Diffusion Generative Model for Categorical Data Modeling

Presenter : Florence Regol
supervised by Prof. Mark Coates

McGill University - Compnet Lab
Bellairs Workshop

December 13th

Overview

1. Categorical data modeling.
2. Generative diffusion model.
3. Diffusion model for categorical data.

Categorical data modeling

Categorical Data

Categorical data definition

- One of K **categories** $\{A, B, \ldots\}, |\{A, B, \ldots\}| = K$.
- No intrinsic **order** in $\{A, B, \ldots\}$.
- Sequence or vector of S categorical variables :

$$\mathbf{x} = [x_1, x_2, \ldots x_S], \quad x_i \in \{A, B, \ldots\}.$$

Categorical Data - Example

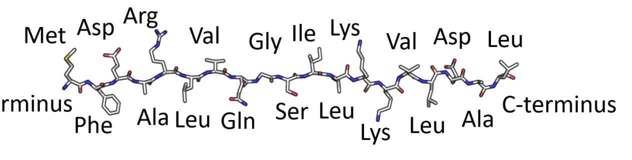
Generating text

$\mathbf{x} =$ [Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras ornare rutrum dapibus. Proin eu ullamcorper ex. Ut fringilla dolor eget elit tincidunt, ... semper et luctus vel, dictum vitae ligula. Maecenas tristique vulputate libero ac molestie. Nam commodo nunc turpis, ac convall] $\sim p_\theta(\mathbf{X})$

$x_i \in \{a, b, c, d, \ldots, w, x, y, z,'', ., -, \ldots\}$, set of **characters**.
$S =$ length of sentence or text.

Categorical Data - Example

Generating proteins



$\sim p_\theta(\mathbf{X})$

Primary protein structure

credit: cropped, Shafee. (2007) Summary of protein structure (primary, secondary, tertiary, and quaternary) using the example of pcna.

$x_i \in \{Ala, Arg, \ldots, Val\}$, set of **amino acids.**

$S =$ size of the protein.

Generative Model Categorical Data

Problem formulation

- Multivariate R.V. $\mathbf{X} \sim p(\mathbf{X})$. $(S \times K)$
- Given dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}, \mathbf{X}_i \overset{\text{iid}}{\sim} p(\mathbf{X})$
- Task $\rightarrow$ learn a generative distribution $p_\theta(\mathbf{X})$ for $\mathbf{X}$.

Generative Model Categorical Data

Performance metric

Negative log likelihood

$$
\begin{aligned}
NLL =& \frac{1}{N}\sum_{i=1}^{N} -\log\big(p_\theta(\mathbf{x}_i)\big) \quad \mathbf{x}_i \in \mathcal{D} \\
\approx& -\mathbb{E}_{p(\mathbf{X})}[\log p_\theta(\mathbf{X})] \quad \text{(cross entropy)}
\end{aligned}
$$

Metric for sample quality $s(\cdot)$

$$
\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s(\mathbf{x}_i) \quad \mathbf{x}_i \sim p_\theta(\mathbf{X})
$$

Generative Model Categorical Data

Challenges

- Non-differentiability $\rightarrow$ hard to optimize.
- No natural ordering $\rightarrow$ can't readily apply continuous approach with **thresholding** methods.

Solution : Learn in **continuous space**,
then **map back** to the nominal space.

Generative diffusion model

Generative Model Overview

Generative Models	+	−
GANs	fast & high res. sampling	no *NLL*/ sample diversity
VAE	diverse/ good *NLL*	sample quality
AR	diverse/ good *NLL*	efficiency
Norm. Flows	diverse/ good *NLL*	sample quality/efficiency
Energy Based (Diffusion)	high res./qual./diverse sample	*NLL*

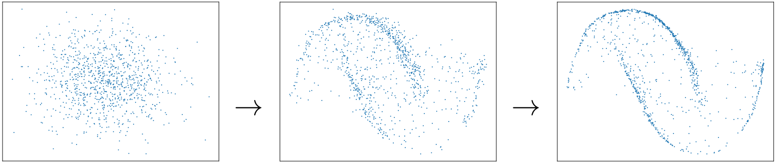
Based on [1]. No clear winner, all have **trade-offs**.

[1]S. Bond-Taylor et al., "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," 2021

Diffusion Model

Diffusion model summary:

- Stochastically transform **data** $\rightarrow$ **noise**.
- Learn to **reconstruct data** from noise.
- Both use **multiple small steps**.



$\mathbf{Z}^T$ $\qquad \ldots, \mathbf{Z}^{t+1}, \mathbf{Z}^t, \mathbf{Z}^{t-1}, \ldots$ $\qquad \mathbf{Z}^0$

Introduced by (J. Sohl-Dickstein et al., 2015.) [2].
Popularized by (J. Ho et al., 2020) [3].

Diffusion Model

- $\mathbf{Z}^0 \to$ data $\mathbf{Z} \in \mathbb{R}^d$.
- $\dots, \mathbf{Z}^{t+1}, \mathbf{Z}^t, \mathbf{Z}^{t-1}, \cdots \to$ latent R.V.
- $\mathbf{Z}^T \to \mathcal{N}(\mathbf{0}, \mathbf{1})$ random noise.

$p^{diff}(\mathbf{Z})$ is a **latent variable model**:

$$\begin{aligned} p_\phi^{diff}(\mathbf{Z}) &\stackrel{\text{m}}{=} f_\phi(\mathbf{Z}^0) \quad (\mathbf{Z} = \mathbf{Z}^0) \\ &= \int f_\phi(\mathbf{Z}^{0:T}) d\mathbf{Z}^{1:T}. \end{aligned}$$

Diffusion Model

Data $\rightarrow$ **noise** : Fixed **Gaussian Markov chain** forward process.

$$q(\mathbf{Z}^t|\mathbf{Z}^{t-1}) = \mathcal{N}(\mathbf{Z}^t; \sqrt{1-\beta_t}\mathbf{Z}^{t-1}, \beta_t \mathbf{I})$$

Increasing noise : $\beta_i < \beta_{i+1} \in (0, 1)$.

$$q(\mathbf{Z}^{0:T}) = q(\mathbf{Z}^0) \prod_{t=1}^{T} q(\mathbf{Z}^i|\mathbf{Z}^{t-1})$$

Closed form $q(\mathbf{Z}^t|\mathbf{Z}^0)$:

$$q(\mathbf{Z}^t|\mathbf{Z}^0) = \mathcal{N}(\mathbf{Z}^t; \sqrt{\bar{\alpha}_t}\mathbf{Z}^0, (1-\bar{\alpha}_t)\mathbf{I})$$

Increasing noise : $(1-\bar{\alpha}_t), \quad \bar{\alpha}_t = \prod_{t=1}^{T}(1-\beta_t)$

Diffusion Model

Noise $\rightarrow$ **data** : **Learned** Gaussian Markov chain.

$$f_\phi(\mathbf{Z}^{t-1}|\mathbf{Z}^t) = \mathcal{N}(\mathbf{Z}^{t-1}; \mu_\phi(\mathbf{Z}^t, t), \sigma_\phi(\mathbf{Z}^t, t))$$

$$f_\phi(\mathbf{Z}^{0:T}) = f(\mathbf{Z}^T) \prod_{t=1}^{T} f_\phi(\mathbf{Z}^{t-1}|\mathbf{Z}^t)$$

$$f(\mathbf{Z}^T) = \mathcal{N}(\mathbf{Z}^T; \mathbf{0}, \mathbf{1})$$

$$\mu_\phi(\mathbf{Z}^t, t) = NN_\phi^1(\mathbf{Z}^t, t)$$
$$\sigma_\phi(\mathbf{Z}^t, t) = NN_\phi^2(\mathbf{Z}^t, t) \text{ , or fixed.}$$

Optimization

$$p_\phi^{diff}(\mathbf{Z}) = f_\phi(\mathbf{Z}^0) = \int f_\phi(\mathbf{Z}^{0:T}) d\mathbf{Z}^{1:T}$$

Variational inference:

$$\begin{aligned}
\log p_\phi^{diff}(\mathbf{Z}) &= \log \left(\int \frac{q(\mathbf{Z}^{1:T}|\mathbf{Z}^0)}{q(\mathbf{Z}^{1:T}|\mathbf{Z}^0)} f_\phi(\mathbf{Z}^{0:T}) d\mathbf{Z}^{1:T} \right) \\
&\geq \mathbb{E}_{q(\mathbf{Z}^{1:T}|\mathbf{Z}^0)} \Big[\log f_\phi(\mathbf{Z}^{0:T}) - \log(q(\mathbf{Z}^{1:T}|\mathbf{Z}^0)) \Big] \\
&\triangleq -\mathcal{L}_\phi^{diff}(\mathbf{Z})
\end{aligned}$$

Diffusion Model - Optimization

- Rearrange $\mathcal{L}_{\phi}^{diff}(\mathbf{Z})$ as a sum of **Gaussian KL/likelihood**:

$$\mathcal{L}_{\phi}^{diff}(\mathbf{Z}) = \mathbb{E}_{q(\mathbf{Z}^{1:T}|\mathbf{Z}^{0})}[L_0 + \sum_{t=2}^{T} L_t + L_T].$$

- $L_0 = \mathcal{D}_{KL}(q(\mathbf{Z}^T|\mathbf{Z}^0)||f(\mathbf{Z}^T))$ ensures **data $\rightarrow$ noise.**
- $L_t = \mathcal{D}_{KL}(q(\mathbf{Z}^t|\mathbf{Z}^{t-1}, \mathbf{Z}^0)||f_{\phi}(\mathbf{Z}^t|\mathbf{Z}^{t-1}))$ undo step noise.
- $L_T = \log f_{\phi}(\mathbf{Z}^0|\mathbf{Z}^1)$ likelihood.

In practice, a **reweighted** version $\mathcal{L}_{\phi}^{simple}$ is **stochastically** optimized.

Categorical diffusion model - CDM

Recall Challenges

- Non-differentiability $\rightarrow$ hard to optimize.
- No natural ordering $\rightarrow$ can't readily apply continuous approach with thresholding methods.

Solution : Learn in **continuous space**,
then **map back** to the nominal space.

$$p(\mathbf{X}) = \int p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})d\mathbf{Z}.$$

CDM - Baselines

SOTA baselines

- Categorical normalizing flows **CNF** [4]

$$p(\mathbf{X}) = \int p_{decoder}(\mathbf{X}|\mathbf{Z})p_{flow}(\mathbf{Z})d\mathbf{Z}.$$

- Argmax flows [5]

$$p(\mathbf{X}) = \int p_{argmax}(\mathbf{X}|\mathbf{Z})p_{flow}(\mathbf{Z})d\mathbf{Z}.$$

Other baselines

- **Discretized** generative models (NF [6], GANs [7]).
- **Dequantization** (NF [8]).

Categorical Diffusion Model

$$p_{\theta,\phi}(\mathbf{X}) = \int p_\theta(\mathbf{X}|\mathbf{Z}) p_\phi^{diff}(\mathbf{Z}) d\mathbf{Z}$$

Push complexity to $p_\phi^{diff}(\mathbf{Z})$, encoder /decoder as in [4] $\rightarrow$

- **Encoder** : Factorized simple model $q_\theta(\mathbf{Z}|\mathbf{X})$ $\mathbf{Z} \in \mathbb{R}^{d*K}$:

$$q_\theta(\mathbf{Z}|\mathbf{X}) = \prod_{i=1}^{S} Logistic(\mathbf{Z}_i|\mu_\theta(X_i), \sigma_\theta(X_i)).$$

- **Decoder** : Bayes rule $p_\theta(\mathbf{X}|\mathbf{Z})$:

$$p_\theta(\mathbf{X}|\mathbf{Z}) = \frac{\tilde{p}(\mathbf{X}) q_\theta(\mathbf{Z}|\mathbf{X})}{\sum_{\mathbf{X}' \in \{A,B,\dots\}} \tilde{p}(\mathbf{X}') q_\theta(\mathbf{Z}|\mathbf{X}')} \quad \tilde{p}(\cdot) = \frac{1}{K}.$$

Optimization

- **Intractable** $p_{\theta,\phi}(\mathbf{X}) = \int p_\theta(\mathbf{X}|\mathbf{Z}) p_\phi^{diff}(\mathbf{Z}) d\mathbf{Z}$
- $\rightarrow$ **variational inference**

$$\log\left(p_{\theta,\phi}(\mathbf{X})\right) \geq \mathbb{E}_{q_\theta(\mathbf{Z}|\mathbf{X})}\left[\log\left(p_\phi^{diff}(\mathbf{Z})\right) + \log\left(\frac{p_\theta(\mathbf{X}|\mathbf{Z})}{q_\theta(\mathbf{Z}|\mathbf{X})}\right)\right]$$

- **Intractable** $\log(p_\phi^{diff}(\mathbf{Z}))$
- $\rightarrow$ **variational inference** previously shown:

$$\log p_\phi^{diff}(\mathbf{Z}) \geq -\mathcal{L}_\phi^{diff}(\mathbf{Z})$$

Final objective

$$
\begin{aligned}
\log\left(p_{\theta,\phi}(\mathbf{X})\right) &= \log\left(\int p_\theta(\mathbf{X}|\mathbf{Z})\, p_\phi^{diff}(\mathbf{Z})\, d\mathbf{Z}\right) \\
&\geq \mathbb{E}_{q_\theta(\mathbf{Z}|\mathbf{X})}\left[\log\left(p_\phi^{diff}(\mathbf{Z})\right) + \log\left(\frac{p_\theta(\mathbf{X}|\mathbf{Z})}{q_\theta(\mathbf{Z}|\mathbf{X})}\right)\right] \\
&\geq \mathbb{E}_{q_\theta(\mathbf{Z}|\mathbf{X})}\left[-\mathcal{L}_\phi^{diff}(\mathbf{Z})\right) + \log\left(\frac{p_\theta(\mathbf{X}|\mathbf{Z})}{q_\theta(\mathbf{Z}|\mathbf{X})}\right] \\
\text{In practice :}\ & \mathbb{E}_{q_\theta(\mathbf{Z}|\mathbf{X})}\left[-\lambda\mathcal{L}_\phi^{simple}(\mathbf{Z}) + \log\left(\frac{p_\theta(\mathbf{X}|\mathbf{Z})}{q_\theta(\mathbf{Z}|\mathbf{X})}\right)\right]
\end{aligned}
$$

$\lambda \quad \propto \quad \mathcal{L}_\phi^{diff} / \mathcal{L}_\phi^{simple}$ works best.

Experiment

Synthetic Experiment - Experiment details

Permutation dataset

$$p(\mathbf{X}) = \begin{cases} \frac{1}{K!} & \text{if } \mathbf{X} \in \mathcal{S}(K) \\ 0 & \text{o.w.} \end{cases}$$

Performance metric sample quality $s(\cdot)$:

$$\hat{p}_{valid} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}[\mathbf{x}_i \in \mathcal{S}(K)] \quad \mathbf{x}_i \sim p_\theta(\mathbf{X})$$

Experiment details

- **CDM** vs **CNF**
- Hyperparameter tuning $p(\mathbf{Z})$ & learning rate
- Fixed $p_\theta(\mathbf{X}|\mathbf{Z}), q_\theta(\mathbf{Z}|\mathbf{X})$
- $K = 3, S = 3$, $\mathbf{Z} \in \mathbb{R}^6$

Synthetic Experiment - Results

	uniform	**CNF**	**CDM**
$(\uparrow)\hat{p}_{valid}, M = 1000$	0.22	0.96	**0.99**
$(\downarrow)NLL$	3.29	$\leq$ **0.64**	0.98
time for training	-	5x	1
time for sampling	-	1.5x	1
num parameter	-	10x	1

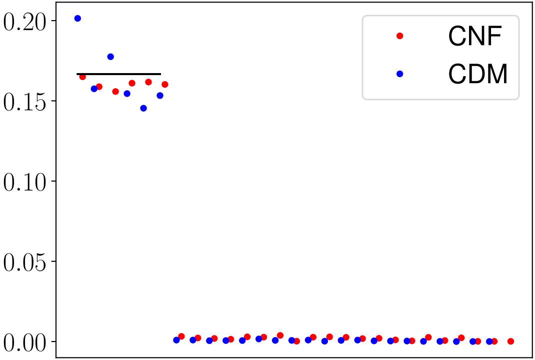
20 trials, (statistically significant at the 5% level using a Wilcoxon signed rank test).

Consistent with literature

- **CNF** needs **more parameters** and is **slower**
- **CNF** reaches better **likelihood**
- **CDM** reaches better **sample quality**

Synthetic Experiment - Results

Empirical $\hat{p}_\phi(\mathbf{X})$ estimated from 10,000 samples.



- **CNF** hard to **set 0** $\rightarrow$ lower $\hat{p}_{valid}$.
- **CDM** hard to set **uniform support**.

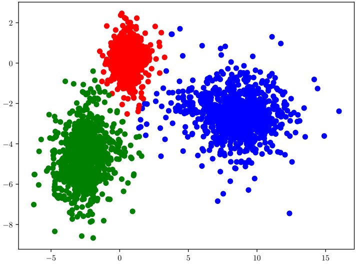
Synthetic Experiment

Why **CNF** has trouble setting points **outside the support** to 0?
Possible explanation:

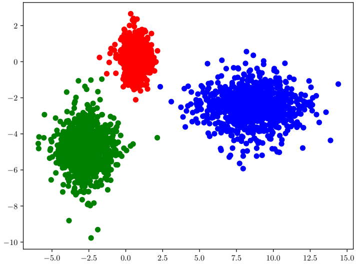
- Complexity in $p(\mathbf{Z}) \rightarrow$ **multimodal** $p(\mathbf{Z})$.
- Known topological issue of **NF** $\rightarrow$ **separate volume** [9, 10].
- Solution $\rightarrow$ introduce **stochasticity** in the process [10].

Synthetic Experiment

Visualization of **latent Z** of **CNF**



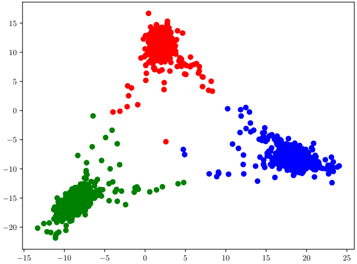
$\mathbf{Z} \sim p_{\phi}^{flow}(\mathbf{Z})$



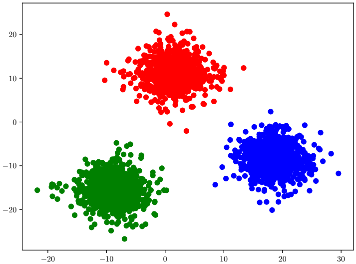
$\mathbf{Z} \sim q_{\theta}(\mathbf{Z}|\mathbf{X})$ (data)

Synthetic Experiment

Visualization of **latent Z** of **CDM**



$\mathbf{Z} \sim p_\phi^{diff}(\mathbf{Z})$



$\mathbf{Z} \sim q_\theta(\mathbf{Z}|\mathbf{X})$ (data)

References

[1] S. Bond-Taylor, A. Leach, Y. Long, and C. Willcocks, “Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models,” *arXiv preprint arXiv:2103.04922*, 2021.

[2] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Proc. Int. Conf. Machine Learning (ICML)*, vol. 37, 2015, pp. 2256–2265.

[3] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proc. Adv. Neural Info. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[4] P. Lippe and E. Gavves, “Categorical normalizing flows via continuous transformations,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021.

[5] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling, “Argmax flows: Learning categorical distributions with normalizing flows,” in *Proc. Adv. in Approximate Bayesian Inference*, 2021.

[6] D. Tran, K. Vafa, K. Agrawal, L. Dinh, and B. Poole, “Discrete flows: Invertible generative models of discrete data,” in *Proc. Adv. Neural Info. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[7] R. D. Camino, C. A. Hammerschmidt, and R. State, “Generating multi-categorical samples with generative adversarial networks,” *arXiv preprint arXiv:1807.01202*, 2018.

[8] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel, “Flow++: Improving flow-based generative models with variational dequantization and architecture design,” in *Proc. Int. Conf. Machine Learning (ICML)*, vol. 97, 2019, pp. 2722–2730.

[9] R. Cornish, A. Caterini, G. Deligiannidis, and A. Doucet, “Relaxing bijectivity constraints with continuously indexed normalising flows,” in *Proc. Int. Conf. Machine Learning (ICML)*, vol. 119, 2020, pp. 2133–2143.

[10] H. Wu, J. Köhler, and F. Noe, “Stochastic normalizing flows,” in *Proc. Adv. Neural Info. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 5933–5944.

Questions