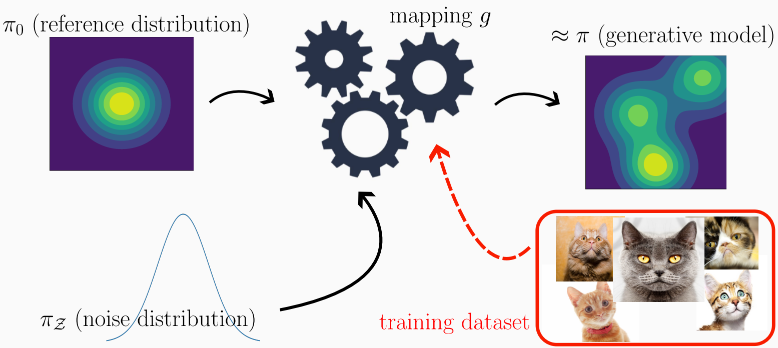
Introduction to diffusion models

Valentin De Bortoli
joint work with: James Thornton, Jeremy Heng, Yuyang Shi, Andrew Campbell, Arnaud Doucet

June 13, 2023

What is generative modeling?

- **Generative modeling**: Given a **dataset** of samples from a distribution π how to obtain **new samples** from π?
- **A general approach**:
  - Sample X_0 from π_0 (reference distribution).
  - Sample Z from $\pi_{\mathcal{Z}}$ (noise distribution).
  - Push with $g(X_0, Z) \to$ approximate sample from π.



Why generative modeling?

- Application in **computational biology**: Senior et al. (2020).
  - **Amino-acid sequence** to **3D structure**.
  - Cryo-Electron Microscopy or crystallography = experimental techniques to determine the shape of the protein.
  - Crystallizing a protein is a real challenge Avanzato et al. (2019).
  - Competition to predict structure: **C**ritical **A**ssessment of protein **S**tructure **P**rediction.
- **Conditional generative modeling**.

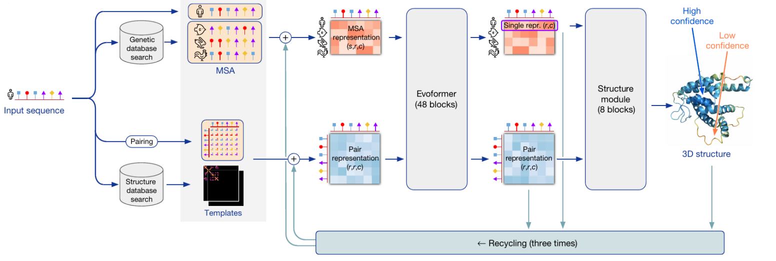
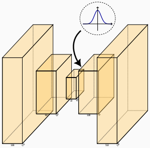


Image extracted from Senior et al. (2020).

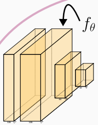
A myriad of models

Variational AutoEncoder



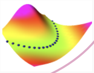
Kingma et al. (2014)
Rezende et al. (2014)
Ranganath et al. (2016)
Vahdat et al. (2021)

Energy-Based Model



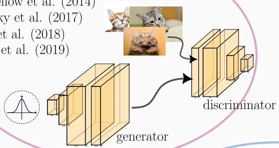
Zhu et al. (1998)
LeCun et al. (2006)
Hinton et al. (2006)
Du et al. (2019)

$$\frac{\exp[-f_\theta(x)]}{\int \exp[-f_\theta(\tilde{x})]\mathrm{d}\tilde{x}}$$



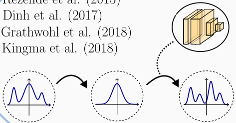
Generative Adversarial Network

Goodfellow et al. (2014)
Arjovsky et al. (2017)
Brock et al. (2018)
Karras et al. (2019)



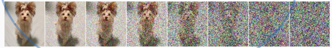
Normalizing Flow

Rezende et al. (2015)
Dinh et al. (2017)
Grathwohl et al. (2018)
Kingma et al. (2018)



Denoising Diffusion Model

Song et al. (2019)
Ho et al. (2020)
Vahdat et al. (2021)



Beyond generative modeling: transfer tasks (1/3)

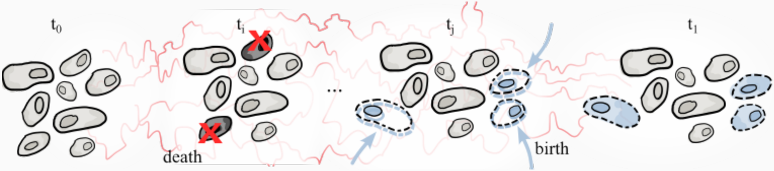
- In **generative modeling**:
  - **initial distribution** is Gaussian $N(0, \mathrm{Id})$,
  - **target distribution** is a data distribution.
- In **unpaired transfer tasks**:
  - **initial distribution** is a data distribution.
  - **target distribution** is *another* data distribution.
  - Not necessary **paired** training examples.
- Different **goals**:
  - **generative modeling**: quality of generated samples.
  - **transfer task**: quality of samples and properties of the **coupling**.

Style transfer. Image extracted from Su et al. (2022).

Beyond generative modeling: transfer tasks (2/3)

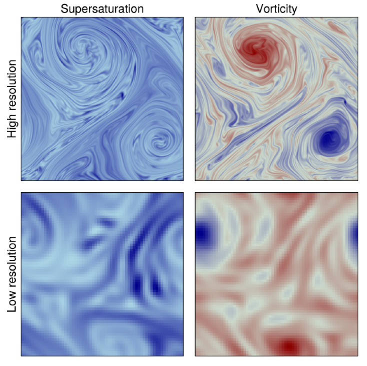
- Application in **biology**:
  - **Tracking** cell population (treatment effect).
  - Cannot track **individual** particles (internal/external influences).
  - Observation at different **discrete** times.

- Goal: reconstruction of the dynamics (**Optimal transport** based)
  - JKO-NET Bunne et al. (2022).
  - Conditional flow matching Tong et al. (2023).



Beyond generative modeling: transfer tasks (3/3)

- Application in **climate science**:
  - **Downscaling**: high resolution data from low resolution ones.
  - This is a **super resolution** task.
  - No **paired datasets** of high and low resolutions exist.



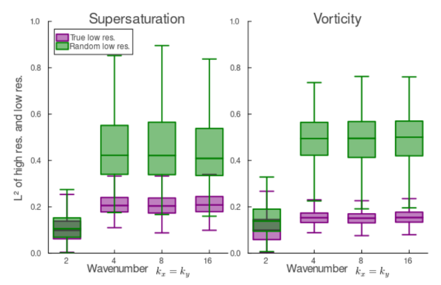


Image extracted from Bischoff and Deck (2023).

Generative Modeling: the rise of diffusion models

Time-reversal of diffusions

- **Forward decomposition**: $p(x_{0:N}) = p_0(x_0) \prod_{k=0}^{N-1} p_{k+1|k}(x_{k+1}|x_k)$.

- **Backward decomposition**: $p(x_{0:N}) = p_N(x_N) \prod_{k=0}^{N-1} p_{k|k+1}(x_k|x_{k+1})$.

Video extracted from Song and Ermon (2019).

Approximate time reversal

¿How to approximate the backward decomposition?

- **Backward decomposition**: $p(x_{0:N}) = p_N(x_N) \prod_{k=0}^{N-1} p_{k|k+1}(x_k|x_{k+1})$.
  - How to compute $p_{k|k+1}(x_k|x_{k+1}) = p_{k+1|k}(x_{k+1}|x_k)p_k(x_k)/p_{k+1}(x_{k+1})$?
  - In practice $p_{k+1|k} = \mathrm{N}(x_k - \gamma x_k, \sqrt{2\gamma}\mathrm{Id})$ is **Gaussian**.
  - (**Discretization** of $\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t$ (**Ornstein-Ulhenbeck**))
  - $p_{k|k+1}$ is approximately Gaussian

$$\boxed{p_{k|k+1} = \mathrm{N}(x_{k+1} + \gamma x_{k+1} + 2\gamma \nabla \log p_{k+1}(x_{k+1}), \sqrt{2\gamma}\,\mathrm{Id}).}$$

¿How to compute the score term?

- **Score matching** techniques: Vincent (2011); Hyvärinen (2005)

$$\nabla \log p_{k+1}(x_{k+1}) = \mathbb{E}_{p_{0|k+1}}[\nabla \log p_{k+1|0}(x_{k+1}|X_0)].$$

  - **Loss function**: $\boxed{\ell(\mathbf{s}_{k+1}) = \mathbb{E}[\|\mathbf{s}_{k+1}(X_{k+1}) - \nabla \log p_{k+1|0}(X_{k+1}|X_0)\|^2].}$
  - Algorithm: replace $\nabla \log p_{k+1}$ by $\mathbf{s}_{k+1}$.

An application: text-to-image

- Text-to-image: Imagen, DALL-E 2, Stable Diffusion, Midjourney, EDiff.

An extremely angry bird. A cute corgi lives in a house made out of sushi.

- **CLIP** (**C**ontrastive **L**anguage–**I**mage **P**re-training) guidance.

From Discrete to Continuous-Time

- First pointed out in (Song et al., 2021). The Markov chain is a Euler discretization of the **Ornstein-Ulhenbeck**

$$\boxed{\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \quad \mathbf{X}_0 \sim p_{\mathrm{data}}.}$$

- The **reverse-time process** $(\mathbf{Y}_t)_{t\in[0,T]} = (\mathbf{X}_{T-t})_{t\in[0,T]}$ satisfies (Haussmann et al., 1986) (Conforti et al., 2021)

$$\boxed{\mathrm{d}\mathbf{Y}_t = \{\mathbf{Y}_t + 2\nabla \log p_{T-t}(\mathbf{Y}_t)\}\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \quad \mathbf{Y}_0 \sim p_T.}$$

- Connection with a continuous ELBO in (Huang et al., 2021) (Durkan & Song, 2021) using **Feynman-Kac** and **Girsanov** theorem.

$$\boxed{\log(p_T(x_T)) \geq -\int_0^T \mathbb{E}[\| s_\theta(T-t, \mathbf{Y}_t) - \nabla \log p_t(\mathbf{Y}_t)\|^2]\mathrm{d}t.}$$

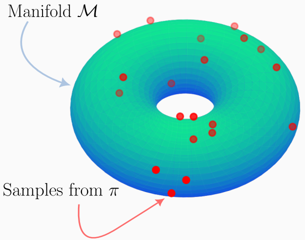
Convergence of diffusion models ($\hat{\pi}$)

Under dissipativity conditions (D.B et al., 2021[1])

- $\|\mathbf{s}_t(x) - \nabla \log p_t(x)\| \leq \mathtt{M}$.
- π admits a density p and $\langle \nabla \log p(x), x \rangle \leq -\mathtt{m}\|x\|^2 + \mathtt{c}$.

Then, there exists $A \geq 0$ such that

$$\|\pi - \hat{\pi}\|_{\mathrm{TV}} \leq A(\underbrace{\exp[-T]}_{\text{forward convergence}} + \exp[T](\underbrace{\gamma^{1/2}}_{\text{discretization}} + \underbrace{\mathtt{M}}_{\text{score approximation}}))$$



Under the manifold hypothesis (D.B., 2022[2])

- π is supported on a compact manifold $\mathcal{M}$.

Then there exists $A \geq 0$ such that

$$\mathbf{W}_1(\pi, \hat{\pi}) \leq A(\exp[-T] + \gamma^{1/2} + \mathtt{M}).$$

[1] **D.B.**, Thornton, Heng, Doucet – Diffusion Schrödinger Bridge – NeurIPS 2021
[2] **D.B.** – Convergence of diffusion models under manifold hypotheses – TMLR 2022

Bridge matching for paired transfer tasks

Paired transfer task

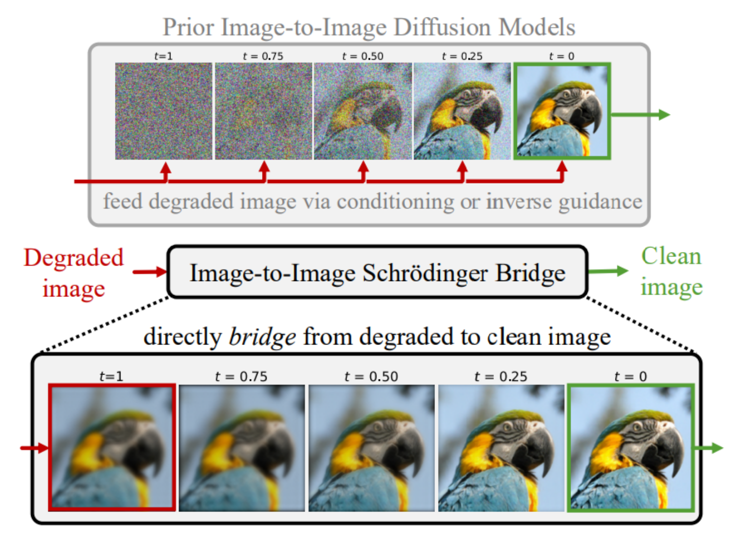


Image extracted from Liu et al. (2023).

Bridge matching

- $(\mathbf{X}_0, \mathbf{X}_T)$ (corrupted, clean) pair in the **inverse problem** example.
- **Training**:
  - Pick $(\mathbf{X}_0, \mathbf{X}_T)$.
  - Draw a sample $\mathbf{X}_t$ with a **Brownian bridge**
  - Learn the **Markov** dynamics closest to $(\mathbf{X}_t)_{t\in[0,T]}$
- **Inference**:
  - Draw $\mathbf{X}_0$ corrupted (no access to $\mathbf{X}_T$)
  - Sample from the learned dynamics
  - Get an approximation of $\mathbf{X}_T$

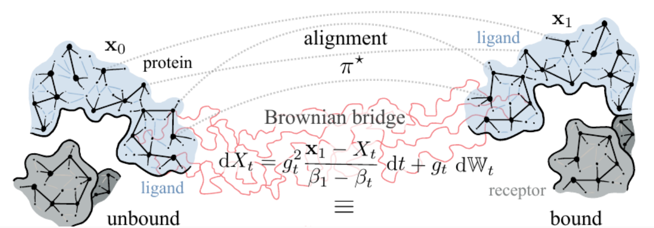


Image extracted from Somnath et al. (2023).

Markovian Projection

- **Path measure** $\mathbb{P} = \mathbb{P}_{0,T}\mathbb{Q}_{|0,T}$ with $\mathbb{Q}_{|0,T}$ associated with

$$\boxed{\mathrm{d}\mathbf{X}_t = b_t(\mathbf{X}_t, x_0, x_T)\mathrm{d}t + \sigma \mathrm{d}\mathbf{B}_t,} \qquad \text{Brownian bridge ex: } \mathrm{d}\mathbf{X}_t = \tfrac{x_T - \mathbf{X}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t.$$

- **Sampling** from $\mathbb{P}$:
  - Sample $(\mathbf{X}_0, \mathbf{X}_t) \sim \mathbb{P}_{0,T}$.
  - Sample $(\mathbf{X}_t)_{t\in[0,T]} \sim \mathbb{P}$ from $\mathrm{d}\mathbf{X}_t = b_t(\mathbf{X}_t, \mathbf{X}_0, \mathbf{X}_T)\mathrm{d}t + \sigma\mathrm{d}\mathbf{B}_t$.
- **Markovian projection**:
  - Sample $\mathbf{X}_0 \sim \mathbb{P}_0$
  - Sample $(\mathbf{X}_t)_{t\in[0,T]} \sim \mathbb{P}$ from $\mathrm{d}\mathbf{X}_t = \mathbb{E}_{0,T|t}[b_t(\mathbf{X}_t, \mathbf{X}_0, \mathbf{X}_T)|\mathbf{X}_t]\mathrm{d}t + \sigma\mathrm{d}\mathbf{B}_t$.
  - We define $\mathrm{proj}_{\mathcal{M}}(\mathbb{P}) \sim (\mathbf{X}_t)_{t\in[0,T]}$.
- **Properties**:
  - **Projection**: $\mathrm{proj}_{\mathcal{M}}(\mathbb{P}) = \mathrm{argmin}\{\mathrm{KL}(\mathbb{P}|\mathbb{M})\ ;\ \mathbb{M} \text{ is Markov}\}$.
  - **Mimicking marginals**: for any $t \in [0, T]$, $\mathbb{P}_t = \mathrm{proj}_{\mathcal{M}}(\mathbb{P})_t$.
- In the probability literature Gyöngy (1986); Brunick and Shreve (2013).

Schrödinger Bridges for general transfer tasks

Revisiting Generative Modeling using Schrödinger Bridges

- The **Schrödinger Bridge (SB) problem** is a classical problem appearing in applied mathematics, optimal transport and probability.
  - Consider a **reference density** $p(x_{0:N})$, find $\pi^\star(x_{0:N})$ such that

    $\pi^\star$ distribution on $(\mathbb{R}^d)^{N+1}$

    $$\boxed{\pi^\star = \arg\min\{\mathrm{KL}(\pi|p) : \pi_0 = p_{\text{data}},\ \pi_N = p_{\text{prior}}\}.}$$

  - **Goal:** If $\pi^\star$ is available: $X_N \sim p_{\text{prior}}$ and $X_k \sim \pi^\star_{k|k+1}(\cdot|X_{k+1})$.

- **Static formulation**: $\pi^\star(x_{0:N}) = \pi^{\mathrm{s},\star}(x_0, x_N)p_{|0,N}(x_{1:N-1}|x_0, x_N)$ where
  - Variational form:

    $\pi^{s,\star}$ distribution on $(\mathbb{R}^d)^2$

    $$\boxed{\pi^{\mathrm{s},\star} = \arg\min\{\mathrm{KL}(\pi^{\mathrm{s}}|p_{0,N}) : \pi^{\mathrm{s}}_0 = p_{\text{data}},\ \pi^{\mathrm{s}}_N = p_{\text{prior}}\}.}$$

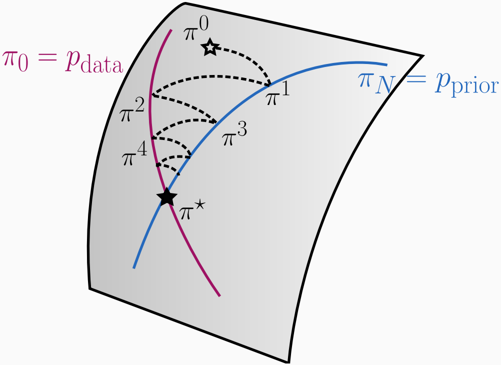
  - In its static form the Schrödinger Bridge is a special case of **entropic optimal transport**, see Mikami (2004).

The Iterative Proportional Fitting algorithm

- The SB problem can be solved using **Iterative Proportional Fitting (IPF)** Sinkhorn and Knopp (1967); Fortet (1940), i.e. set $\pi^0 = p$ and for $n \in \mathbb{N}$

$$\pi^{2n+1} = \operatorname{argmin}\{\mathrm{KL}(\pi|\pi^{2n}),\ \ \pi_N = p_{\text{prior}}\},$$
$$\pi^{2n+2} = \operatorname{argmin}\{\mathrm{KL}(\pi|\pi^{2n+1}),\ \ \pi_0 = p_{\text{data}}\}.$$

- This is akin to **alternative projection** in a Euclidean setting.
- $\lim_{n\to+\infty} \pi^n = \pi^\star$ under regularity conditions.



Continuous Schrödinger Bridge

- **Continuous-time** Schrödinger Bridge problem:

$$\boxed{\mathbb{P}^\star = \operatorname{argmin}\{\mathrm{KL}(\mathbb{P}|\mathbb{Q})\ ;\ \mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T],\mathbb{R}^d)),\ \mathbb{P}_0 = \mu_0,\ \mathbb{P}_T = \mu_1\}.}$$

  - $\mathbb{Q}, \mathbb{P}$ are **path measures**.
  - $\mathbb{Q}$ is a Markov **reference measure** (for instance $(\mathbf{B}_t)_{t \in [0,T]}$).

- **Properties of $\mathbb{P}^\star$**:

  - $\mathbb{Q}$ associated with $(\mathbf{B}_t)_{t \in [0,T]}$, $\mathbb{P}^\star_{0,T}$ **entropic OT** (reg. $1/T$).
  - Link with **static Schrödinger Bridge** $\mathbb{P}^\star = \pi^{s,\star}\mathbb{Q}_{|0,T}$.

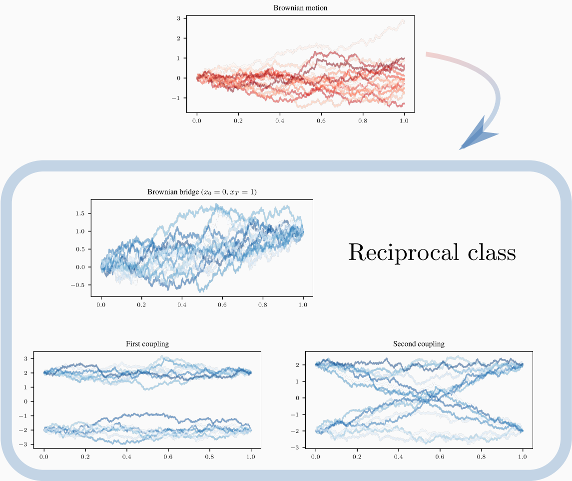
- **Continuous-time** IPF:

$$\mathbb{P}^{2n+1} = \operatorname{argmin}\{\mathrm{KL}(\mathbb{P}|\mathbb{P}^{2n}),\ \ \mathbb{P}_T = \mu_1\},$$
$$\mathbb{P}^{2n+2} = \operatorname{argmin}\{\mathrm{KL}(\mathbb{P}|\mathbb{P}^{2n+1}),\ \ \mathbb{P}_0 = \mu_0\}.$$

- Next: a property of $\mathbb{P}^\star$ and new numerical scheme.

Reciprocal class

- **Reciprocal class** of $\mathbb{Q}$ ($\mathcal{R}_{\mathbb{Q}}$), Léonard et al. (2014):
  - $\mathbb{Q}_{|0,T}$ is the **bridge measure** associated with $\mathbb{Q}$.
  - $\mathcal{R}_{\mathbb{Q}}$ is the set of path measures with **same bridge measure** as $\mathbb{Q}$.



A new scheme: Iterative Markovian Fitting

A characterization of the Schrödinger Bridge Léonard (2014)

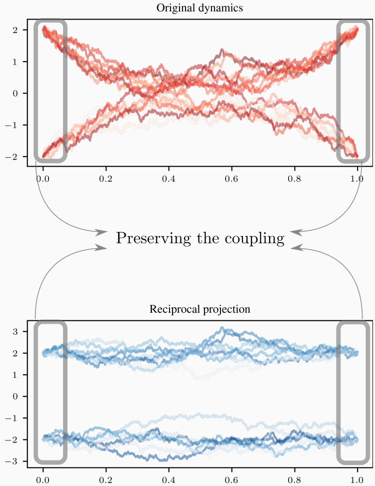
Under mild assumptions, $\mathbb{P}^\star$ is the *only* path measure such that:

- $\mathbb{P}^\star$ is Markov.
- $\mathbb{P}^\star$ is in the reciprocal class of $\mathbb{Q}$, $\mathbb{P}^\star \in \mathcal{R}(\mathbb{Q})$.
- $\mathbb{P}^\star_T = \mu_1$.
- $\mathbb{P}^\star_0 = \mu_0$.

- The **Iterative Proportional Fitting (IPF)**:
  - **Alternate projections** on $\mathbb{P}_1 = \mu_1$ and $\mathbb{P}_0 = \mu_0$.
  - **Preserve the properties** $\mathbb{P}$ is Markov and $\mathbb{P} \in \mathcal{R}(\mathbb{Q})$.
- **!NEW!** The **Iterative Markovian Fitting (IMF)**:
  - **Preserve the properties** $\mathbb{P}_1 = \mu_1$ and $\mathbb{P}_0 = \mu_0$.
  - **Alternate projections** on $\mathbb{P}$ is Markov and $\mathbb{P} \in \mathcal{R}(\mathbb{Q})$.

Reciprocal projection

- **Reciprocal projection** $\mathrm{proj}_{\mathcal{R}(\mathbb{Q})}(\mathbb{P}) = \mathbb{P}_{0,T}\mathbb{Q}_{|0,T}$
  - **Projection**: $\mathrm{proj}_{\mathcal{R}(\mathbb{Q})}(\mathbb{P}) = \mathrm{argmin}\{\mathrm{KL}(\mathbb{Q}|\mathbb{M}) \ ; \ \mathbb{M} \in \mathcal{R}(\mathbb{Q})\}$
  - **Marginals**: $\mathrm{proj}_{\mathcal{R}(\mathbb{Q})}(\mathbb{P})_0 = \mathbb{P}_0$, $\mathrm{proj}_{\mathcal{R}(\mathbb{Q})}(\mathbb{P})_T = \mathbb{P}_T$



Markovian Projection

- **Path measure** $\mathbb{P} = \mathbb{P}_{0,T}\mathbb{Q}_{|0,T}$ with $\mathbb{Q}_{|0,T}$ associated with

$\boxed{\mathrm{d}\mathbf{X}_t = b_t(\mathbf{X}_t, x_0, x_T)\mathrm{d}t + \sigma \mathrm{d}\mathbf{B}_t,}$ Brownian bridge ex: $\mathrm{d}\mathbf{X}_t = \frac{x_T - \mathbf{X}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t$.

- **Sampling** from $\mathbb{P}$:
  - Sample $(\mathbf{X}_0, \mathbf{X}_t) \sim \mathbb{P}_{0,T}$.
  - Sample $(\mathbf{X}_t)_{t\in[0,T]} \sim \mathbb{P}$ from $\mathrm{d}\mathbf{X}_t = b_t(\mathbf{X}_t, \mathbf{X}_0, \mathbf{X}_T)\mathrm{d}t + \sigma \mathrm{d}\mathbf{B}_t$.
- **Markovian projection**:
  - Sample $\mathbf{X}_0 \sim \mathbb{P}_0$
  - Sample $(\mathbf{X}_t)_{t\in[0,T]} \sim \mathbb{P}$ from $\mathrm{d}\mathbf{X}_t = \mathbb{E}_{0,T|t}[b_t(\mathbf{X}_t, \mathbf{X}_0, \mathbf{X}_T)|\mathbf{X}_t]\mathrm{d}t + \sigma \mathrm{d}\mathbf{B}_t$.
  - We define $\mathrm{proj}_{\mathcal{M}}(\mathbb{P}) \sim (\mathbf{X}_t)_{t\in[0,T]}$.
- **Properties**:
  - **Projection**: $\mathrm{proj}_{\mathcal{M}}(\mathbb{P}) = \mathrm{argmin}\{\mathrm{KL}(\mathbb{P}|\mathbb{M})\ ;\ \mathbb{M} \text{is Markov}\}$.
  - **Mimicking marginals**: for any $t \in [0, T]$, $\mathbb{P}_t = \mathrm{proj}_{\mathcal{M}}(\mathbb{P})_t$.
- In the probability literature Gyöngy (1986); Brunick and Shreve (2013).

IMF versus IPF

Iterative Markovian Fitting

- Alternative projection on:
  - Markov measures.
  - Reciprocal class of $\mathbb{Q}$.
- Preserving properties:
  - $\mathbb{P}_0 = \mu_0$.
  - $\mathbb{P}_1 = \mu_1$.
- Theoretical analysis: Shi et al. (2023); Peluchetti (2023).
- Dynamic implementation: **D**iffusion **S**chrödinger **B**ridge **M**atching (DSBM)
- Links with flow/bridge matching

Iterative Proportional Fitting

- Alternative projection on:
  - $\mathbb{P}_0 = \mu_0$.
  - $\mathbb{P}_1 = \mu_1$.
- Preserving properties:
  - Markov measures.
  - Reciprocal class of $\mathbb{Q}$.
- Theoretical analysis: Léonard (2019); Ruschendorf (1995)...
- Dynamic implementation: **D**iffusion **S**chrödinger **B**ridge (DSB)
- Links with diffusion models

Diffusion Schrödinger Bridge Matching

Practical Markovian projection

- Implementing **reciprocal projection** is easy.
- Bottleneck: **Markovian projection**.

Forward/Backward Markovian projection Shi et al. (2023)

Let $\mathbb{P} = \mathbb{P}_{0,T}\mathbb{Q}_{|0,T}$ ($\mathbb{Q}_{|0,T}$ Brownian bridge), $\mathrm{proj}_{\mathcal{M}}(\mathbb{P})$ is given by $(\mathbf{X}_t)_{t\in[0,T]}$

$$\boxed{\mathrm{d}\mathbf{X}_t = \frac{\mathbb{E}_{T|t}[\mathbf{X}_T|\mathbf{X}_t]-\mathbf{X}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \qquad \mathbf{X}_0 \sim \mathbb{P}_0.} \tag{Forward}$$

but also $(\mathbf{Y}_{T-t})_{t\in[0,T]}$

$$\boxed{\mathrm{d}\mathbf{Y}_t = \frac{\mathbb{E}_{0|t}[\mathbf{Y}_T|\mathbf{Y}_t]-\mathbf{Y}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \qquad \mathbf{Y}_0 \sim \mathbb{P}_T.} \tag{Backward}$$

- **Forward** and **Backward** representations.
- In practice:
  - Bias **accumulates** along the trajectory, i.e $\mathcal{L}(\mathbf{X}_T) \approx \mathbb{P}_T$, $\mathcal{L}(\mathbf{Y}_T) \approx \mathbb{P}_0$.
  - **Alternating** between forward/backward projection removes the bias.

Loss functions

- **Forward representation**:

$$\boxed{\mathrm{d}\mathbf{X}_t = \frac{\mathbb{E}_{T|t}[\mathbf{X}_T|\mathbf{X}_t]-\mathbf{X}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \qquad \mathbf{X}_0 \sim \mathbb{P}_0.} \tag{Forward}$$

- **Backward representation**:

$$\boxed{\mathrm{d}\mathbf{Y}_t = \frac{\mathbb{E}_{0|T-t}[\mathbf{Y}_T|\mathbf{Y}_t]-\mathbf{Y}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \qquad \mathbf{Y}_0 \sim \mathbb{P}_T.} \tag{Backward}$$

- **Neural networks** x_T^θ, x_0^Ψ with **loss functions**

$$\mathcal{L}(\theta) = \int_0^T \mathbb{E}_{t,T}[\|\mathbf{X}_T - x_T^\theta(t,\mathbf{X}_t)\|^2]\mathrm{d}t, \qquad x_T^\theta(t,x_t) \approx \mathbb{E}_{T|t}[\mathbf{X}_T|\mathbf{X}_t = x_t],$$
$$\mathcal{L}(\Psi) = \int_0^T \mathbb{E}_{0,t}[\|\mathbf{X}_0 - x_0^\Psi(t,\mathbf{X}_t)\|^2]\mathrm{d}t, \qquad x_0^\Psi(t,x_t) \approx \mathbb{E}_{0|T-t}[\mathbf{X}_0|\mathbf{X}_t = x_t].$$

- **Practical forward representation**:

$$\boxed{\mathrm{d}\mathbf{X}_t = \frac{x_T^\theta(t,\mathbf{X}_t)-\mathbf{X}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \qquad \mathbf{X}_0 \sim \mathbb{P}_0.}$$

- **Practical backward representation**:

$$\boxed{\mathrm{d}\mathbf{Y}_t = \frac{x_0^\Psi(T-t,\mathbf{Y}_t)-\mathbf{Y}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t, \qquad \mathbf{Y}_0 \sim \mathbb{P}_T.}$$

One cycle (4 IMF iterations)

Algorithm 1: One IMF cycle

Sample $(\mathbf{X}_t)_{t\in[0,T]} \sim \mathbb{P}^0$

Extract $(\mathbf{X}_0, \mathbf{X}_T)$

Get Brownian bridge with end points $(\mathbf{X}_0, \mathbf{X}_T)$, $(\mathbf{X}_t)_{t\in[0,T]} \sim \mathbb{P}^1$

Compute loss $\mathcal{L}(\theta)$

Update x_T^θ

Sample from $\mathrm{d}\mathbf{X}_t = \frac{x_T^\theta(t,\mathbf{X}_t)-\mathbf{X}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t$, $\mathbf{X}_0 \sim \mathbb{P}_0$, $(\mathbf{X}_t)_{t\in[0,T]} \sim \mathbb{P}^2$

Extract $(\mathbf{X}_0, \mathbf{X}_T)$

Get Brownian bridge with end points $(\mathbf{X}_0, \mathbf{X}_T)$, $(\mathbf{X}_t)_{t\in[0,T]} \sim \mathbb{P}^3$

Compute loss $\mathcal{L}(\Psi)$

Update x_0^Ψ

Sample from $\mathrm{d}\mathbf{Y}_t = \frac{x_0^\Psi(T-t,\mathbf{Y}_t)-\mathbf{Y}_t}{T-t}\mathrm{d}t + \mathrm{d}\mathbf{B}_t$, $\mathbf{Y}_0 \sim \mathbb{P}_{T\cdot}$, $(\mathbf{Y}_{T-t})_{t\in[0,T]} \sim \mathbb{P}^4$

- **Reciprocal projection**: $\mathbb{P}^1 = \mathrm{proj}_{\mathcal{R}(\mathbb{Q})}(\mathbb{P}^0)$, $\mathbb{P}^3 = \mathrm{proj}_{\mathcal{R}(\mathbb{Q})}(\mathbb{P}^2)$.
- **Markovian projection**: $\mathbb{P}^2 = \mathrm{proj}_{\mathcal{M}}(\mathbb{P}^1)$, $\mathbb{P}^4 = \mathrm{proj}_{\mathcal{M}}(\mathbb{P}^3)$.
  - $\mathbb{P}^2$ has a forward representation, no $\mathbb{P}^2_0 = \mu_0$.
  - $\mathbb{P}^4$ has a backward representation, no $\mathbb{P}^4_T = \mu_1$.
- Full algorithm: loop $\mathbb{P}^0 \leftarrow \mathbb{P}^4$.

Link with existing literature

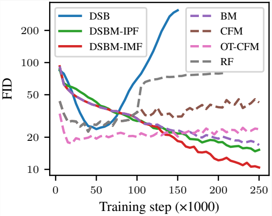
- IMF counterpart of **Diffusion Schrödinger Bridge** De Bortoli et al. (2021)
  - Can be seen as **improved numerics** (cache loader).
  - No bias accumulation on the bridge measure.

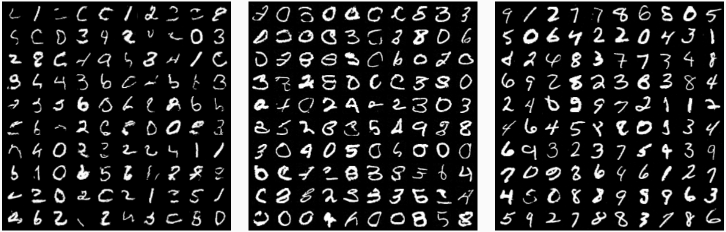
- Losses resemble **flow matching** losses:
  - Deterministic Lipman et al. (2022); Tong et al. (2023); Chen and Lipman (2023).
  - Bridge matching Liu et al. (2022b).
  - **Stochastic interpolants** Albergo et al. (2023).
  - Conditional Bridge matching Liu et al. (2023); Somnath et al. (2023).

- Iterated flow matching, **Rectified flow** Liu et al. (2022a)
  - Deterministic limit.
  - Forward and backward Markovian projection.
  - Concurrent work Peluchetti (2023).

Experiments

- Influence of **initial coupling**:
  - $\mathbb{P}^0 = \mu_0 \mathbb{Q}_{|0}$, **DSBM-IPF**.
  - $\mathbb{P}^0 = (\mu_0 \otimes \mu_T)\mathbb{Q}_{|0,T}$, **DSBM-IMF**.
- Comparison on MNIST:
  - Better than **flow matching** methods.
  - DSB **accumulates bias**.





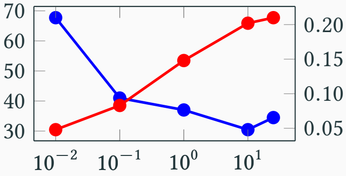
(a) OT-CFM **(b)** DSB **(c)** DSBM-IPF

Influence of the reference measure

- We always choose $\mathbb{Q}$ associated with $(\sigma \mathbf{B}_t)_{t \in [0,T]}$. Influence of σ:
  - Small σ: **better transfer**, **harder to learn** (higher FID).
  - high σ: **worse transfer**, **easier to learn** (lower FID).

- From *male* to *female*.
- $\sigma \in \{0.01, 0.1, 1, 10\}$ (initial samples left).

- Metrics (lower=better):
  - **LPIP** (similarity measure).
  - **FID** (quality measure).

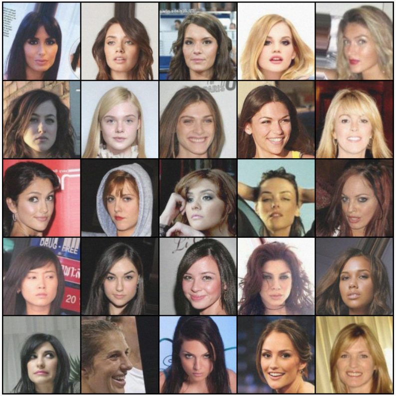


Celeba 128×128

- Male to female.

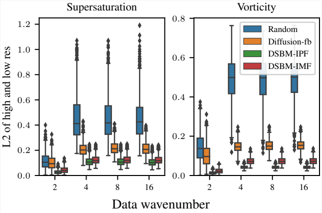
Celeba 128×128

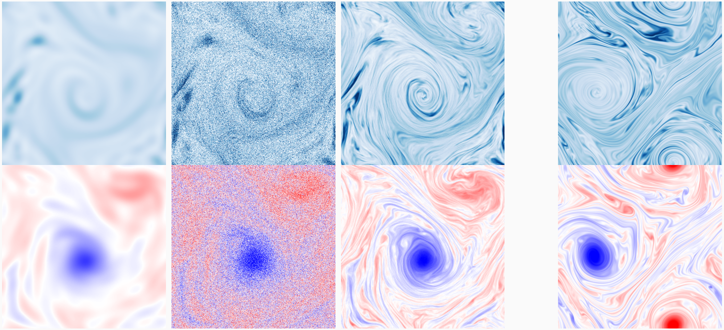
- Female to male.



Downscaling task

- Same setting as Bischoff and Deck (2023).
- **Super resolution** task.
- Quality measure (frequency histogram).
- Similarity measure (ℓ_2 with upscaling).

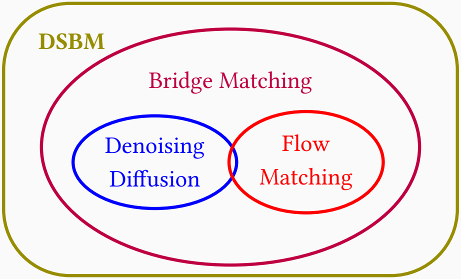




Conclusion

Conclusion

- **Methodology** to solve **Schrödinger Bridge**: **Iterative Markovian Fitting** (IMF).
- **Numerics** to solve IMF: **Diffusion Schrödinger Bridge Matching** (DSBM).
- Links with **optimal transport**, **optimal control**.
- Better numerical properties than De Bortoli et al. (2021).



References

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Victoria A Avanzato, Kasopefoluwa Y Oguntuyo, Marina Escalera-Zamudio, Bernardo Gutierrez, Michael Golden, Sergei L Kosakovsky Pond, Rhys Pryce, Thomas S Walter, Jeffrey Seow, Katie J Doores, et al. A structural basis for antibody-mediated neutralization of nipah virus reveals a site of vulnerability at the fusion glycoprotein apex. *Proceedings of the National Academy of Sciences*, 116(50):25057–25067, 2019.

Tobias Bischoff and Katherine Deck. Unpaired downscaling of fluid flows with diffusion bridges. *arXiv preprint arXiv:2305.01822*, 2023.

References ii

Gerard Brunick and Steven Shreve. Mimicking an Itô process by a solution of a stochastic differential equation. *The Annals of Applied Probability*, 23(4): 1584–1628, 2013.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, pages 2021–12, 2021.

Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022.

Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. *arXiv preprint arXiv:2104.07708*, 2021.

Ricky TQ Chen and Yaron Lipman. Riemannian flow matching on general geometries. *arXiv preprint arXiv:2302.03660*, 2023.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34, 2021.

Robert Fortet. Résolution d'un système d'équations de M. Schrödinger. *Journal de Mathématiques Pures et Appliqués*, 1:83–105, 1940.

István Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an Itô differential. *Probability Theory and Related Fields*, 71:501–516, 1986.

Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205, 1986.

Chin-Wei Huang, Jae Hyun Lim, and Aaron C Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34, 2021.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems-A*, 34(4):1533–1574, 2014.

Christian Léonard. Revisiting Fortet's proof of existence of a solution to the Schrödinger system. *arXiv preprint arXiv:1904.13211*, 2019.

Christian Léonard, Sylvie Rœlly, Jean-Claude Zambrini, et al. Reciprocal processes: a measure-theoretical point of view. *Probability Surveys*, 11:237–269, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I2sb: Image-to-image schrödinger bridge. *arXiv*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022a.

References v

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699*, 2022b.

Toshio Mikami. Monge's problem with a quadratic cost by the zero-noise limit of h-path processes. *Probability theory and related fields*, 129(2):245–260, 2004.

Stefano Peluchetti. Diffusion bridge mixture transports, schrodinger bridge problems and generative modeling. *arXiv preprint arXiv:2304.00917*, 2023.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Ludger Ruschendorf. Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, pages 1160–1174, 1995.

References vi

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrodinger bridge matching. *arXiv preprint arXiv:2303.16852*, 2023.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Vignesh Ram Somnath, Matteo Pariset, Ya-Ping Hsieh, Maria Rodriguez Martinez, Andreas Krause, and Charlotte Bunne. Aligned diffusion schrodinger bridges. *arXiv preprint arXiv:2302.11419*, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34, 2021.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *The Eleventh International Conference on Learning Representations*, 2022.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Approximating Backward Transitions

- We restrict ourselves to discretized **Ornstein-Ulhenbeck** processes

$$p_{k+1|k}(x_{k+1}|x_k) = \mathcal{N}(x_{k+1}; x_k - \gamma x_k, \sqrt{\gamma}\mathrm{Id}),$$

($\gamma > 0$ is close to 0)

- Using a Taylor expansion we get

$$p_{k|k+1}(x_k|x_{k+1}) = p_{k+1|k}(x_{k+1}|x_k)\exp[\log p_k(x_k) - \log p_{k+1}(x_{k+1})]$$
$$\approx \mathcal{N}(x_k; x_{k+1} + \gamma x_{k+1} + 2\gamma \underbrace{\nabla \log p_{k+1}(x_{k+1})}_{\text{Stein score}}, \sqrt{2\gamma}\mathrm{Id}).$$

- The **Stein score** is not available but using that $p_{k+1}(x_{k+1}) = \int p_0(x_0) p_{k+1|0}(x_{k+1}|x_0)\mathrm{d}x_0$, we get that

$$\boxed{\nabla \log p_{k+1}(x_{k+1}) = \mathbb{E}_{p_{0|k+1}}[\nabla_{x_{k+1}} \log p_{k+1|0}(x_{k+1}|X_0)].}$$

Estimating the Scores using Score Matching

- **Conditional expectation** $\rightarrow$ **Regression problem**

$$s_{k+1} = \text{argmin}_s \ \mathbb{E}_{p_{0,k+1}}[||s(X_{k+1}) - \nabla_{x_{k+1}} \log p_{k+1|0}(X_{k+1}|X_0)||^2].$$

- In practice, we restrict ourselves to **neural networks** and estimate all scores simultaneously i.e. $s_{\theta^\star}(k, x_k) \approx \nabla \log p_k(x_k)$ where

$$\boxed{\theta^\star \approx \text{argmin}_\theta \sum_{k=1}^{N} \mathbb{E}_{p_{0,k}}[||s_\theta(k, X_k) - \nabla_{x_k} \log p_{k|0}(X_k|X_0)||^2],}$$

- If $\log p_{k+1|0}(x_{k+1}|x_0)$ is not available, then use

$$\nabla \log p_{k+1}(x_{k+1}) = \mathbb{E}_{p_{k|k+1}}[\nabla_{x_{k+1}} \log p_{k+1|k}(x_{k+1}|X_k)]$$

- Can also be derived from a **continuous-time** perspective (time-reversal of diffusion, Feynman-Kac formula) and can be seen as ELBO (Huang et al., 2021).
- Yet another approach goes fully variational (Ho et al., 2020).

Sketch of the proof

- The central decomposition

$$
\begin{aligned}
||\mathcal{L}(X_0) - p_{\mathrm{data}}||_{\mathrm{TV}} &= ||p_{\mathrm{prior}}\hat{\mathrm{R}}_N - p_{\mathrm{data}}||_{\mathrm{TV}} \\
&= ||p_{\mathrm{prior}}\hat{\mathrm{R}}_N - p_T \mathrm{Q}_T||_{\mathrm{TV}} \\
&\leq ||p_{\mathrm{prior}}\hat{\mathrm{R}}_N - p_{\mathrm{prior}}\mathrm{Q}_T||_{\mathrm{TV}} + ||p_T \mathrm{Q}_T - p_{\mathrm{prior}}\mathrm{Q}_T||_{\mathrm{TV}} \\
&\leq ||p_{\mathrm{prior}}\hat{\mathrm{R}}_N - p_{\mathrm{prior}}\mathrm{Q}_T||_{\mathrm{TV}} + ||p_{\mathrm{data}}\mathrm{P}_T - p_{\mathrm{prior}}||_{\mathrm{TV}},
\end{aligned}
$$

  where
  - $(\mathrm{P}_t)_{t\geq 0}$ is the **forward** Ornstein-Ulhenbeck semi-group,
  - $(\mathrm{Q}_t)_{t\geq 0}$ is the **backward** Ornstein-Ulhenbeck semi-group,
  - $(\hat{\mathrm{R}}_n)_{n\in\{1,\ldots,N\}}$ is the iterated kernel associated with the backward Markov chain.

- $||p_{\mathrm{prior}}\hat{\mathrm{R}}_N - p_{\mathrm{prior}}\mathrm{Q}_T||_{\mathrm{TV}}$: **approximation error** $\rightarrow$ Girsanov theorem.

- $||p_{\mathrm{data}}\mathrm{P}_T - p_{\mathrm{prior}}||_{\mathrm{TV}}$: **geometric ergodicity** of Ornstein-Ulhenbeck.

Reverse process on a compact manifold

- The **Brownian motion** is defined as a process $(\mathbf{B}_t^{\mathcal{M}})_{t \geq 0}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\mathbf{M}_t^f)_{t \geq 0}$ is a martingale where for any $t \geq 0$

$$\boxed{\mathbf{M}_t^f = f(\mathbf{B}_t^{\mathcal{M}}) - f(\mathbf{B}_0^{\mathcal{M}}) - \int_0^t (1/2)\Delta_{\mathcal{M}}(f)(\mathbf{B}_s^{\mathcal{M}}) \mathrm{d}s.}$$

- The **reverse process** is given by $(\mathbf{Y}_t)_{t \in [0,T]}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\mathbf{M}_t^f)_{t \geq 0}$ is a martingale where for any $t \in [0, T]$

$$\boxed{\mathbf{M}_t^f = f(\mathbf{Y}_t) - f(\mathbf{Y}_0) - \int_0^t \{\langle \nabla \log p_t(\mathbf{X}_s), \nabla f(\mathbf{Y}_s)\rangle_{\mathcal{M}} + (1/2)\Delta_{\mathcal{M}}(f)(\mathbf{Y}_s)\} \mathrm{d}s.}$$

- This is an extension of **reversal** results (Haussmann et al., 1986) (Conforti et al., 2021).

- **Take-home message:** The formula is the same except that **gradients, scalar product and Laplacian** are considered w.r.t. the underlying metric.

Sampling on a manifold

- How to sample from the process $(\mathbf{Y}_t)_{t\in[0,T]}$ (approximately)?
- Equivalent of the **Euler-Maruyama** discretization is the **Geodesic Random Walk** (GRW)

Definition of GRW

Let X_0^γ be a M-valued random variable. For any $\gamma > 0$, we define $(X_n^\gamma)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}$,

$$\boxed{X_{n+1}^\gamma = \exp_{X_n^\gamma}\left(\gamma\{b(X_n^\gamma) + (1/\sqrt{\gamma})(V_{n+1} - b(X_n^\gamma))\}\right).}$$

where $(V_n)_{n\in\mathbb{N}}$ is a sequence of M-valued random variables such that for any $n \in \mathbb{N}$, V_{n+1} has distribution $\nu_{X_n^\gamma}$ conditionally to X_n^γ (mean $b(X_n^\gamma)$, covariance $\Sigma(X_n^\gamma)$).

- **Weakly converges** towards the diffusion $\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \Sigma(\mathbf{X}_t)\mathrm{d}\mathbf{B}_t^{\mathsf{M}}$ for small stepsizes γ.
- Hard to obtain **quantitative results** (coupling techniques in Riemannian setting).

Perspectives & Challenges

Plan

Some challenges:

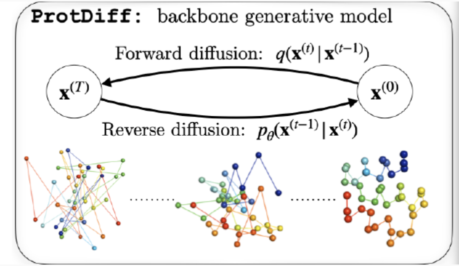
- **Scaling up** Diffusion Schrodinger Bridge and protein applications.
- Particle evolution and **probabilistic splines**.
- **Theoretical understanding** of diffusion models and other projects.

Scaling up and protein applications

- To be competitive: access to large **GPU infrastructure**.

ImageNet 512×512							
BigGAN-deep [5]			256-512	8.43	8.13	**0.88**	0.29
ADM-G (4360K), ADM-U (1050K)	1878	36	1914	**3.85**	**5.86**	0.84	**0.53**
ADM-G (500K), ADM-U (100K)	189	9*	**198**	7.59	6.84	0.84	**0.53**

- More than **200** V100 days to train one SoTA diffusion model on ImageNet 512×512.
- Importance of the scaling for:
  - **Image processing** (realistic outputs, interaction with language models...)
  - **Protein Modeling** (long proteins...) (image from Trippe et al. (2022))



Particle evolution and spline

- For **population evolution**, one Schrödinger bridge is not enough.
- **Multiple snapshots**, can we consider multiple Schrödinger bridges?
- How can we impose some regularity in the **probabilistic structure**?
  - **Spline** in probabilistic spaces (Chen et al. (2018))
  - Efficient combination with Diffusion Schrödinger Bridges.

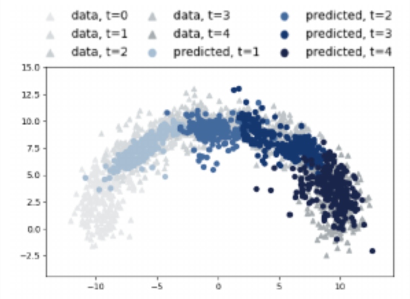


Image extracted from Bunne et al. (2022)

Theoretical understanding of diffusion models & other projects

- A lot of **open questions**:
  - Role of the **manifold hypothesis**.
  - Role of the **Empirical measure**.
  - And what about **multimodal** behavior?

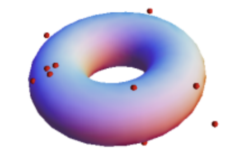
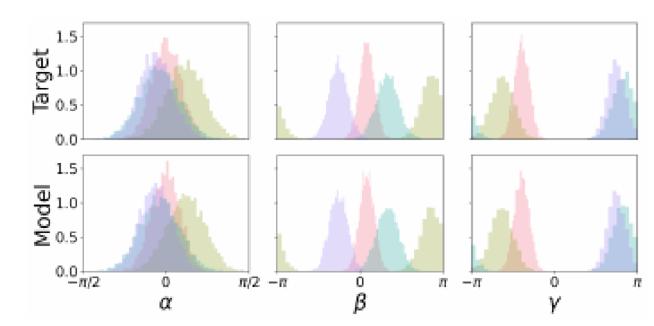


Image extracted from Fefferman et al. (2015)

- Other projects
  - **VAE** as **entropic regularization**
  - Interpretation of **Transformers** with **category theory** tools.

Some results on $SO_3(\mathbb{R})$

- An illustration: targeting **multimodal distributions** on $SO_3(\mathbb{R})$.



Method	$M = 16$		$M = 32$	
	log-likelihood	NFE	log-likelihood	NFE
Moser Flow	$0.85_{\pm 0.03}$	$2.3_{\pm 0.5}$	$0.17_{\pm 0.03}$	$2.3_{\pm 0.9}$
Exp-wrapped SGM	$\mathbf{0.87}_{\pm \mathbf{0.04}}$	$0.5_{\pm 0.1}$	$0.16_{\pm 0.03}$	$0.5_{\pm 0.0}$
RSGM	$\mathbf{0.89}_{\pm \mathbf{0.03}}$	$\mathbf{0.1}_{\pm \mathbf{0.0}}$	$\mathbf{0.20}_{\pm \mathbf{0.03}}$	$\mathbf{0.1}_{\pm \mathbf{0.0}}$

Motivation

- Many datasets do *not* lie on a **Euclidean space**.
- We need to include a **geometric prior**:
  - **Protein modeling** (Boomsma et al., 2008; Hamelryck et al., 2006; Mardia et al., 2008; Shapovalov and Dunbrack Jr, 2011; Mardia et al., 2007).
  - **Geological sciences** (Karpatne et al., 2018; Peel et al., 2001).
  - **Robotics** (Feiten et al., 2013; Senanayake and Ramos, 2018).

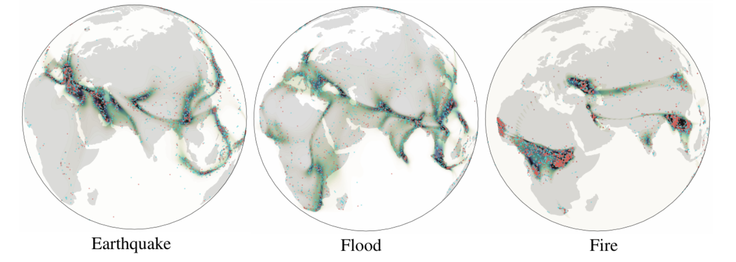


Image extracted from Mathieu et al., 2020.

Noising process on a compact manifold

- To define a **score-based generative modeling** we need to define a **noising process**
  - In **Euclidean spaces** we choose a **Ornstein-Ulhenbeck** process.
  - In **Riemannian manifold** we choose a **Brownian motion**.
- In the **Euclidean** setting the **Ornstein-Ulhenbeck** process converges towards a unit Gaussian.
- In the *compact* **Riemannian manifold** setting the **Brownian motion** converges towards the uniform distribution.

Geometric ergodicity (Urakawa, 2006, Proposition 2.6)

For any $t > 0$, P_t admits a density $p_{t|0}$ w.r.t. p_{ref} and $p_{\mathrm{ref}}\mathrm{P}_t = p_{\mathrm{ref}}$, i.e. p_{ref} is an invariant measure for $(\mathrm{P}_t)_{t\geq 0}$. In addition, if there exists $C, \alpha \geq 0$ such that $p_{t|0}(x|x) \leq Ct^{-\alpha/2}$ for any $t \in (0, 1]$ and any $x \in \mathcal{M}$ then for any $p_0 \in \mathcal{P}(\mathcal{M})$ and for any $t \geq 1/2$ we have

$$\|p_0\mathrm{P}_t - p_{\mathrm{ref}}\|_{\mathrm{TV}} \leq C^{1/2}\mathrm{e}^{\lambda_1/2}\mathrm{e}^{-\lambda_1 t},$$

where λ_1 is the first non-negative eigenvalue of $-\Delta_{\mathcal{M}}$ in $\mathrm{L}^2(p_{\mathrm{ref}})$.

Reverse process on a compact manifold

- The **Brownian motion** is defined as a process $(\mathbf{B}_t^{\mathcal{M}})_{t \geq 0}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\mathbf{M}_t^f)_{t \geq 0}$ is a martingale where for any $t \geq 0$

$$\mathbf{M}_t^f = f(\mathbf{B}_t^{\mathcal{M}}) - f(\mathbf{B}_0^{\mathcal{M}}) - \int_0^t (1/2) \Delta_{\mathcal{M}}(f)(\mathbf{B}_s^{\mathcal{M}}) \mathrm{d}s.$$

- The **reverse process** is given by $(\mathbf{Y}_t)_{t \in [0,T]}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\mathbf{M}_t^f)_{t \geq 0}$ is a martingale where for any $t \in [0, T]$

$$\boxed{\mathbf{M}_t^f = f(\mathbf{Y}_t) - f(\mathbf{Y}_0) - \int_0^t \{ \langle \nabla_{\mathcal{M}} \log p_t(\mathbf{X}_s), \nabla_{\mathcal{M}} f(\mathbf{Y}_s) \rangle_{\mathcal{M}} + (1/2) \Delta_{\mathcal{M}}(f)(\mathbf{Y}_s) \} \mathrm{d}s.}$$

- This is an extension of **reversal** results (Haussmann et al., 1986) (Conforti et al., 2021).
- The formula is the same except that **gradients, scalar product and Laplacian** are considered w.r.t. the underlying metric.

Sampling on a manifold

- How to sample from the process $(bfY_t)_{t\in[0,T]}$ (approximately)?
- Equivalent of the **Euler-Maruyama** discretization is the **Geodesic Random Walk** (GRW)

Definition of GRW

Let X_0^γ be a M-valued random variable. For any $\gamma > 0$, we define $(X_n^\gamma)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}$,
$X_{n+1}^\gamma = \exp_{X_n^\gamma}\left(\gamma\{b(X_n^\gamma) + (1/\sqrt{\gamma})(V_{n+1} - b(X_n^\gamma))\}\right)$, where $(V_n)_{n\in\mathbb{N}}$ is a sequence of M-valued random variables such that for any $n \in \mathbb{N}$, V_{n+1} has distribution $\nu_{X_n^\gamma}$ conditionally to X_n^γ (mean $b(X_n^\gamma)$, covariance $\Sigma(X_n^\gamma)$).

Convergence of GRW (Jorgensen, 1975, Theorem 2.1)

Under mild conditions on M, for any $t \geq 0$, $f \in \mathrm{C}(\mathsf{M})$ we have that $\lim_{\gamma\to 0} |\mathbb{E}[f(X_{t/\gamma}^\gamma)] - \mathrm{P}_t[f]| = 0$, where $(\mathrm{P}_t)_{t\geq 0}$ is the semi-group associated with the infinitesimal generator $\mathcal{A} : \mathrm{C}^\infty(\mathsf{M}) \to \mathrm{C}^\infty(\mathsf{M})$ given for any $f \in \mathrm{C}^\infty(\mathsf{M})$ by $\mathcal{A}(f) = \langle b, \nabla f\rangle_{\mathsf{M}} + \frac{1}{2}\langle \Sigma, \nabla^2 f\rangle_{\mathsf{M}}$.

- Hard to obtain **quantitative results** (coupling techniques fail).

Loss function

- We need to estimate $\nabla \log p_t$.
- Same as **Euclidean** case, $\nabla \log p_t(x_t) = \mathbb{E}[\nabla \log p_{t|0}(\mathbf{X}_t|\mathbf{X}_0)|\mathbf{X}_t = x_t]$.
- Extra difficulty, $\nabla \log p_{t|0}$ is *not* available in **close form**.
- Two possibilities to circumvent this issue:
  - Use the **divergence theorem**

$$\boxed{\nabla \log p_t = \mathrm{argmin}_s\{(1/2)\|s(\mathbf{B}_t^{\mathcal{M}})\|^2 + \mathbb{E}[\mathrm{div}(s)(\mathbf{B}_t^{\mathcal{M}})]\}.}$$

  - Use **approximation** of $\nabla \log p_{t|0}$ (Varadhan approximation and series expansion).

$$\boxed{\nabla \log p_t = \mathrm{argmin}_s\{\mathbb{E}[\|s(\mathbf{B}_t^{\mathcal{M}}) - \nabla \log p_{t|0}(\mathbf{B}_t^{\mathcal{M}}|\mathbf{B}_0^{\mathcal{M}})\|^2]\}.}$$

Euclidean VS compact Riemannian

- **Riemannian score-based generative modeling** (RSGM)
  - Sample from the **forward dynamics**.
  - Train the **score network**.
  - Sample from the **backward dynamics** (initialized at the uniform distribution).
- Differences between the **Euclidean setting** and the **compact manifold setting**.

Ingredient \ Space	Euclidean	Compact manifold
Forward process	Ornstein–Ulhenbeck	Brownian motion
Easy-to-sample distribution	Gaussian	Uniform
Time reversal	(Cattiaux et al., 2021)	This paper
Sampling of the forward process	Direct	Geodesic Random Walk
Sampling of the backward process	Euler–Maruyama	Geodesic Random Walk

Table 1: Differences between SGM on Euclidean spaces and RSGM on compact Riemannian manifolds.

Extension to Schrödinger bridges

- We can extend the **Schrödinger bridge** framework to the manifold setting.
- Difficulty: considering an equivalent of the **mean-matching** technique on manifold (divergence form).

Implicit mean-matching loss

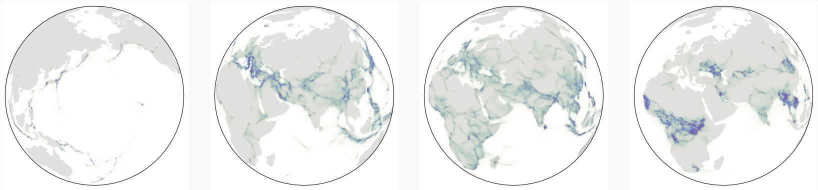
Let $(\mathbf{X}_t)_{t\in[0,T]}$ be a M-valued process with distribution $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T],\mathsf{M}))$ such that for any $t \in [0,T]$, $\mathbf{X}_t$ admits a positive density $p_t \in \mathrm{C}^\infty(\mathsf{M})$ w.r.t. p_{ref}. Let $s : [0,T] \to \mathcal{X}\mathcal{M}$. For any $t \in [0,T]$ and $x \in \mathsf{M}$, let

$$\boxed{b(t,x) = -f(t,x) + g(t,\mathbf{X}_t)^2 \nabla \log p_t(x).}$$

Then, for any $t \in [0,T]$, we have that

$$\boxed{b(t,\cdot) = \operatorname{argmin}_r \{\mathbb{E}[\tfrac{1}{2}\|f(t,\mathbf{X}_t) + r(\mathbf{X}_t)\|^2 + g(t,\mathbf{X}_t)^2 \mathrm{div}(r)(\mathbf{X}_t)]\}.}$$

Application



Learned density on Volcano/Earthquake/Flood/Fire datasets.

	Earthquake	Flood	Fire
Mixture of Kent	$0.33_{\pm 0.05}$	$0.73_{\pm 0.07}$	$-1.18_{\pm 0.06}$
Riemannian CNF	$0.19_{\pm 0.04}$	$0.90_{\pm 0.03}$	$-0.66_{\pm 0.05}$
Moser Flow	$-0.09_{\pm 0.02}$	$0.62_{\pm 0.04}$	$-1.03_{\pm 0.03}$
Stereographic Score-Based	$-0.04_{\pm 0.11}$	$1.31_{\pm 0.16}$	$0.28_{\pm 0.20}$
Riemannian Score-Based	$\mathbf{-0.21}_{\pm 0.03}$	$\mathbf{0.52}_{\pm 0.02}$	$\mathbf{-1.24}_{\pm 0.07}$
Dataset size	6120	4875	12809

Table 2: Negative log-likelihood scores for each method on the earth and climate science datasets. Bold indicates best results (up to statistical significance). Means and standard deviations are computed over 5 different runs.

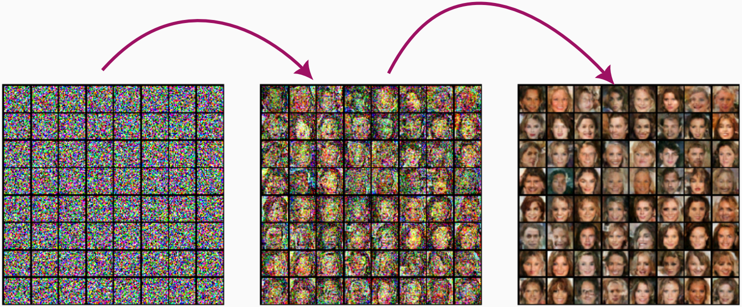
Why generative modeling? (1/2)

- Application in **meteorology**: Ravuri et al. (2021).
  - Prediction of rain in the next 2 hours: **nowcasting**.
  - Solving physical PDEs: **planet scale** predictions days ahead.
  - Struggle for **high resolution** predictions on short time ranges.

- Access to a lot of high quality data: **conditional GAN**.

Image extracted from Ravuri et al. (2021).

Some visual results



Dataset interpolation