

CS886: Theory and Applications of Generative Models

Lec 03: Flow

Yaoliang Yu



UNIVERSITY OF
WATERLOO

FACULTY OF MATHEMATICS
**DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE**

September 28, 2026

Fokker-Planck-Kolmogorov (FPK) Equation

- Recall that formally, $\frac{d\mathcal{P}_{s,t}}{dt} = \mathcal{P}_{s,t}\mathcal{L}_t$, hence

$$\begin{aligned}\langle f; \partial_t p(s, \mathbf{x}, t, \cdot) \rangle &= \partial_t \langle f; p(s, \mathbf{x}, t, \cdot) \rangle = \mathcal{P}_{s,t} \mathcal{L}_t f(\mathbf{x}) = \langle \mathcal{L}_t f; p(s, \mathbf{x}, t, \cdot) \rangle \\ &= \langle f; \mathcal{L}_t^* p(s, \mathbf{x}, t, \cdot) \rangle\end{aligned}$$

- For a diffusion process:

$$\partial_t p = \mathcal{L}_t^* p = - \langle \nabla, p \mathbf{b} \rangle + \frac{1}{2} \langle \nabla^2, p A \rangle \quad (\text{forward})$$

$$- \partial_s p = \mathcal{L}_s p = \langle \mathbf{b}, \nabla p \rangle + \frac{1}{2} \langle A, \nabla^2 p \rangle \quad (\text{backward})$$

FPK for SDE

Consider the stochastic differential equation

$$d\mathbf{X}_t = \mathbf{b}_t(\mathbf{X}_t) dt + G_t(\mathbf{X}_t) dB_t, \quad \mathcal{L}_t f(\mathbf{x}) := \langle \mathbf{b}_t(\mathbf{x}), \nabla f(\mathbf{x}) \rangle + \frac{1}{2} \langle G_t(\mathbf{x}) G_t(\mathbf{x})^\top, \nabla^2 f(\mathbf{x}) \rangle$$

Apply Itô's formula to obtain

$$df(\mathbf{X}_t) = \langle \mathbf{b}_t(\mathbf{X}_t), \nabla f(\mathbf{X}_t) \rangle + \frac{1}{2} \langle G_t(\mathbf{X}_t) G_t^\top(\mathbf{X}_t), \nabla^2 f(\mathbf{X}_t) \rangle dt + \nabla f(\mathbf{X}_t) \cdot G_t(\mathbf{X}_t) dB_t.$$

Take expectation from s at position $\mathbf{x}$ to t on both sides:

$$\begin{aligned} d\mathbb{E}[f(\mathbf{X}_t)] &= \mathbb{E} \left[\langle \mathbf{b}_t(\mathbf{X}_t), \nabla f(\mathbf{X}_t) \rangle + \frac{1}{2} \langle G_t(\mathbf{X}_t) G_t^\top(\mathbf{X}_t), \nabla^2 f(\mathbf{X}_t) \rangle \right] dt \\ \partial_t \langle f; p(s, \mathbf{x}, t, \mathbf{y}) \rangle &= \langle f; \partial_t p(s, \mathbf{x}, t, \mathbf{y}) \rangle = \langle \mathcal{L}_t f; p(s, \mathbf{x}, t, \mathbf{y}) \rangle = \langle f; \mathcal{L}_t^* p(s, \mathbf{x}, t, \mathbf{y}) \rangle \end{aligned}$$

$$\partial_t p = \mathcal{L}_t^* p = -\langle \nabla, p \mathbf{b} \rangle + \frac{1}{2} \langle \nabla^2, p G G^\top \rangle$$

$$\partial_s p = \mathcal{L}_s p = \langle \mathbf{b}, \nabla p \rangle + \frac{1}{2} \langle G G^\top, \nabla^2 p \rangle$$

- Continuity equation when $G = 0$

ODE $\Leftrightarrow$ SDE

$$d\mathbf{x}_t = \mathbf{v}_t(\mathbf{x}_t) dt$$

$$d\mathbf{X}_t = \mathbf{v}_t(\mathbf{X}_t) dt + \sigma_t d\mathbf{B}_t$$

- Any ODE is a (trivial) SDE with $\sigma_t \equiv 0$
- Conversely, any SDE is “equivalent” (meaning marginal preserving) to an ODE:

$$\mathbf{v}_t \leftarrow \mathbf{v}_t - \frac{1}{2}\sigma_t^2 \nabla_{\mathbf{x}} \log p_t$$

- The **score function** plays an important role:

$$\mathbf{s}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x}) := \nabla_{\mathbf{x}} \log p(\mathbf{x})$$

Path Space and Path Measure



- $\mathbf{x} = (\mathbf{x}_t)_{t \in [0,1]}$ is a path
- Disintegration (fancy name for joint = marginal $\times$ conditional):

$$p(\mathbf{x}) = p_0(\mathbf{x}_0) \cdot p(\mathbf{x}|\mathbf{x}_0) = p_1(\mathbf{x}_1) \cdot p(\mathbf{x}|\mathbf{x}_1)$$

Time Reversal



$$dX_t = \mathbf{v}_t(X_t) dt + \sigma_t dB_t$$

$$d\bar{X}_t = [\mathbf{v}_t(\bar{X}_t) - \sigma_t^2 \nabla \log p_t(\bar{X}_t)] dt + \sigma_t d\bar{B}_t$$

H. Föllmer. "An entropy approach to the time reversal of diffusion processes". In: *Stochastic Differential Systems Filtering and Control*. Springer, 1985, pp. 156–163.

Likelihood Ratio

$$dX_t = \mathbf{u}_t(X_t) dt + \sigma_t(X_t) dB_t$$

$$dZ_t = \mathbf{v}_t(Z_t) dt + \sigma_t(Z_t) dB_t$$

Theorem: Girsanov's Change of Variable

$$\frac{dX}{dZ}(X) = \exp \left[\int \left(\frac{\mathbf{u}_t(X_t) - \mathbf{v}_t(X_t)}{\sigma_t(X_t)} \right) dB_t + \frac{1}{2} \int \left\| \frac{\mathbf{u}_t(X_t) - \mathbf{v}_t(X_t)}{\sigma_t(X_t)} \right\|^2 dt \right]$$

Under mild conditions:

$$\text{KL}(X||Z) := \mathbb{E}_X \left[\log \frac{dX}{dZ} \right] = \text{KL}(X_0||Z_0) + \frac{1}{2} \int \mathbb{E} \left\| \frac{\mathbf{u}_t(X_t) - \mathbf{v}_t(X_t)}{\sigma_t(X_t)} \right\|^2 dt$$

Diffusion Model

$$dX_t = \mathbf{0} dt + \sigma_t dB_t$$

$$d\bar{X}_t = -\sigma_t^2 \nabla \log p_t(\bar{X}_t) dt + \sigma_t d\bar{B}_t$$

$$d\bar{Z}_t = -\sigma_t^2 \mathbf{s}_t(\bar{Z}_t; \boldsymbol{\theta}) dt + \sigma_t d\bar{B}_t$$

- Applying Girsanov's change of variable:

$$\text{KL}(\bar{X} \parallel \bar{Z}) = \cancel{\text{KL}(\bar{X}_1 \parallel \bar{Z}_1)} + \frac{1}{2} \int \mathbb{E}[\sigma_t^2 \|\nabla \log p_t(\bar{X}_t) - \mathbf{s}_t(\bar{X}_t; \boldsymbol{\theta})\|^2] dt$$

– this is nothing but (weighted) score matching!

- Can sample $\bar{X}_t = X_{1-t}$ by simulating the forward SDE:

$$X_0 \sim p_{\text{data}}, \quad dX_t = \mathbf{0} dt + \sigma_t dB_t$$

- But, how to avoid the true score $\nabla \log p_t$ in the objective?

Denoising Score Matching

Theorem: Tweedie's Formula

Let $X = Y + Z$ where $Y \perp Z \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Then,

$$\nabla \log p_t(\mathbf{x}) = -\Sigma^{-1} \cdot \mathbb{E}[Z|X = \mathbf{x}]$$

$$dX_t = \sigma_t dB_t \implies \boxed{X_t = X_0 + \epsilon_t}, \quad \text{where } \epsilon_t \sim \mathcal{N}(\mathbf{0}; \varsigma_t^2 I), \quad \varsigma_t^2 = \int_0^t \sigma_s^2 ds$$

$$\implies \nabla \log p_t(\mathbf{x}_t) = \mathbb{E} \left[\frac{\mathbf{x}_0 - \mathbf{x}_t}{\varsigma_t^2} \mid X_t = \mathbf{x}_t \right]$$

$$2\text{KL}(X \| Z) = \int \mathbb{E}[\sigma_t^2 \|\nabla \log p_t(X_t) - \mathbf{s}_t(X_t; \theta)\|^2] dt$$

$$= \int \mathbb{E}[\sigma_t^2 \|\mathbb{E} \left[\frac{\mathbf{x}_0 - \mathbf{x}_t}{\varsigma_t^2} \mid X_t \right] - \mathbf{s}_t(X_t; \theta)\|^2] dt \equiv \boxed{\mathbb{E}_{t, X_0, X_t} \left[\sigma_t^2 \left\| \frac{\mathbf{x}_0 - \mathbf{x}_t}{\varsigma_t^2} - \mathbf{s}_t(X_t; \theta) \right\|^2 \right]}$$

Score-based Diffusion Generative Models

$$d\mathbf{X}_t = \mathbf{v}_t(\mathbf{X}_t) dt + \sigma_t d\mathbf{B}_t$$

$$\mathbf{X}_{t+1} \approx \mathbf{X}_t + \eta_t \cdot \mathbf{v}_t(\mathbf{X}_t) + \sigma_t \sqrt{\eta_t} \cdot \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- Key is to estimate the score $\mathbf{s}_t(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_t$
- Apply denoising score matching: $\mathbb{E}[\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{z}) \mid \mathbf{x}] = \nabla_{\mathbf{x}} \log p(\mathbf{x})$

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t \sim \mu, (\mathbf{X}_t, \mathbf{X}_0) \sim p(\mathbf{x}_t, \mathbf{x}_0)} \lambda_t \|\mathbf{s}_t(\mathbf{X}_t; \boldsymbol{\theta}) - \nabla_{\mathbf{x}} \log p(\mathbf{X}_t | \mathbf{X}_0)\|_2^2$$

- $\mathbf{X}_0 \sim p_{\text{data}}(\mathbf{x})$, the data density
- $p(\mathbf{x}_t | \mathbf{x}_0)$ can be derived from the forward SDE, in closed-form if $\mathbf{v}_t$ is affine

Simple or Sophisticated?

Consider two stochastic processes and their reverse processes:

$$\begin{aligned}d\mathbf{X}_t &= d\mathbf{B}_t, \quad \mathbf{X}_0 \sim p_0, & d\bar{\mathbf{X}}_t &= -\nabla \log p_t(\bar{\mathbf{X}}_t) + d\mathbf{B}_t, \\d\mathbf{Z}_t &= d\mathbf{B}_t, \quad \mathbf{Z}_0 \sim q_0, & d\bar{\mathbf{Z}}_t &= -\nabla \log q_t(\bar{\mathbf{Z}}_t) + d\mathbf{B}_t.\end{aligned}$$

Then, applying Girsanov's theorem twice we obtain

$$\text{KL}(\mathbf{X} \parallel \mathbf{Z}) = \text{KL}(p_0 \parallel q_0) + 0 = \text{KL}(\bar{\mathbf{X}} \parallel \bar{\mathbf{Z}}) = \text{KL}(\mathbf{X}_T \parallel \mathbf{Z}_T) + \frac{1}{2} \int_0^T \mathbb{E}[\|\nabla \log p_t(\mathbf{X}_t) - \nabla \log q_t(\mathbf{X}_t)\|_2^2] dt,$$

Thus,

$$\lim_{T \rightarrow \infty} \text{KL}(\mathbf{X}_T \parallel \mathbf{Z}_T) \rightarrow 0 \implies \text{KL}(p_0 \parallel q_0) = \frac{1}{2} \int_0^\infty \mathbb{E}[\|\nabla \log p_t(\mathbf{X}_t) - \nabla \log q_t(\mathbf{X}_t)\|_2^2] dt.$$

Flow

Neural Ordinary Differential Equations (ODE)

$$\dot{\mathbf{X}}_t = \mathbf{v}_t(\mathbf{X}_t; \mathbf{w}), \quad \mathbf{X}_0 \sim p_{\text{noise}}, \quad \mathbf{X}_{t+1} \approx \mathbf{X}_t + \eta_t \cdot \mathbf{v}_t(\mathbf{X}_t)$$

- Continuous change-of-variable formula:

$$\frac{d \log p_t(\mathbf{x}_t)}{dt} = - \langle \nabla; \mathbf{v}_t(\mathbf{x}_t) \rangle$$

$$\frac{d \log p_t(\mathbf{x}_t)}{dt} = \frac{1}{p_t} [\partial_t p_t + \langle \nabla p_t; \mathbf{v}_t \rangle] = \frac{1}{p_t} [- \langle \nabla; p_t \mathbf{v}_t \rangle + \langle \nabla p_t; \mathbf{v}_t \rangle] = - \langle \nabla; \mathbf{v}_t(\mathbf{x}_t) \rangle$$

- Can compute the **log-likelihood along the trajectory** of $\mathbf{x}_t$:

$$\begin{cases} \dot{\mathbf{x}}_t &= \mathbf{v}_t(\mathbf{x}_t, \mathbf{u}) \\ \dot{l}_t &= - \langle \nabla_{\mathbf{x}}; \mathbf{v}_t(\mathbf{x}_t, \mathbf{u}) \rangle \end{cases}, \text{ with initial conditions } \mathbf{x}_0 \text{ and } l_1.$$

MLE for Neural ODE

- Learn the vector field $\mathbf{v}_t(\cdot; \mathbf{w})$ so that $X_1 \sim p_{\text{data}}$

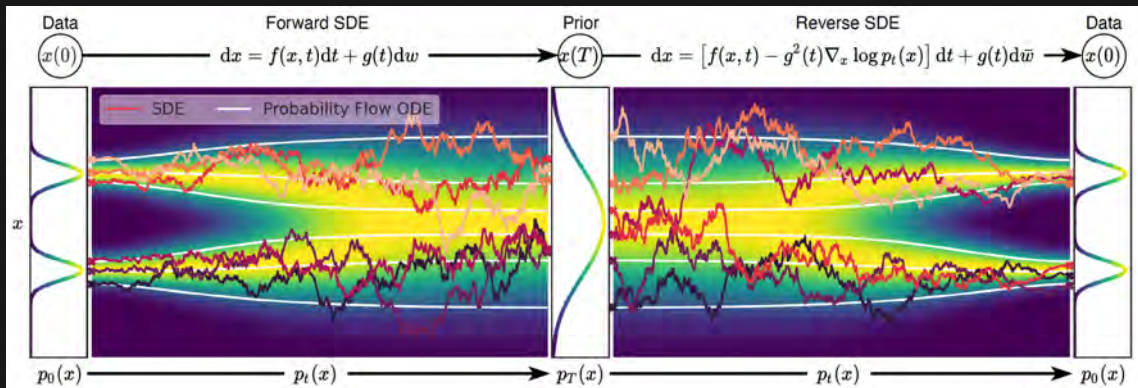
Representation

Can every p_{data} be obtained through some vector field $\mathbf{v}_t$?

- Maximum likelihood estimation:

$$\min \text{KL}(p_{\text{data}} \| p_1) = \text{KL}(q_0 \| p_{\text{noise}})$$

- p_1 is obtained by running ODE forward with initialization p_{noise}
- q_0 is obtained by running ODE backward with initialization p_{data}



Y. Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*. 2021.

Dynamic Optimal Transport

- Let ρ_0 and ρ_1 be two given densities, with 2-Wasserstein distance

$$\mathbb{W}_2^2(\rho_0, \rho_1) = \min_{T: T_{\#}\rho_0 = \rho_1} \mathbb{E}_{\mathbf{X} \sim \rho_0} \|\mathbf{T}\mathbf{X} - \mathbf{X}\|_2^2, \quad \text{where} \quad \rho_0(\mathbf{x}) = \rho_1(\mathbf{T}\mathbf{x}) \cdot \det \mathbf{T}'\mathbf{x}$$

– $\exists$ a (unique) $\mathbf{T} = \nabla \psi$ for some convex ψ

- Consider the ODE:

$$\dot{\mathbf{X}}_t(\mathbf{x}) = \mathbf{b}_t(\mathbf{X}_t), \quad \text{where} \quad \mathbf{X}_0 \sim p_0$$

– densities $\mathbf{X}_t \sim p_t$ satisfy the continuity equation and boundary conditions:

$$\partial_t p_t = -\langle \nabla, p_t \mathbf{b}_t \rangle, \quad p_0 = \rho_0, \quad p_1 = \rho_1$$

DOT (cont')

CLAIM:

$$W_2^2(\rho_0, \rho_1) = \min_{p_t} \int_0^1 \|\partial_t p_t\|_{p_t}^2 dt, \quad \text{s.t.} \quad p_0 = \rho_0, p_1 = \rho_1, \quad \text{where}$$
$$\|\partial_t p_t\|_{p_t}^2 = \min_{\varphi_t} \int p_t(\mathbf{x}) \|\nabla \varphi_t(\mathbf{x})\|_2^2 d\mathbf{x}, \quad \text{s.t.} \quad \partial_t p_t = -\langle \nabla, p_t \nabla \varphi_t \rangle,$$

- Strikingly similar to the **geodesic** distance on a **Riemannian manifold**!
- The optimal $\{p_t\}$ acts as “geodesic” connecting two densities ρ_0 and ρ_1

$$\begin{aligned}
\int \int_0^1 p_t(\mathbf{x}) \|\mathbf{b}_t(\mathbf{x})\|_2^2 dt d\mathbf{x} &= \int \int_0^1 \|\mathbf{b}_t(\mathbf{X}_t(\mathbf{x}))\|_2^2 dt \cdot p_0(\mathbf{x}) d\mathbf{x} = \int \int_0^1 \|\dot{\mathbf{X}}_t(\mathbf{x})\|_2^2 dt \cdot p_0(\mathbf{x}) d\mathbf{x} \\
&\geq \int \left\| \int_0^1 \dot{\mathbf{X}}_t(\mathbf{x}) dt \right\|_2^2 \cdot p_0(\mathbf{x}) d\mathbf{x} = \int p_0(\mathbf{x}) \|\mathbf{X}_1 - \mathbf{x}\|_2^2 d\mathbf{x} \\
&\geq \mathbb{W}_2^2(\rho_0, \rho_1)
\end{aligned}$$

- The first inequality is attained if $\mathbf{X}_t$ is affine in t
- The second is attained if $\mathbf{X}_1(\mathbf{x}) = \nabla\psi(\mathbf{x})$ for some convex ψ
- Thus, minimum is attained iff $\mathbf{X}_t(\mathbf{x}) = (1-t)\mathbf{x} + t\nabla\psi(\mathbf{x})$ and

$$\mathbf{b}_t(\mathbf{X}_t) = \dot{\mathbf{X}}_t(\mathbf{x}) = (\nabla\psi - \text{Id})(\mathbf{x}) = (\nabla\psi - \text{Id}) \circ [\text{Id} + t(\nabla\psi - \text{Id})]^{-1}(\mathbf{X}_t)$$

Summary

$$\begin{aligned} W_2^2(\rho_0, \rho_1) &= \min_{p_t, \mathbf{b}_t \text{ s.t. } \partial_t p_t = -\langle \nabla, p_t \mathbf{b}_t \rangle} \int \int_0^1 p_t(\mathbf{x}) \|\mathbf{b}_t(\mathbf{x})\|_2^2 dt d\mathbf{x} \\ &= \min_{p_t: p_0 = \rho_0, p_1 = \rho_1} \int_0^1 \min_{\mathbf{b}_t: \partial_t p_t = -\langle \nabla, p_t \mathbf{b}_t \rangle} \int p_t(\mathbf{x}) \|\mathbf{b}_t(\mathbf{x})\|_2^2 d\mathbf{x} dt \\ &= \min_{p_t: p_0 = \rho_0, p_1 = \rho_1} \int_0^1 \min_{\varphi_t: \partial_t p_t = -\langle \nabla, p_t \nabla \varphi_t \rangle} \int p_t(\mathbf{x}) \|\nabla \varphi_t(\mathbf{x})\|_2^2 d\mathbf{x} dt \end{aligned}$$

- The optimal vector field $\mathbf{b}_t$ is potential (gradient of convex function)
- $\exists$ vector field $\mathbf{b}_t$ so that $\dot{X}_t = \mathbf{b}_t(X_t)$ connects any pair ρ_0 and ρ_1
- In fact, such vector field exists for any (absolutely continuous) trajectory $\{p_t\}$!

Calculus of Variation

Alternatively, we may examine the minimizer $\mathbf{b}_t$ by perturbing its objective:

$$\int p_t(\mathbf{x}) \|\mathbf{b}_t(\mathbf{x}) + \epsilon \mathbf{c}_t(\mathbf{x})\|_2^2 d\mathbf{x}, \quad \text{s.t.} \quad \partial_t p_t = -\langle \nabla, p_t \mathbf{b}_t \rangle - \epsilon \langle \nabla, p_t \mathbf{c}_t \rangle.$$

- To maintain the continuity equation we must have $\langle \nabla, p_t \mathbf{c}_t \rangle = 0$
- Take derivative w.r.t. ϵ at 0:

$$\frac{d}{d\epsilon} \Big|_{\epsilon=0} = 2 \int \langle p_t \mathbf{c}_t, \mathbf{b}_t \rangle d\mathbf{x} = 0$$

Thus, $\mathbf{b}_t \perp (p_t \mathbf{c}_t)$ for any $p_t \mathbf{c}_t$ such that $\langle \nabla, p_t \mathbf{c}_t \rangle = 0$. According to the [Helmholtz decomposition](#), there exists a potential function φ_t such that the minimizer $\mathbf{b}_t = \nabla \varphi_t$.

Action Matching

Given a trajectory $\{p_t\}$, learn a potential field $\nabla\psi_t$ to solve the continuity equation:

$$\partial_t p_t = -\langle \nabla; p_t \nabla \psi_t \rangle, \quad \dot{\mathbf{X}}_t = \nabla \psi_t(\mathbf{X}_t)$$

- We already know such vector fields exist (albeit not necessarily potential)
- Match it:

$$\min_{\psi} \mathbb{E} \|\nabla \psi_t(\mathbf{X}_t) - \mathbf{b}_t(\mathbf{X}_t)\|_2^2$$

- Don't know $\mathbf{b}_t$... Integration by parts comes to rescue!

$$\begin{aligned}
\int \|\nabla\psi_t(\mathbf{x}) - \mathbf{b}_t(\mathbf{x})\|_2^2 \cdot p_t(\mathbf{x}) \, d\mathbf{x} &\equiv \int \|\nabla\psi_t(\mathbf{x})\|_2^2 \cdot p_t(\mathbf{x}) \, d\mathbf{x} - 2 \int \langle \nabla\psi_t(\mathbf{x}); p_t(\mathbf{x})\mathbf{b}_t(\mathbf{x}) \rangle \, d\mathbf{x} \\
&= \int \|\nabla\psi_t(\mathbf{x})\|_2^2 \cdot p_t(\mathbf{x}) \, d\mathbf{x} + 2 \int \langle \psi_t(\mathbf{x}); \nabla(p_t(\mathbf{x})\mathbf{b}_t(\mathbf{x})) \rangle \, d\mathbf{x} \\
&= \int \|\nabla\psi_t(\mathbf{x})\|_2^2 \cdot p_t(\mathbf{x}) \, d\mathbf{x} - 2 \int \psi_t(\mathbf{x}) \cdot \partial_t p_t \, d\mathbf{x}
\end{aligned}$$

Integrating over t :

$$\begin{aligned}
\mathbb{E}\|\nabla\psi_t(\mathbf{X}_t) - \mathbf{b}_t(\mathbf{X}_t)\|_2^2 &\equiv \mathbb{E}\|\nabla\psi_t(\mathbf{X}_t)\|_2^2 - 2 \int_0^1 \int \psi_t(\mathbf{x}) \cdot \partial_t p_t \, d\mathbf{x} \, dt \\
&= \boxed{\mathbb{E}\|\nabla\psi_t(\mathbf{X}_t)\|_2^2 - 2 \int_0^1 \int \partial_t \psi_t(\mathbf{x}) \cdot p_t \, d\mathbf{x} \, dt - 2 \int [\psi_1(\mathbf{x})p_1(\mathbf{x}) - \psi_0(\mathbf{x})p_0(\mathbf{x})] \, d\mathbf{x}}
\end{aligned}$$

- $\nabla\psi_t$ is essentially the potential part of $\mathbf{b}_t$

Preserving the Marginals

Consider the most general SDE:

$$dX_t(\omega) = \mathbf{u}_t(\omega) dt + \sigma_t(\omega) dB_t(\omega)$$

Theorem: Mimicking

Under mild conditions, the process X_t admits representation

$$dX_t(\omega) = \mathbf{v}_t(X(\omega)) dt + \sigma_t(X(\omega)) dB_t(\omega),$$

where $\mathbf{v}_t(X) = \mathbb{E}[\mathbf{u}_t | X_s, s \leq t]$ and $\sigma_t(X(\omega)) = \sigma_t(\omega)$. Moreover, $\exists$ a process

$$d\tilde{X}_t = \tilde{\mathbf{v}}_t(\tilde{X}_t) dt + \tilde{\sigma}_t(\tilde{X}_t) dB_t$$

such that for all t , $\tilde{X}_t \sim X_t$, where $\tilde{\mathbf{v}}_t(\mathbf{x}) := \mathbb{E}[\mathbf{u}_t | X_t = \mathbf{x}]$ and $\tilde{\sigma}_t^2(\mathbf{x}) := \mathbb{E}[\sigma_t^2 | X_t = \mathbf{x}]$.

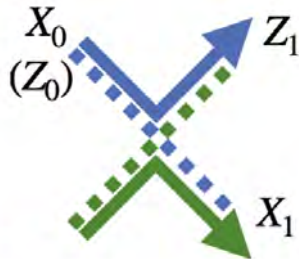
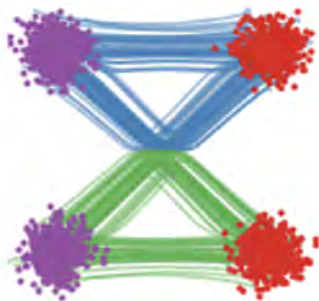
Rectified Flow

- Sample a **pair** $(X_0, X_1) \sim \pi$ and interpolate $X_t = (1 - t)X_0 + tX_1$
- Approximate the velocity field by a neural net $\mathbf{u}_t(\cdot)$:

$$\min_{\mathbf{u}_t(\cdot)} \mathbb{E}[\|\mathbf{u}_t(X_t) - \mathbb{E}(\dot{X}_t | X_t)\|_2^2] = \boxed{\min_{\mathbf{u}_t(\cdot)} \mathbb{E}[\|\mathbf{u}_t(X_t) - \dot{X}_t\|_2^2]} - \mathbb{E}[\|\dot{X}_t - \mathbb{E}(\dot{X}_t | X_t)\|_2^2]$$

- Solve the ODE $\dot{X}_t = \mathbf{u}_t(X_t)$ with initial condition X_0 (or end condition X_1)
 - effectively update the coupling π into a deterministic one that lowers convex transport cost
- Reflow: repeat the above
- Distill: $T(X_0) := X_0 + \mathbf{u}_0(X_0)$ where $\mathbf{u}_0$ minimizes the loss

$$\min_{\mathbf{u}_0(\cdot)} \mathbb{E}[\|X_1 - X_0 - \mathbf{u}_0(X_0)\|_2^2]$$



(a) Linear interpolation

(b) Rectified flow Z_t

$$X_t = tX_1 + (1 - t)X_0 \quad \text{induced by } (X_0, X_1)$$

- Interpolation process X_t is non-causal, possibly crossing



(c) Linear interpolation

$$Z_t = tZ_1 + (1 - t)Z_0$$

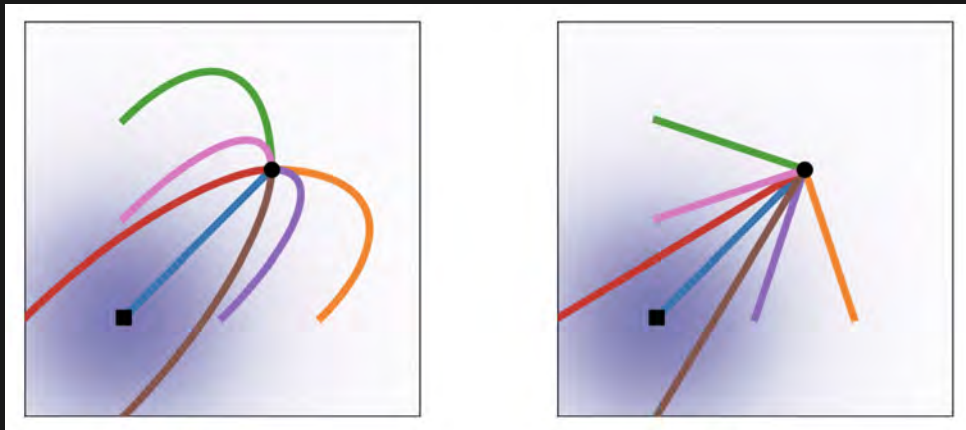
(d) Rectified flow Z'_t

induced by (Z_0, Z_1)

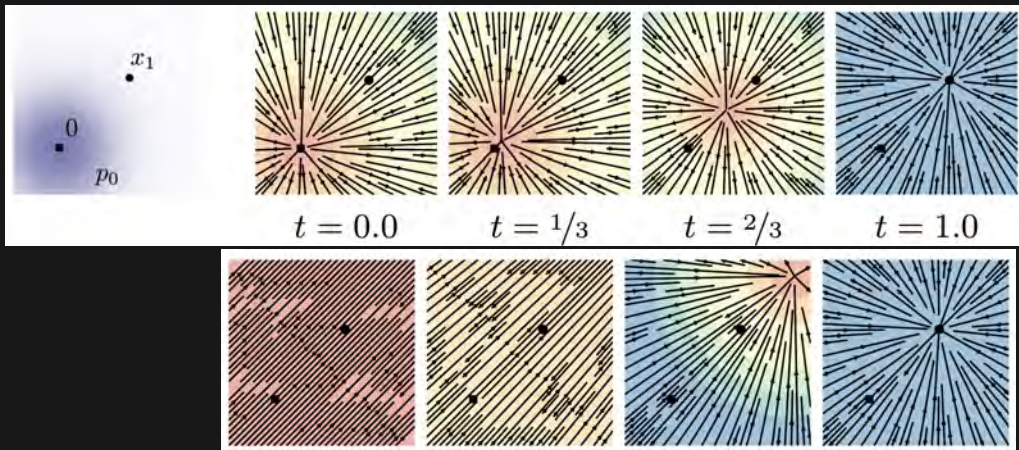
- After rewiring: causal, non-crossing

Flow Matching

$$\mathbf{X}_t = \mathbf{m}_t(\mathbf{X}_{data}) + \sigma_t(\mathbf{X}_{data}) \cdot \boldsymbol{\epsilon}, \quad \text{where} \quad \mathbf{X}_{data} \perp \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I)$$



$$\mathbf{s}_t(\mathbf{x}_t \mid \mathbf{x}_{data}) := \nabla \log p_t(\mathbf{x}_t \mid \mathbf{x}_{data}) = -\frac{1}{\sigma_t^2(\mathbf{x}_{data})}[\mathbf{x}_t - \mathbf{m}_t(\mathbf{x}_{data})]$$



$$\mathbf{v}_t(\mathbf{x}_t, \mathbf{x}_{data}) = \dot{\mathbf{x}}_t = \dot{\mathbf{m}}_t(\mathbf{x}_{data}) + \frac{\dot{\sigma}_t(\mathbf{x}_{data})}{\sigma_t(\mathbf{x}_{data})}[\mathbf{x}_t - \mathbf{m}_t(\mathbf{x}_{data})]$$

$$\mathbf{x}_t = \boxed{t\mathbf{x}_{data} + (1-t)\boldsymbol{\epsilon}} + \sigma t \boldsymbol{\epsilon}$$

$$\mathbf{v}_t(\mathbf{x}_t, \mathbf{x}_{data}) = \mathbf{x}_{data} + \frac{\sigma - 1}{1 - t + \sigma t}(\mathbf{x}_t - t\mathbf{x}_{data}) = \mathbf{x}_{data} - (1 - \sigma)\boldsymbol{\epsilon} = \dot{\mathbf{x}}_t$$

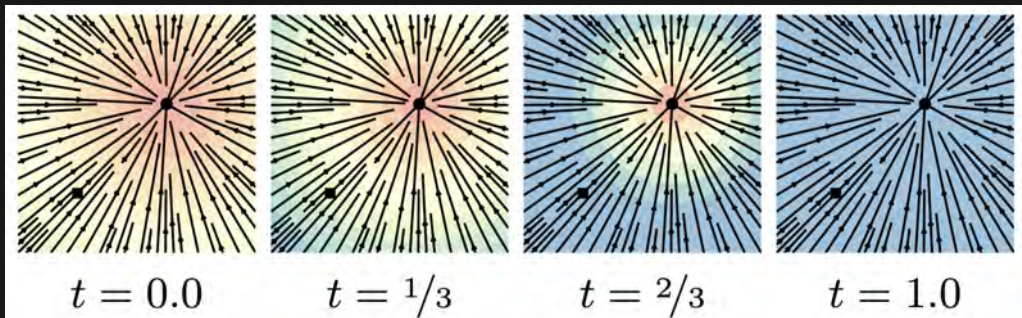
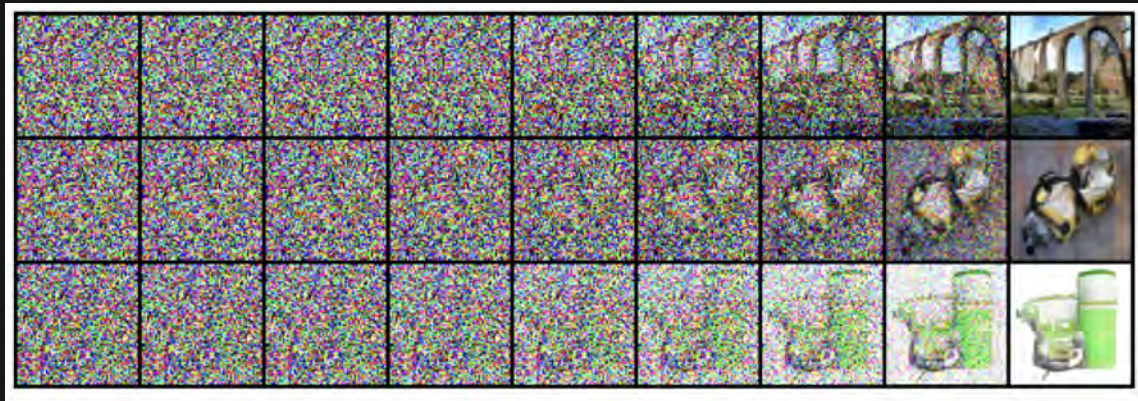
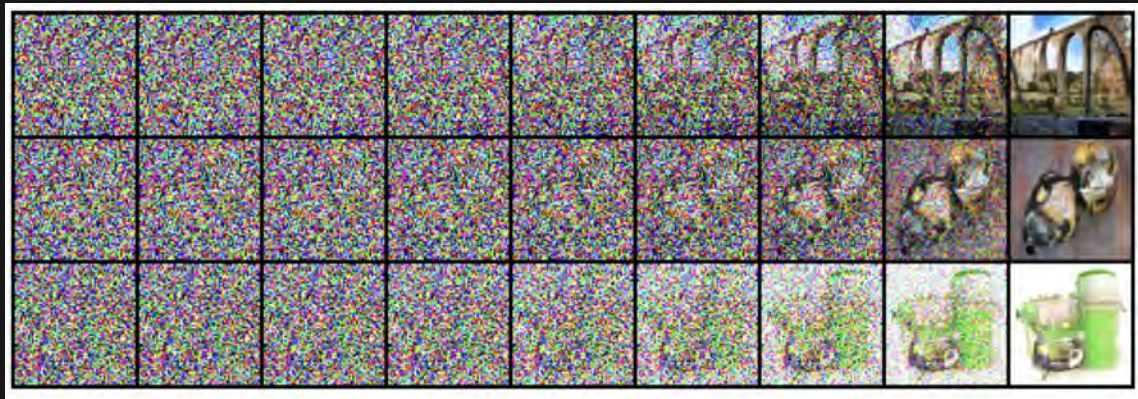


Image Generation: Score Matching with Diffusion Path



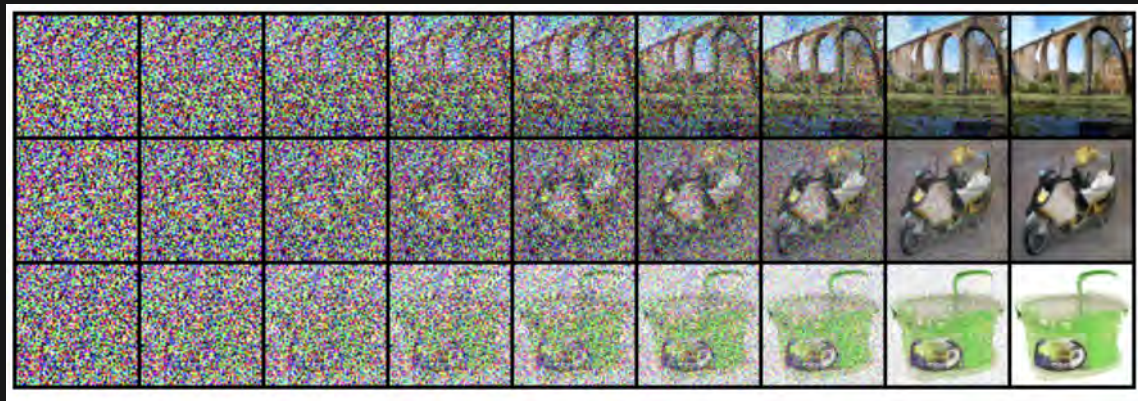
Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. "Flow Matching for Generative Modeling". In: *International Conference on Learning Representations*. 2023.

Image Generation: Flow Matching with Diffusion Path



Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. "Flow Matching for Generative Modeling". In: *International Conference on Learning Representations*. 2023.

Image Generation: Flow Matching with Straight Path



Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. "Flow Matching for Generative Modeling". In: *International Conference on Learning Representations*. 2023.

Let's Build Bridges

Doob's h -transform

Let X_t be a Markov process with transition kernel $p(s, \mathbf{x}, t, d\mathbf{y})$. Fix an end time T .

- Define

$$h(t, \mathbf{y}) := p(t, \mathbf{y}, T, A) = \Pr(X_T \in A | X_t = \mathbf{y})$$

- Conditioned on the event $X_T \in A$, we have a new Markov kernel:

$$\begin{aligned}\tilde{p}(s, \mathbf{x}, t, B) &:= \Pr(X_t \in B | X_s = \mathbf{x}, X_T \in A) = \frac{\Pr(X_t \in B, X_T \in A | X_s = \mathbf{x})}{\Pr(X_T \in A | X_s = \mathbf{x})} \\ &= \int_B h(t, \mathbf{y}) p(s, \mathbf{x}, t, d\mathbf{y}) / h(s, \mathbf{x}),\end{aligned}$$

or written more succinctly in terms of the balance equation:

$$h(s, \mathbf{x}) \tilde{p}(s, \mathbf{x}, t, d\mathbf{y}) = p(s, \mathbf{x}, t, d\mathbf{y}) h(t, \mathbf{y}), \quad i.e., \quad \tilde{p} = h^{-1} p h.$$

More Generally

- Everything remains true for any **positive h that is consistent**:

$$\mathbb{E}[h(t, \mathbf{X}_t) | \mathbf{X}_s] = h(s, \mathbf{X}_s), \quad i.e., \quad h = \mathcal{P}_{s,t}h.$$

- In particular, for Markov densities we may set

$$h(t, \mathbf{y}) := p(t, \mathbf{y}, T, \mathbf{x}_T), \quad h(T, \mathbf{y}) = p(T, \mathbf{y}, T, \mathbf{x}_T) = \delta_{\mathbf{x}_T}(\mathbf{y}),$$

where $\mathbf{x}_T$ is a fixed value that the process ends at.

h -transform for Diffusions

Consider the diffusion process

$$d\mathbf{X}_t = \mathbf{b}_t(\mathbf{X}_t) dt + G_t(\mathbf{X}_t) dB_t, \quad \mathcal{L}f = \langle \nabla f, \mathbf{b} \rangle + \frac{1}{2} \langle \nabla^2 f, A \rangle, \quad A = G_t G_t^\top$$

The generator of the conditional Markov kernel $\tilde{p}$ is

$$\tilde{\mathcal{L}}_s f = \frac{1}{h} (\partial_s + \mathcal{L}_s)(hf) = \langle \nabla \log h \otimes \nabla f, A \rangle + \mathcal{L}_s f = \langle \nabla f, A \nabla \log h \rangle + \mathcal{L}_s f,$$

corresponding to the SDE:

$$d\mathbf{X}_t = [\mathbf{b}_t(\mathbf{X}_t) + G_t(\mathbf{X}_t) G_t(\mathbf{X}_t)^\top \nabla \log h(t, \mathbf{X}_t)] dt + G_t(\mathbf{X}_t) dB_t.$$

Brownian Bridge

Consider a Brownian motion with transition density p :

$$\forall t \in [0, T], \quad dX_t = dB_t, \quad X_0 = 0$$

- Define

$$h(t, \mathbf{x}) := p(t, \mathbf{x}, T, \mathbf{0}) \propto \exp \left(-\frac{\|\mathbf{0} - \mathbf{x}\|^2}{2(T-t)} \right)$$

- Thus, $\nabla \log h(t, \mathbf{x}) = -\frac{\mathbf{x}}{T-t}$
- The conditional process corresponds to the SDE:

$$dX_t = -\frac{X_t}{T-t} dt + dB_t,$$

namely the Brownian bridge.

Schrodinger Bridge

$$\min_{Q: Q_0=\mu, Q_1=\nu} \text{KL}(Q\|P)$$

- P is the reversible Brownian motion: $P = d\mathbf{x}_0 \cdot \mathbf{B}_{\mathbf{x}_0}$