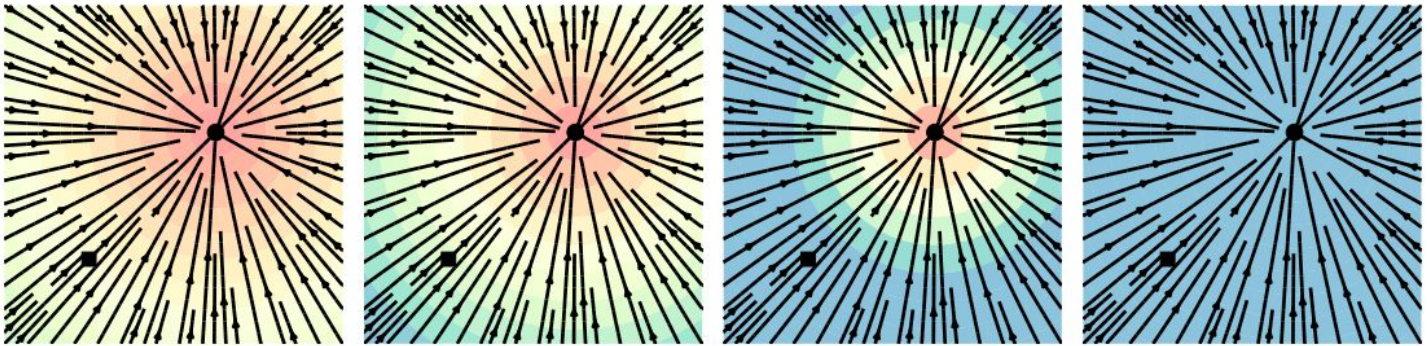
Flow Matching



$t = 0.0$ $t = 1/3$ $t = 2/3$ $t = 1.0$

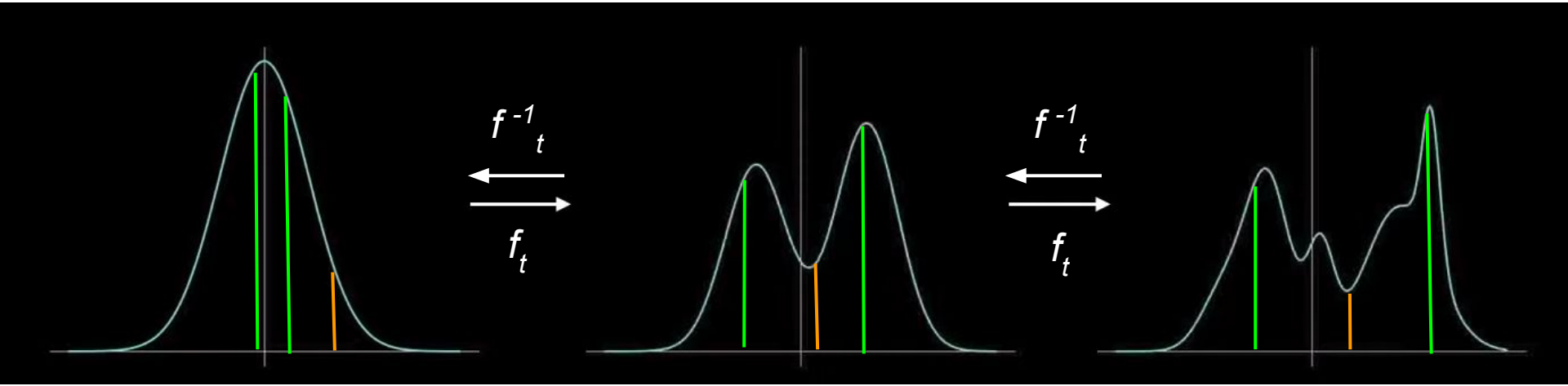
Feb 25

Preface: Normalizing Flows

- Let’s learn a map from a simple distribution to a complex distribution
  - Simple dist. $p_0(x)$ should be known and easy to calculate
  - “Complex” distribution would be your data distribution $p_1(x)$

- Bijective please!
  - Train by MLE by mapping complex -> simple
  - Sample by mapping simple -> complex

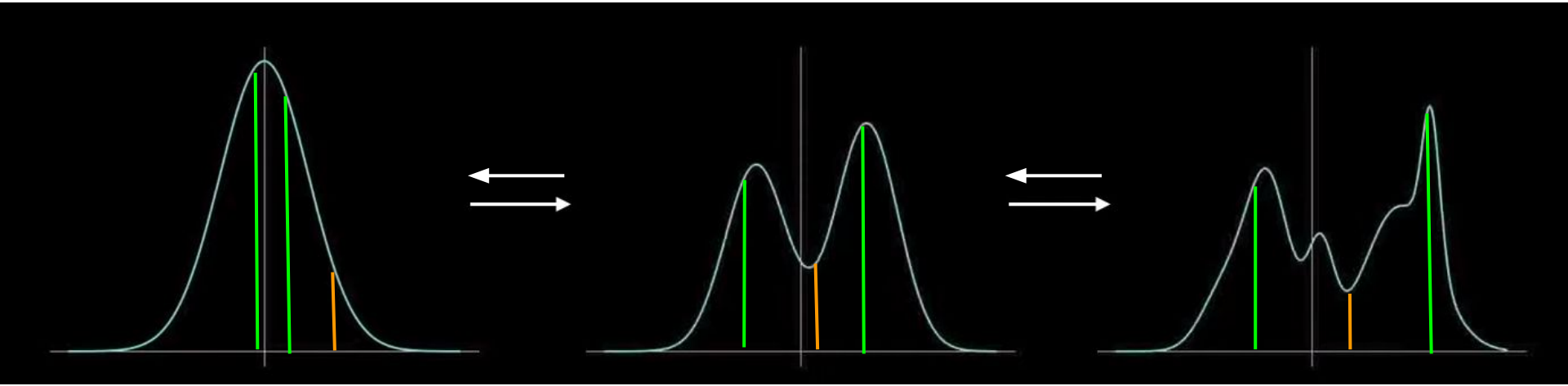
Normalizing Flows: Intuition



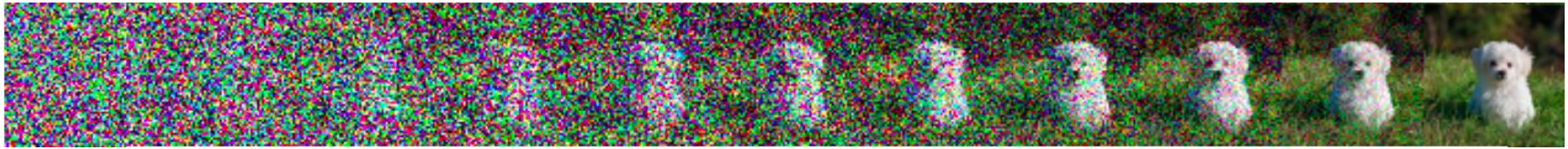
Simple dist. $p_0(x)$ Intermediate dist. $p_t(x)$ Complex dist. $p_1(x)$

1) We can sample from here
2) We can do MLE here (ie. $p_0(x; \theta)$)

Preface: Normalizing Flows

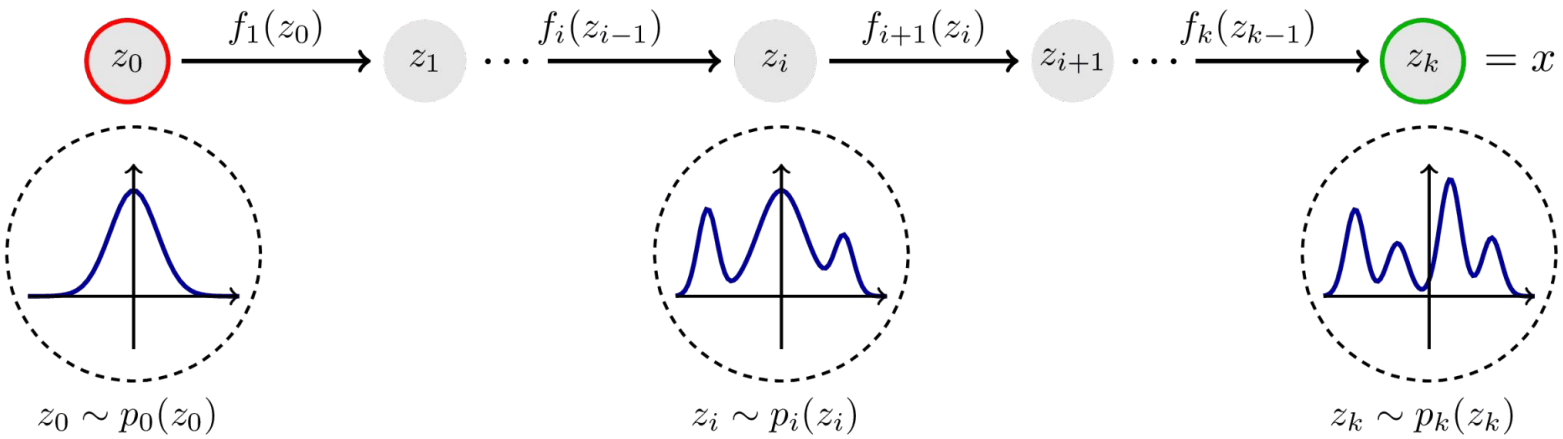


Simple dist. $p_0(x)$ Intermediate dist. $p_t(x)$ Complex dist. $p_1(x)$



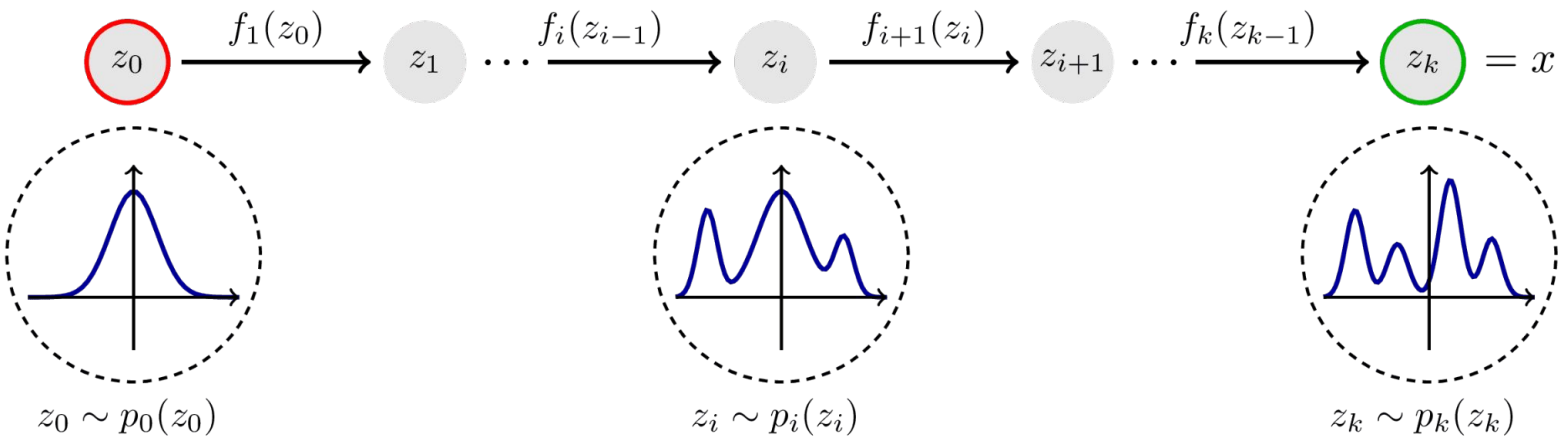
Normalizing Flows: Sampling Formulation

- Sampling: $f_\theta(p_0(x))$
  - Typically f_θ is composed of a chain of functions (sound familiar?)
  - We can just assume that $f = f_1 \circ \ldots \circ f_{t-1} \circ f_t$



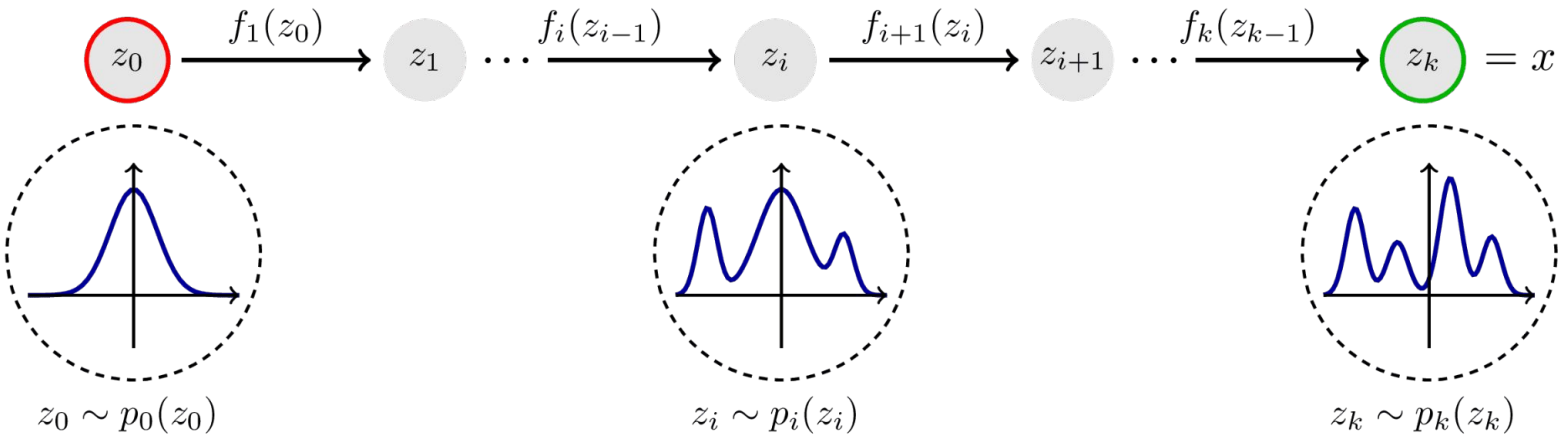
Normalizing Flows: Training

- Just map your data backwards and perform MLE !
- $p_1(x) = p_0(f^{-1}_{\theta}(x))$



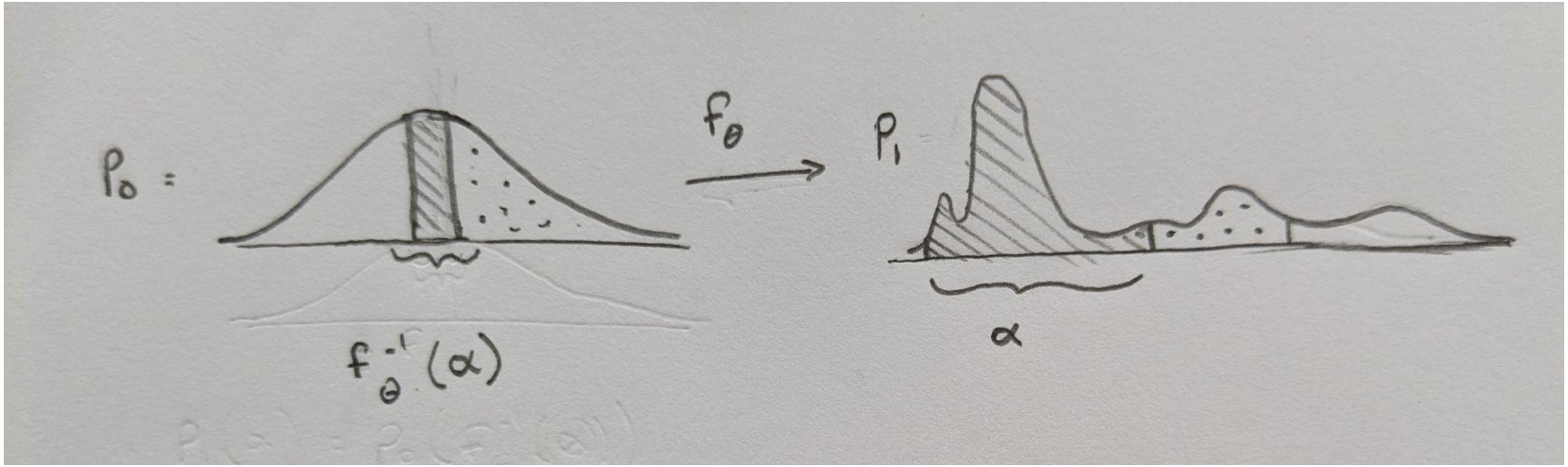
Normalizing Flows: Training

- Just map your data backwards and perform MLE
- $p_1(x) = p_0(f^{-1}_{\theta}(x))$!
  - Hold on… this isn't quite right (it's almost right though)

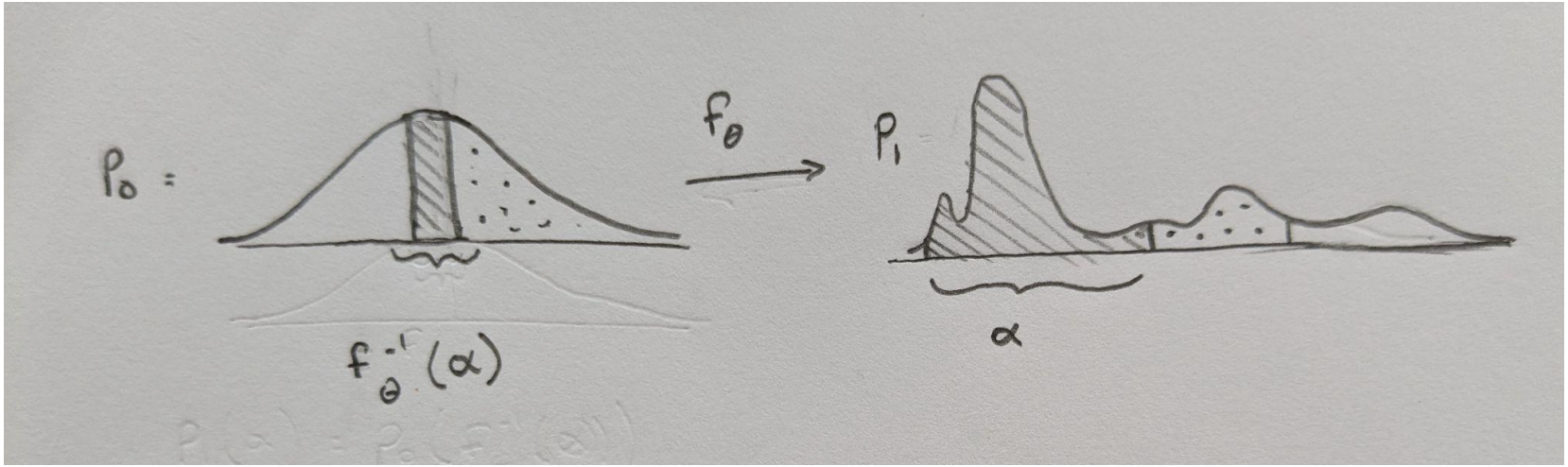


Why is this wrong: $p_1(x) = p_0(f^{-1}_{\theta}(x))$?

Why is this wrong: $p_1(x) = p_0(f^{-1}_{\theta}(x))$?



Why is this wrong: $p_1(x) = p_0(f^{-1}_{\theta}(x))$?



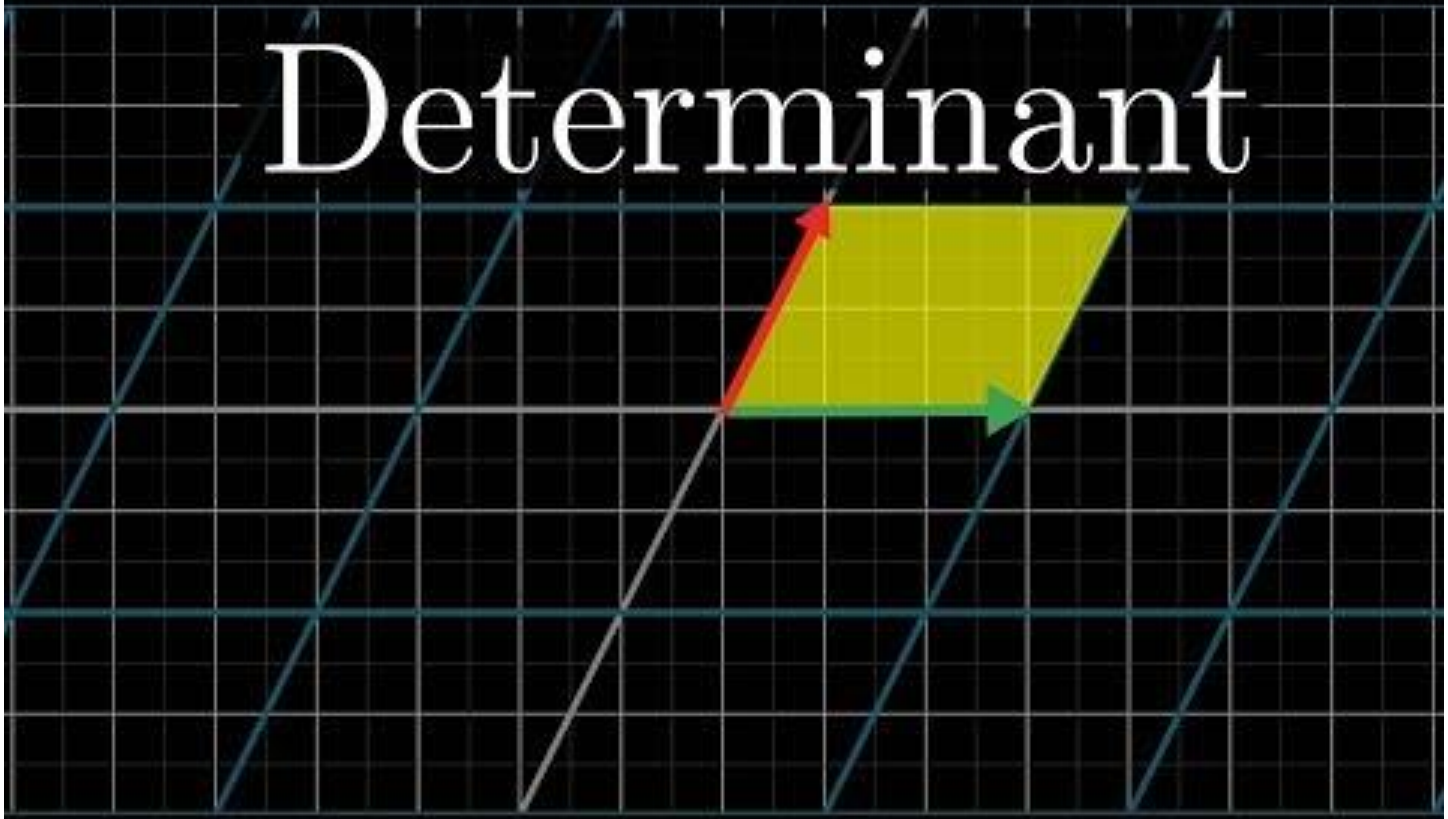
Answer: If we squish / stretch the data the probability comes out wrong

Correct Equation: (Respect the squish)

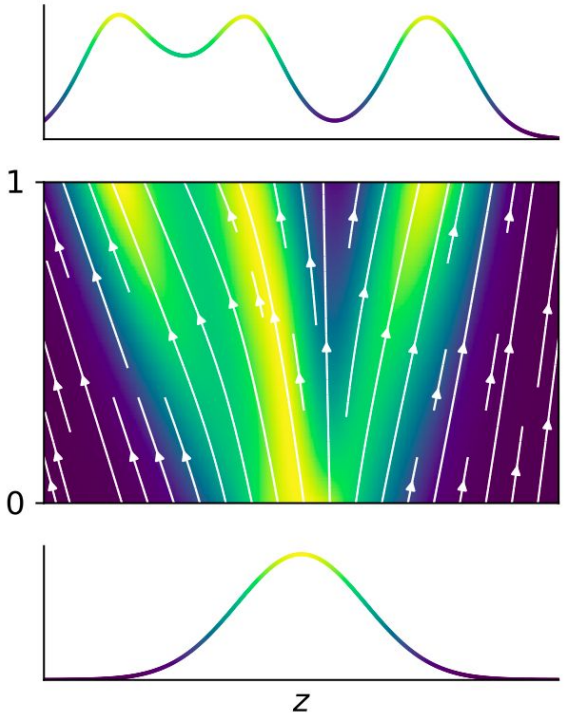
- *x* is *n* dimensional, so the derivative thing is a *jacobian*
- The jacobian is how the change in each x_i effects each $f_i(x)$ (square matrix)
  - The determinant measures volume

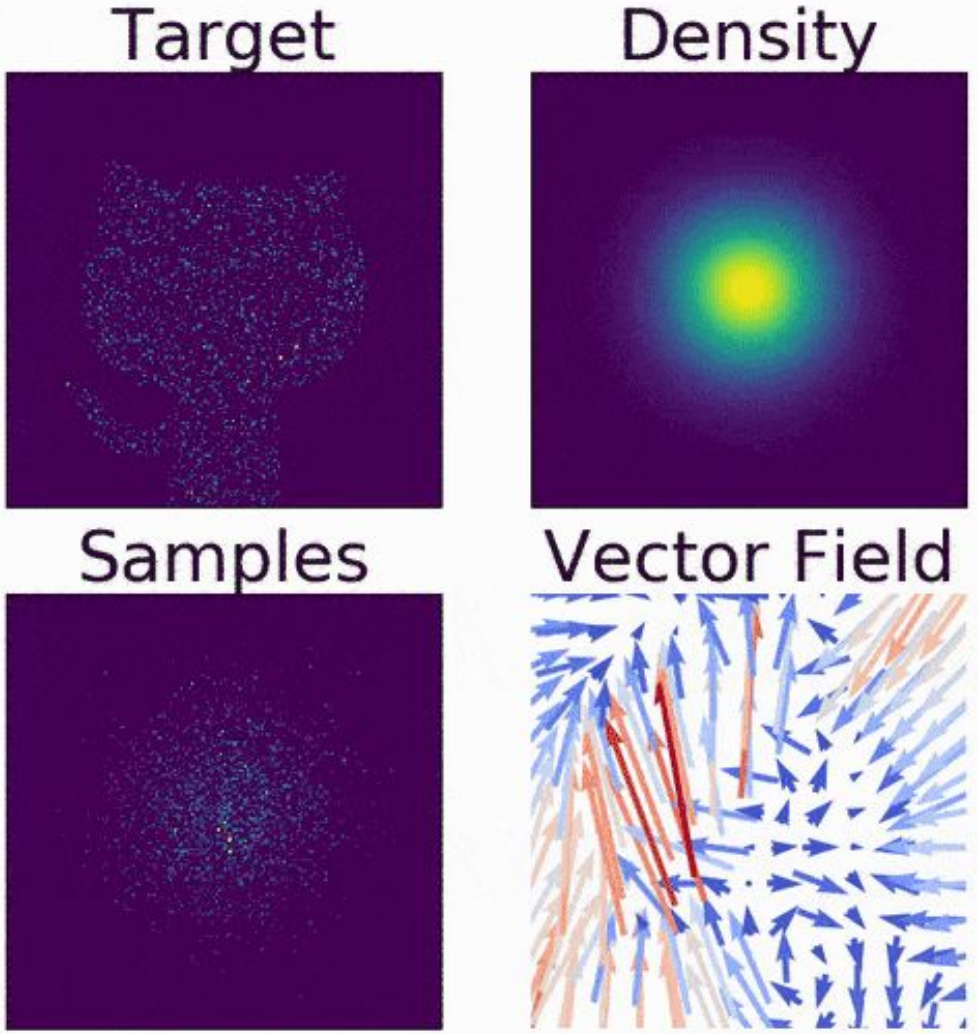
$$p_1(x) = p_0(f_\theta^{-1}) \left| \det \left(\frac{\partial f_\theta^{-1}}{\partial x} \right) \right|$$

Determinant



Continuous Normalizing Flows





Flow Matching: Basics

- Points in space to be pushed around
- Vector field that pushes the points

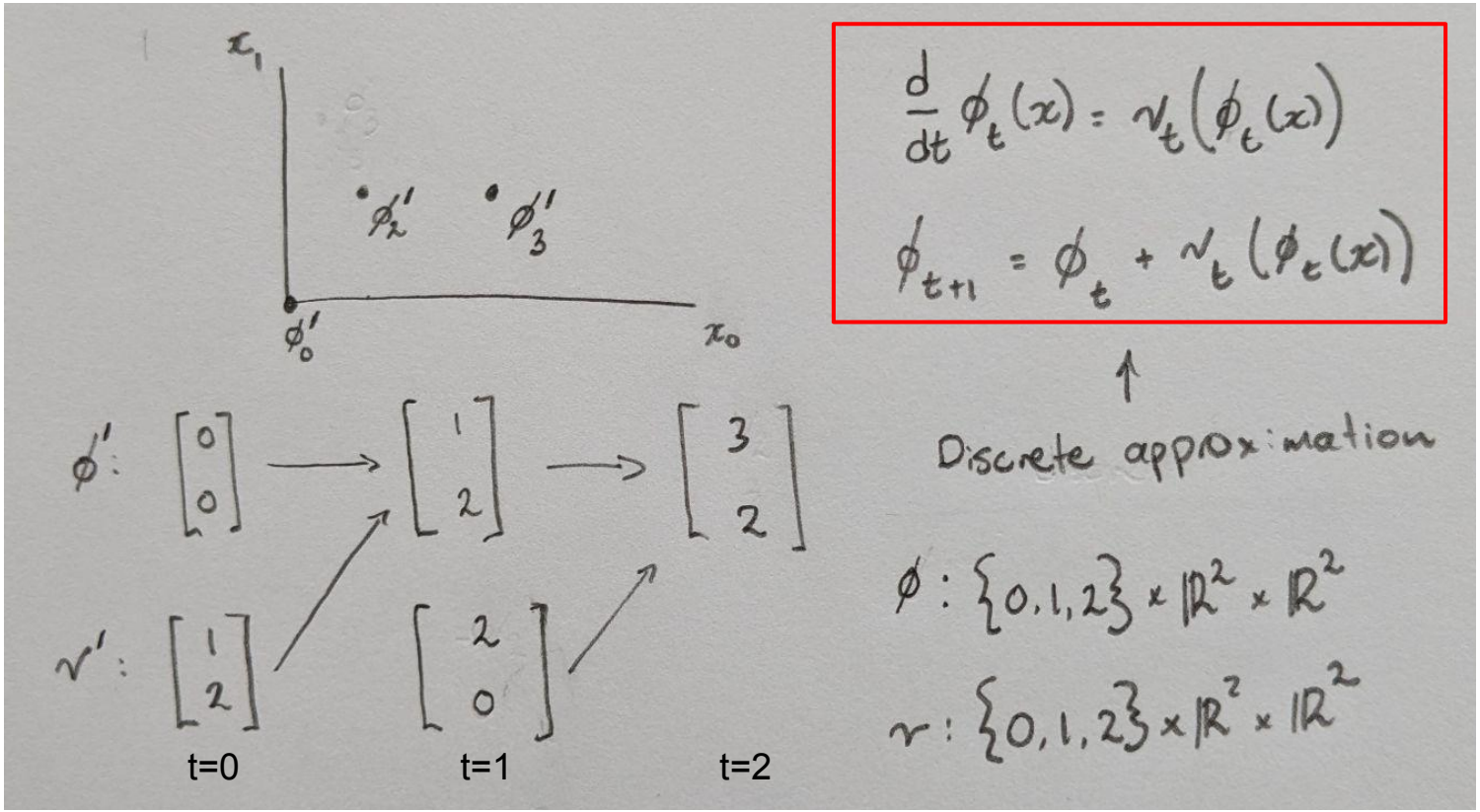
$$\phi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$$

$$v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$$

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x))$$

$$\phi_0(x) = x$$

Flow Matching: Basics



Flow Matching: Pop Quiz

- Lets use p_0 as a simple distribution (eg. Gaussian)
  - We then know the probability of $\phi_{t=0}$

- What is the probability p_t of points ϕ_t ??

Reminder:

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x))$$

$$\phi_0(x) = x$$

Flow Matching: Pop Quiz

- Lets use p_0 as a simple distribution (eg. Gaussian)
  - We then know the probability of $\phi_{t=0}$

- What is the probability p_t of points ϕ_t ??

Reminder:

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x))$$

$$\phi_0(x) = x$$

$$p_t = [\phi_t]_* p_0$$

$$[\phi_t]_* p_0(x) = p_0(\phi_t^{-1}(x)) \det \left[\frac{\partial \phi_t^{-1}}{\partial x}(x) \right]$$

Flow Matching: Goal

- Learn the vector field that pushes points around
  - Ok, but how??
  - We know nothing about p_t or u_t…

Reminder:

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x))$$
$$\phi_0(x) = x$$

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,p_t(x)} \left\| v_t(x) - u_t(x) \right\|^2$$

Break up the Problem

- q(x) is our unknown data distribution
- Let's look for a single sample x_1
  - $p_t(x \mid x_1)$ is the conditional probability path
  - Similar to forward process in diffusion

Break up the Problem

- q(x) is our unknown data distribution
- Let's look for a single sample x_1
  - $p_t(x \mid x_1)$ is the conditional probability path
  - Similar to forward process in diffusion

$$p_t(x) = \int p_t(x|x_1)q(x_1)dx_1$$

$$p_1(x) = \int p_1(x|x_1)q(x_1)dx_1 \approx q(x)$$

Break up the Problem

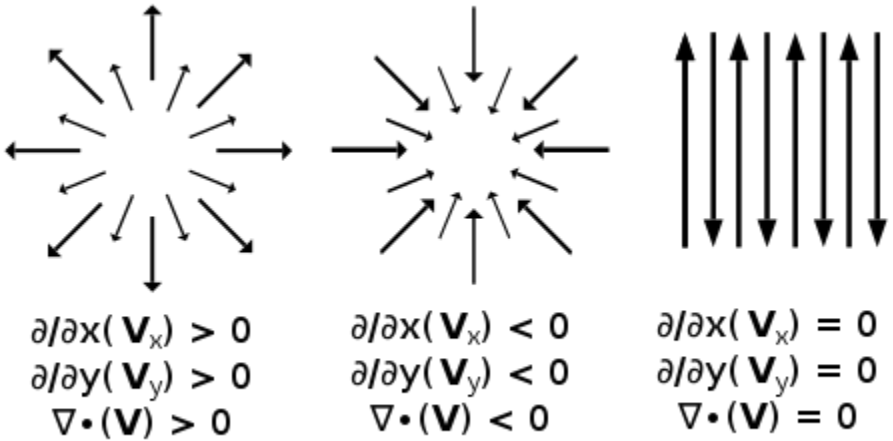
- Let’s follow this idea further
- We can define a marginal vector field (that pushes points)
  - Not obvious (to me) why this form is correct…

$$u_t(x) = \int u_t(x|x_1) \frac{p_t(x|x_1)q(x_1)}{p_t(x)} dx_1$$

Intuition

- These conditionals are connected to marginals via the **continuity equation**!
  - Change in density is divergence (think water flow)

$$\frac{\partial \rho}{\partial t} + \nabla \bullet \left(\rho \vec{u} \right) = 0$$



Divergence

More Powerful Intuition

- These conditionals are connected to marginals via the **continuity equation**!
  - Change in density is divergence (think water flow)

$$\frac{\partial \rho}{\partial t} + \nabla \bullet \left(\rho \vec{u} \right) = 0$$

$$\frac{d}{dt} p_t(x) = \int \left(\frac{d}{dt} p_t(x|x_1) \right) q(x_1) dx_1 = - \int \operatorname{div}\Big(u_t(x|x_1) p_t(x|x_1) \Big) q(x_1) dx_1$$

$$= -\operatorname{div}\Big(\int u_t(x|x_1) p_t(x|x_1) q(x_1) dx_1 \Big) = -\operatorname{div}\Big(u_t(x) p_t(x) \Big),$$

$$u_t(x) = \int u_t(x|x_1) \frac{p_t(x|x_1) q(x_1)}{p_t(x)} dx_1$$

Final Hard Step

- These have identical gradients:
  - Just like in diffusion, we use the path of a single data point
  - Now we just need to define $p_t(x|x_1)$ and $u(x|x_1)$

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,p_t(x)} \left\| v_t(x) - u_t(x) \right\|^2$$

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)} \left\| v_t(x) - u_t(x|x_1) \right\|^2$$

Choosing Conditional Probability Flow

- We can choose u_t that generates p_t
- Distributions are gaussians that are pushed through space

$$\frac{d}{dt}\psi_t(x) = u_t(\psi_t(x)|x_1)$$

$$[\psi_t]_* p(x) = p_t(x|x_1)$$

$$p_t(x|x_1) = \mathcal{N}(x \mid \mu_t(x_1), \sigma_t(x_1)^2 I)$$

- You can get diffusion paths from this formulation

Choosing Conditional Probability Flow

- We can choose any path
- Distributions are gaussians that are pushed through space

$$\frac{d}{dt}\psi_t(x) = u_t(\psi_t(x)|x_1)$$

$$[\psi_t]_* p(x) = p_t(x|x_1)$$

$$p_t(x|x_1) = \mathcal{N}(x \mid \mu_t(x_1), \sigma_t(x_1)^2 I)$$

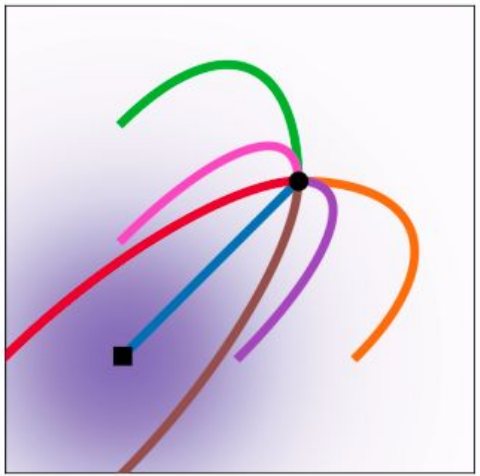
- You can get diffusion paths from this formulation

$$\psi_t(x) = \sigma_t(x_1)x + \mu_t(x_1)$$

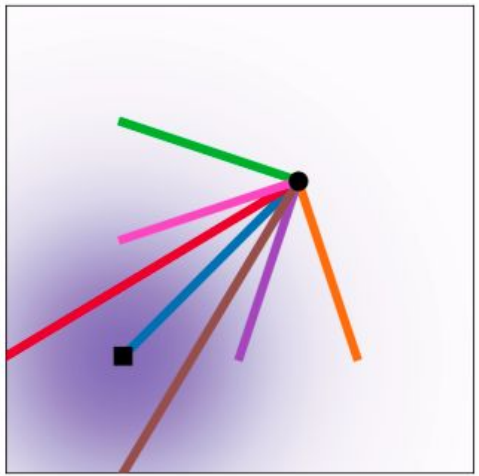
$$u_t(x|x_1) = \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}(x - \mu_t(x_1)) + \mu_t'(x_1)$$

A Special Case

- Linear Flow $\mu_t(x) = tx_1, \text{ and } \sigma_t(x) = 1 - (1 - \sigma_{\min})t.$
  - This is actually the optimal transport displacement map!

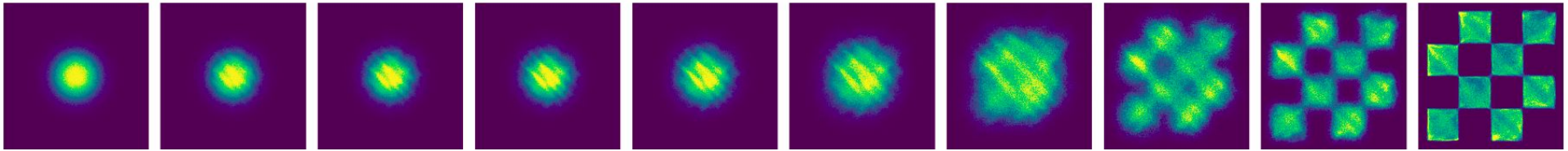


Diffusion

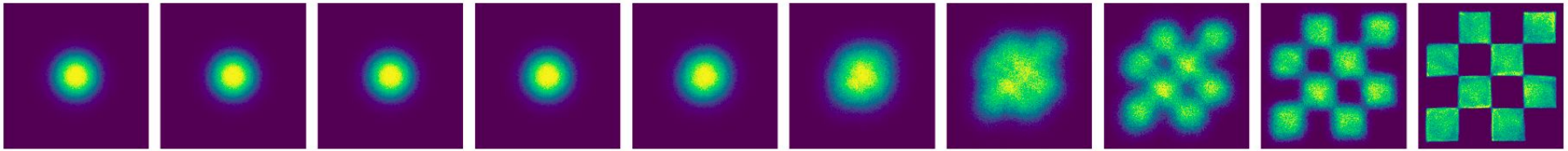


OT

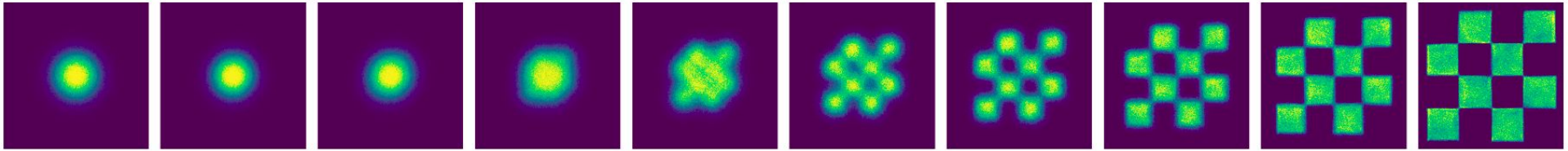
Comparison on Toy Data



Score matching w/ Diffusion



Flow Matching w/ Diffusion



Flow Matching w/ OT

Comparison on Harder Benchmarks

Model	CIFAR-10			ImageNet 32×32			ImageNet 64×64		
	NLL↓	FID↓	NFE↓	NLL↓	FID↓	NFE↓	NLL↓	FID↓	NFE↓
Ablations									
DDPM	3.12	7.48	274	3.54	6.99	262	3.32	17.36	264
Score Matching	3.16	19.94	242	3.56	5.68	178	3.40	19.74	441
ScoreFlow	3.09	20.78	428	3.55	14.14	195	3.36	24.95	601
Ours									
FM w/ Diffusion	3.10	8.06	183	3.54	6.37	193	3.33	16.88	187
FM w/ OT	**2.99**	**6.35**	**142**	**3.53**	**5.02**	**122**	**3.31**	**14.45**	**138**

Comparison on Harder Benchmarks

	ImageNet 128×128	
Model	NLL↓	FID↓
MGAN (Hoang et al., 2018)	–	58.9
PacGAN2 (Lin et al., 2018)	–	57.5
Logo-GAN-AE (Sage et al., 2018)	–	50.9
Self-cond. GAN (Lučić et al., 2019)	–	41.7
Uncond. BigGAN (Lučić et al., 2019)	–	25.3
PGMGAN (Armandpour et al., 2021)	–	21.7
FM ^{w}/ OT	**2.90**	**20.9**

Break up the Problem: Intuition Attempt

- Vector field is integral of weighted conditional vector field (for all data)
- Weighted by:
  - prob. that a point is pushed to *x* by starting at some x_1 divided by prob that it gets there any other possible way
  - Ie. more “plausible” conditional vector fields get more weight.

$$u_t(x) = \int u_t(x|x_1) \frac{p_t(x|x_1)q(x_1)}{p_t(x)} dx_1$$