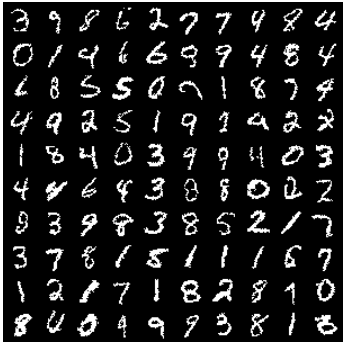
Simplex-to-Euclidean Bijections for Categorical Flow Matching

Bernardo Williams [1], Victor M. Yeom-Song [2,3], Marcelo Hartmann [1], Arto Klami [1]

[1]University of Helsinki
[2]Aalto University
[3]ELLIS Institute Finland

The Problem: Discrete Data Generation



Binary images ($K = 2$ categories)

```
ACCCGCGACAGACACCAAA
CTGGCGCTGCCCTGTGGGA
CGCGGACACTCGATCTCGG
AGCTCATTGCACGCTGCTG
CGGTTCAGGTCGGGGAACG
CCAGCTGCGCAGCGCGAAG
CACGATGGGCCAGCTAATG
GTCGGGGCATGGACCAATT
GGCGTCGACGGGGGCGGGA
TCGGGGAAGGTATATAAGC
CAGCCGGGCGGCCCGGGGC
```

DNA sequences ($K = 4$ categories)

the founder of modern anarchist theory in what is property proudhon answers with the famous accusation property is theft in this work he opposed the institution of decreed property propri t where owners have complete rights to use and abuse their property as they wish such as exploiting workers for profit in its place proudhon supported what he called possession individuals can have limited rights to use resources capital and goods in accordance with principles of equality and justice proudhon s vision of anarchy which he called mutualism mutuellisme involved an exchange economy where individuals and groups could trade the products of their labor using labor notes which represented the amount of working time

Text ($K = 27$ categories)

The Problem: Discrete data generation

- K Categorical data are vertices of the simplex:

$$c \in \{1, .., K\},$$

$$\Delta^D := \{\boldsymbol{x} \in \mathbb{R}_+^K : \sum_{k=1}^{K} x_k = 1\}$$

$$\boldsymbol{e}_c := (0, \dots \underbrace{, 1,}_{c\text{-th entry}} \dots, 0) \in \Delta^D$$

- **Problem:** The simplex has boundaries and **non-Euclidean** geometry e.g. Fisher.

Goal. Reuse *Euclidean generative generative models* while respecting Simplex geometry.

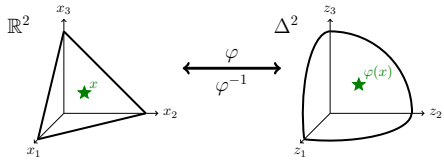


Figure: The Fisher information geometry is the sphere.

Method: Move the Modeling Problem to Euclidean Space

1. Stochastic interpolation

$$\boldsymbol{x} = \lambda \boldsymbol{e}_c + (1-\lambda)\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathrm{Dirichlet}(\alpha).$$

2. Apply a smooth bijection

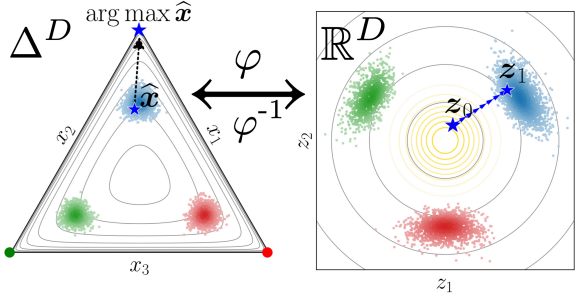
$$\boldsymbol{z}_1 = \varphi(\boldsymbol{x}), \qquad \varphi : \mathring{\Delta}^D \longrightarrow \mathbb{R}^D.$$

3. Train flow matching

$$\boldsymbol{z}_t = (1-t)\boldsymbol{z}_0 + t\boldsymbol{z}_1, \qquad \dot{\boldsymbol{z}}_t = \boldsymbol{z}_1 - \boldsymbol{z}_0,$$

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}\left[\|\boldsymbol{v}_\theta(t, \boldsymbol{z}_t) - \dot{\boldsymbol{z}}_t\|^2\right].$$

Sampling. Draw $\boldsymbol{z}_0 \sim p_0$, solve ODE $\dot{\boldsymbol{z}}_t = \boldsymbol{v}_\theta(t, \boldsymbol{z}_t)$, map back with $\hat{\boldsymbol{x}} = \varphi^{-1}(\boldsymbol{z}_1)$, then use $\arg\max_k \hat{\boldsymbol{x}}_k$.



Method details

Bijection 1: Isometric Logratio transform (ILR) $\boldsymbol{H}$ Helmert matrix.

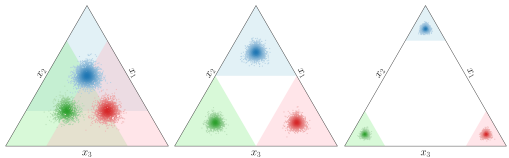
$$\varphi : \Delta^D \to \mathbb{R}^D,\ \boldsymbol{x} \mapsto \boldsymbol{z} = \boldsymbol{H} \log \boldsymbol{x},$$
$$\varphi^{-1} : \mathbb{R}^D \to \Delta^D,\ \boldsymbol{z} \mapsto \boldsymbol{x} = \operatorname{softmax}(\boldsymbol{H}^\top \boldsymbol{z}).$$

Isometric to $(\mathbb{R}^D, \|\cdot\|_2)$.

Bijection 2: Stick-Breaking (SB)
(coordinates are centered before)

$$\varphi : \mathring{\Delta}^D \to \mathbb{R}^D, z_k = \log\left(x_k / \big(1 - \sum_{i=1}^{k} x_i\big) \right),$$
$$\varphi^{-1} : \mathbb{R}^D \to \mathring{\Delta}^D, x_k = \sigma(z_k) \prod_{i=1}^{k-1} \big(1 - \sigma(z_i)\big).$$

Exact recovery:
Interpolation: $\boldsymbol{x} = \lambda \boldsymbol{e}_c + (1-\lambda)\boldsymbol{\varepsilon}$,



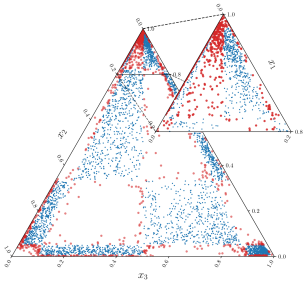
$\lambda = \frac{1}{4}$ $\lambda = \frac{1}{2}$ $\lambda = \frac{3}{4}$

if $\lambda \geq 1/2$ recover original category

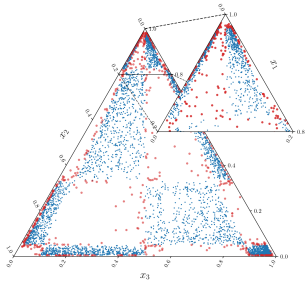
$$c = \arg\max_k x_k.$$

[0] J. J. Egozcue et al., *Isometric logratio transformations for compositional data analysis*, 2003.

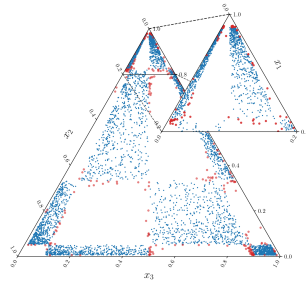
Empirical results



(a) LinearFM, 25.9%



(b) SFM, 12.9%



(c) Ours, 5.4%

Dataset	Metric	Baseline	Ours
BMNIST	FID	SFM 4.62	ILR **4.36**
DNA	SP-MSE	SFM .0258	SB **.0214**
Text8	NLL	SFM 6.85	ILR **6.81**

Pattern. Our approach beats simplex-based methods on discrete-data generation.

Takeaway

FM-$\mathring{\Delta}^D$ makes discrete-data generation look like continuous flow matching.

- Stochastic interpolation turns vertices into recoverable interior points.
- Smooth bijections move the simplex interior to $\mathbb{R}^D$.
- Existing Euclidean generative models can be reused directly.
- Examples include diffusion models, consistency models, flow maps, and mean flows.

See you at Poster 187!

arXiv:
2510.27480

Link to the paper