Discrete Diffusion Models

March 21, 2024

Notation: Given a vector v, we denote $[v]_i$ as the i-th entry of v. Similarly, given a matrix A, we denote $[A]_{ij}$ as the (i,j)-th entry of A.

1 Discrete Data Distribution and Diffusion

We consider data X_0 follows a distribution P_{data} supported on a discrete set $\mathcal{X} = \{v_1, \dots, v_K\}$, where K is the cardinality. Here v_k can be a high-dimensional vector. For example, each entry of $v_k \in \mathbb{R}^d$ can take values in $\{1, \dots, M\}$ and therefore, we have $K = M^d$. We associate a probability vector $p_{\text{data}} \in [0,1]^K$ from the K-dimensional simplex to P_{data}, where the k-th entry of p_{data} is the probability $[p_{\text{data}}]_k = \mathbb{P}(X_0 = v_k)$ and $\sum_{k=1}^K [p_{\text{data}}]_k = 1$. We adopt the convention that capital letter P denotes a discrete distribution and lower-case letter p denotes its probability vector.

Given a pre-collected data set consisting of n sample points $\{x_1, \dots, x_n\}$, we aim to device a discrete diffusion model for generating new data mimicking the unknown data distribution P_{data}.

Forward Process on Discrete Support Analogous to the Gaussian noise corruption for continuous data distributions, we derive a continuous-time Markov process for corrupt the discrete data distribution P_{data}. Let $Q_t \in \mathbb{R}^{K \times K}$ be a time-dependent transition matrix for $t \in [0, T]$. Here T is a sufficiently large terminal time. At time t, we denote the corrupted discrete distribution as P_t with a probability vector p_t. The forward evolution of p_t is described as

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = Q_t p_t \quad \text{with} \quad p_0 = p_{\text{data}}. \tag{1}$$

The terminal distribution of (1) is P_T, subject to the choice of the transition matrix Q_t. There are flexible choices on Q_t, yet they should ensure that the stationary distribution P_∞ of (1) is relatively simple and P_T is close to P_∞. In the following, we review several design examples of Q_t in the existing literature.

- **Uniform**. Setting $\beta_t \in [0,1]$, we choose

  $$Q_t = (1 - \beta_t) I + \frac{\beta_t}{K} \cdot \mathbf{1}\mathbf{1}^\top,$$

  where $\mathbf{1}$ is a vector of ones. The corresponding stationary distribution is uniform on $\mathcal{X}$.

- **Absorbing**. We augment $\mathcal{X}$ by an absorbing state v_{K+1} and denote $\mathcal{X}_{\text{aug}} = \mathcal{X} \cup \{x_{K+1}\}$. Accordingly, all the probability vectors in (1) is augmented by one dimension. With $\beta_t \in [0,1]$, we choose

  $$Q_t = (1 - \beta_t) I + \frac{\beta_t}{K} \cdot \mathbf{1} e_{K+1}^\top,$$

  where e_{K+1} is the one-hot vector with only the $(K+1)$-th entry being one. The stationary distribution is a point mass on the absorbing state v_{K+1}.

- **Discretized Gaussian**. Setting $\beta_t \in [0,1]$ again, we choose

  $$[Q_t]_{ij} = \begin{cases} \frac{1}{Z_t} \exp\left(-\frac{4|i-j|^2}{(K-1)^2 \beta_t}\right) & \text{if } i \neq j \\ 1 - \sum_{s \neq j} [Q_t]_{sj} & \text{if } i = j \end{cases},$$

where $Z_t = 1 + 2\sum_{m=1}^{K-1} \exp\left(-\frac{4m^2}{(K-1)^2\beta_t}\right)$ is a normalization constant.

More examples can be found in https://arxiv.org/pdf/2107.03006.pdf.

Solving (1), we can show the marginal distribution satisfies

$$p_t = \exp\left(\int_0^t Q_\tau \mathrm{d}\tau\right) p_{\text{data}}.$$

Proof. We take the ansatz $p_t = \exp\left(\int_0^t Q_\tau \mathrm{d}\tau\right) u_t$, where $u_t \in \mathbb{R}^K$ is a time-varying vector. Taking derivative of p_t with respect to t yields

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = Q_t \exp\left(\int_0^t Q_\tau \mathrm{d}\tau\right) u_t + \exp\left(\int_0^t Q_\tau \mathrm{d}\tau\right) \frac{\mathrm{d}u_t}{\mathrm{d}t}.$$

Comparing with (1), we deduce $\frac{\mathrm{d}u_t}{\mathrm{d}t} = 0$ for all t, which implies u_t is a constant vector. Therefore, we simplify $p_t = \exp(\int_0^t Q_\tau \mathrm{d}\tau) u_0$. The unknown u_0 can be determined by taking $t = 0$, where we obtain $u_0 = p_{\text{data}}$. □

Backward Process on Discrete Support Similar to the Gaussian noise corruption for continuous distributions, the forward process (1) assumes a backward process:

$$\frac{\mathrm{d}q_t}{\mathrm{d}t} = \bar{Q}_t q_t \quad \text{with} \quad [\bar{Q}_t]_{ij} = \begin{cases} \frac{[p_{T-t}]_i}{[p_{T-t}]_j}[Q_{T-t}]_{ji} & \text{if } i \neq j \\ -\sum_{s\neq i}[Q_{T-t}]_{is}\frac{[p_{T-t}]_s}{[p_{T-t}]_i} & \text{if } i = j \end{cases}. \tag{2}$$

Here q_t matches p_{T-t}, as a time reversal of the forward process.

Proof. We consider a time $t \in (0, T)$ and an infinitesimal increment $\delta > 0$. Denote X_t and $\bar{X}_t$ as the random processes corresponding to the forward and backward processes, respectively. We have $\bar{X}_t \overset{\mathrm{d}}{=} X_{T-t}$ in the distribution. We examine the conditional transition probability

$$\begin{aligned}\mathbb{P}\left(\bar{X}_{t+\delta} = v_i | \bar{X}_t = v_j\right) &= \frac{\mathbb{P}\left(\bar{X}_{t+\delta} = v_i, \bar{X}_t = v_j\right)}{\mathbb{P}\left(\bar{X}_t = v_j\right)} \\ &= \frac{\mathbb{P}\left(\bar{X}_t = v_j | \bar{X}_{t+\delta} = v_i\right)\mathbb{P}\left(\bar{X}_{t+\delta} = v_i\right)}{\mathbb{P}\left(\bar{X}_t = v_j\right)}.\end{aligned}$$

We take derivative of $\mathbb{P}\left(\bar{X}_{t+\delta} = v_i | \bar{X}_t = v_j\right)$ with respect to δ, which gives rise to

$$\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{P}\left(\bar{X}_{t+\delta} = v_i | \bar{X}_t = v_j\right) = \underbrace{\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{P}\left(\bar{X}_t = v_j | \bar{X}_{t+\delta} = v_i\right)\frac{\mathbb{P}\left(\bar{X}_{t+\delta} = v_i\right)}{\mathbb{P}\left(\bar{X}_t = v_j\right)}}_{A_t(\delta)} + \underbrace{\mathbb{P}\left(\bar{X}_t = v_j | \bar{X}_{t+\delta} = v_i\right)\frac{\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{P}\left(\bar{X}_{t+\delta} = v_i\right)}{\mathbb{P}\left(\bar{X}_t = v_j\right)}}_{B_t(\delta)}.$$

In term $A_t(\delta)$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{P}\left(\bar{X}_t = v_j | \bar{X}_{t+\delta} = v_i\right) &= \frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{P}\left(X_{T-t} = v_j | X_{T-t-\delta} = v_i\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}\delta}\left[\exp\left(\int_{T-t-\delta}^{T-t} Q_\tau \mathrm{d}\tau\right) e_i\right]_j \\ &= [Q_{T-t-\delta}]_{ji}.\end{aligned} \tag{3}$$

Sending δ to 0, we obtain

$$\lim_{\delta\to 0^+} A_t(\delta) = [Q_{T-t}]_{ji}\frac{[p_{T-t}]_i}{[p_{T-t}]_j}.$$

In term $B_t(\delta)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{P}\left(\bar{X}_{t+\delta}=v_i\right)=\frac{\mathrm{d}}{\mathrm{d}\delta}[p_{T-t-\delta}]_i=-\left[Q_{T-t-\delta}\ p_{T-t-\delta}\right]_i,$$

by the forward equation (1). In the case of $i\neq j$, sending δ to 0, we deduce

$$\lim_{\delta\to 0^+}B_t(\delta)=\lim_{\delta\to 0^+}\mathbb{P}\left(\bar{X}_t=v_j|\bar{X}_{t+\delta}=v_i\right)\frac{-\left[Q_{T-t-\delta}\ p_{T-t-\delta}\right]_i}{[p_{T-t}]_j}=0,$$

since the first transition probability converges to 0 if $i\neq j$ by (3) and the remaining terms are finite. When $i=j$, we have

$$\lim_{\delta\to 0^+}B_t(\delta)=-\frac{[Q_{T-t}\ p_{T-t}]_i}{[p_{T-t}]_i}=-\frac{\sum_s[Q_{T-t}]_{is}[p_{T-t}]_s}{[p_{T-t}]_i}=-[Q_{T-t}]_{ii}-\sum_{s\neq i}[Q_{T-t}]_{is}\frac{[p_{T-t}]_s}{[p_{T-t}]_i}.$$

Now summing up $A_t(\delta)$ and $B_t(\delta)$, we derive

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{P}\left(\bar{X}_{t+\delta}=v_i|\bar{X}_t=v_j\right)&=\begin{cases}[Q_{T-t}]_{ji}\frac{[p_{T-t}]_i}{[p_{T-t}]_j} & \text{if } i\neq j\\ [Q_{T-t}]_{ii}-[Q_{T-t}]_{ii}-\sum_{s\neq i}[Q_{T-t}]_{is}\frac{[p_{T-t}]_s}{[p_{T-t}]_i} & \text{if } i=j\end{cases}\\&=\begin{cases}[Q_{T-t}]_{ji}\frac{[p_{T-t}]_i}{[p_{T-t}]_j} & \text{if } i\neq j\\ -\sum_{s\neq i}[Q_{T-t}]_{is}\frac{[p_{T-t}]_s}{[p_{T-t}]_i} & \text{if } i=j\end{cases}.\end{aligned}$$

Lastly, we observe that $[\bar{Q}_t]_{ij}=\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{P}\left(\bar{X}_{t+\delta}=v_i|\bar{X}_t=v_j\right)$ and in one row, the diagonal element of $\bar{Q}_t$ is the negative sum of the off-diagonal entries. □

2 Discrete Score Estimation

We observe from the backward process (2) that to generate new samples, we only need to estimate the ratios $\frac{[p_t]_i}{[p_t]_j}$ for all $i,j\in\{1,\dots,K\}$ and $t\in[0,T]$. We can view this probability ratio as an analogy to the score function in the continuous distribution setting. Accordingly, we term these probability ratios as discrete score functions.

A rather naïve idea of estimating the ratios is

$$\underset{s}{\operatorname{argmin}}\ \mathbb{E}_{v_i\sim p_t}\left[\sum_{i\neq j}\left([s(v_i,t)]_j-\frac{[p_t]_i}{[p_t]_j}\right)^2\right],\tag{4}$$

which resembles the conventional continuous distribution score matching (replacing the score $\nabla\log p_t$ by the ratio $\frac{[p_t]_i}{[p_t]_j}$). Here s is a trainable neural network. Objective (4) is known as the concrete score matching in https://arxiv.org/pdf/2211.00802.pdf. As pointed out in https://arxiv.org/pdf/2310.16834.pdf, although theoretically plausible, the empirical performance of (4) suffers due to the fact that the quadratic loss does not enforce s to be positive.

As an alternative, score entropy was proposed in https://arxiv.org/pdf/2310.16834.pdf to preserve the positiveness of the learned ratios.

3 Questions

- Understand how discrete diffusion models estimate data distributions. Several key factors to take into consideration:
  1. The cardinality of the support and how to mitigate the influence.
  2. The range of ratios.

3. Design of the forward process and their corresponding impact.

- Any difference among discrete score estimation methods?
- Any connection or advantage of applying discrete diffusion to graphical models compared to standard continuous diffusion? https://arxiv.org/pdf/2309.11420.pdf.