

# Generative Flows II

Gabriele Steidl

TU Berlin



Jannis Chemseddine



Gregor Kornhardt

Gefördert durch:



|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

## I. Flow Matching for Continuous Distributions

### I.1 Motivation

### I.2 Flow Matching

### I.3 Numerical Examples

## ADD Diffusion Models



Figure 1: Flow matching provides a continuous transition between two distributions. The resulting algorithmic dynamics, statistical inference, and the evidence of physical and topological structures.

## II. Flow Matching for Discrete Distributions

### II.1 Motivation

### II.2 Flow Matching

### II.3 Numerical Examples

Ref: Chemseddine, Kornhardt, St: Spherical flows for sampling categorical distributions, NeuRips 2026, Spotlight talk, 0.95% of papers

## III. SOL: Measuring Gaps between Text Distributions

by Double Sliced Wasserstein Metrics

Ref: Chemseddine, Kornhardt, St: SOL ..., ArXiv 2026

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

## II.1 Motivation

### ◆ Sudoku (Reasoning models)

|                       |                       |
|-----------------------|-----------------------|
| 5 3 .   . 7 .   . . . | 5 3 4   6 7 8   9 1 2 |
| 6 . .   1 9 5   . . . | 6 7 2   1 9 5   3 4 8 |
| . 9 8   . . .   . 6 . | 1 9 8   3 4 2   5 6 7 |
| +-----+-----+         |                       |
| 8 . .   . 6 .   . . 3 | 8 5 9   7 6 1   4 2 3 |
| 4 . .   8 . 3   . . 1 | 4 2 6   8 5 3   7 9 1 |
| 7 . .   . 2 .   . . 6 | 7 1 3   9 2 4   8 5 6 |
| +-----+-----+         |                       |
| . 6 .   . . .   2 8 . | 9 6 1   5 3 7   2 8 4 |
| . . .   4 1 9   . . 5 | 2 8 7   4 1 9   6 3 5 |
| . . .   . 8 .   . 7 9 | 3 4 5   2 8 6   1 7 9 |

### ◆ LLMs for text generation (LM1B), OpenWebText (large!)

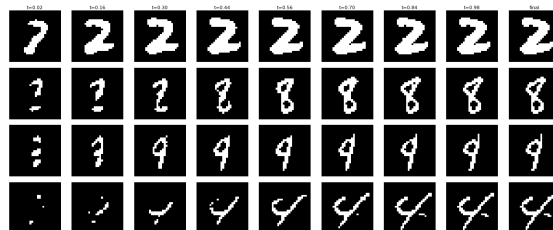
[CLS] ll be able to connect alongside these wind steps! [CLS] " not to see people understand anything that i ' m going to do. [CLS] the bulldogs fell to within 24 - 19 late in the second half. [CLS] the agency needed an alternate place to get the balance on the contact lists. [CLS] and at a

### ◆ MATH reasoning (GSM8K via TinyGSM)

PROMPT:  
A new program had 60 downloads in the first month. The number of downloads in the second month was three times as many as the downloads in the first month, but then reduced by 30% in the third month. How many downloads did the program have total over the three months?

```
month1 = 60
month2 = month1 * 3
month3 = month2 * 0.7
total_downloads = month1 + month2 + month3
result = total_downloads
return result
```

### ◆ binary MNIST



1 2

3 4

5 6

7 8

9 10

11 12

13 14

15 16

17 18

19 20

21 22

23 24

25 26

27 28

29 30

31 32

33 34

35

## II.1 Motivation

**Prevailing approach:** autoregressive decoding with sequential text generation

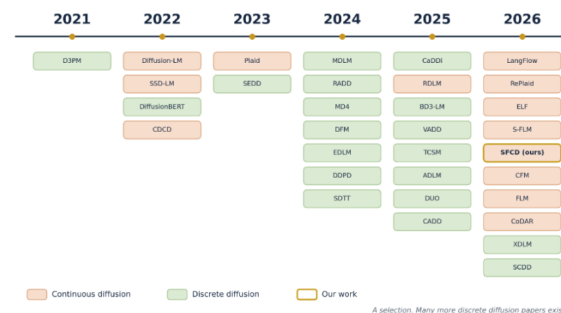
$$p(\mathbf{Y}) = \prod_{l=1}^L p(Y^l \mid Y^1, \dots, Y^{l-1}).$$

Decode left to right,  $L$  sequential steps per generation.

**Alternative:** diffusion or flow-based modeling with entire text generation

**Two groups of methods:**

- ◆ Discrete diffusion models operating directly on the finite state space
- ◆ Continuous diffusion methods: embed tokens from discrete space into a continuous one and define the noise process there



From Liang, Langflow blog, 2026

Refs discrete diffusion: Austin et al. 2023; Campbell et al. 2022; Lou et al. (2024); Sahoo et al. 2024

CDCD: Dieleman et al. (2022), Deschenaux, Gulcehere 2026

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

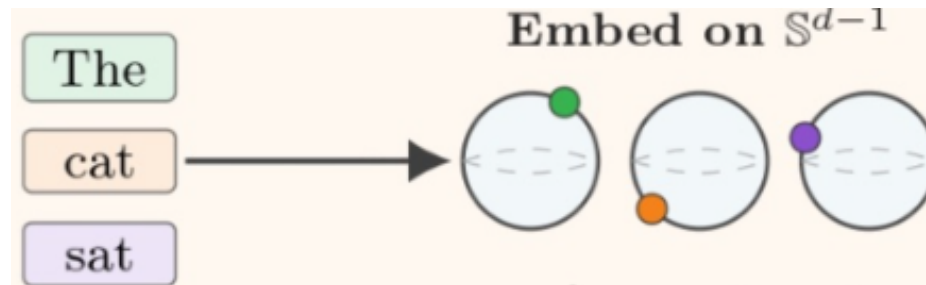
**Setting:** probability space:  $(\Omega, \mathcal{A}, \mathbb{P})$

discrete vocabulary of tokens:  $\mathcal{V} := \{1, \dots, N\}$

discrete random variable:  $\mathbf{Y} = (Y^1, \dots, Y^L) : \Omega \rightarrow \mathcal{V}^L$

- ◆ binary MNIST:  $N = 2$  and  $L = 28^2$ ,  $d = 3$
- ◆ Sudoku :  $N = 10$  with 10 = mask and  $L = 81$ ,  $d = 10$
- ◆ language modeling:  $N = 30522$ ,  $L = 128$ ,  $d = 784$

**Embedding (bijection)**  $\iota : \mathcal{V} \rightarrow \mathcal{W} := \{w_1, \dots, w_N\} \subset \mathbb{S}^{d-1}$



**Task:** Sample from  $\mathbf{W} : \Omega \rightarrow \mathcal{W}^L$  with

$$p_{\text{data}}(\mathbf{w}) := \mathbb{P}(\mathbf{W} = \mathbf{w}) = \mathbb{P}(\mathbf{Y} = \mathbf{y})$$

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

## I. Flow Matching for Continuous Distributions

### I.1 Motivation

### I.2 Flow Matching

### I.3 Numerical Examples

## II. Flow Matching for Discrete Distributions

### II.1 Motivation

### II.2 Flow Matching

### II.3 Numerical Examples

1 2

3 4

5 6

7 8

9 10

11 12

13 14

15 16

17 18

19 20

21 22

23 24

25 26

27 28

29 30

31 32

33 34

35

## II.2 Flow Matching

**Theorem** Let  $p_t : I \rightarrow \mathcal{P}(\mathcal{M})$  be a curve of probability densities on  $\mathcal{M}$ ,  $I = (0, 1)$ , and  $v : I \times \mathcal{M} \rightarrow \mathcal{TM}$  be a sufficiently smooth Borel-measurable vector field. The pair  $(p_t, v_t)$  satisfies the **continuity equation**

$$\partial_t p_t + \operatorname{div}_{\mathcal{M}}(p_t v_t) = 0$$

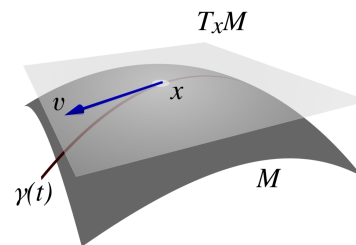
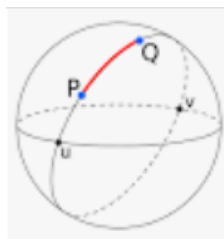
if and only if there exists a solution  $\Phi : I \times \mathcal{M} \rightarrow \mathcal{M}$  of the **flow ODE**

$$\partial_t \Phi(t, x) = v_t(\Phi(t, x)), \quad \Phi(0, x) = x$$

with  $p_t = \Phi(t, \cdot)_{\#} p_0$ .

**Here:**  $\mathcal{M} = \mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$

- ◆ geodesic distance:  $d(x, z) := \theta = \arccos(\langle x, z \rangle)$
- ◆ tangent space at a point  $x \in \mathbb{S}^{d-1}$ :  $\mathcal{T}_x \mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \langle v, x \rangle = 0\}$
- ◆ orthogonal projection onto  $\mathcal{T}_x \mathbb{S}^{d-1}$ :  $P_x(w) := w - \langle w, x \rangle x, \quad w \in \mathbb{R}^d$
- ◆ Riemannian gradient:  $\nabla_{\mathbb{S}^{d-1}} g(x) = P_x(\nabla g(x)) = \nabla g(x) - \langle \nabla g(x), x \rangle x$
- ◆ divergence:  $\operatorname{div}_{\mathbb{S}^{d-1}} u(x) = \operatorname{tr}(\nabla u(x)) - x^\top \nabla u(x) x$



|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

**Three step procedure:** latent distribution = uniform  $X_0 \sim \mathcal{U}(\mathbb{S}^{d-1})$

1. **conditional** flows on  $\mathcal{M}$
2. **conditional** flows on  $\mathcal{M}^L$
3. whole marginal flow on  $\mathcal{M}^L$

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

# 1. Conditional Flows on $\mathcal{M}$

Consider **fixed position**  $l \in \{1, \dots, L\}$  and drop the superscript

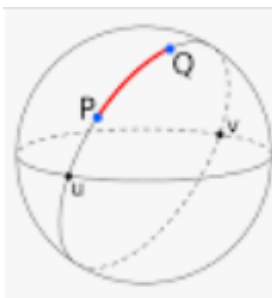
For each  $w \in \mathcal{W}$ , we **aim to construct** a path of conditional random variables  $(X_t \mid W = w) \sim p_t(x \mid w)$  which together with  $v_t(x \mid w)$  fulfill **continuity equation**

$$\partial_t p_t(x \mid w) + \operatorname{div}_{\mathcal{M}}(p_t(x \mid w) v_t(x \mid w)) = 0, \quad t \in (0, 1).$$

**First Idea: Spherical linear interpolation (slerp):**  $\theta := \arccos(\langle x_0, x_1 \rangle)$

$$X_t = \frac{\sin((1-t)\theta)}{\sin \theta} X_0 + \frac{\sin(t\theta)}{\sin \theta} X_1$$

$$(X_t \mid X_1 = w) = \frac{\sin((1-t)\theta)}{\sin \theta} X_0 + \frac{\sin(t\theta)}{\sin \theta} w$$



**Drawbacks:** no analytic expression for  $X_t \sim p_t$  (zero along the flow)

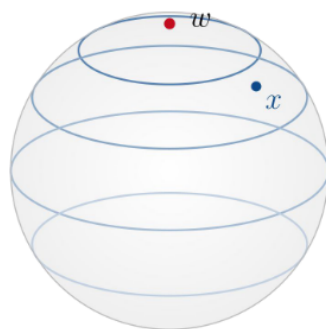
|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

A **density**  $p : \mathbb{S}^{d-1} \rightarrow \mathbb{R}_{\geq 0}$  is called **radially symmetric around**  $w \in \mathbb{S}^{d-1}$ , if there exists a function  $\bar{p} : [-1, 1] \rightarrow \mathbb{R}$  such that

$$p(x) = \bar{p}(\langle w, x \rangle) \quad \text{for all } x \in \mathbb{S}^{d-1}.$$

It relies only on  $S = \langle w, X \rangle \in [-1, 1]$  which has up to a constant the density

$$f(s) := (1 - s^2)^{(d-3)/2} \bar{p}(s), \quad s \in (-1, 1).$$



A **vector field**  $v : \mathbb{S}^{d-1} \rightarrow \mathcal{T}_x \mathbb{S}^{d-1}$  is **radially symmetric around**  $w \in \mathbb{S}^{d-1}$ , if there exists  $\psi : [-1, 1] \rightarrow \mathbb{R}$  such that

$$v(x) = \psi(\langle w, x \rangle) P_x(w).$$

1 2

3 4

5 6

7 8

9 10

11 12

13 14

15 16

17 18

19 20

21 22

23 24

25 26

27 28

29 30

31 32

33 34

35

**Theorem** Let  $(p_t, v_t)$ ,  $t \in I$  be radially symmetric with  $(f_t, \psi_t)$  as above. Then  $(p_t, v_t)$ ,  $t \in I$  satisfy the **continuity equation**

$$\partial_t p_t + \operatorname{div}_{\mathbb{S}^{d-1}}(p_t v_t) = 0, \quad t \in I, x \in \mathbb{S}^{d-1} \setminus \{\pm w\}$$

if and only if  $(f_t, \psi_t)$ ,  $t \in I$  satisfy the **one-dimensional flux equation**

$$\partial_t f_t + \partial_s (f_t \psi_t \cdot (1 - s^2)) = 0, \quad s \in (-1, 1).$$

◆  $p_t \implies f_t(s) = (1 - s^2)^{(d-3)/2} \bar{p}_t(s) \implies \psi_t(s) \text{ from flux eq.}$

1 2

3 4

5 6

7 8

9 10

11 12

13 14

15 16

17 18

19 20

21 22

23 24

25 26

27 28

29 30

31 32

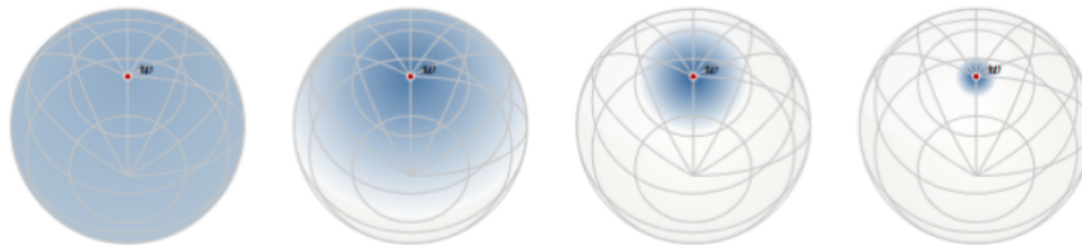
33 34

35

## 1. von Mises-Fisher distribution (vMF)

$$p_t(\cdot|w) := C_d(\kappa_t) \exp(\kappa_t \langle w, x \rangle), \quad C_d(\kappa) := \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} \mathcal{I}_{d/2-1}(\kappa)},$$

with the modified Bessel function of the first kind  $\mathcal{I}_{d/2-1}$  and a **schedule function**  $\kappa_t : [0, 1] \rightarrow [0, \infty)$  which is a continuously differentiable, strictly monotone increasing and  $\kappa_0 = 0$



**Theorem**  $\psi_t$  is the **unique**  $C[-1, 1]$  solution of the **linear ODE**

$$(1 - s^2) \psi'_t + \left( \kappa_t(1 - s^2) - (d - 1)s \right) \psi_t = (A_d(\kappa_t) - s) \dot{\kappa}_t, \quad s \in (-1, 1).$$

with **mean resultant length** of the vMF distribution is the Bessel ratio

$$A_d(\kappa) := \mathbb{E}_{x \sim \text{vMF}(w, \kappa)}[\langle w, x \rangle] = \frac{\mathcal{I}_{d/2}(\kappa)}{\mathcal{I}_{d/2-1}(\kappa)},$$

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

## 2. Power Spherical Distribution

$$p_t(\cdot \mid w) := C(\alpha_t) (1 + \langle w, \cdot \rangle)^{\alpha_t}$$

with

$$C(\alpha)^{-1} = 2^{d-1+\alpha} \pi^{(d-1)/2} \frac{\Gamma(\alpha + \frac{d-1}{2})}{\Gamma(\alpha + d - 1)},$$

**Theorem**  $\psi_t$  is the **unique**  $C[-1, 1]$  solution of the **linear ODE**

$$(1-s^2) \psi_t' + [\alpha_t(1-s) - (d-1)s] \psi_t = -\dot{\alpha}_t \left[ \log \frac{1+s}{2} + h(d-1+\alpha_t) - h\left(\frac{d-1}{2} + \alpha_t\right) \right],$$

where  $h = (\log \Gamma)'$  is the digamma function.

## 3. Spherical Cauchy Distribution

$$p_t(x \mid w) = \frac{1}{\omega_{d-1}} \frac{(1 - a_t^2)^{(d-1)/2}}{(1 - a_t \langle w, x \rangle)^{d-1}}$$

with

$$\omega_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$$

**Theorem** The unique  $C[-1, 1]$  solution  $\psi_t$  of the corresponding linear ODE is given by

$$\psi_t(s) = \dot{\beta}_t = \frac{\dot{a}_t}{1 - a_t^2}, \quad \beta_t := \log \frac{1 + a_t}{1 - a_t} \quad a_t := \tanh \beta_t$$

◆  $\psi_t$  does not depend on  $s$

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

Three step procedure:

1. Conditional flows on  $\mathcal{M}$
2. Conditional flows on  $\mathcal{M}^L$
3. Whole marginal flow on  $\mathcal{M}^L$

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

## 2. Conditional Flows on $\mathcal{M}^L$

|                               |    |
|-------------------------------|----|
| Technische Universität Berlin |    |
| 1                             | 2  |
| 3                             | 4  |
| 5                             | 6  |
| 7                             | 8  |
| 9                             | 10 |
| 11                            | 12 |
| 13                            | 14 |
| 15                            | 16 |
| 17                            | 18 |
| 19                            | 20 |
| 21                            | 22 |
| 23                            | 24 |
| 25                            | 26 |
| 27                            | 28 |
| 29                            | 30 |
| 31                            | 32 |
| 33                            | 34 |
| 35                            |    |

**Proposition** Let  $\mathbf{w} = (w^l)_{l=1}^L \in \mathcal{W}^L$  be fixed.

Let  $p_t^l(\cdot \mid w^l), v_t^l(\cdot \mid w^l), l = 1, \dots, L$  fulfill continuity equation on  $\mathcal{M}$ . Then

$$p_t(\mathbf{x} \mid \mathbf{w}) := \prod_{l=1}^L p_t^l(x^l \mid w^l), \quad \mathbf{x} = (x^1, \dots, x^L),$$

together with

$$v_t(\mathbf{x} \mid \mathbf{w}) := (v_t^1(x^1 \mid w^1), \dots, v_t^L(x^L \mid w^L))$$

satisfies continuity equation on  $\mathcal{M}^L$ .

Refs: 1. Markov chains: Holderrieth et al. 2025

2. Discrete flow matching: Lipman et al. 2024

3. General measures: Duong, Chemseddine, Friz, St. SIAM Reviews 2026

Three step procedure:

1. Conditional flows on  $\mathcal{M}$
2. Conditional flows on  $\mathcal{M}^L$
3. Whole marginal flow on  $\mathcal{M}^L$

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

|                               |    |
|-------------------------------|----|
| Technische Universität Berlin |    |
| 1                             | 2  |
| 3                             | 4  |
| 5                             | 6  |
| 7                             | 8  |
| 9                             | 10 |
| 11                            | 12 |
| 13                            | 14 |
| 15                            | 16 |
| 17                            | 18 |
| 19                            | 20 |
| 21                            | 22 |
| 23                            | 24 |
| 25                            | 26 |
| 27                            | 28 |
| 29                            | 30 |
| 31                            | 32 |
| 33                            | 34 |
| 35                            |    |

### 3. Whole Marginal Flows on $\mathcal{M}^L$

By the law of total probability the unconditional probability and vector field  $(p_t, v_t)$  fulfill the **continuity equation** on  $\mathcal{M}^L$ :

$$p_t(\mathbf{x}) = \sum_{\mathbf{w} \in \mathcal{W}^L} p_t(\mathbf{x} \mid \mathbf{w}) p_{\text{data}}(\mathbf{w})$$

$$v_t(\mathbf{x}) = \sum_{\mathbf{w} \in \mathcal{W}^L} v_t(\mathbf{x} \mid \mathbf{w}) p_t(\mathbf{w} \mid \mathbf{x})$$

By our special construction, the components  $v_t = (v_t^l)_{l=1}^L$  simplify to

$$\begin{aligned}
 v_t^l(\mathbf{x}) &= \sum_{\mathbf{w} \in \mathcal{W}^L} v_t^l(x^l \mid w^l) p_t(\mathbf{w} \mid \mathbf{x}) = \sum_{w^l \in \mathcal{W}} v_t^l(x^l \mid w^l) \sum_{\substack{\mathbf{w} \in \mathcal{W}^L \\ \mathbf{w}^l = w^l}} p_t(\mathbf{w} \mid \mathbf{x}) \\
 &= \sum_{w^l \in \mathcal{W}} v_t^l(x^l \mid w^l) \underbrace{p_t^l(w^l \mid \mathbf{x})}_{\text{to learn}}
 \end{aligned}$$

|                               |    |
|-------------------------------|----|
| Technische Universität Berlin |    |
| 1                             | 2  |
| 3                             | 4  |
| 5                             | 6  |
| 7                             | 8  |
| 9                             | 10 |
| 11                            | 12 |
| 13                            | 14 |
| 15                            | 16 |
| 17                            | 18 |
| 19                            | 20 |
| 21                            | 22 |
| 23                            | 24 |
| 25                            | 26 |
| 27                            | 28 |
| 29                            | 30 |
| 31                            | 32 |
| 33                            | 34 |
| 35                            |    |

## Kullback–Leibler divergence

$$\text{KL}(p||q) = \sum_{k=1}^N \left( p(k) \log p(k) \underbrace{- p(k) \log q(k)}_{\text{cross entropy (CE)}} \right),$$

Training:  $p_{\theta,t}$  with componentwise cross-entropy loss

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{l=1}^L \mathbb{E}_{t \sim \mathcal{U}(0,1), \mathbf{x}_t \sim p_t} \left[ \text{CE}(p_t^l(w^l | \mathbf{x}_t), p_{\theta,t}^l(w^l | \mathbf{x}_t)) \right] \\ &= - \sum_{l=1}^L \mathbb{E}_{t \sim \mathcal{U}(0,1), \mathbf{w} \sim p_{\text{data}}, \mathbf{x}_t \sim p_t(\cdot | \mathbf{w})} \left[ \log p_{t,\theta}^l(w^l | \mathbf{x}_t) \right] \end{aligned}$$

Here  $\mathbf{x}_t$  is accessible via

$$p_t(\mathbf{x} | \mathbf{w}) := \prod_{l=1}^L p_t^l(x^l | w^l), \quad \mathbf{x} = (x^1, \dots, x^L).$$

Backbone network:

$T_{\theta,t} : \mathcal{M}^L \rightarrow (\mathbb{R}^d)^L, \mathbf{x} \mapsto \hat{\mathbf{x}} = (\hat{x}^1, \dots, \hat{x}^L) := T_{\theta,t}(\mathbf{x})$  and define

$$p_{\theta,t}^l(w_k | \mathbf{x}) = p_{\theta}^l(w_k | \hat{x}^l) := \frac{\exp(s_k^l)}{\sum_{j=1}^N \exp(s_j^l)}, \quad s_k^l := \langle w_k, \hat{x}^l \rangle + b_k$$

Finally we get the velocity field for the ODE by

$$v_{\theta,t}^l(\mathbf{x}) = \sum_{w^l \in \mathcal{W}} v_t^l(x^l | w^l) p_{\theta,t}^l(w^l | \mathbf{x})$$

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

|                               |    |
|-------------------------------|----|
| Technische Universität Berlin |    |
| 1                             | 2  |
| 3                             | 4  |
| 5                             | 6  |
| 7                             | 8  |
| 9                             | 10 |
| 11                            | 12 |
| 13                            | 14 |
| 15                            | 16 |
| 17                            | 18 |
| 19                            | 20 |
| 21                            | 22 |
| 23                            | 24 |
| 25                            | 26 |
| 27                            | 28 |
| 29                            | 30 |
| 31                            | 32 |
| 33                            | 34 |
| 35                            |    |

1. **Flow ODE.** The associated flow ODE

$$\partial_t \Phi^l(t, \mathbf{x}) = v_{\theta,t}^l(\Phi(t, \mathbf{x})), \quad \Phi^l(0, \mathbf{x}) = x^l \sim \mathcal{U}(\mathbb{S}^{d-1}), \quad l = 1, \dots, L,$$

transports  $p_0 \sim \mathcal{U}(\mathbb{S}^{(d-1)L})$  to  $p_1 = p_{\text{data}}$ .

- ◆ equations are coupled through the posterior  $p_t^l(w_k|\mathbf{x})$ , which depends on the full state  $\mathbf{x} = (x^1, \dots, x^L)$

2. **Predictor-corrector.** The score provides a mechanism to do corrector steps during inference.

After advancing the state by one ODE step, we apply  $k$  **Langevin diffusion** steps:

The Langevin dynamics on  $(\mathbb{S}^{d-1})^L$  read per position as

$$\begin{aligned} dX_\tau^l &= \nabla_{\mathbb{S}^{d-1}, x^l} \log p_t(\mathbf{X}_\tau) d\tau + \sqrt{2} dB_\tau^l \\ &\approx \kappa_t \sum_{k=1}^N p_{\theta,t}^l(w_k|\mathbf{X}_\tau) P_{x^l}(w_k) d\tau + \sqrt{2} dB_\tau^l, \end{aligned}$$

where  $B_\tau^l$  is Brownian motion on  $\mathbb{S}^{d-1}$  at position  $l$ , independent across positions.

3. **Decoding.** At the terminal state  $\mathbf{x}_1$ , we pass through the backbone to obtain  $\hat{\mathbf{x}} = T_{\theta,1}(\mathbf{x}_1)$  and decode the token at each position  $l \in \{1, \dots, L\}$  as

$$\hat{y}^l := \arg \max_{k=1, \dots, N} s_k^l, \quad s_k^l = \langle w_k, \hat{x}^l \rangle + b_k.$$

## I. Flow Matching for Continuous Distributions

### I.1 Motivation

### I.2 Flow Matching

### I.3 Numerical Examples

## II. Flow Matching for Discrete Distributions

### II.1 Motivation

### II.2 Flow Matching

### II.3 Numerical Examples

1 2

3 4

5 6

7 8

9 10

11 12

13 14

15 16

17 18

19 20

21 22

23 24

25 26

27 28

29 30

31 32

33 34

35

Evaluation sets of vMF path on

- ◆ Sudoku
- ◆ One Billion Word Benchmark (LM1B)
- ◆ MATH Reasoning

Comparison:

1. vMF path
2. slerp geodesic interpolation
3. variance-preserving (VP) given by  $(X_t|W = w) = (1 - t)Z + tw$
4. variance-exploding (VE) given by  $(X_t|W = w) = w + \sigma_t Z$  with  $\sigma_t \in [0, \sigma_{\max}]$

ODE sampler:

- ◆ Euler integrator of the flow ODE (Flow ODE) with n predictor steps.
- ◆ predictor-corrector scheme (PC), in which each predictor step is followed by k Langevin corrector steps
- ◆ self-conditioning

Backbone: DiT-style transformer of Peebles & Xie (2023) with adaLN conditioning.

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

# Sudoku

| Method           | ODE          | PC          |
|------------------|--------------|-------------|
| vMF              | <b>65.84</b> | <b>74.5</b> |
| Geodesic         | 49.4         | -           |
| VP               | 53.4         | 69.1        |
| VE               | 48.0         | 46.5        |
| Masked diffusion | 35.1         |             |

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

Baseline: masked diffusion (MDM), Sahoo et al. 2024

Refs: For non-generative neural network approaches on Sudoku-Extreme, see  
Wang et al. 2025;

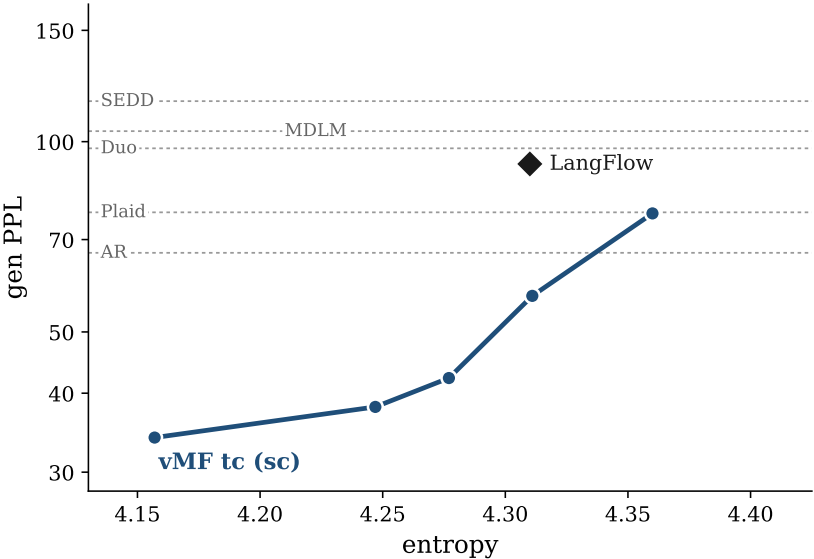
Jolicoeur-Martineau 2025;

[Kornhardt, Chemseddine, Wald, St.: Self-aware Markov-models for discrete reasoning, 2026](#); **99%  
correctness**

# One Billion Word Benchmark

◆ generation perplexity (PPL), entropy ( $H \sim 4.3$ )

| Path     | PPL ↓ | $H$  | ODE (sc)    | vMF                | PPL ↓       | $H$  |
|----------|-------|------|-------------|--------------------|-------------|------|
| vMF      | 171.8 | 4.34 | <b>85.9</b> | ODE(sc)            | 93.7        | 4.35 |
| Geodesic | 167.9 | 4.34 | 95.1        | $PC_{H \geq 4.30}$ | <b>57.0</b> | 4.31 |
| VP       | 140.4 | 4.34 | -           | $PC_{H \geq 4.25}$ | 39.0        | 4.26 |
| VE       | 149.7 | 4.34 | 140.1       | $PC_{H \geq 4.20}$ | 38.1        | 4.25 |



Reference baselines (Gen. PPL):

| AR Transf. | Plaid | LangFlow | FLM  | MDLM  | SEDD  |
|------------|-------|----------|------|-------|-------|
| 66.7       | 77.3  | 92.2     | 96.9 | 103.9 | 115.9 |

Refs: AR (sequentiell), Langflow Chen et al. 2026, Plaid Gulrajani & Hashimoto (2023), MDLM Sahoo et al. (2024), Duo Sahoo et al. (2025), SEDD Lou et al. (2024)

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

| Method              | ODE (%) | PC (%)      |
|---------------------|---------|-------------|
| AR (sample)         | 53.9    | –           |
| AR (greedy)         | 63.3    | –           |
| MDLM                | 18.0    | –           |
| DUO                 | 17.2    | –           |
| FLM                 | 0.3     | –           |
| S-FLM (top- $k=1$ ) | 18.0    | –           |
| vMF                 | 23.7    | <b>43.5</b> |
| Geodesic            | 14.9    | –           |
| VP                  | 18.4    | 28.5        |
| VE                  | 3.5     | 4.1         |

Reference first part: Deschenaux, Gulcehre 2026

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

## I. Flow Matching for Continuous Distributions

### I.1 Motivation

### I.2 Flow Matching

### ADD Diffusion Models

### I.3 Numerical Examples

## II. Flow Matching for Discrete Distributions

### II.1 Motivation

### II.2 Flow Matching

### II.3 Numerical Examples

## III. SOL: Measuring Gaps between Text Distributions by Double Sliced Wasserstein Metrics

1 2

3 4

5 6

7 8

9 10

11 12

13 14

15 16

17 18

19 20

21 22

23 24

25 26

27 28

29 30

31 32

33 34

35

A *causal decoder-only transformer*  $\mathbf{H}^\theta$  computes a hidden state at each position, which depends only on the preceding states, resulting in

$$\mathbf{H}^\theta : \mathcal{V}^{\leq L} \rightarrow \mathbb{R}^{d \times L}, \quad \mathbf{x} \mapsto (h_{\mathbf{x} \leq 1}^\theta, \dots, h_{\mathbf{x} \leq L}^\theta).$$

From the output layer, we get the probabilities of the next token recursively:

$$p_\theta(x_l \mid \mathbf{x}_{\leq l-1}) := \text{softmax}(W_\theta h_{\mathbf{x} \leq l-1}^\theta + b_\theta)$$

A transformer is used as an autoregressive generative model by factorizing the probability of a sequence  $\mathbf{x} = (x_1, \dots, x_L)$  as

$$p_\theta(\mathbf{x}) = \prod_{l=1}^L p_\theta(x_l \mid \mathbf{x}_{\leq l-1}),$$

and generating a new sequence is done by iteratively evaluating the model  $p_\theta$  and sampling from it.

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

The distribution of human-texts  $p_{\text{data}} \in \mathcal{P}(\mathcal{V}^{\leq L})$  is a distribution over token sequences.

Use the learned causal transformer  $\mathbf{H}^\theta$ , and introduce the associated *token-cloud distribution* by

$$T^\theta : \mathcal{V}^{\leq L} \rightarrow \mathcal{P}_2(\mathbb{R}^d), \quad \mathbf{x} \mapsto \frac{1}{|\mathbf{x}|} \sum_{l=1}^{|\mathbf{x}|} \delta_{h_{\mathbf{x} \leq l}^\theta}.$$

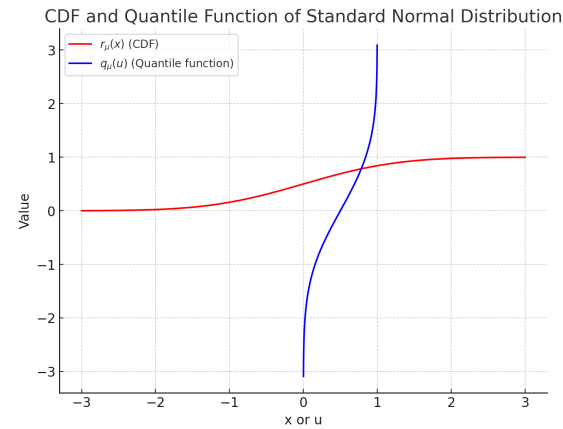
This allows us to push forward  $p_{\text{data}} \in \mathcal{P}(\mathcal{V}^{\leq L})$  to

$$T_{\#}^\theta p_{\text{data}} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$$

In general Wasserstein distances are hard to compute. An exception is  $\mathcal{P}_2(\mathbb{R})$ .

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

# Sliced Wasserstein Distance



For measures on the real line  $\mu \in \mathcal{P}_2(\mathbb{R})$ , using the isometric embedding

$$Q : \mathcal{P}_2(\mathbb{R}) \rightarrow L_2([0, 1]), \quad \mu \mapsto Q_\mu \quad \text{quantile function,}$$

the Wasserstein distance becomes

$$W_2^2(\mu, \nu) = W^2(\mu, \nu) = \int_0^1 |Q_\mu(s) - Q_\nu(s)|^2 ds, \quad \mu, \nu \in \mathcal{P}_2(\mathbb{R}).$$

For empirical measures

$$\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}, \quad x_i \in \mathbb{R}$$

the computation of  $W$  reduces to the sorting of real numbers in  $O(n \log n)$  operations.

1 2

3 4

5 6

7 8

9 10

11 12

13 14

15 16

17 18

19 20

21 22

23 24

25 26

27 28

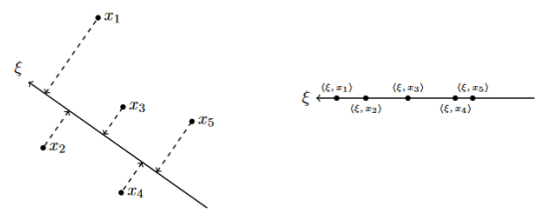
29 30

31 32

33 34

35

# Sliced Wasserstein Distance



|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |

|                               |    |
|-------------------------------|----|
| Technische Universität Berlin |    |
| 1                             | 2  |
| 3                             | 4  |
| 5                             | 6  |
| 7                             | 8  |
| 9                             | 10 |
| 11                            | 12 |
| 13                            | 14 |
| 15                            | 16 |
| 17                            | 18 |
| 19                            | 20 |
| 21                            | 22 |
| 23                            | 24 |
| 25                            | 26 |
| 27                            | 28 |
| 29                            | 30 |
| 31                            | 32 |
| 33                            | 34 |
| 35                            |    |

For  $\xi \in \mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \mid \|x\| = 1\}$  define the **slicing operator** and its push-forward by

$$S_\xi: \mathbb{R}^d \rightarrow \mathbb{R}, \quad x \mapsto \langle \xi, x \rangle,$$

$$\mathbf{e}_\xi: \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathcal{P}_2(\mathbb{R}), \quad \mu \mapsto S_{\xi,\#} \mu,$$

The **sliced Wasserstein-2 distance** reads as

$$SW^2(\mu, \nu) := \int_{\mathbb{S}^{d-1}} W^2(\mathbf{e}_\xi \mu, \mathbf{e}_\xi \nu) \, d\mathcal{U}(\mathbb{S}^{d-1})(\xi),$$

Apply  $\mathbf{e}_\xi: \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathcal{P}(\mathbb{R})$  and  $Q: \mathcal{P}_2(\mathbb{R}) \rightarrow L_2([0, 1])$  to map

$$Q \circ \mathbf{e}_\xi: \mathcal{P}_2(\mathbb{R}^d) \rightarrow L_2([0, 1]),$$

$$(Q \circ \mathbf{e}_\xi)_\#: \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)) \rightarrow \mathcal{P}_2(L_2([0, 1]))$$

# Double-Sliced Wasserstein distance

|                               |    |
|-------------------------------|----|
| Technische Universität Berlin |    |
| 1                             | 2  |
| 3                             | 4  |
| 5                             | 6  |
| 7                             | 8  |
| 9                             | 10 |
| 11                            | 12 |
| 13                            | 14 |
| 15                            | 16 |
| 17                            | 18 |
| 19                            | 20 |
| 21                            | 22 |
| 23                            | 24 |
| 25                            | 26 |
| 27                            | 28 |
| 29                            | 30 |
| 31                            | 32 |
| 33                            | 34 |
| 35                            |    |

For  $g \in L_2([0, 1])$ , define **slicing operator on  $\mathcal{P}_2(L_2[0, 1])$**  by

$$\begin{aligned} \mathcal{S}_g : L_2([0, 1]) &\rightarrow \mathbb{R}, & f &\mapsto \langle f, g \rangle_{L_2[0,1]}, \\ \mathcal{S}_{g,\#} : \mathcal{P}_2(L_2([0, 1])) &\rightarrow \mathcal{P}_2(\mathbb{R}), & \eta &\mapsto \mathcal{S}_{g,\#}\eta. \end{aligned}$$

The **sliced Wasserstein distance** on  $\mathcal{P}_2(L_2([0, 1]))$  is defined by

$$\text{SW}^2(\eta, \zeta \mid \gamma) := \int_{L_2} W^2(\mathcal{S}_{g,\#}\eta, \mathcal{S}_{g,\#}\zeta) \, d\gamma(g)$$

with Gaussian reference measure  $\gamma \in \mathcal{P}_2(L_2([0, 1]))$

◆ For realization, use the correspondence between Gaussian measures and processes

Combination: For  $\xi \in \mathbb{S}^{d-1}$  and  $g \in L_2([0, 1])$ , define

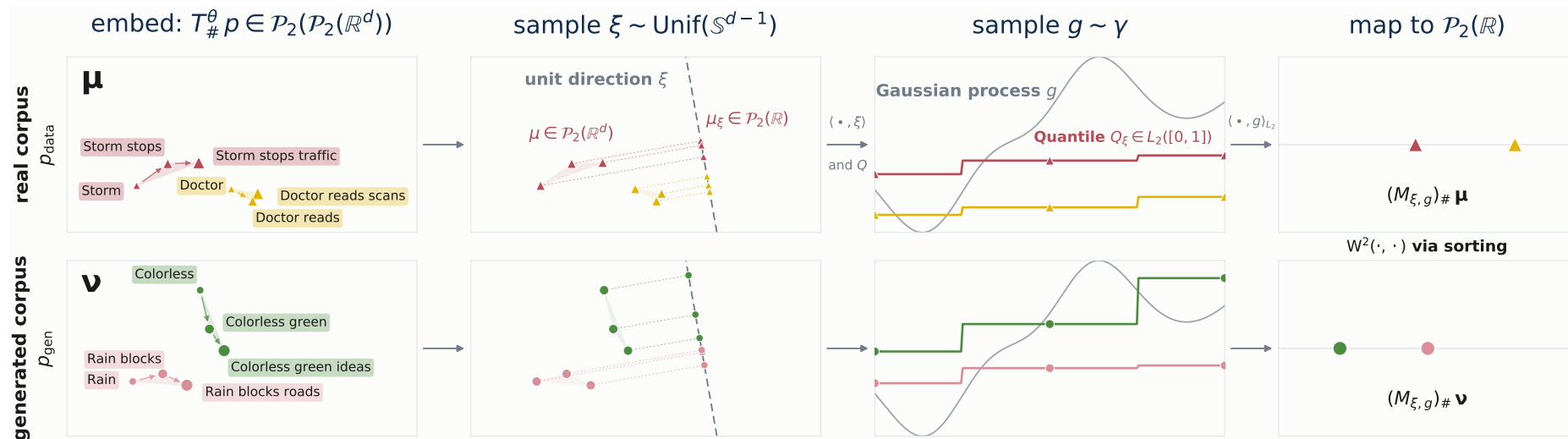
$$\begin{aligned} M_{\xi,g} : \mathcal{P}_2(\mathbb{R}^d) &\rightarrow \mathbb{R}, & \mu &\mapsto \left\langle Q_{\mathbf{e}_\xi \mu}, g \right\rangle_{L_2}, \\ (M_{\xi,g})_{\#} : \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)) &\rightarrow \mathcal{P}_2(\mathbb{R}), & \boldsymbol{\mu} &\mapsto (M_{\xi,g})_{\#}\boldsymbol{\mu} \end{aligned}$$

Now we can define the **double sliced Wasserstein distance** on meta-measures by

$$\begin{aligned} \text{DSW}^2(\boldsymbol{\mu}, \boldsymbol{\nu} \mid \gamma) &:= \int_{\mathbb{S}^{d-1}} \text{SW}^2((Q \circ \mathbf{e}_\xi)_{\#}\boldsymbol{\mu}, (Q \circ \mathbf{e}_\xi)_{\#}\boldsymbol{\nu} \mid \gamma) \, d\mathcal{U}(\mathbb{S}^{d-1})(\xi) \\ &= \mathbb{E}_{\xi \sim \mathcal{U}(\mathbb{S}^{d-1}), g \sim \gamma} \left[ W^2((M_{\xi,g})_{\#}\boldsymbol{\mu}, (M_{\xi,g})_{\#}\boldsymbol{\nu}) \right]. \end{aligned}$$

## New distance between text distributions

$$\text{SOTT}_\theta(p_{\text{data}}, p_{\text{gen}}) := \text{DSW}(T_{\#}^\theta p_{\text{data}}, T_{\#}^\theta p_{\text{gen}} \mid \gamma).$$



1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

| Model    | OLMo ↓       | Dream ↓      |
|----------|--------------|--------------|
| Real OWT | 0.75         | 1.97         |
| AR       | <b>7.12</b>  | <b>21.02</b> |
| DUO      | 11.42        | 29.65        |
| MDLM     | 13.57        | 31.83        |
| FLM      | 19.25        | 45.85        |
| vMF      | <b>10.31</b> | <b>29.10</b> |
| VP       | 10.68        | 30.13        |

SOTT of different models on OpenWebText using 5,000 samples

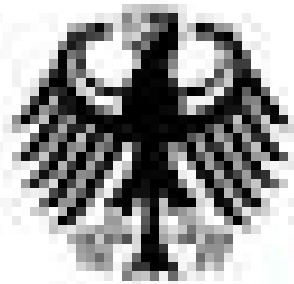
Many thanks for you attention!

|    |    |
|----|----|
| 1  | 2  |
| 3  | 4  |
| 5  | 6  |
| 7  | 8  |
| 9  | 10 |
| 11 | 12 |
| 13 | 14 |
| 15 | 16 |
| 17 | 18 |
| 19 | 20 |
| 21 | 22 |
| 23 | 24 |
| 25 | 26 |
| 27 | 28 |
| 29 | 30 |
| 31 | 32 |
| 33 | 34 |
| 35 |    |





Gefördert durch:



Bundesministerium  
für Forschung, Technologie  
und Raumfahrt



**Fellow**

**Ambassador**

Funded under Germany's Excellence Strategy by  
**DFG** Deutsche  
Forschungsgemeinschaft



Figure 1: Evolving point clouds and their continuum limit. Measure transportation provides a common framework for modeling algorithmic dynamics, statistical inference, and the evolution of physical and biological systems.



[CLS] ll be able to connect alongside these wind steps! [CLS] " not to see people understand anything that i ' m going to do. [CLS] the bulldogs fell to within 24 - 19 late in the second half. [CLS] the agency needed an alternate place to get the balance on the contact lists. [CLS] and at a

PROMPT:

A new program had 60 downloads in the first month. The number of downloads in the second month was three times as many as the downloads in the first month,

but then reduced by 30% in the third month. How many downloads did the program have total over the three months?

```
month1 = 60
month2 = month1 * 3
month3 = month2 * 0.7
total_downloads = month1 + month2 + month3
result = total_downloads
return result
```

t=0.02

t=0.16

t=0.30

t=0.44

t=0.56

t=0.70

t=0.84

t=0.98

final

7

2

2

2

2

2

2

2

2

?

?

?

8

8

8

8

8

8

:

9

9

9

9

9

9

9

9

.

.

.

.

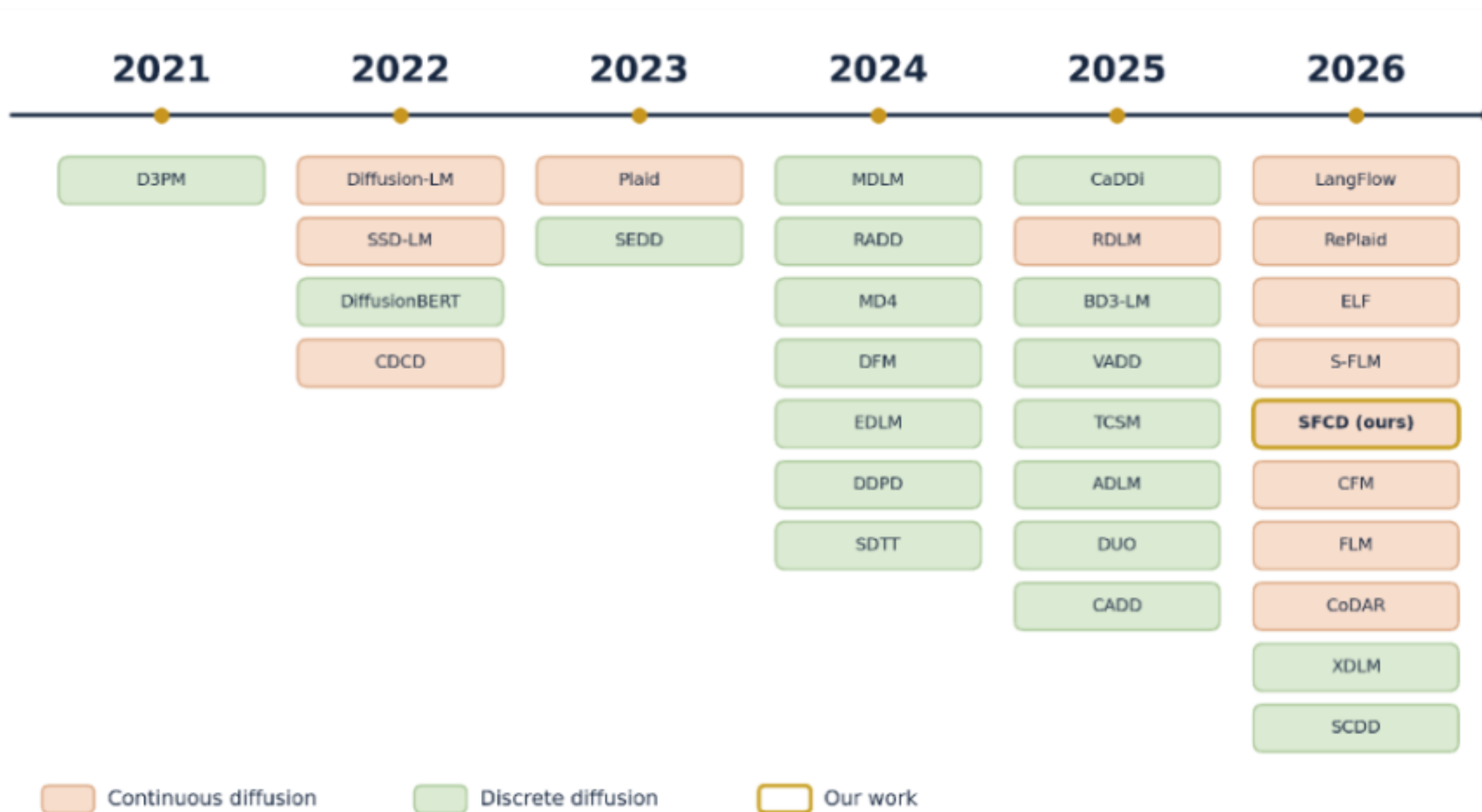
.

.

.

.

.



*A selection. Many more discrete diffusion papers exist.*

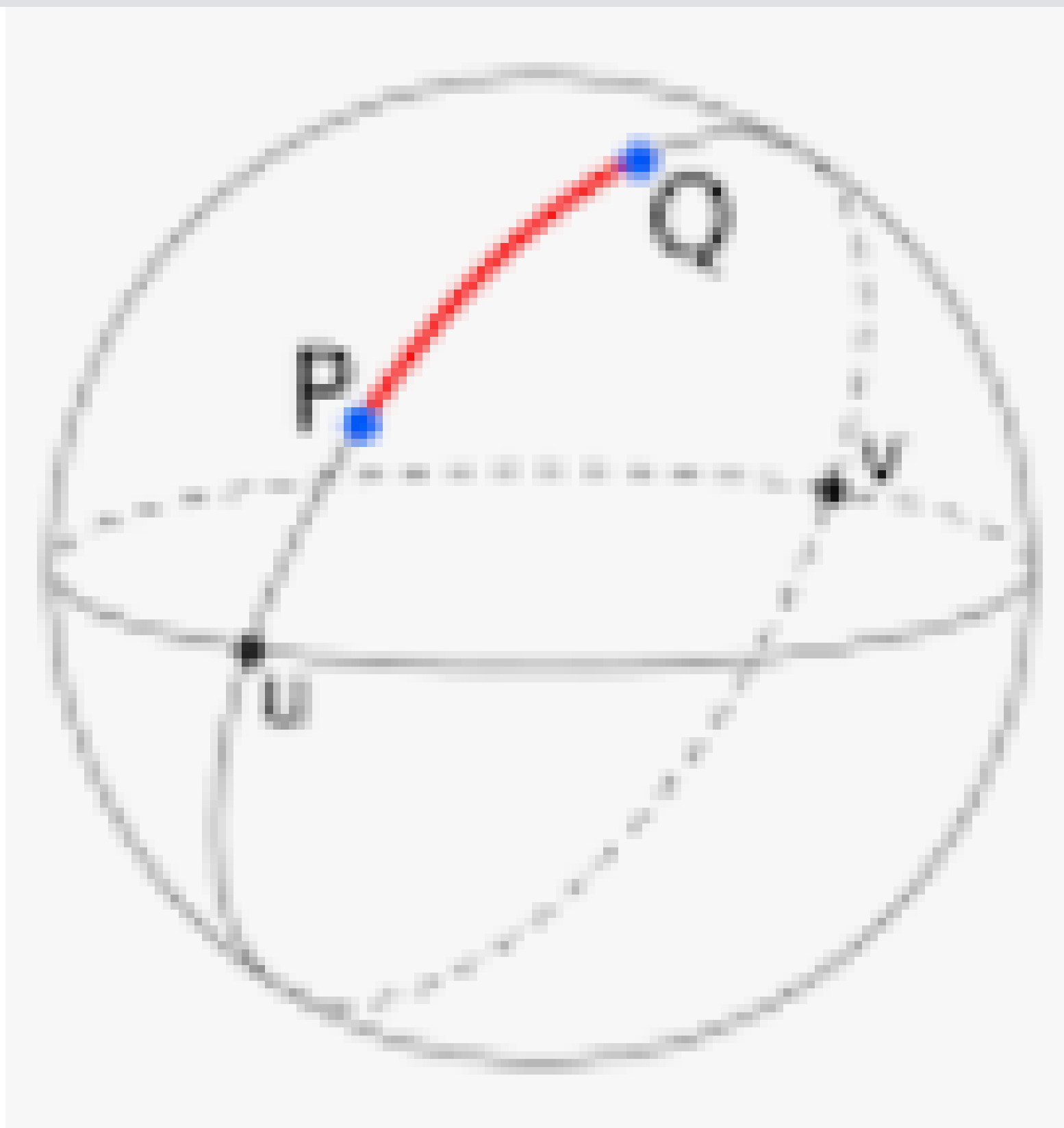
The

cat

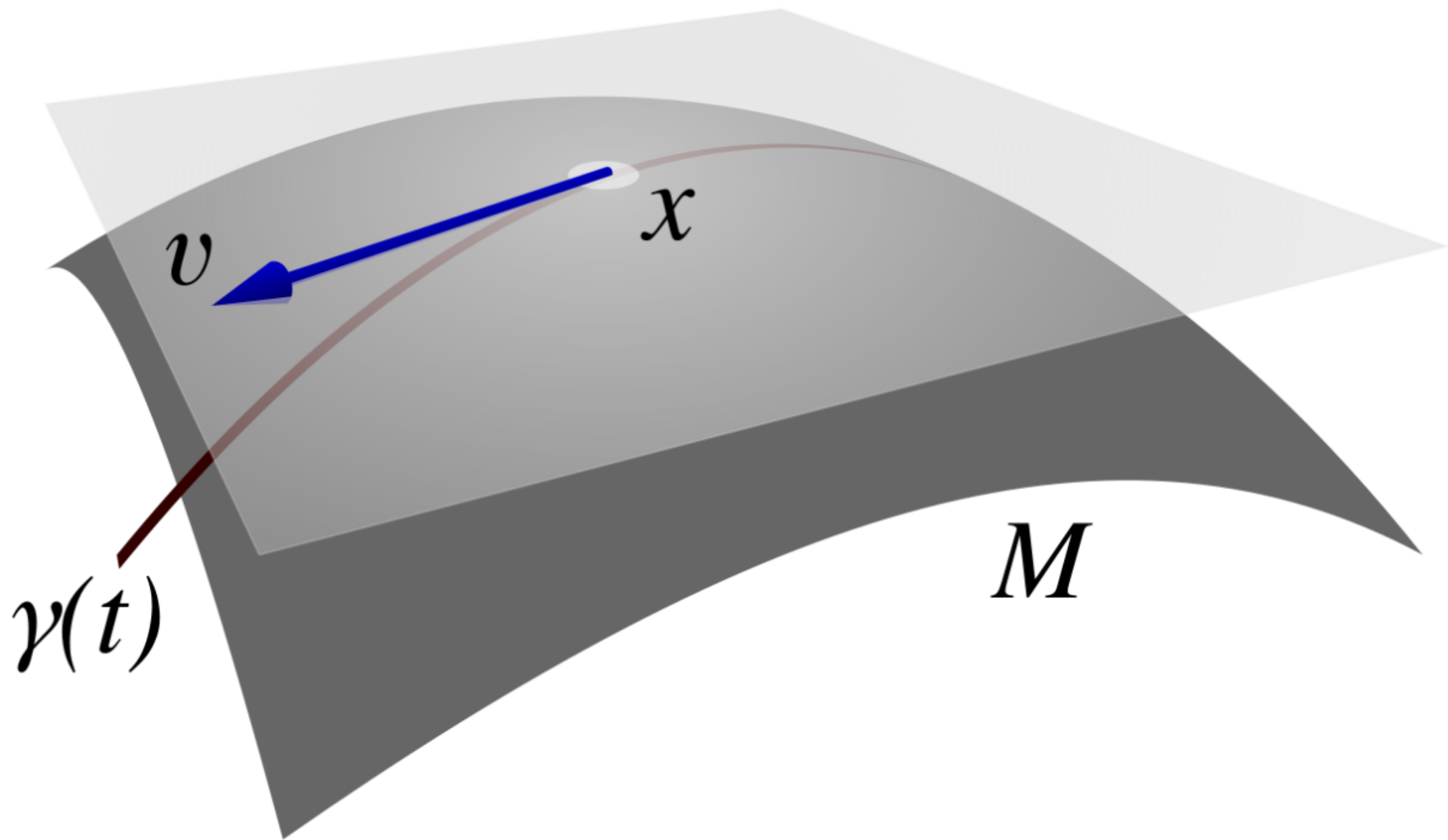
sat

Embed on  $\mathbb{S}^{d-1}$





$T_x M$

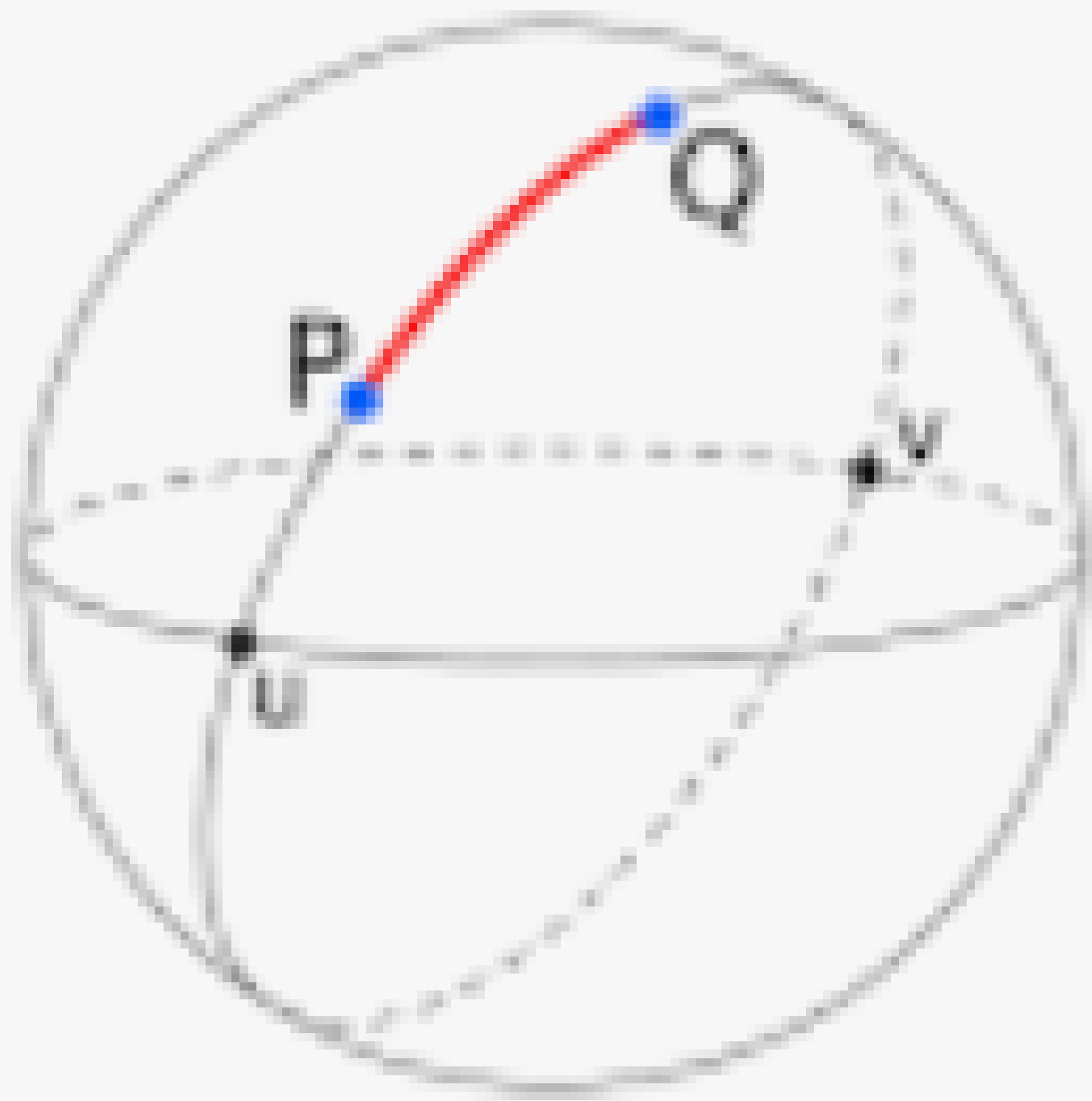


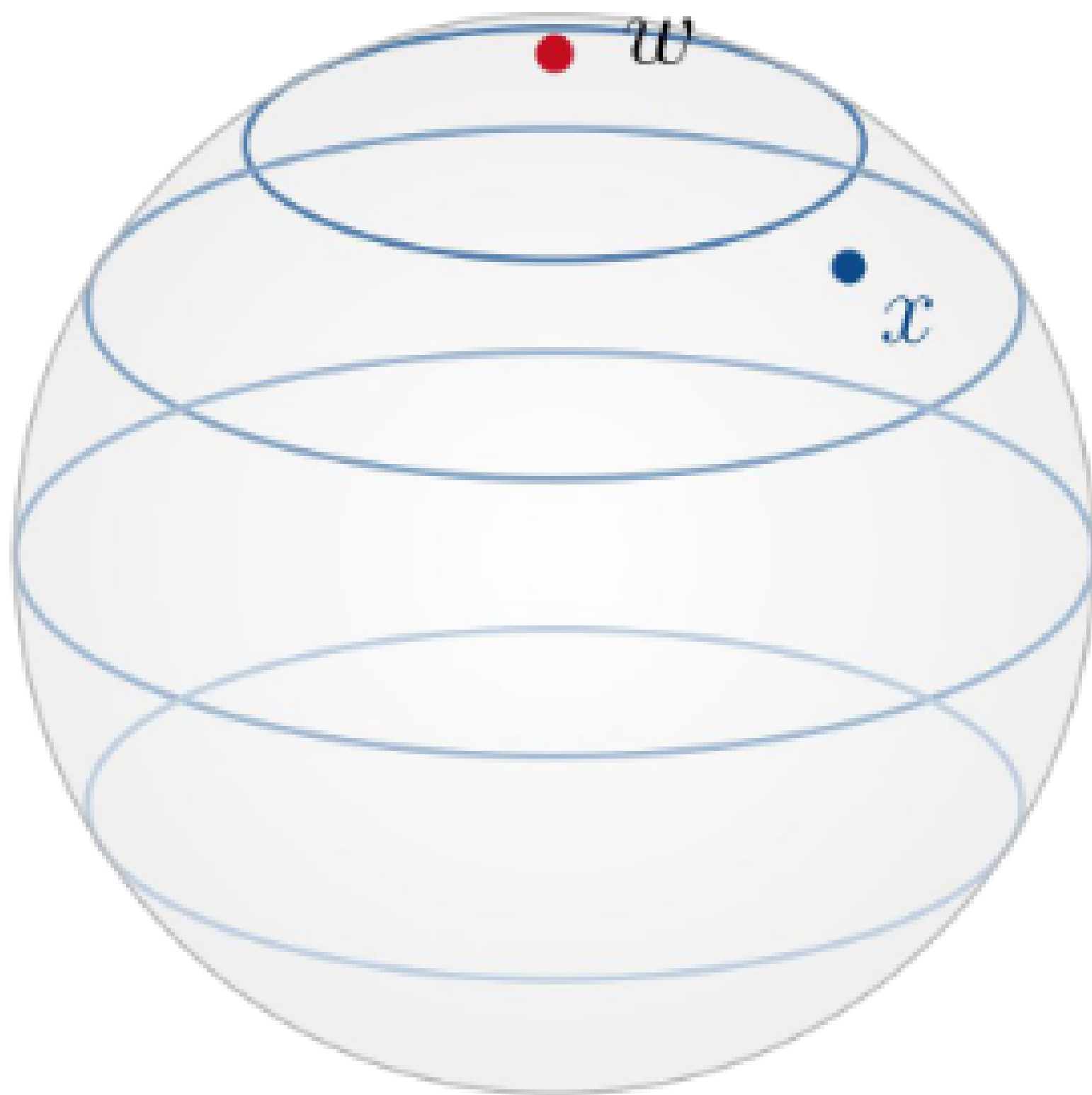
$M$

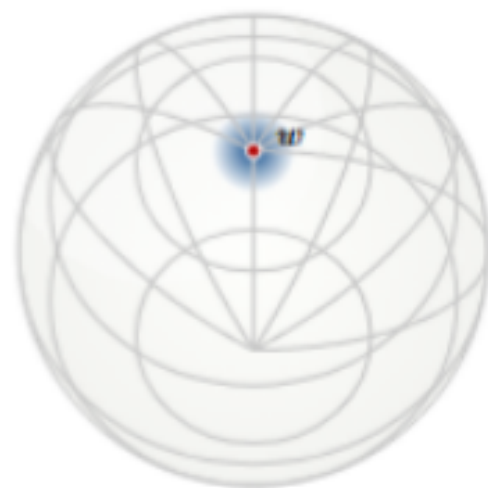
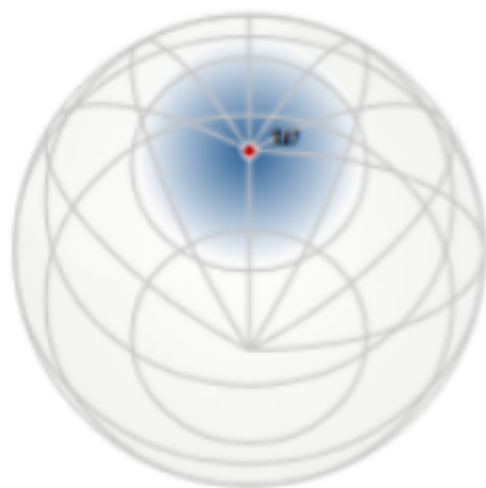
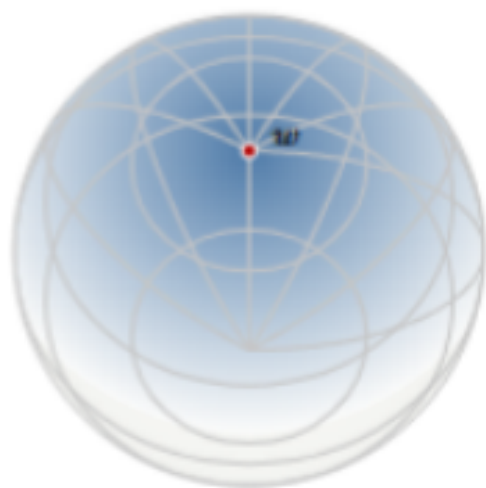
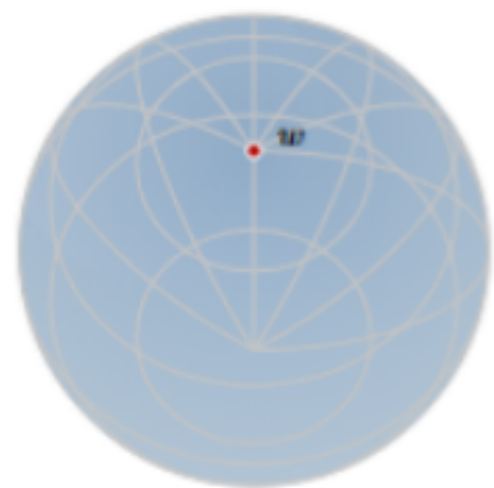
$\gamma(t)$

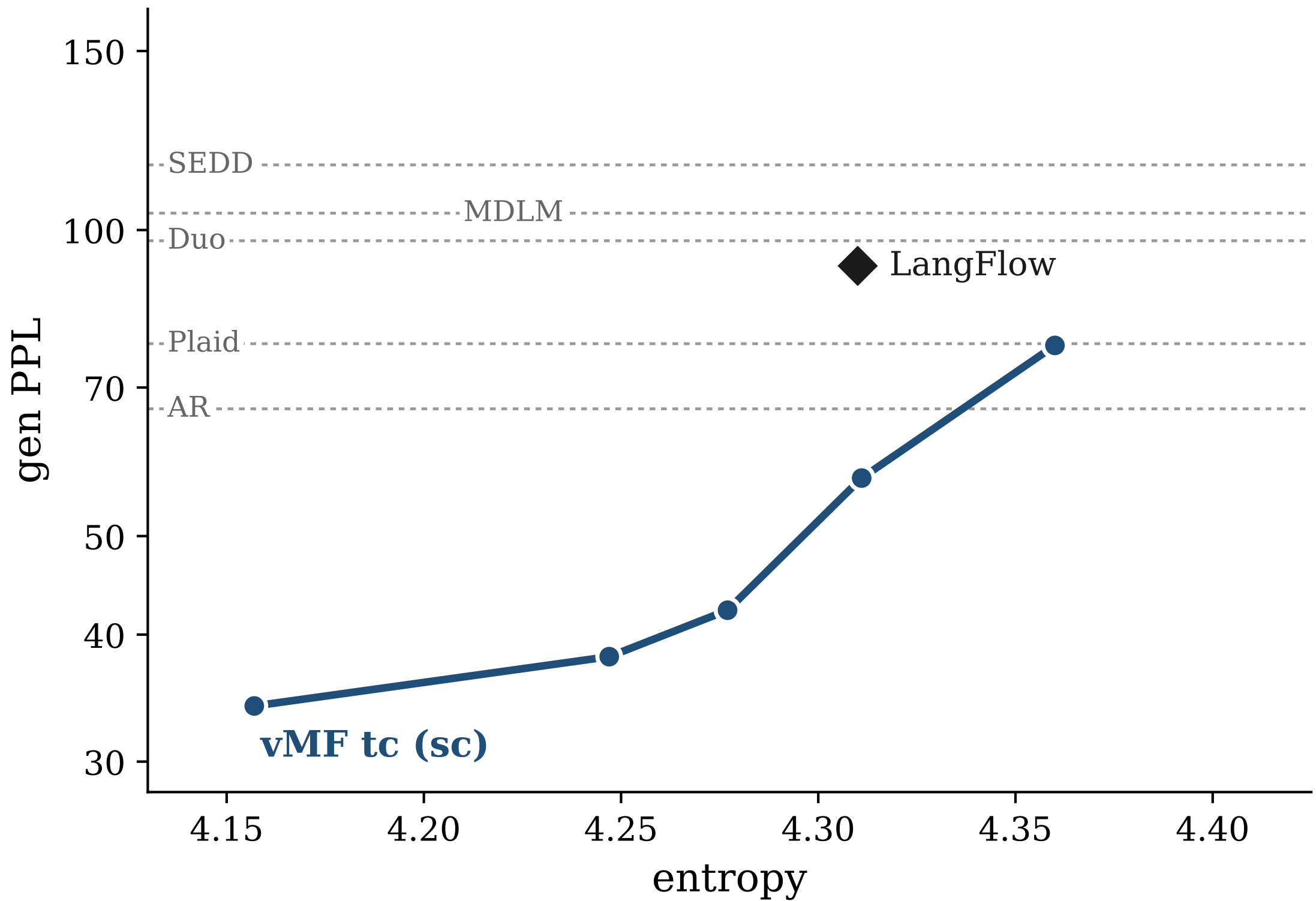
$v$

$x$





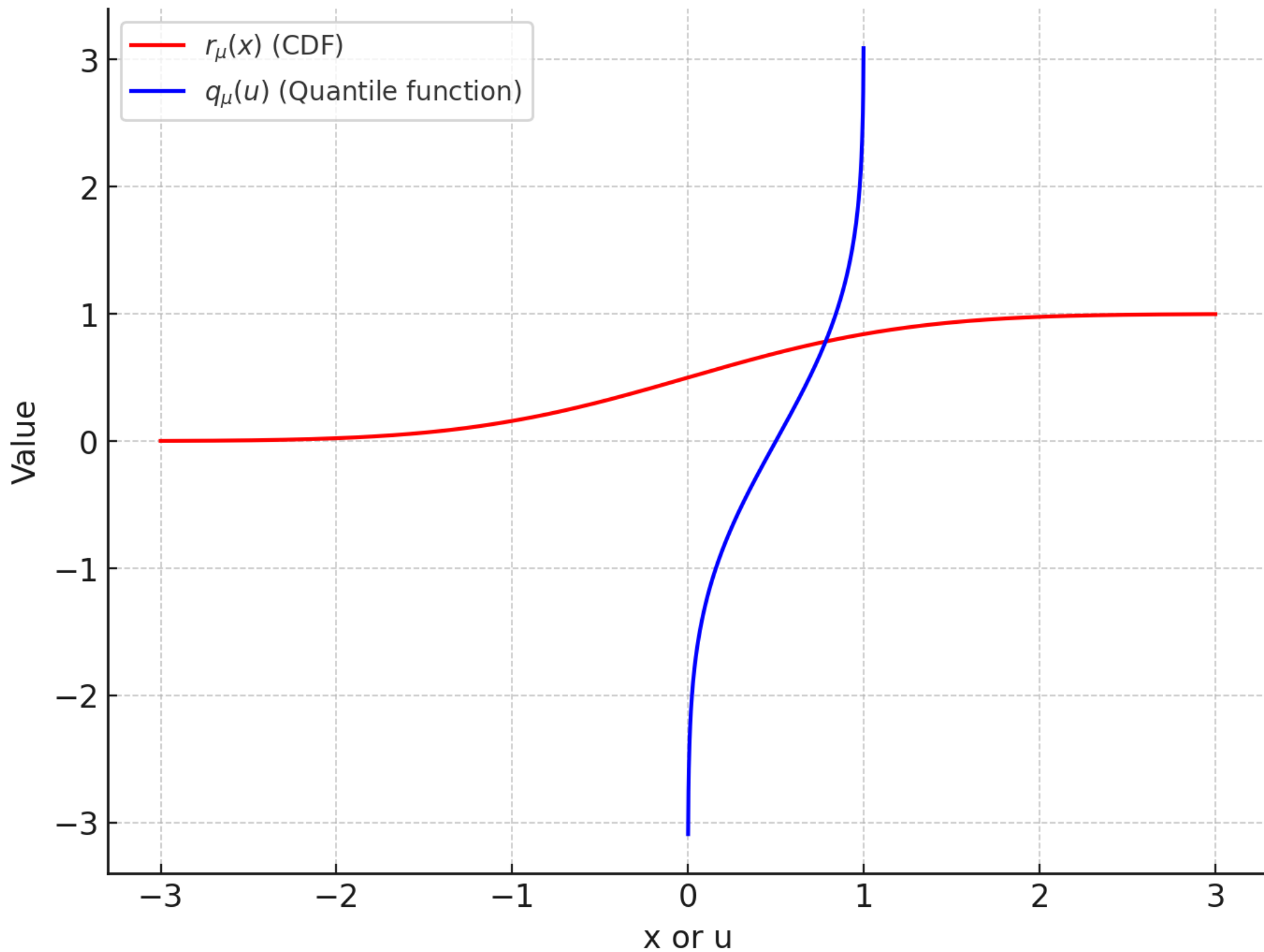


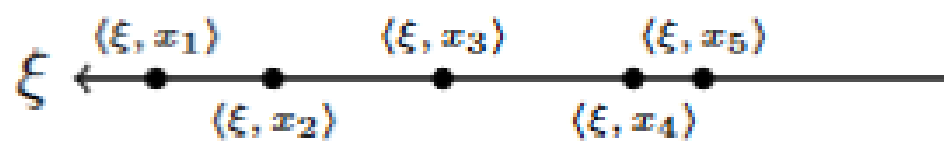
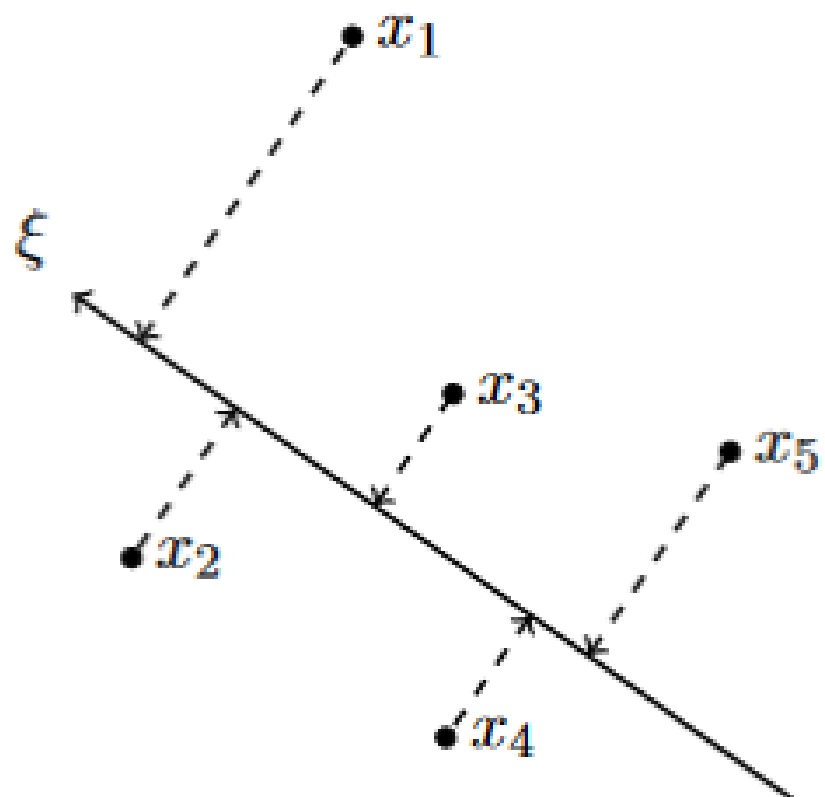


*Reference baselines (Gen. PPL):*

| AR Transf. | Plaid | LangFlow | FLM  | MDLM  | SEDD  |
|------------|-------|----------|------|-------|-------|
| 66.7       | 77.3  | 92.2     | 96.9 | 103.9 | 115.9 |

# CDF and Quantile Function of Standard Normal Distribution



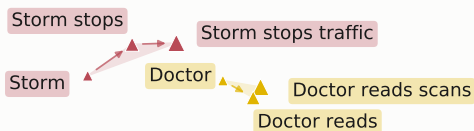


real corpus

$\rho_{\text{data}}$

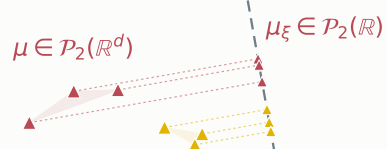
embed:  $T_{\#}^{\theta} p \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$

$\mu$



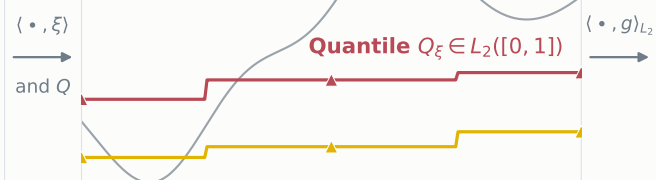
sample  $\xi \sim \text{Unif}(\mathbb{S}^{d-1})$

unit direction  $\xi$



sample  $g \sim \gamma$

Gaussian process  $g$



map to  $\mathcal{P}_2(\mathbb{R})$

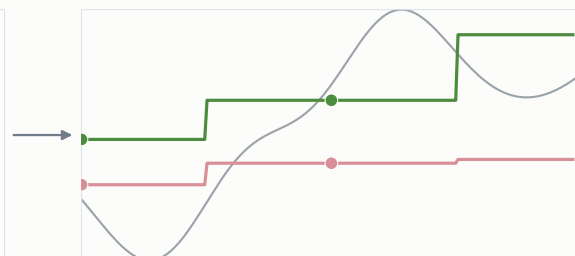
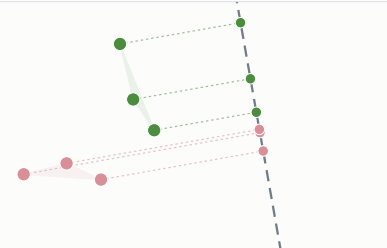
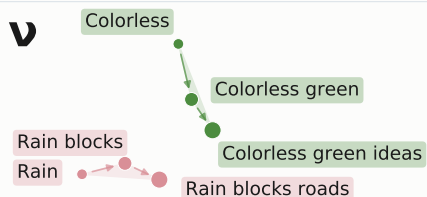
$(M_{\xi, g})_{\#} \mu$

$W^2(\cdot, \cdot)$  via sorting

generated corpus

$\rho_{\text{gen}}$

$\nu$



$(M_{\xi, g})_{\#} \nu$