Deep Generative Models

Lecture 14

Roman Isachenko

Moscow Institute of Physics and Technology
Yandex School of Data Analysis

2025, Autumn

Recap of Previous Lecture

Outline

Discrete or Continuous Diffusion Models?

Reminder: Diffusion models define a forward corruption process and a reverse denoising process. Previously, we studied diffusion models with continuous states $\mathbf{x}(t) \in \mathbb{R}^m$.

Continuous state space

- **Discrete time** $t \in \{0, 1, \ldots, T\} \;\Rightarrow\;$ **DDPM / NCSN**.
- **Continuous time** $t \in [0, 1] \;\Rightarrow\;$ **Score-based SDE models**.

Discrete or Continuous Diffusion Models?

Reminder: Diffusion models define a forward corruption process and a reverse denoising process. Previously, we studied diffusion models with continuous states $\mathbf{x}(t) \in \mathbb{R}^m$.

Continuous state space

- **Discrete time** $t \in \{0, 1, \ldots, T\}$ $\Rightarrow$ **DDPM / NCSN**.
- **Continuous time** $t \in [0, 1]$ $\Rightarrow$ **Score-based SDE models**.

Now we turn to diffusion over discrete-value states $\mathbf{x}(t) \in \{1, \ldots, K\}^m$.

Discrete state space

- **Discrete time** $t \in \{0, 1, \ldots, T\}$.
- **Continuous time** $t \in [0, 1]$.

Let's discuss why we need discrete diffusion models.

Why Discrete Diffusion Models?

While autoregressive (AR) models dominate discrete-data domains (e.g., text or sequences), they have fundamental limitations.

https://aaronlou.com/blog/2024/discrete-diffusion/

Why Discrete Diffusion Models?

While autoregressive (AR) models dominate discrete-data domains (e.g., text or sequences), they have fundamental limitations.

Key advantages of discrete diffusion

- **Parallel generation:** diffusion enables sampling all tokens simultaneously, unlike AR's strictly left-to-right process.

https://aaronlou.com/blog/2024/discrete-diffusion/

Why Discrete Diffusion Models?

While autoregressive (AR) models dominate discrete-data domains (e.g., text or sequences), they have fundamental limitations.

Key advantages of discrete diffusion

- **Parallel generation:** diffusion enables sampling all tokens simultaneously, unlike AR's strictly left-to-right process.
- **Flexible infilling:**. diffusion can mask arbitrary parts of a sequence and reconstruct them, rather than generating only from prefix to suffix.

https://aaronlou.com/blog/2024/discrete-diffusion/

Why Discrete Diffusion Models?

While autoregressive (AR) models dominate discrete-data domains (e.g., text or sequences), they have fundamental limitations.

Key advantages of discrete diffusion

- **Parallel generation:** diffusion enables sampling all tokens simultaneously, unlike AR's strictly left-to-right process.
- **Flexible infilling:**. diffusion can mask arbitrary parts of a sequence and reconstruct them, rather than generating only from prefix to suffix.
- **Robustness:** diffusion avoids the "exposure bias" caused by teacher forcing in AR training.

https://aaronlou.com/blog/2024/discrete-diffusion/

Why Discrete Diffusion Models?

While autoregressive (AR) models dominate discrete-data domains (e.g., text or sequences), they have fundamental limitations.

Key advantages of discrete diffusion

- **Parallel generation:** diffusion enables sampling all tokens simultaneously, unlike AR's strictly left-to-right process.
- **Flexible infilling:**. diffusion can mask arbitrary parts of a sequence and reconstruct them, rather than generating only from prefix to suffix.
- **Robustness:** diffusion avoids the "exposure bias" caused by teacher forcing in AR training.
- **Unified framework:** diffusion generalizes naturally to discrete domains that do not suit continuous Gaussian noise.

https://aaronlou.com/blog/2024/discrete-diffusion/

Forward Discrete Process

Continuous Diffusion Markov Chain

In continuous diffusion, the forward Markov chain is defined by progressively corrupting data with Gaussian noise:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}).$$

Austin J. et al. Structured denoising diffusion models in discrete state-spaces, 2021.

Forward Discrete Process

Continuous Diffusion Markov Chain

In continuous diffusion, the forward Markov chain is defined by progressively corrupting data with Gaussian noise:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}).$$

Discrete Diffusion Markov Chain

For discrete data, we instead define a Markov chain over categorical states:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathrm{Categorical}(\mathbf{Q}_t \mathbf{x}_{t-1}),$$

where $\mathbf{Q}_t \in [0,1]^{K\times K}$ is a **transition matrix** where each column gives transition probabilities from one state to all others, and columns sum to 1:

$$[\mathbf{Q}_t]_{ij} = q(x_t = i | x_{t-1} = j), \qquad \sum_{i=1}^{K} [\mathbf{Q}_t]_{ij} = 1.$$

Austin J. et al. Structured denoising diffusion models in discrete state-spaces, 2021.

Forward Process over Time

- The forward diffusion gradually destroys information through repeated random transitions.

Forward Process over Time

- The forward diffusion gradually destroys information through repeated random transitions.
- Applying transitions t times yields a marginal distribution:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathrm{Categorical}(\mathbf{Q}_{1:t}\mathbf{x}_0), \qquad \mathbf{Q}_{1:t} = \mathbf{Q}_t\mathbf{Q}_{t-1}\cdots\mathbf{Q}_1.$$

Forward Process over Time

- The forward diffusion gradually destroys information through repeated random transitions.
- Applying transitions t times yields a marginal distribution:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathrm{Categorical}(\mathbf{Q}_{1:t}\mathbf{x}_0), \qquad \mathbf{Q}_{1:t} = \mathbf{Q}_t\mathbf{Q}_{t-1}\cdots\mathbf{Q}_1.$$

- This process drives the data towards a stationary distribution as $t \to T$.

Forward Process over Time

- The forward diffusion gradually destroys information through repeated random transitions.
- Applying transitions t times yields a marginal distribution:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathrm{Categorical}(\mathbf{Q}_{1:t}\mathbf{x}_0), \qquad \mathbf{Q}_{1:t} = \mathbf{Q}_t\mathbf{Q}_{t-1}\cdots\mathbf{Q}_1.$$

- This process drives the data towards a stationary distribution as $t \to T$.
- To achieve this behavior, we design the transition matrices $\mathbf{Q}_t$ appropriately.

Designing the Transition Matrix Q_t

- The choice of $\mathbf{Q}_t$ determines how information is erased and what the stationary distribution becomes.

Austin J. et al. Structured denoising diffusion models in discrete state-spaces, 2021.

Designing the Transition Matrix Q_t

- The choice of $\mathbf{Q}_t$ determines how information is erased and what the stationary distribution becomes.
- **Uniform diffusion**

$$\mathbf{Q}_t = (1 - \beta_t)\mathbf{I} + \beta_t \mathbf{1}\mathbf{1}^\top.$$

  Each token is replaced by a uniformly random symbol with probability β_t. The stationary distribution is uniform noise.

Austin J. et al. Structured denoising diffusion models in discrete state-spaces, 2021.

Designing the Transition Matrix Q_t

- The choice of $\mathbf{Q}_t$ determines how information is erased and what the stationary distribution becomes.
- **Uniform diffusion**

$$\mathbf{Q}_t = (1 - \beta_t)\mathbf{I} + \beta_t \mathbf{1}\mathbf{1}^\top.$$

  Each token is replaced by a uniformly random symbol with probability β_t. The stationary distribution is uniform noise.
- **Absorbing diffusion**

$$\mathbf{Q}_t = (1 - \beta_t)\mathbf{I} + \beta_t \, \mathbf{e}_m \mathbf{1}^\top.$$

  Tokens are gradually replaced by a special mask m; the stationary distribution is fully masked.

Austin J. et al. Structured denoising diffusion models in discrete state-spaces, 2021.

NOT READY

Summary

-