Gabriele Steidl, Jannis Chemseddine, TU Berlin

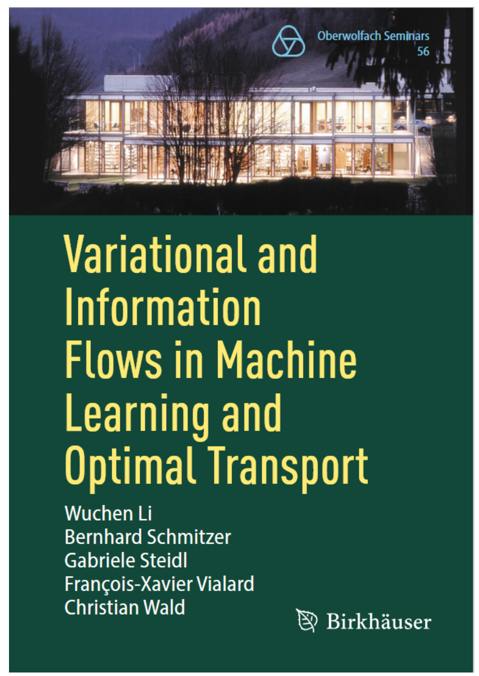
Lecture 1: Optimal Transport
Lecture 2: Generative Flows
Lecture 3: Bayesian Inverse Problems
Lecture 4: Experiments

Berlin Mathematics Research Center MATH+

Funded under Germany's Excellence Strategy by Deutsche Forschungsgemeinschaft

2. Generative Flows

1. Basic Idea of Generative Flows
2. Absolutely Continuous Curves in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$
3. Flow Matching
4. A Glimpse to Score-based Diffusion



Refs: Wald/St.: Flow Matching: Markov kernels, transport plans and stochastic processes, MFO Seminar Series, Springer 2025

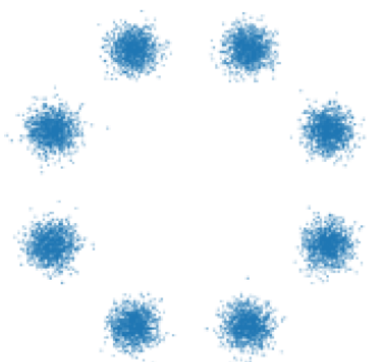
1. Basic Idea of Generative Flows

Aim: Sampling from a probability distribution μ_X (target distribution) given:

- some samples (oft many = data hungry)
- probability density function $p_X(x) = \exp(-\psi(x))/C$ (Boltzmann density) with or without normalizing factor C

It is only easy to sample from

- 1d distributions (cdf)
- $d \gg 1$ uniform or Gaussian "friendly" distributions μ_Z (latent distribution)



$d = 2$

$d = 28^2 = 784$

Refs: J. Chemseddine, C. Wald, R. Duong, St: Neural sampling from Boltzmann densities: Fisher-Rao curves in the Wasserstein geometry, ICLR 2025

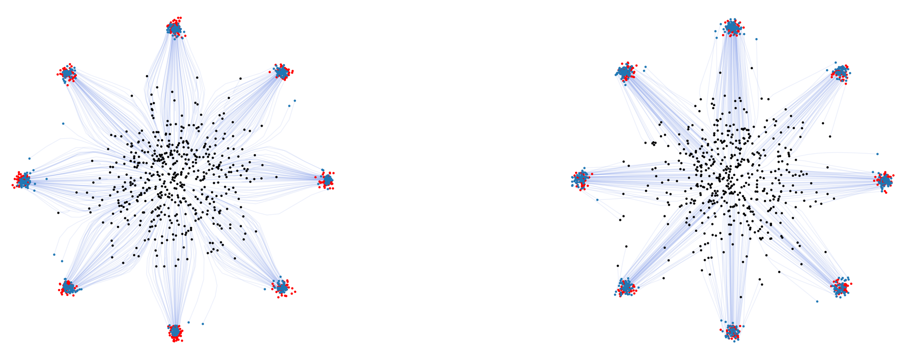
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53

Flow Based Generative Models

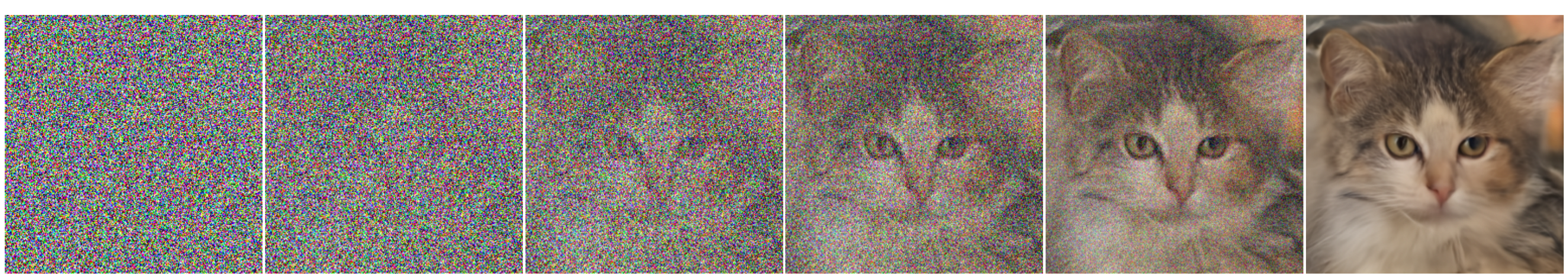
Idea: Learn a curve from μ_Z to μ_X

Basic methods:

- score-based diffusion models (Sohl-Dickstein 2015; Song/Ermon et al. 2019)
- flow matching (Lipman et al. 2022/23, Liu et al. 2022/23, Albergho et al. 2023)
- consistency models, e.g. inductive moment matching (Zhou et al. ICLM 2025, Boffi et al. 2025)



Trajectories of points via flow matching from different couplings, $d = 2$



Single trajectory for a flow matching model trained on cat images, $d = 256^2$

1. Basic Idea of Generative Flows
2. Absolutely Continuous Curves in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$
3. Flow Matching
4. A Glimpse to Score-based Diffusion

Absolutely Continuous Curves in Metric Spaces

In a complete metric space (X, d), a curve $\gamma : I \to X$ is called absolutely continuous ($AC^2(\mathbb{R}^d)$), if there exists as $g \in L_2(I)$ such that

$$\mathrm{d}(\gamma(s), \gamma(t)) \leq \int_s^t g(r)\,\mathrm{d}r \quad \text{for} \quad s \leq t,\ s, t \in I.$$

Then there exists a.e. the metric derivative (speed) which is indeed the smallest function g in the above inequality

$$|\gamma'|(t) := \lim_{s \to t} \frac{\mathrm{d}(\gamma(t), \gamma(s))}{t - s}.$$

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 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53

- A narrowly continuous curve $\mu_t : I = [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ absolutely continuous, if there exists a Borel measurable vector field $v : I \times \mathbb{R}^d \to \mathbb{R}^d$ with $\|v_t\|_{L_2(\mu_t)} \in L_2(I)$ such that (μ_t, v_t) satisfies the continuity equation (CE)

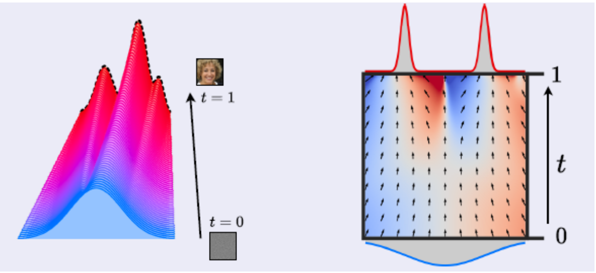
$$\partial_t \mu_t + \nabla_x \cdot (\mu_t v_t) = 0$$

- If μ_t has a smooth density p_t, then

$$\partial_t p_t + \nabla_x \cdot (p_t v_t) = 0$$

- otherwise (CE) is meant in a weak sense, i.e. for all $\varphi \in C_c^\infty((a,b) \times \mathbb{R}^d)$:

$$\int_I \int_{\mathbb{R}^d} \partial_t \varphi \, \mathrm{d}\mu(t) \mathrm{d}t + \int_I \int_{\mathbb{R}^d} \langle \nabla_x \varphi, v \rangle \, \mathrm{d}\mu_t(x) \mathrm{d}t = 0$$



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 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53

Absolutely Continuous Curves in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

- Let (μ_t, v_t) fulfill (CE) and

$$\int_I \sup_{x\in B} \|v_t(x)\| + \mathrm{Lip}(v_t, B)\,\mathrm{d}t < \infty \quad \forall \text{compact } B \subset \mathcal{B}(\mathbb{R}^d).$$

Then there exists a solution $\phi : I \times \mathbb{R}^d \to \mathbb{R}^d$ of the ODE

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

and $\phi(t,\cdot)_\sharp \mu_0 = \mu_0 \circ \phi(t,\cdot)^{-1} = \mu_t$.

Also the other direction is true under some conditions.

Assume that ϕ satisfies the ODE for some locally bounded vector field v_t on $I = (a, b)$. For $\varphi \in C_c^\infty((a, b) \times \mathbb{R}^d)$, we can compute

$$\begin{aligned}
0 &= \varphi(b, \phi(b, x)) - \varphi(a, \phi(a, x)) \\
&= \int_{\mathbb{R}^d} \varphi(b, \phi(b, x)) - \varphi(a, \phi(a, x)) \mathrm{d}\mu_0 \\
&= \int_{\mathbb{R}^d} \int_a^b \frac{\mathrm{d}}{\mathrm{d}t}(\varphi(t, \phi(t, x))) \, \mathrm{d}t \mathrm{d}\mu_0 \\
&= \int_{\mathbb{R}^d} \int_a^b \langle \nabla_x \varphi\big(t, \phi(t, x)\big), v_t\big(t, \phi(t, x)\big)\rangle + (\partial_t \varphi)\big(t, \phi(t, x)\big) \, \mathrm{d}t \mathrm{d}\mu_0 \\
&= \int_a^b \int_{\mathbb{R}^d} \langle \nabla_x \varphi, v_t\rangle + \partial_t \varphi \, \mathrm{d}[\phi(t, \cdot)_\sharp \mu_0] \mathrm{d}t
\end{aligned}$$

Thus $\mu_t := \phi(t, \cdot)_\sharp \mu_0$ satisfies (CE) (in a weak sense).

Summary

- A narrowly continuous curve $\mu_t : I = [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ absolutely continuous, if there exists a Borel measurable vector field $v : I \times \mathbb{R}^d \to \mathbb{R}^d$ with $\|v_t\|_{L_2(\mu_t)} \in L_2(I)$ such that (μ_t, v_t) satisfies the continuity equation (CE)

$$\partial_t \mu_t + \nabla_x \cdot (\mu_t v_t) = 0$$

- Let (μ_t, v_t) as above and

$$\int_I \sup_{x \in B} \|v_t(x)\| + \mathrm{Lip}(v_t, B)\,\mathrm{d}t < \infty \quad \forall \text{compact } B \subset \mathcal{B}(\mathbb{R}^d).$$

  Then there exists a solution $\phi : I \times \mathbb{R}^d \to \mathbb{R}^d$ of the ODE

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

  and $\phi(t,\cdot)_\sharp \mu_0 = \mu_0 \circ \phi(t,\cdot)^{-1} = \mu_t$.
- If a nice person gifts you such a velocity field v_t, then use your favorite ODE solver to transport samples from $\mu_0 = \mu_X$ to samples of $\mu_1 = \mu_Z$ or backwards using $-v_{1-t}$.

All you need is v_t !

2. Generative Flows

How to Learn a Velocity Field?

Task: Learn/approximate v_t by neural network v_t^θ!

Loss function for learning:

$$\mathcal{L}(\theta) := \mathbb{E}_{t\sim\mathcal{U}(0,T),x\sim\mu_t}\Big[\left\|v_t^\theta(x) - v_t(x)\right\|^2\Big],$$

not accessible 😢.

Idea: Make it conditional since up to a constant

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,T),\, y\sim\mu_1,\, x\sim\mu_t(\cdot|y)}\left[\left\|v_t^\theta(x) - v_t^y(x)\right\|^2\right]$$

🙂 We will see that conditional velocity fields v_t^y belonging to a conditional flow $\mu_t(\cdot|y)$ starting in δ_y are available in an analytic form in certain cases e.g. from couplings of μ_X and μ_Z (optimal transport)

All you need is conditional v_t^y !

Rewriting the Loss Function

Random variables: $X_0 \sim P_{X_0} = \mu_0$, $X_1 \sim P_{X_1} = \mu_1$, $X_t \sim P_{X_t} = \mu_t$, $\mu_t(\cdot|y) = P_{X_t|X_1=y}$

$$P_{X_t,X_1} = P_{X_t|X_1=y} \times_y P_{X_1} = P_{X_1|X_t=x} \times_x P_{X_t}$$

Theorem. Let $(\mu_t(\cdot|y), v_t^y)$ satisfy the (CE), then the following (μ_t, v_t) also satisfy the (CE):

$$\mu_t = \pi^1_\#(\mu_t(\cdot|y) \times_y \mu_1), \qquad p_t(x) = \int p_t^y(x) \mathrm{d}\mu_1(y) \tag{1}$$

$$v_t(x) = \int_{\mathbb{R}^d} v_t^y(x)\, \mathrm{d}P_{X_1|X_t=x}(y) \tag{2}$$

Rewriting the loss function:

$$\mathcal{L}(\theta) := \mathbb{E}_{t\sim\mathcal{U}(0,T),x\sim\mu_t}\left[\left\|v_t^\theta(x) - v_t(x)\right\|^2\right] = \mathbb{E}\left[\|v_t^\theta\|^2\right] - 2\mathbb{E}\left[\langle v_t, v_t^\theta\rangle\right] + \underbrace{\mathbb{E}\left[\|v_t\|^2\right]}_{\text{const}}$$

Consider first and second term separately.

1.

$$
\begin{aligned}
\mathbb{E}\left[\|v_t^\theta(x)\|^2\right] &= \int_0^1 \int_{\mathbb{R}^d} \|v_t^\theta(x)\|^2 \mathrm{d}\mu_t(x)\mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|v_t^\theta(x)\|^2 \mathrm{d}P_{X_1|X_t=x}(y)\mathrm{d}P_{X_t}(x)\mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d\times\mathbb{R}^d} \|v_t^\theta(x)\|^2 \mathrm{d}P_{X_t,X_1}(x,y)\mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d\times\mathbb{R}^d} \|v_t^\theta(x)\|^2 \mathrm{d}P_{X_t|X_1=y}(x)\mathrm{d}P_{X_1}(y)\mathrm{d}t \\
&= \mathbb{E}_{t\sim\mathcal{U}(0,T),\, {\color{red}y}\sim\mu_1,\, x\sim{\color{red}\mu_t(\cdot|y)}}\left[\|v_t^\theta(x))\|^2\right]
\end{aligned}
$$

2.

$$
\begin{aligned}
\mathbb{E}_{(t,x)\sim\mu_t\times_t\mathrm{d}t}\left[\langle v_t^\theta(x), v_t(x)\rangle\right] &= \int_0^1 \int_{\mathbb{R}^d} \left\langle v_t^\theta(x), \int_{\mathbb{R}^d} v_t^y(x)\mathrm{d}P_{X_1|X_t=x}(y)\right\rangle \mathrm{d}P_{X_t}(x)\mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle v_t^\theta(x), v_t^y(x)\rangle \mathrm{d}P_{X_t,X_1}(x,y)\mathrm{d}t \\
&= \mathbb{E}_{t\sim\mathcal{U}(0,T),\, {\color{red}y}\sim\mu_1,\, x\sim{\color{red}\mu_t(\cdot|y)}}\left[\langle v_t^\theta(x), v_t^y(x)\rangle\right]
\end{aligned}
$$

How we can conditioned velocity fields?

- from couplings/linear processes
- from special processes:

  Duong, Chemseddine, Friz, Steidl: Telegrapher's generative model via Kac flows

  Chemseddine, Kornhardt, Duong, Steidl: Adapting noise to data: generative flows from 1D processes

Conditioned velocity fields from couplings

Curve ends: $\mu_0 = \mu_Z$, $\mu_1 = \mu_X$

Arbitrary coupling: $\alpha \in \Pi(\mu_0, \mu_1)$, $(X_0, X_1) \sim \alpha$

Interpolation map: $e_t(x, y) = (1-t)x + ty$, $t \in [0, 1]$

Theorem The pair (μ_t, v_t) induced by α, i.e.

$$\mu_t := e_{t,\sharp}\alpha \qquad X_t = (1-t)X_0 + tX_1 \sim \mu_t$$

$$v_t\mu_t := e_{t,\sharp}[(y-x)\alpha] \quad \text{for a.e. } t \in [0,1].$$

satisfies the (CE). Further,

$$\mu_t^y = e_{t,\sharp}(P_{X_t|X_1=y} \times \delta_y),$$

$$v_t^y = \frac{y-x}{1-t}$$

fulfill the (CE) and are related to μ_t and v_t by (1) and (2), resp.

Example

Example: Let $\alpha = \mu_0 \times \mu_1$ and $\mu_0 = \mathcal{N}(0, I_d)$. Then (μ_t, v_t) above is given by $\mu_t = p_t \mathrm{d}x$:

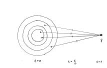
$$p_t(x) = C_t \int_{\mathbb{R}^d} \mathrm{e}^{-\frac{\|x-ty\|^2}{2(1-t)^2}} \mathrm{d}\mu_1(y) \quad \text{with} \quad C_t := \left(2\pi(1-t)^2\right)^{-\frac{d}{2}}$$

$$v_t = \frac{1-t}{t} \underbrace{\nabla \log p_t}_{\text{score}} + \frac{x}{t}$$

and thus for $\mu_1 = \delta_y$,

$$p_t(x|y) = C_t \, \mathrm{e}^{-\frac{\|x-ty\|^2}{2(1-t)^2}}$$

$$v_t^y(x) = \frac{y-x}{1-t}$$



and $\mu_t(\cdot|y) = (e_t)_\sharp \alpha = T^y_{t,\sharp}\mu_0 \quad$ with $\quad T_t^y(x) := (1-t)x + ty$

ODE with $T_t^y = \phi_t$:

$$\partial_t T_t^y(x) = v_t^y(T_t^y(x)) = \frac{y - ((1-t)x + ty)}{1-t} = y - x$$

Example

1. For $\mu_t := e_{t,\sharp}\alpha$ we get with $x := (z - yt)/(1-t)$:

$$\int \varphi \,\mathrm{d}\mu_t = \int \varphi \,\mathrm{d}e_{t,\sharp}\alpha = \int \varphi\left(e_t(x,y)\right)\,\mathrm{d}\alpha(x,y)$$

$$= (2\pi)^{-\frac{d}{2}} \int \varphi\left((1-t)x + ty\right) \mathrm{e}^{-\frac{\|x\|^2}{2}} \,\mathrm{d}x\mathrm{d}\mu_1(y),$$

$$= \left(2\pi(1-t)^2\right)^{-\frac{d}{2}} \int \varphi(z)\mathrm{e}^{-\frac{\|z-ty\|^2}{2(1-t)^2}} \,\mathrm{d}\mu_1(y)\mathrm{d}z.$$

2. Further we verify for $f_t := \frac{1-t}{t}\log p_t + \frac{1}{2t}\|\cdot\|^2$ and $\mu_t = p_t \mathrm{d}x$ that

$$\begin{aligned}
\int_{\mathbb{R}^d} \varphi\,\nabla f_t\,\mathrm{d}\mu_t &= \tfrac{1-t}{t}\int_{\mathbb{R}^d}\varphi(x)\frac{\nabla p_t(x)}{p_t(x)}p_t(x)\,\mathrm{d}x + \frac{1}{t}\int_{\mathbb{R}^d}\varphi(x)\,x\,p_t(x)\,\mathrm{d}x\\
&= \frac{1}{t(1-t)^{d+1}(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d\times\mathbb{R}^d}\varphi(x)(ty-x)\mathrm{e}^{-\frac{\|x-ty\|^2}{2(1-t)^2}}\,\mathrm{d}\mu_1(y)\,\mathrm{d}x\\
&\quad + \frac{1}{t(1-t)^{d}(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d\times\mathbb{R}^d}\varphi(x)x\mathrm{e}^{-\frac{\|x-ty\|^2}{2(1-t)^2}}\,\mathrm{d}\mu_1(y)\mathrm{d}x\\
&= \frac{1}{(1-t)^{d+1}(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d\times\mathbb{R}^d}\varphi(x)(y-x)\mathrm{e}^{-\frac{\|x-ty\|^2}{2(1-t)^2}}\,\mathrm{d}\mu_1(y)\,\mathrm{d}x.
\end{aligned}$$

On the other hand

$$\begin{aligned}
\int_{\mathbb{R}^d}\varphi\,v_t\,\mathrm{d}\mu_t &= \int_{\mathbb{R}^d}\varphi\,\mathrm{d}e_{t,\sharp}(y-x)\alpha\\
&= \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d\times\mathbb{R}^d}\varphi(e_t(x,y))(y-x)\mathrm{e}^{-\frac{\|x\|^2}{2}}\,\mathrm{d}\mu_1(y)\mathrm{d}x\\
&= \frac{1}{(1-t)^{d+1}(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d\times\mathbb{R}^d}\varphi(z)(y-z)\mathrm{e}^{-\frac{\|z-ty\|^2}{2(1-t)^2}}\,\mathrm{d}\mu_1(y)\mathrm{d}z
\end{aligned}$$

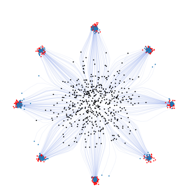
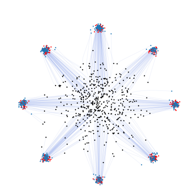
Loss function

Recall $\mu_t = P_{X_t} = (e_t)_\# \mu_0 = (e_t)_\# P_{X_0}$, $\alpha = P_{X_0,X_1}$, $v_t^y = \frac{y-x}{1-t}$

$$
\begin{aligned}
\mathcal{L}(\theta) &:= \mathbb{E}_{t\sim\mathcal{U}(0,T),\, y\sim\mu_1,\, x\sim\mu_t(\cdot|y)} \left[\left\| v_t^\theta(x) - v_t^y(x) \right\|^2 \right] \\
&= \int_0^1 \int_{\mathbb{R}^d\times\mathbb{R}^d} \left\| v_t^\theta(x) - v_t^y(x) \right\|^2 \mathrm{d}P_{X_t|X_1=y}(x)\mathrm{d}P_{X_1}(y)\mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d\times\mathbb{R}^d} \left\| v_t^\theta(x) - v_t^y(x) \right\|^2 \mathrm{d}P_{X_t,X_1}(x,y)\mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d\times\mathbb{R}^d} \left\| v_t^\theta(e_t(x,y)) - v_t^y(e_t(x,y)) \right\|^2 \mathrm{d}P_{X_0,X_1}(x,y)\mathrm{d}t \\
&= \mathbb{E}_{t\sim\mathcal{U}(0,T),\, (x,y)\sim\alpha}\left[\| v_t^\theta(e_t(x,y)) - (y-x)\|^2 \right]
\end{aligned}
$$

All you need is a coupling α !

- independent coupling: $\alpha = \mu_0 \times \mu_1$
- optimal coupling: α minimizer of $W_2^2(\mu_0,\mu_1)$ with advantage that $\alpha = (\mathrm{Id}, T)_\sharp \mu_0$ is induced (often) by a Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$ and we have straight trajectories $\phi_t(x) = T_t(x) := (1-t)x + tT(x)$

Let $X_t \in L^2(\Omega, \mathbb{R}^d, \mathbb{P})$ be continuously differentiable in every $t \in [0,1]$. Let

$$\mu_t := X_{t,\sharp}\mathbb{P} = P_{X_t}, \quad \text{and} \quad v_t := \mathbb{E}[\partial_t X_t | X_t = \cdot].$$

- (μ_t, v_t) fulfills the continuity equation

Example: Let $X_t = (1-t)X_0 + tX_1$ for independent random variables X_0, X_1, where $X_0 \sim \mathbb{N}(0, I_d)$. Then $P_{X_t|X_1=y} = C_t \, \mathrm{e}^{-\frac{\|x-ty\|^2}{2(1-t)^2}} \mathrm{d}x$ and we get

$$v_t(x) = \mathbb{E}[\partial_t X_t | X_t = x] = \mathbb{E}[X_1 - X_0 | X_t = x] = \mathbb{E}\left[\frac{X_1}{1-t} - \frac{X_t}{1-t}\middle| X_t = x\right]$$

Using Tweedie's (Miyasawa's) formula

$$\nabla_x \log p_t(x) = \frac{t\mathbb{E}[X_1|X_t = x] - x}{(1-t)^2}$$

$$\mathbb{E}[X_1|X_t = x] = \frac{(1-t)^2}{t}\nabla_x \log p_t(x) + \frac{x}{t}$$

we finally get with $\mathbb{E}[X_t|X_t] = X_t$ that

$$\begin{aligned} v_t(x) &= \frac{1-t}{t}\nabla_x \log p_t(x) + \frac{x}{t(1-t)} - \frac{x}{1-t} \\ &= \frac{1-t}{t}\nabla_x \log p_t(x) + \frac{x}{t}, \end{aligned}$$

Remark: Tweedie's Formula

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 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53

Consider

$$y = x + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma^2 I)$$

The observation density is related to the prior via Bayes' Theorem:

$$p(y) = \int p(y|x)p(x)\,\mathrm{d}x = \int g(y-x)p(x)\,\mathrm{d}x, \quad g(z) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\|z\|^2/(2\sigma^2)}.$$

For a given observation y, the minimum mean squared error (MMSE) estimator is given as

$$\hat{x}(y) = \int x\,p(x|y)\,\mathrm{d}x = \int x\,\frac{p(y|x)p(x)}{p(y)}\,\mathrm{d}x.$$

The MMSE can be computed using Tweedie's (Miyasawa's) formula

$$\hat{x}(y) = y + \sigma^2 \nabla_y \log p(y),$$

which can be seen directly by computing

$$\nabla_y p(y) = \frac{1}{\sigma^2} \int (x-y) g(y-x) p(x)\,\mathrm{d}x = \frac{1}{\sigma^2} \int (x-y) p(y,x)\,\mathrm{d}x,$$

and then multiplying both sides by $\sigma^2/p(y)$ gives

$$\frac{\sigma^2 \nabla_y p(y)}{p(y)} = \int x\,p(x|y)\,\mathrm{d}x - \int y\,p(x|y)\,\mathrm{d}x = \hat{x}(y) - y.$$

Forward SDE computes $(X_t)_{t\in[0,T]}$ by

$$\mathrm{d}X_t = f(t, X_t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t, \quad X_0 \sim P_{\text{data}},$$

where W_t is a standard Brownian motion.

Example: Usual choice is

$$f(t, X_t) := -\frac{1}{2}\beta_t X_t \quad \text{and} \quad g(t) := \sqrt{\beta_t}$$

with a positive, increasing so-called "time schedule" β_t. Then SDE becomes linear

$$\mathrm{d}X_t = -\frac{1}{2}\beta_t X_t\,\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}W_t, \quad X_0 \sim P_{\text{data}},$$

which has the closed form solution

$$X_t = \sqrt{1 - e^{-h(t)}}Z + e^{\frac{-h(t)}{2}}X_0, \quad h(t) := \int_0^t \beta_s\,\mathrm{d}s, \qquad Z \sim \mathcal{N}(0, I_d)$$

Hence, X_t reaches Z for $t \to \infty$.

Reverse SDE becomes

$$\mathrm{d}Y_t = \Big(- f(T-t, Y_t) + g(T-t)^2 \underbrace{\nabla \log p_{X_{T-t}}(Y_t)}_{\text{score}} \Big)\,\mathrm{d}t + g(T-t)\,\mathrm{d}W_t, \quad Y_0 \sim X_T.$$

Approximation of score by a neural network s^θ:

$$\begin{aligned}\mathcal{L}(\theta) &:= \mathbb{E}_{t\sim\mathcal{L}_{[0,T]}}\mathbb{E}_{x\sim P_{X_t}}\left[\|s_t^\theta(x) - \nabla\log p_{X_t}(x)\|^2\right]\\ &= \mathbb{E}_{t\sim\mathcal{L}_{[0,T]},x_0\sim X_0,x\sim P_{X_t|X_0=x_0}}\left[\|s_t^\theta(x) - \nabla\log p_{X_t|X_0=x_0}(x)\|^2\right] + \text{const}\end{aligned}$$

Example: (continued)

$$P_{X_t|X_0=x_0} = \mathcal{N}\big(b_t\, x_0,\, (1-b_t^2)\, I_d\big) \quad \text{with} \quad b_t = \mathrm{e}^{-\frac{h(t)}{2}}$$

This implies

$$\nabla\log p_{X_t|X_0=x_0}(x) = \nabla\left(-\tfrac{1}{2(1-b_t^2)}\|x - b_t x_0\|^2\right) = -\frac{x - b_t x_0}{1-b_t^2}.$$

Plugging this into the loss function, we get

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim\mathcal{L}_{[0,T]},x_0\sim X_0,x\sim P_{X_t|X_0=x_0}}\left[\left\|s_t^\theta(x) + \frac{x - b_t x_0}{1-b_t^2}\right\|^2\right].$$

Once the score is computed, we can use it in the reverse SDE starting with the $Z\sim\mathcal{N}(0, I_d)$ which approximates X_T for T large enough.

Fokker Planck Equation

For the forward PDE

$$\mathrm{d}X_t = f(t, X_t)\,\mathrm{d}t + g(t)\,\mathrm{d}W_t, \quad X_0 \sim P_{\text{data}},$$

the corresponding densities $p_t = p_{X_t}$ fulfill the Fokker-Planck equation (CE)

$$\begin{aligned} \partial_t p_t(x) &= -\nabla \cdot (f(t,x)p_t(x)) + g(t)^2 \Delta p_t(x) \\ &= -\nabla \cdot \Big(\underbrace{\big(f(t,x) - g(t)^2 \nabla_x \log p_t(x)\big)}_{v_t} p_t(x) \Big) \end{aligned}$$

Wasserstein gradient flow for minimizing $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$

Choosing $v_t \in -\partial \mathcal{F}(\mu_t) = -\nabla \delta \mathcal{F}(\mu_t)$ leads to the (CE)

$$\partial_t p_t = \nabla \cdot \big(p_t \, \nabla \delta \mathcal{F}(\mu_t)\big)$$

- $\mathcal{F}_\nu(\mu) = \mathrm{KL}(\mu, \nu) = \int_{\mathbb{R}^d} \log p(x) p(x)\,\mathrm{d}x - \int_{\mathbb{R}^d} \log q(x) p(x)\,\mathrm{d}x$, $\mu = p\lambda$, $\nu = q\lambda$
  Then
  $$\delta \mathcal{F}(p_t) = 1 + \log p_t - \log q, \quad \nabla \delta \mathcal{F}(p_t) = \frac{1}{p_t} \nabla p_t - \log q$$
  so that
  $$\partial_t p_t = \Delta p_t - \nabla \cdot (p_t \nabla \log q)$$
  which is the above Fokker-Planck equation for $f(t,x) = \nabla \log q(x)$ and $g(t) = 1$.

For $q(x) = C^{-1}\mathrm{e}^{-\beta\Psi(x)}$ we obtain the

Overdamped Langevin SDE:

$$X(0) = X^0 \sim p^0$$
$$\mathrm{d}X_t = -\nabla\Psi(X_t)d\tau + \sqrt{2\beta^{-1}}\mathrm{d}W_t$$

Euler-Maruyama forward step with step size η:

$$X_k := X_{k-1} - \eta\nabla\Psi(X_{k-1}) + \sqrt{2\beta^{-1}\eta}\,\xi_k, \quad \xi_k \sim \mathcal{N}(0, I_d)$$

GENERATIVE FLOWS FROM 1D PROCESSES

Gabriele Steidl

TU Berlin

Joint work with

Jannis Chemseddine, Richard Duong, Gregor Kornhardt and Peter Friz




Berlin Mathematics Research Center MATH+

Theoretical Foundations of Deep Learning

Funded under Germany's Excellence Strategy by DFG Deutsche Forschungsgemeinschaft

Task: similar as in diffusion models

- Learn transport of samples from $\mu_0 = \mu_X$ to samples of $\mu_1 = \mu_Z$ (target to noise)
  Learn/approximate v_t by neural network v_t^θ!
- Inference: use Backward flow with $-v_{1-t}$ in ODE (noise to target)

Loss function for learning:

$$\mathcal{L}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0,T),\, x_0 \sim \mu_0,\, x \sim \mu_t(\cdot|x_0)} \left[\left\| v_t^\theta(x) - v_t^{x_0} \right\|^2 \right]$$

🙂 **All you need is conditional $v_t^{x_0}$!**

Recall:

- We have already seen how to get $v_t^{x_0}$ from couplings, resp. linear stochastic processes.
- So far our latent distribution was a standard Gaussian.

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 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53

Outline

1. From One-Dimensional to Multi-Dimensional Flows
2. 1D Flows Besides Diffusion
3. Conclusions

Construction Plan

1. One-dimensional noise: construct flows in 1D fulfilling

$$\partial_t \mu_t + \partial_x(\mu_t v_t) = 0, \qquad \mu_0 = \delta_0 \quad \text{(CE)}$$

2. Multi-dimensional noise: set up a multi-dimensional flow starting in

$$\mu_0 = \delta_0, \quad 0 \in \mathbb{R}^d$$

3. Incorporating the data: find a multi-dimensional flow starting in

$$\mu_0 = \delta_{x_0}, \quad x_0 \sim \mu_0$$

Motivation: Wiener process $\mathbf{W}_t = (W_t^1, ..., W_t^d) \in \mathbb{R}^d$ consists of independent, identically distributed 1D components

Proposition

For $i = 1, \ldots, d$, let (μ_t^i, v_t^i) fulfill the one-dimensional CE and the μ_t^i are mutually independent. Then

$$\mu_t(x) = \prod_{i=1}^{d} \mu_t^i(x^i)$$

has process $\mathbf{Y}_t := (Y_t^1, \ldots, Y_t^d) \sim \mu_t$ and satisfies a multi-dimensional CE with a velocity field

$$v_t(x) := \left(v_t^1(x^1), \ldots, v_t^d(x^d)\right).$$

Step 3: Incorporating the Data

Done: $(\mathbf{Y}_t)_t$ starting in $\mathbf{Y}_0 = 0 \in \mathbb{R}^d$ with $v_t = v_t^{\mathbf{Y}}(\cdot \mid 0)$

Task: $(\mathbf{X}_t)_t$ which can start in any sample x_0 from μ_0. Let $\mathbf{X_0} \sim \mu_0$.

Mean-reverting process: $\mathbf{X}_t := f(t)\,\mathbf{X_0} + \mathbf{Y}_{g(t)}$, $t \in [0,1]$

with smooth scheduling functions f, g satisfying

$$f(0) = 1, \quad f(1) = 0 \quad \text{and} \quad g(0) = 0, \quad g(1) = 1$$

Conditional velocity field of $\mathbf{X}_t$:

$$\begin{aligned} v_t^{\mathbf{X}}(x \mid x_0) &= \mathbb{E}\big[\dot{\mathbf{X}}_t | \mathbf{X}_t = x,\ \mathbf{X_0} = x_0\big] \\ &= \mathbb{E}\big[\dot{f}(t)\,x_0 + \dot{g}(t)\,\dot{\mathbf{Y}}_{g(t)} \;\big|\; \mathbf{Y}_{g(t)} = x - f(t)x_0\big] \\ &= \dot{f}(t)\,x_0 + \dot{g}(t)\,v_{g(t)}^{\mathbf{Y}}\big(x - f(t)x_0 | 0\big). \end{aligned}$$

Remark (Generalization of FM and Diffusion):

$$\mathbf{X}_t = f(t)\,\mathbf{X_0} \; + \; \mathbf{Y}_{g(t)}$$
$$\mathbf{X}_t^{\mathsf{GF}} = \alpha_t \mathbf{X}_0 + \sigma_t \mathbf{X}_1, \qquad \mathbf{X}_1 \sim \mathcal{N}(0, I_d)$$

- $f(t) := \alpha_t$, $g(t) := \sigma_t^2$, $\mathbf{Y}_t$ standard Brownian motion: $\mathbf{X}_t^{\mathsf{GF}} \overset{d}{=} \mathbf{X}_t$
- FM: $f(t) := 1 - t$, $g(t) := t^2$
- Diffusion (Ohrenstein-Uhlenbeck):

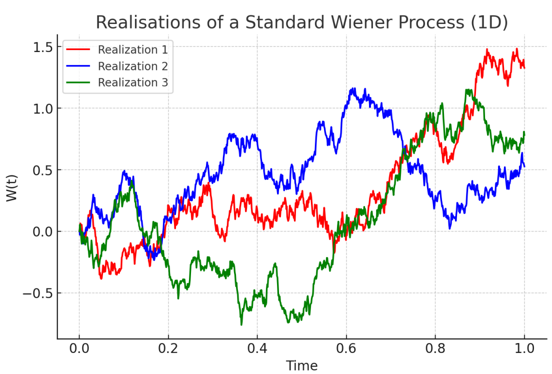
  $f(t) := \exp\left(-\frac{h(t)}{2}\right)$, $g(t) := 1 - \exp\left(-h(t)\right)$,

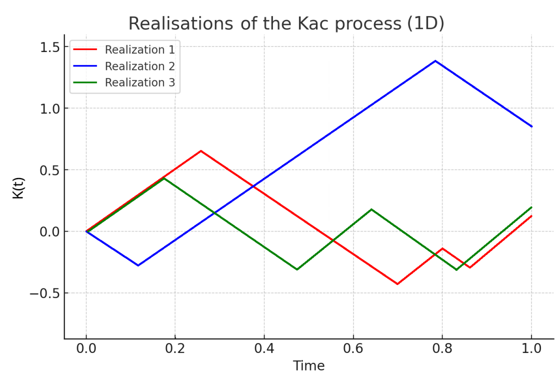
  where $h(t) := \int_0^t \beta_{\min} + s(\beta_{\max} - \beta_{\min})$ṣ

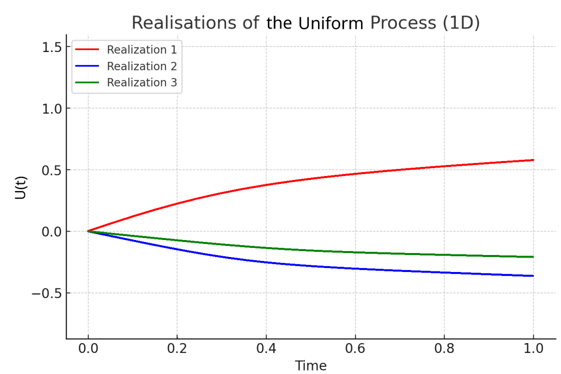
Outline

2.1. Prescribed Processes

1. Wiener process W_t $-$ diffusion equation
2. Kac process K_t $-$ damped wave equation
3. Uniform process U_t $-$ gradient flow of the maximum mean discrepancy functional $\mathcal{F}_\nu = \mathrm{MMD}_K(\cdot, \nu)$ with negative distance kernel
   $K(x, y) = -|x - y|$
   and $\nu = \mathcal{U}(-b, b)$, i.e.
   (μ_t, v_t) fulfill the continuity equation with $v_t \in -\partial \mathcal{F}_\nu(\mu_t)$







1. Wiener Process and Diffusion Equation

Probability density flow p_t:

$$\partial_t p_t = \nabla \cdot (p_t \underbrace{\frac{1}{2}\nabla \log p_t}_{v_t}) = \frac{1}{2}\Delta p_t, \quad t \in (0,1], \qquad \lim_{t \downarrow 0} p_t = \delta_0,$$

Analytic solution:

$$p_t(x) = (2\pi t)^{-\frac{d}{2}} \mathrm{e}^{-\frac{\|x\|^2}{2t}}, \quad p_1 = \mathcal{N}(0, I_d)$$

Velocity field:

$$v_t(x) = -\frac{1}{2}\nabla \log p_t = \frac{x}{2t}$$

Reverse velocity field:

$$-v_{1-t}(x) = -\frac{x}{2(1-t)} \quad \longrightarrow \quad \|v_{1-t}\|^2_{L_2(\mu_t)} = \frac{d}{4(1-t)} \notin L_2(0,1)$$ 😥

Instability issues caused by this explosion at times close to the target Kim et al. ICLR 2022 and similar with drifts (Fokker-Planck equation) Pidstrigach, NeuRips 2022;
may be be avoided by e.g. time truncations

Refs: Mate/Fleuret, TMLR 2023, Maurais/Marzuok ICML 2024, Chemsedine, Wald, Duong, St. ICLR 2025, Berner, Richter et al. 2024, ICLR 2025)

Motivation

- It appears to be unphysical that particles in diffusion can travel with unbounded velocities, contradicting Einstein's principles of relativity. On the other hand, diffusion (mathematical) can be seen as limit of the damped wave equation with infinite damping and infinite propagation speed.

Let us consider the damped wave equation!

- What other physical PDEs can be used in generative AIs?

  - Xu, Liu, Tegmark, Jakkola, Poisson flows, NeuRips 2022
  - Liu, Lu, Xu, Jaakkola, Tengmark, GenPhys: from physical processes to generative models, 2023

Refs: Duong, Chemseddine, Friz, St., Telegraphers generative models via Kac flows, 2025

Last week follow up paper: DestillKac, Ermon et al. 2025

Kac Process (Mark Kac 1974)

Particle starts at $x_0 = 0$ to the right and moves with step Δt and velocity $c > 0$ in either positive or negative direction, where

- same direction with probability $1 - a\Delta t$
- reverse direction with the probability $a\Delta t$ $\quad \rightarrow$ random variable $\varepsilon_k \in \{\pm 1\}$

Particle displacement S_n after n steps (random variable):

$$S_n = c\Delta t(1 + \varepsilon_1 + \varepsilon_1\varepsilon_2 + \ldots + \varepsilon_1 \ldots \varepsilon_n) = c\Delta t \sum_{k=0}^{n} (-1)^{N_k}$$

where $N_k \sim B(k, a\Delta t)$ number of reversals up to step k

- Continuous limit for $n \to \infty$ such that $na\Delta t \to at$, then $N_n \to \mathrm{Poi}(at)$

Number of reversals $N(t)$ up to time t is given by a

homogeneous Poisson point process with rate a:

i) $N(0) = 0$;

ii) Independent increments of $N(t)$

iii) $N(t) - N(s) \sim \mathrm{Poi}\big(a(t-s)\big)$ for all $0 \leq s < t$.

Kac Process

Particle displacement:

$$S_t = c \int_0^t (-1)^{N(s)} \text{ṣ}$$

Kac process starting in 0: $K_t := \mathrm{B}_{\frac{1}{2}} S_t, \qquad \mathrm{B}_{\frac{1}{2}} \sim B(1, \frac{1}{2})$

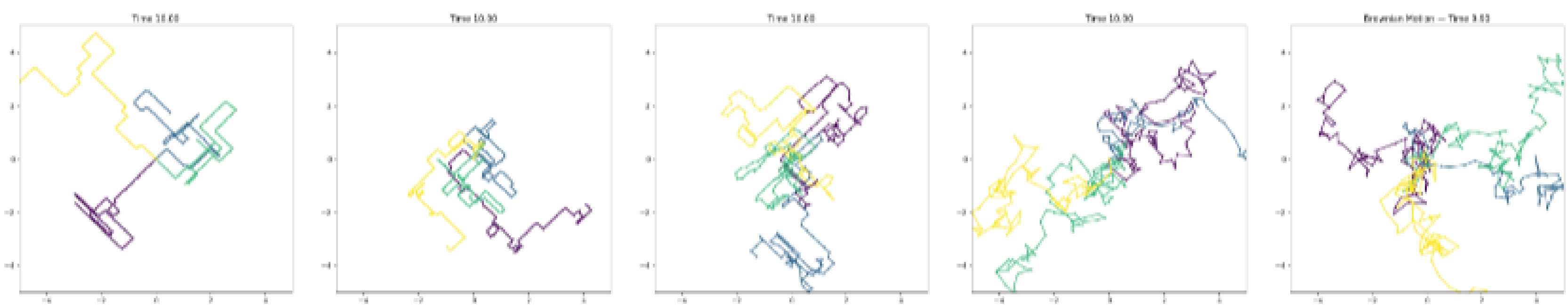


Figure 1: Paths of the componentwise Kac walk in 2D, simulated until time $T = 10$ with damping/velocity parameters $(a, c) = (1, 1), (2, 1), (4, 2), (25, 5)$, and a standard Brownian motion (right).

Refs.: Zhang et al., Revisiting Kac's method: A Monte Carlo algorithm for solving the telegrapher's equations, 2018

Theorem. The probability distribution flow of $(K_t)_t$ is

$$\mu_t(x) = \frac{1}{2}e^{-at}\big(\delta_0(x+ct) + \delta_0(x-ct)\big) + \tilde{p}_t(x),$$

with the absolutely continuous part

$$\tilde{p}_t(x) := \frac{1}{2}e^{-at}\Big(\beta ct \frac{I_0'(\beta r_t(x))}{r_t(x)} + \beta I_0(\beta r_t(x))\Big) 1_{[-ct,ct]}(x),$$

where

$$r_t(x) := \sqrt{c^2t^2 - x^2}$$

and $\beta := \frac{a}{c}$, and I_0 is 0-th *modified Bessel function of first kind*. It is the generalized solution of the damped wave equation/telegrapher equation

$$\begin{aligned} \partial_{tt}u(t,x) + 2a\,\partial_t u(t,x) &= c^2 \Delta_x u(t,x), \\ u(0,x) &= \delta_0(x), \\ \partial_t u(0,x) &= 0. \end{aligned}$$

- $a > 0$ damping coefficient
- $c > 0$ velocity of the wave front

Theorem. Let $f_0 = \delta_0$. Then μ_t associated to $u(t,\cdot)$ solves CE

$$\partial_t \mu_t = -\partial_x(\mu_t v_t)$$

in a weak sense with velocity field

$$v_t(x) := \begin{cases} \frac{x}{t + \frac{r_t(x)}{c} \frac{I_0(\beta r_t(x))}{I_0'(\beta r_t(x))}} & \text{if } \quad x \in (-ct, ct), \\ c & \text{if } \quad x = ct, \\ -c & \text{if } \quad x = -ct, \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

It holds $\|v_t\|^2_{L_2(\mu_t)} \leq c^2$, and hence $\|v_t\|_{L_2(\mu_t)} \in L_2(0,1)$. 🙂

- Compare to diffusion starting in δ_0: $\quad v_t(x) = \frac{x}{2t}$
- Damped wave equation is in general no longer mass-conserving in dimension $d \geq 3$ Refs: Tautz,Lerche 2016

Theorem (Kac Insertion Methods) Refs: Griego/Hersh 1971

For any $f_0 \in H^2(\mathbb{R}^d)$, let $w_c(t,x)$ solve the undamped wave equation

$$\partial_{tt} w(t,x) = c^2 \Delta w(t,x), \quad x \in \mathbb{R}^d,\ t > 0,$$
$$w(0,x) = f_0(x), \qquad \partial_t w(0,x) = 0.$$

Then, the functions defined by

$$u(t,x) := \mathbb{E}\big[w_c(c^{-1} S_t, x)\big]$$
$$h(t,x) := \mathbb{E}\left[w_1\left(\sigma W_t, x\right)\right],$$

multi-dimensional damped wave equation, resp. the heat equation

$$\partial_t h(t,x) = \frac{\sigma^2}{2} \Delta h(t,x), \quad x \in \mathbb{R}^d,\ t > 0,$$
$$h(0,x) = f_0(x).$$

Corollary. The solution $u^{a,c}(t,\cdot)$ of the damped heat equation converges to the solution $h(t,\cdot)$ of the diffusion equation for $a, c \to \infty$ with fixed $\sigma^2 = \frac{c^2}{a}$.

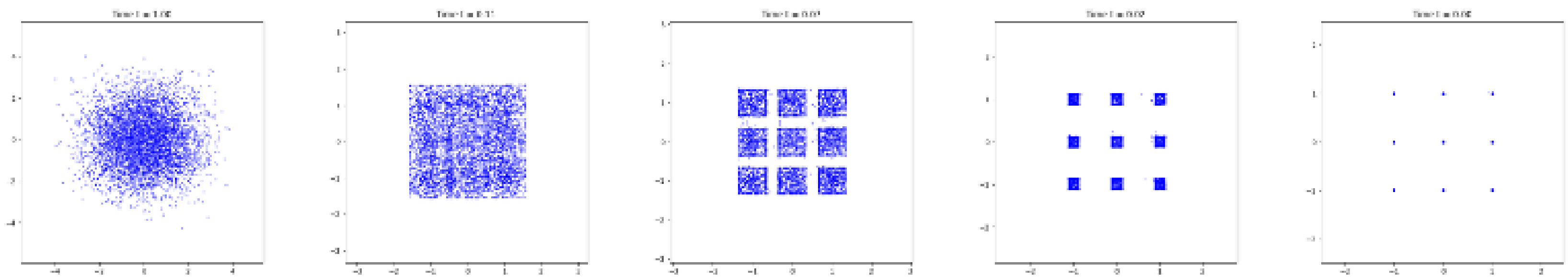
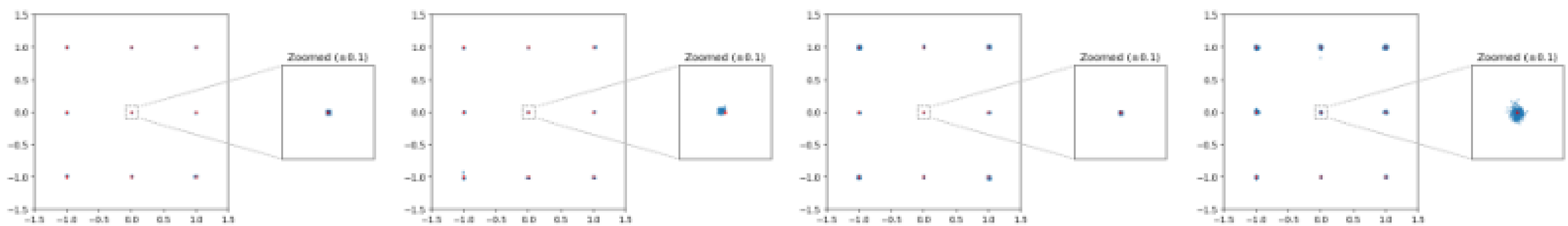


Figure 4: Backward evolution of the *learned* Kac flow for $(a, c) = (25, 5)$, see also Figure 5.



$(a, c) = (25, 5)$
80k iterations

$(a, c) = (64, 8)$
280k iterations

$(a, c) = (1600, 40)$
460k iterations

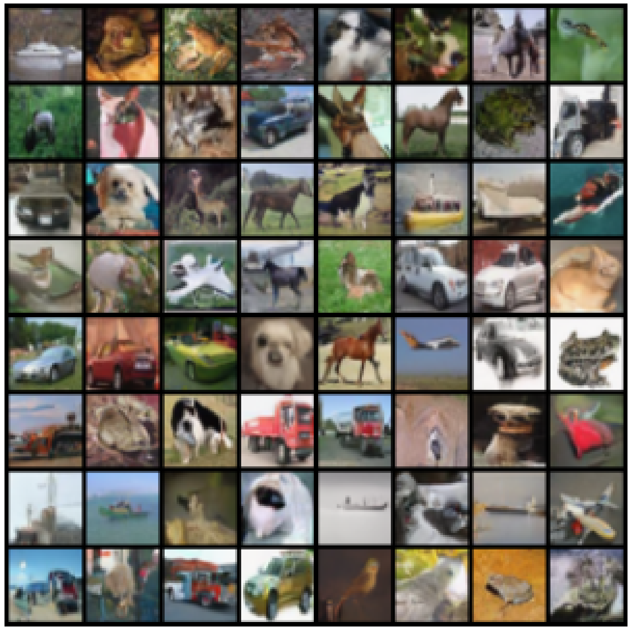
diffusion
420k iterations

Figure 5: Generated samples (blue) vs. ground truth (red) at the indicated iteration for each model. The Kac model can precisely recover the small modes, while the diffusion model creates "blobs".

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 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53

Numerical Example

Schedule: $g(t) = t^2$			
Method	**FID**	**Method**	**FID**
$a = 900, c = 10$	**7.26**	$a = 100, c = 10$	10.01
$a = 900, c = 20$	7.46	$a = 25, c = 1$	8.60
$a = 900, c = 30$	8.05	$a = 25, c = 2$	8.65
$a = 100, c = 1$	11.41	$a = 25, c = 3$	9.15
$a = 100, c = 3$	7.77	$a = 25, c = 4$	9.56
$a = 100, c = 5$	7.73	$a = 25, c = 5$	10.70
$a = 100, c = 7$	8.58	FM (our impl.)	7.59



$(a, c) = (900, 10)$

3. Uniform Process and MMD Gradient Flow

Gradient flow of the maximum mean discrepancy functional $\mathcal{F}_\nu = \mathrm{MMD}_K(\cdot,\nu)$ with negative distance kernel $K(x,y) = -|x-y|$ and $\nu = \mathcal{U}(-b,b)$, i.e. (μ_t, v_t) fulfill the continuity equation with

$$v_t \in -\partial \mathcal{F}_\nu(\mu_t)$$

Theorem The Wasserstein gradient flow μ_t of $\mathcal{F}_\nu$ starting in $\mu_0 = \delta_0$ towards the uniform distribution $\nu = \mathcal{U}[-b,b]$ is

$$\mu_t = (1-\exp(-\tfrac{t}{b}))\mathcal{U}[-b,b], \quad t>0,$$

with corresponding velocity field

$$v_t(x) = \frac{x}{b\left(\exp\left(\frac{t}{b}\right)-1\right)}, \quad x \in \mathrm{supp}(\mu_t).$$

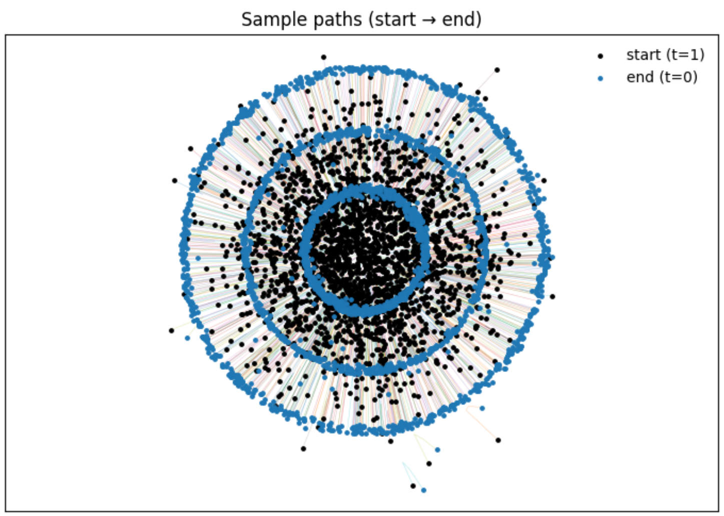
A corresponding (stochastic) process $(U_t)_t$ is given by

$$U_t := b\left(1-\exp\left(-\frac{t}{b}\right)\right)U, \qquad U \sim \mathcal{U}[-1,1].$$

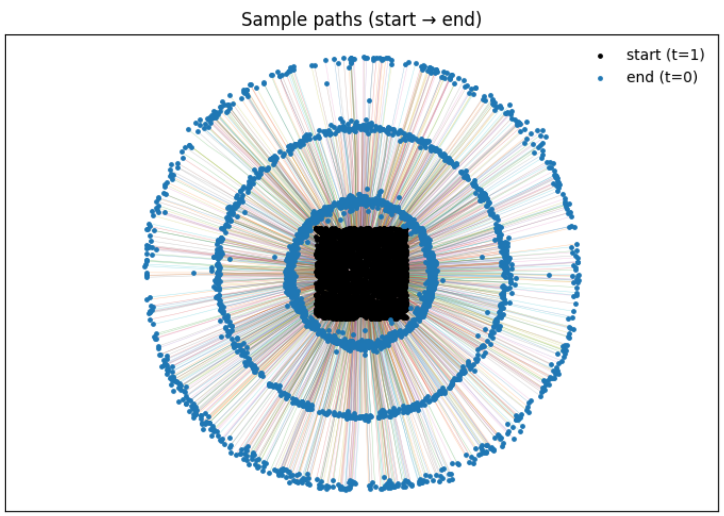
It holds $\|v_t\|^2_{L_2(\mu_t)} = \frac{2b}{3}\exp(-\frac{2t}{b})$, and hence $\|v_t\|_{L_2(\mu_t)} \in L_2(0,1)$. 🙂

Refs: Hertrich, Beinert, Gräf, St., JMAA 2024

Example



IMM + Gaussian

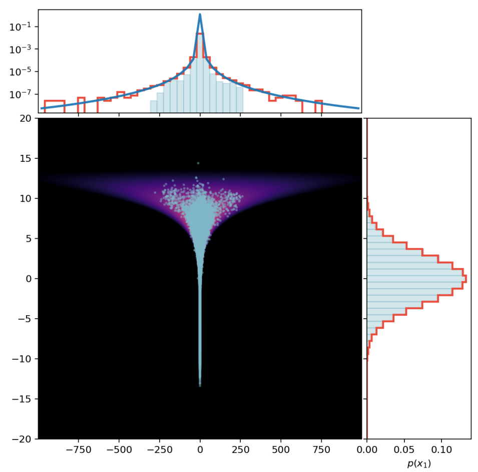


IMM + MMD

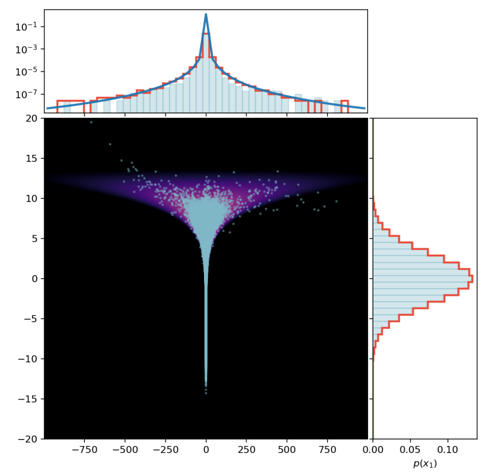
Outline

3.2 Learned Processes

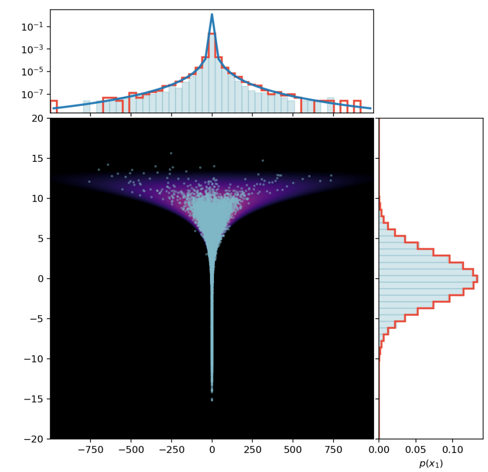
- Besides exploding velocity fields, starting at a Gaussian may encounter difficulties if e.g.
  the target is heavy tailed or multimodal or ...



Gaussian



Student-T



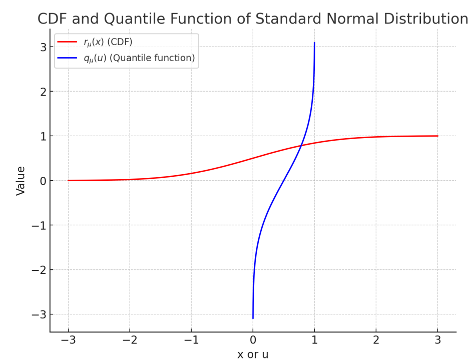
Our (Learned)

Refs:
Hagemann/Neumayer, Stabilizing invertible neural networks using mixture models, Inverse Problems 2021,
Salomea et al., Can push-forward generative models fit multimodal distributions?, Neurips 2022

Pandey et al., Heavy-tailed diffusion models, 2024

Quantile Functions and Measures on $\mathbb{R}$

- Cumulative distribution function: $R_\mu(x) := \mu\big((-\infty, x]\big)$, $x \in \mathbb{R}$
- Quantile function: $Q_\mu(u) := \min\{x \in \mathbb{R} : R_\mu(x) \geq u\}$, $u \in (0,1)$



- Quantile functions form a closed, convex cone in $L_2(0,1)$

$$\mathcal{C} := \{f \in L_2(0,1) : f \text{ increasing } a.e.\}$$

- $\mu \mapsto Q_\mu$ is an isometric embedding of $(\mathcal{P}_2(\mathbb{R}), W_2)$ into $(L_2(0,1), \|\cdot\|_{L_2})$:

$$W_2^2(\mu,\nu) = \int_0^1 \big|Q_\mu(s) - Q_\nu(s)\big|^2 \d s$$

and $Q_\mu \circ U \sim \mu$, resp. $\mu = Q_{\mu,\sharp}\mathcal{L}_{(0,1)}$

Learning Quantile Functions

Idea: Learn quantile functions $\mathbf{Q}_\phi = (Q^i_\phi)_{i=1}^d$ taking their monotonicity into account Refs. monotonicity parameterization: Durkan et al. 2019 and consider

$$\mathbf{X}_t = (1-t)\mathbf{X}_0 + t\mathbf{Q}_\phi(\mathbf{U}) \quad \text{with i.i.d. } U^i \sim \mathcal{U}[0,1]$$

such that the noise $\mathbf{Y}_1 = (Q^i_\phi(U^i))_{i=1}^d$ adapts to the given data from $\mathbf{X}_0$. By construction

$$\text{Law}(\mathbf{Y}_1) \in \mathcal{S} := \{\nu \in \mathcal{P}_2(\mathbb{R}^d) : \nu = \rho \underset{\cdot}{x} \text{ and } \rho = \Pi_{i=1}^d \rho^i\},$$

so that it appears reasonable to ask for

$$\mathcal{E}(\phi) = W_2^2\big(\mu_0, \nu_\phi\big), \quad \nu_\phi := (\mathbf{Q}_\phi)_\# \mathcal{U}([0,1]^d).$$

Loss Function: Take also the velocity field into account

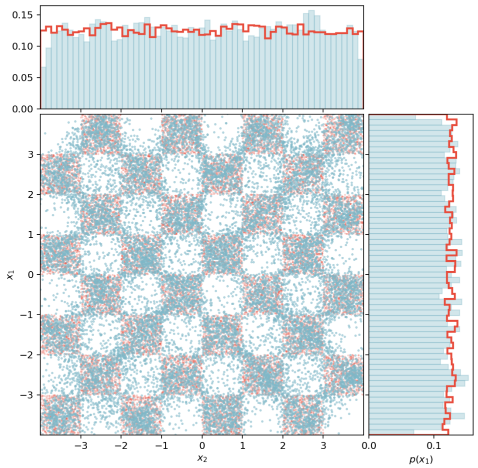
$$\mathcal{L}(\theta,\phi) = \mathcal{E}_{\text{CFM}}(\theta,\phi) + \lambda\,\mathcal{E}(\phi), \quad \lambda > 0,$$

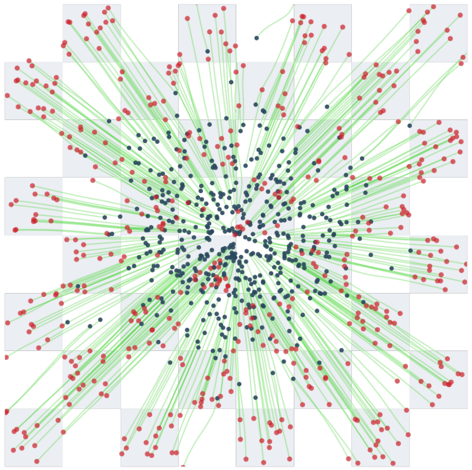
$$\mathcal{E}_{\text{CFM}}(\theta,\phi) = \mathbb{E}_{t\sim\mathcal{U}(0,1),(x_0,x)\sim\pi_\phi}\Big[\big\|v_\theta\big((1-t)x_0 + tx,\, t\big) - (x_0 - x)\big\|_2^2\Big],$$

where $\pi_\phi \in \Pi_o(\mu_0, \nu_\phi)$

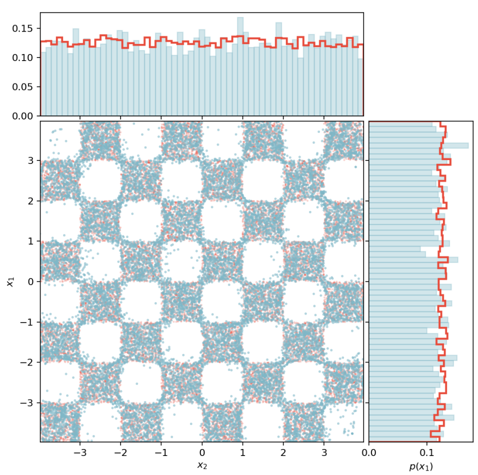
Example: Checkerboard

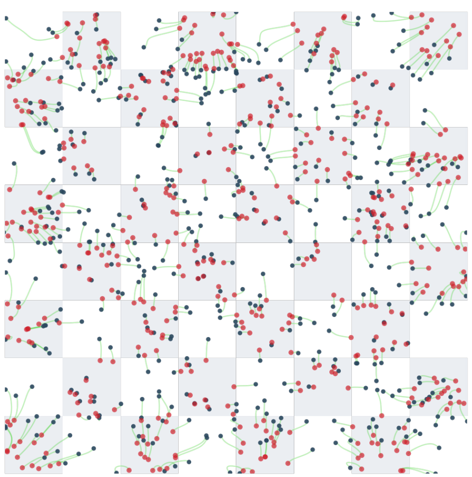
Gaussian:

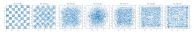




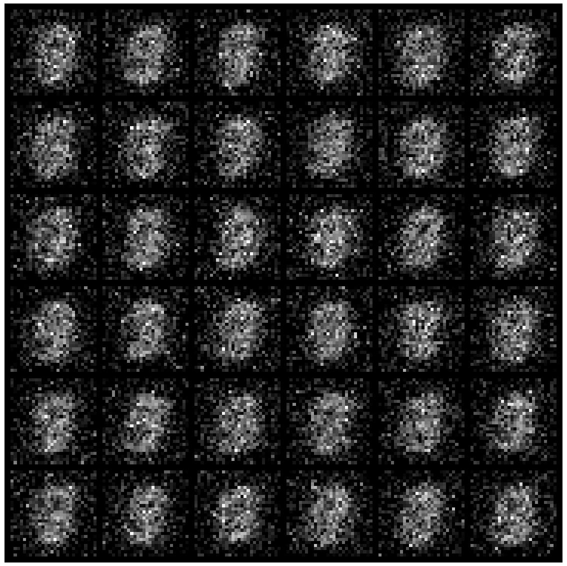
Learned:





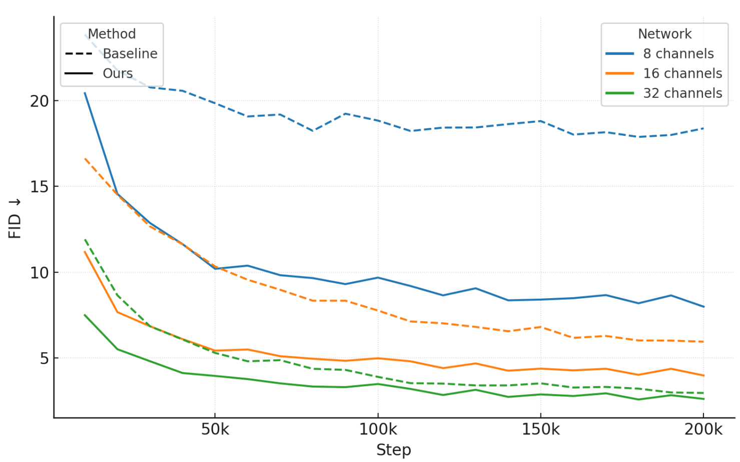
Noise: 

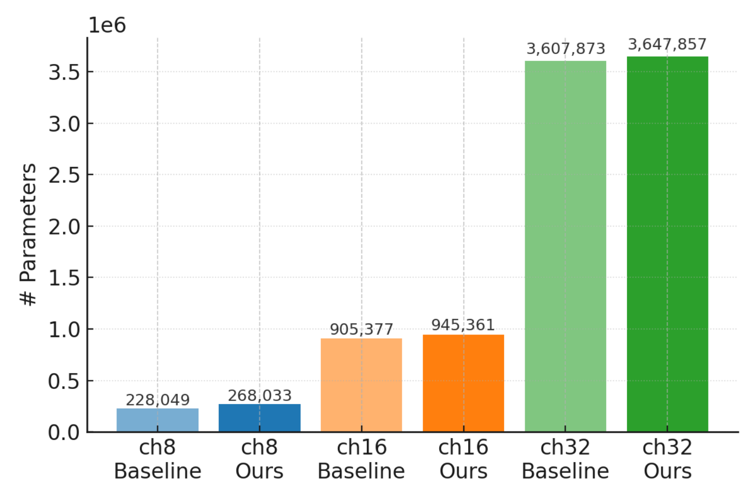
Example: MNIST



Samples from the learned latent

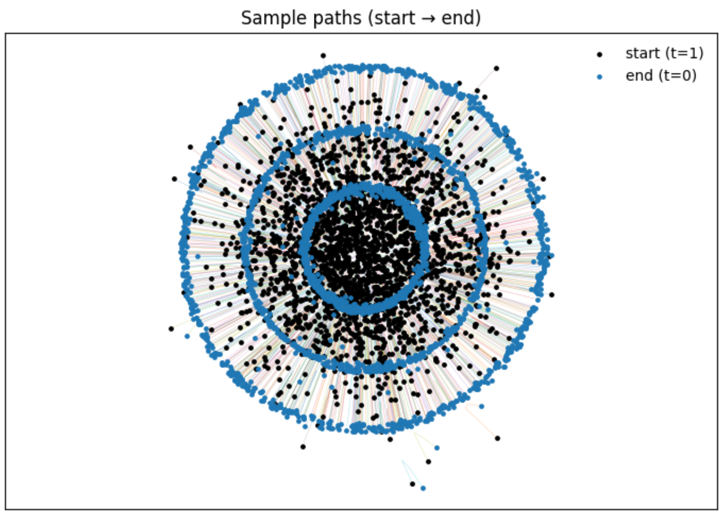
Generated samples



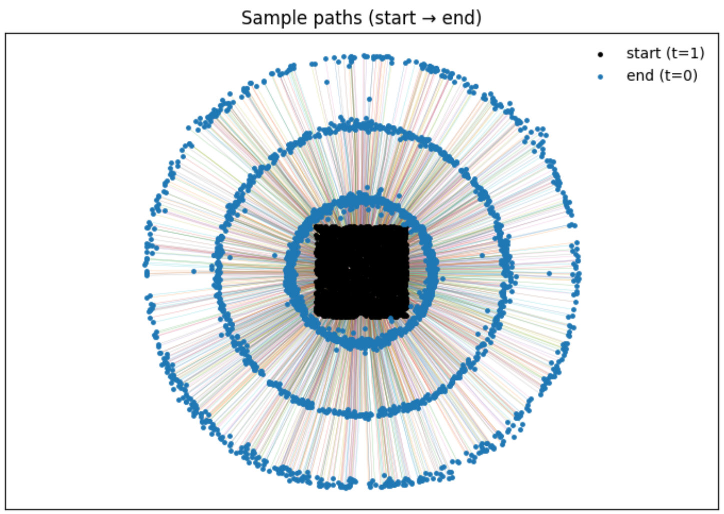


3. Conclusions

- "Quantile sandbox" for building generative models: a unifying theory and a practical toolkit that turns noise selection into a data-driven design element
- Seamless fit into standard objectives including Flow Matching and consistency models, e.g. Inductive Moment Matching (IMM) via so-called quantile interpolants



IMM+ Gaussian



IMM + MMD

- Promising directions for future research
- References:
1. Duong, Chemseddine, Friz, St: Telegrapher's Generative Model via Kac Flows, ArXiv 2025
2. Chemseddine, Kornhardt, Duong, St: Adapting data to noise: generative flows from 1D processes, ArXiv this week

Berlin Mathematics Research Center

MATH+

Funded under Germany's Excellence Strategy by

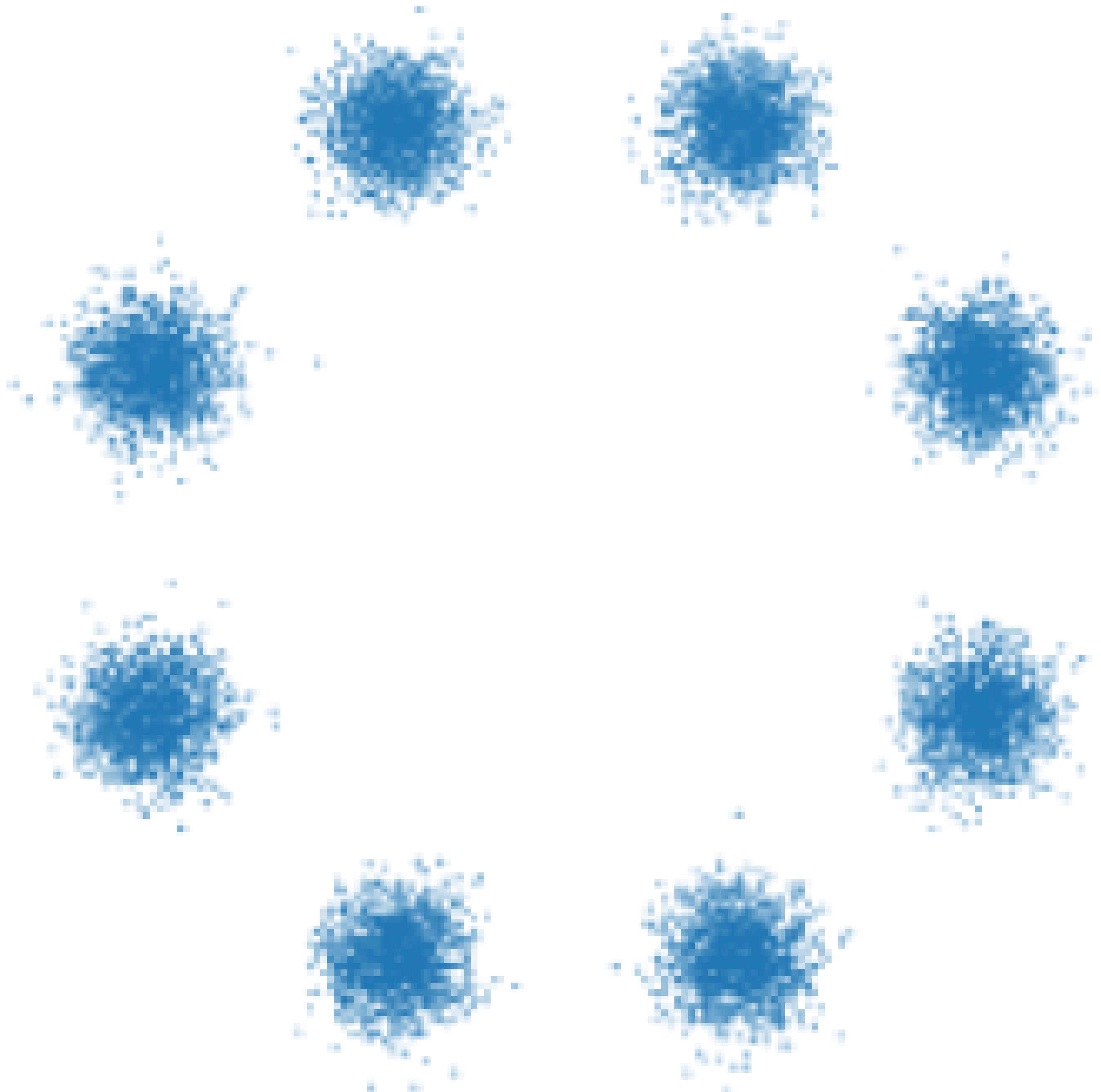
DFG Deutsche Forschungsgemeinschaft

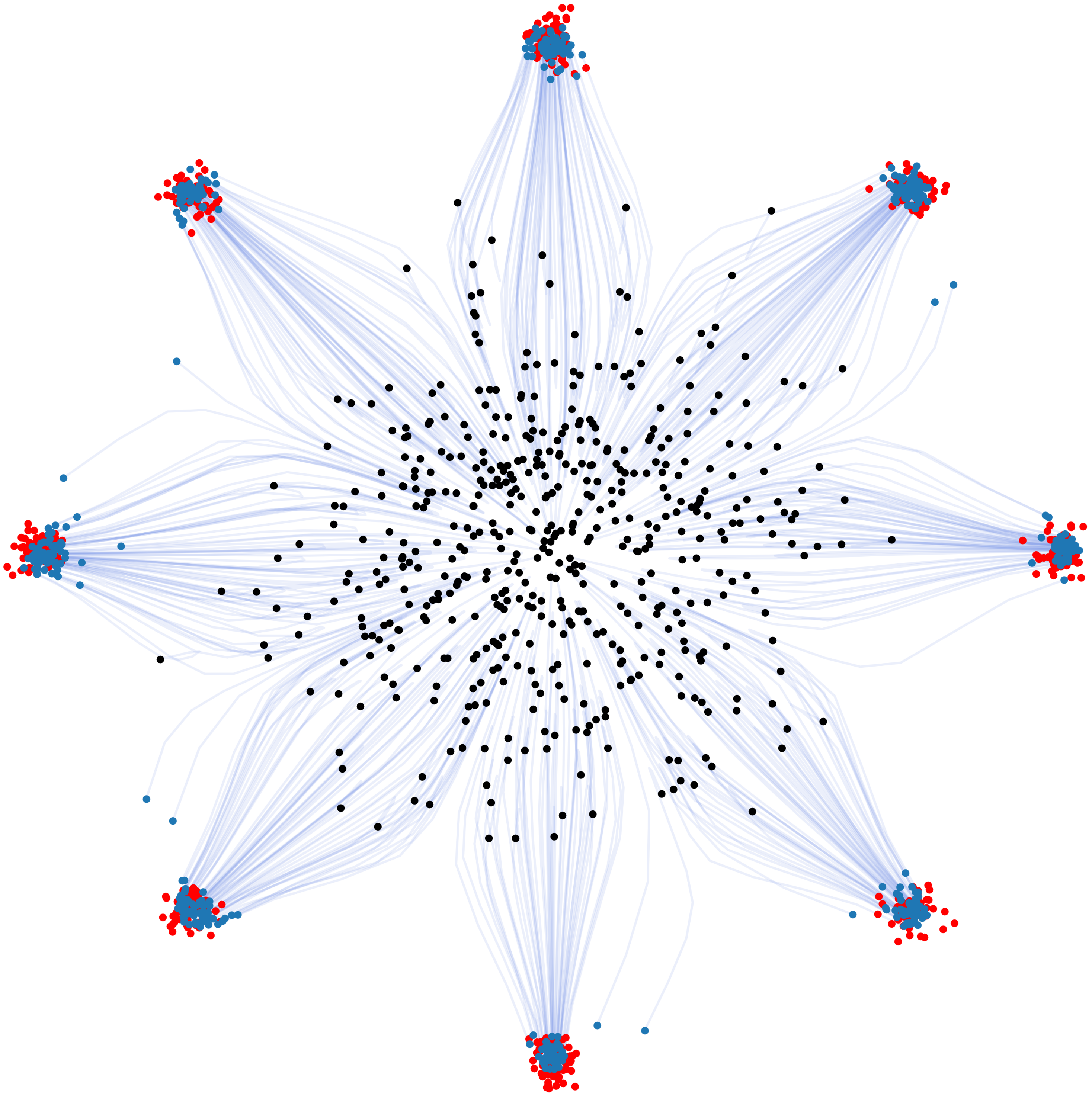
Oberwolfach Seminars
56

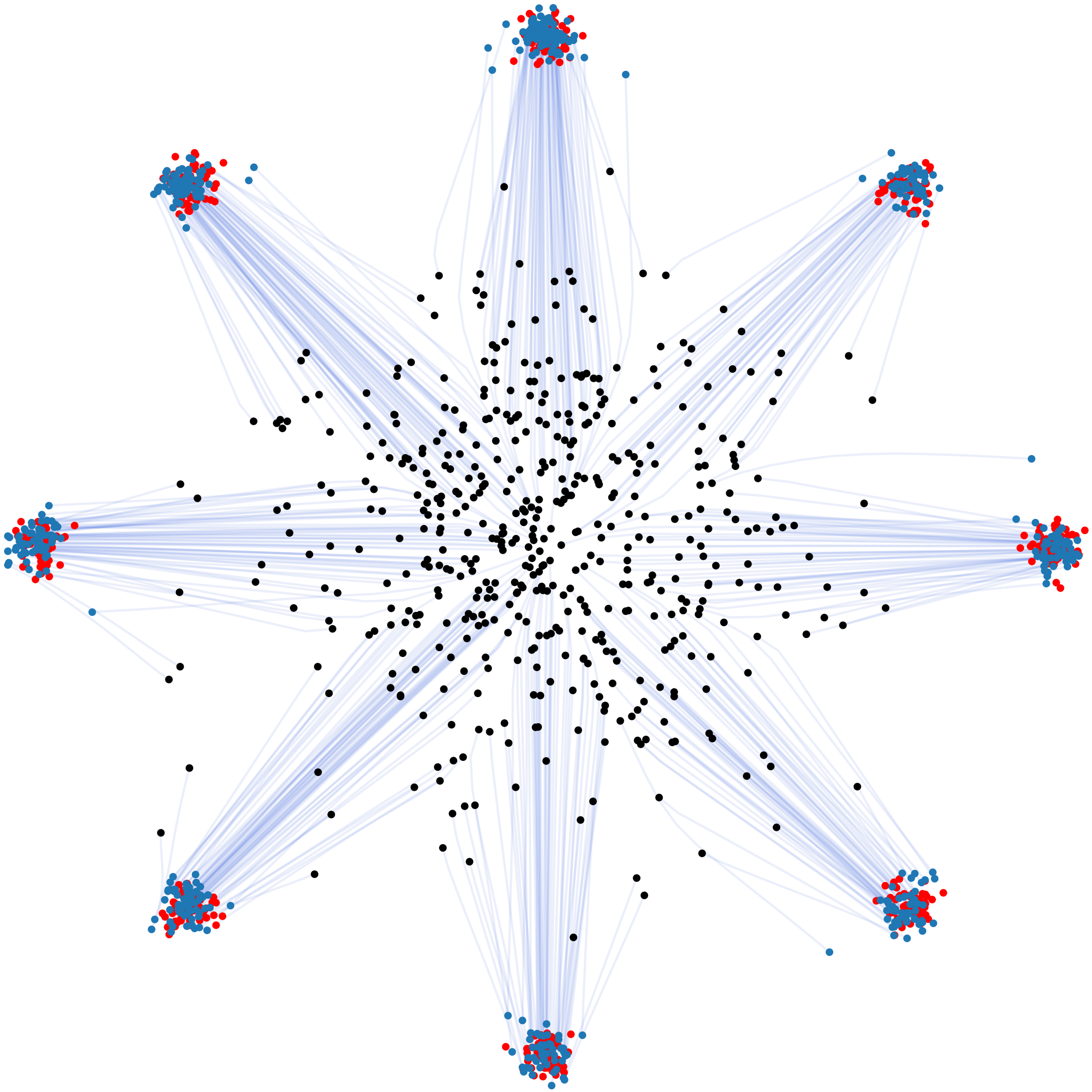
Variational and Information Flows in Machine Learning and Optimal Transport

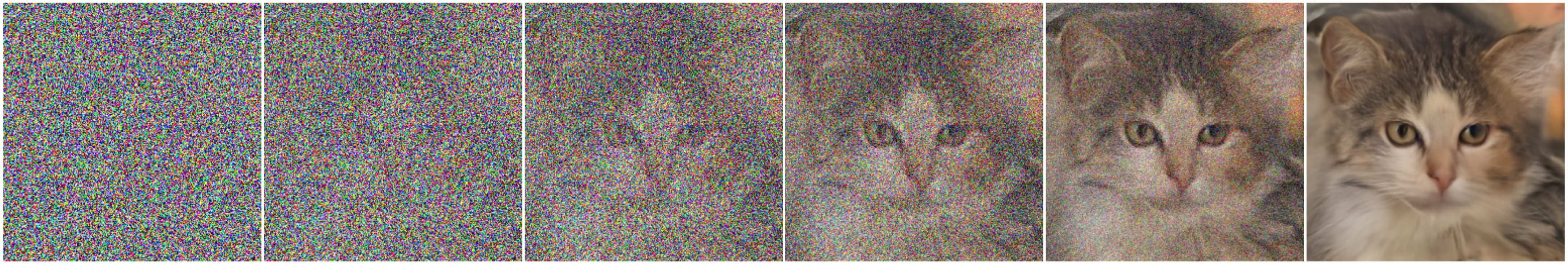
Wuchen Li
Bernhard Schmitzer
Gabriele Steidl
François-Xavier Vialard
Christian Wald

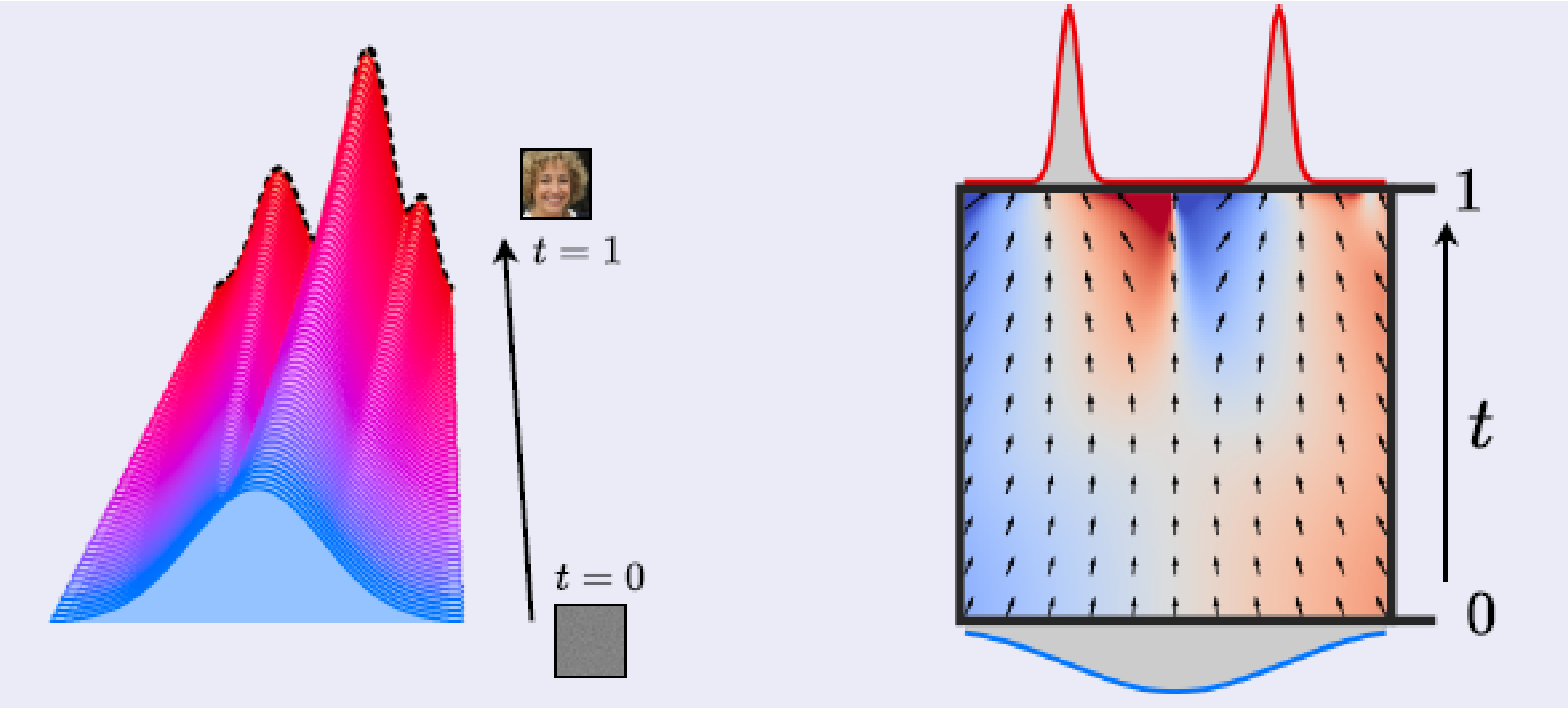
Birkhäuser



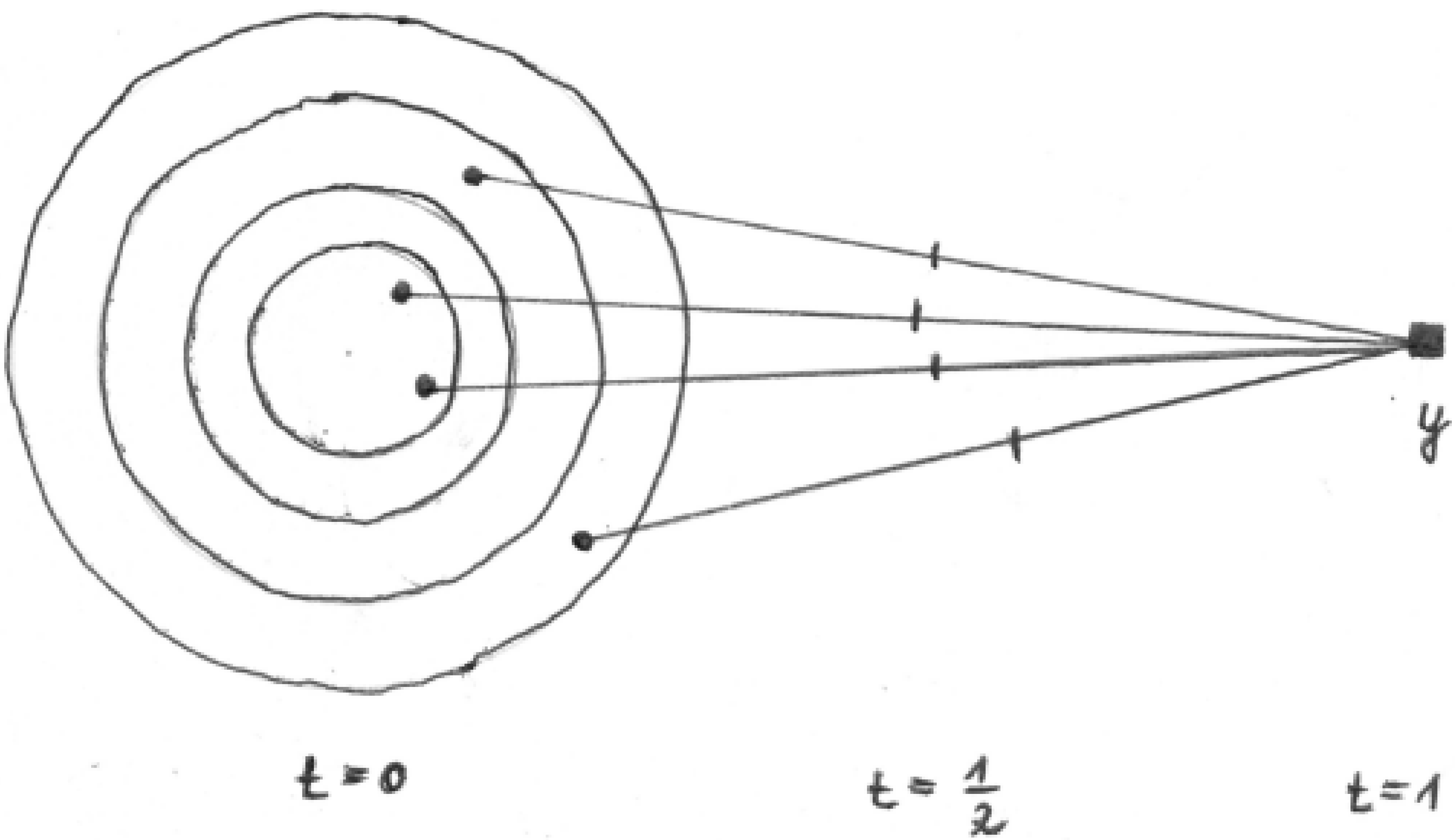


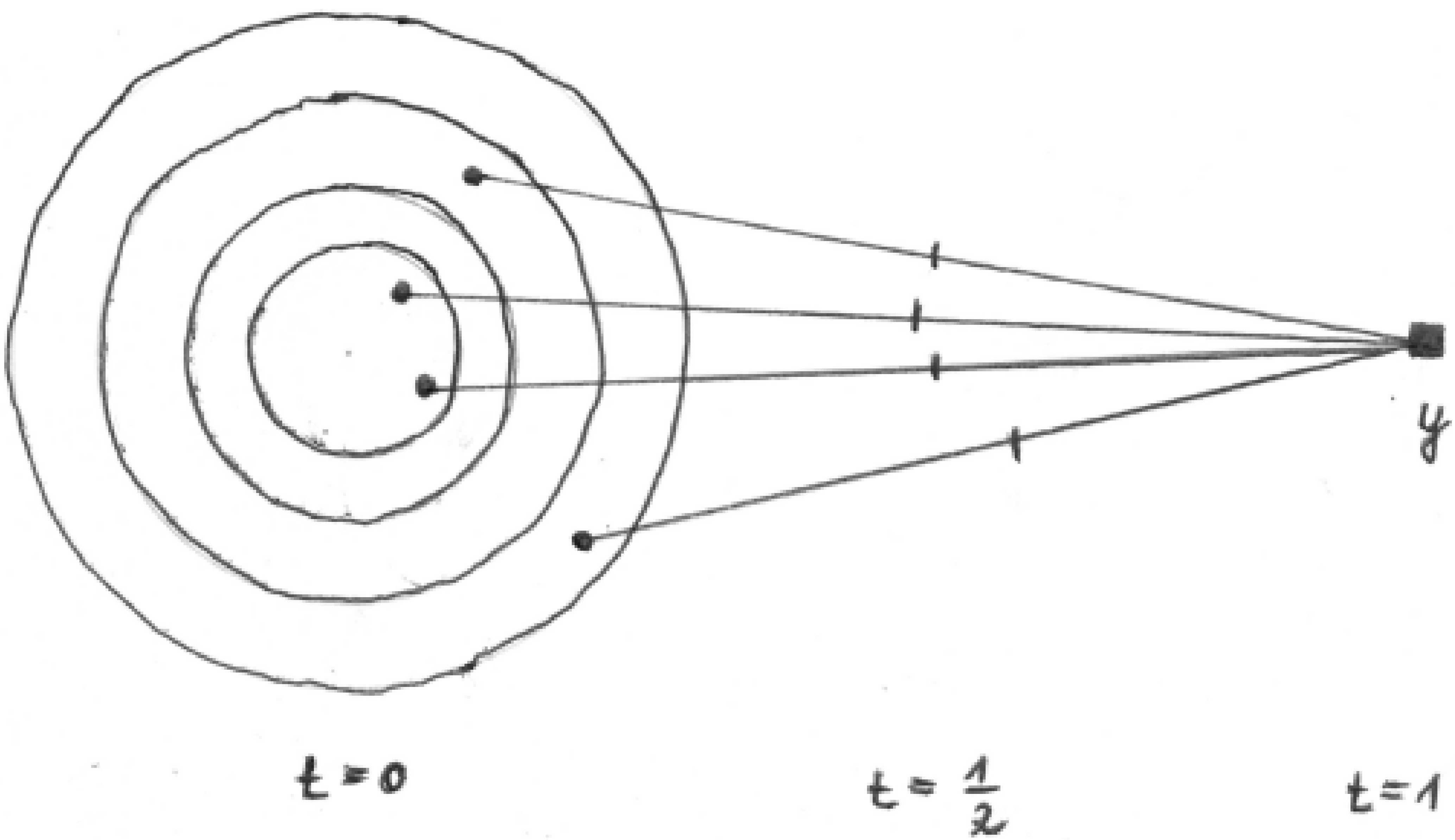


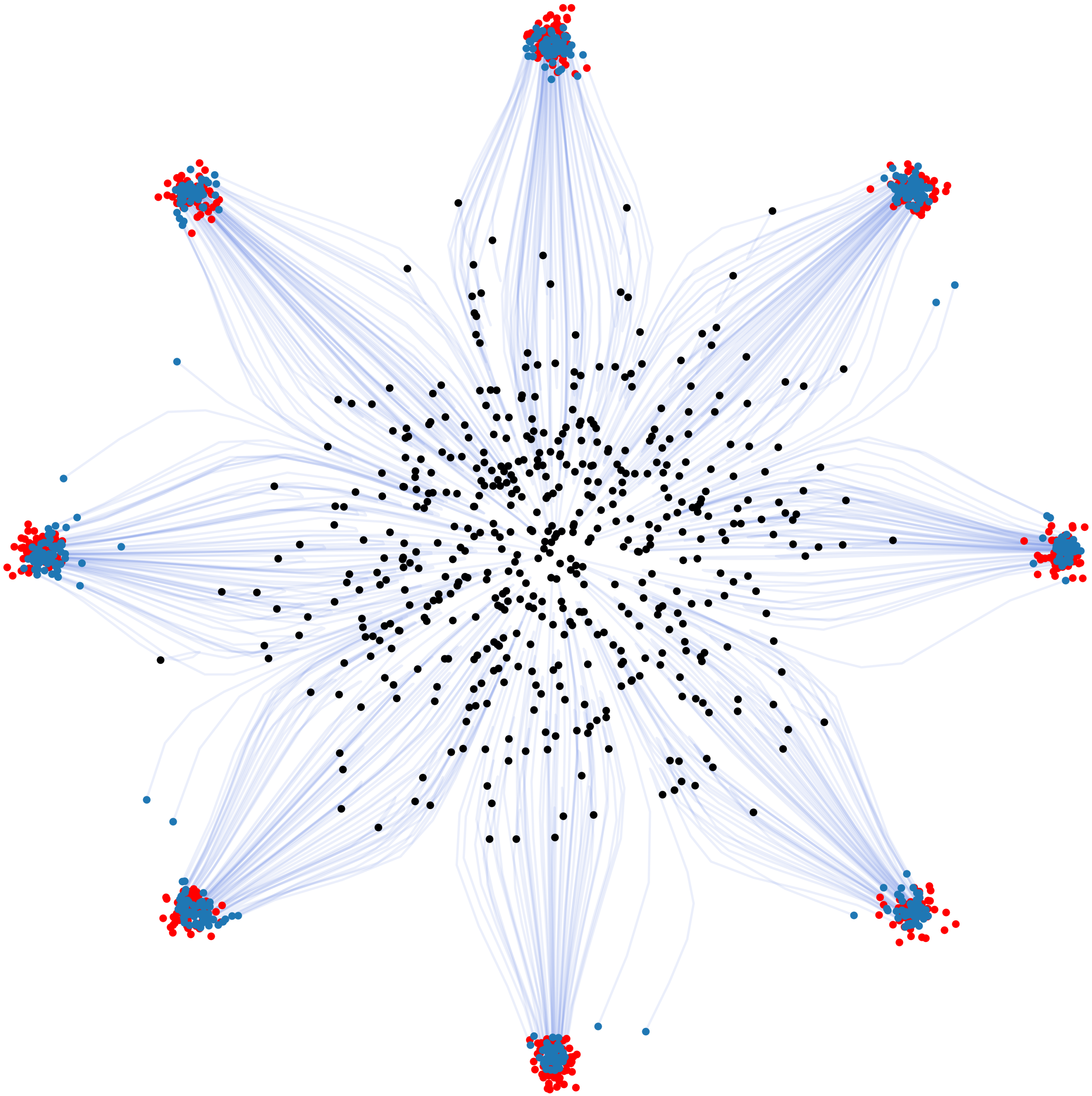


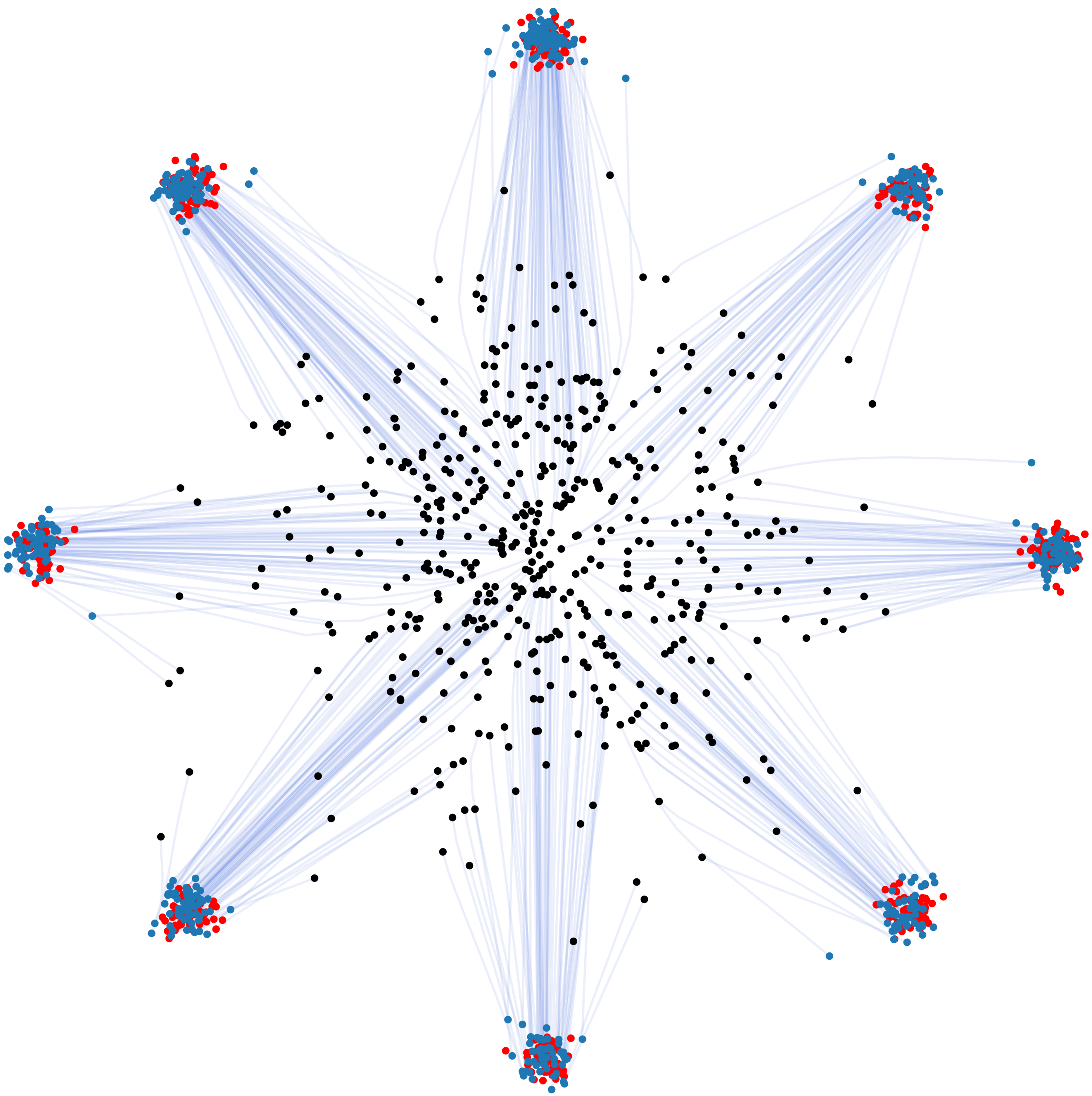


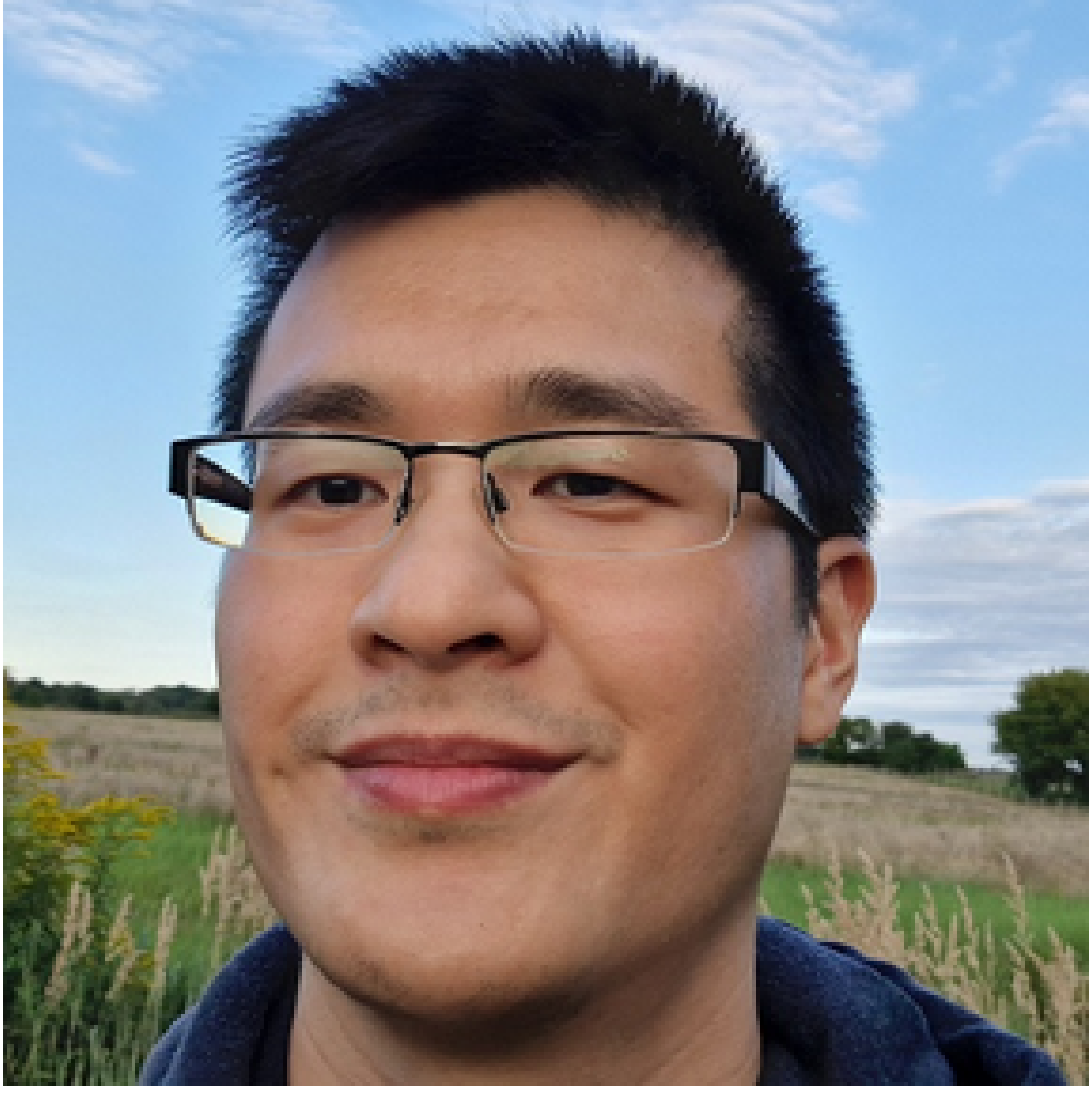
$t = 1$
$t = 0$
1
t
0


y
t = 0
t = 1/2
t = 1


y
t = 0
t = 1/2
t = 1







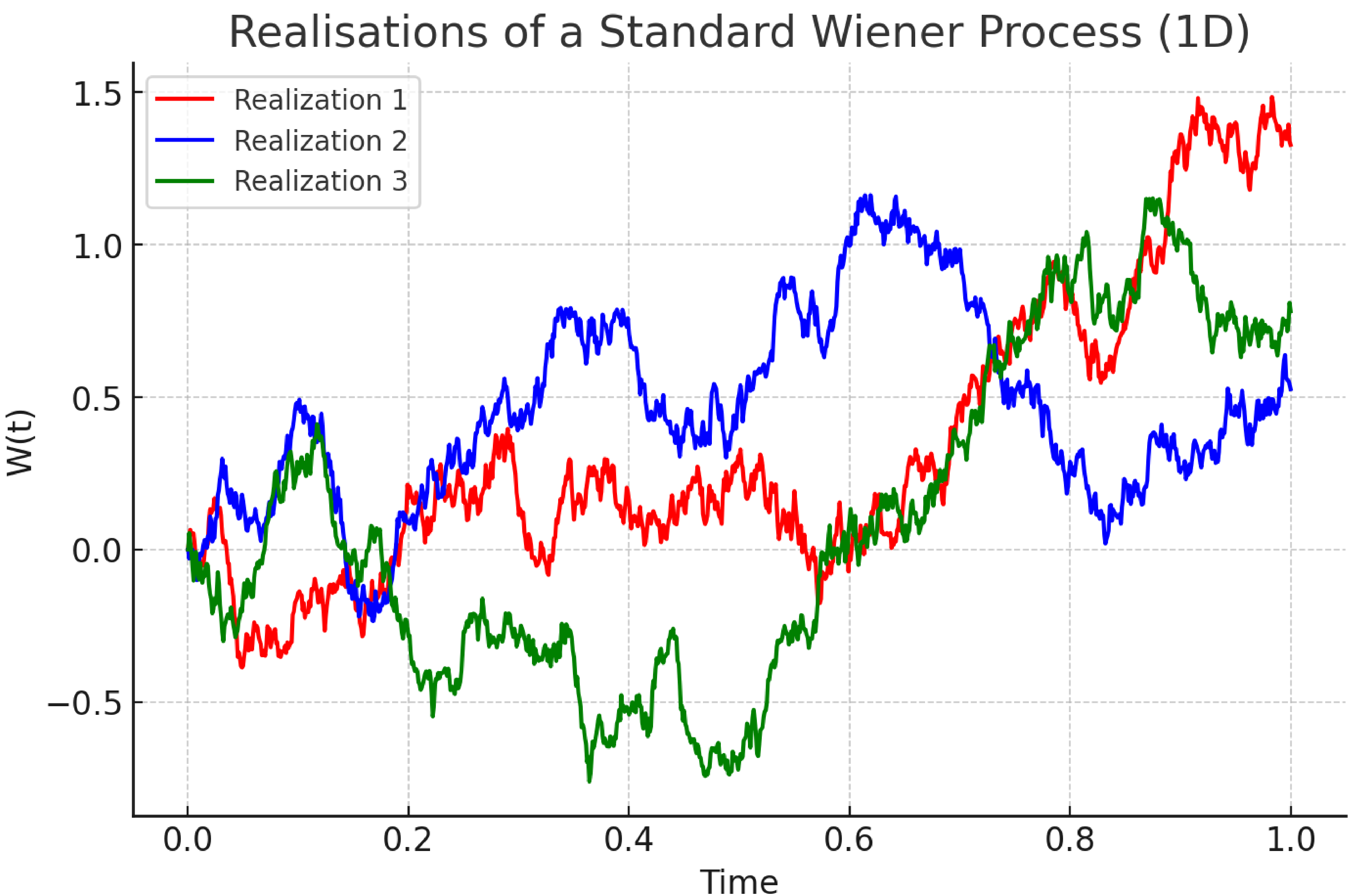
Berlin Mathematics Research Center

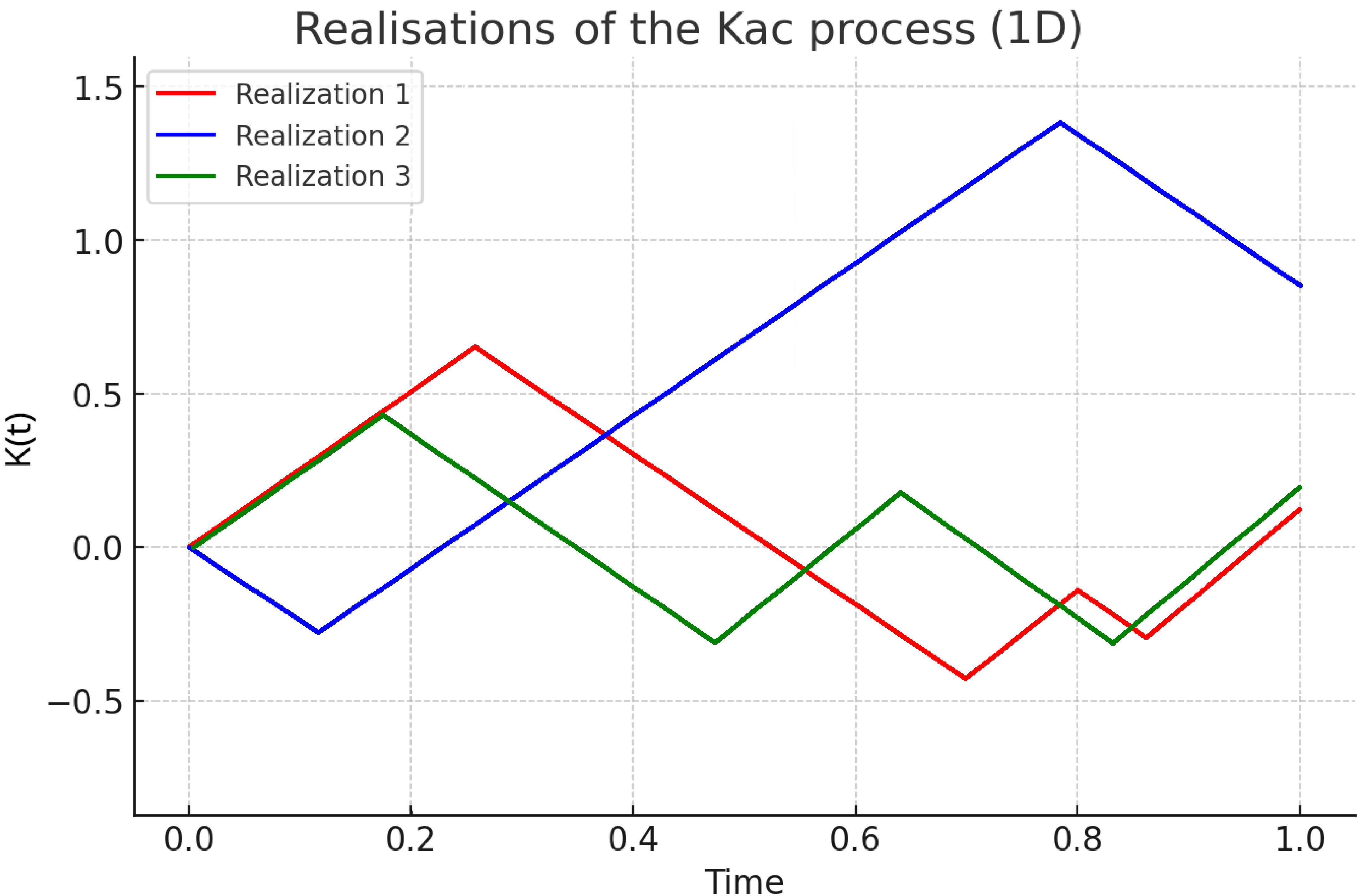
MATH+


Theoretical Foundations of Deep Learning
DFG–funded Priority Program 2298

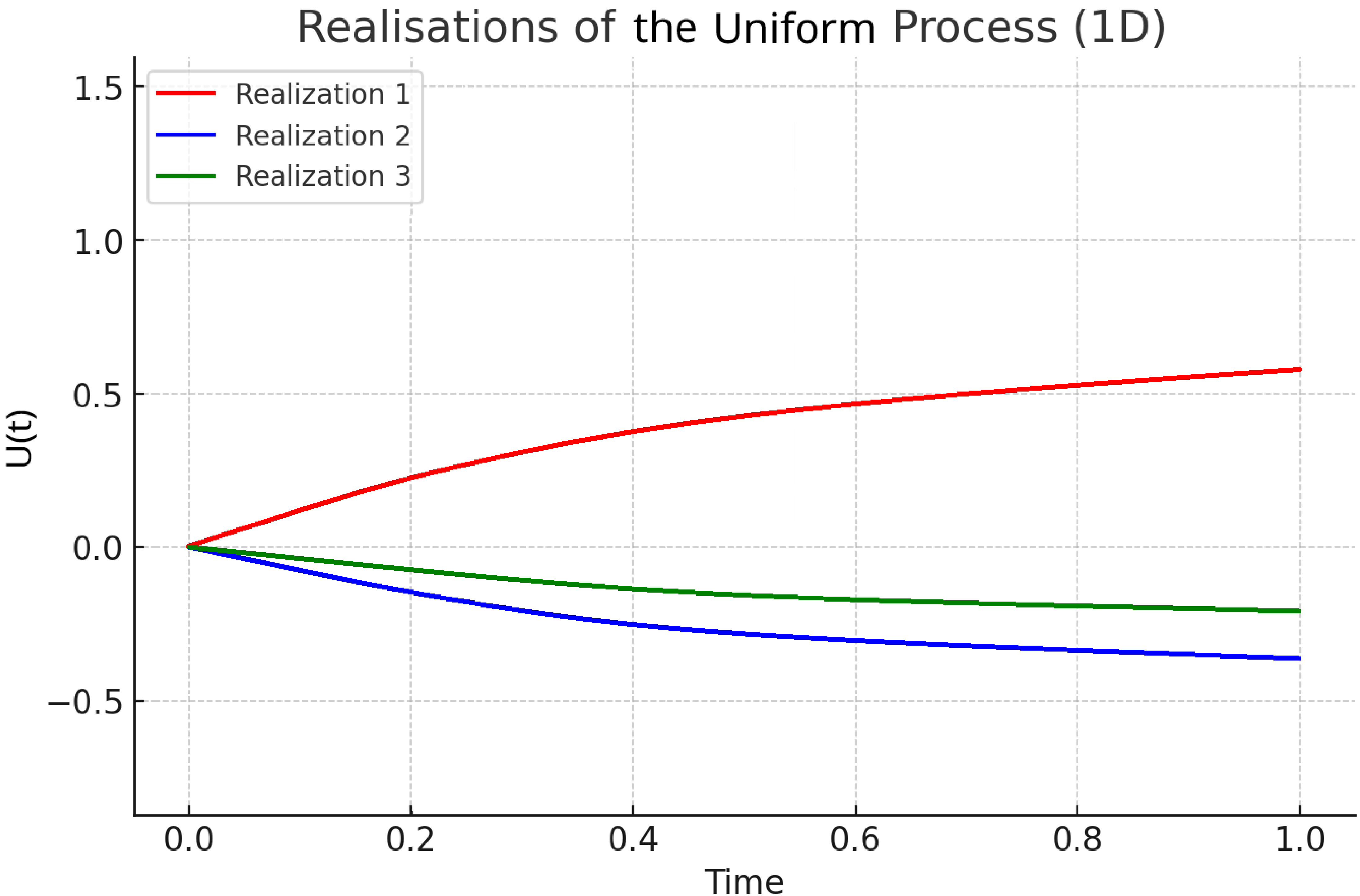
Funded under Germany's Excellence Strategy by

DFG Deutsche Forschungsgemeinschaft


Realisations of a Standard Wiener Process (1D)
Realization 1
Realization 2
Realization 3
W(t)
Time
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0



Realisations of the Kac process (1D)
Realization 1
Realization 2
Realization 3
K(t)
1.5
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0
Time



Realisations of the Uniform Process (1D)
Realization 1
Realization 2
Realization 3
1.5
1.0
0.5
0.0
−0.5
U(t)
0.0
0.2
0.4
0.6
0.8
1.0
Time

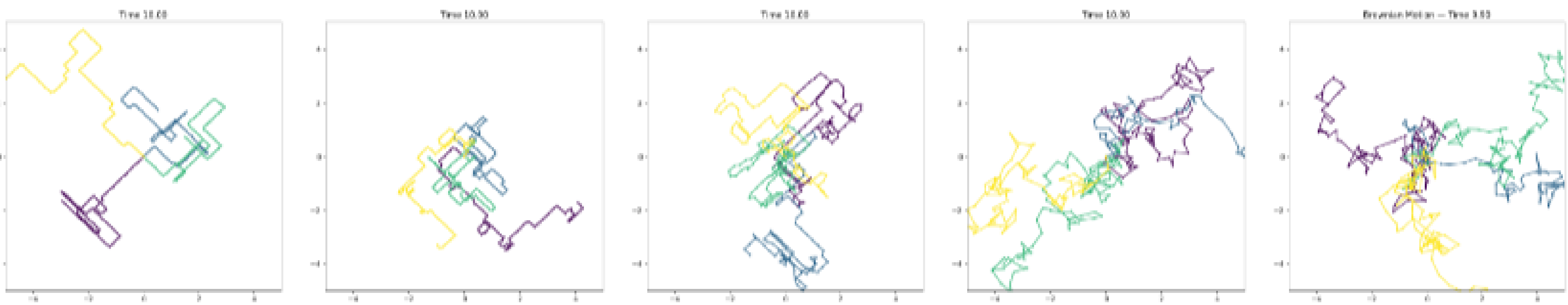


Figure 1: Paths of the componentwise Kac walk in 2D, simulated until time $T = 10$ with damping/velocity parameters $(a, c) = (1, 1), (2, 1), (4, 2), (25, 5)$, and a standard Brownian motion (right).

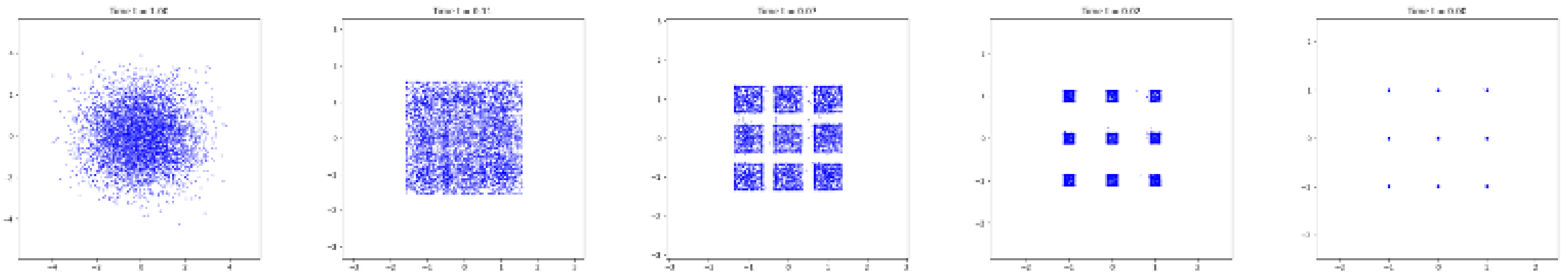
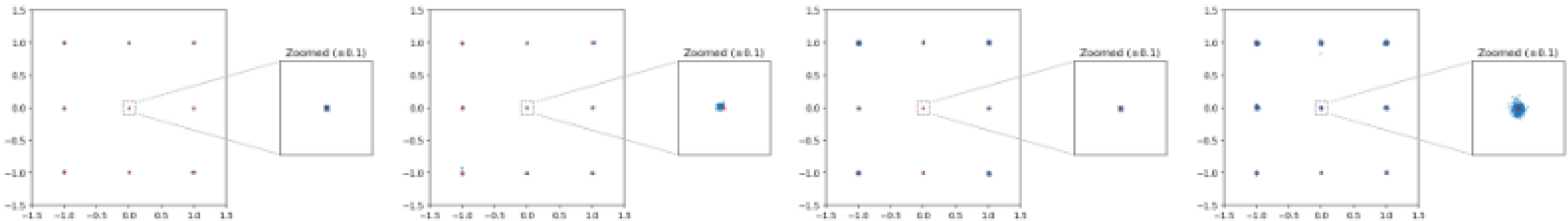


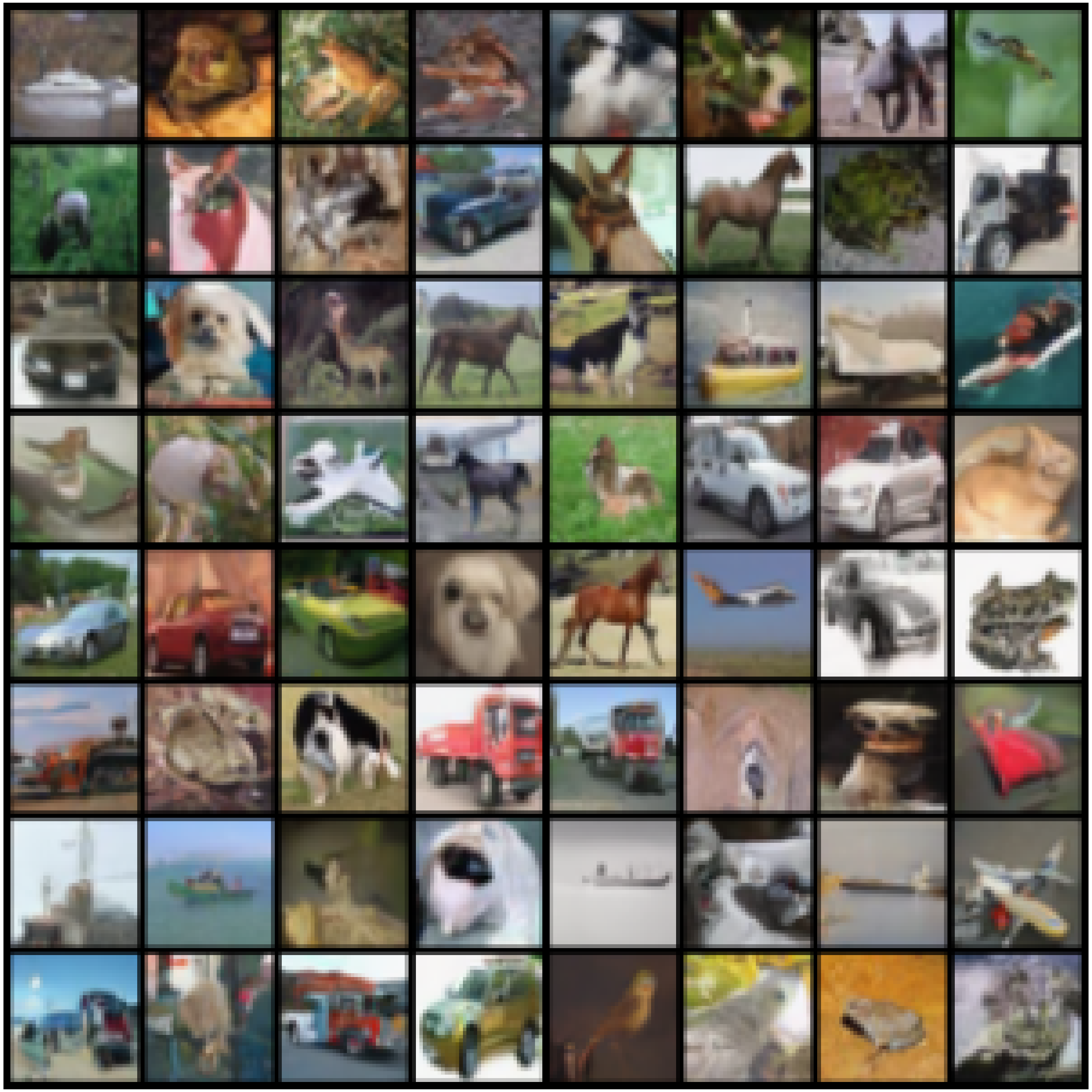
Figure 4: Backward evolution of the *learned* Kac flow for $(a, c) = (25, 5)$, see also Figure 5.



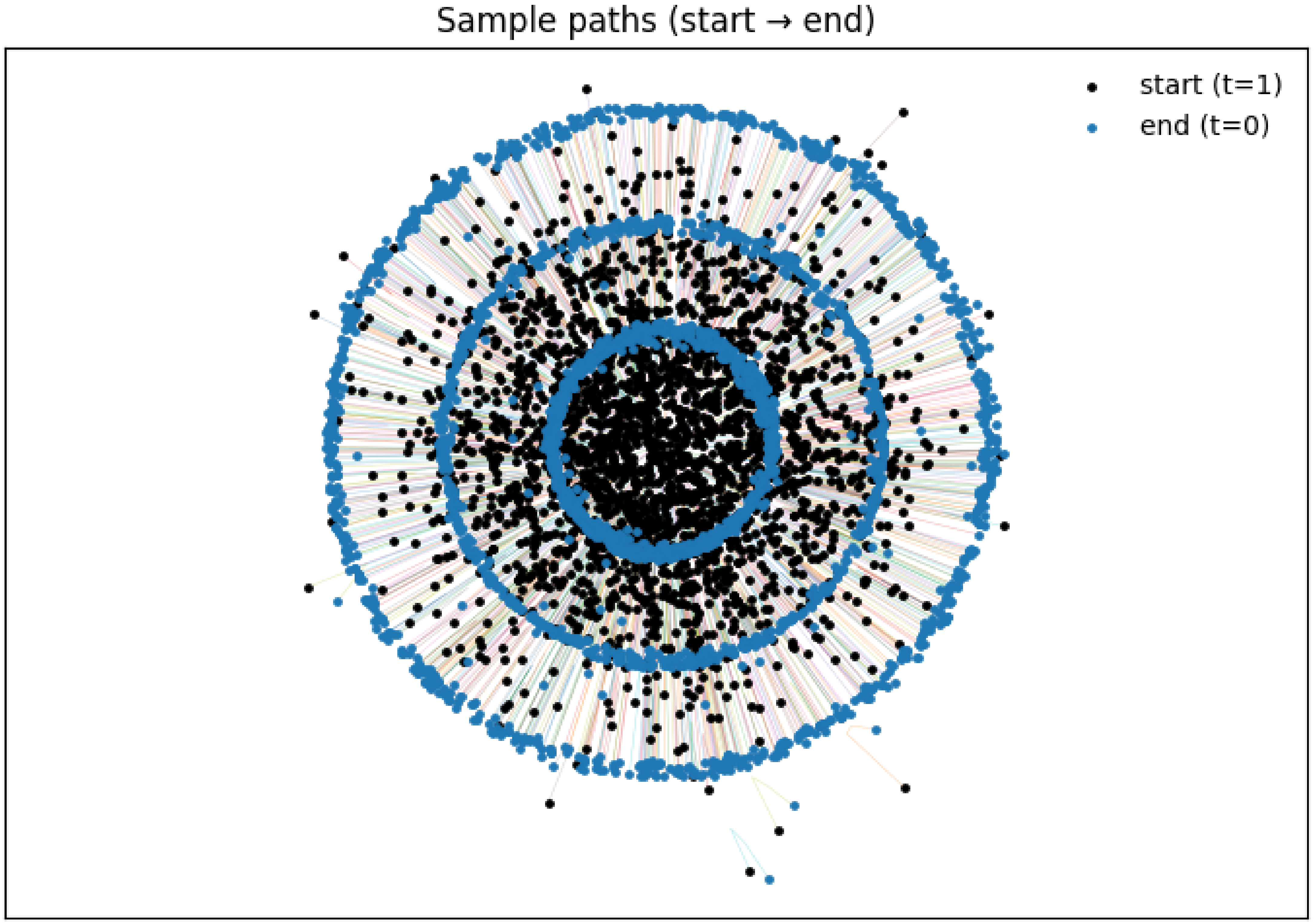
$(a, c) = (25, 5)$ 80k iterations | $(a, c) = (64, 8)$ 280k iterations | $(a, c) = (1600, 40)$ 460k iterations | diffusion 420k iterations

Figure 5: Generated samples (blue) vs. ground truth (red) at the indicated iteration for each model. The Kac model can precisely recover the small modes, while the diffusion model creates "blobs".

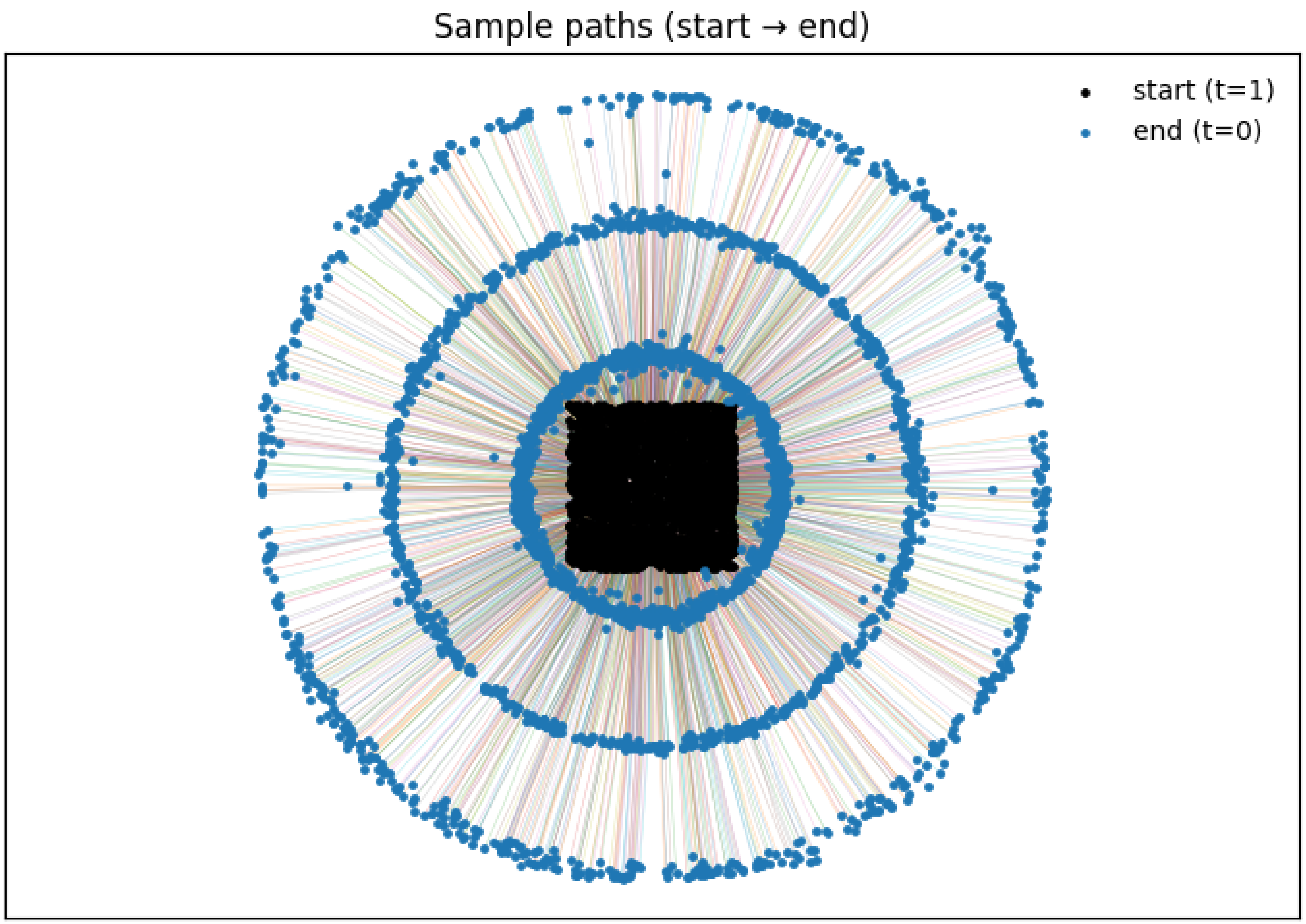
Schedule: $g(t) = t^2$			
Method	**FID**	**Method**	**FID**
$a = 900, c = 10$	**7.26**	$a = 100, c = 10$	10.01
$a = 900, c = 20$	7.46	$a = 25, c = 1$	8.60
$a = 900, c = 30$	8.05	$a = 25, c = 2$	8.65
$a = 100, c = 1$	11.41	$a = 25, c = 3$	9.15
$a = 100, c = 3$	7.77	$a = 25, c = 4$	9.56
$a = 100, c = 5$	7.73	$a = 25, c = 5$	10.70
$a = 100, c = 7$	8.58	FM (our impl.)	7.59

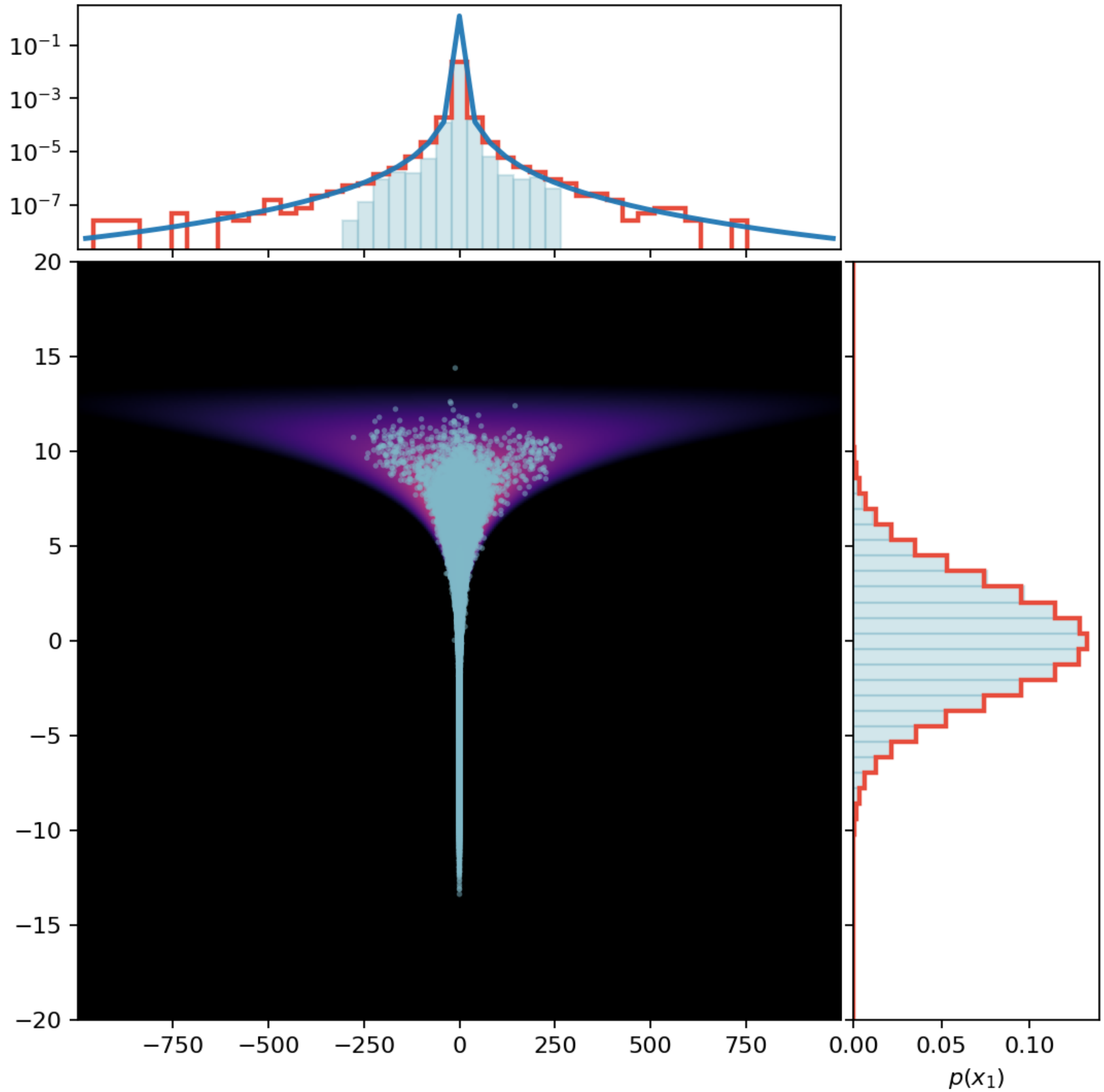


$(a, c) = (900, 10)$

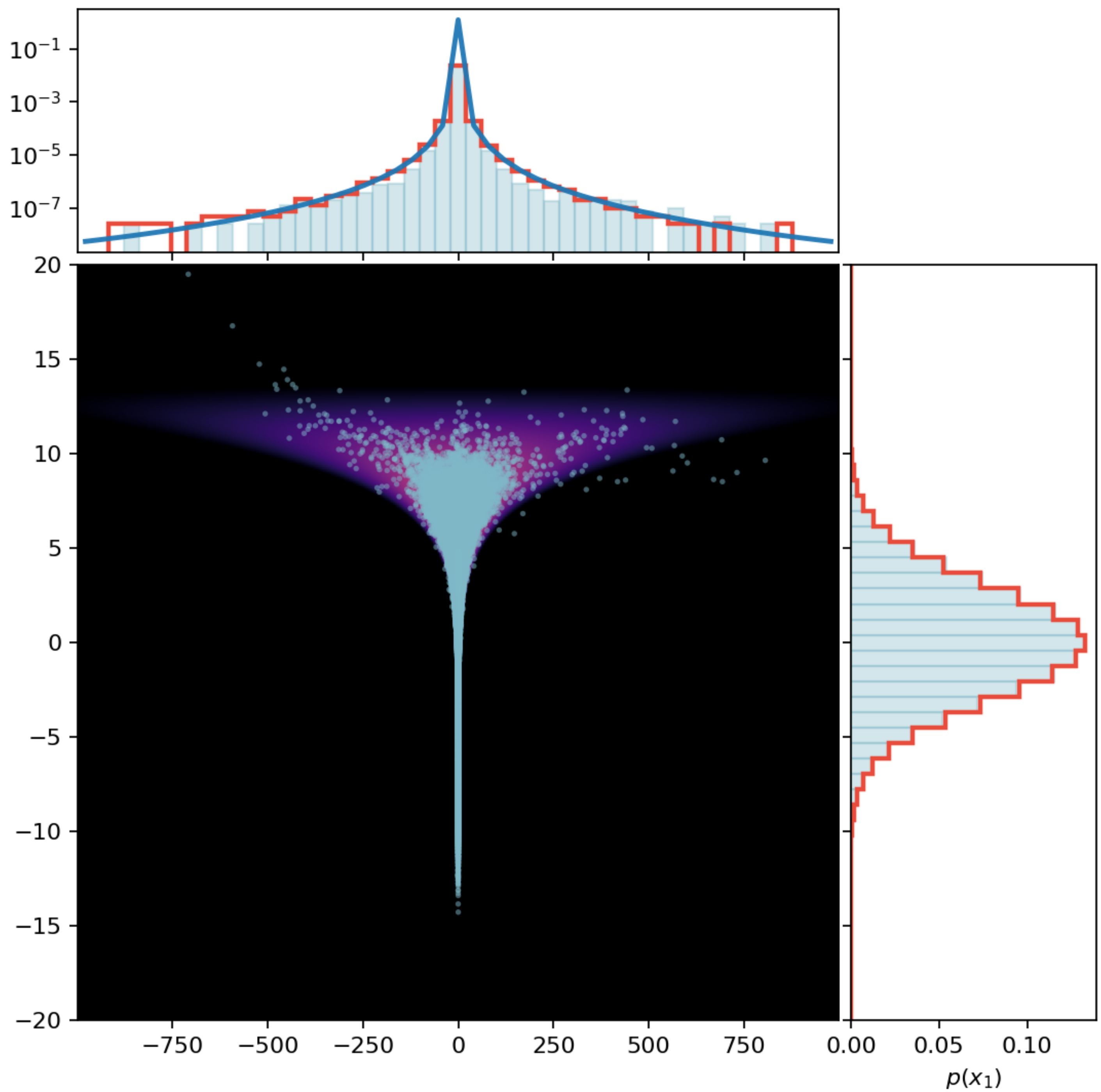


Sample paths (start → end)
start (t=1)
end (t=0)

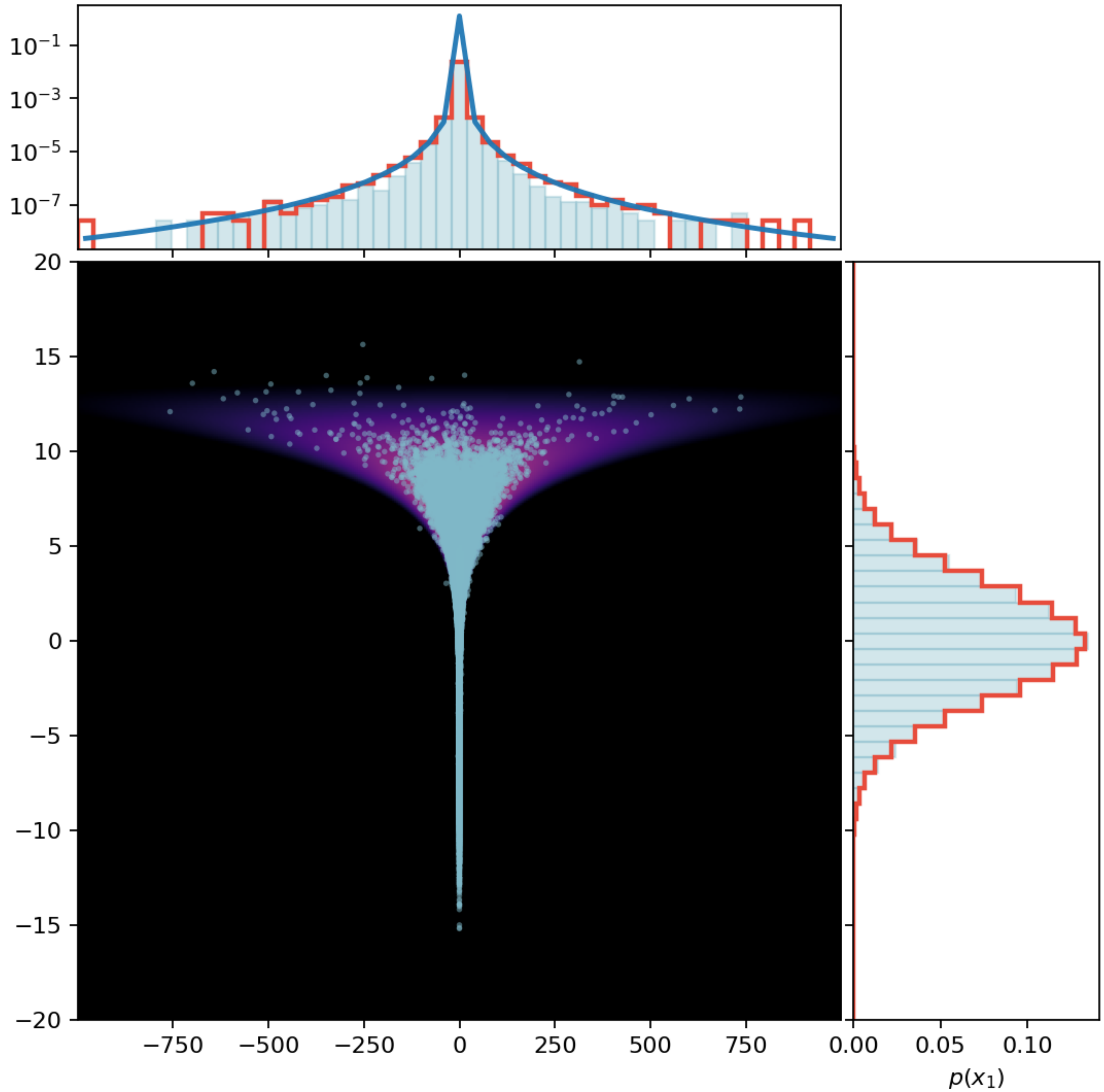

Sample paths (start → end)
start (t=1)
end (t=0)



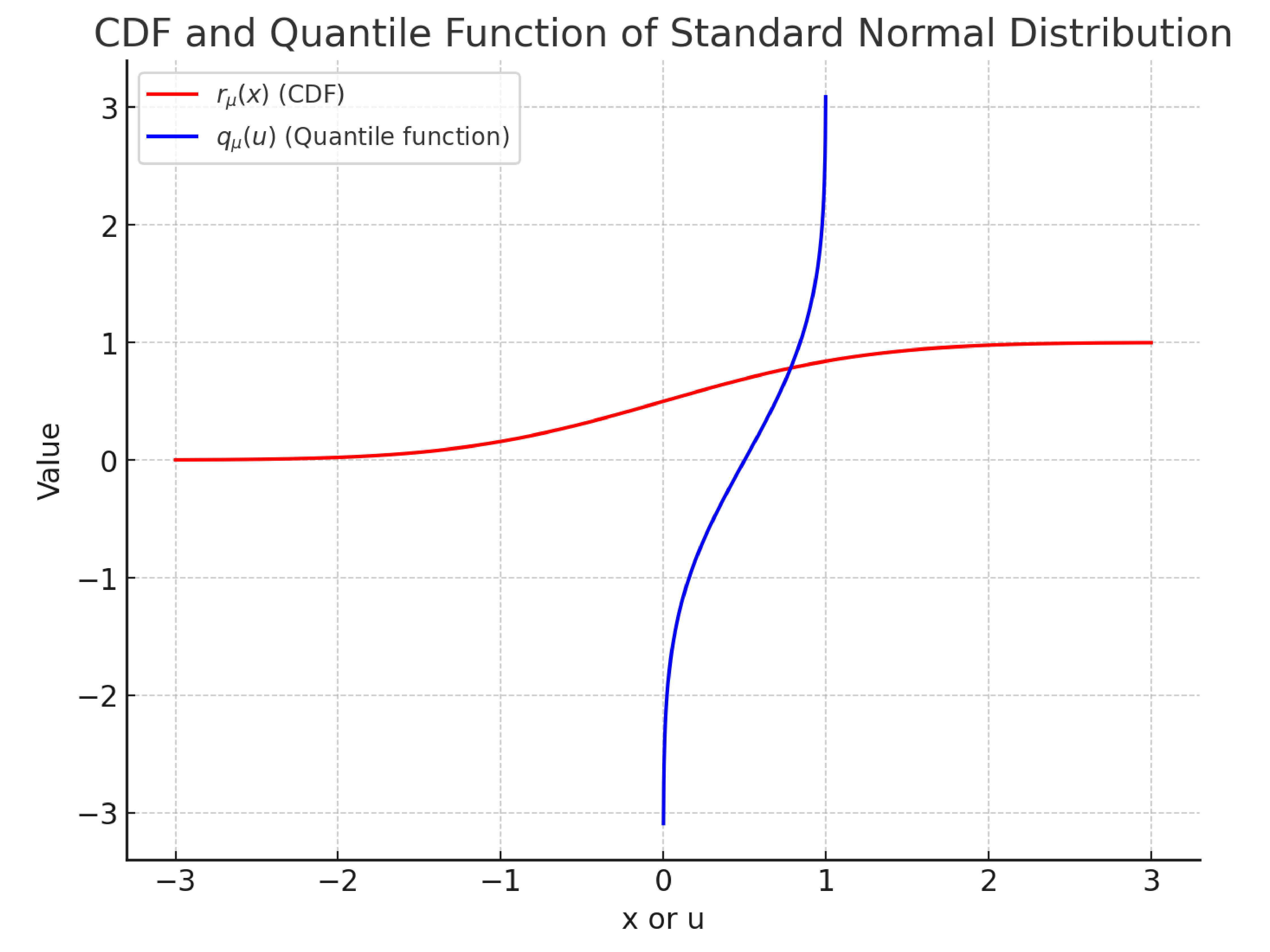
10^{-1}
10^{-3}
10^{-5}
10^{-7}
20
15
10
5
0
−5
−10
−15
−20
−750
−500
−250
0
250
500
750
0.00
0.05
0.10
$p(x_1)$



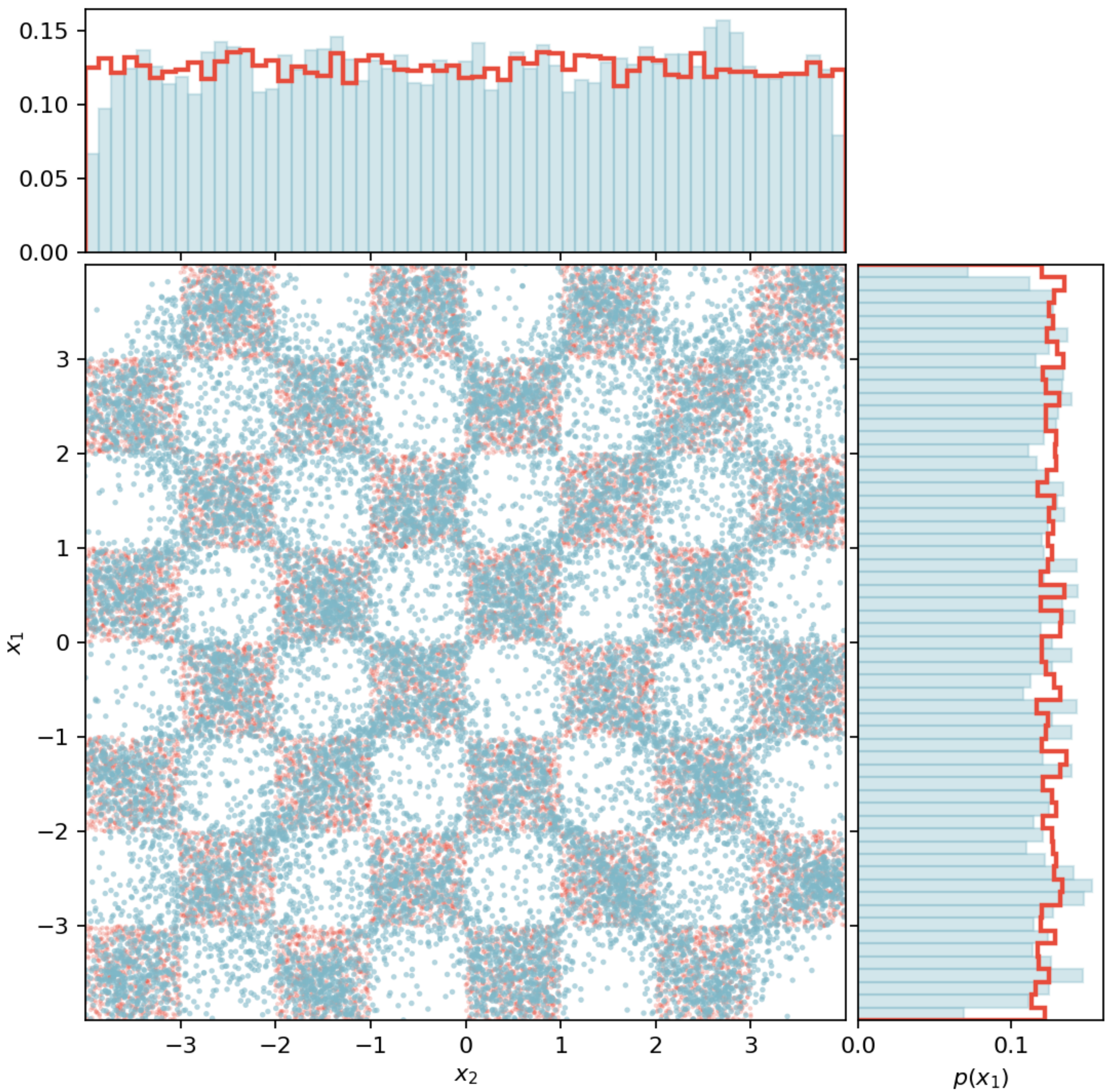
10^{-1}
10^{-3}
10^{-5}
10^{-7}
20
15
10
5
0
−5
−10
−15
−20
−750
−500
−250
0
250
500
750
0.00
0.05
0.10
$p(x_1)$

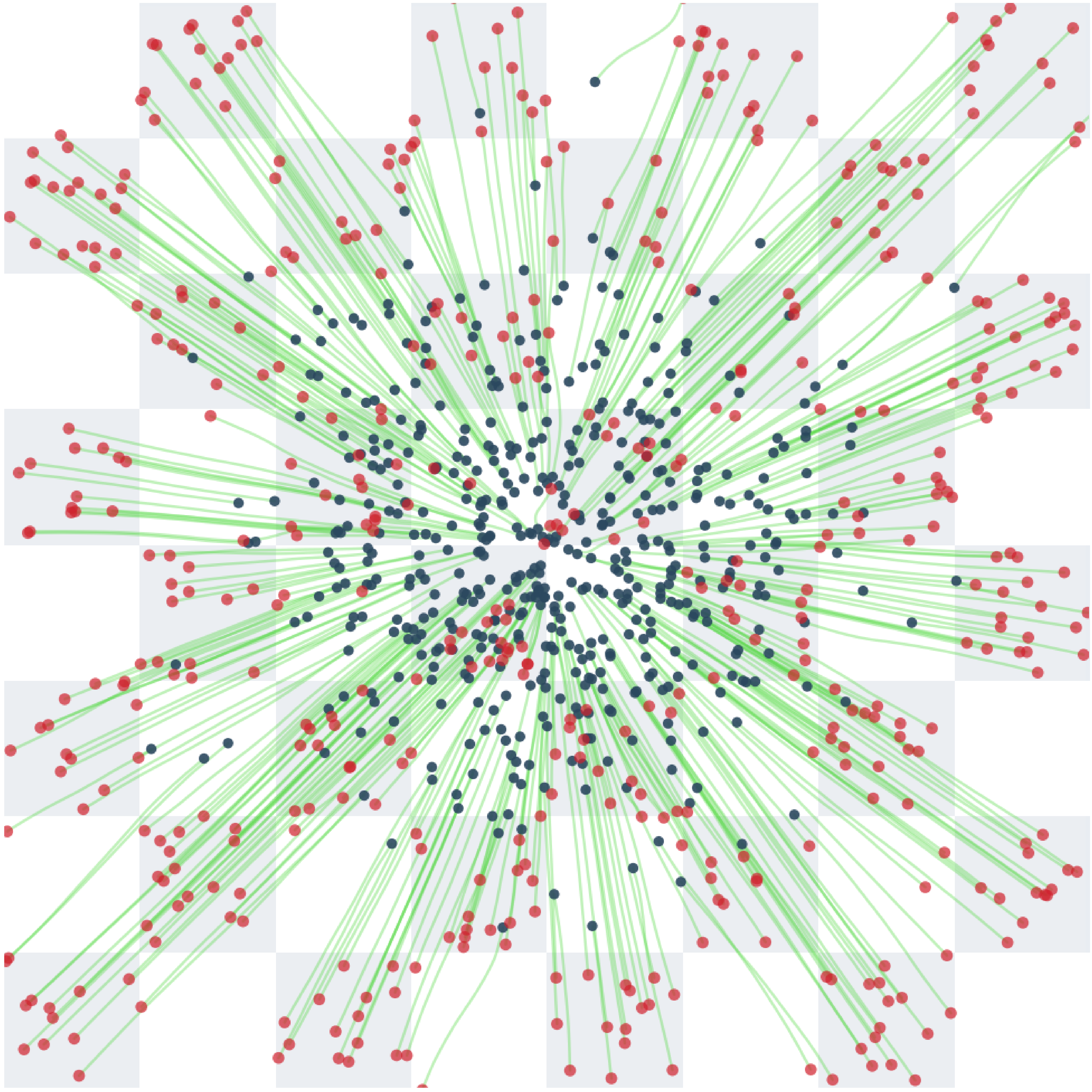


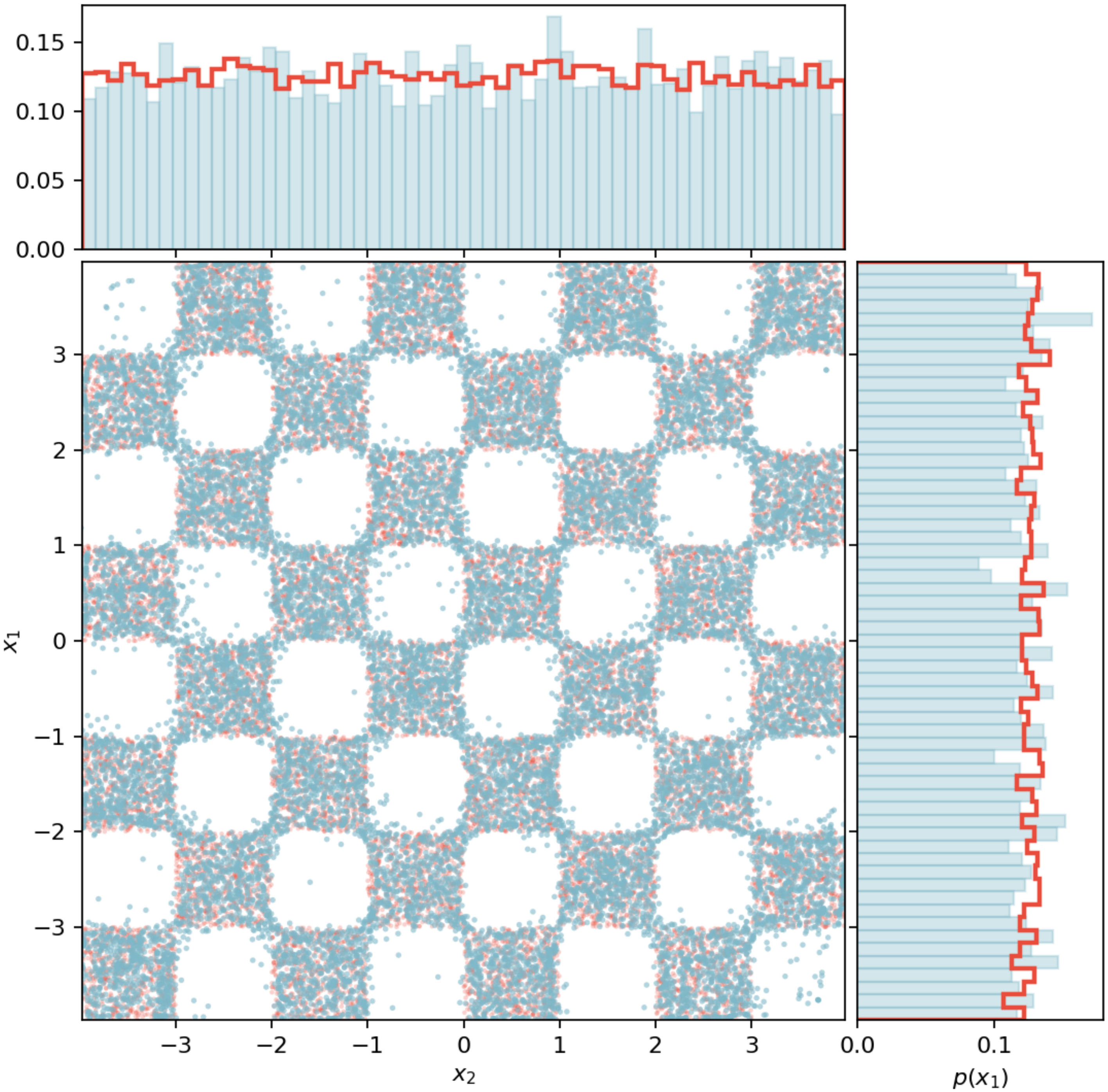
10^{-1}
10^{-3}
10^{-5}
10^{-7}
20
15
10
5
0
−5
−10
−15
−20
−750
−500
−250
0
250
500
750
0.00
0.05
0.10
$p(x_1)$

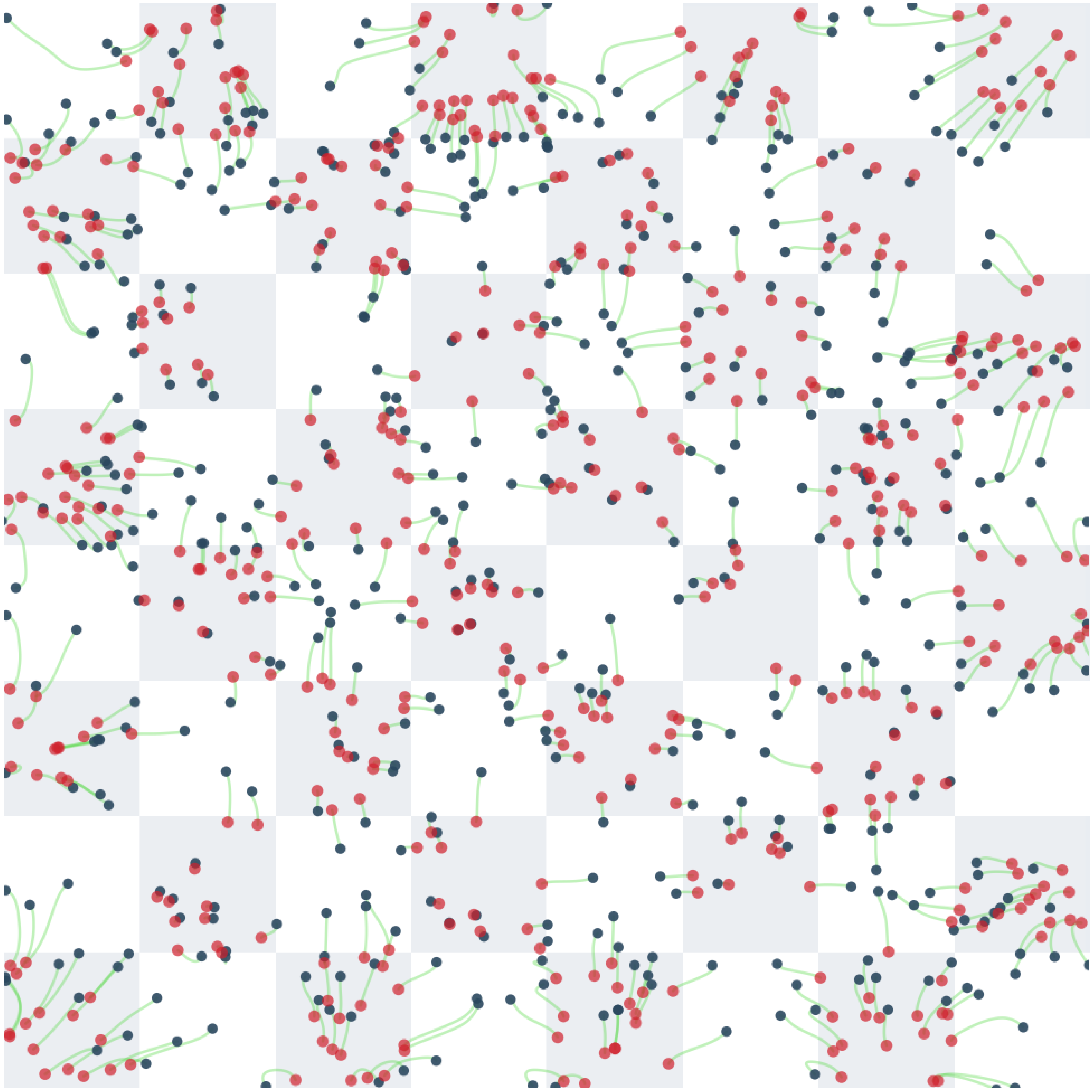


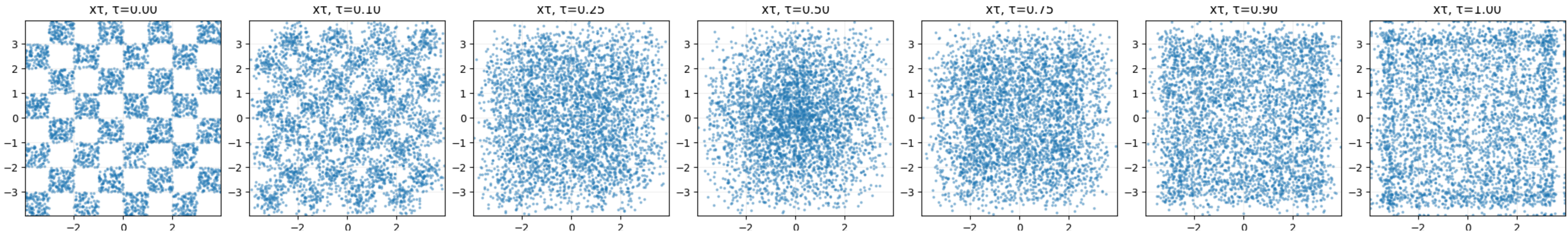
CDF and Quantile Function of Standard Normal Distribution
$r_\mu(x)$ (CDF)
$q_\mu(u)$ (Quantile function)
Value
x or u
−3
−2
−1
0
1
2
3


0.15
0.10
0.05
0.00
3
2
1
0
−1
−2
−3
x_1
−3
−2
−1
0
1
2
3
x_2
0.0
0.1
$p(x_1)$

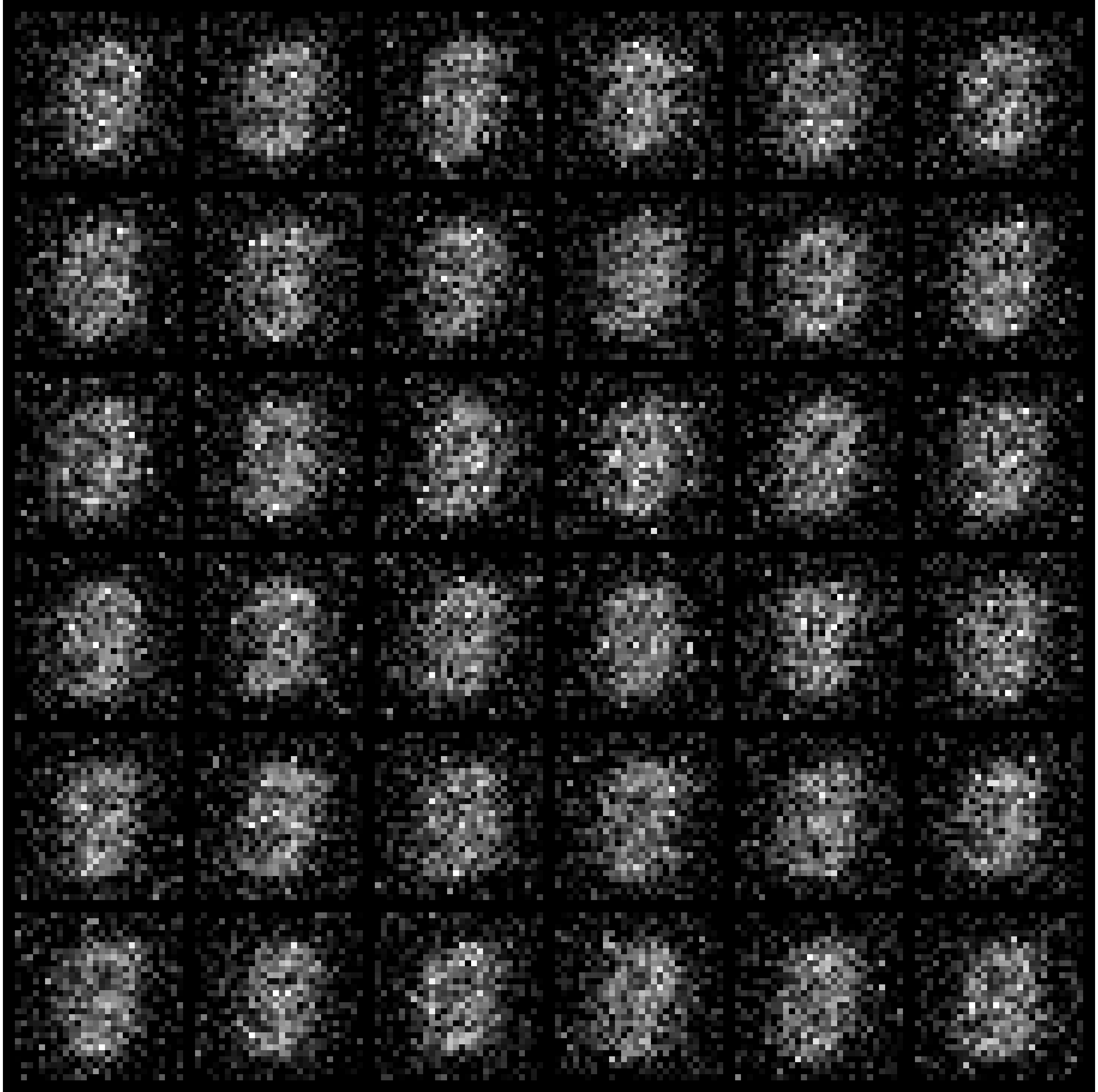


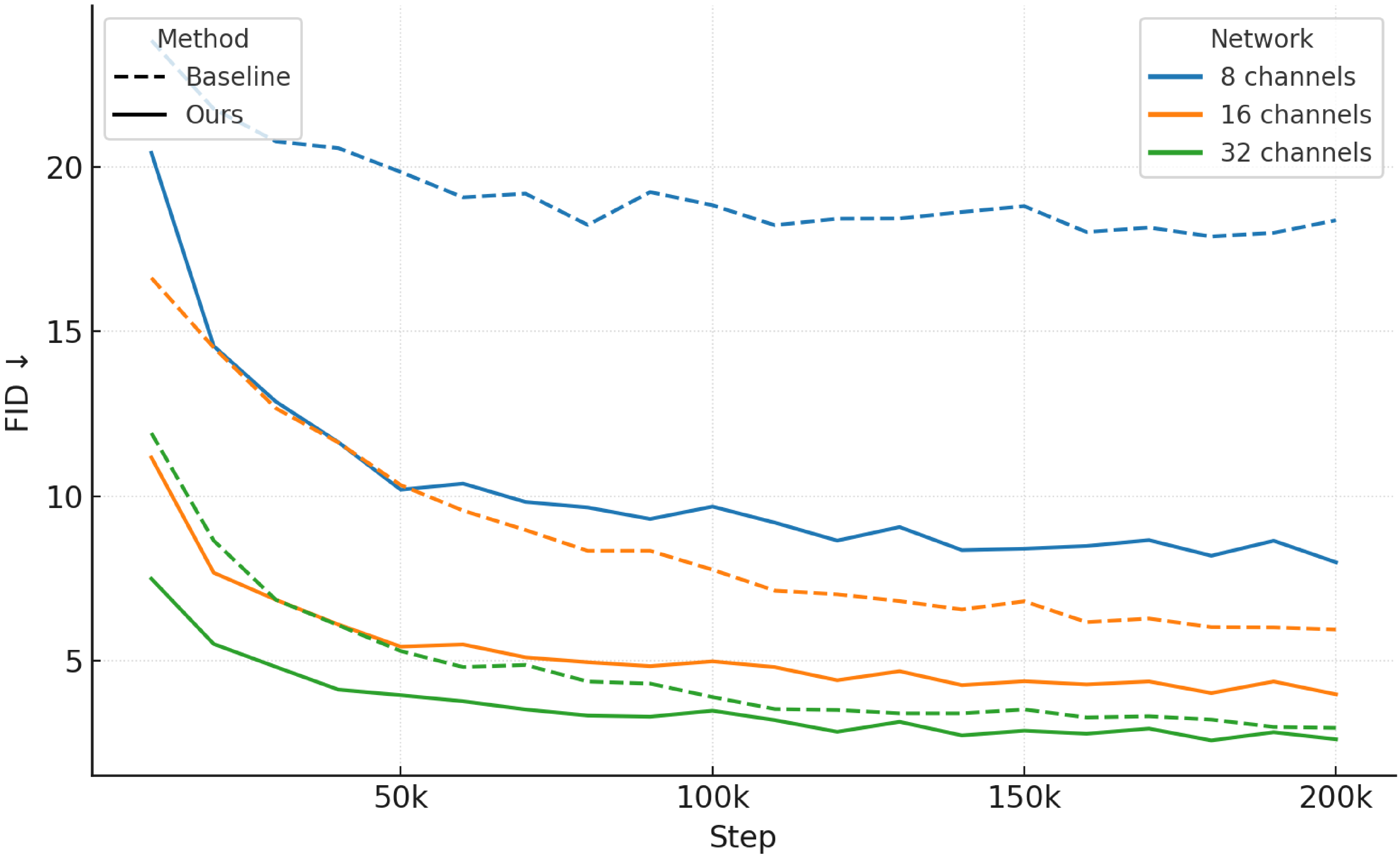

0.15
0.10
0.05
0.00
3
2
1
0
−1
−2
−3
x_1
−3
−2
−1
0
1
2
3
x_2
0.0
0.1
$p(x_1)$

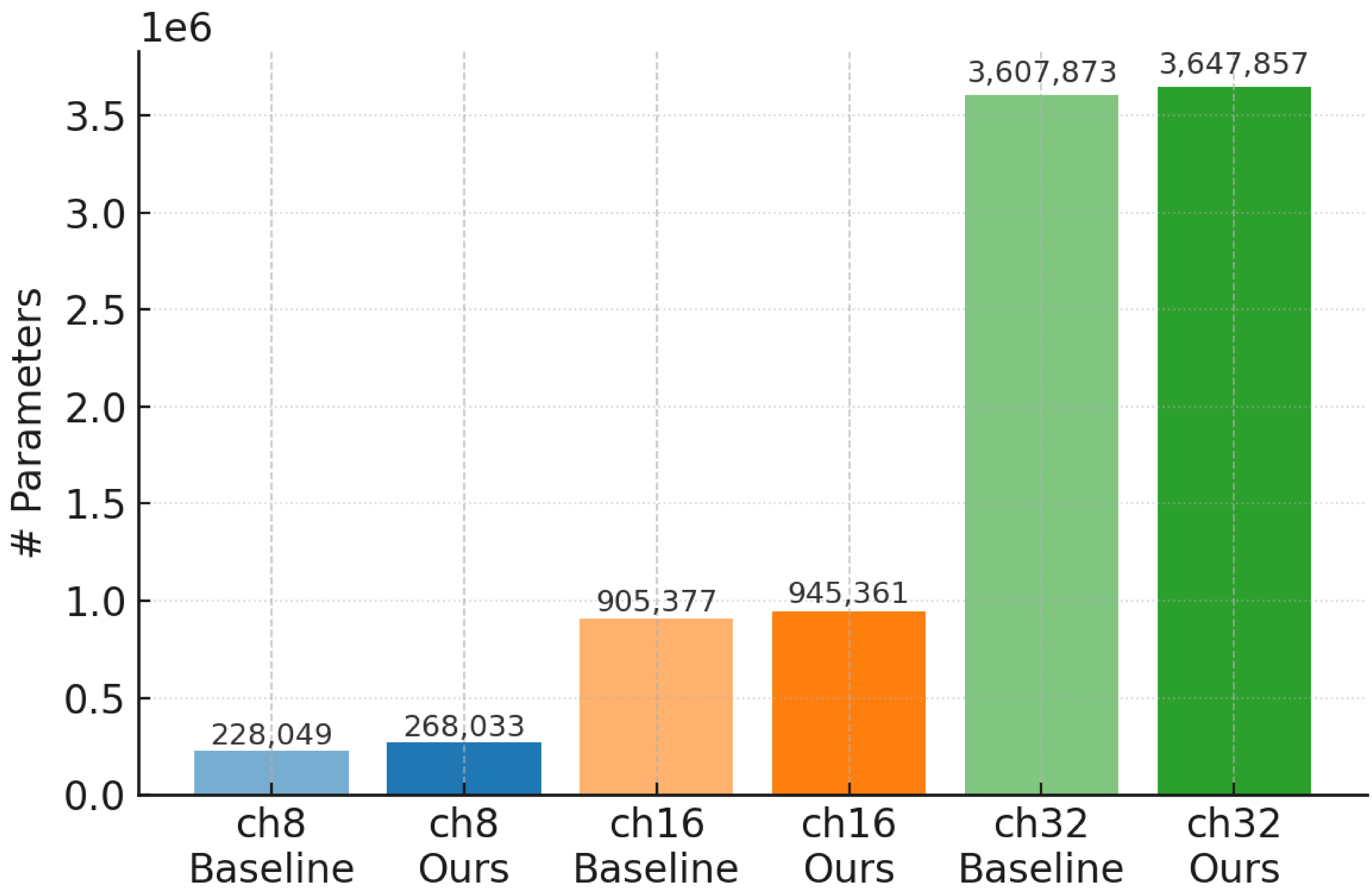




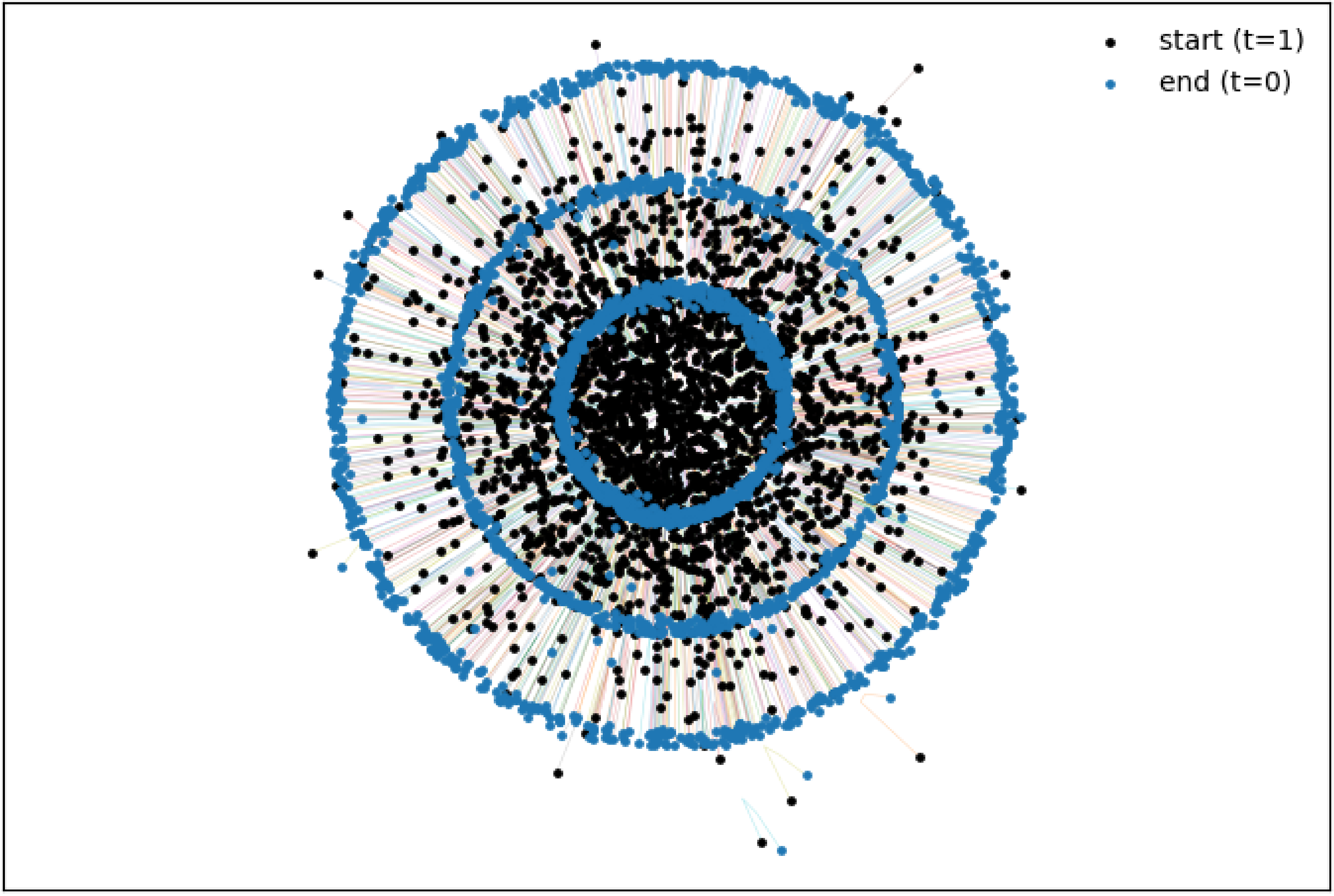
xτ, τ=0.00
xτ, τ=0.10
xτ, τ=0.25
xτ, τ=0.50
xτ, τ=0.75
xτ, τ=0.90
xτ, τ=1.00




Method
Baseline
Ours
Network
8 channels
16 channels
32 channels
FID ↓
20
15
10
5
50k
100k
150k
200k
Step


1e6
Parameters
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
228,049
268,033
905,377
945,361
3,607,873
3,647,857
ch8 Baseline
ch8 Ours
ch16 Baseline
ch16 Ours
ch32 Baseline
ch32 Ours

Sample paths (start → end)



Sample paths (start → end)