Schrödinger Bridge Samplers

Espen Bernton
Columbia University

Joint work with J. Heng, A. Doucet, P. E. Jacob

JB^3, July 9, 2020

Schrödinger Bridge Samplers

(+ a note on exchangeability and optimal transport)

Espen Bernton
Columbia University

Joint work with J. Heng, A. Doucet, P. E. Jacob
(+ joint work with P. Ghosal, M. Nutz)

JBˆ3, July 9, 2020

Outline

- Problem setup and Monte Carlo
- The Schrödinger bridge problem
- Sequential Schrödinger bridge sampling
- Examples and numerical experiments
- Conclusion and future directions

Problem setup

Suppose that π_T is a Lebesgue density on $\mathsf{E} = \mathbb{R}^d$, expressed

$$\pi_T(x) = \frac{\gamma_T(x)}{Z_T}, \qquad Z_T = \int_{\mathsf{E}} \gamma_T(x)\mathrm{d}x.$$

We want to calculate

- expectations with respect to π_T,
- the unknown normalizing constant Z_T.

Can only evaluate γ_T (and later, $\nabla \log \gamma_T$) pointwise.

A stylized Monte Carlo problem

Suppose we can **sample $\boldsymbol{x}_0$ from** and **evaluate the density of $\boldsymbol{\pi}_0$**.

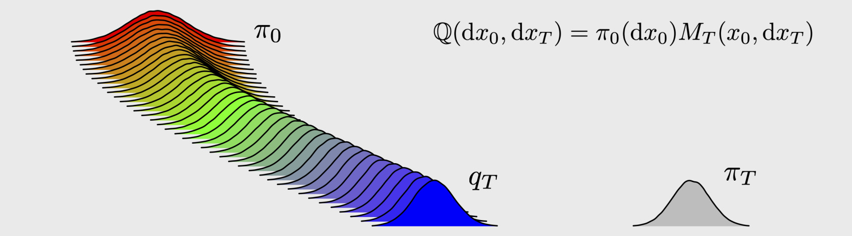
Choose and **sample a Markov kernel** $x_T \sim M_T(x_0, \mathrm{d}x_T)$ such that $q_T = \mathcal{L}(x_T)$ is closer to π_T than π_0.

We **want to use $\boldsymbol{q_T}$** as the proposal in **importance sampling.**

Two challenges:

1. How do we choose M_T?
2. The density of q_T is typically intractable.

A stylized Monte Carlo problem



Two challenges:

1. How do we choose M_T?
2. The density of q_T is typically intractable.

Second challenge

Extend the domain of integration to E^2:

- Define $\mathbb{Q}(\mathbf{d}\boldsymbol{x_0}, \mathbf{d}\boldsymbol{x_T}) = \boldsymbol{\pi_0}(\mathbf{d}\boldsymbol{x_0})\boldsymbol{M_T}(\boldsymbol{x_0}, \mathbf{d}\boldsymbol{x_T})$.
- Choose an auxiliary "backward" kernel L_0 and define the auxiliary target $\mathbb{P}(\mathbf{d}\boldsymbol{x_0}, \mathbf{d}\boldsymbol{x_T}) = \boldsymbol{\pi_T}(\mathbf{d}\boldsymbol{x_T})\boldsymbol{L_0}(\boldsymbol{x_T}, \mathbf{d}\boldsymbol{x_0})$,

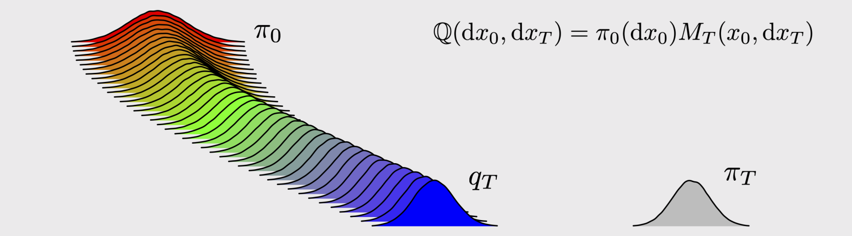
such that $\mathbb{P} \ll \mathbb{Q}$ and $w_{0,T}(x_0, x_T) = \frac{\mathrm{d}L_0 \otimes \gamma_T}{\mathrm{d}\pi_0 \otimes M_T}(x_0, x_T)$ can be evaluated pointwise.

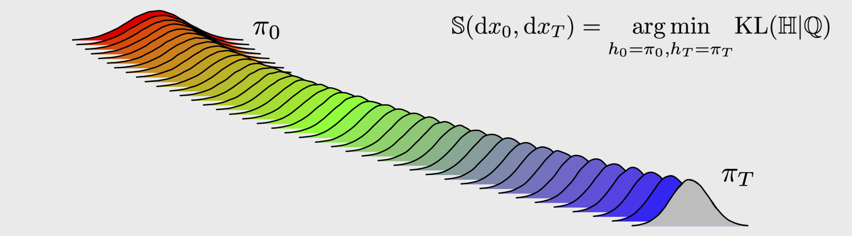
If $(x_0^n, x_T^n) \sim \mathbb{Q}$ and $w_{0,T}^n = w_{0,T}(x_0^n, x_T^n)$, then $\left\{\boldsymbol{x_T^n}, \boldsymbol{w_{0,T}^n}\right\}_{\boldsymbol{n=1}}^{\boldsymbol{N}}$

- is a **weighted sample from $\boldsymbol{\pi_T}$**, and
- $\hat{Z}_T = N^{-1} \sum_{n=1}^{N} w_{0,T}^n$ is an **unbiased estimator of $\boldsymbol{Z_T}$**.

First challenge

Main idea: Approximate $M_T^\star$ corresponding to the **Schrödinger bridge** between π_0 and π_T for a class of kernels.





The Schrödinger bridge problem

Given a **reference distribution** $\mathbb{Q}(\mathrm{d}x_0, \mathrm{d}x_T)$ and **marginal constraints** π_0 and π_T, find

$$\mathbb{S}(\mathrm{d}x_0, \mathrm{d}x_T) = \underset{h_0=\pi_0, h_T=\pi_T}{\operatorname{argmin}} \mathrm{KL}(\mathbb{H}|\mathbb{Q}),$$

<u>Remark:</u>
Consider $\boldsymbol{\mathbb{Q}^{\psi}(\mathrm{d}x_0, \mathrm{d}x_T) = \pi_0(\mathrm{d}x_0) M_T^{\psi}(x_0, \mathrm{d}x_T)}$, where ψ is a strictly positive function, or *policy*, and

$$M_T^{\psi}(x_0, \mathrm{d}x_T) = \frac{\psi(x_T) M_T(x_0, \mathrm{d}x_T)}{\int_{\mathsf{E}} \psi(x_T) M_T(x_0, \mathrm{d}x_T)}.$$

Then, $\mathbb{S}(\mathrm{d}x_0, \mathrm{d}x_T) = \mathbb{Q}^{\psi^\star}(\mathrm{d}x_0, \mathrm{d}x_T)$, where $\boldsymbol{\psi^\star}$ **is the solution to a Schrödinger equation.**

Some notes

- Original formulation by Schrödinger in 1931: **gas with very large number of particles N.**

- The modern formulation is derived by a **large deviations principle** as $N \to \infty$, where the KL is the rate functional.

- Connection to **optimal transport**: Suppose Schrödinger's particles are Brownian with scale σ, denoted $\mathbb{Q}^\sigma$, then

$$\begin{aligned}\lim_{\sigma \to 0} \sigma^2 \mathrm{KL}(\mathbb{S}^\sigma | \mathbb{Q}^\sigma) &= \inf_{\gamma_0 = \pi_0, \gamma_T = \pi_T} \int_{E^2} \|x_0 - x_T\|^2 \gamma(\mathrm{d}x_0, \mathrm{d}x_T) \\ &= \mathcal{W}_2^2(\pi_0, \pi_T).\end{aligned}$$

  - Important in computation, idea behind **entropically regularized optimal transport** (Cuturi, 2013).

- We will use a formulation from **optimal control** which is amenable to computation (Heng et al., 2019).

High-level algorithm to compute $\mathbb{S}(\mathrm{d}x_0, \mathrm{d}x_T)$

Iterative proportional fitting (or Sinkhorn's algorithm):

Let $\mathbb{Q}^{(0)} = \mathbb{Q}$, and for $i \geq 1$, define

$$\mathbb{P}^{(i)}(\mathrm{d}x_0, \mathrm{d}x_T) = \underset{h_T = \pi_T}{\operatorname{argmin}} \ \mathrm{KL}(\mathbb{H}|\mathbb{Q}^{(i-1)}),$$
$$\mathbb{Q}^{(i)}(\mathrm{d}x_0, \mathrm{d}x_T) = \underset{h_0 = \pi_0}{\operatorname{argmin}} \ \mathrm{KL}(\mathbb{H}|\mathbb{P}^{(i)}).$$

Let $\mathbb{S}^{(2i+1)} = \mathbb{P}^{(i+1)}$ and $\mathbb{S}^{(2i)} = \mathbb{Q}^{(i)}$ for any $i \geq 0$.

Remark: Given $\mathbb{Q}$ as the reference, $\mathbb{P}^{(1)}$ is the **optimal auxiliary target** in the sense of Del Moral et al. (2006).

Convergence of iterative proportional fitting

Rüschendorf (1995) shows that if there exists $c > 0$ such that

$$M_T(x_0, \mathrm{d}x_T) \geq c\pi_T(\mathrm{d}x_T), \quad \text{for } \pi_0\text{-a.e. } x_0 \in \mathsf{E},$$

then $\mathbb{S}^{(i)}$ **converges to** $\mathbb{S}$ in KL and TV as $i \to \infty$.

Proposition: For any $\varepsilon > 0$, IPF returns an $\mathbb{S}^{(i)}$ that satisfies

$$\mathrm{KL}(\pi_0|s_0^{(i)}) + \mathrm{KL}(\pi_T|s_T^{(i)}) < \varepsilon$$

in fewer than $\lceil \mathrm{KL}(\mathbb{S}|\mathbb{Q})/\varepsilon \rceil$ iterations.

IPF as policy refinement

Using the ψ-parameterization, it turns out that we can express

$$\mathbb{Q}^{(i)} = \mathbb{Q}^{\psi^{(i)}},$$

for two sequences $\psi^{(i)}$ and $\phi^{(i)}$, satisfying

$$\boldsymbol{\phi^{(i)}(x_T) = \frac{\mathrm{d}\pi_T}{\mathrm{d}q_T^{\psi^{(i-1)}}}(x_T)}, \qquad \psi^{(i)} = \psi^{(i-1)} \cdot \phi^{(i)}.$$

The sequence $\psi^{(i)} \to \psi^{\star}$ as $i \to \infty$.

IPF as policy refinement

For any $\mathbb{H} \ll \mathbb{Q}^\psi$ such that $h_T = \pi_T$, we have that

$$\frac{\mathrm{d}\pi_T}{\mathrm{d}q_T^\psi}(x_T) = \int_{\mathsf{E}} \frac{\mathrm{d}\mathbb{H}}{\mathrm{d}\mathbb{Q}^\psi}(x_0, x_T)\mathbb{Q}^\psi(\mathrm{d}x_0|x_T).$$

If $(x_0, x_T) \sim \mathbb{Q}^\psi$, then, conditional on x_T, we have $x_0 \sim \mathbb{Q}^\psi(\mathrm{d}x_0|x_T)$.

Thus, if $\mathbb{H}(\mathrm{d}x_0, \mathrm{d}x_T) = \pi_T(\mathrm{d}x_T)L_0^\psi(x_T, \mathrm{d}x_0)$, then $\boldsymbol{w_{0,T}^\psi(x_0, x_T)}$ is an **unbiased estimator of** $\boldsymbol{\frac{\mathrm{d}\pi_T}{\mathrm{d}q_T^\psi}(x_T)}$.

- Can borrow ideas from **conditional SMC** to reduce variance.

Approximate IPF

Given $\{(\boldsymbol{x_0^n}, \boldsymbol{x_T^n})\}_{\boldsymbol{n=1}}^{\boldsymbol{N}} \sim \mathbb{Q}^{\hat{\boldsymbol{\psi}}^{(\boldsymbol{i-1})}}$, approximate $\phi^{(i)}$ with

$$\hat{\phi}^{(i)} = \operatorname*{argmin}_{f \in \mathsf{F}} \sum_{n=1}^{N} \left| \log f(x_T^n) - \log R^{\hat{\psi}^{(i-1)}}(x_T^n) \right|^2,$$

- F is a **function class**,
- $R^{\hat{\psi}^{(i-1)}}(x_T)$ is an **estimator** of $\frac{\mathrm{d}\pi_T}{\mathrm{d}q_T^{\hat{\psi}^{(i-1)}}}(x_T)$.

Choice of kernels and function classes

Restriction: Must be able to

- sample from $\mathbb{Q}^{\hat{\boldsymbol{\psi}}^{(i-1)}}$, i.e. **sample from** $M_T^{\hat{\boldsymbol{\psi}}^{(i-1)}}$,
- **evaluate** $\boldsymbol{w}_{\mathbf{0},\boldsymbol{T}}^{\hat{\boldsymbol{\psi}}^{(i-1)}}$ at the points $\left\{(x_0^n, x_T^n)\right\}_{n=1}^N \sim \mathbb{Q}^{\hat{\boldsymbol{\psi}}^{(i-1)}}$.

Important example:

- the kernel $M_T(x_0, \mathrm{d}x_T)$ is **Gaussian**,
- the function class $\log \mathsf{F}$ is the **quadratic forms**,
- approximate the optimal backward kernel $L_0^{(i)}$, in the sense of Del Moral et al. (2006), with similar regressions.

Toy example

Suppose $\boldsymbol{\pi_0 = \mathcal{N}(0, \mathcal{I})}$, $\boldsymbol{\pi_T = \mathcal{N}(\mu_T, \Sigma_T)}$, where

$$\mu_T = (17.9, 17.9), \qquad \Sigma_T = \begin{pmatrix} 0.40 & 0.24 \\ 0.24 & 0.40 \end{pmatrix}$$

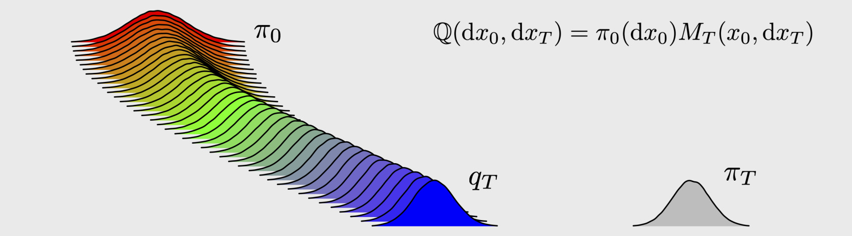
Let M_T be the kernel arising from an **Euler-Maruyama discretization** of the Langevin diffusion

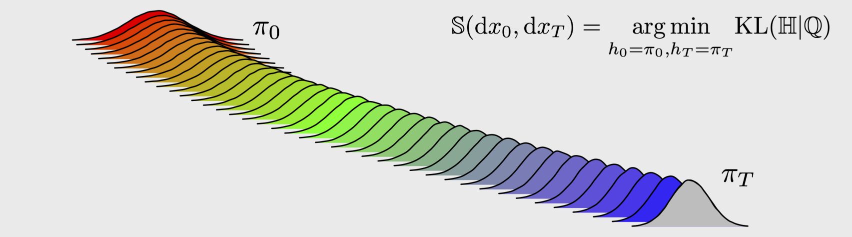
$$\mathrm{d}X_s = \frac{1}{2}\nabla \log \pi_s(X_s)\mathrm{d}s + \mathrm{d}W_s, \quad \text{for } s \in [0, \tau], \quad X_0 \sim \pi_0,$$

where $(\pi_s)_{s\in[0,\tau]}$ is the geometric interpolation of π_0 and π_T.

Suppose we take $\boldsymbol{\tau = 2}$ and **40 steps** of Euler-Maruyama, and $\boldsymbol{i = 5}$ iterations of IPF.

Toy example: Illustration of first marginal





Sequential Schrödinger bridge sampling

Instead of targeting π_T directly, we introduce an **interpolation $\{\pi_t\}_{t=0}^T$**, for example

$$\gamma_t(x_t) = \pi_0(x_t)^{1-\lambda_t}\gamma_T(x_t)^{\lambda_t}, \qquad \pi_t(x_t) = \gamma_t(x_t)/Z_t,$$

where $\{\lambda_t\}_{t=0}^T \subset [0,1]$ is increasing, $\lambda_0 = 0$ and $\lambda_T = 1$.

Introduce a **sequence of Markov kernels $\{M_t\}_{t=1}^T$**, and let

$$\mathbb{Q}(\mathrm{d}x_{0:T}) = \pi_0(\mathrm{d}x_0)\prod_{t=1}^{T} M_t(x_{t-1}, \mathrm{d}x_t).$$

Sequential Schrödinger bridge sampling

Consider the **multi-marginal** Schrödinger bridge problem:

$$\mathbb{S}(\mathrm{d}x_{0:T}) = \underset{h_t=\pi_t,\,\forall\, t\in\{0,\dots,T\}}{\operatorname{argmin}} \mathrm{KL}(\mathbb{H}|\mathbb{Q}).$$

Proposition: Can be solved **sequentially**. Consider the sequence of intermediate problems

$$\begin{aligned}\mathbb{S}_{t-1,t}(\mathrm{d}x_{t-1},\mathrm{d}x_t) &= \underset{h_{t-1}=\pi_{t-1},h_t=\pi_t}{\operatorname{argmin}} \mathrm{KL}(\mathbb{H}_{t-1,t}|\mathbb{Q}_{t-1,t}),\\ &= \pi_{t-1}(\mathrm{d}x_{t-1})\boldsymbol{M_t^{\psi_t^\star}(x_{t-1},\mathrm{d}x_t)}.\end{aligned}$$

Then, $\boldsymbol{\mathbb{S}(\mathrm{d}x_{0:T}) = \pi_0(\mathrm{d}x_0)\prod_{t=1}^T M_t^{\psi_t^\star}(x_{t-1},\mathrm{d}x_t)}$, where $\{\psi_t^\star\}_{t=1}^T$ similarly solve a set of Schrödinger equations.

Algorithm

Initialize $\boldsymbol{\{x_0^n\}_{n=1}^N \sim \pi_0}$. For each $t = 1, \ldots, T$,

- Perform $\boldsymbol{i}$ **iterations of approximate IPF** to obtain $x_t^n \sim M_t^{(i)}(x_{t-1}^n, \mathrm{d}x_t^n)$ and

$$w_{t-1,t}^{(i)}(x_{t-1}^n, x_t^n) = \frac{\mathrm{d}L_{t-1}^{(i)} \otimes \gamma_t}{\mathrm{d}\gamma_{t-1} \otimes M_t^{(i)}}(x_{t-1}^n, x_t^n),$$

  for $n = 1, \ldots, N$.

Return $\boldsymbol{\{(x_T^n, w_{0:T}^n)\}_{n=1}^N}$, where $w_{0:T}^n = \prod_{t=1}^T w_{t-1,t}^{(i)}(x_{t-1}^n, x_t^n)$.

Optional: Add resampling steps.

Generic choice of kernels

For $t = 1, \ldots, T$, let M_t denote the t-th step of the Euler-Maruyama **discretization of Langevin diffusion**:

$$\mathrm{d}X_s = \frac{1}{2}\nabla \log \pi_s(X_s)\mathrm{d}s + \mathrm{d}W_s, \quad \text{for } s \in [0, \tau], \quad X_0 \sim \pi_0.$$

Let $\log \mathsf{F}_t$ be the **quadratic forms**, then $\boldsymbol{M_t^{\psi}}$ **is Gaussian** for every t and ψ.

Can similarly approximate the optimal backward kernels using quadratic forms.

Small step-size regime

For sufficiently large τ and small step size $h > 0$, q_t should provide a **reasonable approximation** of π_t.

For small h, we can also leverage **flexible function classes** by approximating the underlying continuous-time SBP:

$$M_t^{\psi}(x_{t-1}, \mathrm{d}x_t) \approx \mathcal{N}\left(\mathrm{d}x_t; x_{t-1} + \tfrac{h}{2}\nabla \log \pi_t(x_{t-1}) + h\nabla \log \psi_t(x_{t-1}), h\mathcal{I}_d\right).$$

Continuous-time Schrödinger bridge problem:

Find $(\psi_s^\star)_{s\in[0,\tau]}$ such that $\boldsymbol{X_0 \sim \pi_0, X_\tau \sim \pi_T}$,

$$\mathrm{d}X_s = \frac{1}{2}\nabla \log \pi_s(X_s)\mathrm{d}s + \nabla \log \psi_s(X_s)\mathrm{d}s + \mathrm{d}W_s, \quad \text{for } s \in [0, \tau],$$

and $(\psi_s^\star)_{s\in[0,\tau]}$ **minimizes** $\boldsymbol{\int_0^\tau \mathbb{E}\|\nabla \log \psi_s(X_s)\|^2 \mathrm{d}s}$.

Example: Linear Quadratic Gaussian

Prior: $\pi_0(\mathrm{d}x_0) = \mathcal{N}(\mathrm{d}x_0; 0, \mathcal{I})$.

Log-likelihood: $\ell(x) = -(y-x)^\top R^{-1}(y-x)/2$, observation $y \in \mathbb{R}^d$, symmetric positive definite $R \in \mathbb{R}^{2\times 2}$.

Posterior: $\pi_T(\mathrm{d}x_T) = \mathcal{N}(\mathrm{d}x_T; \mu_T, \Sigma_T)$ with $\Sigma_T = \left(\Sigma_0^{-1} + R^{-1}\right)^{-1}$, $\mu_T = \Sigma_T \left(\Sigma_0^{-1}\mu_0 + R^{-1}y\right)$.

Parameters: $y = (8, 8)^\top$, $R_{11} = R_{22} = 1$, $R_{12} = R_{21} = 0.8$.

Example: Linear Quadratic Gaussian

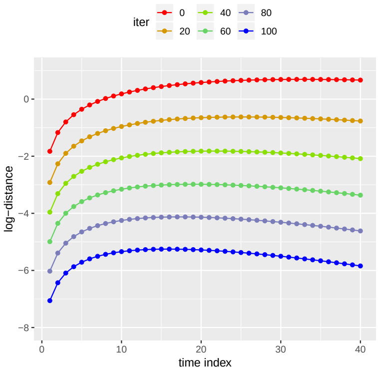
Kernels: Discretized Langevin diffusion with $h = 1/20$.

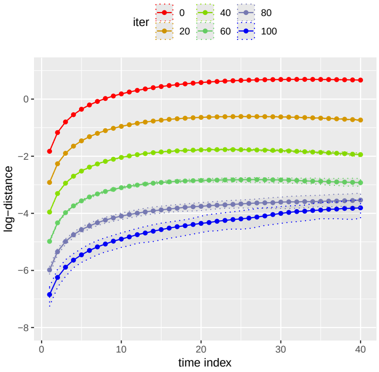
Interpolation: $\tau = 2$, $T = 40$, $\lambda_t = t/T$.

Function classes: If $f \in \mathsf{F}_t$, then $\log f$ is quadratic.

Example: Linear Quadratic Gaussian

Plot: $\log \mathcal{W}_2(\pi_t, q_t^{(i)})$ as a function of t, for different $i \geq 0$.





Left: Exact IPF. Right: SSB with $N = 1,000$.

Example: Linear Quadratic Gaussian

Comparing the reference sampler with the SSB sampler for $N = 1,000$,

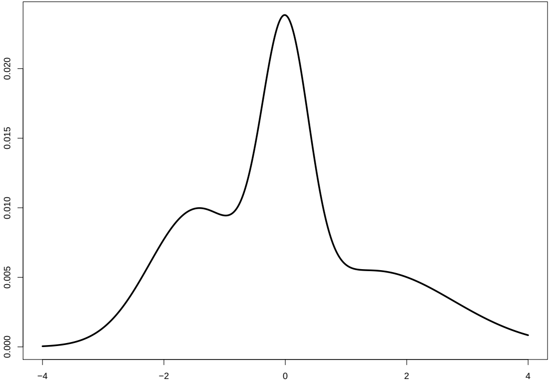
- The MSE of $\log \hat{Z}_T$ obtained with reference sampler was **7396 times higher** than the SSB estimator.
- The wall-clock time consumed by the SSB sampler was **7.4 times higher** than the reference sampler.

SSB about **1,000 times more efficient** in terms of MSE per unit of computation time.

Example: 1D mixture

Target distribution: $\pi_T(\mathrm{d}x_T) = \sum_{i=1}^{p} w_i \mathcal{N}(\mathrm{d}x_T; \mu_i, \sigma_i^2)$.

Parameters: $p = 3$, $\mu = (-1.5, 0, 1.5)$, $\sigma = (0.6, 0.15, 1.8)$, $w = (1/3, 1/3, 1/3)$.



Example: 1D mixture

Kernels: Discretized Langevin diffusion with $h = 1/50$.

Interpolation: $\pi_0(\mathrm{d}x_0) = \mathcal{N}(\mathrm{d}x_0; 0, 50)$, $\tau = 2$, $T = 100$, $\lambda_t = t^2/T^2$.

Function classes: If $f \in \mathsf{F}_t$, then $\log f$ is a cubic smoothing spline with 25 knots, estimated with `smooth.spline` in R.

Example: 1D mixture

For the SSB sampler and reference sampler with $N = 500$,

- The MSE of $\log \hat{Z}_T$ obtained with the reference sampler was **53.4 times higher** than the SSB estimator.
- The wall-clock time consumed by the SSB sampler was **17.8 times higher** than the reference sampler.

SSB about **3 times more efficient** in terms of MSE per unit of computation time.

Example: Logistic regression

Data: Cleveland heart disease database, $M = 297$ individuals, each with $d = 20$ binary and continuous covariates X_m.

Prior: Weakly informative prior from Gelman et al. (2008).

Log-likelihood: $\ell(x) = y^\top Xx - \sum_{m=1}^{M} \log(1 + \exp(x^\top X_m))$, response variable $y \in \{0, 1\}^M$, covariate matrix $X \in \mathbb{R}^{M \times d}$.

Example: Logistic regression

Interpolation: $\tau = 2$, $T = 40$, $\lambda_t = t^2/T^2$.

Kernels: Discretized Langevin diffusion with $h = 1/20$.

Function classes: If $f \in \mathsf{F}_t$, then $\log f(x) = x^\top A x + b^\top x + c$, where $A \in \mathbb{R}^{d \times d}$ is diagonal.

Example: Logistic regression

Over 100 repeated simulations with $N = 4,000$, the average estimates of $\log Z_T$ were

- SSB: $\mathbf{-126.7\ (sd = 0.09)}$,
- Reference: $\mathbf{-130.5\ (sd = 2.7)}$,

Summary

Using the SMC framework, we leverage approximations of **Schrödinger bridges to do Monte Carlo sampling**.

Important features of the algorithm include

- iterative proportional fitting,
- function approximation,
- estimation of normalizing constants and Radon-Nikodym derivatives.

Compared to a well-tuned reference processes, the SSB sampler showed **computational gains** in a few simple examples.

Future directions

Extend the method to **other kinds of kernels**, e.g.

- Gibbs sampling,
- Kernels that utilize model structure in high dimensions.

Many **theoretical aspects** left to consider, e.g.

- Asymptotic properties in N, i and T,
- Behavior of IPF with **misspecified** function classes.

Optimal transport and statistics

Ideas from optimal transport and related literatures has inspired many recent methods and results in statistics.

Relatively small community using **statistical ideas to learn about optimal transport**.

Example: Optimal transport from exchangeability.

Optimal transport from exchangeability

Optimal transport problem:

Given

- marginals μ on X and ν on Y,
- a cost function $c : \mathsf{X} \times \mathsf{Y} \to [0, \infty]$,

solve

$$\min_{\gamma_x = \mu, \gamma_y = \nu} \int_{\mathsf{X} \times \mathsf{Y}} c(x, y) \gamma(\mathrm{d}x, \mathrm{d}y),$$

and find the argmin.

Notably studied by **Monge** (1781) and **Kantorovich** (1942).

Optimal transport from exchangeability

Consider the following scheme:

- sample $z_k = \big[(x_i, y_i)\big]_{i=1}^k \sim (\mu \otimes \nu)^k$,
- find $M(z_k) = \operatorname{argmin}_{\sigma \in \mathcal{S}(k)} \sum_{i=1}^k c(x_i, y_{\sigma(i)})$,
- sample $\bar{\sigma} \sim \mathrm{Unif}\{M(z_k)\}$,
- return $\bar{z}_k = \big[(x_i, y_{\sigma(i)})\big]_{i=1}^k = \big[(\bar{x}_i, \bar{y}_i)\big]_{i=1}^k$.

Define $\boldsymbol{\Gamma_k = \mathcal{L}(\bar{z}_k)}$, which takes values on $\mathcal{C}_k = \{\bar{z}_k : \sigma_{\mathrm{id}} \in M(\bar{z}_k)\}$.

Note that $\hat{\gamma}_{\boldsymbol{\bar{z}_k}} = \frac{1}{k} \sum_{i=1}^k \delta_{(\bar{x}_i, \bar{y}_i)} \in \mathbf{OT}(\boldsymbol{\hat{\mu}_k, \hat{\nu}_k})$.

Optimal transport from exchangeability

For every $k \geq 1$, the rows of $\bar{z}_k \sim \mathit{\Gamma}_k$ are **exchangeable**.

By the Diaconis-Freedman theorem, one can derive a limit of $\mathit{\Gamma}_k$ on $(\mathsf{X} \times \mathsf{Y})^\infty$:

$$\mathit{\Gamma}(A) = \int \gamma^\infty(A) \mathrm{d}\mathcal{L}(\gamma),$$

where $\mathcal{L}(\gamma)$ is the weak limit of $\mathcal{L}(\hat{\gamma}_{\bar{z}_k})$.

By **stability results** on optimal transport, we know that the limit points of $\hat{\gamma}_{\bar{z}_k}$ almost surely belong to $\mathrm{OT}(\mu, \nu)$.

Optimal transport from exchangeability

Hence, $\mathcal{L}(\gamma)$ takes values in $\mathrm{OT}(\mu, \nu)$, and

$$\boldsymbol{\gamma^\star(B) = \int_{\mathbf{OT}(\mu,\nu)} \gamma(B)\mathbf{d}\mathcal{L}(\gamma)}$$

is an optimal transport measure.

Thanks!