Simplex Denoising Models

Modeling Discrete Data in Continuous Space

Assumption

Two motivational assumptions:

1. Deterministic / ODE samplers are good: they enable simple/efficient distillation.
2. Modeling in a continuous space is good for guidance (jumping between states is not).

Please, let's assume this without discussion for now.

Outline

1. Overview of denoising approaches for discrete data.
2. Noising on the Simplex
3. Dirichlet Flow Matching
   - Original Dirichlet Flow Matching (Stark et al.)
   - Unsid (old)
   - Dirichlet-CDF Flow Matching (new)

Overview

Three main families of denoising models for **categorical data**:

1. **Continuous Relaxation**
2. **Discrete Diffusion / Flow Matching**
3. **Simplex Diffusion / Flow Matching**

Note: In all three approaches, we usually represent the data as $x_1 = e_i$

1. Continuous Relaxation

Idea

Encode each K-categorical variable as a *one-hot vector* and treat it as a continuous point in $\mathbb{R}^K$.

- State space: $X_t \in \mathbb{R}^K$: add Gaussian noise in the usual way

\+ Advantages

- Entire continuous-diffusion toolbox applies directly
- Deterministic (DDIM-style) sampling available
- Classifier-free and classifier guidance straightforward

– Drawbacks

- Noising path **uncontrolled**: samples can leave the simplex
- **Dimension mismatch**: K dims for $K-1$ degrees of freedom
- Learned boundary is implicit, not structural

2. Discrete Diffusion / Discrete Flow Matching

Idea

Define the noising process *directly* on the categorical state space.

- State space: $X_t \in [K]$ — noising = sampling from a discrete noisy distribution (e.g. token masking, uniform mixing)

\+ Advantages

- Matches the true structure of discrete data
- State-of-the-art **unconditional** generation quality

– Drawbacks

- Dynamics consist of **discrete jumps** between states
- Sampling is **stochastic**: no deterministic analogue
- **Guidance is difficult**: no gradients on a discrete space

3. Simplex Diffusion / Flow Matching

Idea

Define a *continuous* process directly on the **probability simplex** Δ_{K-1}.

- State space: $X_t \in \Delta_{K-1} = \{p \in \mathbb{R}^K : p_i \geq 0,\ \sum_i p_i = 1\}$ — correct $K-1$ dimensional geometry

+ Advantages

- Continuous $\Rightarrow$ deterministic sampler and toolbox for guidance
- Geometry matches true degrees of freedom
- Clean probabilistic interpretation at every step

− Drawbacks

- More complex to implement than Euclidean methods

Comparison Summary

	Cont. Relaxation	Discrete	Simplex
State space	$\mathbb{R}^K$	$[K]$	Δ_{K-1}
Geometry correct	✗	✓	✓
ODE samplers available	✓	✗	✓

Take-away

Discrete Flow Matching is often better for unconditional generation, but at the cost of discontinuities in denoising.

Noising on the Simplex: Design Choices

Linear interpolant approach:

$$x_t = \alpha_t x_1 + (1 - \alpha_t) x_0,$$

$p_t(\cdot \mid x_1)$ does not have full support on $\Delta_{K=1}$
$x_1 \sim p_{data}$, $x_0 \sim \mathcal{U}(\Delta_{K-1})$

$t = 1$ $t = 0.75$ $t = 0.5$ $t = 0.25$ $t = 0$

Dirichlet Flow Matching

Parametric approach:

$$x_t = \mathrm{Dir}(x; \mathbf{1} + \alpha_t x_1)$$

with $\alpha_t \in [0, \infty)$.

$t \approx 1$ $t = 0.75$ $t = 0.5$ $t = 0.25$ $t = 0$

Dirichlet Flow Matching: Motivations

$\Rightarrow$ **Dirichlet Probability Path**: analytically tractable noising schedule entirely on Δ_K.

Why the Dirichlet Distribution

- Dirichlet is the conjugate prior for categorical distribution: Bayesians are happy.
- Naturally recover the **Uniform** over Δ_{K-1} for $\alpha_t = 0$.
- **Aggregation property:**
  - $x \sim \mathrm{Dir}(\boldsymbol{\alpha}) \implies$
  - $x' = (x_1, \ldots, x_i + x_j, \ldots, x_K) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_K)$

Dirichlet Flow Matching: Overview

Goal

Goal: Learn to sample from the conditional distribution:

$$x_{t+h} \sim p(x_{t+h} \mid x_t)$$

Standard Flow Matching Learn the **marginal vector field**:

$$x_{t+h} = x_t + h\, v_t(x_t)$$

Dirichlet Flow Matchin: Parametrization

Parametrization

Instead of v_t directly, we learn

$$P_\theta(x_1 \mid x_t)$$

i.e. a **categorical** over the 'clean' token.

From this, derive the velocity via the *conditional expectation*:

$$v_t = \mathbb{E}_{x_1 \sim P_\theta(x_t \mid x_t)}\big[u_t(x \mid x_1)\big]$$

Dirichlet Flow Matching: Vector Fields

Vector Fields

Marginal:

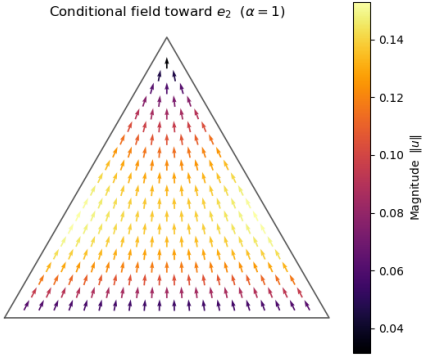
$$v_t = \mathbb{E}_{x_1 \sim P_\theta(x_t|x_t)}\big[u_t(x \mid x_1 = e_i)\big]$$

Conditional:

$$u_t(x \mid x_1 = e_i) = C(x_i, t)(e_i - x)$$

$$C(x_i, \alpha_t) = -\tilde{I}_{x_i}(\alpha_t{+}1, K{-}1)\frac{\mathcal{B}(\alpha_t + 1, K - 1)}{(1 - x_i)^{K-1} x_i^t}$$

with $\tilde{I}_x(a, b) = \frac{\partial}{\partial a} I_x(a, b)$



Problem with Dirichlet Flow Matching

Problem: $u_t(x \mid x_1)$ is not linear in t.

$$x_t + h u_t + \varepsilon \sim p\left(x_{t+h} \mid x_t, x_1\right)$$

- Integrating with a finite-step ODE solver incurs a discretization error.
- Error terms accumulates over steps.
- Numerically unstable.

Can we do better?

Can We Do Better?

Two alternative strategies to address Dirichlet FM's limitations:

1. UNSID

Unrestained **Si**mplex **D**enoising

Exploit *conditional independence* to decouple time steps.

2. CDF-Dirichlet Flow Matching

Use the monotonic map between consecutive conditionals.

Yields exact conditionals and deterministic transport.

Both avoid the compounding error of Dirichlet FM.

Reformulation: What We Actually Need

The one-step distribution factorises as:

$$p(x_{t+h} \mid x_t) \;=\; \sum_{x_1} p(x_{t+h} \mid x_1, x_t) P(x_1 \mid x_t)$$

What we learn

$$P(x_1 \mid x_t) \;\approx\; P_\theta(x_1 \mid x_t)$$

A categorical distribution over the *clean* data given the current position.

What we still need

$$p(x_{t+h} \mid x_1, x_t)$$

Two design options.

UNSID: Unconditional Sampling with Independent Denoising

Key assumption: conditional independence

$$p(x_s \mid x_1, x_t) = p(x_s \mid x_1) \quad \forall\, s \neq t$$

Samples are not dependent across time.

Consequence:

$$x_{t+h} \sim p(x_{t+h} \mid x_t) = \sum_{x_1} p(x_{t+h} \mid x_1) P_\theta(x_1 \mid x_t)$$

Interpretation: Sample $\hat{x}_1 \sim P^\theta(\cdot \mid x_t)$, then re-noise from the probability path directly.

+ Advantages

- Uses the **exact** denoising kernel under the true distribution.
- **No compounding** denoising errors.
- Simple to implement; **fast** inference.

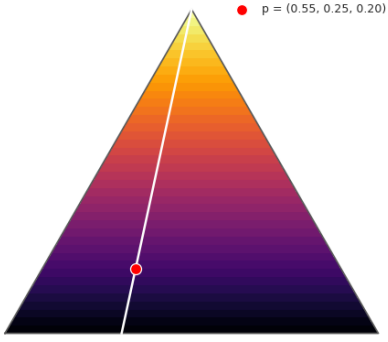
− Drawbacks

- Sampling is **stochastic**: hard to distill
- Guidance faces **jumps** between states.

CDF-Flow Matching

Motivation

We want to find an exact flow map, i.e., a flow that does not involve an approximate discretization step.

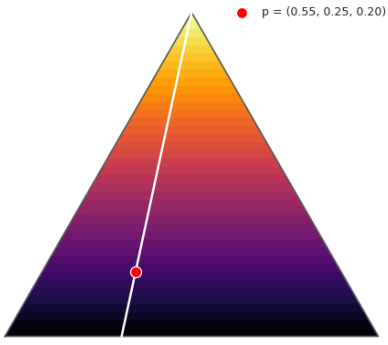


Key Observation (New)

Beta Conditional

Conditioned on $x_1 = e_i$, the conditional path corresponding to the i-th coordinate satisfies $(p(x_{i,t} \mid x_1 = e_i))$:

$$x_{i,t} \sim \mathrm{Beta}(\alpha_t,\ K-1)$$



CDF Flow Matching: Summary

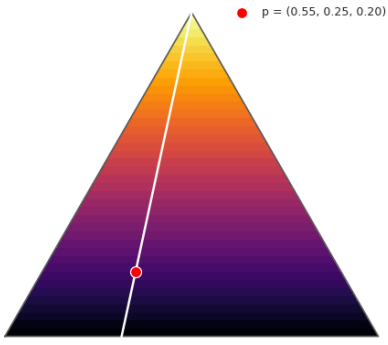
Rank-preserving pushforward

The unique monotone (1D) map between $\mathrm{Beta}(\alpha_t, K-1)$ and $\mathrm{Beta}(\alpha_{t+h}, K-1)$:

$$x_{i,t+h} = I^{-1}\Big(I(x_t;\ \alpha_t, K-1);\ \alpha_{t+h}, K-1\Big)$$

Where $I(x : a, b)$ is the regularized incomplete beta function, aka the Beta CDF.

$$x_{j,t+h} = \frac{1 - x_{j,t+h}}{1 - x_{j,t}} x_{j,t}, j \neq i$$



Dirichlet-CDF Flow Matching: Summary

+ Advantages

- **Exact** conditional transport:
- **Deterministic** trajectories $\Rightarrow$ distillation-friendly
- **Continuous** paths on Δ_K throughout
- **Faster** than Dirichlet FM.

− Limitations

- I^{-1} has **no closed form**: requires numerical evaluation (as Dirichlet FM)
- Exact Conditional Flow Map $\neq$ Exact marginal flow map.

Generalized Dirichlet Flow Matching

Two alternative strategies to address Dirichlet FM's limitations:

1. UNSID

Maximizes the entropy of $x_{t+h}^{\text{UNDSID}} \sim p(x_{t+h} \mid x_t, x_1)$

2. CDF-Dirichlet Flow Matching

Minimizes the entropy of $x_{t+h}^{\text{DFM}} \sim p(x_{t+h} \mid x_t, x_1)$

By interpolating the two, we obtain a generalized Dirichlet flow matching:

$$x_{t+h}^{\text{GEN}} = \lambda\, x_{t+h}^{\text{UNDSID}} + (1-\lambda) x_{t+h}^{\text{DFM}}$$

$\lambda = 0 \implies$ ODE. $\lambda = 1 \implies$ Stochastic

Preliminary Results

Method	MSE	NFE
D3PM-uniform*	.0375	100
DDSM*	.0334	100
Language Model	.0333	1024
Linear FM	.0281	100
Dirichlet FM	.0269	100
CDF-Dirichlet	0.263	100
UNSID	0.256	100

Table: Conditioned on a transcription profile, each method is tasked to generate a DNA sequence with that profile. The MSE is between the predicted regulatory activity of the designed sequence and the ground truth sequence