Flow Matching for Generative Modeling

Revant Mahajan, Liu Dai, Yucheng Mao, Alexiy Buynitsky

Key Words

Optimal Transport

Generative Modeling

Flow Matching

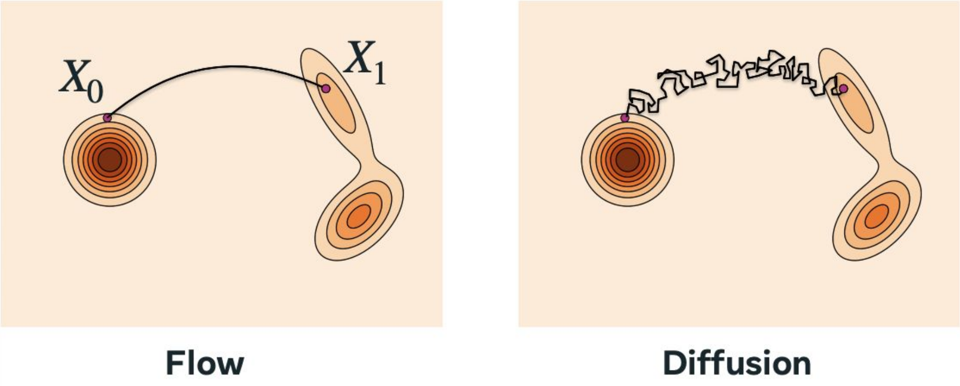
CNF

Introduction

The Goal of Generative Modelling

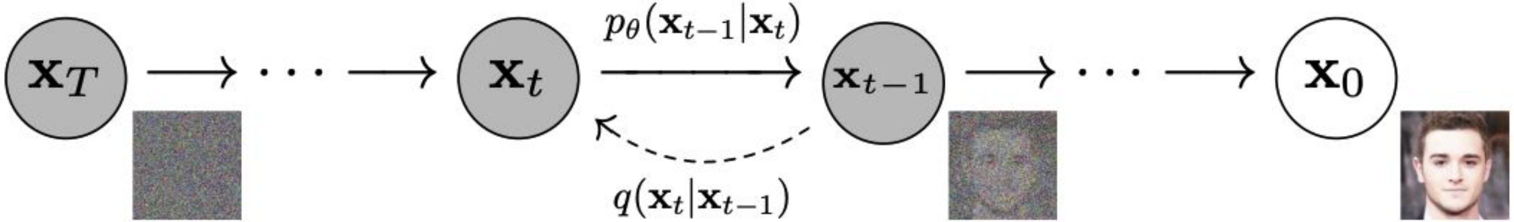
Modeled as a Continuous-time Markov Process:

$$X_{t+h} = \Phi_{t+h|t}(X_t)$$



https://www.calibraint.com/blog/beginners-guide-to-diffusion-models
https://neurips.cc/virtual/2024/tutorial/99531

Generative Modelling with Diffusion



Forward Process:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

Backward Process:

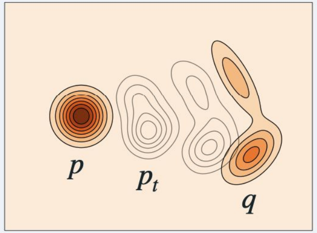
$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

https://arxiv.org/pdf/2006.11239

Preliminaries

Marginal probability Path:

$$X_t \sim p_t$$

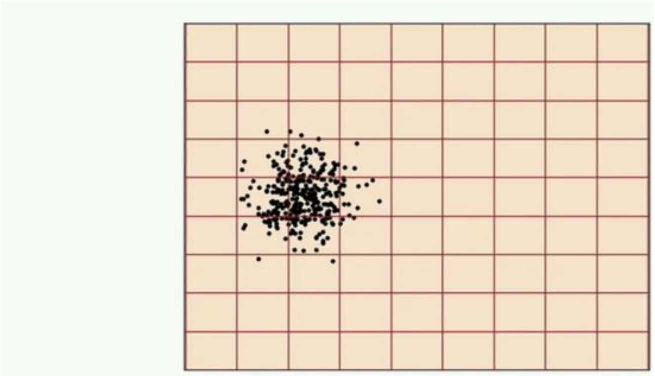


Key Property:

$$\int p_t(x)dx = 1$$

Warping Function / Flow:

$$X_t = \phi_t(X_0)$$



Key Properties:

- Bijective function (its invertible)
- Markovian assumption:

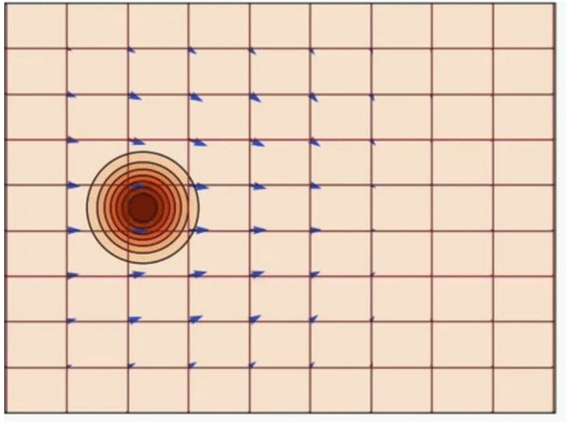
$$X_{t+h} = \phi_{t+h|t}(X_t)$$

https://neurips.cc/virtual/2024/tutorial/99531

Flow as ODE

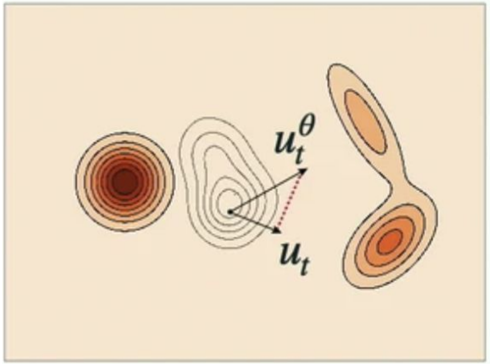
Parametrize Flow through velocity as ODE:

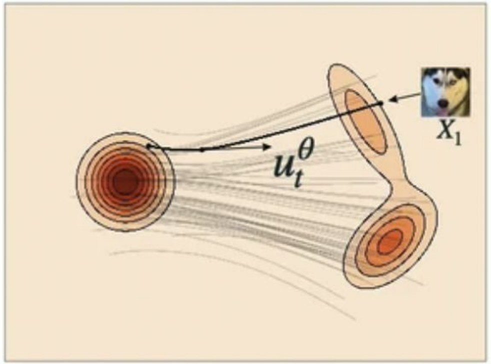
$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x))$$



Flow matching:

1. Train velocity u_t^θ that generates p_t with $p_0 = p$ and $p_1 = q$
2. Sample $X_0 \sim p$ from source distribution, then solve ODE





https://neurips.cc/virtual/2024/tutorial/99531

Flow Matching Objective:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,p_t(x)} \mid\mid u_t^\theta(x) - u_t(x) \mid\mid^2$$

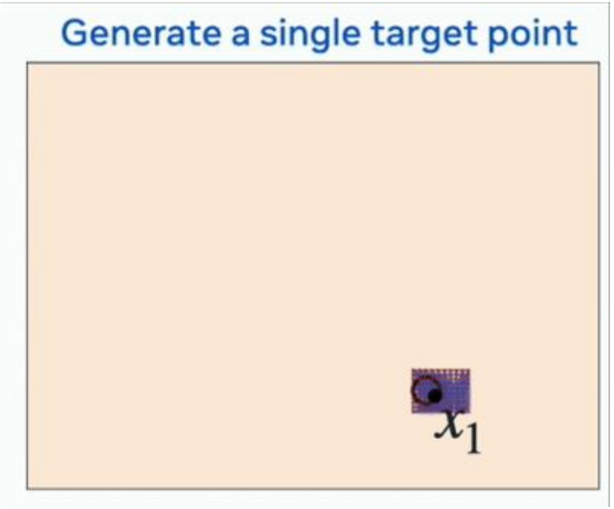
Have no prior knowledge of what p_t and $u_t(x)$

1. p_t : many choices of probability paths that satisfy $p_1 \approx q$

2. u_t : don’t have access to a closed form that generates p_t

How to construct p_t and u_t ?

Solution: Build from the simpler parts

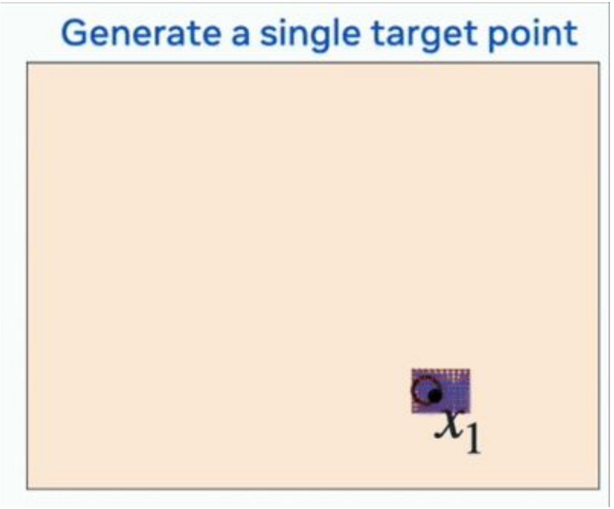


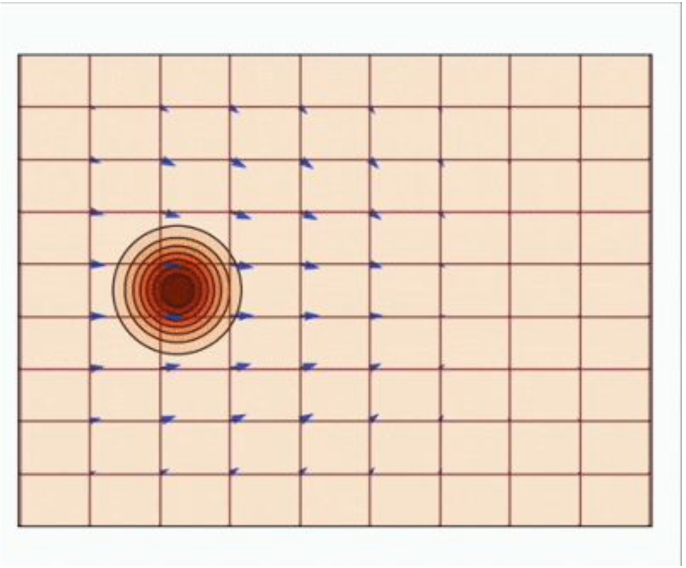
$$p_t(x \mid x_1)$$

$$u_t(x \mid x_1)$$

https://neurips.cc/virtual/2024/tutorial/99531

The marginal vector field generate the marginal probability path





$p_t(x \mid x_1)$ $\longrightarrow$ $p_t(x) = \mathbb{E}_{X_1} p_{t|1}(x \mid X_1)$

Average

$u_t(x \mid x_1)$ $\longrightarrow$ $u_t(x) = \mathbb{E}[\, u_t(X_t \mid X_1) \mid X_t = x \,]$

https://neurips.cc/virtual/2024/tutorial/99531

Conditional Flow Matching can give you Flow Matching

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

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

$$\nabla_\theta \mathcal{L}_{\text{FM}}(\theta) = \nabla_\theta \mathcal{L}_{\text{CFM}}(\theta)$$

https://neurips.cc/virtual/2024/tutorial/99531

The Choice of Conditional Probability Path in Flow Matching

Why

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

1. The **ground-truth** vector field $u_t(x|x_1)$ still remains unknown here.
2. What is a **good** condition path $p_t(x|x_1)$

Conditional Probability Path

In paper, we use a general Gaussian Distribution to sample X from it:

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

When **t=0**, $\mu_0(x_1) = 0$ and $\sigma_0(x_1) = 1$ so that at timestamp 0, the sample distribution is standard Gaussian Distribution.

When **t=1**, $\mu_1(x_1) = x_1$ and $\sigma_1(x_1) = \sigma_{\text{min}}$ where the standard deviation is small enough that the sample distribution is a **concentrated Gaussian distribution** centered at x1.

Conditional Probability Path

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

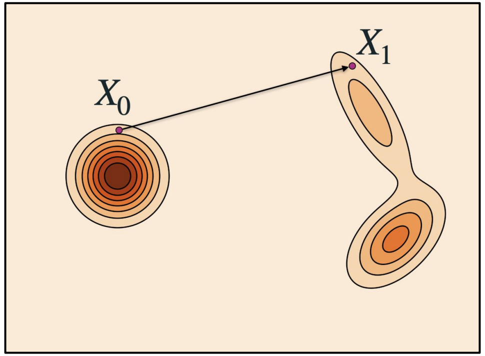
Reparameterizing (Mentioned in VAE lecture)

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

Vector Field

Data

Gaussian Noise



https://neurips.cc/virtual/2024/tutorial/99531

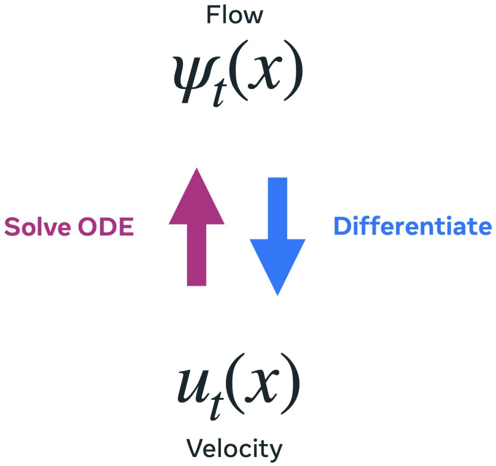
Ground Truth Vector Field

We already know: Ground-Truth Vector Path:

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

Differentiate:

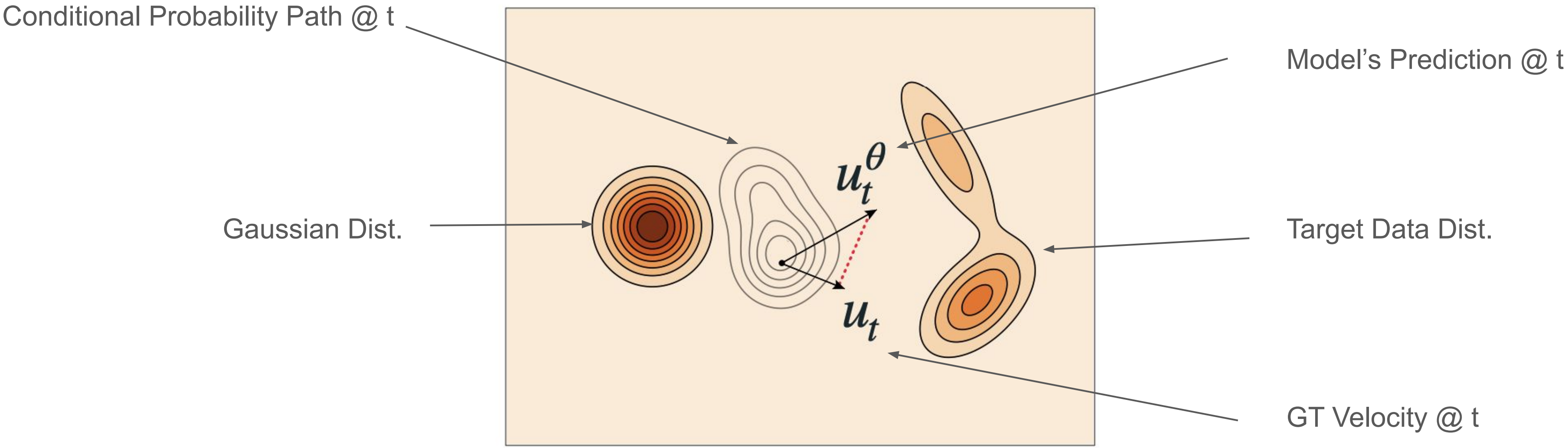
$$\frac{d}{dt}\psi_t(x) = u_t(\psi_t(x)|x_1)$$



Ground Truth Vector Field

Training Objective:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,q(x_1),p(x_0)} \left\| v_t(\psi_t(x_0)) - \frac{d}{dt}\psi_t(x_0) \right\|^2$$



https://neurips.cc/virtual/2024/tutorial/99531

Path1: Diffusion Path (Variance Exploding & Variance Preserving)

Both VE and VP can be called as diffusion process:
[Score-Based Generative Modeling through Stochastic Differential Equations]

Variance Exploding Path (Known as Score Matching):

$$p_t(x) = \mathcal{N}(x | x_1, \sigma_{1-t}^2 I)$$

Keep the original data distribution, scale up the variance of each Conditional Probability Path

Variance Preserving Path (Known as DDPM):

$$p_t(x|x_1) = \mathcal{N}(x \,|\, \alpha_{1-t} x_1, \left(1 - \alpha_{1-t}^2\right) I)$$

Here: $\alpha_t = e^{-\frac{1}{2}T(t)}, T(t) = \int_0^t \beta(s) ds$, beta is a schedule function

“Squeeze” the original data distribution, keep the variance of the Conditional Probability Path to be the same

Path 2: Optimal Transport

For Optimal Transport, we set: $\mu_t(x) = tx_1$ and $\sigma_t(x) = 1 - (1 - \sigma_{\min})t$. Here, $t \in [0, 1]$ the path can be rewritten as:

$$\psi_t(x) = (1 - (1 - \sigma_{\min})t)x + tx_1$$

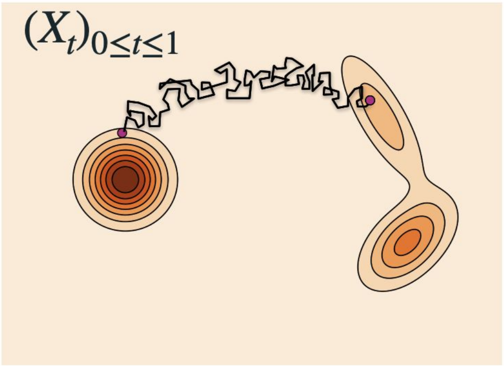
Main Difference:
VE, VP path need to **add Gaussian Noise** in each timestep
OT path do **interpolation** between Gaussian and Ground-Truth data.

A even more easy path (Rectified Flow):

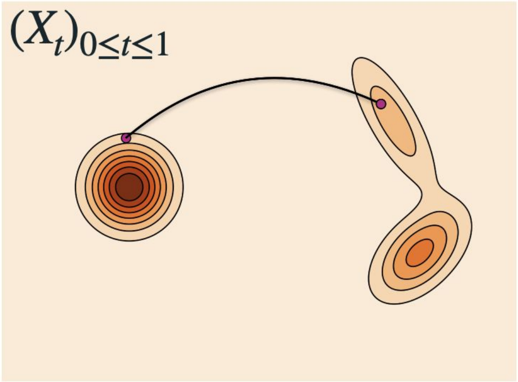
$$\psi_t(x \mid x_1) = tx_1 + (1 - t)x$$

Diffusion Path vs. OT Path

- Simpler Training Objective
- Faster Sampling Speed
- Stabilize Training



Diffusion



Flow

Experiment Results

Generation Quality

	CIFAR-10			ImageNet 32×32			ImageNet 64×64		
Model	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**

	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**

NLL: likelihood; FID: image quality; NFE: evaluation time

Datasets:

- CIFAR-10
- ImageNet at resolution 32/64/128

Baselines:

- DDPM
- Score Matching / Score Flow
- Flow Matching w/ Diffusion Sampling

Sampling Efficiency

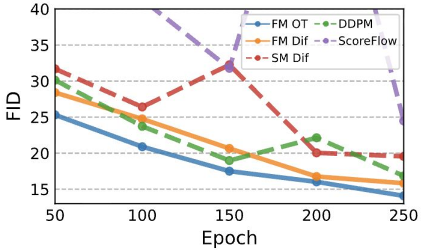


Figure 5: Image quality during training, ImageNet 64×64.

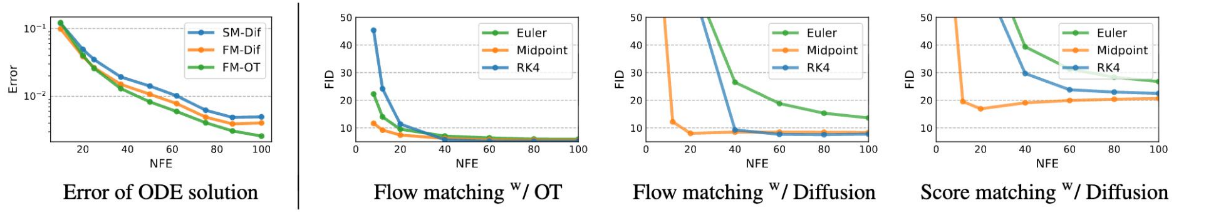


Figure 7: Flow Matching, especially when using OT paths, allows us to use fewer evaluations for sampling while retaining similar numerical error (left) and sample quality (right). Results are shown for models trained on ImageNet 32×32, and numerical errors are for the midpoint scheme.

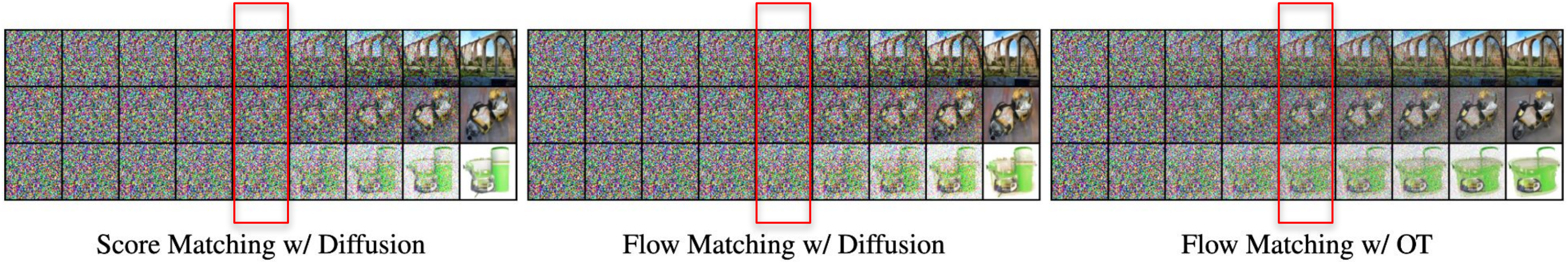


Figure 6: Sample paths from the same initial noise with models trained on ImageNet 64×64. The OT path reduces noise roughly linearly, while diffusion paths visibly remove noise only towards the end of the path. Note also the differences between the generated images.

Conclusion

Strength & Weakness

Strength:

- Concise and more generalizable framework
- Simulation-free method to train CNFs
- Faster sampling during inference

Weakness:

- Finding the “best” path is still left to be an open problem
- Flow matching is not as expressive as diffusion process (diffusion formulation offers a larger design space)
- Complexity of Marginal Vector Field: Condition Vector Field is the best and the easiest "simulation" of Marginal Vector Field

Improvement & Application

Follow-ups:

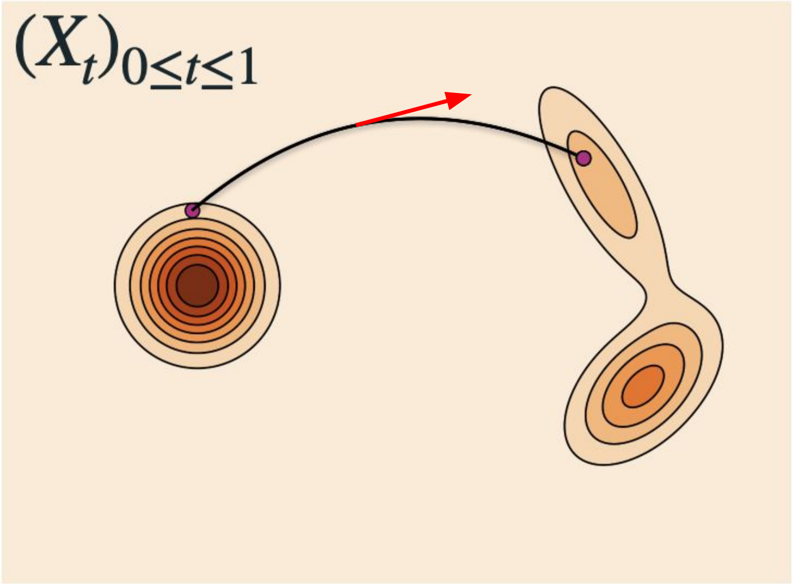
- Rectified Flow (concurrent work)
- Minibatch Optimal Transport

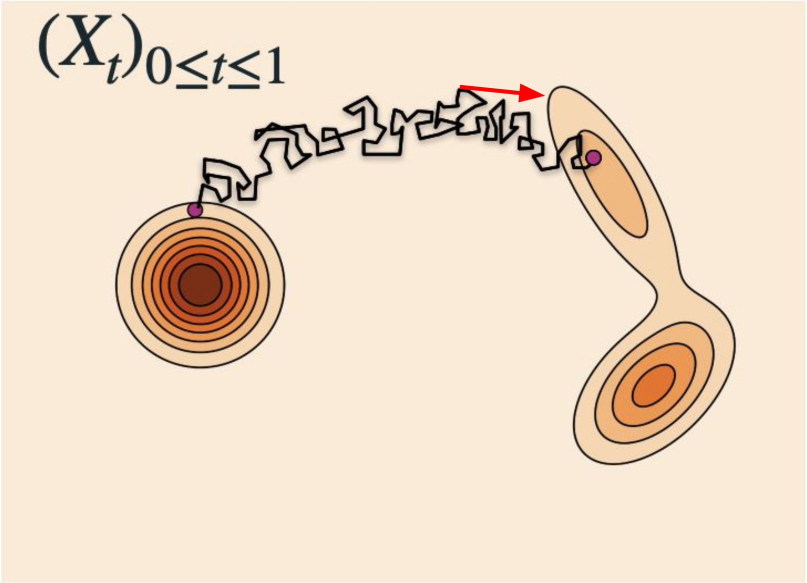
Applications:

- Stable Diffusion 3.0
- PI policy for Robotics

Piazza Discussion

1. Difference between FM and v-prediction





2. Why Use Diffusion After Flow Matching?

- Maturity of the ecosystem (SD2.1, CogVideox….)
- Under some limited data(**limited of data scale and diversity**) setting