About Flow Matching generative models

Anatole Gallouët

December 2025

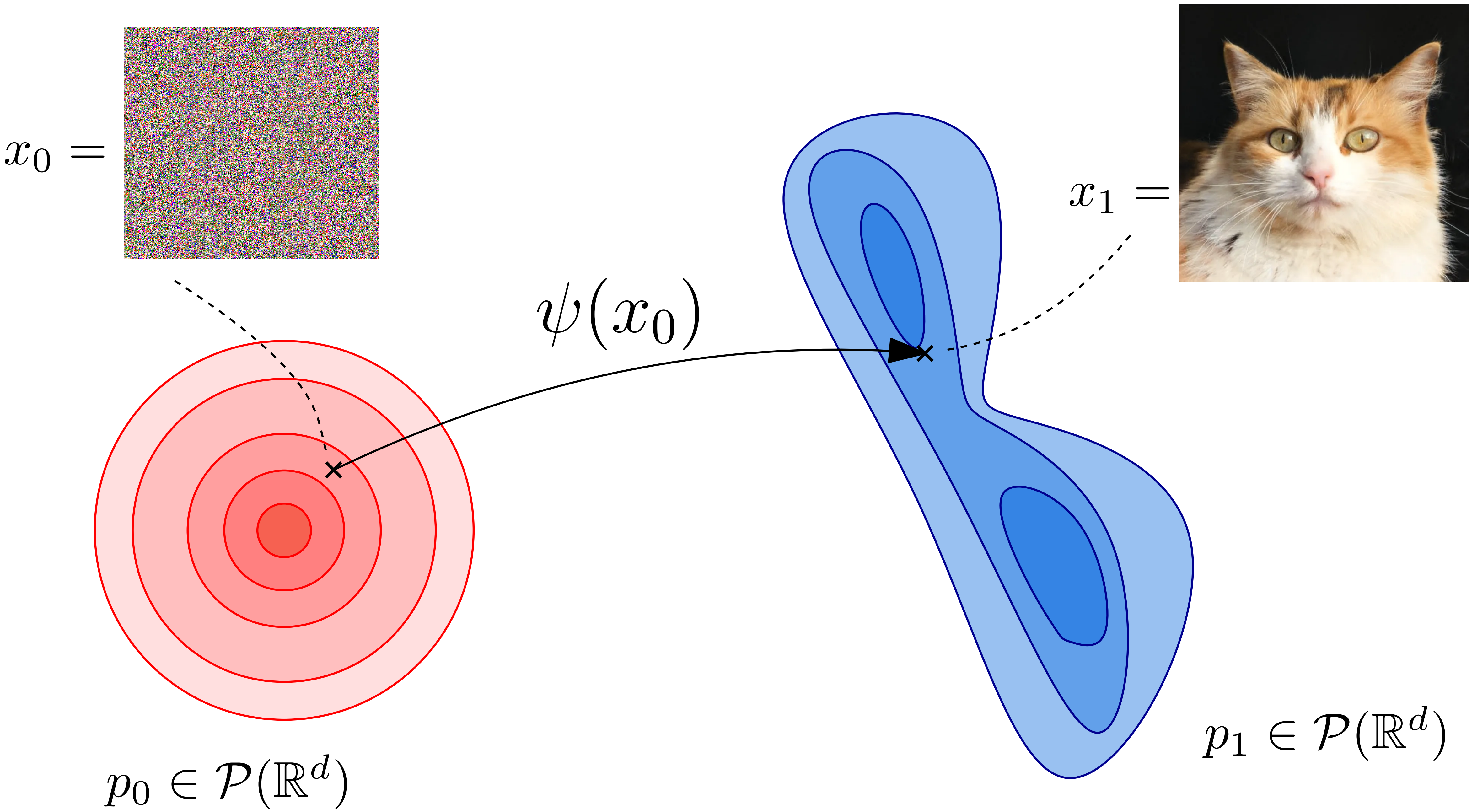
Bibliography

- *[Lipman et al. "Flow matching for generative modeling." ICLR 2023]*
- *[Tong et al. "Improving and generalizing flow-based generative models with minibatch optimal transport." TMLR (2024)]*
- *[Gagneux et al. "A visual dive into conditional flow matching." The Fourth Blogpost Track at ICLR 2025.]*

(Non exhaustive list...)

Introduction

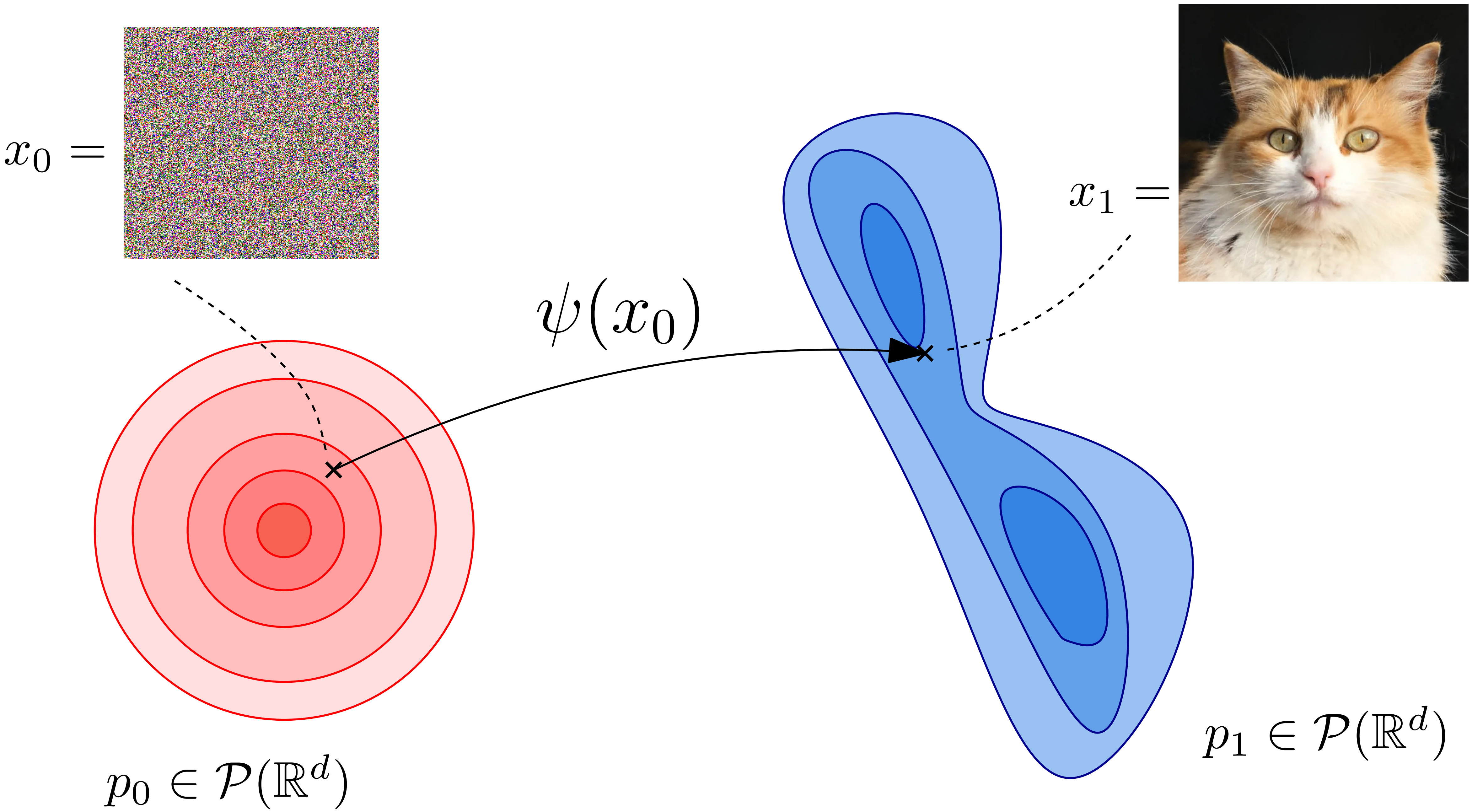
Generative modeling



Goal: Find ψ such that for $x_0 \sim p_0$, $\psi(x_0) \sim p_1$.

Introduction

Generative modeling



Goal: Find ψ such that for $x_0 \sim p_0$, $\psi(x_0) \sim p_1$.

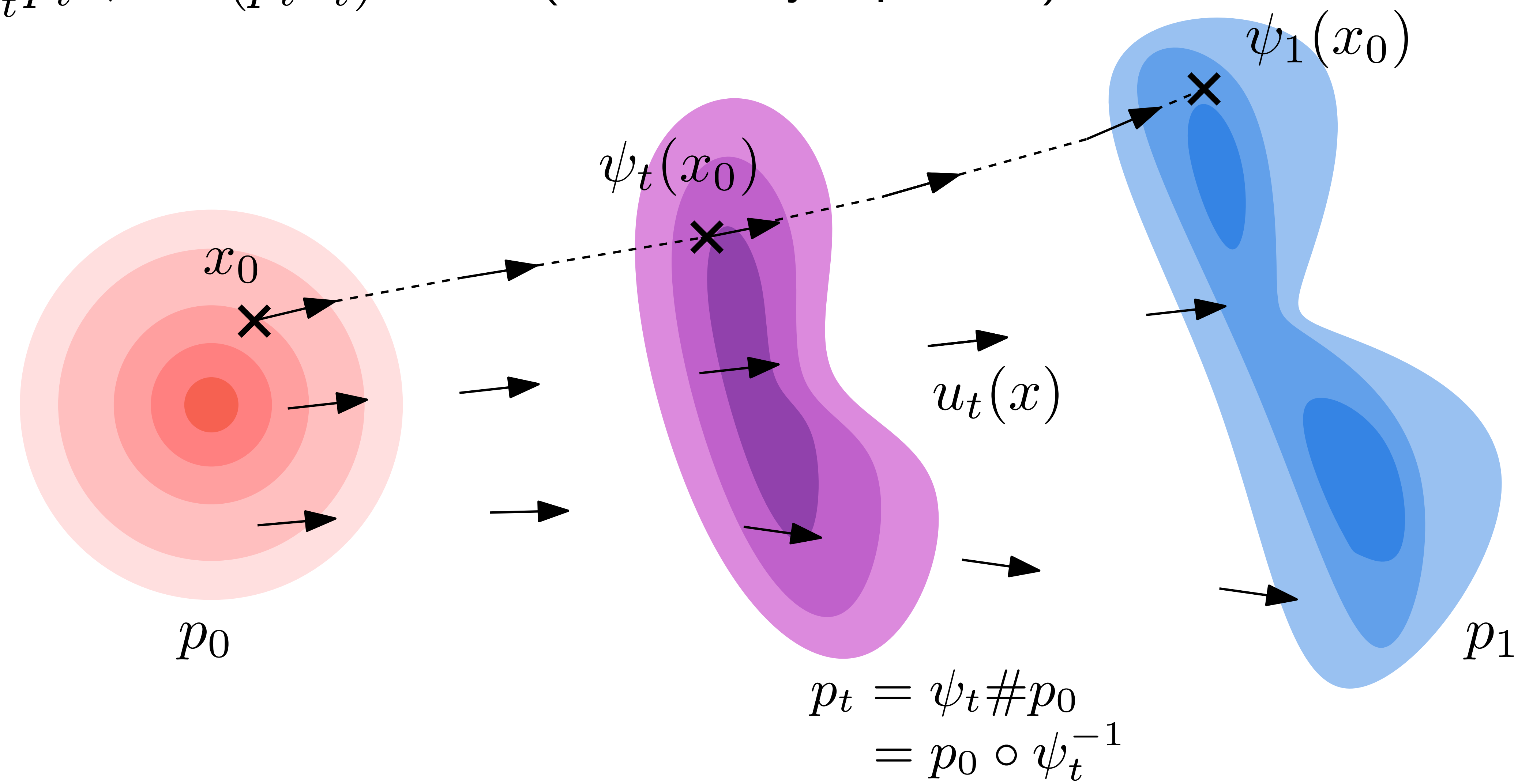
Problem: Dimension d is big, and p_1 is only known from data.

ODEs and probability flows

- Velocity field $u : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$
- Flow map $\psi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$
- Probability path $p_t \in \mathcal{P}(\mathbb{R}^d)$.

$$\forall x \in \mathbb{R}^d \begin{cases} \psi_0(x) = x \\ \frac{\partial}{\partial t}\psi_t(x) = u_t(\psi_t(x)) \end{cases} \qquad \text{(Flow ODE)}$$

$$\frac{\partial}{\partial t}p_t + \operatorname{div}(p_t u_t) = 0 \quad \text{(Continuity equation)}$$



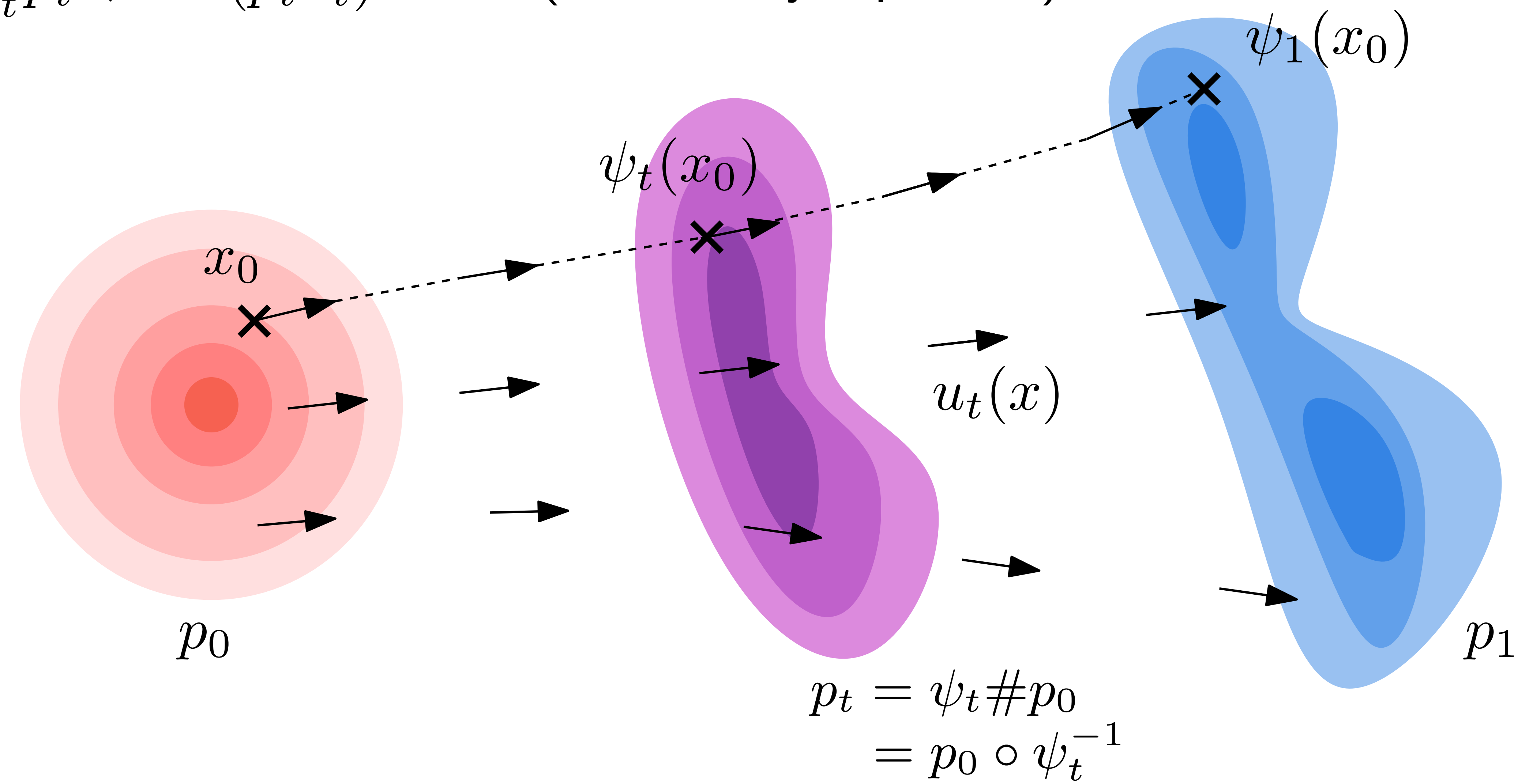
ODEs and probability flows

- Velocity field $u : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$
- Flow map $\psi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$
- Probability path $p_t \in \mathcal{P}(\mathbb{R}^d)$.

Question :
Can I find a $u_t(x)$ to follow (p_t) between 0 and 1.

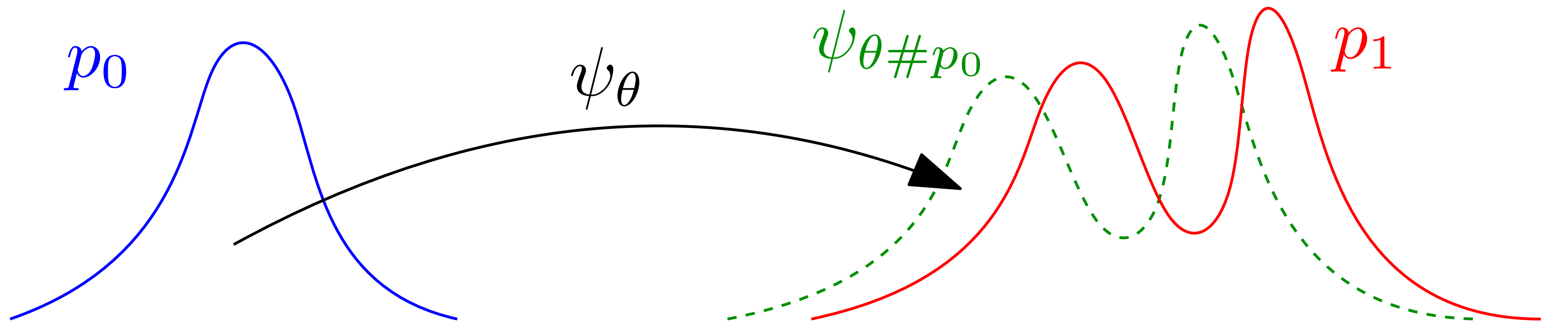
$$\forall x \in \mathbb{R}^d \begin{cases} \psi_0(x) = x \\ \frac{\partial}{\partial t}\psi_t(x) = u_t(\psi_t(x)) \end{cases} \quad \text{(Flow ODE)}$$

$$\frac{\partial}{\partial t}p_t + \mathrm{div}(p_t u_t) = 0 \quad \text{(Continuity equation)}$$



(Continuous) Normalizing flows

Main idea: Minimize KL divergence or maximize log-likelihood



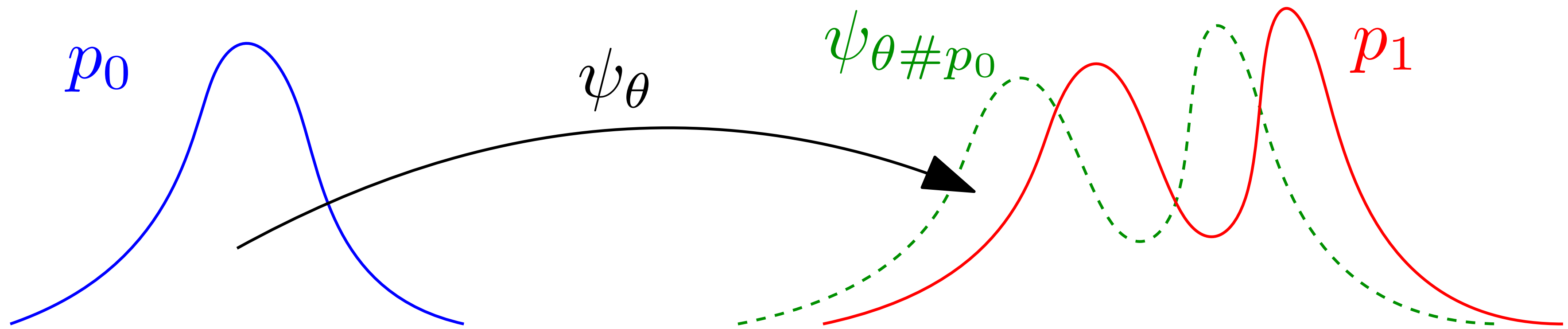
$$\mathrm{KL}(p_1|\psi_{\theta\#p_0}) = -\mathbb{E}_{x\sim p_1}(\log(\psi_{\theta\#p_0}(x)) + cst$$

$$= -\mathbb{E}_{x\sim p_1}[\log(p_0(\psi_\theta^{-1}(x)) + \log(\det(J_{\psi_\theta^{-1}}(x))] + cst$$

(change of variable)

(Continuous) Normalizing flows

Main idea: Minimize KL divergence or maximize log-likelihood



$$\mathrm{KL}(p_1|\psi_{\theta\#p_0}) = -\mathbb{E}_{x\sim p_1}(\log(\psi_{\theta\#p_0}(x)) + cst$$

$$= -\mathbb{E}_{x\sim p_1}[\log(p_0(\psi_\theta^{-1}(x)) + \log(\det(J_{\psi_\theta^{-1}}(x))] + cst$$

- The map ψ_θ has to be invertible.
- We need to have access to $J_{\psi_\theta^{-1}}$.

Choice : $\psi_\theta = \underbrace{f_K \circ \cdots \circ f_1}$

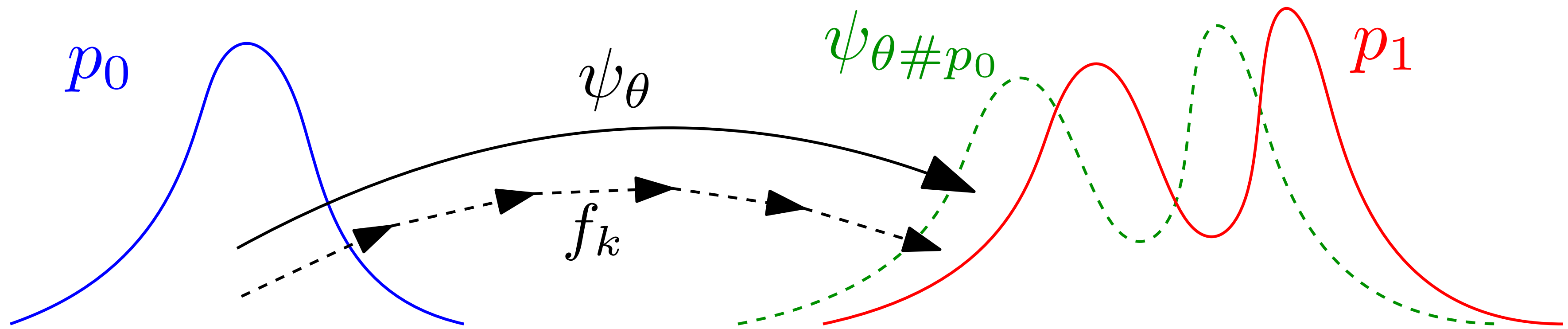
f_k simple and invertible

$$\log(p_0(\psi_\theta^{-1}) = \log(p_0(f_1^{-1} \circ \cdots \circ f_K^{-1}))$$

$$\log(\det(J_{\psi_\theta^{-1}}) = \sum_k \log(\det(J_{f_k^{-1}}))$$

(Continuous) Normalizing flows

Main idea: Minimize KL divergence or maximize log-likelihood



In particular: $\psi_\theta = f_K \circ \cdots \circ f_1$ with $f_k(x) = x + h_k(x)$

Then
$$x_k = f_k(x_{k-1}) = x_{k-1} + h_k(x_{k-1}) = x_{k-1} + \frac{1}{K} u_{k-1}(x_{k-1})$$

For suitable h_k and u

So $x_K = \psi_\theta(x_0)$ is the Euler discretization of

$$\begin{cases} x(0) = x_0 \\ \frac{\partial}{\partial t} x(t) = u_t(x(t)) \end{cases}$$

(Flow ODE) ← (invertible by backward integration)

CNF: Train directly $v_\theta(t, x)$ maximizing $-\int_0^1 \mathrm{div}(v_\theta(x(t), t))$ (log-likelihood)

Advantages and limitations

- Regular Normalizing Flows have limited architectures.
- Continuous Normalizing Flows have some advantages:
  - Less restrictive: choose any u Lipschitz in space ad cont. in time.
  - Inversion is easier (only integrate from 1 to 0).
  - Likelihood easier to compute, no log of determinant.
- CNF are unstable in high dimension:
  - Training with log-likelihood does not scale well to high dimension.
  - There is an infinite number of probability path p_t.

Conditional Flow Matching: Fix a specific vector field & probabilty path

Conditional Flow Matching

Goal: Train $v_\theta(t,x)$ with $\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,p_t(x)} \|v_\theta(t,x) - u_t(x)\|^2$

untractable

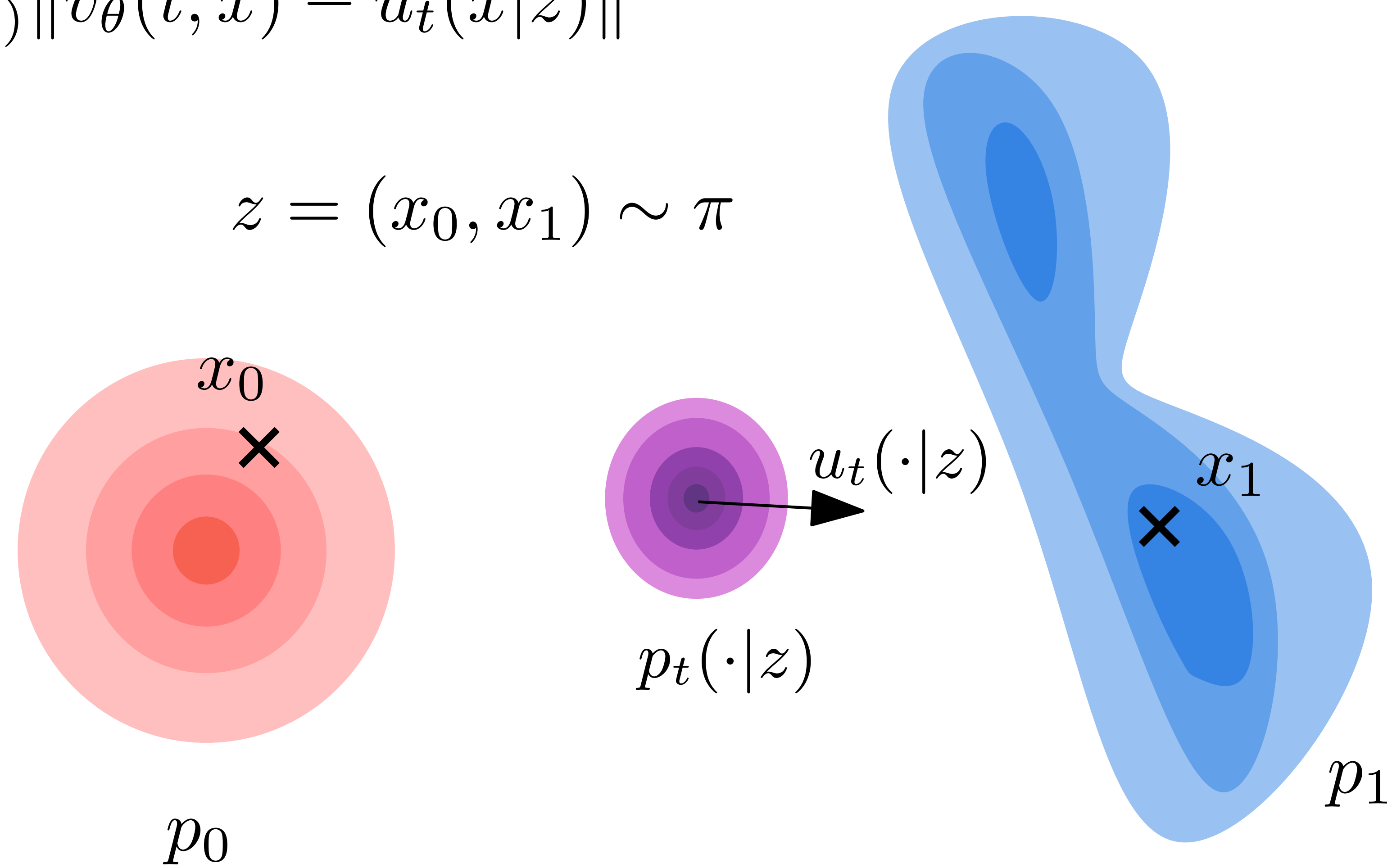
Conditioning:

- Choose $\pi(z) \in \Pi(p_0, p_1)$ — $\pi(\cdot, \Omega) = p_0$, $\pi(\Omega, \cdot) = p_1$
- Choose cond. path $p_t(x|z)$ — Defines a unique $u_t(x|z)$ via (continuity eq.)
- $\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\pi(z),p_t(x|z)} \|v_\theta(t,x) - u_t(x|z)\|^2$

Example :

- $\pi = p_0 \otimes p_1$
- $p_t(x|z) = \delta_{(1-t)x_0 + tx_1}$

Then $u_t(x|z) = x_1 - x_0$.



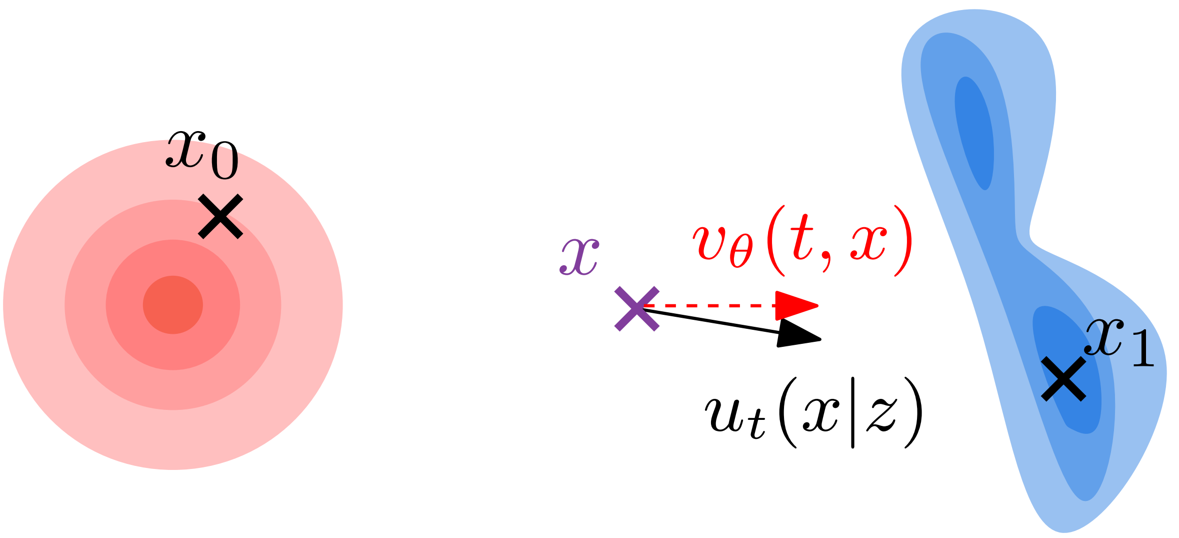
Conditional Flow Matching

Theorem : For any $z \sim \pi$ with $\pi \in \Pi(p_0, p_1)$, the expectancies of the probability path $p_t(x) = \mathbb{E}_{\pi(z)}[p_t(x|z)]$ and of the vector field $u_t(x) = \mathbb{E}_{\pi(z)}[u_t(x|z)]$ are solution of the same continuity equation.

Theorem : (FM and CFM loss are equivalent) *[Lipman et. al 23]*

$$\begin{aligned}\mathcal{L}_{\mathrm{CFM}}(\theta) &= \mathbb{E}_{t\sim\mathcal{U}(0,1),\pi(z),p_t(x|z)}\|v_\theta(t,x) - u_t(x|z)\|^2 \\ &= \mathbb{E}_{t\sim\mathcal{U}(0,1),p_t(x)}\|v_\theta(t,x) - u_t(x)\|^2 + \mathrm{cst} \\ &= \mathcal{L}_{\mathrm{FM}}(\theta) + \mathrm{cst}\end{aligned}$$

- Training: minimize $\mathcal{L}_{\mathrm{CFM}}$ with $\begin{cases} t \sim \mathcal{U}(0,1) \\ x_0 \sim p_0 \\ x_1 \sim p_{\mathrm{data}} \end{cases}$
- Sampling: $x_{k+1} = x_k + \frac{1}{N} v_\theta(t_k, x_k)$

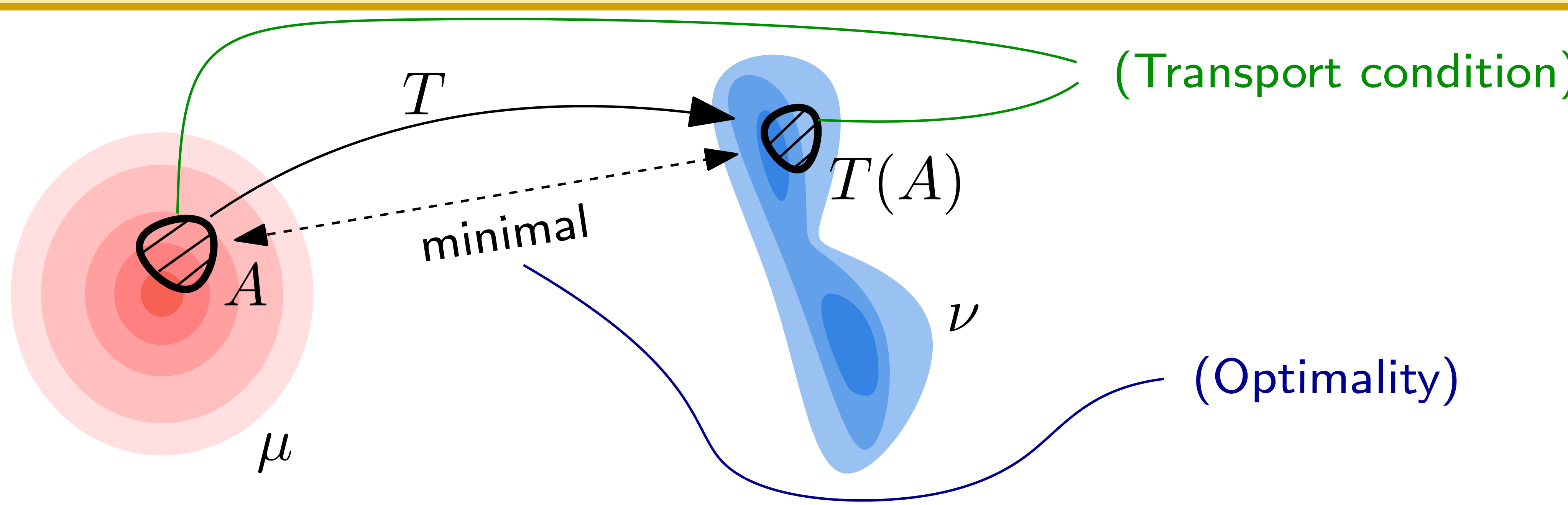


What about Optimal Transport ?

Definition: (Wasserstein distance)
The Wasserstein distance between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$\mathrm{W}_p^p(\mu, \nu) = \inf_{T_{\#\mu}=\nu} \int_{\mathbb{R}^d} \|x - T(x)\|_p^p \,\mathrm{d}\,\mu(x)$$

where $T_{\#\mu} = \mu \circ T^{-1}$ is the pushforward measure of μ by T.

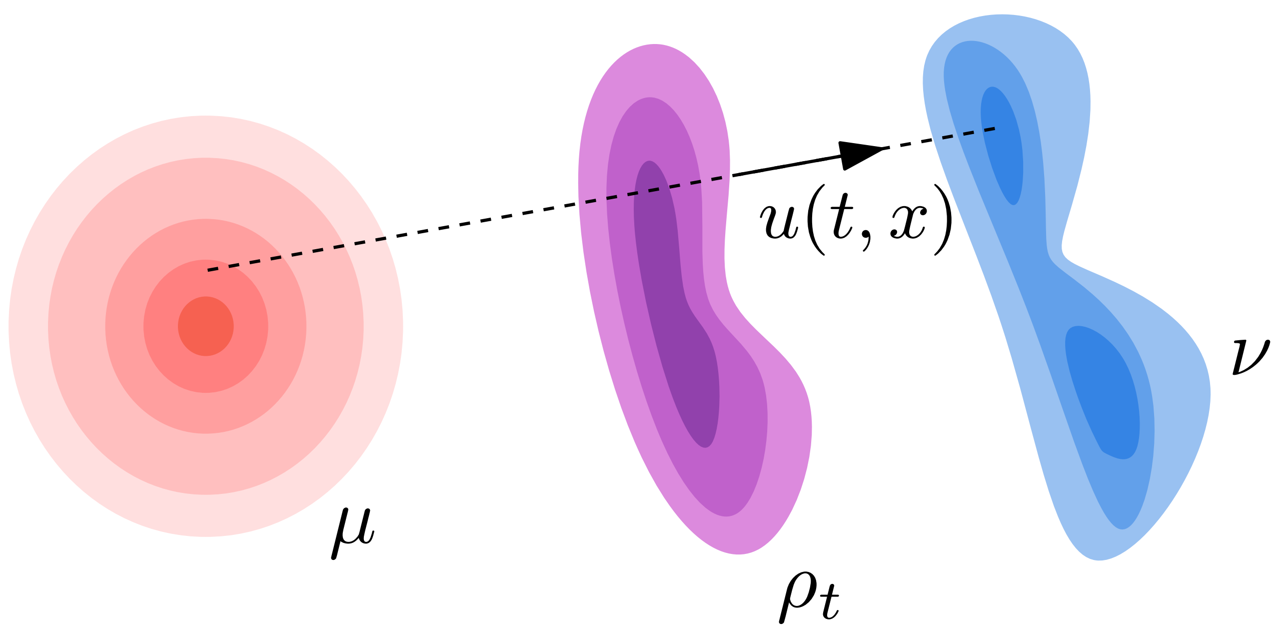


What about Optimal Transport ?

Definition: (Wasserstein distance)
The Wasserstein distance between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$\mathrm{W}_p^p(\mu, \nu) = \inf_{T_{\#\mu}=\nu} \int_{\mathbb{R}^d} \|x - T(x)\|_p^p \,\mathrm{d}\,\mu(x)$$

where $T_{\#\mu} = \mu \circ T^{-1}$ is the pushforward measure of μ by T.



Definition: (Benamou-Brenier or dynamic formulation of OT)

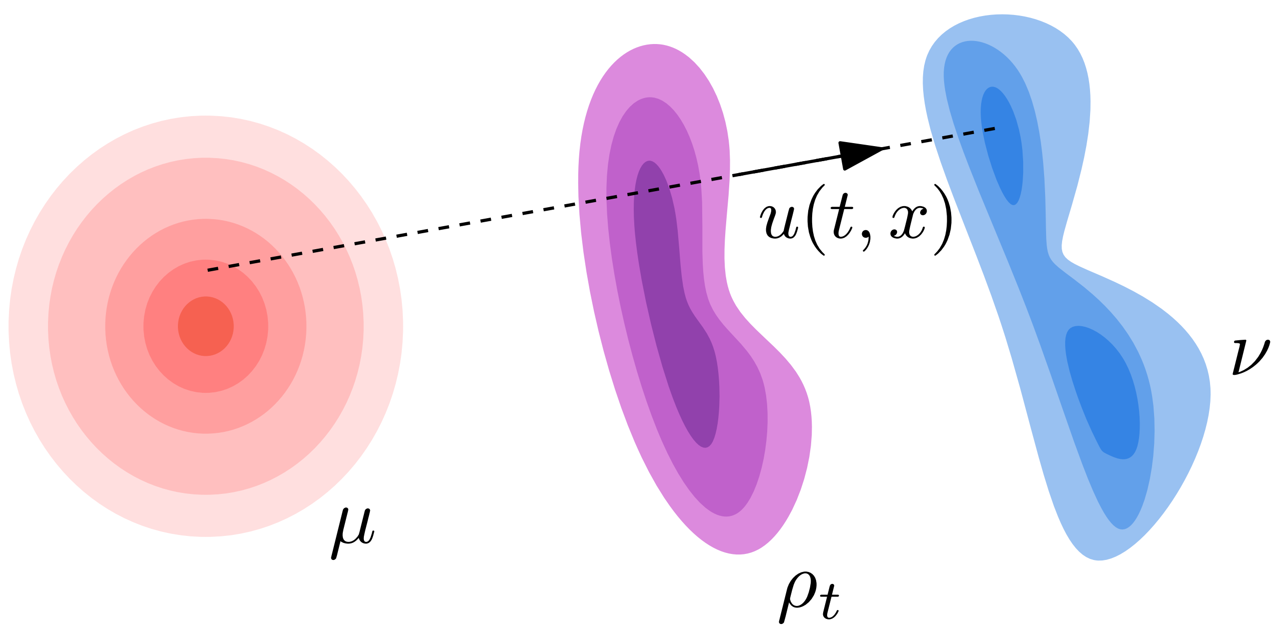
$$\mathrm{W}_2^2(\mu, \nu) = \inf_u \int_{\mathbb{R}^d} \int_0^1 \|u(t,x)\|^2 \,\mathrm{d}\,\rho_t(x) \qquad \begin{cases} \rho_0 = \mu, \ \rho_1 = \nu \\ \frac{\partial}{\partial t}\rho_t + \mathrm{div}(\rho_t u(t,\cdot)) = 0 \end{cases}$$

What about Optimal Transport ?

Definition: (Wasserstein distance)
The Wasserstein distance between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$\mathrm{W}_p^p(\mu, \nu) = \inf_{T_{\#\mu}=\nu} \int_{\mathbb{R}^d} \|x - T(x)\|_p^p \,\mathrm{d}\,\mu(x)$$

where $T_{\#\mu} = \mu \circ T^{-1}$ is the pushforward measure of μ by T.



Optimality condition:

Trajectories are straight

Definition: (Benamou-Brenier or dynamic formulation of OT)

$$\mathrm{W}_2^2(\mu, \nu) = \inf_u \int_{\mathbb{R}^d} \int_0^1 \|u(t,x)\|^2 \,\mathrm{d}\,\rho_t(x) \quad \begin{cases} \rho_0 = \mu, \ \rho_1 = \nu \\ \frac{\partial}{\partial t}\rho_t + \mathrm{div}(\rho_t u(t,\cdot)) = 0 \end{cases}$$

Optimality: minimize energy

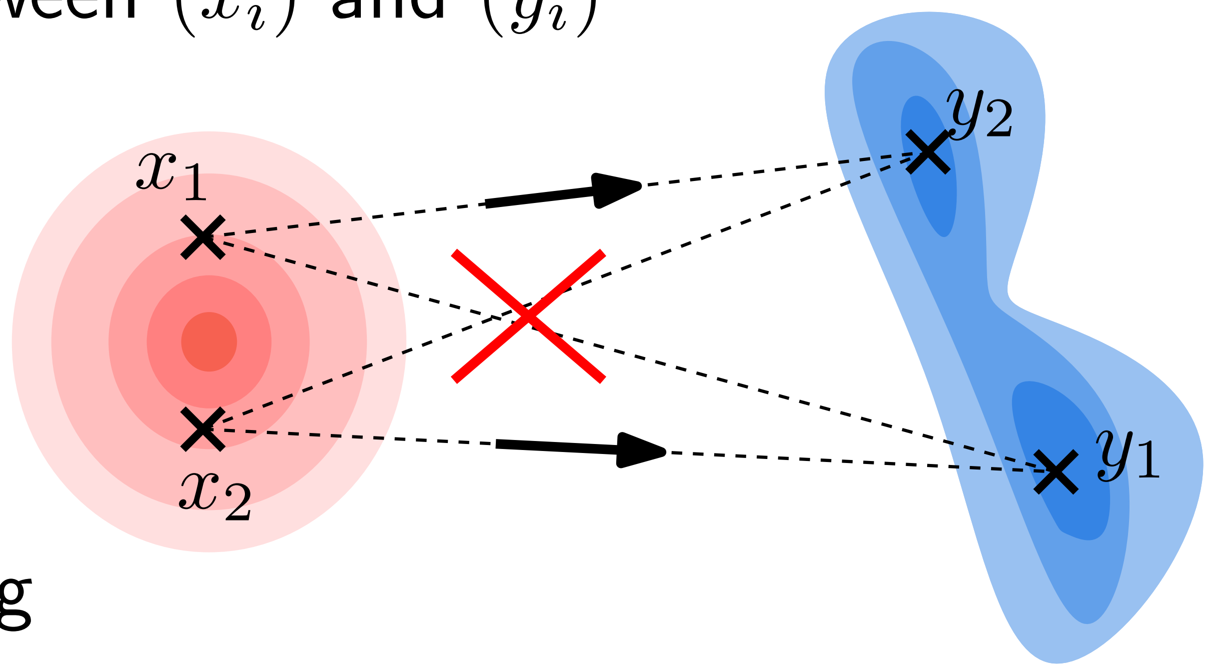
Flow Matching framework

Optimal transport flow matching

- Optimal transport flow is the ideal candidate for FM

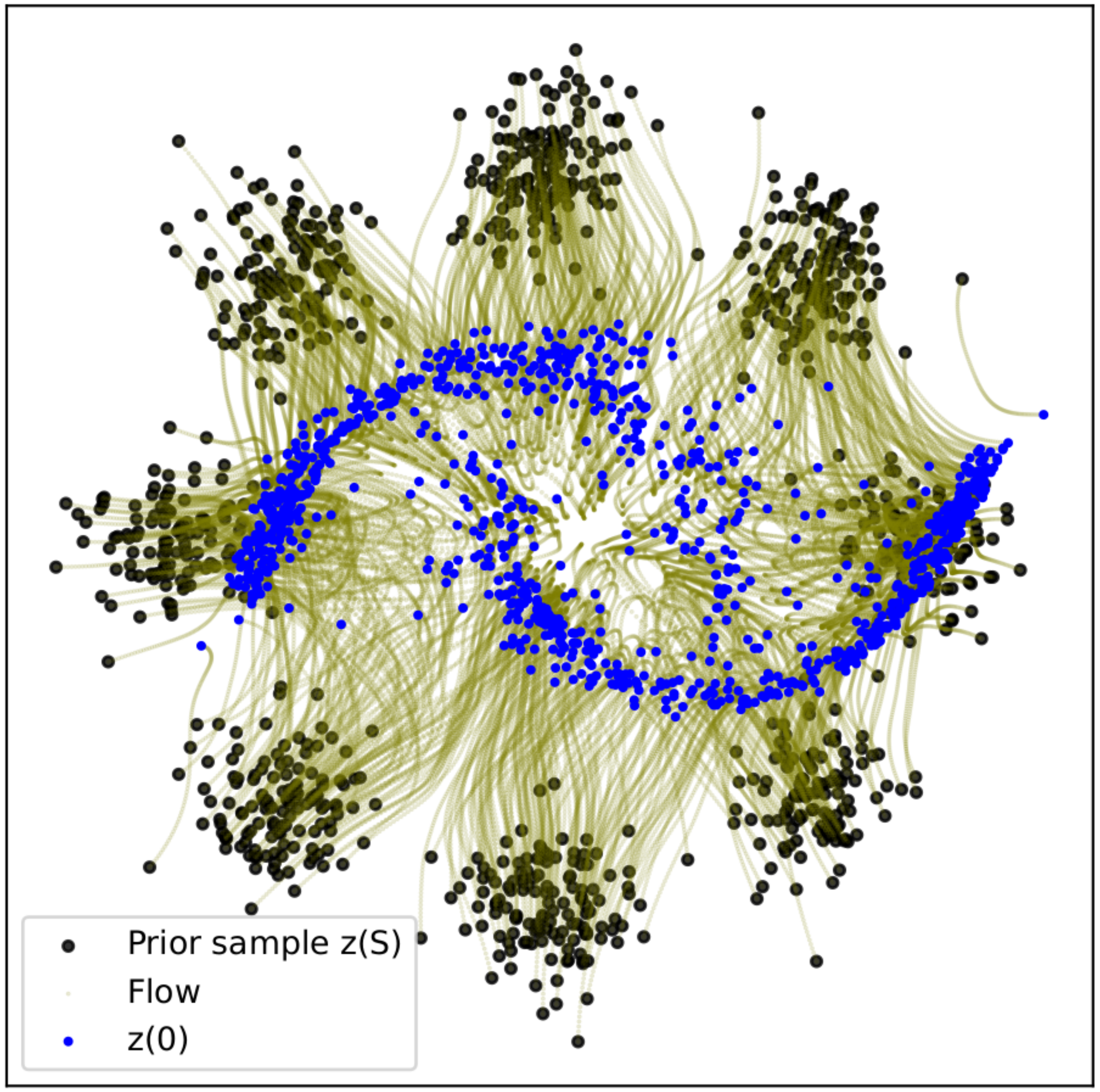
  ↳ Impossible in practice, bad high dimension scaling, $\mathrm{O}(n^3)$.

- One solution: minibatch OT *[Tong et. al 23, Pooladian et. al 23]*
  - Sample $(x_i)_1^n \sim p_0$ and $(y_i)_1^n \sim p_{\text{data}}$.
  - Compute optimal transport between (x_i) and (y_i)
  - Train v_θ to match $y_i - x_i$.

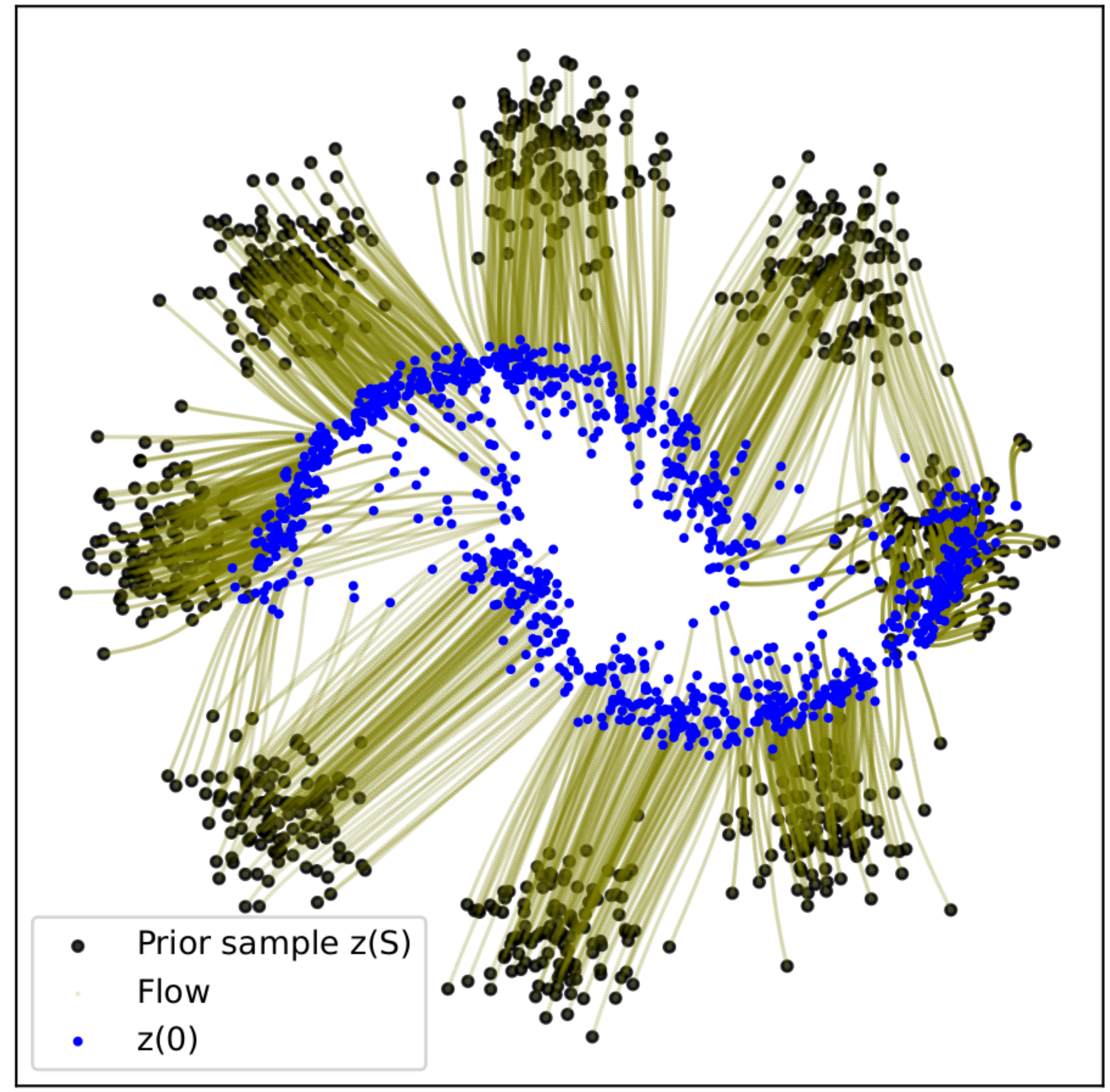


- Straighter trajectories
- Faster integration and thus sampling
- Limitation: Minibatch OT is not a great approximation in high dimension.

Comparison in 2d



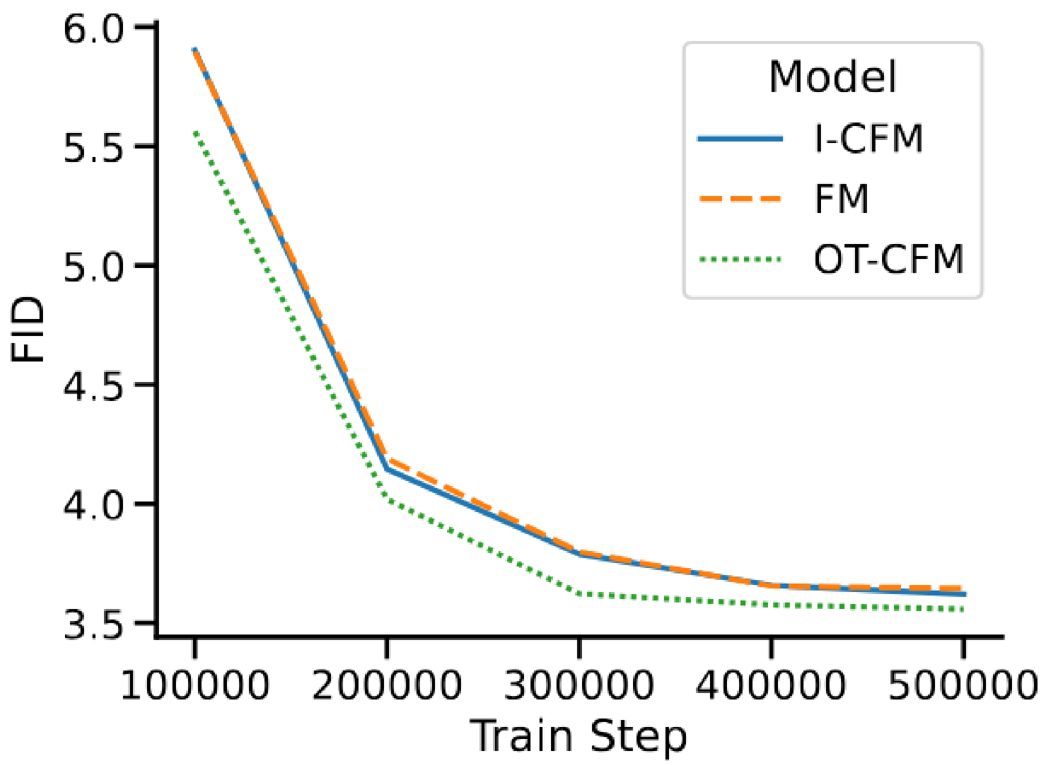
Independant sampling



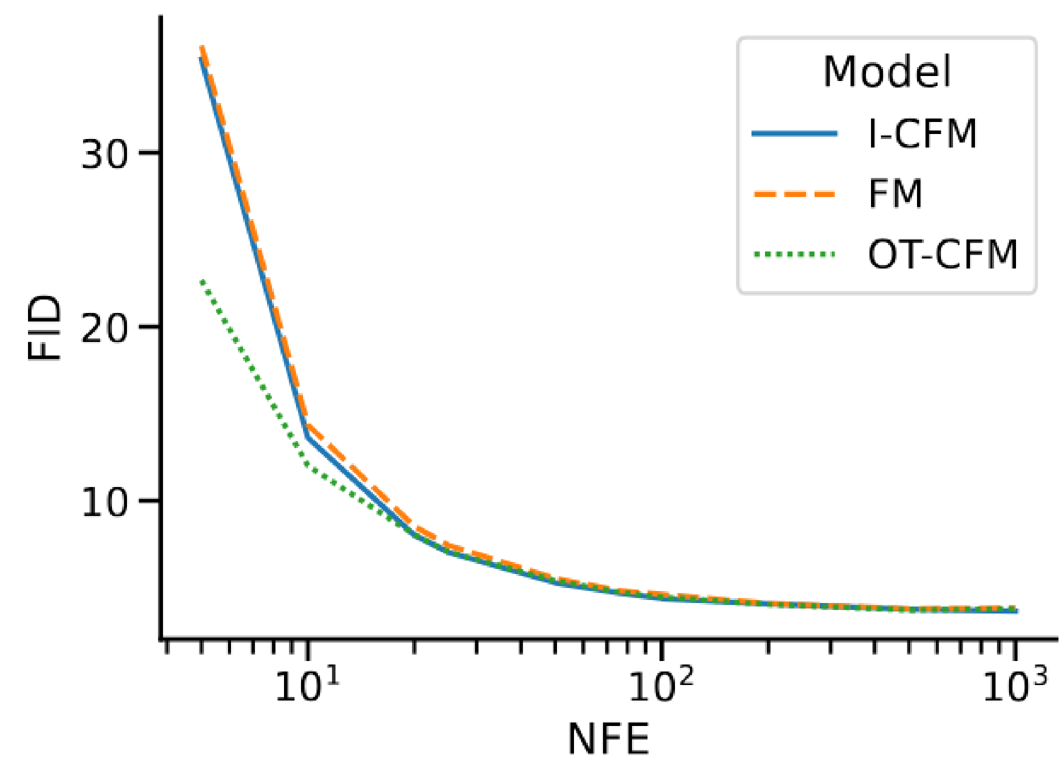
Minibatch OT sampling

Higher dimension: CIFAR10

- Experiments on CIFAR10 *[Tong et al. 23]*



FID along training with DoPri 5 solver

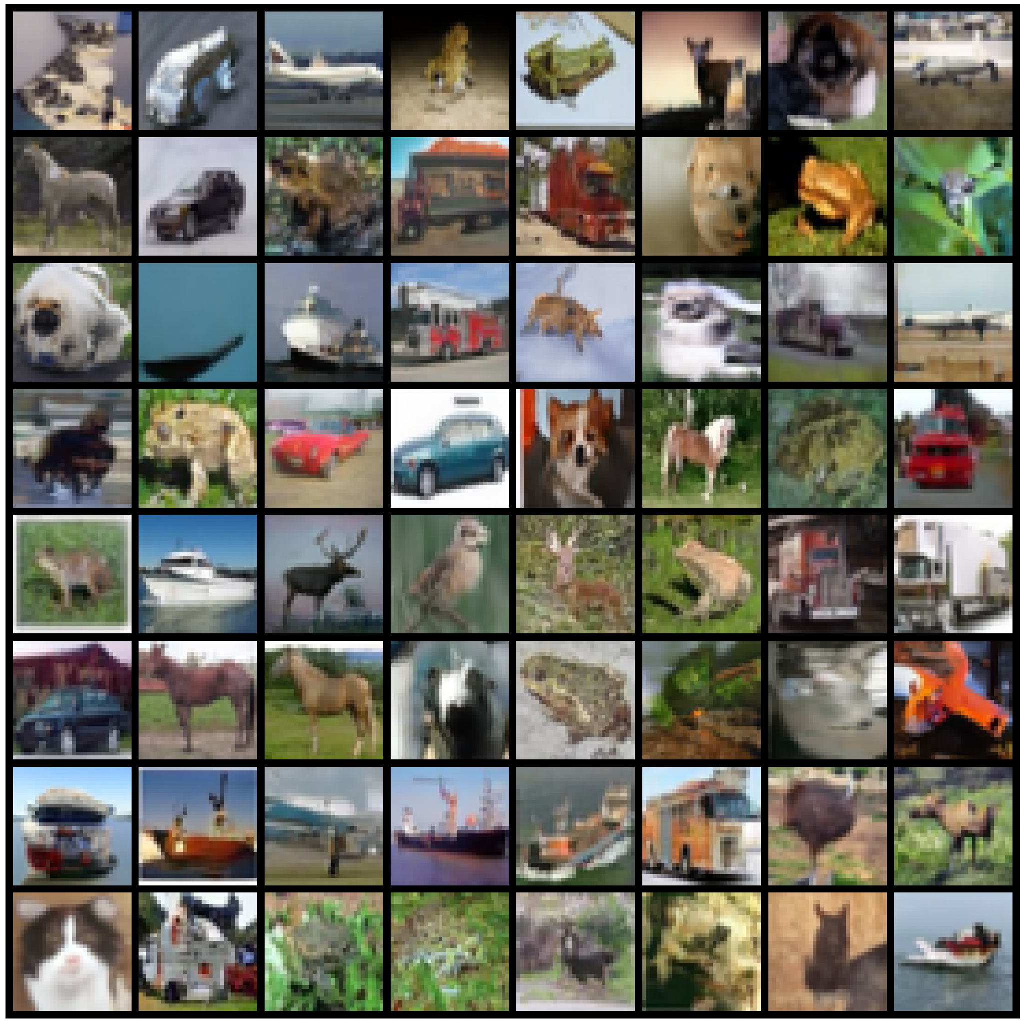


FID for euler solver with different NFE after 400k training steps

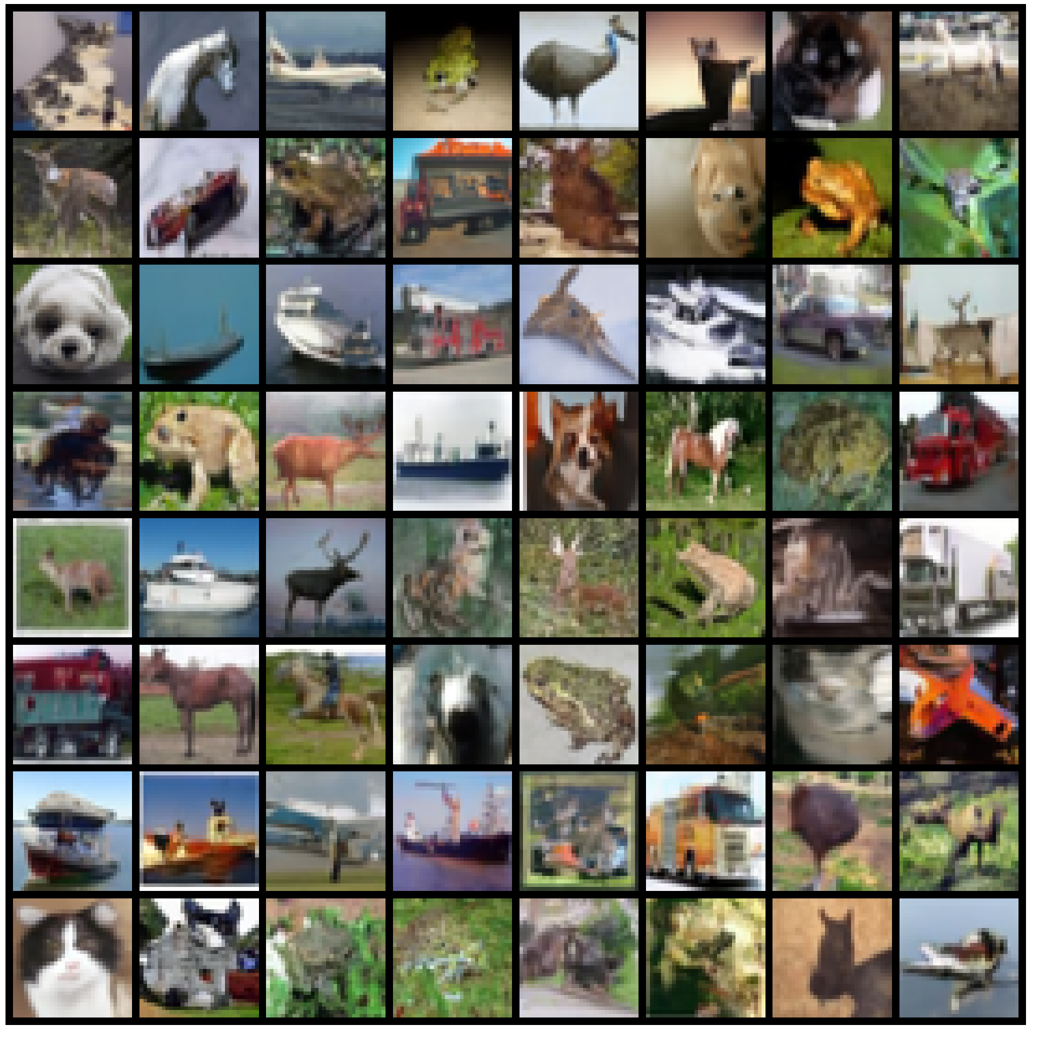
- FID is better for minibatch OT when few NFE
- Not very convincing in high dimension (small diff, no std dev)

Higher dimension: CIFAR10

Generated CIFAR10 samples:



ICFM



OTCFM

Conclusion and Remarks

- Conditional Flow Matching is a new simple and efficient framework for generative modeling.

- Minibatch optimal transport leads to straighter flows but is not really better in high dimension.

- An open question: Why does flow matching generalizes well the dataset ?

  [Gagneux et al. "The Generation Phases of Flow Matching: a Denoising Perspective." (2025).]