Discrete Markov Bridge: Bridging Variational Inference and Discrete Diffusion

Hengli Li, Yuxuan Wang, Song-Chun Zhu, Ying Nian Wu, Zilong Zheng

Peking University, Beijing Institute for General AI, UCLA

June 13, 2025

Outline

Motivation

Generative modeling seeks to estimate an underlying data distribution μ and generate new samples from it.

Diffusion models have shown strong performance for both continuous and discrete data, but:

- Discrete diffusion models typically use fixed-rate transition matrices (e.g., uniform, absorbing).
- Fixed matrices limit the expressiveness of latent representations and the design space.
- Variational methods (e.g., VAEs) are highly expressive due to adaptive latent encodings.

Goal: Combine the flexibility of variational inference with the tractability of discrete diffusion.

Limitations of Prior Discrete Diffusion

- Existing discrete diffusion models use non-learnable, static transition matrices.
- This restricts the model's ability to adapt to complex data.
- Only simple matrices (Absorb, Uniform) are computationally feasible.
- No adaptive latent variable structure as in VAEs.

Key Question: Can we design a discrete diffusion model with a learnable, adaptive transition matrix and a variational structure?

Overview of Discrete Markov Bridge (DMB)

- DMB is a two-stage, bidirectional learning framework:
  1. **Matrix-learning (Forward Bridge):** Learn a parameterized rate transition matrix to map data to a latent (prior) distribution.
  2. **Score-learning (Backward Bridge):** Learn a neural network to estimate probability ratios (concrete scores) for reverse generation.
- Integrates variational inference (ELBO) with discrete diffusion.
- Enables richer latent representations and improved modeling flexibility.

Diagram:

$$\mu \xrightarrow{\text{Matrix-learning}} p_T \xrightarrow{\text{Score-learning}} \hat{\mu}$$

Continuous-Time Discrete Markov Chain

State space: $X = \{1, 2, \ldots, n\}$
Markov process: $\{X(t) : t \in \mathbb{R}, X(t) \in X\}$
Transition probability: $p_{t+s|t}(y|x) = \mathbb{P}(X(t+s) = y|X(t) = x)$
Marginal: $p_t(x) = \mathbb{P}(X(t) = x)$
Rate matrix:

$$q_t(x, y) = \frac{d}{ds} p_{t+s|t}(y|x)\Big|_{s=0}$$

Forward Kolmogorov equation:

$$\frac{d}{dt} p_t = p_t Q(t)$$

where $Q(t)$ is a rate matrix: rows sum to zero, off-diagonals non-negative.

Reversibility and Backward Process

Reversibility theorem:
Given the forward equation $\frac{d}{dt}p_t = p_t Q(t)$, there exists a reverse process:

$$\frac{d}{dt}p_{T-t} = p_{T-t}\hat{Q}(T-t)$$

where

$$\hat{Q}(t)_{x,y} = \frac{p_t(y)}{p_t(x)}Q(t)_{y,x}$$

Implication: If we can estimate the probability ratios $\frac{p_t(y)}{p_t(x)}$, we can construct the reverse process and sample from the data distribution.

Matrix-learning: Adaptive Transition Matrix

Key innovation: Learn a parameterized, diagonalizable rate matrix Q_α with parameters α.

- Q_α must have non-negative off-diagonals, rows sum to zero.
- Diagonalizability enables efficient computation of matrix exponentials.
- The forward process is:

$$\frac{d}{dt}p_t = p_t Q_\alpha(t), \quad p_0 \approx \mu$$

- The terminal distribution p_T serves as the learned prior.

Matrix-learning Objective

Objective: Minimize the KL divergence between the forward conditional and the terminal prior:

$$J_Q = \mathbb{E}_\mu \left[\mathrm{KL}(p_{T|0;\alpha} \| p_{T;\alpha}) \right]$$

where $p_{T|0;\alpha}$ is the conditional at time T given data, and $p_{T;\alpha}$ is the marginal at T. **Example:** For $X = \{A, B, C\}$, Q_α can be learned so that

the process at T yields a desired prior (e.g., uniform or learned).

Diagonalizability and Computational Efficiency

- Diagonalizable Q_α allows efficient computation of $\exp(\sigma Q_\alpha)$ for transition probabilities.
- Structured or sparse parameterizations reduce memory and computation cost.
- Enables scaling to large state spaces (e.g., text sequences).

Score-learning: Neural Score Estimation

Goal: Estimate the ratio

$$s_\theta(x, t, y) \approx \frac{p_{t|0}(y|x_0)}{p_{t|0}(x|x_0)}$$

for all $x, y \in X$.

- s_θ is parameterized by a neural network.
- Used to construct the reverse rate matrix:

$$\hat{Q}(t)_{x,y} = s_\theta(x, t, y) Q_\alpha(t)_{y,x}$$

- Enables the backward process to reconstruct the data distribution from the latent.

Loss Function and Training

Variational objective: Continuous-time Evidence Lower Bound (ELBO):

$$\mathrm{ELBO} = \mathbb{E}_{x_0 \sim \mu} \left[\mathbb{E}_{x_T \sim p_{T|0;\alpha}(\cdot|x_0)} \log p_{0|T;\theta}(x_0|x_T) - \mathrm{KL}(p_{T|0;\alpha}(\cdot|x_0) \| p_T) \right]$$

- $p_{0|T;\theta}$: backward process conditional (decoding).
- $p_{T|0;\alpha}$: forward process conditional (encoding).
- Jointly optimize α (transition matrix) and θ (score network) to maximize ELBO.

Theoretical Guarantees

- DMB provides formal guarantees for reversibility and convergence of the learning process.
- Under mild conditions, the learned model converges to the true data distribution as model capacity and data increase.
- The ELBO structure ensures principled likelihood-based training and evaluation.

Efficient Computation and Space Complexity

- Diagonalizable Q_α allows fast computation of matrix exponentials.
- Structured/sparse parameterizations (e.g., block-diagonal, low-rank) reduce memory cost.
- DMB scales to large vocabularies and sequence lengths, unlike prior models with dense $n \times n$ matrices.

Text Modeling: Text8 Dataset

- DMB achieves an ELBO of 1.38 on Text8, outperforming SEDD and other baselines.
- Demonstrates effective modeling of complex natural language distributions.

Image Modeling: CIFAR-10

- DMB integrated with VQ-VAE achieves results comparable to DDPM on CIFAR-10.
- Shows generality across data modalities (text, images).

Ablation and Analysis

- Both Matrix-learning and Score-learning are essential for optimal performance.
- Learned transition matrices adapt to data structure.
- Score network accurately reconstructs data from latent.

Advantages and Limitations

Advantages:

- Learnable transition matrix enables richer latent representations.
- Variational structure provides principled likelihood estimation.
- Scalable to large state spaces.

Limitations:

- Computational overhead for very large vocabularies.
- Further exploration needed for other data types (e.g., graphs).

Conclusion

- Discrete Markov Bridge (DMB) unifies variational inference and discrete diffusion.
- Introduces learnable, adaptive transition matrices for flexible modeling.
- Achieves state-of-the-art results on text and image tasks.
- Opens new avenues for discrete generative modeling research.

Code: `https://github.com/Henry839/Discrete-Markov-Bridge`

Notation Summary

- $X = \{1, \ldots, n\}$: Discrete state space
- μ: Data distribution
- p_t: Distribution at time t
- Q_α: Learnable rate matrix
- $s_\theta(x, t, y)$: Neural score estimator
- $p_{T|0;\alpha}$: Forward conditional
- $p_{0|T;\theta}$: Backward conditional
- ELBO: Evidence Lower Bound

Example Computation

Suppose: $X = \{A, B\}$, $\mu = (0.7, 0.3)$, and

$$Q_\alpha = \begin{bmatrix} -0.5 & 0.5 \\ 0.4 & -0.4 \end{bmatrix}$$

At time t, $p_t = (0.6, 0.4)$.

Reverse matrix:

$$\hat{Q}(t)_{A,B} = \frac{0.4}{0.6} \times 0.4 \approx 0.267$$

$$\hat{Q}(t)_{B,A} = \frac{0.6}{0.4} \times 0.5 = 0.75$$

The neural network is trained to estimate these ratios for all x, y.

References

Discrete Markov Bridge, Hengli Li et al., arXiv:2505.19752 (2025)
https://arxiv.org/abs/2505.19752