Optimal Transport for Generative Modeling: Dynamic Views

Gianluca Covini

Presentation for the exam of *Modern Applied Machine Learning*

Based on: Tong et al., *Improving and Generalizing Flow-Based Generative Models with Minibatch OT*, TMLR 2024 · Shi et al., *Diffusion Schrödinger Bridge Matching*, NeurIPS 2023 · Peluchetti, *Diffusion Bridge Mixture Transports, Schrödinger Bridge Problems and Generative Modeling*, JMLR 2023

Generative Modeling as Transport

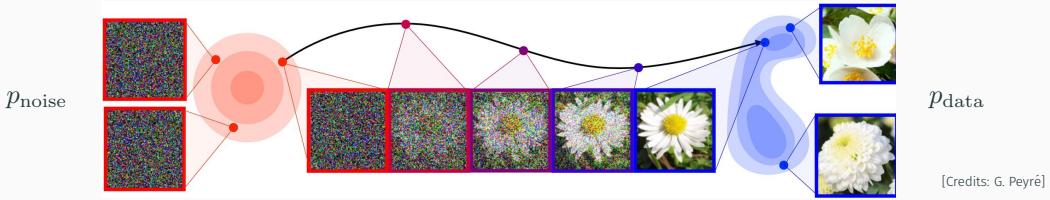
Generative modeling

Given a data distribution p_{data} on $\mathbb{R}^d$, learn how to sample from it starting from a simple reference distribution p_{noise}:

$$x_0 \sim p_{\text{noise}} \qquad \Longrightarrow \qquad x_T \sim p_{\text{data}}.$$

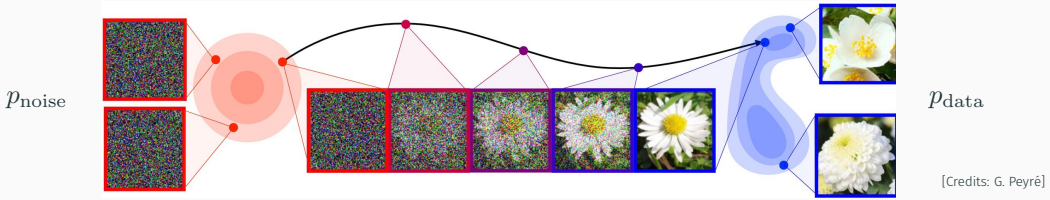
Transport viewpoint: A generative model can be seen as a mechanism transporting p_{noise} into p_{data}:

$$p_{\text{noise}} \xrightarrow{\text{dynamics (ODE/SDE)}} p_{\text{data}}.$$



[Credits: G. Peyré]

Generative Modeling as Transport



Transport viewpoint: A generative model can be seen as a mechanism transporting p_{noise} into p_{data}:

$$p_{\text{noise}} \xrightarrow{\text{dynamics (ODE/SDE)}} p_{\text{data}}.$$

ODE $\rightarrow$ Flow matching

$$\mathrm{d}x_t = u_t(x_t)\mathrm{d}t$$

deterministic transport / flows

$$\partial_t p_t(x_t) + \nabla \cdot (p_t(x_t)u_t(x_t)) = 0$$

SDE $\rightarrow$ Bridge matching

$$\mathrm{d}x_t = f_t(x_t)\mathrm{d}t + \sigma \mathrm{d}w_t$$

stochastic transport / diffusions

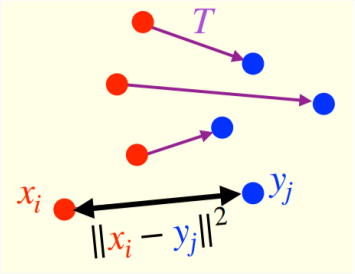
$$\partial_t p_t(x_t) = -\nabla \cdot (f_t(x_t)p_t(x_t)) + \tfrac{1}{2}\sigma^2 \Delta p_t(x_t)$$

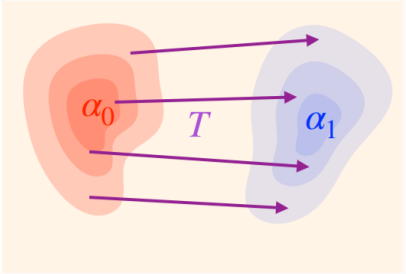
A Primer on Optimal Transport

Static optimal transport

Given two distributions $p_{\text{noise}}, p_{\text{data}} \in \mathcal{P}(\mathbb{R}^d)$, OT looks for an optimal coupling:

$$W_2^2(p_{\text{noise}}, p_{\text{data}}) = \inf_{\pi \in \Pi(p_{\text{noise}}, p_{\text{data}})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x_1 - x_0\|^2 \mathrm{d}\pi(x_0, x_1).$$





[Credits: G. Peyré]

A Primer on Optimal Transport

Static optimal transport

Given two distributions $p_{\text{noise}}, p_{\text{data}} \in \mathcal{P}(\mathbb{R}^d)$, OT looks for an optimal coupling:

$$W_2^2(p_{\text{noise}}, p_{\text{data}}) = \inf_{\pi \in \Pi(p_{\text{noise}}, p_{\text{data}})} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x_1 - x_0\|^2 \mathrm{d}\pi(x_0, x_1).$$

$\pi = p_{\text{noise}} \otimes p_{\text{data}}$ independent coupling

- any noise paired with any data
- Crossed paths and inefficient transport

$\pi = \pi_{\text{OT}}$ optimal coupling

- nearby noise paired with nearby data
- straight paths, efficient transport
- easier to integrate

Learning a generative model means learning an approximate transport dynamics from p_{noise} to p_{data}.

Dynamic Optimal Transport

From couplings to paths

The dynamic OT problem seeks the velocity field of minimal kinetic energy:

$$W_2^2(p_{\text{noise}}, p_{\text{data}}) = \inf_{(p_t, u_t)} \int_0^1 \int_{\mathbb{R}^d} \|u_t(x)\|^2 \mathrm{d}p_t(x)\mathrm{d}t,$$

subject to $\partial_t p_t + \nabla \cdot (p_t u_t) = 0, \quad p_{t=0} = p_{\text{noise}}, \quad p_{t=1} = p_{\text{data}}.$

Flow matching

Learn ODE dynamic $\mathrm{d}x_t = u_t(x_t)\mathrm{d}t$, i.e. $\partial_t p_t + \nabla \cdot (p_t u_t) = 0$. Regress on

$$\mathcal{L}_{FM}(\theta) := \mathbb{E}_{\substack{t \sim U(0,1) \\ x \sim p_t(x)}} \left[\|v_\theta(t, x) - u_t(x)\|^2 \right]$$

Problem: it is difficult to sample from p_t.

Dynamic Optimal Transport

From couplings to paths

The dynamic OT problem seeks the velocity field of minimal kinetic energy:

$$W_2^2(p_{\text{noise}}, p_{\text{data}}) = \inf_{(p_t, u_t)} \int_0^1 \int_{\mathbb{R}^d} \|u_t(x)\|^2 \mathrm{d}p_t(x)\mathrm{d}t,$$

subject to $\partial_t p_t + \nabla \cdot (p_t u_t) = 0, \quad p_{t=0} = p_{\text{noise}}, \quad p_{t=1} = p_{\text{data}}.$

Conditional flow matching [Tong et al., TMLR 2024]

Learn ODE dynamic $\mathrm{d}x_t = u_t(x_t)\mathrm{d}t$, i.e. $\partial_t p_t + \nabla \cdot (p_t u_t) = 0$.

Conditional probabilities are easier to sample: condition on starting and arriving points x_0, x_1 of the path.

$$\mathcal{L}_{CFM}(\theta) := \mathbb{E}_{\substack{t\sim U(0,1) \\ (x_0,x_1)\sim\pi \\ x|(x_0,x_1)\sim p_t(x|x_0,x_1)}} \left[\|v_\theta(t,x) - u_t(x \mid x_0, x_1)\|^2\right]$$

Theorem: It is equivalent to minimize $\mathcal{L}_{FM}$

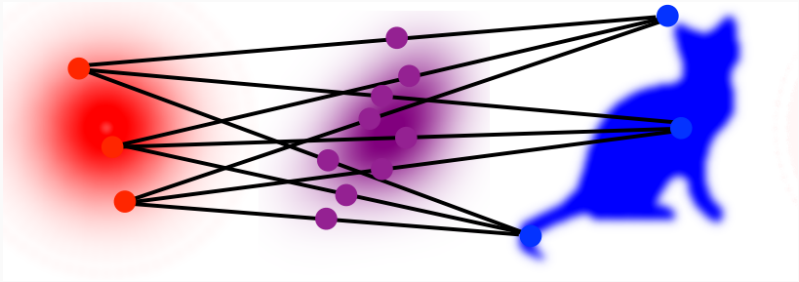
OT-Conditional Flow Matching

Which path?

- **Conditional distribution:** $(x_0, x_1) \sim \pi$, with $x_0 \sim p_{\text{noise}}$, $x_1 \sim p_{\text{data}}$.
- **Conditional interpolation:**

$$x_t = (1-t)x_0 + tx_1, \qquad u_t(x_t \mid x_0, x_1) = x_1 - x_0.$$

p_{noise}



p_{data}

[Credits: G. Peyré]

OT-Conditional Flow Matching

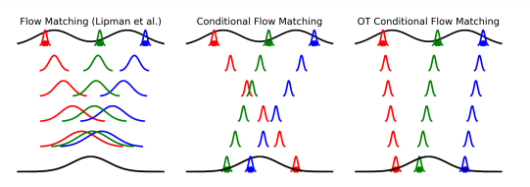
Which path?

- **Conditional distribution:** $(x_0, x_1) \sim \pi$, with $x_0 \sim p_{\text{noise}}$, $x_1 \sim p_{\text{data}}$.
- **Conditional interpolation:**

$$x_t = (1-t)x_0 + tx_1, \qquad u_t(x_t \mid x_0, x_1) = x_1 - x_0.$$

OT-CFM: choose $\pi = \pi_{\text{OT}}$

OT coupling $\Rightarrow$ straighter paths $\Rightarrow$ fewer neural function evaluations (NFE) at inference.



[Credits: Tong et al.]

OT-CFM: Simulation-Free Training

Dynamic OT: $W_2^2(p_0, p_1) = \inf_{\substack{(p_t, u_t) \\ \partial_t p_t + \nabla \cdot (p_t u_t) = 0 \\ p_{t=0} = p_{\text{noise}}, p_{t=1} = p_{\text{data}}}} \int_0^1 \int_{\mathbb{R}^d} \|u_t(x)\|^2 \mathrm{d}p_t(x) \mathrm{d}t$

Choosing π_{OT} and linear interpolants recovers the dynamic OT solution.

Algorithm

- Sample $(x_0, x_1) \sim \pi_{\text{OT}}$ *(minibatch OT coupling)*
- Sample $t \sim \text{Uniform}[0, 1]$, compute $x_t = (1 - t)x_0 + tx_1$
- Gradient step on $\left\| v_\theta(x_t, t) - (x_1 - x_0) \right\|^2$

Inference: integrate $\mathrm{d}x_t = v_\theta(x_t, t)\, \mathrm{d}t$ from 0 to 1

OT-CFM in Practice

Minibatch OT coupling

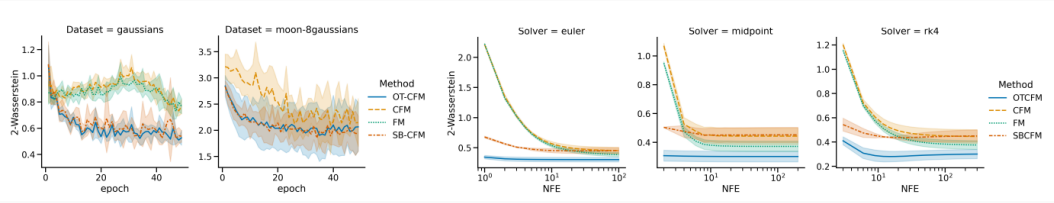
Exact OT between all samples is too expensive. Instead, at each SGD step:

$$\{x_0^i\}_{i=1}^B \sim p_{\text{noise}}, \qquad \{x_1^j\}_{j=1}^B \sim p_{\text{data}}.$$

Solve a discrete OT problem inside the batch:

$$\pi_B \in \underset{\pi \in \Pi(\hat{p}^B_{\text{noise}}, \hat{p}^B_{\text{data}})}{\arg\min} \sum_{i,j} \pi_{ij} \, \|x_1^j - x_0^i\|^2.$$

Then sample pairs $(x_0, x_1) \sim \pi_B$.



From Deterministic to Stochastic: Adding Entropy

Entropic optimal transport (static)

$$\mathrm{OT}_\sigma(p_0, p_1) := \inf_\pi \Big\{ \int \|x_1 - x_0\|^2 \mathrm{d}\pi + \sigma^2 \mathrm{KL}(\pi \mid p_0 \otimes p_1) \Big\}.$$

Schrödinger Bridge (dynamic)

$$p^{\mathrm{SB}}_{0:T} = \underset{p:\, p_{t=0}=p_{\mathrm{noise}},\, p_{t=T}=p_{\mathrm{data}}}{\arg\min} \mathrm{KL}(p_{0:T} \mid q_{0:T}),$$

where $q_{0:T}$ is a reference diffusion, e.g. $\mathrm{d}x_t = \sigma \mathrm{d}w_t$.

Key connection

The Schrödinger Bridge **is** the solution to the dynamic entropic OT problem. As $\sigma \to 0$, it reduces to classical OT (and to OT-CFM).

DSBM [Shi et al., NeurIPS 2023]: learn the SB between p_{noise} and p_{data} numerically

Recovers **OT-CFM** in the limit $\sigma \to 0$, learning entropy-regularised stochastic transports.

Iterative Markovian Fitting (IMF)

Markovian process ($p_{0:T} \in \mathcal{M}$)

Fully described by forward or backward drift:

$$\mathrm{d}x_t = b_t(x_t)\mathrm{d}t + \sigma \mathrm{d}w_t.$$

$\Rightarrow$ Drift b_t can be **learned by a NN**.

Reciprocal process ($p_{0:T} \in \mathcal{R}(q_{0:T})$)

Conditioned on its endpoints, has the same bridge distributions as $q_{0:T}$:

$p_{|0,T} = q_{|0,T}$.

$\Rightarrow$ **Pins marginals** $p_{\mathrm{noise}}, p_{\mathrm{data}}$ at every step.

$p^{\mathrm{SB}}_{0:T}$ is the unique process that is *both* Markovian and reciprocal. **IMF alternates** between the two projections to approach it:

IMF — Iterative Markovian Fitting [Shi et al., NeurIPS 2023 · Peluchetti, JMLR 2023]

$$p^{2n+1}_{0:T} = \mathrm{proj}_{\mathcal{M}}(p^{2n}_{0:T}) \qquad \text{learn a drift — } \textbf{supervised regression}$$

$$p^{2n+2}_{0:T} = \mathrm{proj}_{\mathcal{R}(q_{0:T})}(p^{2n+1}_{0:T}) \qquad \text{pin endpoints — sample bridges from } q_{0:T}$$

Endpoint marginals $p_{\mathrm{noise}}, p_{\mathrm{data}}$ are **preserved at every iterate**.

Diffusion Schrödinger Bridge Matching

IMF in theory [Gentiloni Silveri et al., NeurIPS, 2025]

Under log-concavity, for $p^{SB}_{0:T}$ the optimal bridge and $p^{(n)}_{0:T}$ the IMF iterates:

$$\mathrm{KL}\Big(p^{(n)}_{0:T} \mid p^{SB}_{0:T}\Big) \leq \gamma^n \,\mathrm{KL}\Big(p^{(0)}_{0:T} \mid p^{SB}_{0:T}\Big), \qquad 0 < \gamma < 1.$$

$\Rightarrow$ IMF **converges exponentially fast**.

DSBM: computing $\mathrm{proj}_{\mathcal{M}}$ in practice [Shi et al.]

Learn forward *and* backward drifts alternately:

$$\mathrm{d}x_t = b^+_t(x_t)\mathrm{d}t + \sigma \mathrm{d}w_t, \qquad \mathrm{d}x_t = b^-_t(x_t)\mathrm{d}t + \sigma \mathrm{d}\bar{w}_t,$$

by minimising $\mathcal{L}_{\mathrm{DSBM}}(\theta) = \mathbb{E}\Big[\|b_\theta(x_t,t) - b^\star_t(x_t)\|^2\Big]$.

Diffusion Schrödinger Bridge Matching

DSBM: computing $\mathrm{proj}_{\mathcal{M}}$ in practice [Shi et al.]

Learn forward *and* backward drifts alternately:

$$\mathrm{d}x_t = b_t^+(x_t)\mathrm{d}t + \sigma \mathrm{d}w_t, \qquad \mathrm{d}x_t = b_t^-(x_t)\mathrm{d}t + \sigma \mathrm{d}\bar{w}_t,$$

by minimising $\mathcal{L}_{\mathrm{DSBM}}(\theta) = \mathbb{E}\Big[\|b_\theta(x_t,t) - b_t^\star(x_t)\|^2\Big]$.

Algorithm (outer loop)

1. Sample $(x_0, x_1) \sim p_{0,T}^{(n)}$; deduce reference bridge $q_{|0,T}$ *(reciprocal step)*
2. Learn b_θ^+ and b_θ^- by regressing $\mathcal{L}_{\mathrm{DSBM}}$ *(Markov step)*
3. Simulate SDE with $b_\theta \;\rightarrow$ new coupling $p_{0:T}^{(n+1)}$; repeat

Results

From Shi et al. *DSBM* paper,

Datasets:	moons	scurve	8gaussians	moons-8gaussians
DSBM-IMF+	0.123 ± 0.014	0.130 ± 0.025	0.276 ± 0.030	0.802 ± 0.172
OT-CFM	0.111 ± 0.005	0.102 ± 0.013	0.253 ± 0.040	0.716 ± 0.187

Table 1: 2D experiments - 2-Wasserstein distance (lower is better)

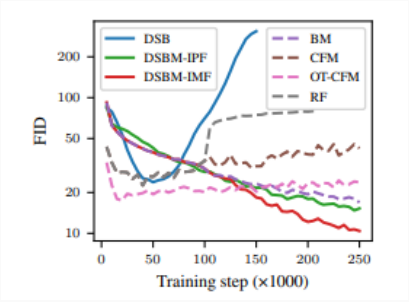
Results

From Shi et al. *DSBM* paper,

Datasets:	moons	scurve	8gaussians	moons-8gaussians
DSBM-IMF+	0.123 ± 0.014	0.130 ± 0.025	0.276 ± 0.030	0.802 ± 0.172
OT-CFM	0.111 ± 0.005	0.102 ± 0.013	0.253 ± 0.040	0.716 ± 0.187

Table 1: 2D experiments - 2-Wasserstein distance (lower is better)

Image generation (MNIST): DSBM-IMF achieves FID (Fréchet inception distance) ≈ 10 at $250k$ steps, outperforming OT-CFM (FID ≈ 20)



Conclusion

OT-CFM [Tong et al.]

$$\mathrm{d}x_t = u_t(x_t)\mathrm{d}t$$

- **(Dynamic) OT** used to provide straight paths between p_{noise} and p_{data}
- Simulation-free; fewer NFE at inference
- **Best W_2 on all 2D benchmarks**

DSBM [Shi et al.]

$$\mathrm{d}x_t = f(x_t)\mathrm{d}t + \sigma \mathrm{d}w_t$$

- **(Entropic) OT** used to learn optimal stochastic dynamic from p_{noise} to p_{data}
- Retrieve OT-CFM for $\sigma \to 0$
- **Better performances on real-world datasets (stochasticity helps)**

Conclusion

OT-CFM [Tong et al.]

$$\mathrm{d}x_t = u_t(x_t)\mathrm{d}t$$

- **(Dynamic) OT** used to provide straight paths between p_{noise} and p_{data}
- Simulation-free; fewer NFE at inference
- **Best W_2 on all 2D benchmarks**

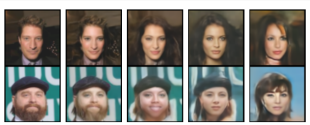
DSBM [Shi et al.]

$$\mathrm{d}x_t = f(x_t)\mathrm{d}t + \sigma \mathrm{d}w_t$$

- **(Entropic) OT** used to learn optimal stochastic dynamic from p_{noise} to p_{data}
- Retrieve OT-CFM for $\sigma \to 0$
- **Better performances on real-world datasets (stochasticity helps)**

[Credits: Alvarez-Melis, Fusi]

Thank you for your attention!



[Credits: Shi et al.]

Full results

	2-Wasserstein (Euler 20 steps)				*Path energy*			
Dataset	moons	scurve	8gaussians	moons-8gaussians	moons	scurve	8gaussians	moons-8gaussians
DSBM-IPF	0.140±0.006	0.140±0.024	0.315±0.079	*0.812±0.092*	1.598±0.034	*2.110±0.059*	14.91±0.310	42.16±1.026
DSBM-IMF	0.144±0.024	0.145±0.037	0.338±0.091	0.838±0.098	**1.580±0.036**	**2.092±0.053**	**14.81±0.255**	**41.00±1.495**
DSBM-IMF+	**0.123±0.014**	**0.130±0.025**	*0.276±0.030*	**0.802±0.172**	*1.594±0.043*	2.116±0.018	*14.88±0.252*	*41.09±1.206*
DSB	0.190±0.049	0.272±0.065	0.411±0.084	0.987±0.324	-	-	-	-
SB-CFM	*0.129±0.024*	*0.136±0.030*	**0.238±0.044**	0.843±0.079	1.649±0.035	2.144±0.044	15.08±0.209	45.69±0.661
FM	0.212±0.025	0.161±0.033	0.351±0.066	-	2.227±0.056	2.950±0.074	18.12±0.416	-
CFM	0.215±0.028	0.171±0.023	0.370±0.049	*1.285±0.314*	2.391±0.043	3.071±0.026	18.00±0.090	116.5±2.633
RF	*0.129±0.022*	*0.126±0.019*	*0.267±0.041*	1.522±0.304	*1.185±0.052*	*1.633±0.074*	**14.84±0.441**	*37.61±3.906*
OT-CFM	**0.111±0.005**	**0.102±0.013**	**0.253±0.040**	**0.716±0.187**	**1.178±0.020**	**1.577±0.036**	*15.10±0.215*	**30.50±0.626**

Table 2: Sampling quality as measured by 2-Wasserstein distance and path energy for the 2D experiments. ±1 SD over 5 seeds. Best values are in bold and second best are italicized.