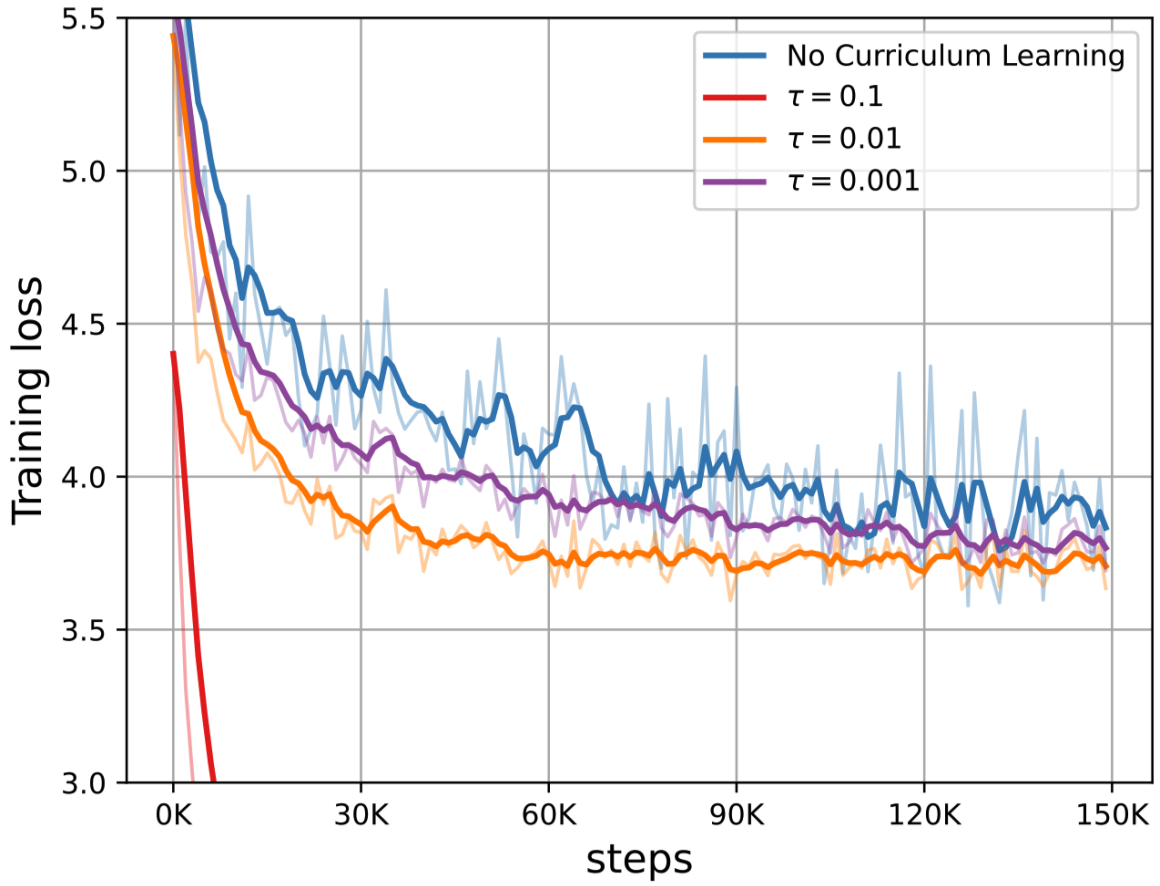


Table 1: Test perplexities (PPL; ↓) on LM1B. *Reported in He et al. (2022). Best uniform/Gaussian diffusion value is bolded. ¶Denotes the dataset didn't incorporate sentence packing. †Reported in Arriola et al. (2025). For diffusion models, we report the bound on the likelihood. Best diffusion value is underlined. ‡Denotes retrained models.

	LM1B¶	LM1B	OWT
Autoregressive			
Transformer‡	22.3	22.8†	17.5
Diffusion (absorbing state)			
BERT-Mouth* (Wang & Cho, 2019)	-	142.9	-
D3PM Absorb (Austin et al., 2021)	-	76.9	-
DiffusionBert (He et al., 2022)	-	63.8	-
SEDD Absorb‡ (Lou et al., 2023)	32.7	-	24.1
MDLM (Sahoo et al., 2024a)	27.0	31.8†	23.2
Diffusion (Uniform-state / Gaussian)			
D3PM Uniform (Austin et al., 2021)	-	137.9	-
Diffusion-LM* (Li et al., 2022)	-	118.6	-
SEDD Uniform (Lou et al., 2023)	40.3	-	29.7
UDLM‡ (Schiff et al., 2025)	31.3	36.7	27.4
Duo (Ours)	**29.9**	**33.7**	**25.2**

Improved Training

The evaluation of models trained on the OWT dataset on 7 other datasets shows the strong performance of **Duo**, even compared to ARMs in 3 datasets.

Table 2: Zero-shot perplexities (↓) of models trained for 1M steps on OWT. All perplexities for diffusion models are upper bounds. † Taken from Sahoo et al. (2024a). ¶ Taken from (Lou et al., 2023) models were trained for 1.3Msteps as opposed to the baselines that were trained for 1Msteps. All perplexities for diffusion models are upper bounds. Best uniform / Gaussian diffusion values are **bolded** and diffusion values better than AR are underlined. ‡ denotes retrained model.

	PTB	Wikitext	LM1B	Lambada	AG News	Pubmed	Arxiv
Autoregressive							
Transformer†	82.05	25.75	51.25	51.28	52.09	49.01	41.73
Diffusion (absorbing state)							
SEDD Absorb†	100.09	34.28	68.20	49.86	62.09	44.53	38.48
D3PM Absorb¶	200.82	50.86	138.92	93.47	-	-	-
MDLM†	95.26	32.83	67.01	47.52	61.15	41.89	37.37
Diffusion (Uniform-state / Gaussian)							
SEDD Uniform‡	105.51	41.10	82.62	57.29	82.64	55.89	50.86
Plaid¶	142.60	50.86	91.12	57.28	-	-	-
UDLM‡	112.82	39.42	77.59	53.57	80.96	50.98	44.08
Duo (Ours)	**89.35**	**33.57**	**73.86**	**49.78**	**67.81**	**44.48**	**40.39**

Improved Sampling

Summary

- **Duo** generates higher-quality samples than all previous diffusion models;
- Combining DCD with the Greedy-Tail sampler **reduces sampling steps by two orders of magnitude**;
- **Duo+DCD** outperforms a distilled MDLM model (esp. in low NFEs).

Setup

- **Duo** trained on the OWT dataset is distilled via DCD;
- MDLM is distilled with SDTT [Deschenaux & Gulcehre, ICLR 2025];
- The models are distilled over $\mathrm{N} = 5$ **distillation rounds**;
- The discretization step size δ begins with $1/512$ and is doubled every $M = 10\mathrm{K}$ steps;
- Sample quality is measured with **GPT-2 Generative Perplexity**;
- Sample diversity is measured with **average sequence entropy**.

Improved Sampling

Without DCD, **Duo** achieves lower generative perplexities compared to both uniform and absorbing state diffusion models.

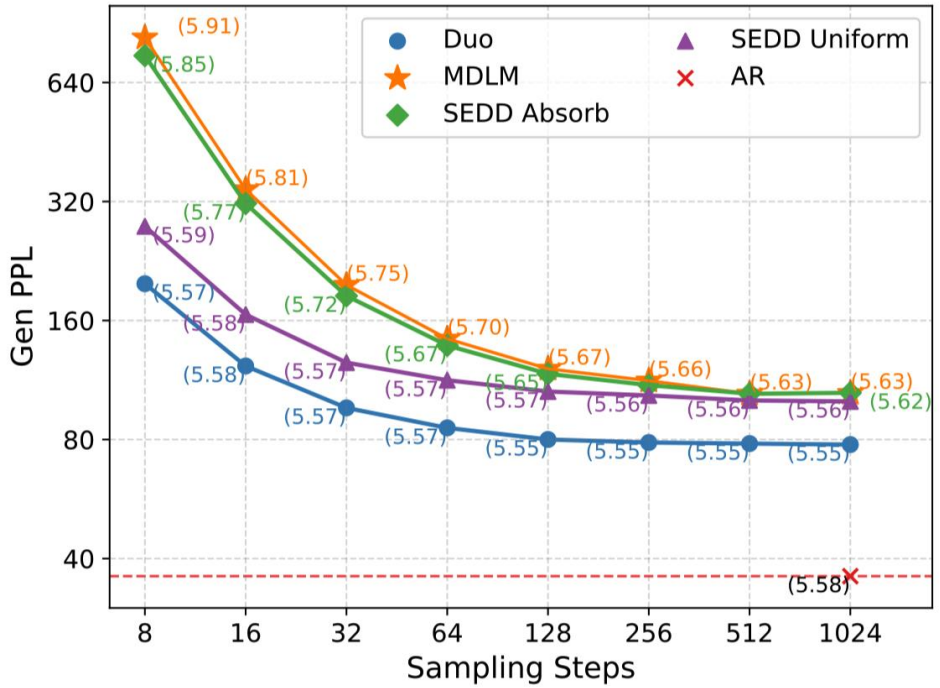


Figure 9: Sample quality comparison using Gen PPL (↓) between Duo (ours), MDLM, SEDD (Absorb / Uniform), and AR. Values in brackets indicate sample entropy (↑). Among USDMs, Duo achieves lower Gen PPL than SEDD-Uniform, indicating higher sample quality. Compared to MDMs, Duo yields lower Gen PPL with a slight trade-off in entropy. Exact quantitative numbers for Gen PPL can be found in Table 4.

Improved Sampling

Using the Greedy-Tail sampler, **Duo+DCD** performs on par with MDLM+SDTT at high NFEs and outperforms it at lower NFEs (but at the cost of sample diversity).

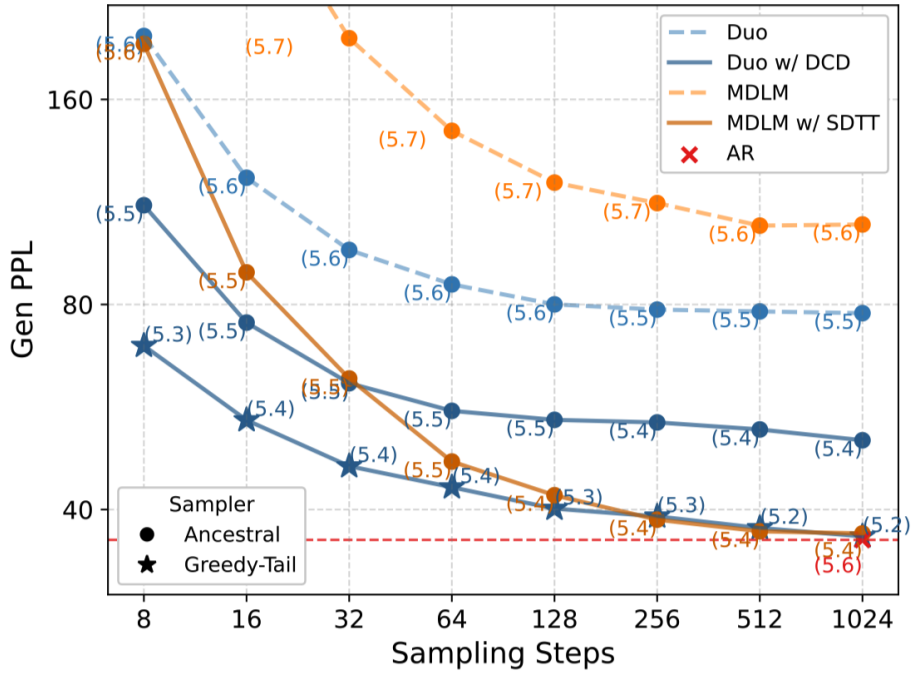


Figure 3: Sample quality comparison of Duo vs. MDLM. Duo outperforms MDLM in Gen PPL (↓) for base models and in low-NFE regime after 5 distillation rounds.

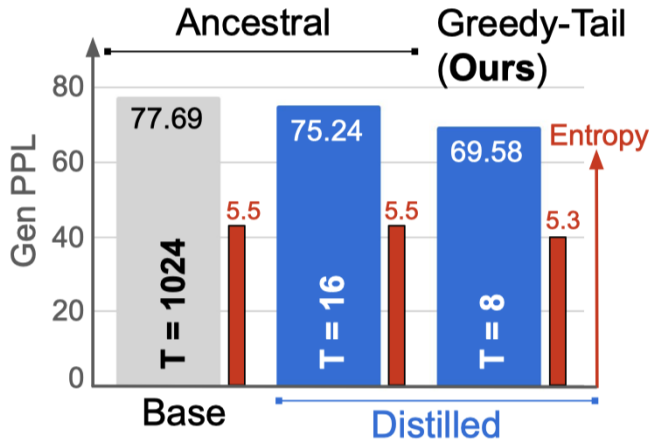


Figure 4: Sample quality comparison between the base Duo model and Duo distilled for 5 rounds using our DCD algorithm. The distilled model matches the base model's sample quality in just 16 steps (vs. 1024) with ancestral sampling. With our Greedy-Tail sampler, sampling steps can be further reduced to 8, achieving slightly better Gen PPL and lower entropy.

Conclusion

This paper

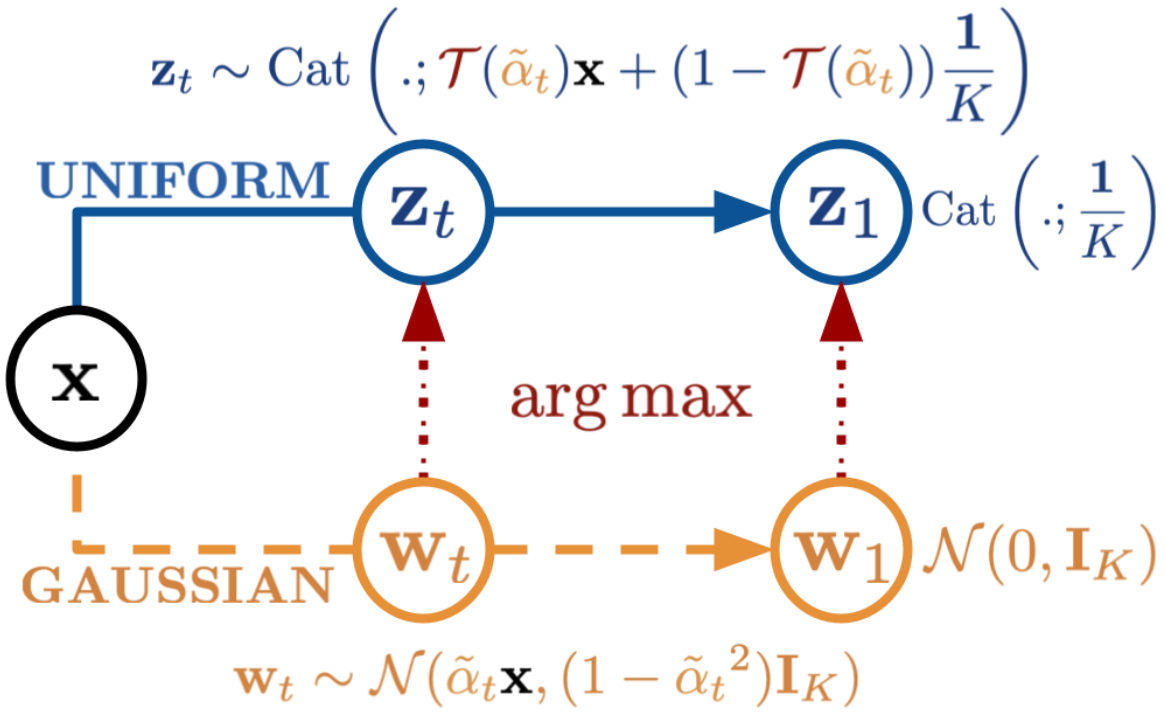
- Establishes **a theoretical connection** between continuous-state Gaussian diffusion models and discrete-space Uniform-state diffusion models;
- Leverages the duality to design **a curriculum learning strategy** that **doubles the training speed**;
- Enables **consistency distillation**, yielding up to a **two-order-of-magnitude speedup** in sampling;
- Demonstrates that **USDMs can surpass MDMs in low NFE regimes**, thanks to their self-correcting properties.

Food for Thought

- The discovered duality is about forward processes. **Can we establish a similar connection for reverse processes**, as well as continuous-state discrete diffusion models?
- While the authors claim that USDMs outperform MDMs in low NFE regimes, **correction techniques (*e.g.*, predictor-corrector, DDPD) are not considered in evaluations**.
- With the recent surge of KV caching techniques for MDMs (e.g., Eso-LMs) that greatly accelerate inference, is it still worthwhile to explore USDMs?

Thank You

The Diffusion Duality



Seungwoo Yoo

KAIST

https://dvelopery0115.github.io