From Optimization to Sampling, with a Touch of Schrödinger Bridge

Qijia Jiang

October 2023

Stanford ISL Colloquium

Outline

Hope to share a few complementary perspectives on the sampling problem:

$$\pi \propto e^{-f} \text{ target density known up to normalizing constant}$$

Outline

Hope to share a few complementary perspectives on the sampling problem:

$$\pi \propto e^{-f} \text{ target density known up to normalizing constant}$$

Meta-principle

Design a process that gradually transform

$$\text{simple} \to \text{complicated distribution}\,.$$

Outline

Hope to share a few complementary perspectives on the sampling problem:

$$\pi \propto e^{-f} \text{ target density known up to normalizing constant}$$

Meta-principle

Design a process that gradually transform

$$\text{simple} \to \text{complicated distribution}\,.$$

A few stops

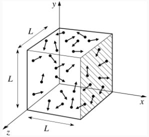
1. Optimization interpretation through PDE lens: Mirror Langevin for sampling under more general geometry
2. Borrow ideas from generative modeling: optimal stochastic control / optimal transport to steer a trajectory from ν to π using machine learning
3. ~~Traditional MCMC: Mixing of Hamiltonian Monte Carlo~~

Mirror Langevin under Isoperimetry

(Unadjusted) Overdamped Langevin

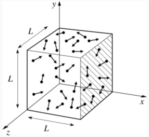
- SDE with gradient of potential as drift

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{1}$$



(Unadjusted) Overdamped Langevin

- SDE with gradient of potential as drift



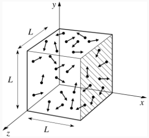
$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{1}$$

- (Unique) Invariant measure is $\pi \propto e^{-f}$ under mild assumptions

(Unadjusted) Overdamped Langevin

- SDE with gradient of potential as drift

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{1}$$

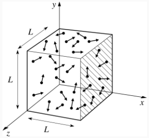


- (Unique) Invariant measure is $\pi \propto e^{-f}$ under mild assumptions
- Can be seen from PDE representation of density (Fokker-Planck equation)

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\pi} \right)$$

(Unadjusted) Overdamped Langevin

- SDE with gradient of potential as drift



$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{1}$$

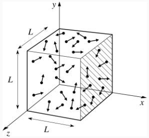
- (Unique) Invariant measure is $\pi \propto e^{-f}$ under mild assumptions
- Can be seen from PDE representation of density (Fokker-Planck equation)

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\pi}\right)$$

- Probability flow ODE:

$$\dot{X}_t = \nabla \log \rho_t(X_t) - \nabla \log \pi(X_t) \leftarrow \text{interacting particle system (need } \hat{\rho}_t)$$

(Unadjusted) Overdamped Langevin



- SDE with gradient of potential as drift

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{1}$$

- (Unique) Invariant measure is $\pi \propto e^{-f}$ under mild assumptions
- Can be seen from PDE representation of density (Fokker-Planck equation)

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\pi}\right)$$

- Probability flow ODE:

$$\dot{X}_t = \nabla \log \rho_t(X_t) - \nabla \log \pi(X_t) \leftarrow \text{interacting particle system (need } \hat{\rho}_t)$$

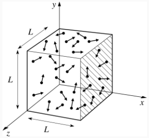
- To implement, Euler discretization using the classical MCMC SDE perspective (1):

$$x_{k+1} = x_k - h \cdot \nabla f(x_k) + \sqrt{2h} \cdot z_{k+1}$$

converges to $\pi_h \neq \pi$ but $\pi_h \to \pi$ as $h \to 0$. Can add Metropolis Hastings accept/reject for the proposal to correct for the bias

(Unadjusted) Overdamped Langevin

- SDE with gradient of potential as drift



$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{1}$$

- (Unique) Invariant measure is $\pi \propto e^{-f}$ under mild assumptions
- Can be seen from PDE representation of density (Fokker-Planck equation)

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\pi} \right)$$

- Probability flow ODE:

$$\dot{X}_t = \nabla \log \rho_t(X_t) - \nabla \log \pi(X_t) \leftarrow \text{interacting particle system (need } \hat{\rho}_t)$$

- To implement, Euler discretization using the classical MCMC SDE perspective (1):

$$x_{k+1} = x_k - h \cdot \nabla f(x_k) + \sqrt{2h} \cdot z_{k+1}$$

  converges to $\pi_h \neq \pi$ but $\pi_h \to \pi$ as $h \to 0$. Can add Metropolis Hastings accept/reject for the proposal to correct for the bias
- Convergence $\mathcal{O}(\text{poly}(\frac{1}{\epsilon}, d, \kappa))$ under various assumptions/metrics well known by now

JKO Scheme - I

There is a connection to *deterministic* optimization, but one has to lift the formalism to the space of probability measures $\mathcal{P}(\mathbb{R}^d)$.

[JKO '98] Density $X_t \sim \rho_t$ along SDE dynamics (1) follows gradient flow of minimizing KL functional with Wasserstein-2 metric in the space of probability measures

$$\text{``}\dot{\rho}_t = -\nabla_{W_2} KL(\rho_t \| \pi)\text{''}$$

JKO Scheme - I

[JKO '98] Density $X_t \sim \rho_t$ along SDE dynamics (1) follows gradient flow of minimizing KL functional with Wasserstein-2 metric in the space of probability measures

$$\boxed{\text{“}\dot{\rho}_t = -\nabla_{W_2} KL(\rho_t \| \pi)\text{”}}$$

What the paper actually described was an iterative scheme

$$\rho_{k+1} = \arg\min_{\rho} \int \rho \log \frac{\rho}{\pi} dx + \frac{1}{2h} W_2^2(\rho, \rho_k) \tag{2}$$

taking stepsize $h \to 0$, the discrete update trace out a curve $(\rho_t)_t$ in $\mathcal{P}(\mathbb{R}^d)$, which solves the Fokker-Planck PDE for Langevin:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t = \nabla \cdot (\rho_t \underbrace{\nabla \log \frac{\rho_t}{\pi}}_{\text{“}\nabla_{W_2} KL\text{”}})$$

JKO Scheme - I

[JKO '98] Density $X_t \sim \rho_t$ along SDE dynamics (1) follows gradient flow of minimizing KL functional with Wasserstein-2 metric in the space of probability measures

$$\boxed{\text{“}\dot{\rho}_t = -\nabla_{W_2} KL(\rho_t \| \pi)\text{”}}$$

What the paper actually described was an iterative scheme

$$\rho_{k+1} = \arg\min_{\rho} \int \rho \log \frac{\rho}{\pi} dx + \frac{1}{2h} W_2^2(\rho, \rho_k) \tag{2}$$

taking stepsize $h \to 0$, the discrete update trace out a curve $(\rho_t)_t$ in $\mathcal{P}(\mathbb{R}^d)$, which solves the Fokker-Planck PDE for Langevin:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t = \nabla \cdot (\rho_t \underbrace{\nabla \log \frac{\rho_t}{\pi}}_{\text{“}\nabla_{W_2} KL\text{”}})$$

But Wasserstein gradient flow (2) isn't implementable as evolution of density. PDE typically easier for analysis, SDE easier for simulation.

JKO Scheme - II

Many developments since, tremendous implications for PDEs, calculus of variations, optimal transport etc.,

... but doesn't seem to be very well known in MCMC/statistical computation community

JKO Scheme - II

One immediate consequence (of the GF perspective)

ϕ-Log-Sobolev inequality: For all ρ, and some strongly convex function ϕ, π satisfies

$$\int \rho(x) \log \frac{\rho(x)}{\pi(x)} dx \leq \frac{1}{2\alpha} \int \rho(x) \left\| \nabla \log \frac{\rho(x)}{\pi(x)} \right\|^2_{[\nabla^2 \phi(x)]^{-1}} dx \tag{3}$$

becomes gradient-domination condition (implies no local minima):

$$\text{suboptimality gap in obj function} \leq 1/2\alpha \cdot \|\text{gradient}\|^2$$

$\Rightarrow$ linear convergence in KL in continuous time $\mathcal{O}(\log(\frac{1}{\epsilon}))$ for Langevin dynamics when taking $\phi = \frac{1}{2}\|\cdot\|^2$.

JKO Scheme - II

One immediate consequence (of the GF perspective)

ϕ-Log-Sobolev inequality: For all ρ, and some strongly convex function ϕ, π satisfies

$$\int \rho(x) \log \frac{\rho(x)}{\pi(x)} dx \leq \frac{1}{2\alpha} \int \rho(x) \left\| \nabla \log \frac{\rho(x)}{\pi(x)} \right\|^2_{[\nabla^2 \phi(x)]^{-1}} dx \tag{3}$$

becomes gradient-domination condition (implies no local minima):

$$\text{suboptimality gap in obj function} \leq 1/2\alpha \cdot \|\text{gradient}\|^2$$

$\Rightarrow$ linear convergence in KL in continuous time $\mathcal{O}(\log(\frac{1}{\epsilon}))$ for Langevin dynamics when taking $\phi = \frac{1}{2}\|\cdot\|^2$.

Assumption (3) is more robust / weaker than $\nabla^2 f \succ 0$ (tail condition on target π).

Motivating Q: Is there a mirror descent analogue of Langevin? Convergence?

Brief Optimization Reminder: Mirror Flow and Mirror Descent

Mirror flow (in dual space) for $\nabla^2\phi \succ 0$:

$$dY_t = -\nabla f(X_t)dt, \quad Y_t = \nabla\phi(X_t) \tag{4}$$

same as (in primal space) Riemannian gradient flow:

$$dX_t = -\underbrace{(\nabla^2\phi(X_t))^{-1}\nabla f(X_t)}_{\text{grad } f \text{ under metric } \nabla^2\phi} dt \tag{5}$$

precondition for local geometry through choice of mirror map ϕ.

Brief Optimization Reminder: Mirror Flow and Mirror Descent

Mirror flow (in dual space) for $\nabla^2\phi \succ 0$:

$$dY_t = -\nabla f(X_t)dt, \quad Y_t = \nabla\phi(X_t) \tag{4}$$

same as (in primal space) Riemannian gradient flow:

$$dX_t = -\underbrace{(\nabla^2\phi(X_t))^{-1}\nabla f(X_t)}_{\text{grad } f \text{ under metric } \nabla^2\phi} dt \tag{5}$$

precondition for local geometry through choice of mirror map ϕ.

Mirror descent discretizes (4):

$$\begin{aligned} x_{k+1} &= \nabla\phi^*(\nabla\phi(x_k) - h_{k+1}\nabla f(x_k)) \\ &= \arg\min_x \langle x, \nabla f(x_k)\rangle + h_{k+1}^{-1}D_\phi(x, x_k) \end{aligned}$$

NB: No need to evaluate higher order derivatives of $\phi(\cdot)$, in contrast to discretizing in primal space.

Brief Optimization Reminder: Mirror Flow and Mirror Descent

Mirror flow (in dual space) for $\nabla^2\phi \succ 0$:

$$dY_t = -\nabla f(X_t)dt, \quad Y_t = \nabla\phi(X_t) \tag{4}$$

same as (in primal space) Riemannian gradient flow:

$$dX_t = -\underbrace{(\nabla^2\phi(X_t))^{-1}\nabla f(X_t)}_{\text{grad } f \text{ under metric } \nabla^2\phi} dt \tag{5}$$

precondition for local geometry through choice of mirror map ϕ.

Mirror descent discretizes (4):

$$\begin{aligned} x_{k+1} &= \nabla\phi^*(\nabla\phi(x_k) - h_{k+1}\nabla f(x_k)) \\ &= \operatorname*{arg\,min}_x \langle x, \nabla f(x_k)\rangle + h_{k+1}^{-1} D_\phi(x, x_k) \end{aligned}$$

NB: No need to evaluate higher order derivatives of $\phi(\cdot)$, in contrast to discretizing in primal space.

Common choices: $\phi(x) = \|x\|_2^2/2$ GD; $\phi(x) = -\sum_i x_i \log(x_i)$, which gives multiplicative weight update. When $\phi = f$ Newton's method.

Mirror Langevin and Naive Discretization

Mirror Langevin continuous dynamics in primal space:

$$dX_t = (\nabla \cdot (\nabla^2 \phi(X_t)^{-1}) - \nabla^2 \phi(X_t)^{-1} \nabla f(X_t))dt + \sqrt{2\nabla^2\phi(X_t)^{-1}} dW_t \,.$$

Fokker-Planck PDE for the primal variable X (stationary at $\rho_t = \pi$, has GF interpretation with tangent vector grad $f = -[\nabla^2\phi]^{-1}\nabla_{W_2} KL$):

$$\boxed{\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t [\nabla^2 \phi]^{-1} \nabla \log \frac{\rho_t}{\pi} \right)}$$

Corresponding to SDE in dual space ("Wasserstein mirror flow"):

$$dY_t = -\nabla f(\nabla \phi^*(Y_t))dt + \sqrt{2[\nabla^2\phi^*(Y_t)]^{-1}} dW_t, \quad Y_t = \nabla\phi(X_t)$$

Mirror Langevin and Naive Discretization

Corresponding to SDE in dual space ("Wasserstein mirror flow"):

$$dY_t = -\nabla f(\nabla \phi^*(Y_t))dt + \sqrt{2[\nabla^2 \phi^*(Y_t)]^{-1}} dW_t, \quad Y_t = \nabla \phi(X_t)$$

Euler-Maruyama [Zhang, Peyré et al. '20]

$$x_{k+1} = \nabla \phi^* \left(\nabla \phi(x_k) - h_{k+1} \nabla f(x_k) + \sqrt{2h_{k+1}[\nabla^2 \phi(x_k)]} \cdot z_{k+1} \right)$$

Can invert $\nabla \phi^*$ numerically, i.e., convex optimization. $\nabla \phi^*(x) = \arg\max_y \; x^\top y - \phi(y)$

Mirror Langevin and Naive Discretization

Corresponding to SDE in dual space ("Wasserstein mirror flow"):

$$dY_t = -\nabla f(\nabla \phi^*(Y_t))dt + \sqrt{2[\nabla^2 \phi^*(Y_t)]^{-1}}dW_t, \quad Y_t = \nabla \phi(X_t)$$

Euler-Maruyama [Zhang, Peyré et al. '20]

$$x_{k+1} = \nabla \phi^* \left(\nabla \phi(x_k) - h_{k+1} \nabla f(x_k) + \sqrt{2h_{k+1}[\nabla^2 \phi(x_k)]} \cdot z_{k+1} \right)$$

Can invert $\nabla \phi^*$ numerically, i.e., convex optimization. $\nabla \phi^*(x) = \arg\max_y \; x^\top y - \phi(y)$

There is a splitting somewhat inspired by the JKO perspective:

$$KL(\rho \| \pi) = \int \rho \log \frac{\rho}{\pi} dx = \mathrm{NegEnt}(\rho) + \mathbb{E}_\rho[f] =: (a) + (b)$$

Maximizing entropy part (a) is solvable by Brownian motion; minimizing f part (b) involves gradient flow on potential f.

A better (implementable) scheme

Recall: $dY_t = -\nabla f(X_t)dt + \sqrt{2[\nabla^2\phi^*(Y_t)]^{-1}}dW_t, \quad Y_t = \nabla\phi(X_t)$

Alternative Schemes (discretize objective but not geometry)

Forward: $(\dot{X}_t = -\nabla f(X_t) \Rightarrow x_{k+1} = x_k - h \cdot \nabla f(x_k))$

$$\begin{cases} x_{k+1/2} = \arg\min_v \; h\nabla f(x_k)^\top v + D_\phi(v, x_k) = \nabla\phi^*(\nabla\phi(x_k) - h\nabla f(x_k)) \\ \text{solve } dy_t = \sqrt{2[\nabla^2\phi^*(y_t)]^{-1}}dW_t \;\text{ for }\; y_0 = \nabla\phi(x_{k+1/2}) \qquad (\clubsuit) \\ x_{k+1} = \nabla\phi^*(y_h) \end{cases}$$

Brownian motion part ($\clubsuit$) can be solved approximately with EM.

A better (implementable) scheme

Recall: $dY_t = -\nabla f(X_t)dt + \sqrt{2[\nabla^2 \phi^*(Y_t)]^{-1}}dW_t, \quad Y_t = \nabla\phi(X_t)$

Alternative Schemes (discretize objective but not geometry)

Forward: $(\dot{X}_t = -\nabla f(X_t) \Rightarrow x_{k+1} = x_k - h \cdot \nabla f(x_k))$

$$\begin{cases} x_{k+1/2} = \arg\min_v \; h\nabla f(x_k)^\top v + D_\phi(v, x_k) = \nabla\phi^*(\nabla\phi(x_k) - h\nabla f(x_k)) \\ \text{solve } dy_t = \sqrt{2[\nabla^2\phi^*(y_t)]^{-1}}dW_t \;\text{ for }\; y_0 = \nabla\phi(x_{k+1/2}) \qquad (\clubsuit) \\ x_{k+1} = \nabla\phi^*(y_h) \end{cases}$$

Backward: $(x_{k+1} = x_k - h \cdot \nabla f(x_{k+1}) = \arg\min_x f(x) + 1/2h \cdot \|x - x_k\|_2^2)$

$$\begin{cases} \text{solve } dy_t = \sqrt{2[\nabla^2\phi^*(y_t)]^{-1}}dW_t \;\text{ for }\; y_0 = \nabla\phi(x_k) \qquad (\clubsuit) \\ x_{k+1} = \arg\min_v hf(v) + \phi(v) - y_h^\top v \\ \qquad \Leftrightarrow x_{k+1} = \nabla\phi^*(y_h - h\nabla f(x_{k+1})) \end{cases}$$

Brownian motion part ($\clubsuit$) can be solved approximately with EM.

What this particular analysis suggests [1]

In KL divergence [1]

- EM has irreducible bias w.r.t diminishing step size h
- Forward discretization has slower rate and requires stronger assumption for convergence (Hessian stability of ϕ)
- Backward discretization requires somewhat weaker assumption and has faster rate ($\mathcal{O}(1/\epsilon)$ vs. $\mathcal{O}(1/\sqrt{\epsilon})$)

[1] Required assumption: relative smoothness, mirror log-Sobolev

What this particular analysis suggests [1]

In KL divergence

- EM has irreducible bias w.r.t diminishing step size h
- Forward discretization has slower rate and requires stronger assumption for convergence (Hessian stability of ϕ)
- Backward discretization requires somewhat weaker assumption and has faster rate ($\mathcal{O}(1/\epsilon)$ vs. $\mathcal{O}(1/\sqrt{\epsilon})$)

Proof use interpolation argument for the discrete updates $\rightsquigarrow$ construct another SDE agreeing with the update at time $t = h, 2h, \ldots,$ e.g., for EM update

$$d\tilde{Y}_t = -\nabla f(\nabla \phi^*(\tilde{Y}_0))dt + \sqrt{2(\nabla^2 \phi^*(\tilde{Y}_0))^{-1}} dW_t\,, \quad t \in [0, h]$$

$\rightsquigarrow$ corresponds to **perturbed density PDE** for $\tilde{X}_t$ whose asymptotic bias and contraction we can bound with a recursion on $KL(\rho_h \| \pi)$ using relative smoothness assumptions etc.

Pathspace forward-backward perspective: Schrödinger Bridge

Diffusion Generative Modeling and Time Reversal SDE - I

Goal

Given many samples from a complex distribution π, generate more samples from it. One does not have access to analytical expression for π, i.e., can't compute ∇.

Diffusion Generative Modeling and Time Reversal SDE - I

With two path measures represented as (π is target, ν simple e.g., $\mathcal{N}(0, I)$)

$$dX_t = \sigma u_t(X_t)dt + \sigma \overrightarrow{dW_t}, \; X_0 \sim \nu \Rightarrow (X_t)_t \sim \overrightarrow{\mathbb{P}}^{\nu,\sigma u} \tag{6}$$

$$dX_t = \sigma v_t(X_t)dt + \sigma \overleftarrow{dW_t}, \; X_T \sim \pi \Rightarrow (X_t)_t \sim \overleftarrow{\mathbb{P}}^{\pi,\sigma v} \tag{7}$$

Interested in learning drifts u, v such that $D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\sigma u} || \overleftarrow{\mathbb{P}}^{\pi,\sigma v}) = 0$ or $D_{KL}(\overleftarrow{\mathbb{P}}^{\pi,\sigma v} || \overrightarrow{\mathbb{P}}^{\nu,\sigma u}) = 0$:

$$\boxed{\nu(x) \underset{\mathbb{P}^{\pi,\sigma v}}{\overset{\mathbb{P}^{\nu,\sigma u}}{\rightleftarrows}} \pi(x)}$$

Diffusion Generative Modeling and Time Reversal SDE - I

With two path measures represented as (π is target, ν simple e.g., $\mathcal{N}(0, I)$)

$$dX_t = \sigma u_t(X_t)dt + \sigma \overrightarrow{dW_t},\ X_0 \sim \nu \Rightarrow (X_t)_t \sim \overrightarrow{\mathbb{P}}^{\nu,\sigma u} \tag{6}$$

$$dX_t = \sigma v_t(X_t)dt + \sigma \overleftarrow{dW_t},\ X_T \sim \pi \Rightarrow (X_t)_t \sim \overleftarrow{\mathbb{P}}^{\pi,\sigma v} \tag{7}$$

Interested in learning drifts u, v such that $D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\sigma u} || \overleftarrow{\mathbb{P}}^{\pi,\sigma v}) = 0$ or $D_{KL}(\overleftarrow{\mathbb{P}}^{\pi,\sigma v} || \overrightarrow{\mathbb{P}}^{\nu,\sigma u}) = 0$:

$$\boxed{\nu(x) \underset{\mathbb{P}^{\pi,\sigma v}}{\overset{\mathbb{P}^{\nu,\sigma u}}{\rightleftarrows}} \pi(x)}$$

Such forward/backward process is not unique but u^* can be used to transport ν to π.

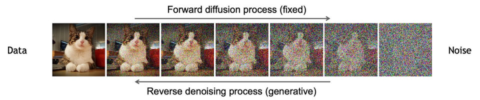


Figure 1: Diffusion generative modeling

Diffusion Generative Modeling and Time Reversal SDE - II

Useful Fact

Path measures $\overrightarrow{\mathbb{P}}^{\nu,\sigma u} = \overleftarrow{\mathbb{P}}^{\pi,\sigma v}$ iff $\sigma v_t(X_t) = \sigma u_t(X_t) - \sigma^2 \nabla \log(\overleftarrow{\mathbb{P}}_t^{\pi,\sigma v})\ \forall t$.

$$\nu(x) \quad \underset{\mathbb{P}^{\pi,\sigma v}}{\overset{\mathbb{P}^{\nu,\sigma u} = \mathbb{P}^{\nu,\sigma v + \sigma^2 \nabla \log \rho_t}}{\rightleftarrows}} \quad \pi(x)$$

Generative modeling fix one part of the process $\overleftarrow{\mathbb{P}}^{\pi,\sigma v}$ (e.g., OU) and learn the other using data from $\pi \rightsquigarrow$ score matching loss:

$$\arg\min_s D_{KL}(\overleftarrow{\mathbb{P}}^{\pi,\sigma v} \| \overrightarrow{\mathbb{P}}^{\nu,\sigma v + \sigma^2 s}) = \arg\min_s \mathbb{E}_{\overleftarrow{\mathbb{P}}^{\pi,\sigma v}} \left[\int_0^T \frac{\sigma^2}{2} \|s_t(X_t)\|^2 \, dt + \sigma^2 \int_0^T \nabla \cdot s_t(X_t) \, dt \right]$$

where $s^* = \nabla \log \rho_t$ score function upon convergence (impose time-reversal consistency).

Diffusion Generative Modeling and Time Reversal SDE - II

Useful Fact

Path measures $\overrightarrow{\mathbb{P}}^{\nu,\sigma u} = \overleftarrow{\mathbb{P}}^{\pi,\sigma v}$ iff $\sigma v_t(X_t) = \sigma u_t(X_t) - \sigma^2 \nabla \log(\overleftarrow{\mathbb{P}}_t^{\pi,\sigma v})\ \forall t$.

$$\nu(x) \underset{\mathbb{P}^{\pi,\sigma v}}{\overset{\mathbb{P}^{\nu,\sigma u} = \mathbb{P}^{\nu,\sigma v + \sigma^2 \nabla \log \rho_t}}{\rightleftarrows}} \pi(x)$$

Generative modeling fix one part of the process $\overleftarrow{\mathbb{P}}^{\pi,\sigma v}$ (e.g., OU) and learn the other using data from $\pi \rightsquigarrow$ score matching loss:

$$\arg\min_s D_{KL}(\overleftarrow{\mathbb{P}}^{\pi,\sigma v} \| \overrightarrow{\mathbb{P}}^{\nu,\sigma v + \sigma^2 s}) = \arg\min_s \mathbb{E}_{\overleftarrow{\mathbb{P}}^{\pi,\sigma v}} \left[\int_0^T \frac{\sigma^2}{2} \|s_t(X_t)\|^2 \, dt + \sigma^2 \int_0^T \nabla \cdot s_t(X_t) \, dt \right]$$

where $s^* = \nabla \log \rho_t$ score function upon convergence (impose time-reversal consistency).

Generating new samples from π is easy once we know $\hat{u} = v + \sigma \hat{s}$. But errors from (1) initialization $\overleftarrow{\mathbb{P}}_0^{\pi,\sigma v} \neq \nu$; (2) discretization of SDE; (3) estimator $\hat{u}$, i.e., $\mathbb{E} \to \sum$.

Sampling from a Path-wise perspective

Sampling can also be viewed as learning a transition path $(\rho_t)_t$, but we don't have samples from π: (1) do reverse KL to enforce the marginal; (2) introduce a reference process that facilitate likelihood-ratio calculation in $D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\sigma u} \| \overleftarrow{\mathbb{P}}^{\pi,\sigma v})$

$$\textbf{sampling: } \nu(x) \underset{\mathbb{P}^{\pi,\sigma \mathrm{v}}}{\overset{\mathbb{P}^{\nu,\sigma \mathrm{u}}}{\rightleftarrows}} \pi(x) \quad \text{and} \quad \textbf{reference: } \nu(x) \underset{\mathbb{P}^{\eta,\sigma \mathrm{v}}}{\overset{\mathbb{P}^{\nu,\sigma \mathrm{r}}}{\rightleftarrows}} \eta(x)$$

To obtain tractable estimate of KL to train u for sampling: ref can be OU in equilibrium

Sampling from a Path-wise perspective

Sampling can also be viewed as learning a transition path $(\rho_t)_t$, but we don't have samples from π: (1) do reverse KL to enforce the marginal; (2) introduce a reference process that facilitate likelihood-ratio calculation in $D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\sigma u} \| \overleftarrow{\mathbb{P}}^{\pi,\sigma v})$

$$\textbf{sampling: } \nu(x) \xrightleftharpoons[\mathbb{P}^{\pi,\sigma \mathbf{v}}]{\mathbb{P}^{\nu,\sigma \mathbf{u}}} \pi(x) \quad \text{and} \quad \textbf{reference: } \nu(x) \xrightleftharpoons[\mathbb{P}^{\eta,\sigma \mathbf{v}}]{\mathbb{P}^{\nu,\sigma \mathbf{r}}} \eta(x)$$

To obtain tractable estimate of KL to train u for sampling: ref can be OU in equilibrium

$$D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\sigma u} \| \overleftarrow{\mathbb{P}}^{\pi,\sigma v}) = \mathbb{E}_{\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}\left[\log\left(\frac{d\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}{d\overleftarrow{\mathbb{P}}^{\pi,\sigma v}}\right)\right] = \mathbb{E}_{\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}\left[\log\left(\frac{d\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}{d\overrightarrow{\mathbb{P}}^{\nu,\sigma r}} \frac{d\overleftarrow{\mathbb{P}}^{\eta,\sigma v}}{d\overleftarrow{\mathbb{P}}^{\pi,\sigma v}}\right)\right]$$

Sampling from a Path-wise perspective

Sampling can also be viewed as learning a transition path $(\rho_t)_t$, but we don't have samples from π: (1) do reverse KL to enforce the marginal; (2) introduce a reference process that facilitate likelihood-ratio calculation in $D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\sigma u} \| \overleftarrow{\mathbb{P}}^{\pi,\sigma v})$

$$\textbf{sampling: } \nu(x) \underset{\mathbb{P}^{\pi,\sigma v}}{\overset{\mathbb{P}^{\nu,\sigma u}}{\rightleftarrows}} \pi(x) \quad \text{and} \quad \textbf{reference: } \nu(x) \underset{\mathbb{P}^{\eta,\sigma v}}{\overset{\mathbb{P}^{\nu,\sigma r}}{\rightleftarrows}} \eta(x)$$

To obtain tractable estimate of KL to train u for sampling: ref can be OU in equilibrium

$$D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\sigma u} \| \overleftarrow{\mathbb{P}}^{\pi,\sigma v}) = \mathbb{E}_{\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}\left[\log\left(\frac{d\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}{d\overleftarrow{\mathbb{P}}^{\pi,\sigma v}}\right)\right] = \mathbb{E}_{\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}\left[\log\left(\frac{d\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}{d\overrightarrow{\mathbb{P}}^{\nu,\sigma r}} \frac{d\overleftarrow{\mathbb{P}}^{\eta,\sigma v}}{d\overleftarrow{\mathbb{P}}^{\pi,\sigma v}}\right)\right]$$

$$= \mathbb{E}_{X\sim\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}\left[\int_0^T \frac{1}{2}\|u_s(X_s) - r_s(X_s)\|^2 ds + \log\left(\frac{d\eta}{d\pi}\right)(X_T)\right] =: \mathcal{L}_{KL}(u)$$

Sampling from a Path-wise perspective

$$\textbf{sampling: } \nu(x) \underset{\mathbb{P}^{\pi,\sigma\mathrm{v}}}{\overset{\mathbb{P}^{\nu,\sigma\mathrm{u}}}{\rightleftarrows}} \pi(x) \quad \text{and} \quad \textbf{reference: } \nu(x) \underset{\mathbb{P}^{\eta,\sigma\mathrm{v}}}{\overset{\mathbb{P}^{\nu,\sigma\mathrm{r}}}{\rightleftarrows}} \eta(x)$$

To obtain tractable estimate of KL to train u for sampling: ref can be OU in equilibrium

$$D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\sigma u} \| \overleftarrow{\mathbb{P}}^{\pi,\sigma v}) = \mathbb{E}_{\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}\left[\log\left(\frac{d\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}{d\overleftarrow{\mathbb{P}}^{\pi,\sigma v}}\right)\right] = \mathbb{E}_{\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}\left[\log\left(\frac{d\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}{d\overrightarrow{\mathbb{P}}^{\nu,\sigma r}}\frac{d\overleftarrow{\mathbb{P}}^{\eta,\sigma v}}{d\overleftarrow{\mathbb{P}}^{\pi,\sigma v}}\right)\right]$$

$$= \mathbb{E}_{X\sim\overrightarrow{\mathbb{P}}^{\nu,\sigma u}}\left[\int_0^T \frac{1}{2}\|u_s(X_s) - r_s(X_s)\|^2 ds + \log\left(\frac{d\eta}{d\pi}\right)(X_T)\right] =: \mathcal{L}_{KL}(u)$$

$\rightsquigarrow$ we fixed v (therefore r) so solution $\min_u \mathcal{L}_{KL}(u)$ is unique

$\rightsquigarrow$ Algorithm: trajectory rollout with current u, estimate loss & update u, repeat

To sample: run $X_{k+1} = X_k + h\sigma\hat{u}_k(X_k) + \sqrt{h}\sigma z_k$ from $X_0 \sim \nu$.

Schrödinger Bridge - I

Both rely on fixing some aspect of the forward-backward process to impose uniqueness.

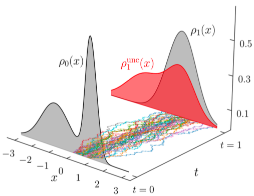
Motivating Question

What if we don't want just one path interpolating from ν to π, but a particular (e.g., in some sense *optimal*) one? After all, there is no special meaning to the reference process in the previous slide beyond convenience.

Schrödinger's thought experiment (30s): Observe $\rho_1(y) \neq \int P(0,x,1,y)\rho_0(x)dx$ for

$$P(0,x,1,y) = (2\pi)^{-d/2}\exp(-\|x-y\|^2/2) \quad \text{the heat kernel}$$

Of the many unlikely ways in which it could have happened, which one is the most likely?

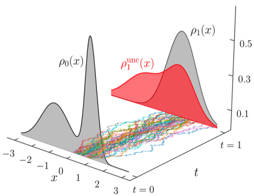


Schrödinger Bridge - I

Schrödinger's thought experiment (30s): Observe $\rho_1(y) \neq \int P(0,x,1,y)\rho_0(x)dx$ for

$$P(0,x,1,y) = (2\pi)^{-d/2}\exp(-\|x-y\|^2/2) \quad \text{the heat kernel}$$

Of the many unlikely ways in which it could have happened, which one is the most likely?



Classical formulation:

$$P^* = \arg\min_{P_0=\nu, P_T=\pi} D_{KL}(P\|Q)$$

for simplicity assume base measure Q admits SDE representation (i.e., Wiener process):

$$dX_t = \sigma dW_t\,, \quad X_0 \sim \nu\,.$$

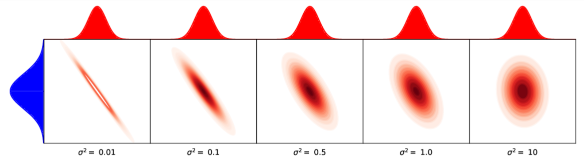
Schrödinger Bridge - II

Will not show but quote the following two additional perspectives on P^* here:

(1) Entropy-regularized optimal transport: the joint distribution P^*_{0T} at time $0, T$ solves

$$\rho^*(x_0, x_T) = \underbrace{\arg \min_{\rho \in \Pi(\nu,\pi)} \int \|x_0 - x_T\|^2 \rho(x_0, x_T)\, dx_0\, dx_T}_{\mathcal{W}_2} + 2\sigma T \int \rho(x_0, x_T) \log \rho(x_0, x_T)\, dx_0\, dx_T$$

Optimal solution takes the form $P^*(x_{0:T}) = \rho^*(x_0, x_T) Q(\cdot | x_0, x_T) \rightsquigarrow$ mixture of pinned diffusions w/ weight ρ^*.



Schrödinger Bridge - II

Will not show but quote the following two additional perspectives on P^* here:

(1) Entropy-regularized optimal transport: the joint distribution P^*_{0T} at time $0, T$ solves

$$\rho^*(x_0, x_T) = \underbrace{\arg\min_{\rho\in\Pi(\nu,\pi)} \int \|x_0 - x_T\|^2 \rho(x_0, x_T)\, dx_0\, dx_T}_{\mathcal{W}_2} + 2\sigma T \int \rho(x_0, x_T) \log \rho(x_0, x_T)\, dx_0\, dx_T$$

Optimal solution takes the form $P^*(x_{0:T}) = \rho^*(x_0, x_T) Q(\cdot | x_0, x_T) \rightsquigarrow$ mixture of pinned diffusions w/ weight ρ^*.

Challenge

Classical method for solving EOT: Sinkhorn iterative projection

$$P^{(1)} = \arg\min_{Q\in\mathcal{P}(\nu,\cdot)} D_{KL}(Q\|P^{(0)}) \propto P^{(0)}\nu, \quad P^{(0)} = \arg\min_{Q\in\mathcal{P}(\cdot,\pi)} D_{KL}(Q\|P^{(1)}) \propto P^{(1)}\pi$$

but we don't have samples from π so can't implement second part as it is.

Schrödinger Bridge - III

(2) Stochastic optimal control: P^* is driven by the control u^*, which solves

$$\inf_u \ \mathbb{E}_u\left[\int_0^T \frac{1}{2}\|u_t(X_t)\|^2 dt\right]$$

$$\text{s.t. } dX_t = \sigma u_t(X_t)dt + \sigma dW_t, X_0 \sim \nu, X_T \sim \pi$$

$\rightsquigarrow$ minimum control effort steering ν to π. Dynamics reaches target in finite time.

Schrödinger Bridge - III

(2) Stochastic optimal control: P^* is driven by the control u^*, which solves

$$\inf_u \ \mathbb{E}_u \left[\int_0^T \frac{1}{2} \|u_t(X_t)\|^2 dt \right]$$

$$\text{s.t. } dX_t = \sigma u_t(X_t)dt + \sigma dW_t, X_0 \sim \nu, X_T \sim \pi$$

$\rightsquigarrow$ minimum control effort steering ν to π. Dynamics reaches target in finite time.

Challenge

Would like to avoid solving high-dimensional HJB PDEs for the control $\rightsquigarrow$ it is not clear it's much cheaper than performing high-dimensional sampling

Sampling as an optimal control / transport over path-space [3]

Add regularizer to D_{KL} $\rightsquigarrow$ This imposes uniqueness, terminal marginals, and fulfills a reversible noising/denoising function in an "optimal" way: (a two-parameter loss)

$$\arg\min_{\nabla u, \nabla v} D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\nabla u} \| \overleftarrow{\mathbb{P}}^{\pi,-\nabla v}) \to \mathbb{E}\left[\int_0^T \frac{1}{2}\|\nabla u_t(X_t) + \nabla v_t(X_t)\|^2 + \Delta v_t(X_t) dt + \log \frac{\nu(X_0)}{\pi(X_T)}\right]$$
$$+\mathrm{Var}\left(u_T(X_T) - u_0(X_0) + \frac{1}{2}\int_0^T \|\nabla u_t\|^2(X_t)\, dt - \int_0^T \nabla u_t(X_t)^\top dW_t\right) \qquad (\spadesuit)$$

Regularizer on the forward/backward control $\nabla u, \nabla v$ can be done in various ways.

$$dX_t = \nabla u_t(X_t) dt + \overrightarrow{dW_t}, X_0 \sim \nu\,,$$
$$dX_t = -\nabla v_t(X_t) dt + \overleftarrow{dW_t},\ X_T \sim \pi\,.$$

Sampling as an optimal control / transport over path-space [3]

Add regularizer to D_{KL} $\rightsquigarrow$ This imposes uniqueness, terminal marginals, and fulfills a reversible noising/denoising function in an "optimal" way: (a two-parameter loss)

$$\arg\min_{\nabla u, \nabla v} D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\nabla u} \| \overleftarrow{\mathbb{P}}^{\pi,-\nabla v}) \to \mathbb{E}\left[\int_0^T \frac{1}{2}\|\nabla u_t(X_t) + \nabla v_t(X_t)\|^2 + \Delta v_t(X_t) dt + \log\frac{\nu(X_0)}{\pi(X_T)}\right]$$
$$+\mathrm{Var}\left(u_T(X_T) - u_0(X_0) + \frac{1}{2}\int_0^T \|\nabla u_t\|^2(X_t)\,dt - \int_0^T \nabla u_t(X_t)^\top dW_t\right) \qquad (\spadesuit)$$

Regularizer on the forward/backward control $\nabla u, \nabla v$ can be done in various ways.

($\spadesuit$) exploit the important *factorization* property for the optimal coupling:

$$\rho^*(x_0, x_T) = f(x_0)\rho^0(x_0, x_T)g(x_T)$$

$\to$ we know how f, g behave for SB, and these are also related to controls/drifts $\nabla u, \nabla v$ in the SDE $\rightsquigarrow$ try to optimize w.r.t those criteria

Sampling as an optimal control / transport over path-space [3]

Add regularizer to D_{KL} $\rightsquigarrow$ This imposes uniqueness, terminal marginals, and fulfills a reversible noising/denoising function in an “optimal” way: (a two-parameter loss)

$$\arg\min_{\nabla u, \nabla v} D_{KL}(\overrightarrow{\mathbb{P}}^{\nu,\nabla u} \| \overleftarrow{\mathbb{P}}^{\pi,-\nabla v}) \to \mathbb{E}\left[\int_0^T \frac{1}{2}\|\nabla u_t(X_t) + \nabla v_t(X_t)\|^2 + \Delta v_t(X_t) dt + \log \frac{\nu(X_0)}{\pi(X_T)}\right]$$

$$+\mathrm{Var}\left(u_T(X_T) - u_0(X_0) + \frac{1}{2}\int_0^T \|\nabla u_t\|^2(X_t)\, dt - \int_0^T \nabla u_t(X_t)^\top dW_t\right) \qquad (\spadesuit)$$

Regularizer on the forward/backward control $\nabla u, \nabla v$ can be done in various ways.

Algorithm: Alternate between:

(1) simulate trajectory with current control ∇u;

(2) optimize/update the neural-network parameterized controls $\nabla u, \nabla v$ with the estimated loss above. $\rightsquigarrow$ if loss = 0, the controls found must be optimal.

Some Generalizations Possible

- Deterministic ODE dynamics for sampling:

$$dX_t = \frac{1}{2}(\nabla \log \hat{u}_t(X_t) - \nabla \log \hat{v}_t(X_t))\, dt$$

- Importance sampling to correct for imperfect controls:

$$\mathbb{E}_u[g(X_T)w^u(X_T)] = \mathbb{E}_{u*}[g(X_T)] = \mathbb{E}_\pi[g]$$

- Path-wise divergence doesn't have to be KL, e.g., log-variance divergence.
- Estimation of normalizing constant Z for π available.

Conclusion

Sampling from un-normalized density, interestingly,

- Connects to many things - wasn't expecting such when embarking on the journey but turned out to be a nice surprise

Conclusion

Sampling from un-normalized density, interestingly,

- Connects to many things - wasn't expecting such when embarking on the journey but turned out to be a nice surprise
- Different perspectives are helpful, from classical to contemporary, with physics seems to be scattered throughout

Conclusion

Sampling from un-normalized density, interestingly,

- Connects to many things - wasn't expecting such when embarking on the journey but turned out to be a nice surprise
- Different perspectives are helpful, from classical to contemporary, with physics seems to be scattered throughout
- Open questions abound, from the most theoretical to the most practical, and everything in between

Conclusion

Sampling from un-normalized density, interestingly,

- Connects to many things - wasn't expecting such when embarking on the journey but turned out to be a nice surprise
- Different perspectives are helpful, from classical to contemporary, with physics seems to be scattered throughout
- Open questions abound, from the most theoretical to the most practical, and everything in between
- Various scientific pursuits crucially rely on good numerical sampling procedure: lattice QCD, molecular dynamics etc.,

Thanks!
Questions?

References i

Q. Jiang.
Mirror Langevin Monte Carlo: the Case Under Isoperimetry, 2021.

Q. Jiang.
On the Dissipation of Ideal Hamiltonian Monte Carlo Sampler, 2022.

Q. Jiang.
Control, Transport and Sampling: The Benefit of a Reference Process, 2023.

Dissipation of Hamiltonian Monte Carlo Sampler

Unifying framework

- Overdamped Langevin is the high-friction limit of the underdamped Langevin:

$$dX_t = V_t dt$$

$$dV_t = -\nabla f(X_t)dt \overbrace{-\gamma V_t dt + \sqrt{2\gamma} dW_t}$$

- Invariant measure $\propto e^{-f(X)-\frac{1}{2}\|V\|^2} = \pi(X) \otimes \mathcal{N}(0, I) \rightarrow$ take marginal over X
- Conservation of Hamiltonian $H(X, V) = f(X) + \frac{1}{2}\|V\|^2$ along

$$\dot{X}_t = \frac{\partial H}{\partial v} = V_t, \quad \dot{V}_t = -\frac{\partial H}{\partial x} = -\nabla f(X_t) \Leftarrow \text{define flow map: } \phi_t(X_0, V_0) = (X_t, V_t) \tag{8}$$

- Ornstein-Uhlenbeck process

$$dV_t = -\gamma V_t dt + \sqrt{2\gamma} dW_t \tag{9}$$

can be integrated exactly as $V_t = e^{-\gamma t} V_0 + \sqrt{1 - e^{-2\gamma t}} Z$
- The term γ introduces damping, related to fluctuation-dissipation
- Dynamics can be used for optimizing $(X_t, V_t) \rightarrow (X^*, 0)$ if no stochasticity
- Second-order SDE with friction / memory, better mixing with proper discretization

HMC and Ergodicity

Classical HMC:

1. Follow deterministic flow ϕ_t (8) for time $T \leftarrow$ known from practice (NUTS) that performance very sensitive to trajectory length T and hard to tune
2. Redraw the velocity $V_t = Z \sim \mathcal{N}(0, I)$

Imagine an ensemble of particles (take potential $f(x) = \|x\|^2$): what if we initialize at stationarity (most around center)? What if not?

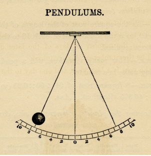


Figure 2: The importance of refreshment as illustrated by a harmonic oscillator.

Ergodic: unique invariant measure (initial ρ_0 is eventually forgotten), or equivalently

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T f(x_t)dt = \int_{\mathbb{R}^d} f(x)\pi(x)dx$$

Dissipation of the Dynamics - I

Two extremes:

- T too short: short deterministic dynamics $\rightsquigarrow$ diffusive behavior

- T too long: periodic therefore we backtrack on the progress made

Assuming smooth, strongly-convex potential, i.e.,

$$\mu \cdot I \preceq \nabla^2 f \preceq L \cdot I, \quad \kappa := L/\mu$$

Previous work show for $T = 1/\sqrt{L}$, mixing time is $\mathcal{O}(\kappa \log(1/\epsilon))$ and this is tight in W_2.

What if we do partial refreshment (as inspired by under-damped Langevin)?

1. Follow deterministic flow ϕ_t for time T
2. Redraw the velocity $V_t = \eta V_0 + \sqrt{1-\eta^2} Z$

What if we randomize the integration time?

1. Follow deterministic flow ϕ_t for time $T \sim \text{Pois}(\lambda) \leftarrow$ jump process
2. Redraw the velocity $V_t = Z$

Dissipation of the Dynamics - II

Key Observation [2]

Both give improved performance by $\sqrt{\kappa}$ factor, and the crucial quantity is

$$\lambda^{-1}(1-\eta^2) \approx \sqrt{\mu}$$

with either $\lambda^{-1} = \sqrt{\mu}, \eta = 0$ or $1-\eta^2 = \sqrt{\mu}/\sqrt{L}, \lambda^{-1} = \sqrt{L}$, which compared to classical

$$\lambda^{-1}(1-\eta^2) \approx \sqrt{L}$$

when $\eta = 0, \lambda^{-1} = \sqrt{L}$ can be much smaller, i.e., more memory.

Proof use (synchronous) coupling of two chains X, Y, challenge is using the right Lyapunov function over extended state-space $\mathbb{R}^{2d}$ for contraction such that

(1) $\mathbb{E}[d(X_{k+1}, Y_{k+1})] \le e^{-c}\mathbb{E}[d(X_k, Y_k)]$; (2) $c_1\|X-Y\|^2 \le d(X,Y) \le c_2\|X-Y\|^2$.

Unlike over-damped Langevin, second-order dynamics require smoothness $\nabla^2 f \preceq L \cdot I$ for convergence even in continuous time.

Discretized Algorithm

Symplectic integrator simulate long trajectory w/o incur much err (preserve phase space volume) $\rightsquigarrow$ for Gaussian, there's a "shadow Hamiltonian" the discrete dynamics preserve $\rightsquigarrow$ invariant measure is another Gaussian with shifted mean $\rightsquigarrow$ bias $\mathcal{O}(\sqrt{d}h^2)$ in W_2

One gradient call, leapfrog (i.e., Verlet) is composition of trapezoidal & implicit midpoint:

$$
\begin{aligned}
x_{k+1/2} &= x_k + h/2 \cdot v_k \\
v_{k+1} &= v_k - h \cdot \nabla f(x_{k+1/2}) \\
x_{k+1} &= x_{k+1/2} + h/2 \cdot v_{k+1}
\end{aligned}
$$

Update can be rewritten as ($x_{k+1/2} = x_k + h/2 \cdot v_k$)

$$
\begin{aligned}
v_{k+1} &= v_k - h \cdot \nabla f(x_{k+1/2}) \\
x_{k+1} &= x_k + h \cdot v_k - h^2/2 \cdot \nabla f(x_{k+1/2})
\end{aligned} \tag{10}
$$

a randomized choice of $x_{k+1/2}$ (i.e., random stepsize h) better (smaller asymptotic bias).

Can perform adjustment based on energy error $H(X_T, V_T) - H(X_0, V_0)$.

Comparisons

A general splitting scheme (free parameters K, h, η recover different algorithms, $T := Kh$):

$$O: \; v_k = \eta/2 \cdot v_k + \sqrt{1-(\eta/2)^2} \cdot Z$$

$$\text{K-times, deterministic}\begin{cases} A: \; x_{k+1/2} = x_k + h/2 \cdot v_k \\ B: \; v_{k+1} = v_k - h \cdot \nabla f(x_{k+1/2}) \\ A: \; x_{k+1} = x_{k+1/2} + h/2 \cdot v_{k+1} \end{cases}$$

$$O: \; v_{k+1} = \eta/2 \cdot v_{k+1} + \sqrt{1-(\eta/2)^2} \cdot Z'$$

- Underdamped Langevin: h determined by bias of ABA part, $K = 1$ so $T = h$, $\eta = e^{-\gamma h} \approx 1 - \gamma h$, γ needs to be sufficiently large for contraction
- HMC: h determined by bias of ABA part, $K = T/h$ steps of leapfrog, η can be either 0 or not, depending on how we pick T (upper bound on $\eta + T$ to contract)
- Overdamped Langevin: From (10) can see by picking $\eta = 0$ and $K = 1$, equivalent to overdamped Langevin with stepsize $h^2/2$ and full refreshment, i.e., $v_k \sim \mathcal{N}(0, I)$ always