Generative modeling via Schrödinger bridge (basics on Schrödinger bridge)

Valentin De Bortoli

March 21, 2022

Summary of the previous lecture (1/4)

- In the previous lecture we developed some **theory** for **score-based generative modeling**:
  - Continuous **time-reversal**.
  - **Approximation theorem**.
  - Connection with **Normalizing Flows**.
  - **Accelerations** of SGMs.
- Recall the basics of **SGM**:
  - Sample a **forward trajectory**, noising the distribution.

$$\boxed{X_{k+1} = X_k - \gamma X_k + \sqrt{2\gamma} Z_{k+1} \ .}$$

  - Sample a **backward trajectory** via **ancestral sampling**.

$$\boxed{X_k = X_{k+1} + \gamma\{X_{k+1} + \mathbf{s}_\theta(k\gamma, X_{k+1})\} + \sqrt{2\gamma} Z_{k+1} \ .}$$

  - Backward sampling relies on learning the **score** (**score-matching**)

$$\boxed{\mathbf{s}_{\theta^\star}(k\gamma, \cdot) = \arg\min_\theta \{\mathbb{E}[\|\mathbf{s}_\theta(k\gamma, X_k) - \nabla \log p_{k|0}(X_k|X_0)\|^2] \ : \ f \in \mathrm{L}^2(p_k)\} \ .}$$

Summary of the previous lecture (2/4)

Convergence of diffusion models (De Bortoli et al., 2021)

- Assume there exists $M \geq 0$ such that for any $t \in [0, T]$ and $x \in \mathbb{R}^d$

$$||\mathbf{s}_{\theta^\star}(t, x) - \nabla \log p_t(x)|| \leq M \ ,$$

  with $\mathbf{s}_{\theta^\star} \in C([0, T] \times \mathbb{R}^d, \mathbb{R}^d)$ and regularity conditions on the density of π w.r.t. the Lebesgue measure and its gradients.

- Then there exist $B, C, D \geq 0$ s.t. for any $N \in \mathbb{N}$ and $\{\gamma_k\}_{k=1}^N$ the following hold:

$$\boxed{||\mathcal{L}(Y_N) - \pi||_{\mathrm{TV}} \leq B \exp[-T] + C(M + \gamma^{1/2}) \exp[DT] \ .}$$

  where $T = N\gamma$.

- **A few remarks**:
  - The assumption on π is *not* satisfied if π defined on a **manifold** of $\mathbb{R}^d$ with dimension $p < d$.
  - The approximation assumption is strong and could be **relaxed**.
  - The term $\exp[DT]$ can be improved and turned into a **polynomial dependency**.

Summary of the previous lecture (3/4)

- Having a **deterministic** model is useful for:
  - **Likelihood computation**
  - **Interpolation**
  - **Temperature scaling**
- We can explore the **latent structure**.

Figure 1: Interpolation with ODE. Image extracted from Song et al. (2020).

Summary of the previous lecture (4/4)

- For **high-quality** image sampling **vanilla** SGMs are notably **slow**.

A critical drawback of these models is that they require many iterations to produce a high quality sample. For DDPMs, this is because that the generative process (from noise to data) approximates the reverse of the forward *diffusion process* (from data to noise), which could have thousands of steps; iterating over all the steps is required to produce a single sample, which is much slower compared to GANs, which only needs one pass through a network. For example, it takes around 20

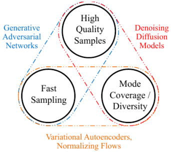
control the generation sample. To obtain high-quality synthesis, a large number of denoising steps is used (i.e. 1000 steps). A notable property of the diffusion process is a closed-form formulation of

network). Although very powerful, score-based models generate data through an undesirably long iterative process; meanwhile, other state-of-the-art methods such as GANs generate data from a single forward pass of a neural network. Increasing the speed of the generative process is thus an active area of research.

denoises the samples under the fixed noise schedule. However, DDPMs often need hundreds-to-thousands of denoising steps (each involving a feedforward pass of a large neural network) to achieve

However, GANs are typically much more efficient than DDPMs at generation time, often requiring a single forward pass through the generator network, whereas DDPMs require hundreds of forward passes through a U-Net model. Instead of learning a generator directly, DDPMs learn to convert

A major downside to score-based generative models is that they require performing expensive MCMC sampling, often with a thousand steps or more. As a result, they can be up to three orders of magnitude slower than GANs, which only require a single network evaluation. To address this issue, Denoising Diffusion Implicit Models, or DDIMs, have been



Outline of the course

- We introduce basics **Schrödinger bridges**.
- **Goal of the course**:
  - Introduce the **Schrödinger bridge (SB) problem**.
  - Present **algorithms** to solve the SB problem.
- **Outline of the course**
  - A **dynamic** and **static** Schrödinger bridges.
  - Convergence of the **Sinkhorn** algorithm.

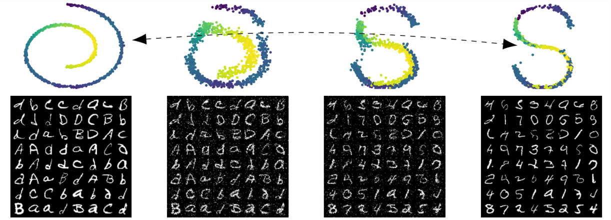


Figure 2: A Schrödinger Bridge between two data distributions. Image extracted from De Bortoli et al. (2021).

The Schrödinger Bridge Problem

Outline of the section

- In this section:
  - We present **generative modeling** via **Schrödinger Bridge** (SB).
  - We introduce **dynamic** and **static** SB.
  - We draw links with **regularized Optimal Transport** (OT).

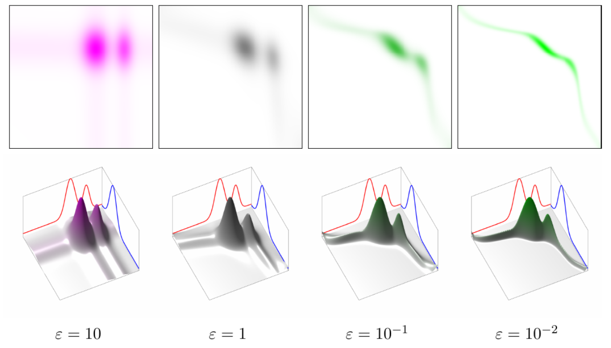


Figure 3: Entropic regularized OT. Image extracted from Peyré et al. (2019).

Generative modeling and Schrödinger bridges

The dynamical setting

- Problem introduced by Schrödinger (1932).
  - Particles follow a **Brownian motion**.
  - At $t = T$ the **observed distribution** is different from a Brownian evolution.
  - What was the **most likely** evolution?
- A first **dynamical** formulation:

$$\boxed{\pi^\star = \arg\min\{\mathrm{KL}(\pi|\pi^0) \ : \ \pi \in \mathcal{P}((\mathbb{R}^d)^N), \pi_0 = \nu_0, \ \pi_N = \nu_1\} \ ,}$$

- where:
  - $\pi^0 \in \mathcal{P}((\mathbb{R}^d)^N)$ is a **reference measure**.
  - $\nu_i \in \mathcal{P}(\mathbb{R}^d)$ are **extremal conditions** $i \in \{0, 1\}$.
- $\pi^\star$ is the **"closest"** measure to π^0 such that its **initial** and **terminal** conditions are fixed.
- The problem is said to be **dynamical** because it is defined on the **state-space** $(\mathbb{R}^d)^{N+1}$.
- We will later see a **static** formulation.

Generative modeling and Schrödinger bridge

- Recall that the **dynamical** formulation is given by

$$\boxed{\pi^\star = \arg\min\{\mathrm{KL}(\pi|\pi^0) \;:\; \pi \in \mathcal{P}((\mathbb{R}^d)^N), \pi_0 = \nu_0,\; \pi_N = \nu_1\}\,,}$$

- Link with **generative modeling**:
  - $\pi^0 \in \mathcal{P}((\mathbb{R}^d)^N)$ is the discretization of the **Ornstein-Ulhenbeck** process.
  - ν_0 is the **data distribution**.
  - $\nu_1 = \mathrm{N}(0, \mathrm{Id})$ is the **easy-to-sample** distribution.
- Contrary to classical SGM we do not require $\pi_N \approx \nu_1$ ($N \gg 1$ in vanilla SGM).
- In **Schrödinger bridges** this condition is **imposed**.

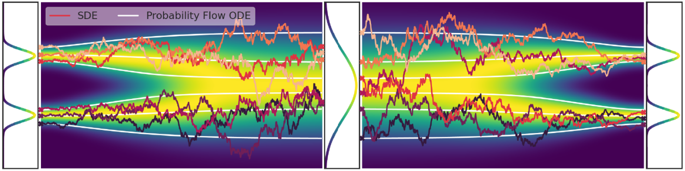


Figure 4: Noising and generative processes in SGM. Image extracted from Song et al. (2020).

The continuous dynamical setting

- The **discrete dynamical** formulation is given by

$$\boxed{\pi^\star = \arg\min\{\mathrm{KL}(\pi|\pi^0) \ : \ \pi \in \mathcal{P}((\mathbb{R}^d)^N), \pi_0 = \nu_0, \ \pi_N = \nu_1\} \ ,}$$

- We can also state the problem in **continuous** time:
    - We replace $\mathcal{P}((\mathbb{R}^d)^N)$ by $\mathcal{P}(\mathcal{C})$.
    - $\mathcal{C} = \mathrm{C}([0, T], \mathbb{R}^d)$, with the topology given by $\|\cdot\|_\infty$.
    - Technical point: $\mathcal{C}$ is a **Polish space**.
- The **continuous dynamical** formulation is given by

$$\boxed{\Pi^\star = \arg\min\{\mathrm{KL}(\Pi|\Pi^0) \ : \ \Pi \in \mathcal{P}(\mathcal{C}), \Pi_0 = \nu_0, \ \Pi_T = \nu_1\} \ ,}$$

    - $\Pi^0 \in \mathcal{P}((\mathbb{R}^d)^N)$ is a **reference measure**.
    - $\nu_i \in \mathcal{P}(\mathbb{R}^d)$ are **extremal conditions** $i \in \{0, 1\}$.
- The **discrete formulation** can be seen as a discretization of the **continuous formulation**.

The static setting

- We have seen two different **dynamical** settings:
  - The **discrete** formulation.
  - The **continuous** formulation.
- We now present the **static** formulation.

$$\boxed{\pi^{\star,s} = \arg\min\{\mathrm{KL}(\pi|\pi^0_{0,N}) \ : \ \pi \in \mathcal{P}((\mathbb{R}^d)^2), \pi_0 = \nu_0, \ \pi_1 = \nu_1\} \ ,}$$

- where:
  - $\pi^0_{0,N} \in \mathcal{P}((\mathbb{R}^d)^2)$ is a **reference measure**.
  - $\nu_i \in \mathcal{P}(\mathbb{R}^d)$ are **extremal conditions** $i \in \{0, 1\}$.
  - This amounts to finding the **coupling** the "closest" to $\pi^0_{0,N}$ w.r.t. the Kullback-Leibler divergence.
- We will see that these formulations are **equivalent**, when $\pi^0_{0,N}$ is the marginal of π^0 at time $\{0, N\}$.

Basics on disintegration

- Let X, Y be **Polish spaces**.
- Let $\mathbb{P} \in \mathcal{P}(\mathsf{X})$ and $\phi : \mathsf{X} \to \mathsf{Y}$ a measurable mapping.
- Let $\mathbb{P}_\phi = \phi_\# \mathbb{P}$ (in particular, $\mathbb{P}_\phi \in \mathcal{P}(\mathsf{Y})$).
- There exists $\mathrm{R}_{\mathbb{P},\phi}$ a **Markov kernel**, i.e.
  - For any $y \in \mathsf{Y}$, $\mathrm{R}_{\mathbb{P},\phi}(y, \cdot) \in \mathcal{P}(\mathsf{X})$.
  - For any $\mathsf{A} \in \mathcal{B}(\mathsf{X})$, $\mathrm{R}_{\mathbb{P},\phi}(\cdot, \mathsf{A}) : \mathsf{Y} \to [0, 1]$ is measurable.
  - We have the **disintegration formula**

$$\boxed{\mathbb{P}(\mathsf{A}) = \int_{\mathsf{Y}} \mathrm{R}_{\mathbb{P},\phi}(y, \mathsf{A}) \mathrm{d}\mathbb{P}_\phi(y) \ .}$$

- Example: if $\mathsf{X} = \mathbb{R}^d \times \mathbb{R}^d$, $\mathsf{Y} = \mathbb{R}^d$ and $\phi(x_1, x_2) = x_1$. Assume that $\mathbb{P}$ admits a positive density w.r.t. the Lebesgue measure. In this case:
  - $\mathbb{P}_\phi$ is the **marginal** w.r.t. the first component with density $p(x_1)$
  - $\mathrm{R}_{\mathbb{P},\phi}$ is the **conditional** probability of the second component given the first with density $p(x_2|x_1)$.
  - The previous formula then simply states that $p(x_1, x_2) = p(x_2|x_1)p(x_1)$.

The chain rule formula

- Using the **disintegration of the measure** we have the following result.

Chain rule for the Kullback-Leibler divergence Léonard (2014)

- Let X, Y be **Polish spaces**.
- Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathsf{X})$, $\phi : \mathsf{X} \to \mathsf{Y}$ measurable. Then, we have

$$\boxed{\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \mathrm{KL}(\mathbb{P}_\phi|\mathbb{Q}_\phi) + \int_{\mathsf{Y}} \mathrm{KL}(\mathrm{R}_{\mathbb{P},\phi}|\mathrm{R}_{\mathbb{Q},\phi}) \mathrm{d}\mathbb{P}_\phi(y) \ .}$$

- Proof with positive densities (assuming that all quantities are finite) and $\phi(x_0, x_1) = x_0$

$$\begin{aligned}\mathrm{KL}(\mathbb{P}|\mathbb{Q}) &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \log(p(x_0, x_1)/q(x_0, x_1)) p(x_0, x_1) \mathrm{d}x_0 \mathrm{d}x_1 \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \log(p(x_0) p(x_1|x_0)/\{q(x_0) q(x_1|x_0)\}) p(x_0, x_1) \mathrm{d}x_0 \mathrm{d}x_1 \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} \log(p(x_0)/q(x_0)) p(x_0) \mathrm{d}x_0 \\ &\quad + \int_{\mathbb{R}^d} (\int_{\mathbb{R}^d} \log(p(x_1|x_0)/q(x_1|x_0)) p(x_1|x_0) \mathrm{d}x_1) p(x_0) \mathrm{d}x_0 \ .\end{aligned}$$

- This formula is **key** for the analysis of Schrödinger bridges.

Equivalence between static and dynamic (1/2)

- Recall the **discrete dynamical** formulation

$$\boxed{\pi^\star = \arg\min\{\mathrm{KL}(\pi|\pi^0) \ : \ \pi \in \mathcal{P}((\mathbb{R}^d)^N), \pi_0 = \nu_0, \ \pi_N = \nu_1\} \ ,}$$

- Recall the **static** formulation

$$\boxed{\pi^{\star,\mathrm{s}} = \arg\min\{\mathrm{KL}(\pi|\pi^0_{0,N}) \ : \ \pi \in \mathcal{P}((\mathbb{R}^d)^2), \pi_0 = \nu_0, \ \pi_1 = \nu_1\} \ ,}$$

- Apply the **chain rule** formula with $\phi(x_{0:N}) = (x_0, x_N)$,

$$\mathrm{KL}(\pi|\pi^0) = \mathrm{KL}(\pi_{0,N}|\pi^0_{0,N}) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathrm{KL}(\mathrm{R}_{\pi,\phi}|\mathrm{R}_{\pi^0,\phi})\mathrm{d}\pi_{0,N}(x_0, x_N) \ .$$

- To minimize the RHS term under $\pi_0 = \nu_0$ and $\pi_N = \nu_1$, we can set $\mathrm{R}_{\pi,\phi} = \mathrm{R}_{\pi^0,\phi}$.
- We have that $\pi^\star = \pi^\star_{0,N}\mathrm{R}_{\pi^0,\phi}$, with $\pi^\star_{0,N}$ solution of the **static problem**, i.e.

$$\boxed{\pi^\star = \pi^{\star,\mathrm{s}}\mathrm{R}_{\pi^0,\phi} \ .}$$

Equivalence between static and dynamic (2/2)

- This equivalence gives us a way to sample from $\pi^\star$:
  - **Sample** (x_0, x_N) from $\pi^{\star,s}$.
  - Sample from the **bridge** associated with π^0 and **extremal conditions** x_0, x_N.

Video extracted from a tweet by Lenaïc Chizat.

The potential approach

Information geometry

- We start with a **projection** result by Csiszár (1975).

Projection for the Kullback-Leibler divergence Csiszár (1975)

- Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and $\mathsf{F} = \{f_i \;:\; i \in \mathsf{I}\}$ a set of real-valued measurable functions.
- Let $\mathbb{P}^0 \in \mathcal{P}(\mathsf{X})$ and let $\mathcal{P}_\mathsf{F}(\mathsf{X}) = \{\mathbb{P} \in \mathcal{P}(\mathsf{X}) \;:\; \sup_\mathsf{F} \int_\mathsf{X} |f(x)| \mathrm{d}\mathbb{P}(x) < +\infty\}$.
- Let $\mathsf{A} = \{a_i \;:\; i \in \mathsf{I}\}$ and

$$\boxed{\mathcal{P}_{\mathsf{F},\mathsf{A}}(\mathsf{X}) = \{\mathbb{P} \in \mathcal{P}_\mathsf{F}(\mathsf{X}) \;:\; \textstyle\int_\mathsf{X} f_i(x) \mathrm{d}\mathbb{P}(x) = a_i, \text{ for any } i \in \mathsf{I}\}\ .}$$

- Assume that there exists $\mathbb{Q} \in \mathcal{P}_{\mathsf{F},\mathsf{A}}$ such that $\mathrm{KL}(\mathbb{Q}|\mathbb{P}^0) < +\infty$.
- Then $\mathbb{P}^\star = \arg\min\{\mathrm{KL}(\mathbb{P}|\mathbb{P}^0) \;:\; \mathbb{P} \in \mathcal{P}_{\mathsf{F},\mathsf{A}}(\mathsf{X})\}$ exists is unique and there exist:
  - $g \in \bar{\mathsf{F}}$ (closure in $\mathrm{L}^1(\mathbb{P}^\star)$), $C \geq 0$,
  - N with $\mathbb{P}^\star(\mathsf{N}) = 0$,
- such that for any $x \in \mathsf{N}$, $(\mathrm{d}\mathbb{P}^\star/\mathrm{d}\mathbb{P}^0)(x) = 0$ and for any $x \in \mathsf{X}\backslash\mathsf{N}$

$$\boxed{(\mathrm{d}\mathbb{P}^\star/\mathrm{d}\mathbb{P}^0)(x) = C\exp[g(x)]\ .}$$

Exponential model

- A first case of application of the theorem: **maximum entropy models**.
- In this case $|\mathsf{I}| < +\infty$ (**finite** family of constraints).
- We get that (if $\mathbb{P}^0 \ll \mathbb{P}^\star$) for any $x \in \mathsf{X}$

$$\boxed{(\mathrm{d}\mathbb{P}^\star/\mathrm{d}\mathbb{P}^0)(x) = \exp[\langle \theta^\star, f(x)\rangle] / \int_{\mathsf{X}} \exp[\langle \theta^\star, f(\tilde{x})\rangle] \mathrm{d}\mathbb{P}^0(\tilde{x}) \ .}$$

- In the previous lectures we showed that $\theta^\star \in \mathbb{R}^{|\mathsf{I}|}$ could be interpreted as **dual parameters**.
- In particular, under mild conditions, they can be obtain by solving the following optimization problem

$$\boxed{\theta^\star = \arg\min\{\log(\textstyle\int_{\mathsf{X}} \exp[\langle \theta, f(\tilde{x})\rangle] \mathrm{d}\mathbb{P}^0(\tilde{x})) \ : \ \theta \in \mathbb{R}^{|\mathsf{I}|}\} \ .}$$

- We obtain a family of (linear) **exponential models** (macrocanonical models).

Schrödinger Bridges as projections

- We are going to see that the **static** Schrödinger Bridge problem can be seen as a **projection**.
- We set the following:
  - $\mathsf{X} = (\mathbb{R}^d)^2$, $\mathbb{P}^0 = \pi^0_{0,N} \in \mathcal{P}(\mathsf{X})$.
  - $\mathsf{F} = \{f_0 \oplus f_1 \ : \ f_i \in \mathrm{L}^1(\nu_i), \ i \in \{0,1\}\}$.
  - $\mathsf{A} = \{\int_{\mathbb{R}^d} f_0(x)\mathrm{d}\nu_0(x) + \int_{\mathbb{R}^d} f_1(x)\mathrm{d}\nu_1(x) \ : \ f_i \in \mathrm{L}^1(\nu_i), \ i \in \{0,1\}\}$.
- We obtain that $\mathcal{P}_{\mathsf{F},\mathsf{A}}(\mathsf{X}) = \{\pi \in \mathcal{P}((\mathbb{R}^d)^2) \ : \ \pi_0 = \nu_0, \ \pi_1 = \nu_1\}$.
- Hence, we get that

$$\boxed{\arg\min\{\mathrm{KL}(\pi|\pi^0_{0,N}) \ : \ \pi_0 = \nu_0, \ \pi_1 = \nu_1\} = \arg\min\{\mathrm{KL}(\pi|\mathbb{P}^0) \ : \ \pi \in \mathcal{P}_{\mathsf{F},\mathsf{A}}(\mathsf{X})\} \ .}$$

- Assuming that $\mathrm{KL}(\nu_0 \otimes \nu_1|\mathbb{P}^0) < +\infty$ we can apply the **projection theorem** Csiszár (1975) and $\pi^{\star,\mathrm{s}} = \arg\min\{\mathrm{KL}(\pi|\pi^0_{0,N}) \ : \ \pi_0 = \nu_0, \ \pi_1 = \nu_1\}$ exists is unique and there exist:
  - $g \in \bar{\mathsf{F}}$ (closure in $\mathrm{L}^1(\mathbb{P}^\star)$), $C \geq 0$,
  - N with $\mathbb{P}^\star(\mathsf{N}) = 0$,
- such that for any $(x,y) \in \mathsf{N}$, $(\mathrm{d}\pi^{\star,\mathrm{s}}/\mathrm{d}\pi^0_{0,N})(x,y) = 0$ and for any $(x,y) \in \mathsf{X}\backslash\mathsf{N}$

$$\boxed{(\mathrm{d}\pi^{\star,\mathrm{s}}/\mathrm{d}\pi^0_{0,N})(x,y) = C\exp[g(x,y)] \ .}$$

Optimal potential (1/2)

- Assuming that $\mathrm{KL}(\nu_0 \otimes \nu_1 | \mathbb{P}^0) < +\infty$ we have that there exist:
  - $g \in \bar{\mathsf{F}}$ (closure in $\mathrm{L}^1(\mathbb{P}^\star)$), $C \geq 0$,
  - N with $\mathbb{P}^\star(\mathsf{N}) = 0$,
- such that for any $(x, y) \in \mathsf{N}$, $(\mathrm{d}\pi^{\star,s} / \mathrm{d}\pi^0_{0,N})(x, y) = 0$ and for any $(x, y) \in \mathsf{X} \backslash \mathsf{N}$

$$\boxed{(\mathrm{d}\pi^{\star,s} / \mathrm{d}\pi^0_{0,N})(x, y) = C \exp[g(x, y)] \ .}$$

- What is the **form** of g?

Optimal potential Rüschendorf and Thomsen (1993)

- Assume that $\mathrm{KL}(\nu_0 \otimes \nu_1 | \pi^0_{0,N}) < +\infty$, then there exists g_0, g_1 measurable and N with $\pi^{\star,s}(\mathsf{N}) = 0$ such that for any $(x, y) \in \mathsf{N}$, $(\mathrm{d}\pi^{\star,s} / \mathrm{d}\pi^0)(x, y) = 0$. In addition, for any $(x, y) \in (\mathbb{R}^d)^2 \backslash \mathsf{N}$ we have

$$\boxed{(\mathrm{d}\pi^{\star,s} / \mathrm{d}\pi^0_{0,N})(x, y) = C \exp[g_0(x)] \exp[g_1(y)] \ .}$$

- We have a **factorized** structure.
- We have shown that under **mild conditions** this structure is **necessary**.

Optimal potential (2/2)

- Under a slightly **stronger assumption** we have the following theorem.

Optimal potential Nutz (2021)

- Assume that $\mathrm{KL}(\nu_0 \otimes \nu_1 | \pi^0_{0,N}) < +\infty$ and that $\pi^0_{0,N} \ll \nu_0 \otimes \nu_1$.
- Then $\pi^{\star,s} = \arg\min\{\mathrm{KL}(\pi|\pi^0_{0,N}) \ : \ \pi_0 = \nu_0, \ \pi_1 = \nu_1\}$ exists is unique and there exist g_0, g_1 such that for any $x, y \in \mathbb{R}^d$

$$\boxed{(\mathrm{d}\pi^{\star,s}/\mathrm{d}\pi^0)(x,y) = \exp[g_0(x) + g_1(y)] / \int_{(\mathbb{R}^d)^2} \exp[g_0(\tilde{x}) + g_1(\tilde{y})] \mathrm{d}\pi^0(\tilde{x}, \tilde{y}) \ .}$$

- If there exists π, g_0, g_1 such that for any $x, y \in \mathbb{R}^d$

$$\boxed{(\mathrm{d}\pi/\mathrm{d}\pi^0)(x,y) = \exp[g_0(x) + g_1(y)] / \int_{(\mathbb{R}^d)^2} \exp[g_0(\tilde{x}) + g_1(\tilde{y})] \mathrm{d}\pi^0(\tilde{x}, \tilde{y}) \ ,}$$

and $\pi_0 = \nu_0$, $\pi_1 = \nu_1$, then $\pi = \pi^{\star,s}$.

- How to find the **potentials** g_0, g_1?
- These potentials satisfy a system of **coupled equations**.
- A modern overview of **properties of Schrödinger bridges** Nutz (2021).

Schrödinger equations

- Under mild assumptions we have that

$$\boxed{(\mathrm{d}\pi^{\star,s}/\mathrm{d}\pi^0)(x,y) = \exp[g_0(x) + g_1(y)]\ .}$$

- We recall that such a **decomposition** is **necessary** and **sufficient**.
- **Agreement** with the marginals: for any $\mathsf{A}, \mathsf{B} \in \mathcal{B}(\mathbb{R}^d)$

$$\nu_0(\mathsf{A}) = \int_{\mathsf{A}\times\mathbb{R}^d} \exp[g_0(x) + g_1(y)]\mathrm{d}\pi^0(x,y)\ ,$$
$$\nu_1(\mathsf{B}) = \int_{\mathbb{R}^d\times\mathsf{B}} \exp[g_0(x) + g_1(y)]\mathrm{d}\pi^0(x,y)\ .$$

- These equations are called the **Schrödinger equations**.
- This a **coupled** system of equations.
- We will see that the **Sinkhorn algorithm** iteratively solves these equations.
- First proof of existence of such potentials by Fortet (see Léonard (2019) for a recent presentation and survey).

Discrete Dynamic potentials and twisted kernels

- Under mild assumptions we have

$$\boxed{(\mathrm{d}\pi^{\star,\mathrm{s}}/\mathrm{d}\pi^0_{0,N})(x,y) = f_0(x) f_1(y)\ .}$$

- We also have $\pi^\star = \pi^{\star,\mathrm{s}} \mathrm{R}_{\pi^0,\phi}$, with $\phi(x_{0:N}) = (x_0, x_N)$.
- **Combining** these two results we get that for any $x_{0:N} \in (\mathbb{R}^d)^{N+1}$

$$\boxed{(\mathrm{d}\pi^\star/\mathrm{d}\pi^0)(x_{0:N}) = f_0(x_0) f_N(x_N)\ .}$$

- Denote $f_0^0 = f_0$, $f_1^N = f_1$ and define for any $\ell \in \{1, \ldots, N\}$

$$f_0^\ell(x_\ell) = \int_{\mathbb{R}^d} f_0^{\ell-1}(x_{\ell-1}) \pi^0_{\ell|\ell-1}(x_\ell|x_{\ell-1}) \mathrm{d}x_{\ell-1}\ ,$$
$$f_1^\ell(x_\ell) = \int_{\mathbb{R}^d} f_1^{\ell+1}(x_{\ell+1}) \pi^0_{\ell+1|\ell}(x_{\ell+1}|x_\ell) \mathrm{d}x_{\ell+1}\ .$$

- We get that for any $k, \ell \in \{0, \ldots, N\}$ with $k \leq \ell$

$$\boxed{(\mathrm{d}\pi^\star_{k:\ell}/\mathrm{d}\pi^0_{k:\ell})(x_{k:\ell}) = f_0^k(x_k) f_1^\ell(x_\ell)\ .}$$

- In particular, we get that for any $k \in \{0, \ldots, N-1\}$

$$\boxed{\pi^\star(x_{k+1}|x_k) = \pi^0(x_{k+1}|x_k) f_1^{k+1}(x_{k+1}) / f_1^k(x_1^k)\ .}$$

- We obtain **twisted kernels**. This is a discrete **Doob *h*-transform**.

Interlude on Doob h-transform (1/2)

- Let $\{\mathrm{P}_{t|s}\}_{s,t\in[0,T],s\leq t}$ a **semi-group** with **infinitesimal generator** $\{\mathscr{A}_u\}_{u\in[0,T]}$, i.e. for any $s, t \in [0, T]$, $s \leq t$ and $\varphi \in \mathrm{C}_c(\mathbb{R}^d)$

$$\boxed{\int_{\mathbb{R}^d} \varphi(x_t)\mathrm{d}\mathrm{P}_{t|s}(x_t, \mathbf{X}_s) = \mathbb{E}[\varphi(\mathbf{X}_t)\,|\mathbf{X}_s] = \int_s^t \mathbb{E}[\mathscr{A}_u(\varphi)(\mathbf{X}_u)\,|\mathbf{X}_s]\mathrm{d}u\ .}$$

- Let $f \in \mathrm{C}^\infty([0, T] \times \mathbb{R}^d)$ such that $\partial_t f_t = -\mathscr{A}_t(f_t)$ (**backward Kolmogorov** equation).
- Define the **twisted** generators $\{\hat{\mathrm{P}}_{t|s}\}_{s,t\in[0,T],s\leq t}$ such that

$$\boxed{\mathrm{d}\hat{\mathrm{P}}_{t|s}(x_t, x_s) = \mathrm{d}\mathrm{P}_{t|s}(x_t, x_s)f_t(x_t)/f_s(x_s)\ .}$$

- Then, $\{\mathrm{P}_{t|s}\}_{s,t\in[0,T],s\leq t}$ a **semi-group** with **infinitesimal generator** $\{\hat{\mathscr{A}}_u\}_{u\in[0,T]}$ such that

$$\boxed{\hat{\mathscr{A}}_u(\varphi) = \mathscr{A}_u(\varphi) + \langle\nabla\varphi, \nabla\log(f_u)\rangle\ .}$$

- This is assuming that $\mathscr{A}_u(\varphi) = \langle b_u, \varphi\rangle + (1/2)\Delta\varphi$.

Interlude on Doob h-transform (2/2)

- Let us prove this fact. Let $s, t \in [0, T]$ with $t \geq s$

$$\mathbb{E}[\varphi(\hat{\mathbf{X}}_t) \,|\hat{\mathbf{X}}_s] = \mathbb{E}[\varphi(\mathbf{X}_t) f_t(\mathbf{X}_t) \,|\mathbf{X}_s]/f_s(\mathbf{X}_s) \ .$$

- We have

$$\begin{aligned}
\mathbb{E}[\varphi(\mathbf{X}_t) f_t(\mathbf{X}_t) \,|\mathbf{X}_s] - \varphi(\mathbf{X}_s) f_s(\mathbf{X}_s) &= \int_s^t \mathbb{E}[\{\mathscr{A}_u(\varphi f_u) + \varphi \partial_u f_u\}(\mathbf{X}_u) \,|\mathbf{X}_s] \mathrm{d}u \\
&= \int_s^t \mathbb{E}[\{\mathscr{A}_u(\varphi) f_u + \langle \nabla \varphi, \nabla f_u \rangle + \varphi \mathscr{A}_u(f_u) + \varphi \partial_u f_u\}(\mathbf{X}_u) \,|\mathbf{X}_s] \mathrm{d}u \\
&= \int_s^t \mathbb{E}[\{\mathscr{A}_u(\varphi) f_u + \langle \nabla \varphi, \nabla f_u \rangle\}(\mathbf{X}_u) \,|\mathbf{X}_s] \mathrm{d}u \\
&= \int_s^t \mathbb{E}[\{\mathscr{A}_u(\varphi) + \langle \nabla \varphi, \nabla \log(f_u) \rangle\}(\mathbf{X}_u) f_u(\mathbf{X}_u) \,|\mathbf{X}_s] \mathrm{d}u \\
&= \int_s^t \mathbb{E}[\hat{\mathscr{A}}_u(\varphi)(\mathbf{X}_u) f_u(\mathbf{X}_u) \,|\mathbf{X}_s] \mathrm{d}u \\
&= f_s(\mathbf{X}_s) \int_s^t \mathbb{E}[\hat{\mathscr{A}}_u(\varphi)(\hat{\mathbf{X}}_u) \,|\hat{\mathbf{X}}_s] \mathrm{d}u \ .
\end{aligned}$$

- Hence, we get that

$$\boxed{\mathbb{E}[\varphi(\hat{\mathbf{X}}_t) \,|\hat{\mathbf{X}}_s] = \varphi(\hat{\mathbf{X}}_s) + \int_s^t \mathbb{E}[\hat{\mathscr{A}}_u(\varphi)(\hat{\mathbf{X}}_u) \,|\hat{\mathbf{X}}_s] \mathrm{d}u \ .}$$

Continuous dynamic potentials

- Back to the **Schrödinger bridge** problem.
- We consider the **continuous** dynamic problem

$$\boxed{\Pi^\star = \arg\min\{\mathrm{KL}(\Pi|\Pi^0) \ : \ \Pi \in \mathcal{P}(\mathcal{C}), \Pi_0 = \nu_0, \ \Pi_T = \nu_1\} \ ,}$$

- Under mild assumptions, we have that for any $\omega \in \mathcal{C}$

$$\boxed{(\mathrm{d}\Pi^\star / \mathrm{d}\Pi^0)(\omega) = f_0(\omega_0) f_T(\omega_T) \ .}$$

- Define for any $t \in [0, T]$

$$f_0^t(\omega_t) = \int_{\mathbb{R}^d} f_0(\omega_0) \Pi^0(\omega_t|\omega_0) \mathrm{d}\omega_0 \ ,$$
$$f_t^t(\omega_t) = \int_{\mathbb{R}^d} f_T(\omega_T) \Pi^0(\omega_T|\omega_t) \mathrm{d}\omega_T \ .$$

- If we denote $\mathrm{P}_{t|s}$ the **semi-group** associate with Π^0 then $\hat{\mathrm{P}}_{t|s}$, the semi-group associated with $\Pi^\star$ is the **Doob h-transform** with twist $\{f_T^t\}_{t \in [0,T]}$.
- In particular if Π^0 is associated with $\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t$ then $\Pi^\star$ is associated with $\mathrm{d}\mathbf{X}_t = \{b(\mathbf{X}_t) + \nabla \log f_T^t(\mathbf{X}_t)\}\mathrm{d}t + \mathrm{d}\mathbf{B}_t$.
- This formulation can be linked with **stochastic control** Dai Pra (1991).

A quick summary

- The **Schrödinger bridge** problem is a **theoretically grounded** framework for **generative modeling**.
- This problem can be formulated in a **dynamical** or **static** setting.
- We show the existence of **potentials** for the solutions.
- These potentials correspond to a **twisting dynamic** in the discrete and continuous-time Schrödinger bridge problem.
- In what follows, we draw a link with **Entropic Regularized Optimal Transport**.

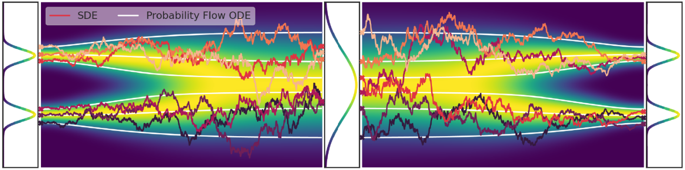


Figure 5: Noising and generative processes in SGM. Image extracted from Song et al. (2020).

Regularized Optimal Transport

Basics on Optimal transport

- Recall that **Optimal transport** corresponds to finding the solution of

$$\boxed{\Lambda^\star = \arg\min\{\int_{(\mathbb{R}^d)^2} c(x,y)\mathrm{d}\Lambda(x,y) \ : \ \Lambda_0 = \nu_0, \ \Lambda_1 = \nu_1\} \ .}$$

  - c is the **cost function**.
  - $\Lambda^\star$ is the **optimal coupling**.
- If $c(x,y) = (1/2)\|x-y\|^2$ and under mild regularity assumptions on ν_0, ν_1 this problem coincides with the **Brenier problem**

$$\boxed{T^\star = \arg\min\{\int_{\mathbb{R}^d} c(x, T(x))\mathrm{d}\nu_0(x) \ : \ T \in \mathrm{L}^2(\nu_0), \ T_\# \nu_0 = \nu_1\} \ .}$$

- We get that $\Lambda^\star = (\mathrm{Id}, T)_\# \nu_0$.

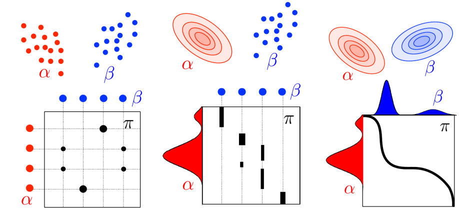


Figure 6: Examples of Optimal Transport. Image extracted from Peyré et al. (2019).

Entropic Regularized Optimal Transport

- **Entropic Regularized Optimal Transport**

$$\boxed{\Lambda_\varepsilon^\star = \arg\min\{\int_{(\mathbb{R}^d)^2} c(x,y)\mathrm{d}\Lambda(x,y) + \varepsilon \mathrm{KL}(\Lambda|\pi_0 \otimes \pi_1) \ : \ \Lambda_0 = \nu_0, \ \Lambda_1 = \nu_1\} \ .}$$

  - $\pi_0, \pi_1 \in \mathcal{P}(\mathbb{R}^d)$.
  - The solution is the same if π_0, π_1 replaced by $\tilde{\pi}_0, \tilde{\pi}_1 \in \mathcal{P}(\mathbb{R}^d)$, see (Peyré et al., 2019, Proposition 4.2).
- This regularization allows for **fast algorithms** in discrete state spaces such as the **Sinkhorn algorithm**.
- Entropic optimal transport plans are **more diffuse**.

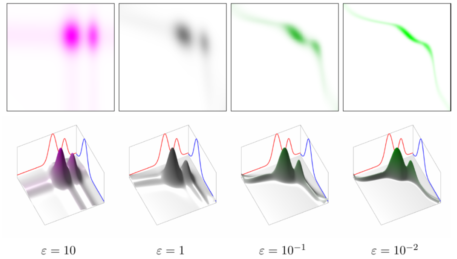


Figure 7: Entropic regularized OT. Image extracted from Peyré et al. (2019).

From Schrödinger Bridge to OT (1/2)

- Recall the **static formulation**

$$\boxed{\pi^{\star,s} = \arg\min\{\mathrm{KL}(\pi|\pi^0_{0,N}) \;:\; \pi \in \mathcal{P}((\mathbb{R}^d)^2), \pi_0 = \nu_0,\ \pi_1 = \nu_1\}\ ,}$$

- Assume that the **reference measure** is of the form

$$\boxed{\mathrm{d}\pi^0_{0,N}(x,y) = (2\pi\varepsilon)^{-d/2}\exp[-\|x-y\|^2/(2\varepsilon)]\mathrm{d}\nu_0(x)\mathrm{d}y\ .}$$

- Note that in the **continuous** setting with is equivalent to choosing a reference measure Π^0 associated with $(\mathbf{B}_{(\varepsilon/T)t})_{t\in[0,T]}$, a time-rescaled **Brownian motion**.
- Let $\pi \in \mathcal{P}((\mathbb{R}^d)^2)$ with $\pi_0 = \nu_0$ and $\pi_1 = \nu_1$. Using the **chain-rule** with $\phi(x,y) = x$ we have

$$\boxed{\mathrm{KL}(\pi|\pi^0_{0,N}) = \mathrm{KL}(\nu_0|\pi^0_{0,N}) + \int_{\mathbb{R}^d}\mathrm{KL}(\mathrm{R}_{\pi,\phi}|\mathrm{R}_{\pi^0_{0,N},\phi})\mathrm{d}\nu_0(x)\ .}$$

- This can be rewritten as

$$\boxed{\mathrm{KL}(\pi|\pi^0_{0,N}) = \int_{\mathbb{R}^d\times\mathbb{R}^d}\log((\mathrm{dR}_{\pi,\phi}/\mathrm{dLeb})(y|x)(2\pi\varepsilon)^{d/2}\exp[\|x-y\|^2/(2\varepsilon)])\mathrm{d}\pi(x,y)\ .}$$

From Schrödinger Bridge to OT (2/2)

- We have

$$\boxed{\mathrm{KL}(\pi|\pi^0_{0,N}) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \log((\mathrm{dR}_{\pi,\phi}/\mathrm{dLeb})(y|x)(2\pi\varepsilon)^{d/2}\exp[\|x-y\|^2/(2\varepsilon)])\mathrm{d}\pi(x,y)\ .}$$

- This can again be written as

$$\boxed{\mathrm{KL}(\pi|\pi^0_{0,N}) = (2\varepsilon)^{-1}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^2\,\mathrm{d}\pi(x,y) + \mathrm{KL}(\pi|\nu_0\otimes\nu_1 + C_\varepsilon\ .)}$$

- Therefore, we have that a **Schrödinger bridge** with reference measure $(\mathbf{B}_{(\varepsilon/T)t})_{t\in[0,T]}$ is equivalent (in its **static formulation**) to the **ε-entropic regularized OT**.

Video extracted from a tweet by Lenaïc Chizat.

A limit theorem

- The following result from Mikami (2004) shows the connection between **Schrödinger bridges** and **Optimal Transport**.

Limits of Schrödinger bridge Mikami (2004)

- Assume that the reference measure is associated with $(\mathbf{B}_{(\varepsilon/T)t})_{t \in [0,T]}$.
- Denote $\pi_\varepsilon^{\star,\mathrm{s}}$ the solution of the **static** Schrödinger bridge.
- Under mild assumptions we have

$$\boxed{\lim_{\varepsilon \to 0} \varepsilon \mathrm{KL}(\pi_\varepsilon^{\star,\mathrm{s}} | \pi_{0,N}^{0,\varepsilon}) = \mathbf{W}_2^2(\nu_0, \nu_1)\ .}$$

- We have that $\lim_{\varepsilon \to 0} \pi_\varepsilon^{\star,\mathrm{s}} = (\mathrm{Id}, T)_{\#}\nu_0$, the Optimal Transport plan w.r.t. the **Wasserstein distance** of order 2.

- What happens if the reference dynamic is *not* a **Brownian motion**?
- If the dynamics is an **Ornstein-Ulhenbeck** process then we still get a **quadratic cost** but instead of $(1/2)\|x - y\|^2$ we get $(1/2)\|x - \mathrm{e}^{-T} y\|^2$.
- Correlate with the intuition that (in the Ornstein-Ulhenbeck setting) when $T \to +\infty$, the Schrödinger bridge is closer to $\nu_0 \otimes \nu_1$.

The Sinkhorn algorithm

Outline of the section

- So far we have introduced the **Schrödinger bridge** in their **static** and **dynamic** formulations.
- We have seen a **potential formulation** and a link with **entropic regularized OT**.
- Most of the time Schrödinger bridges are **untractable**. How can we approximate them?
- We are going to study an **efficient algorithm** to approximate the potentials.

- In this section:
  - Introduction of the **Sinkhorn algorithm**.
  - **Geometric** convergence in the **compact** setting.
  - **Convergence** results in the **non-compact** setting.

Introduction of the algorithm (1/2)

- Recall the **Schrödinger equations**: for any $\mathrm{A}, \mathrm{B} \in \mathcal{B}(\mathbb{R}^d)$ we have

$$\nu_0(\mathrm{A}) = \int_{\mathrm{A} \times \mathbb{R}^d} \exp[g_0(x) + g_1(y)] \mathrm{d}\pi^0(x, y) \;,$$
$$\nu_1(\mathrm{B}) = \int_{\mathbb{R}^d \times \mathrm{B}} \exp[g_0(x) + g_1(y)] \mathrm{d}\pi^0(x, y) \;.$$

- We want to solve these equations in g_0, g_1. In what follows we overload the notations and denote ν_0, ν_1, π^0 the **density** w.r.t. the Lebesgue measure of these probabilities. The **Schrödinger equations** become

$$f_0(x) = \nu_0(x)\left(\int_{\mathbb{R}^d} f_1(y)\pi^0(x, y)\mathrm{d}y\right)^{-1} \;,$$
$$f_1(y) = \nu_1(y)\left(\int_{\mathbb{R}^d} f_0(x)\pi^0(x, y)\mathrm{d}x\right)^{-1} \;.$$

- Start with $f_0^0 = f_1^0 = 1$ and define

$$f_1^{n+1}(y) = \nu_1(y)\left(\int_{\mathbb{R}^d} f_0^n(x)\pi^0(x, y)\mathrm{d}x\right)^{-1} \;,$$
$$f_0^{n+1}(x) = \nu_0(x)\left(\int_{\mathbb{R}^d} f_1^{n+1}(y)\pi^0(x, y)\mathrm{d}y\right)^{-1} \;.$$

- **Iteratively** solve the **system of equations** looking for a **fixed point**.
- This is the **Sinkhorn** algorithm, also sometimes called **Iterative Proportional Fitting** (IPF).

Introduction of the algorithm (2/2)

- We obtain a **sequence of measures** $\pi^{2n}(x,y) = \pi^0(x,y) f_0^n(x) f_1^n(y)$ and $\pi^{2n+1}(x,y) = \pi^0(x,y) f_0^n(x) f_1^{n+1}(y)$.
- Under mild assumptions we have that

$$\pi^{2n+1} = \arg\min\{\mathrm{KL}(\pi|\pi^{2n}) \; : \; \pi \in \mathcal{P}((\mathbb{R}^d)^2),\ \pi_1 = \nu_1\} \ ,$$
$$\pi^{2n+2} = \arg\min\{\mathrm{KL}(\pi|\pi^{2n+1}) \; : \; \pi \in \mathcal{P}((\mathbb{R}^d)^2),\ \pi_0 = \nu_0\} \ .$$

- The **Sinkhorn algorithm** amounts to solving **half-bridges**.
- This is an **alternate projection** scheme w.r.t. the Kullback-Leibler divergence.

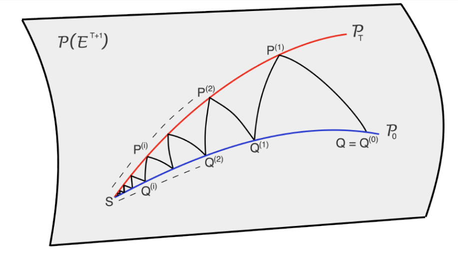


Figure 8: Solving half-bridges. Image extracted from Bernton et al. (2019).

Convergence in the compact case

Geometric convergence

- We are going to restrict ourselves to the **compact** setting.
- Instead of assuming that the distributions are supported on $\mathbb{R}^d$ we assume that they are **supported on a compact set** K.
- The results obtained so far remain true.
- We are going to prove the following theorem

Geometric convergence

- Let $(\pi^n)_{n\in\mathbb{N}}$ be the sequence obtained with the **Sinkhorn** algorithm and $\pi^\star$ the **Schrödinger bridge**. Under mild assumptions, we have

$$\boxed{\mathbf{W}_1(\pi^n, \pi^\star) \leq C\rho^n\ .}$$

- In fact the main result is a **geometric convergence** results on the potentials w.r.t. the **Hilbert-Birkhoff** metric.
- The **compactness** assumption is key.

Hilbert-Birkhoff metric

- Survey on this distance Lemmens and Nussbaum (2012); Kohlberg and Pratt (1982); Bushell (1973).
- Let $(\mathrm{E}, \|\cdot\|)$ be a normed real vector space and $\hat{\mathrm{C}}$ a **cone**:
  - $\hat{\mathrm{C}} \cap (-\hat{\mathrm{C}}) = \{0\}$.
  - $\lambda\hat{\mathrm{C}} \subset \hat{\mathrm{C}}$ for $\lambda \geq 0$.
  - $\hat{\mathrm{C}}$ is convex.
- Let C be a **part of the cone**, i.e. for any $x, y \in \mathrm{C}$, there exist $\alpha, \beta \geq 0$ such that $\alpha x - y \in \hat{\mathrm{C}}$ and $\beta y - x \in \hat{\mathrm{C}}$.
- We define for any $x, y \in \mathrm{C}$

$$M(x,y) = \inf\{\beta \geq 0 \;:\; \beta y - x \in \hat{\mathrm{C}}\} > 0 \;,$$
$$m(x,y) = \sup\{\alpha \geq 0 \;:\; x - \alpha y \in \hat{\mathrm{C}}\} \;.$$

- Finally, we define the **Hilbert-Birkhoff** metric

$$\boxed{d_H(x,y) = \log(M(x,y)/m(x,y)) \;.}$$

- $\tilde{\mathrm{D}} = \{x \in \mathrm{C} \;:\; \|x\| = 1\}$ is such that $(\tilde{\mathrm{D}}, d_H)$ is a **metric** space.

The Birkhoff contraction theorem

- Let $(\mathrm{V}, \|\cdot\|), (\mathrm{V}', \|\cdot\|')$ be two normed real vector spaces and C, C' be **convex parts** of the **cones** $\hat{\mathrm{C}}, \hat{\mathrm{C}}'$ respectively.
- Let $u : \mathrm{V} \to \mathrm{V}'$ be a linear mapping such that $u(\mathrm{C}) \subset \mathrm{C}'$.
- The **projective diameter** of u is given by

$$\boxed{\Delta(u) = \sup\{d_H(u(x), u(y)) \ : \ x, y \in \mathrm{C}, \|x\| = \|y\| = 1\} \ .}$$

- The **Birkhoff contraction ratio** of u is given by

$$\boxed{\kappa(u) = \sup\{\kappa \ : \ d_H(u(x), u(y)) \leq \kappa d_H(x, y), x, y \in \mathrm{C}\} \ .}$$

- Then, we have the following theorem.

Birkhoff contraction theorem Birkhoff (1957)

- Under the previous assumptions on u, we have

$$\boxed{\kappa(u) \leq \tanh(\Delta(u)/4) \ .}$$

In the space of continuous functions

- We have the following proposition.

Hilbert-Birkhoff in continuous spaces

Let Z be a compact space. $\mathsf{F} = [0, +\infty)^{\mathsf{Z}}$ is a cone and $\tilde{\mathsf{F}} = \mathrm{C}(\mathsf{Z}, (0, +\infty))$ is a convex part of F such that for any $\lambda > 0$, $\lambda\tilde{\mathsf{F}} \subset \tilde{\mathsf{F}}$. In addition, we have that for any $f, g \in \tilde{\mathsf{F}}$

$$d_H(f, g) = \log(\|f/g\|_\infty) + \log(\|g/f\|_\infty).$$

- $D : f \mapsto 1/f$ is an **isometry** w.r.t d_H.
- $H_g : f \mapsto (x \mapsto g(x)f(x))$ with $g \in \tilde{\mathsf{F}}$ is also an **isometry**.
- Consider the mapping $E_{k,1}(f)(x) = \int_{\mathbb{R}^d} k(x, y)f(y)\mathrm{d}y$ (with $k \in \mathrm{C}_c(\mathbb{R}^d \times \mathbb{R}^d)$. We are going to compute its **projective diameter**.

$$\boxed{\Delta(E_{k,1}) \leq 2\sup\{d_H(f, 1) \ : \ f \in \tilde{\mathsf{F}}\} = 2\sup\{\log(\sup_{\mathsf{Z}} f / \inf_{\mathsf{Z}} f) \ : \ f \in \tilde{\mathsf{F}}\} \ .}$$

- We find that $\Delta(E_{k,1}) \leq 2\log(\sup_{\mathsf{Z}\times\mathsf{Z}} k / \inf_{\mathsf{Z}\times\mathsf{Z}} k)$. Hence, we get that

$$\boxed{\kappa(E_{k,1}) \leq (\sup_{\mathsf{Z}\times\mathsf{Z}} k - \inf_{\mathsf{Z}\times\mathsf{Z}} k)/(\sup_{\mathsf{Z}\times\mathsf{Z}} k + \inf_{\mathsf{Z}\times\mathsf{Z}} k) \ .}$$

Convergence of the potentials

- Recall that the **Sinkhorn updates** are given by

$$f_1^{n+1}(y) = \nu_1(y)\left(\int_{\mathbb{R}^d} f_0^n(x)\pi^0(x, y)\mathrm{d}x\right)^{-1} ,$$
$$f_0^{n+1}(x) = \nu_0(x)\left(\int_{\mathbb{R}^d} f_1^{n+1}(y)\pi^0(x, y)\mathrm{d}y\right)^{-1} .$$

- The update is given by $H_{\nu_0} \circ D \circ E_{\pi^0,1} \circ H_{\nu_1} \circ D \circ E_{\pi^0,0}$. This is a **contraction**.
- Denoting f_0, f_1 the **Schrödinger potentials**

$$\boxed{d_H(f_0^n, f_0) + d_H(f_1^n, f_1) \leq \rho^n\{d_H(1, f_0) + d_H(1, f_1)\} .}$$

- This convergence result can be found in Chen et al. (2016).
- To obtain the $\mathbf{W}_1$ result we can proceed as in Deligiannidis et al. (2021).
- First results in Sinkhorn and Knopp (1967).

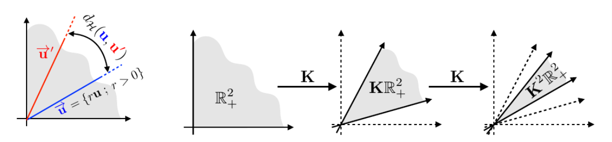


Figure 9: Contraction on cones. Image extracted from Peyré et al. (2019).

Results in the non-compact setting

Extension to non-compact setting?

- So far we have seen that the **Sinkhorn algorithm** converges **exponentially fast** on compact spaces.
- What about the **non-compact** setting?
- First, we have the following convergence result.

Convergence of the Sinkhorn algorithm Nutz (2021)

- Assume that $\int_{\mathbb{R}^d} \exp[r|\log \pi^0(x,y)|] \mathrm{d}(\nu_0 \otimes \nu_1)(x,y) < +\infty$ for some $r > 1$.
- Then $\lim_{n \to +\infty} \mathrm{KL}(\pi^n | \pi^\star) = 0$.

- The **exponential integrability** condition is replaced by an uniformly integrable condition in Ruschendorf (1995).
- We also get the convergence of the **potentials**.
- We are now going to see what kind of **quantitative rates** we can achieve.

A Pythagorean theorem

- This **Pythagorean theorem** was first established by Csiszár (1975) and is at the basis of the **projection theorem**.

Pythagorean theorem

- Let $\mathrm{C} \subset \mathcal{P}(\mathbb{R}^d)$ be a convex set.
- Let $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ and assume that $\mathbb{P}^\star = \arg\min\{\mathrm{KL}(\mathbb{P}|\mathbb{Q}) \ : \ \mathbb{Q} \in \mathrm{C}\}$ exists with $\mathrm{KL}(\mathbb{P}|\mathbb{P}^\star) < +\infty$ (hence is unique).
- Then we have that for any $\mathbb{Q} \in \mathrm{C}$

$$\boxed{\mathrm{KL}(\mathbb{P}|\mathbb{Q}) \geq \mathrm{KL}(\mathbb{P}|\mathbb{P}^\star) + \mathrm{KL}(\mathbb{P}^\star|\mathbb{Q}) \ .}$$

- Assume that $\mathbb{P}^\star$ is an **algebraic interior point**, i.e. for any $\mathbb{Q}_0 \in \mathrm{C}$, there exists $\alpha \in (0,1)$ and $\mathbb{Q}_1 \in \mathrm{C}$ such that $\mathbb{P}^\star = \alpha\mathbb{Q}_0 + (1-\alpha)\mathbb{Q}_1$. Then, we have equality.

- In our **Schrödinger bridge** setting we have

$$\boxed{\mathrm{KL}(\pi^0|\pi^\star) \geq \mathrm{KL}(\pi^0|\pi^1) + \mathrm{KL}(\pi^1|\pi^\star) \ .}$$

- Iterating, we get that

$$\boxed{\mathrm{KL}(\pi^0|\pi^\star) \geq \textstyle\sum_{k=0}^n \mathrm{KL}(\pi^k|\pi^{k+1}) + \mathrm{KL}(\pi^{n+1}|\pi^\star) \ .}$$

Convergence rates

- Additionally we can show that

$$\mathrm{KL}(\pi^k|\pi^{k+1}) \leq \mathrm{KL}(\pi^k|\pi^{k-1}) \, , \qquad \mathrm{KL}(\pi^{k+1}|\pi^k) \leq \mathrm{KL}(\pi^{k-1}|\pi^k) \, .$$

- Combining this with the fact that $\sum_{k\in\mathbb{N}} \mathrm{KL}(\pi^k|\pi^{k+1}) < +\infty$, we get that

$$\boxed{\lim_{n\to+\infty} n\{\mathrm{KL}(\pi_0^n|\nu_0) + \mathrm{KL}(\pi_1^n|\nu_1)\} = 0 \, .}$$

- This is a **quantitative rate** on the convergence of the **marginals**.
- Drawing connections with **Bregman gradient descent** we also have the following result.

Quantitative rate Léger (2021)

- We have the following rate

$$\boxed{\mathrm{KL}(\pi_0^n|\nu_0) + \mathrm{KL}(\pi_1^n|\nu_1) \leq 2\mathrm{KL}(\pi^\star|\pi^0)/n \, .}$$

- If $\pi^\star$ is close to π^0 then the **convergence is faster** (constant is smaller).

Conclusion

Limitation of the potential approach

- Recall that the **dynamical** formulation is given by

$$\boxed{\pi^\star = \arg\min\{\mathrm{KL}(\pi|\pi^0) \ : \ \pi \in \mathcal{P}((\mathbb{R}^d)^N), \pi_0 = \nu_0, \ \pi_N = \nu_1\} \ ,}$$

- Link with **generative modeling**:
  - $\pi^0 \in \mathcal{P}((\mathbb{R}^d)^N)$ is the discretization of the **Ornstein-Ulhenbeck** process.
  - ν_0 is the **data distribution**.
  - $\nu_1 = \mathrm{N}(0, \mathrm{Id})$ is the **easy-to-sample** distribution.
- The **Sinkhorn algorithm** is very efficient in **discrete settings** (matrix operations).

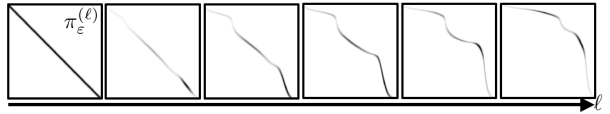


Figure 10: Convergence of the Sinkhorn algorithm. Image extracted from Peyré et al. (2019).

- Limitation of the **Sinkhorn algorithm** for Schrödinger bridges:
  - Learning the **potentials** (dynamic programming).
  - Sampling from **twisted kernels**.

Conclusion

- We have introduced a new **generative modeling** framework.
  - Introduction of **Schrödinger bridges**.
  - Connection with **Optimal transport**.
  - Introduction of the **Sinkhorn algorithm**.
- Next time:
  - Introduction of **Diffusion Schrödinger Bridge**.
  - **Implementation** of DSB.
  - **Extensions** of DSB.

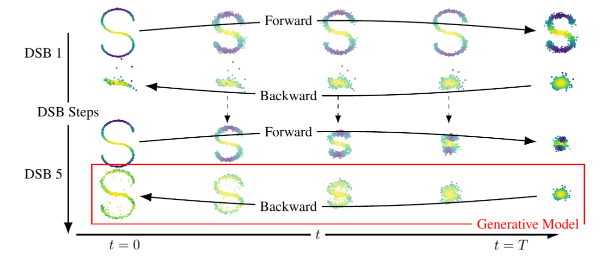


Figure 11: Diffusion Schrödinger Bridge. Image extracted from De Bortoli et al. (2021).

References

Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E Jacob. Schrodinger bridge samplers. *arXiv preprint arXiv:1912.13170*, 2019.

Garrett Birkhoff. Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227, 1957.

Peter J Bushell. Hilbert's metric and positive contraction mappings in a banach space. *Archive for Rational Mechanics and Analysis*, 52(4):330–338, 1973.

Yongxin Chen, Tryphon Georgiou, and Michele Pavon. Entropic and displacement interpolation: a computational approach using the hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396, 2016.

I. Csiszár. *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158, 1975. doi: 10.1214/aop/1176996454. URL https://doi.org/10.1214/aop/1176996454.

Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

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.

George Deligiannidis, Valentin De Bortoli, and Arnaud Doucet. Quantitative uniform stability of the iterative proportional fitting procedure. *arXiv preprint arXiv:2108.08129*, 2021.

Elon Kohlberg and John W Pratt. The contraction mapping approach to the perron-frobenius theory: Why hilbert's metric? *Mathematics of Operations Research*, 7(2):198–210, 1982.

Flavien Léger. A gradient descent perspective on sinkhorn. *Applied Mathematics & Optimization*, 84(2):1843–1855, 2021.

Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

Christian Léonard. Some properties of path measures. In *Séminaire de Probabilités XLVI*, pages 207–230. Springer, 2014.

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

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.

Marcel Nutz. Introduction to entropic optimal transport, 2021.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11 (5-6):355–607, 2019.

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

Ludger Rüschendorf and Wolfgang Thomsen. Note on the schrödinger equation and i-projections. *Statistics & probability letters*, 17(5):369–375, 1993.

References iv

Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pages 269–310, 1932.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.