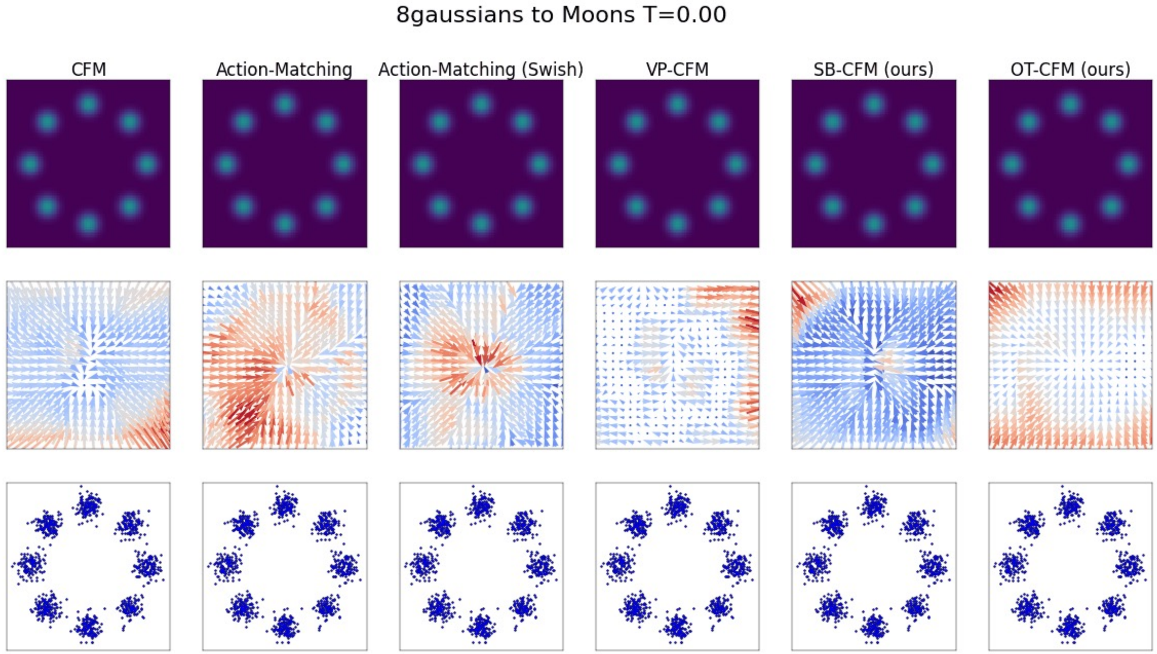
# Deep Generative Models

## 17. OT-CFM


- 국가수리과학연구소 산업수학혁신센터 김민중

# Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport (2023)

- Tong et al.
- Transactions on Machine Learning Research (03/2024)



# Main contribution

- Unifying CFM framework for FM models with arbitrary transport maps
  - This cover CFM, I-CFM, OT-CFM, SB-CFM, UOT-CFM

- Propose a variant of CFM called OT-CFM that approximates dynamic OT via CNFs
  - OT-CFM not only improves the efficiency of training and inference, but also leads to more accurate OT flows than existing neural OT models

# Background: Neural ODE and Optimal transport

- Pair of data distributions (data set) over $\mathbb{R}^d$ with densities $q(\boldsymbol{x}_0)$ and $q(\boldsymbol{x}_1)$ also denote $q_0$ and $q_1$
  - $q_0$: source distribution
  - $q_1$: target distribution

# ODE and Probability flows

- Time dependent vector field $\boldsymbol{u}: [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ defines an ODE:
$$d\boldsymbol{x} = \boldsymbol{u}_t(\boldsymbol{x})dt$$
- $\boldsymbol{u}_t$ called velocity field

- Denote by $\psi_t(\boldsymbol{x})$ the solution of the ODE with initial condition $\psi_0(\boldsymbol{x}) = \boldsymbol{x}$ (called flow)
$$\frac{d\psi_t}{dt}(\boldsymbol{x}) = \boldsymbol{u}_t\big(\psi_t(\boldsymbol{x})\big), \qquad \psi_0(\boldsymbol{x}) = \boldsymbol{x}$$
  - I.e., $\psi_t(\boldsymbol{x})$ is the point $\boldsymbol{x}$ transported along the vector field $\boldsymbol{u}_t$ from time $0$ up to time $t$
  - Equivalence between flow $\psi_t$ and velocity field $\boldsymbol{u}_t$

# ODE and Probability flows

- Probability density $p_0$ over $\mathbb{R}^d$
- Flow $\psi_t$ induces a pushforward

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

- which is the density of points $\boldsymbol{x}_0 \sim p_0$ transported along $\boldsymbol{u}_t$ from time 0 to time $t$

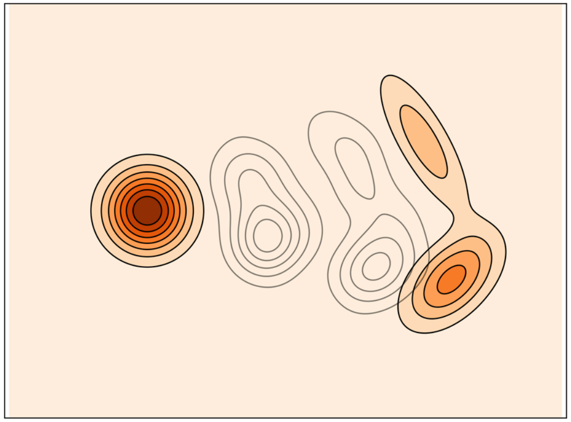
- Time-varying density $p_t$ (probability path) viewed as a function $p: [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ is characterized by continuity equation

$$\frac{\partial p_t}{\partial t} = -\nabla \cdot (p_t \boldsymbol{u}_t)$$

- with initial condition $p_0$

# Goal

$q_0(\boldsymbol{x})$
Source



$q_1(\boldsymbol{x})$
Target

- Construct $\boldsymbol{u}_t$ (or $\psi_t$) such that the resulting probability path $p_t$ governed by the continuity equation:
  - $p_0 \approx q_0$ at time $t = 0$ (source distribution)
  - $p_1 \approx q_1$ at time $t = 1$ (target distribution)

# Approximating ODE with neural networks

- Suppose that $p_t(\boldsymbol{x})$ and the vector field $\boldsymbol{u}_t(\boldsymbol{x})$ which generates $p_t(\boldsymbol{x})$ are known and $p_t(\boldsymbol{x})$ can be tractably sampled

- Let $\boldsymbol{v}_\theta: [0,1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a time-dependent vector filed parametrized by $\theta$
- $\boldsymbol{v}_\theta$ can be regressed to $\boldsymbol{u}$ via the FM loss:

$$\mathcal{L}_{FM}(\theta) \coloneqq E_{t \sim U[0,1], \boldsymbol{x} \sim p_t(\boldsymbol{x})}[\|\boldsymbol{v}_\theta(t, \boldsymbol{x}) - \boldsymbol{u}_t(\boldsymbol{x})\|^2]$$

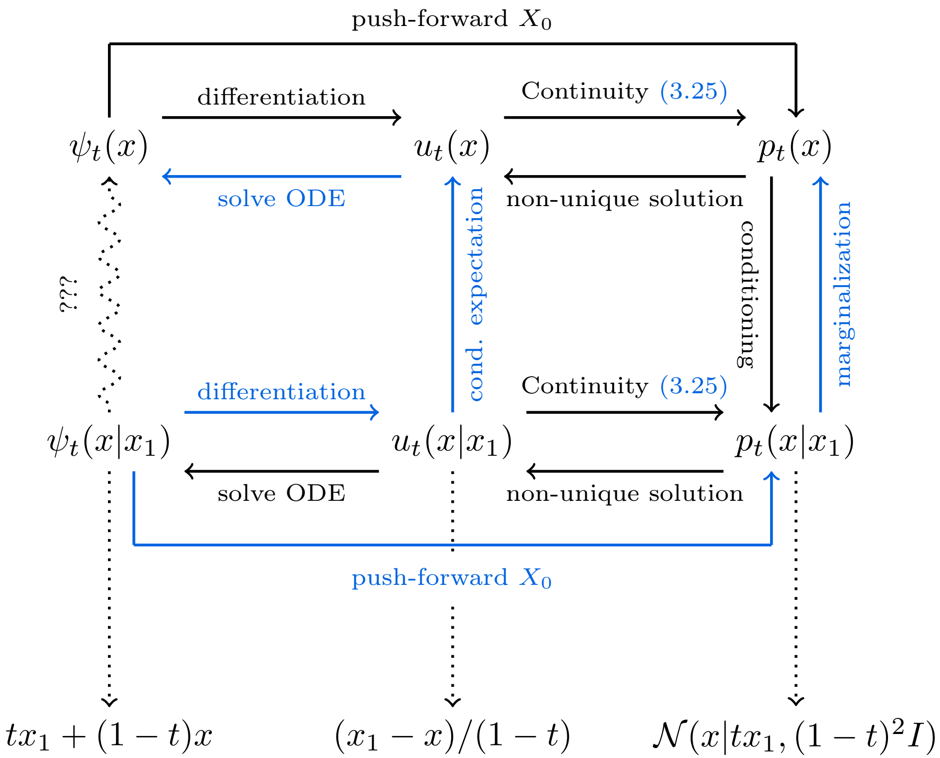
- This objective becomes intractable for general source and target distribution

# The case of Gaussian marginals

- Isotropic $d$-dimensional Gaussian marginal path
$$p_t(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{\mu}_t, \sigma_t^2 \boldsymbol{I})$$
- The flow $\psi_t$ that generates the above Gaussian marginal path is not unique. One simplest flow is
$$\psi_t(x_0) = \boldsymbol{\mu}_t + \frac{\sigma_t}{\sigma_0}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0) \qquad (*)$$
- The unique vector field whose integration map satisfying $(*)$ has the form
$$\boldsymbol{u}_t(\boldsymbol{x}) = \frac{\sigma_t'}{\sigma_t}(\boldsymbol{x} - \boldsymbol{\mu}_t) + \boldsymbol{\mu}_t'$$
- where $\sigma_t'$ and $\boldsymbol{\mu}_t'$ denote the time derivative
- The vector field $\boldsymbol{u}_t$ with initial conditions $N(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I})$ generates $p_t(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{\mu}_t, \sigma_t^2 \boldsymbol{I})$

# Recap: Relation for FM

Flow | Velocity field | Probability path | Boundary conds. | Loss




$p_0 = p$
$p_1 = q$

**Flow Matching (FM)** (4.22)
$D\left(u_t(X_t), u_t^\theta(X_t)\right)$

$p_0 = p$
$p_1 = \delta_{x_1}$

**Conditional FM (CFM)** (4.23)
$D\left(u_t(X_t|X_1), u_t^\theta(X_t)\right)$

$p_0 = \mathcal{N}(0, I)$
$p_1 = \delta_{x_1}$

**OT, Gauss CFM** (2.9)
$\|u_t^\theta(X_t) - (X_1 - X_0)\|^2$

# Static optimal transport

- 2-Wasserstein distance (static OT) between densities $q_0$ and $q_1$ over $\mathbb{R}^d$ w.r.t. Euclidean distance cost $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$

$$W_2(q_0, q_1)^2 \coloneqq \inf_{\pi \in \Pi} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(\boldsymbol{x}, \boldsymbol{y})^2 \, d\pi(\boldsymbol{x}, \boldsymbol{y})$$

- where $\Pi$ denotes the set of all joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are $q_0$ and $q_1$

# Dynamic optimal transport

- Dynamic form of 2-Wasserstein distance is defined by an optimization problem over vector fields $\boldsymbol{u}_t$

$$W_2(q_0, q_1)^2 = \inf_{p_t, \boldsymbol{u}_t} \int_{\mathbb{R}^d} \int_0^1 p_t(\boldsymbol{x}) \|\boldsymbol{u}_t(\boldsymbol{x})\|^2 dt \, d\boldsymbol{x}$$

- with $p_t \geq 0$ and subject to the boundary conditions $p_0 = q_0$, $p_1 = q_1$ and

$$\frac{\partial p_t}{\partial t} = -\nabla \cdot (p_t \boldsymbol{u}_t)$$

- Authors showed that when the true OT plan is available, OT-CFM method approximates dynamic OT

# CFM: Vector fields generating conditional probability paths

- Let $\boldsymbol{z}$ be a conditioning (latent) variable and

$$p_t(\boldsymbol{x}) = \int p_t(\boldsymbol{x}|\boldsymbol{z})q(\boldsymbol{z})d\boldsymbol{z}$$

- where $q(\boldsymbol{z})$ is some distribution over the conditioning variable
- If $p_t(\boldsymbol{x}|\boldsymbol{z})$ is generated by $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})$ from $p_0(\boldsymbol{x}|\boldsymbol{z})$, then

$$\boldsymbol{u}_t(\boldsymbol{x}) \coloneqq E_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\frac{\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})p_t(\boldsymbol{x}|\boldsymbol{z})}{p_{t(\boldsymbol{x})}}\right]$$

- generates the probability path $p_t(\boldsymbol{x})$ under some mild conditions

# CFM: A regression objective for mixtures

- **Given**: conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{z})$ and conditional vector fields $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})$ are known (our design) and simple

- **Goal**: recover the unconditional vector field $\boldsymbol{u}_t(\boldsymbol{x})$ generating the marginal path $p_t(\boldsymbol{x})$

- Exact computation of $\boldsymbol{u}_t(\boldsymbol{x})$ is intractable because $p_t(\boldsymbol{x})$ is difficult to evaluate

# CFM loss

- Let $\boldsymbol{v}_\theta: [0,1]\times\mathbb{R}^d \rightarrow \mathbb{R}^d$ be a time-dependent vector filed parametrized by $\theta$
- Conditional flow matching loss

$$\mathcal{L}_{CFM}(\theta) \coloneqq E_{t\sim U[0,1],\boldsymbol{z}\sim q(\boldsymbol{z}),\boldsymbol{x}\sim p_t(\boldsymbol{x}|\boldsymbol{z})}[||\boldsymbol{v}_\theta(t,\boldsymbol{x}) - \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})\|^2]$$

- CFM loss aims to regress to the marginal vector field $\boldsymbol{u}_t(\boldsymbol{x})$ using
  - samples from the conditional path $p_t(\boldsymbol{x}|\boldsymbol{z})$ and
  - conditional vector fields $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})$

- No direct access to $\boldsymbol{u}_t(\boldsymbol{x})$; estimate it indirectly via regression

# CFM

- **Theorem** If $p_t(\boldsymbol{x}) > 0 \ \forall \boldsymbol{x} \in \mathbb{R}^d$ and $t \in [0,1]$, then
$$\nabla_\theta \mathcal{L}_{FM}(\theta) = \nabla_\theta \mathcal{L}_{CFM}(\theta)$$

# CFM Algorithm

- **Goal**: recover the unconditional vector field $\boldsymbol{u}_t(\boldsymbol{x})$ generating the marginal path $p_t(\boldsymbol{x})$
  $$\mathcal{L}_{CFM}(\theta) \coloneqq E_{t\sim U[0,1],\boldsymbol{z}\sim q(\boldsymbol{z}),\boldsymbol{x}\sim p_t(\boldsymbol{x}|\boldsymbol{z})}[\|\boldsymbol{v}_\theta(t,\boldsymbol{x}) - \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})\|^2]$$
- Requirements:
  - efficiently sample from $q(\boldsymbol{z})$ and $p_t(\boldsymbol{x}|\boldsymbol{z})$
  - efficiently compute $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})$
- Then we can use stochastic CFM objective to regress $\boldsymbol{v}_\theta(t,\boldsymbol{x}) \approx \boldsymbol{u}_t(\boldsymbol{x})$

**Algorithm 1** Conditional Flow Matching

**Input:** Efficiently samplable $q(z)$, $p_t(x|z)$, and computable $u_t(x|z)$ and initial network $v_\theta$.
**while** Training **do**
  $z \sim q(z)$;  $t \sim \mathcal{U}(0,1)$;  $x \sim p_t(x|z)$
  $\mathcal{L}_{\mathrm{CFM}}(\theta) \leftarrow \|v_\theta(t,x) - u_t(x|z)\|^2$
  $\theta \leftarrow \mathrm{Update}(\theta, \nabla_\theta \mathcal{L}_{\mathrm{CFM}}(\theta))$
**return** $v_\theta$

# FM from the Conditional Gaussian path

- Several forms of CFM depending on the choices of $q(\boldsymbol{z})$ and $p_t(\cdot\,|\boldsymbol{z})$ and $\boldsymbol{u}_t(\cdot\,|\boldsymbol{z})$ with

$$p_t(\boldsymbol{x}|\boldsymbol{z}) = N(\boldsymbol{x}|\boldsymbol{\mu}_t(\boldsymbol{z}), \sigma_t(\boldsymbol{z})^2\boldsymbol{I})$$

- Which is generated by $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})$

# CFM: Algorithm overview

1. Define source and target distributions ($q_0$ and $q_1$)
2. Choose the latent distribution $q(\boldsymbol{z})$
3. Define conditional path $p_t(\cdot\,|\boldsymbol{z})$ with
$$p_t(\boldsymbol{x}|\boldsymbol{z}) = N(\boldsymbol{x}|\boldsymbol{\mu}_t(\boldsymbol{z}), \sigma_t(\boldsymbol{z})^2\boldsymbol{I})$$
4. Determine conditional velocity field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})$ which generates $p_t(\boldsymbol{x}|\boldsymbol{z})$

# Variance exploding (Song & Ermon 2019)

$$q(\boldsymbol{z}) = q(\boldsymbol{x}_1) \ \text{target distribution}$$
$$p_t(\boldsymbol{x}|\boldsymbol{x}_1) = N(\boldsymbol{x}|\boldsymbol{x}_1, \sigma_t^2\boldsymbol{I})$$
$$\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{x}_1) = -\frac{{\sigma_t}'}{\sigma_t}(\boldsymbol{x} - \boldsymbol{x}_1)$$

| Noise space | | Data space |
|---|---|---|
| $\boldsymbol{x}_0 \sim N(\boldsymbol{0}, \sigma_0 \boldsymbol{I})$ | | $\boldsymbol{x}_1 \sim q_1(\text{or } p_{data})$ |
| $p_0 \approx N(0, \sigma_0 I)$ | $p_t$ | $p_1 \approx q_1(\text{or } p_{data})$ |

- $\sigma_t^2$ is decreasing function of $t$ with sufficiently large $\sigma_0$ (exploding) and small $\sigma_1$

# Variance preserving (Ho et al. 2020)

$$q(\boldsymbol{z}) = q(\boldsymbol{x}_1) \ \text{target distribution}$$
$$p_t(\boldsymbol{x}|\boldsymbol{x}_1) = N(\boldsymbol{x}|\alpha_t \boldsymbol{x}_1, (1-\alpha_t)\boldsymbol{I})$$

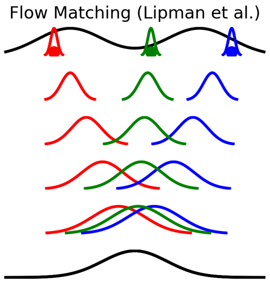
| Noise space | | Data space |
|---|---|---|
| $\boldsymbol{x}_0 \sim N(\boldsymbol{0}, \boldsymbol{I})$ | | $\boldsymbol{x}_1 \sim q_1 (\text{or } p_{data})$ |
| $p_0 \approx N(0, I)$ | $p_t$ | $p_1 \approx q_1 (\text{or } p_{data})$ |

- $\alpha_t$ is increasing function of $t$ with $\alpha_0 = 0$ and $\alpha_1 = 1$

# Flow matching (Lipman et al. 2023)

$$q(\boldsymbol{z}) = q(\boldsymbol{x}_1), \qquad q(\boldsymbol{x}_0) = N(\boldsymbol{x}_0|\boldsymbol{0}, \boldsymbol{I})$$
$$p_t(\boldsymbol{x}|\boldsymbol{x}_1) = N(\boldsymbol{x}|t\boldsymbol{x}_1, (t\sigma - t + 1)^2\boldsymbol{I})$$
$$\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{x}_1) = \frac{1}{1-(1-\sigma)t}(\boldsymbol{x}_1 - (1-\sigma)\boldsymbol{x})$$

- where smoothing constant $\sigma > 0$
- $p_t(\boldsymbol{x}|\boldsymbol{x}_1)$ is a conditional probability path from the standard normal distribution $p_0(\boldsymbol{x}|\boldsymbol{z}) = N(\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{I})$ to a Gaussian distribution centered at $\boldsymbol{x}_1$ with small variance $\sigma$

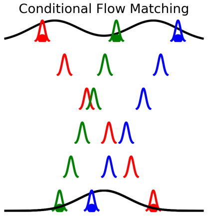


# Independent-CFM

- Let $\boldsymbol{z}$ be a pair of random variables, a source point $\boldsymbol{x}_0 \sim q_0$ and a target point $\boldsymbol{x}_0 \sim q_1$
- Set $q(\boldsymbol{z}) = q(\boldsymbol{x}_0)q(\boldsymbol{x}_1)$ to be the independent coupling and
$$p_t(\boldsymbol{x}|\boldsymbol{z}) = N(\boldsymbol{x}|t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0, \sigma^2 \boldsymbol{I})$$
$$\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}) = (\boldsymbol{x}_1 - \boldsymbol{x}_0)$$
- where smoothing constant $\sigma > 0$
- Note that $q(\boldsymbol{z})$ and $p_t(\boldsymbol{x}|\boldsymbol{z})$ is efficiently sampleable and $\boldsymbol{u}_t$ is efficiently computable, thus gradient descent on $\mathcal{L}_{CFM}$ is also efficient
- As $\sigma \to 0$, the marginal vector field $\boldsymbol{u}_t$ approaches one that transports the distribution $q(\boldsymbol{x}_0)$ to $q(\boldsymbol{x}_1)$

# Remark of I-CFM

- No requirement for $q(\boldsymbol{x}_0)$ to be Gaussian
- The conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{z})$ is an optimal transport path from $p_0(\boldsymbol{x}|\boldsymbol{z})$ to $p_1(\boldsymbol{x}|\boldsymbol{z})$
- However, the marginal path $p_t$ is "not" in general an OT path from $p_0(\boldsymbol{x})$ to $p_1(\boldsymbol{x})$




**Algorithm 2** Simplified Conditional Flow Matching (I-CFM)

**Input:** Empirical or samplable distributions $q_0, q_1$, bandwidth $\sigma$, batchsize $b$, initial network $v_\theta$.
**while** Training **do**
  /* *Sample batches of size b i.i.d. from the datasets*
  $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0); \quad \boldsymbol{x}_1 \sim q_1(\boldsymbol{x}_1)$
  $t \sim \mathcal{U}(0,1)$
  $\mu_t \leftarrow t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0$
  $x \sim \mathcal{N}(\mu_t, \sigma^2 I)$
  $\mathcal{L}_{\text{CFM}}(\theta) \leftarrow \|v_\theta(t,x) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|^2$
  $\theta \leftarrow \text{Update}(\theta, \nabla_\theta \mathcal{L}_{\text{CFM}}(\theta))$
**return** $v_\theta$

# OT-CFM

- Let $\boldsymbol{z}$ be a pair of random variables, a source point $\boldsymbol{x}_0 \sim q_0$ and a target point $\boldsymbol{x}_0 \sim q_1$
- General formulation: previous formulation extends to joint distributions

$$q(\boldsymbol{z}) = q(\boldsymbol{x}_0, \boldsymbol{x}_1)$$

- Allows $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ to be **dependent** and $q(\boldsymbol{x}_0, \boldsymbol{x}_1)$ has marginals $q(\boldsymbol{x}_0)$ and $q(\boldsymbol{x}_1)$
- Set $q(\boldsymbol{z}) = q(\boldsymbol{x}_0, \boldsymbol{x}_1)$to be the 2-Wasserstein optimal transport map $\pi$. I.e.,

$$q(\boldsymbol{z}) = \pi(\boldsymbol{x}_0, \boldsymbol{x}_1)$$

- This method is called OT-CFM

# OT-CFM: Dynamic OT

- Let $q(\boldsymbol{z}) = \pi(\boldsymbol{x}_0, \boldsymbol{x}_1)$
- If

$$p_t(\boldsymbol{x}|\boldsymbol{z}) = N(\boldsymbol{x}|t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0, \sigma^2\boldsymbol{I})$$
$$\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}) = (\boldsymbol{x}_1 - \boldsymbol{x}_0),$$

- then OT-CFM is equivalent to dynamic OT in the following sense

# OT-CFM: Dynamic OT

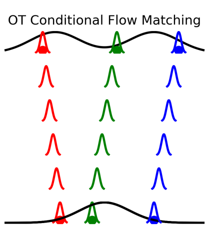
- **Proposition** Under regularity properties of $q_0$, $q_1$ and OT plan $\pi$, as $\sigma^2 \to 0$ the marginal path $p_t$ and vector field $\boldsymbol{u}_t$ minimize the dynamic form of the 2-Wasserstein distance

$$W_2(q_0, q_1)^2 = \inf_{p_t, \boldsymbol{u}_t} \int_{\mathbb{R}^d} \int_0^1 p_t(\boldsymbol{x}) \|\boldsymbol{u}_t(\boldsymbol{x})\|^2 dt\, d\boldsymbol{x}$$

- I.e., $\boldsymbol{u}_t$ solves the dynamic OT between $q_0$ and $q_1$

# Minibatch OT approximation

- The transport plan $\pi$ is difficult to compute and store due to OT's cubic time and quadratic memory complexity in the number of samples
- Therefore, we rely on a minibatch OT approximation
- For each batch of data $\left(\left\{\boldsymbol{x}_0^{(i)}\right\}_{i=1}^{B}, \left\{\boldsymbol{x}_1^{(i)}\right\}_{i=1}^{B}\right)$ seen during training, we sample pairs of points from the joint distribution $\pi_{batch}$




**Algorithm 3** Minibatch OT Conditional Flow Matching (OT-CFM)

**Input:** Empirical or samplable distributions $q_0, q_1$, bandwidth $\sigma$, batch size $b$, initial network $v_\theta$.
**while** Training **do**
  /* *Sample batches of size b i.i.d. from the datasets*
  $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0); \quad \boldsymbol{x}_1 \sim q_1(\boldsymbol{x}_1)$
  $\pi \leftarrow \mathrm{OT}(\boldsymbol{x}_1, \boldsymbol{x}_0)$
  $(\boldsymbol{x}_0, \boldsymbol{x}_1) \sim \pi$
  $\boldsymbol{t} \sim \mathcal{U}(0,1)$
  $\mu_t \leftarrow \boldsymbol{t}\boldsymbol{x}_1 + (1-\boldsymbol{t})\boldsymbol{x}_0$
  $\boldsymbol{x} \sim \mathcal{N}(\mu_t, \sigma^2 I)$
  $\mathcal{L}_{\mathrm{CFM}}(\theta) \leftarrow \|v_\theta(\boldsymbol{t}, \boldsymbol{x}) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|^2$
  $\theta \leftarrow \mathrm{Update}(\theta, \nabla_\theta \mathcal{L}_{\mathrm{CFM}}(\theta))$
**return** $v_\theta$

# Schrödinger bridge CFM (2021 ICML)

- Conditional flow matching method that learns vector fields to interpolate between two distributions via an entropic optimal transport path
- Let $p_{ref}: [0,1]\times\mathbb{R}^d \to \mathbb{R}$ be the time dependent probability path as the standard Wiener process scaled by $\sigma$ with initial-time marginal $p_{ref}(\boldsymbol{x}_0) = q(\boldsymbol{x}_0)$
- The SB problem (Schrödinger 1932) seeks the process $\pi$ that is the closest to $p_{ref}$ s.t. its initial and terminal marginal distributions are $q(\boldsymbol{x}_0)$ and $q(\boldsymbol{x}_1)$ resp. I.e.,

$$\pi^* = \underset{\substack{\pi(\boldsymbol{x}_0)=q(\boldsymbol{x}_0)\\ \pi(\boldsymbol{x}_1)=q(\boldsymbol{x}_1)}}{\operatorname{argmin}} KL\left(\pi \parallel p_{ref}\right)$$

# Schrödinger bridge CFM (2021 ICML)

- Define the joint distribution

$$q(z) = \pi_{2\sigma^2}(\boldsymbol{x}_0, \boldsymbol{x}_1)$$

- where $\pi_{2\sigma^2}$ is the solution of the entropy regularized OT with cost $c(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||$ and entropy regularization $\lambda = 2\sigma^2$

$$W_{2,\lambda}(q_0, q_1)^2 \coloneqq \inf_{\pi_\lambda \in \Pi} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(\boldsymbol{x}, \boldsymbol{y})^2 \, d\pi_\lambda(\boldsymbol{x}, \boldsymbol{y}) - \lambda H(\pi_\lambda)$$

- where $\Pi$ denotes the set of all joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are $q_0$ and $q_1$

# Schrödinger bridge CFM (2021 ICML)

- Set the conditional path distribution to be a Brownian bridge with diffusion scale $\sigma$ between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ with
$$p_t(\boldsymbol{x}|\boldsymbol{z}) = N(\boldsymbol{x}|t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0, t(1-t)\sigma^2\boldsymbol{I})$$
$$\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}) = \frac{1-2t}{2t(1-t)}\big(\boldsymbol{x} - (t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0)\big) + (\boldsymbol{x}_1 - \boldsymbol{x}_0)$$
- where conditional vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})$ generates the probability path $p_t(\boldsymbol{x}|\boldsymbol{z})$
- The solution of the SB is known to be the map which is the solution of the entropically-regularized OT problem
- We recover OT-CFM when $\lambda = 2\sigma^2 \to 0$ and I-CFM when $\lambda \to \infty$

# Schrödinger bridge CFM (2021 ICML)

**Algorithm 4** Minibatch Schrödinger Bridge Conditional Flow Matching (SB-CFM)

**Input:** Empirical or samplable distributions $q_0, q_1$, bandwidth $\sigma$, batch size $b$, initial network $v_\theta$.
**while** Training **do**
  /* *Sample batches of size b i.i.d. from the datasets*
  $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0); \quad \boldsymbol{x}_1 \sim q_1(\boldsymbol{x}_1)$
  $\pi_{2\sigma^2} \leftarrow \mathrm{Sinkhorn}(\boldsymbol{x}_1, \boldsymbol{x}_0, 2\sigma^2)$
  $(\boldsymbol{x}_0, \boldsymbol{x}_1) \sim \pi_{2\sigma^2}$
  $\boldsymbol{t} \sim \mathcal{U}(0,1)$
  $\mu_t \leftarrow \boldsymbol{t}\boldsymbol{x}_1 + (1-\boldsymbol{t})\boldsymbol{x}_0$
  $\boldsymbol{x} \sim \mathcal{N}(\mu_t, \sigma^2 \boldsymbol{t}(1-\boldsymbol{t})I)$
  $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}) \leftarrow \frac{1-2\boldsymbol{t}}{2\boldsymbol{t}(1-\boldsymbol{t})}(\boldsymbol{x} - (\boldsymbol{t}x_1 + (1-\boldsymbol{t})\boldsymbol{x}_0)) + (\boldsymbol{x}_1 - \boldsymbol{x}_0)$
  $\mathcal{L}_{\mathrm{CFM}}(\theta) \leftarrow \|v_\theta(\boldsymbol{t}, \boldsymbol{x}) - \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})\|^2$
  $\theta \leftarrow \mathrm{Update}(\theta, \nabla_\theta \mathcal{L}_{\mathrm{CFM}}(\theta))$
**return** $v_\theta$

# Experiments

Table 2: Comparison of neural optimal transport methods over four distribution pairs ($\mu \pm \sigma$ over five seeds) in terms of fit (2-Wasserstein), optimal transport performance (normalized path energy), and runtime. '—' indicates a method that requires a Gaussian source. Best in **bold**. CFM and RF models are trained on a single CPU core, other baselines are trained with a GPU and two CPUs.

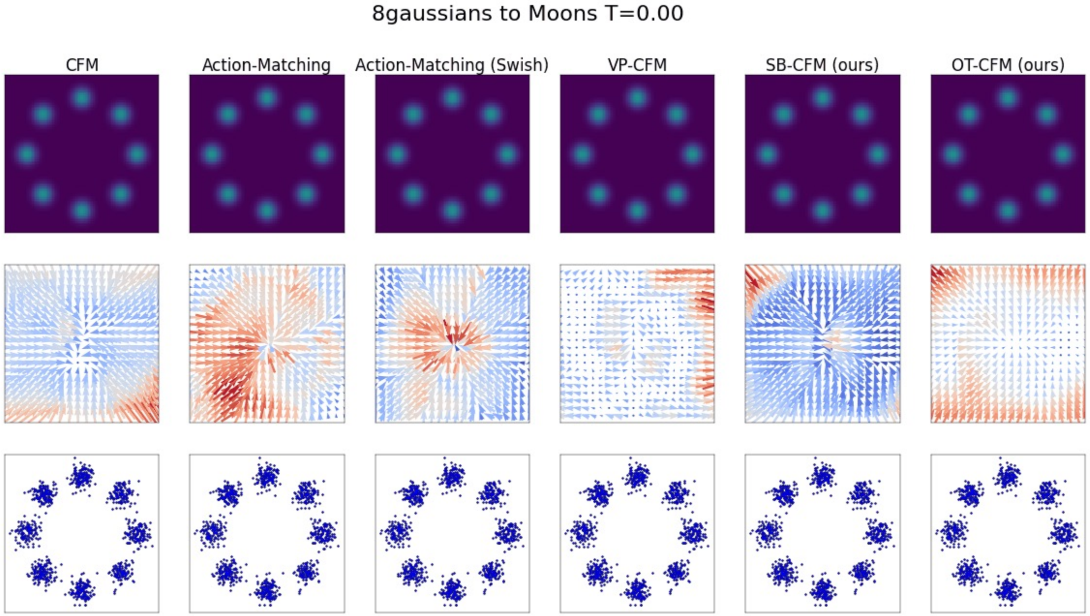
| Dataset → | $\mathcal{N}$→8gaussians | | moons→8gaussians | | $\mathcal{N}$→moons | | $\mathcal{N}$→scurve | | Avg. train time |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm ↓ Metric → | $W_2^2$ | NPE | $W_2^2$ | NPE | $W_2^2$ | NPE | $W_2^2$ | NPE | ($\times 10^3$ s) |
| OT-CFM | 1.262±0.348 | **0.018**±0.014 | **1.923**±0.391 | **0.053**±0.035 | **0.239**±0.048 | 0.087±0.061 | **0.264**±0.093 | **0.027**±0.026 | 1.129±0.335 |
| I-CFM | 1.284±0.384 | 0.222±0.032 | 1.977±0.266 | 2.738±0.181 | 0.338±0.109 | 0.841±0.148 | 0.333±0.060 | 0.867±0.117 | **0.630**±0.365 |
| 2-RF (Liu, 2022) | 1.436±0.344 | 0.069±0.027 | 2.211±0.423 | 0.149±0.101 | 0.278±0.026 | **0.076**±0.067 | 0.395±0.111 | 0.112±0.085 | 0.862±0.166 |
| 3-RF (Liu, 2022) | 1.337±0.367 | 0.055±0.043 | 2.700±0.587 | 0.123±0.112 | 0.305±0.026 | 0.084±0.051 | 0.395±0.082 | 0.129±0.075 | 0.954±0.116 |
| FM (Lipman et al., 2023) | 1.062±0.196 | 0.174±0.030 | — | — | 0.246±0.077 | 0.778±0.144 | 0.377±0.099 | 0.772±0.081 | 0.708±0.370 |
| Reg. CNF (Finlay et al., 2020) | 1.144±0.075 | 0.274±0.060 | — | — | 0.376±0.040 | 0.620±0.088 | 0.581±0.195 | 0.586±0.503 | 8.021±3.288 |
| CNF (Chen et al., 2018) | **1.055**±0.059 | 0.151±0.064 | — | — | 0.387±0.065 | 2.937±1.973 | 0.645±0.343 | 10.548±8.100 | 18.810±12.677 |
| ICNN (Makkuva et al., 2020) | 1.771±0.398 | 0.747±0.029 | 2.193±0.136 | 0.832±0.004 | 0.532±0.046 | 0.267±0.010 | 0.753±0.068 | 0.344±0.045 | 2.912±0.626 |

- Normalized path energy

$$\int_{\mathbb{R}^d} \int_0^1 p_t(\boldsymbol{x}) \|\boldsymbol{u}_t(\boldsymbol{x})\|^2 dt\, d\boldsymbol{x} \approx \frac{1}{T} \sum_{i=1}^{T} \frac{1}{N} \sum_{j=1}^{N} \left\| \boldsymbol{v}_{\boldsymbol{\theta}}\left(t_i, \boldsymbol{x}_{t_i}^{(j)}\right) \right\|^2$$

Where $t_1 = 0$, $t_T = 1$, $\boldsymbol{x}_0^{(j)} \sim q_0$ and $\boldsymbol{x}_{t_{i+1}}^{(j)} = \boldsymbol{x}_{t_i}^{(j)} + \boldsymbol{v}_{\boldsymbol{\theta}}\left(t_i, \boldsymbol{x}_{t_i}^{(j)}\right)$

# Experiments



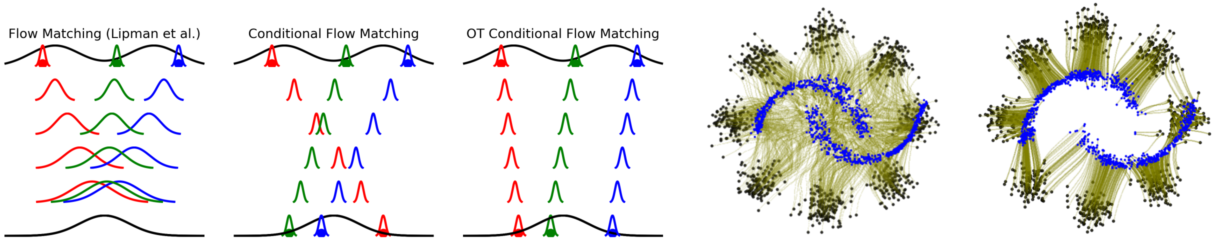




Figure 1: **Left:** Conditional flows from FM (Lipman et al., 2023), I-CFM (§3.2.2), and OT-CFM (§3.2.3). **Right:** Learned flows (green) from moons (blue) to 8gaussians (black) using I-CFM (centre-right) and OT-CFM (far right).

# Thanks