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

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.

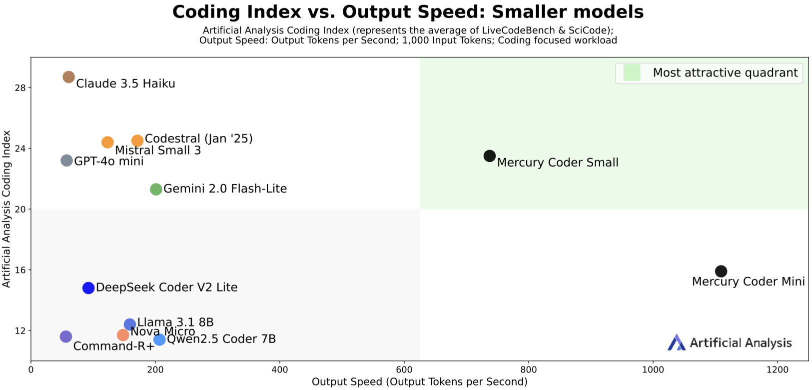
Key advantages of discrete diffusion

- **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/

2025 – Big Bang of Discrete Diffusion Models



Khanna S. et al. Mercury: Ultra-fast language models based on diffusion, 2025.

Outline

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:

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

where each $\mathbf{x}_t \in \{0,1\}^K$ is a **one-hot vector** encoding the categorical state.

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:

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

where each $\mathbf{x}_t \in \{0,1\}^K$ is a **one-hot vector** encoding the categorical state.

What is the transition matrix $\mathbf{Q}_t$?

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

Forward Process over Time

Transition Matrix

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

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

Forward Process over Time

Transition Matrix

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

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

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

Forward Process over Time

Transition Matrix

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

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

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

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

Forward Process over Time

Transition Matrix

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

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

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

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

- As $t \to T$, the process drives the data toward a stationary distribution.

Forward Process over Time

Transition Matrix

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

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

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

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

- As $t \to T$, the process drives the data toward a stationary distribution.
- We design the transition matrices $\mathbf{Q}_t$ to achieve this behavior.

Transition Matrix

- 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.

Transition Matrix

- The choice of $\mathbf{Q}_t$ determines how information is erased and what the stationary distribution becomes.
- $\mathbf{Q}_t$ and $\mathbf{Q}_{1:t}$ should be easy to compute for each t.
- **Uniform diffusion**

$$\mathbf{Q}_t = (1 - \beta_t)\mathbf{I} + \beta_t \mathbf{U}, \qquad \mathbf{U}_{ij} = \frac{1}{K}.$$

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.

Transition Matrix

- The choice of $\mathbf{Q}_t$ determines how information is erased and what the stationary distribution becomes.
- $\mathbf{Q}_t$ and $\mathbf{Q}_{1:t}$ should be easy to compute for each t.
- **Uniform diffusion**

  $$\mathbf{Q}_t = (1 - \beta_t)\mathbf{I} + \beta_t \mathbf{U}, \qquad \mathbf{U}_{ij} = \frac{1}{K}.$$

  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.

Transition Matrix

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

Transition Matrix

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

Uniform Diffusion

$$\mathbf{Q}_t = (1-\beta_t)\mathbf{I} + \beta_t\mathbf{U}, \qquad \mathbf{U}_{ij} = \frac{1}{K}.$$

$$\mathbf{Q}_{1:t} = \bar{\alpha}_t\mathbf{I} + (1-\bar{\alpha}_t)\mathbf{U}, \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s).$$

Transition Matrix

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

Uniform Diffusion

$$\mathbf{Q}_t = (1-\beta_t)\mathbf{I} + \beta_t\mathbf{U}, \qquad \mathbf{U}_{ij} = \tfrac{1}{K}.$$

$$\mathbf{Q}_{1:t} = \bar{\alpha}_t\mathbf{I} + (1-\bar{\alpha}_t)\mathbf{U}, \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s).$$

- Each token retains its original value with prob. $\bar{\alpha}_t$.
- It becomes uniformly random with prob. $(1-\bar{\alpha}_t)$.
- As $t \to T$, the process converges to the stationary uniform distribution.

Transition Matrix

Absorbing Diffusion

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

$$\mathbf{Q}_{1:t} = \bar{\alpha}_t\,\mathbf{I} + (1-\bar{\alpha}_t)\,\mathbf{e}_m\mathbf{1}^\top, \qquad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s).$$

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

Transition Matrix

Absorbing Diffusion

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

$$\mathbf{Q}_{1:t} = \bar{\alpha}_t\, \mathbf{I} + (1-\bar{\alpha}_t)\, \mathbf{e}_m \mathbf{1}^\top, \qquad \bar{\alpha}_t = \prod_{s=1}^{t} (1-\beta_s).$$

- Each token retains its original value with prob. $\bar{\alpha}_t$.
- It becomes $\mathbf{e}_m$ with prob. $(1-\bar{\alpha}_t)$.
- As $t \to T$, all tokens converge to the mask state: $q(\mathbf{x}_T) \approx \mathrm{Cat}(\mathbf{e}_m)$.
- This makes the process analogous to **masked language modeling**.

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

Uniform vs. Absorbing Transition Matrix

Aspect	**Uniform Diffusion**	**Absorbing Diffusion**
$\mathbf{Q}_t$	$(1-\beta_t)\mathbf{I} + \beta_t\mathbf{U}$	$(1-\beta_t)\mathbf{I} + \beta_t\mathbf{e}_m\mathbf{1}^\top$
$\mathbf{Q}_{1:t}$	$\bar{\alpha}_t\mathbf{I} + (1-\bar{\alpha}_t)\mathbf{U}$	$\bar{\alpha}_t\mathbf{I} + (1-\bar{\alpha}_t)\mathbf{e}_m\mathbf{1}^\top$
$\mathbf{Q}_{1:\infty}$	$\mathbf{U}$	$\mathrm{Cat}(\mathbf{e}_m)$
Interpretation	Random replacement	Gradual masking of tokens
Application	Image / symbol diffusion	Text diffusion $\approx$ Masked LM

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

Uniform vs. Absorbing Transition Matrix

Aspect	Uniform Diffusion	Absorbing Diffusion
$\mathbf{Q}_t$	$(1-\beta_t)\mathbf{I}+\beta_t\mathbf{U}$	$(1-\beta_t)\mathbf{I}+\beta_t\mathbf{e}_m\mathbf{1}^\top$
$\mathbf{Q}_{1:t}$	$\bar{\alpha}_t\mathbf{I}+(1-\bar{\alpha}_t)\mathbf{U}$	$\bar{\alpha}_t\mathbf{I}+(1-\bar{\alpha}_t)\mathbf{e}_m\mathbf{1}^\top$
$\mathbf{Q}_{1:\infty}$	$\mathbf{U}$	$\mathrm{Cat}(\mathbf{e}_m)$
Interpretation	Random replacement	Gradual masking of tokens
Application	Image / symbol diffusion	Text diffusion $\approx$ Masked LM

Observation

Both schemes gradually destroy information, but differ in their stationary limit. Absorbing diffusion bridges diffusion and masked-language-model objectives.

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

NOT READY

Reverse Process and Model Parameterization

Goal

Learn a reverse model that reconstructs cleaner data from corrupted inputs:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \approx q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0).$$

- The reverse chain defines the generative process:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t).$$

- We parameterize $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ as a factorized categorical distribution:

$$p_\theta(x_{t-1,i}|x_t) = \mathrm{Cat}(x_{t-1,i}; \pi_\theta(x_t, i, t)),$$

where π_θ are model logits over K symbols.

Austin J. et al., 2021.

Variational Objective (Discrete ELBO)

Evidence Lower Bound

$$\log p_\theta(\mathbf{x}_0) \geq \mathbb{E}_q\Big[\sum_{t=1}^{T} -D_{\mathrm{KL}}\big(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\big)\Big].$$

For categorical transitions, the KL becomes a cross-entropy term:

$$\mathcal{L}_t = \mathbb{E}_{\mathbf{x}_0,t}[-\log p_\theta(x_{t-1}=x_0 \mid \mathbf{x}_t, t)].$$

- Equivalent to predicting the clean token x_0 from a partially noised $\mathbf{x}_t$.
- In practice, the model learns to *denoise* corrupted inputs at multiple noise levels.

Austin J. et al., 2021.

Relation to Masked Language Modeling (MLM)

- In absorbing diffusion, corrupted tokens are replaced by a mask *m*.
- The denoising task becomes identical to predicting masked tokens:

$$\mathcal{L} = \mathbb{E}_{t \sim p(t)} \mathbb{E}_{\mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} [-\log p_\theta(\mathbf{x}_0 | \mathbf{x}_t, t)].$$

- Therefore, discrete diffusion can be seen as a **mixture of MLM objectives** with varying masking rates.
- This view directly connects diffusion LMs to BERT-style training, but provides a principled probabilistic framework.

Austin J. et al., 2021.

Posterior Distribution $q(x_{t-1} \mid x_t, x_0)$

- To train the reverse model, we need the posterior distribution

$$q(x_{t-1} \mid x_t, x_0).$$

- From Bayes' rule (Eq. 3 in the paper):

$$q(x_{t-1} \mid x_t, x_0) = \frac{q(x_t \mid x_{t-1})\, q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)}.$$

- All three terms are tractable:

$$\begin{aligned} q(x_t \mid x_{t-1}) &= \mathrm{Cat}(Q_t[x_{t-1}, :]), \\ q(x_{t-1} \mid x_0) &= \mathrm{Cat}(Q_{1:(t-1)}[x_0, :]), \\ q(x_t \mid x_0) &= \mathrm{Cat}(Q_{1:t}[x_0, :]). \end{aligned}$$

- This expression defines a categorical distribution over x_{t-1}, enabling exact computation of the discrete ELBO.

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

Posterior Distribution $q(x_{t-1} \mid x_t, x_0)$

Exact posterior (Eq. 3 in the paper, TRANSPOSE)

$$q(x_{t-1} \mid x_t, x_0) = \mathrm{Cat}\left(x_{t-1};\ \mathbf{p} = \frac{x_t \mathbf{Q}_t^\top \odot x_0\, \bar{\mathbf{Q}}_{t-1}}{x_0\, \bar{\mathbf{Q}}_t\, x_t^\top}\right).$$

Austin J. et al., Structured Denoising Diffusion Models in Discrete State-Spaces, 2021.

Posterior Distribution $q(x_{t-1} \mid x_t, x_0)$

Exact posterior (Eq. 3 in the paper, TRANSPOSE)

$$q(x_{t-1} \mid x_t, x_0) = \mathrm{Cat}\left(x_{t-1};\ \mathbf{p} = \frac{x_t \mathbf{Q}_t^\top \odot x_0\, \bar{\mathbf{Q}}_{t-1}}{x_0\, \bar{\mathbf{Q}}_t\, x_t^\top}\right).$$

- $\mathbf{Q}_t$ — forward transition at step t.
- $\bar{\mathbf{Q}}_t = \mathbf{Q}_t \mathbf{Q}_{t-1} \cdots \mathbf{Q}_1$ — cumulative transition.
- $\odot$ — elementwise (Hadamard) product.

Austin J. et al., Structured Denoising Diffusion Models in Discrete State-Spaces, 2021.

Summary

-