Flow Matching

Denis Rakitin

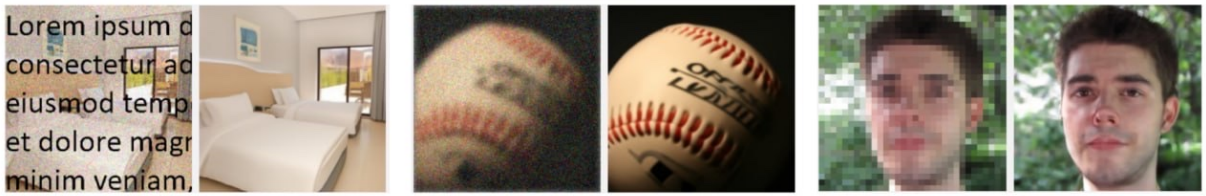
Image 2 Image translation

image credits

https://arxiv.org/pdf/2201.11793
https://arxiv.org/pdf/2207.06635v1
https://arxiv.org/pdf/2201.12220
https://arxiv.org/pdf/1611.07004

Paired:

Inpainting | Deblurring | Super-resolution



Day to Night | Edges to Photo | Labels to Facade

Unpaired:

Handbags2shoes | Cat2wild | Male2female

Diffusion models recap

$X_0 \sim p_{data}$ $\quad dX_t = -\frac{\beta_t}{2} X_t dt + \sqrt{\beta_t}\, dW_t$

$X_t | X_0 \sim \mathcal{N}\left(\sqrt{\alpha_t} X_0,\ (1-\alpha_t) I\right), \quad \alpha_t = \exp\left(-\int_0^t \beta_s ds\right)$

Training: $\int_0^1 w_t \mathbb{E} \| D_\theta(X_t, t) - X_0 \|^2 dt \to \min_\theta$

Score function: $\nabla \log p_t(x) = -\frac{1}{1-\alpha_t}\left(X_t - \sqrt{\alpha_t}\, \mathbb{E}[X_0 | X_t = x]\right)$

Sampling:

$$dX_t^{(b)} = \left(\frac{\beta_{1-t} X_t^{(b)}}{2} + \frac{\beta_{1-t}}{2} \nabla \log p_{1-t}(X_t^{(b)})\right) dt$$

$$dX_t^{(b)} = \left(\frac{\beta_{1-t} X_t^{(b)}}{2} + \beta_{1-t} \nabla \log p_{1-t}(X_t^{(b)})\right) dt + \sqrt{\beta_{1-t}}\, dW_t$$

Observation

$\approx X_1$

VP-case: $X_t = \sqrt{\alpha_t}\, X_0 + \sqrt{1-\alpha_t}\, \varepsilon$

$\approx \sqrt{\alpha_t}\, X_0 + \sqrt{1-\alpha_t}\, X_1 = I_t(X_0, X_1)$

$I_t(x_0, x_1)$ is an interpolant: $I_0(x_0, x_1) = x_0$

$I_1(x_0, x_1) = x_1$.

Maybe interpolate between arbitrary p_0 and p_1?

Flow Matching (Bridge Matching)

Aim: given joint p_{01} with marginals p_0 and p_1, construct ODE/SDE

$$\begin{cases} dY_t = f(Y_t, t)dt \; (+ g(t)dW_t) & \text{such that } Y_1 \sim p_1. \\ Y_0 \sim p_0 \end{cases}$$

Input: process X_t: $(X_0, X_1) \sim p_{0,1}$ and $X_t | X_0, X_1 \sim p_{t|0,1}$ - some interpolation

Plan Assume $p_{t|0,1}$ is generated by ODE/SDE.

Infer the dynamics $p_t(x_t) = \int p_t(x_t | x_0, x_1) p_{01}(x_0, x_1) dx_0 dx_1$

Derivation

$$\frac{\partial}{\partial t} p_{t|0,1}(x|x_0,x_1) = -\frac{\partial}{\partial x}\left(p_{t|0,1}(x|x_0,x_1)\, f(x|x_0,x_1)\right) + \frac{g^2(t)}{2}\frac{\partial^2}{\partial x^2} p_{t|0,1}(x|x_0,x_1)$$

$$\frac{\partial}{\partial t} p_t(x) = \frac{\partial}{\partial t}\int p_t(x|x_0,x_1)\, p_{01}(x_0,x_1)\,dx_0\,dx_1 =$$

$$= \int \frac{\partial}{\partial t} p_t(x|x_0,x_1)\, p_{01}(x_0,x_1)\,dx_0\,dx_1 =$$

$$= \int \left(-\frac{\partial}{\partial x}\left(p_{t|0,1}(x|x_0,x_1)\, f(x|x_0,x_1)\right) + \frac{g^2(t)}{2}\frac{\partial^2}{\partial x^2} p_{t|0,1}(x|x_0,x_1)\right) p_{01}(x_0,x_1)\,dx_0\,dx_1$$

$$= -\frac{\partial}{\partial x}\int f(x|x_0,x_1)\, p_{t|0,1}(x|x_0,x_1)\, p_{01}(x_0,x_1)\,dx_0\,dx_1 + \frac{g^2(t)}{2}\frac{\partial^2}{\partial x^2}\int p_{t|0,1}(x|x_0,x_1)\, p_{01}(x_0,x_1)\,dx_0\,dx_1$$

$$= -\frac{\partial}{\partial x}\left(p_t(x)\int f(x|x_0,x_1)\frac{p_{t|0,1}(x|x_0,x_1)\, p_{01}(x_0,x_1)}{p_t(x)}\,dx_0\,dx_1\right) + \frac{g^2(t)}{2}\frac{\partial^2}{\partial x^2} p_t(x)$$

Alternative derivation

Recall FPE in weak form: for all $\varphi \in C_0^2$ holds

$$\frac{d}{dt}\int \varphi(x)\, p_t(x)\,dx =$$
$$\int \left(\langle \nabla\varphi(x), f(x,t)\rangle + \frac{g^2(t)}{2}\Delta\varphi(x) \right) p_t(x)\,dx$$

I.e. for $dX_t = f(X_t,t)dt + g(t)dW_t$ holds

$$\frac{d}{dt}\mathbb{E}\,\varphi(X_t) = \mathbb{E}\langle \nabla\varphi(X_t), f(X_t,t)\rangle + \frac{g^2(t)}{2}\mathbb{E}\,\Delta\varphi(X_t)$$

Alternative derivation

We know: $p_{t|0,1}$ is generated by FPE with $f(x,t|x_0,x_1)$ and $g(t)$:

$$\frac{d}{dt}\mathbb{E}\left[\varphi(X_t)\,|\,X_0=x_0,X_1=x_1\right] =$$

$$=\mathbb{E}\left[\left\langle\nabla\varphi(X_t),\, f(X_t,t|x_0,x_1)\right\rangle + \frac{g^2(t)}{2}\Delta\varphi(X_t)\,\middle|\,X_0=x_0,X_1=x_1\right]$$

$$\frac{d}{dt}\mathbb{E}\varphi(X_t) = \frac{d}{dt}\mathbb{E}\,\mathbb{E}\left[\varphi(X_t)\,|\,X_0,X_1\right] = \mathbb{E}\frac{d}{dt}\mathbb{E}\left[\varphi(X_t)\,|\,X_0,X_1\right]$$

$$=\mathbb{E}\,\mathbb{E}\left[\left\langle\nabla\varphi(X_t),\, f(X_t,t|X_0,X_1)\right\rangle + \frac{g^2(t)}{2}\Delta\varphi(X_t)\,\middle|\,X_0,X_1\right]$$

$$=\mathbb{E}\,\mathbb{E}\left[\left\langle\nabla\varphi(X_t),\, f(X_t,t|X_0,X_1)\right\rangle\,\middle|\,X_t\right] + \frac{g^2(t)}{2}\mathbb{E}\Delta\varphi(X_t)$$

$$=\mathbb{E}\left\langle\nabla\varphi(X_t),\, \mathbb{E}\left[f(X_t,t|X_0,X_1)\,|\,X_t\right]\right\rangle + \frac{g^2(t)}{2}\mathbb{E}\Delta\varphi(X_t)$$

$\square$

Conditional → unconditional

If $p_{t|0,1}$ is generated by $dX_t^{0,1} = f(X_t^{0,1}, t \mid x_0, x_1)dt + g(t)dW_t$,

then p_t is generated by $dY_t = f(Y_t, t)dt + g(t)dW_t$,

where $f(x,t) = \mathbb{E}\left[f(X_t \mid X_0, X_1) \mid X_t = x \right]$.

As usual, train as

$$\int_0^1 \mathbb{E} \| f_\theta(X_t, t) - f(X_t \mid X_0, X_1) \|^2 dt \to \min_\theta .$$

where $(X_0, X_1) \sim p_{01}$, $X_t \mid X_0, X_1 \sim p_{t|0,1}(\cdot \mid X_0, X_1)$

Examples

ODE $X_t = I_t(X_0, X_1) \Rightarrow dX_t^{0,1} = \frac{\partial}{\partial t} I_t(x_0, x_1)\,dt$

- $X_t = \sqrt{\alpha_t}\, X_0 + \sqrt{1-\alpha_t}\, X_1$ – diffusion
- $X_t = t X_1 + (1-t) X_0$ – straight interpolant

$$dX_t^{0,1} = (x_1 - x_0)\,dt = \left(x_1 - \frac{X_t^{0,1} - t x_1}{1-t}\right)dt = \frac{x_1 - X_t^{0,1}}{1-t}\,dt$$

Training: $\int_0^1 \mathbb{E} \| f_\theta(X_t, t) - (X_1 - X_0) \|^2 dt \to \min_\theta$

$$\Leftrightarrow \quad \int_0^1 \mathbb{E} \left\| f_\theta(X_t, t) - \frac{X_1 - X_t}{1-t} \right\|^2 dt \to \min_\theta$$

Examples

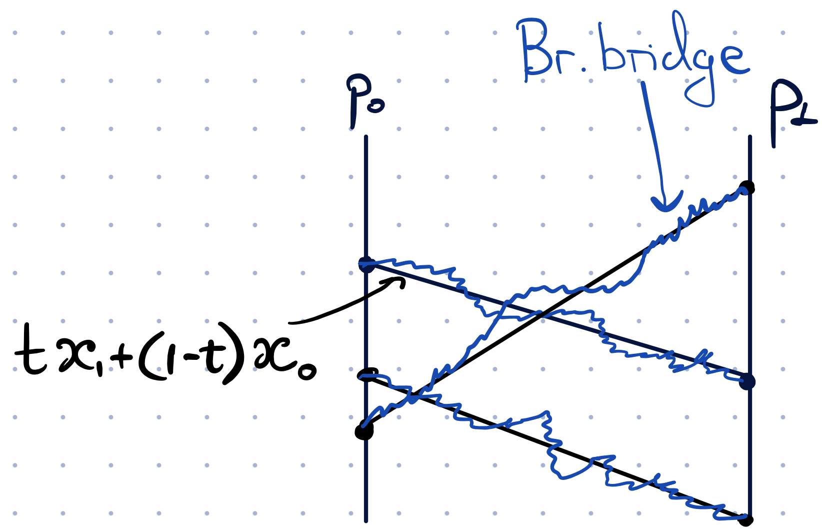
SDE Brownian bridge: $dX_t^{0,1} = \frac{x_1 - X_t^{0,1}}{1-t}dt + \sqrt{\beta}\, dW_t$

(Wiener process, conditioned on x_0 and x_1).

$X_t^{0,1} \sim \mathcal{N}(tx_1 + (1-t)x_0,\ \beta \cdot t(1-t) I)$

Training: $\int_0^1 \mathbb{E} \| f_\theta(X_t, t) - \frac{X_1 - X_t}{1-t} \|^2 dt \to \min_\theta$

Now, $X_t = tX_1 + (1-t)X_0 + \sqrt{\beta t(1-t)} \cdot \varepsilon$.

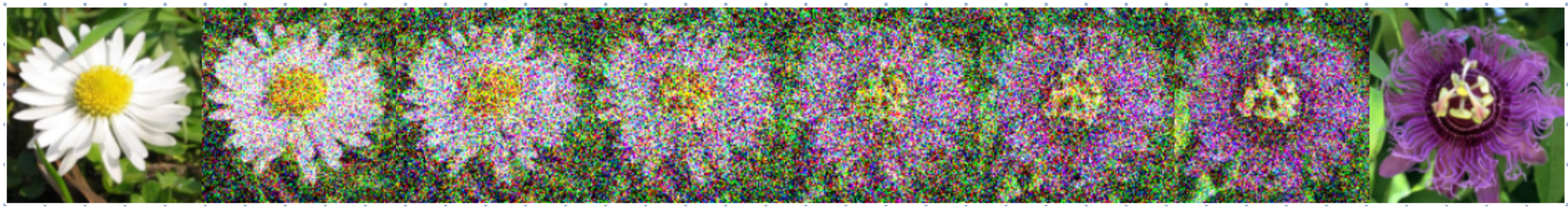


Visualization

$$X_t^{01} = t x_1 + (1-t) x_0 \; ; \; dX_t^{01} = (x_1 - x_0)dt = \frac{x_1 - X_t^{0,1}}{1-t}$$

x_0 x_1

$$dX_t^{01} = \frac{x_1 - X_t^{0,1}}{1-t} dt + \sqrt{\beta}\, dW_t$$



x_0 x_1

Summary

Interpolation process, defined by

$$dX_t^{0,1} = f(X_t^{0,1}, t \mid x_0, x_1)\,dt + g(t)\,dW_t \,, \quad (X_0, X_1) \sim p_{01}.$$

Trained process

$$dY_t = f(Y_t, t)\,dt + g(t)\,dW_t \,, \quad Y_0 \sim p_0.$$

$$f(x,t) = \mathbb{E}\left[f(X_t, t \mid X_0, X_1) \;\middle|\; X_t = x \right]$$

Claim that we proved: $X_t \overset{d}{=} Y_t$.

That means $Y_0 \to Y_1$ translates between p_0 and p_1!

Any problems?

Transport cost

The "map" between p_0 and p_1, defined by (Y_0, Y_1), should possess:

- $Y_0 \sim p_0 \quad Y_1 \sim p_1$
- Y_0 and Y_1 are related : $\mathbb{E}c(Y_0, Y_1)$ is small for some measure of distance c, called the cost function. $\mathbb{E}c(Y_0, Y_1)$ is called **transport cost**.

Rectification

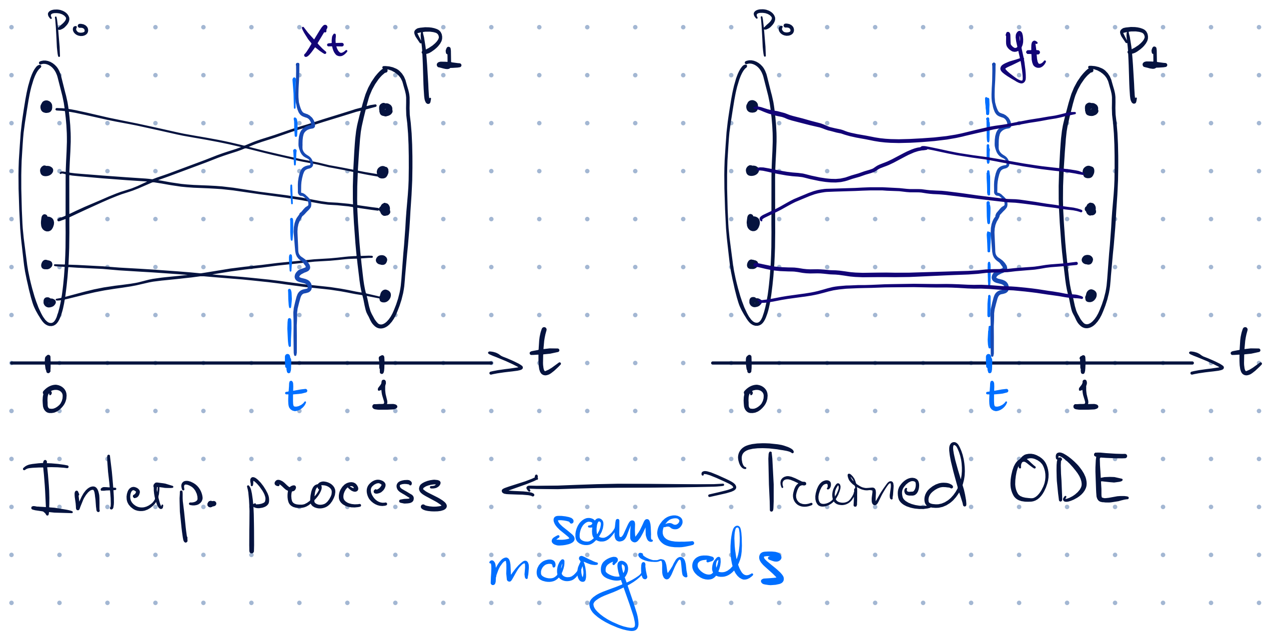
Theorem If $X_t = tX_1 + (1-t)X_0$, and

$dY_t = f(Y_t, t)dt$ with $f(x,t) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$,

then $\mathbb{E}\, c(Y_0, Y_1) \leq \mathbb{E}\, c(X_0, X_1)$ for all

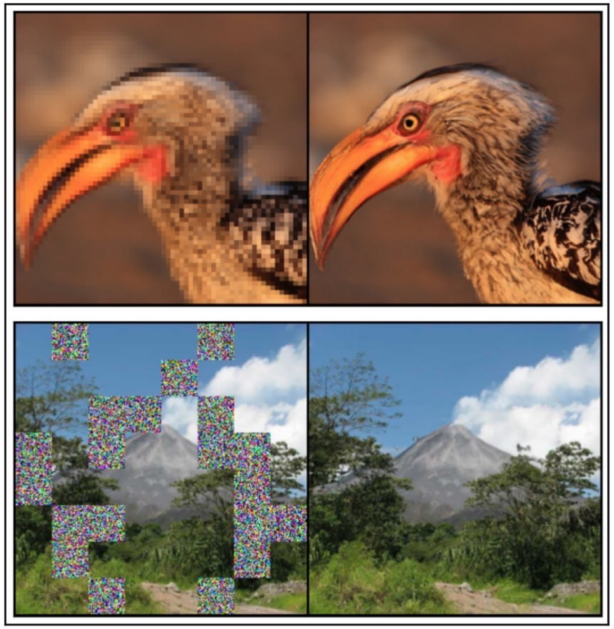
$c(x,y) = f(x-y)$ s.t. f is convex.

Less intersections $\Rightarrow$ less transport cost.

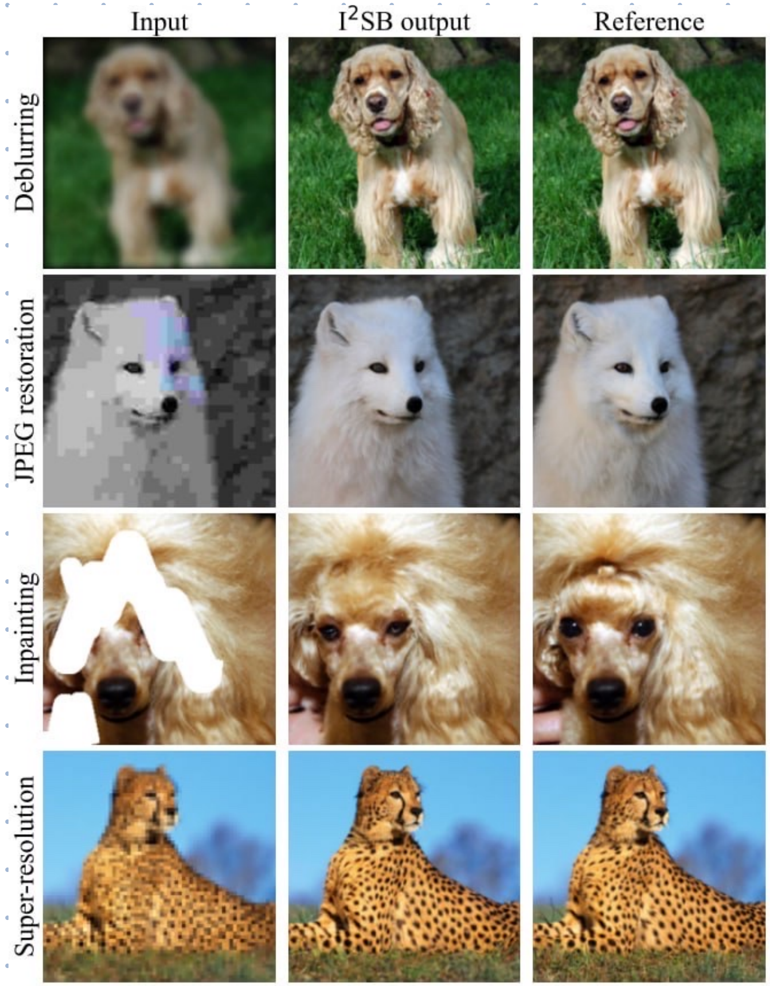


Paired tasks

Corollary: flow matching is applicable for paired (in terms of l_2) data sets.



https://arxiv.org/pdf/2310.03725v2



https://arxiv.org/pdf/2302.05872

Summary

- Flow/Bridge Matching trains ODE/SDE map between p_0 and p_1.
- Generalizes diffusion beyond unconditional generation.
- Straight interpolant reduces transport cost

Further:

- Iterate FM training for reducing cost mult. times (unpaired tasks)
- Straighten lines to reduce computations