Novel perturbations for accelerating Langevin samplers

Konstantinos Spiliopoulos

Department of Mathematics & Statistics, Boston University
Partially supported by NSF-DMS and SIMONS Foundation
Joint Work with Ben Zhang (Umass Amherst) and Youssef Marzouk (MIT)

Motivation: Bayesian inference

Bayesian inference challenges

$$\pi_{\text{posterior}}(x) = \pi(x \mid y) \propto \pi(y \mid x)\pi_{\text{prior}}(x)$$

- Evaluating the likelihood requires computing solution of a complex physical model
- Can only evaluate density up to normalization constant
- Parameters x high dimensional, posterior may be strongly non-Gaussian

Motivation: Approximating expectations

- Random variable X is distributed according to unnormalized *target* density $\pi(x)$ on $\mathbb{R}^d$
- **Goal**: Compute expectations $\mathbb{E}_\pi[f(X)]$

$$\mathbb{E}_\pi[f(X)] = \int_{\mathbb{R}^d} f(x)\pi(x)\mathrm{d}x$$

- **Monte Carlo simulation**: produces i.i.d. *samples* $X_i \sim \pi$, estimate $\mathbb{E}_\pi[f(X)]$

$$\rho \approx \hat{\rho} = \frac{1}{K}\sum_{i=1}^{K} f(X_i)$$

Question: How to produce samples X_i?

Outline

Overdamped Langevin dynamics

$$dX_t = \beta \nabla \log \pi(X_t)dt + \sqrt{2\beta}dW_t$$

- $\pi(x)$ is the (unnormalized) target density on $\mathbb{R}^d$; $\beta > 0$ is the temperature
- Ergodicity: $X_t \sim \pi$ as $t \to \infty$, and

$$\mathbb{E}_\pi[f(X)] = \int_{\mathbb{R}^d} f(x)\pi(x)dx = \lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_t)dt$$

Overdamped Langevin dynamics

$$dX_t = \beta \nabla \log \pi(X_t) dt + \sqrt{2\beta} dW_t$$

- $\pi(x)$ is the (unnormalized) target density on $\mathbb{R}^d$; $\beta > 0$ is the temperature
- Ergodicity: $X_t \sim \pi$ as $t \to \infty$, and

$$\mathbb{E}_\pi [f(X)] = \int_{\mathbb{R}^d} f(x)\pi(x) dx = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(X_t) dt$$

Unadjusted Langevin algorithm (direct discretization of LD)

$$X_{k+1} = X_k + h\beta \nabla \log \pi(X_k) + \sqrt{2\beta h}\xi_k; \quad \xi_k \sim \mathcal{N}(0, \mathbf{I})$$

$$\mathbb{E}_\pi [f(x)] \approx \frac{1}{K} \sum_{k=0}^{K-1} f(X_k)$$

Unadjusted Langevin algorithm guarantees for log-concave densities

Unadjusted Langevin algorithm

$$X_{k+1} = X_k + \beta h \nabla \log \pi(X_k) + \sqrt{2\beta h}\xi_k; \quad \xi_k \sim \mathcal{N}(0, \mathbf{I})$$

$$\mathbb{E}_\pi [f(x)] \approx \frac{1}{K} \sum_{k=0}^{K-1} f(X_k)$$

Guarantees from [Durmus & Moulines 2019]

Let $U(x) = -\log \pi(x)$, assume $U(x) \in \mathcal{C}^2(\mathbb{R}^d)$, $X_k \sim \pi^k$. If

- $U(x)$ is **m-strongly convex**: $\nabla^2 U(x) \succeq m\mathbf{I}$
- $\nabla U(x)$ is **L-Lipschitz**: $\nabla^2 U(x) \preceq L\mathbf{I}$

then $\mathcal{W}_2^2(\pi^k, \pi) \leq C r^k \mathcal{W}_2^2(\pi^0, \pi) + F(m, L, h)$ with $r < 1$, and $F(m, L, h)$ is the bias

Unadjusted Langevin algorithm guarantees for log-concave densities

Guarantees from [Durmus & Moulines 2019]

Let $U(x) = -\log \pi(x)$, assume $U(x) \in \mathcal{C}^2(\mathbb{R}^d)$, $X_k \sim \pi^k$. If

- $U(x)$ is m-strongly convex: $\nabla^2 U(x) \succeq m\mathbf{I}$
- $\nabla U(x)$ is L-Lipschitz: $\nabla^2 U(x) \preceq L\mathbf{I}$

then $\mathcal{W}_2^2(\pi^k, \pi) \leq Cr^k \mathcal{W}_2^2(\pi^0, \pi) + F(m, L, h)$

- What happens if $U(\theta)$ does not satisfy these conditions?
  - Few theoretical guarantees, possible slow convergence

Unadjusted Langevin algorithm guarantees for log-concave densities

Guarantees from [Durmus & Moulines 2019]

Let $U(x) = -\log \pi(x)$, assume $U(x) \in \mathcal{C}^2(\mathbb{R}^d)$, $X_k \sim \pi^k$. If

- $U(x)$ is m-strongly convex: $\nabla^2 U(x) \succeq m\mathbf{I}$
- $\nabla U(x)$ is L-Lipschitz: $\nabla^2 U(x) \preceq L\mathbf{I}$

then $\mathcal{W}_2^2(\pi^k, \pi) \leq Cr^k \mathcal{W}_2^2(\pi^0, \pi) + F(m, L, h)$

- What happens if $U(\theta)$ does not satisfy these conditions?
  - Few theoretical guarantees, possible slow convergence
- **Novel perturbations to Langevin dynamics** can accelerate convergence
- **Transport map ULA** relaxes some conditions, provides some guarantees

Perturbations accelerate convergence of Langevin dynamics

Reversible perturbations (RMLD)

$$dX_t = [\beta \mathbf{B}(X_t)\nabla \log \pi(X_t) + \nabla \cdot \mathbf{B}(X_t)]\, dt + \sqrt{2\beta \mathbf{B}(X_t)} dW_t$$

- $\mathbf{B}(x) = \mathbf{B}(x)^\top$, $\mathbf{B}(x) \succ \mathbf{0}$.

Perturbations accelerate convergence of Langevin dynamics

Reversible perturbations (RMLD)

$$dX_t = [\beta \mathbf{B}(X_t)\nabla \log \pi(X_t) + \nabla \cdot \mathbf{B}(X_t)]\, dt + \sqrt{2\beta \mathbf{B}(X_t)} dW_t$$

- $\mathbf{B}(x) = \mathbf{B}(x)^\top$, $\mathbf{B}(x) \succ \mathbf{0}$. In continuous-time, *no optimal choice of* $\mathbf{B}(x)$, but if $\mathbf{B}(x) - \mathbf{I} \succ \mathbf{0}$, then obtain accelerated convergence [Rey-Bellet & Spiliopoulos 2016]

Perturbations accelerate convergence of Langevin dynamics

Reversible perturbations (RMLD)

$$dX_t = [\beta \mathbf{B}(X_t)\nabla \log \pi(X_t) + \nabla \cdot \mathbf{B}(X_t)]\, dt + \sqrt{2\beta \mathbf{B}(X_t)} dW_t$$

- $\mathbf{B}(x) = \mathbf{B}(x)^\top$, $\mathbf{B}(x) \succ \mathbf{0}$. In continuous-time, *no optimal choice of* $\mathbf{B}(x)$, but if $\mathbf{B}(x) - \mathbf{I} \succ \mathbf{0}$, then obtain accelerated convergence [Rey-Bellet & Spiliopoulos 2016]
- **Riemmanian manifold Langevin dynamics**: $\mathbf{B}(x) = \mathbf{G}(x)^{-1}$, inner product $g_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$, $g_x(u, v) = \langle \mathbf{G}(x)u, v\rangle$. Inspired by information geometry

Perturbations accelerate convergence of Langevin dynamics

Irreversible perturbations (Irr)

$$dX_t = [\beta \nabla \log \pi(X_t) + \gamma(X_t)]\, dt + \sqrt{2\beta} dW_t$$

- Condition on $\gamma(x)$ so that target is held invariant: $\nabla \cdot (\gamma(x)\pi(x)) = 0$.
- Simple choice: $\gamma(x) = \mathbf{D}\nabla \log \pi(x)$, $\mathbf{D} = -\mathbf{D}^\top$.
- In continuous-time, *will always improve convergence* [Rey-Bellet & Spiliopoulos 2016]

Perturbations accelerate convergence of Langevin dynamics

$$\mathrm{d}X_t = [\beta \mathbf{B}(X_t)\nabla \log \pi(X_t) + \nabla \cdot \mathbf{B}(X_t)]\,\mathrm{d}t + \sqrt{2\beta \mathbf{B}(X_t)}\mathrm{d}W_t$$

$$\mathrm{d}X_t = [\beta \nabla \log \pi(X_t) + \gamma(X_t)]\,\mathrm{d}t + \sqrt{2\beta}\mathrm{d}W_t,\ \gamma(x) = \mathbf{D}\nabla \log \pi(x)$$

Theorem [Rey-Bellet & Spiliopoulos 2016]

Let $\mathcal{L}$ and $\mathcal{L}_0$ be the generators of RMLD and OLD. If $\mathbf{B}(x) - \mathbf{I} \succ \mathbf{0}$ or any $\mathbf{D} = -\mathbf{D}^\top$, then

- Spectral gap (leading nonzero eigenvalue of generator) decreases
- Asymptotic variance $\sigma^2(\phi) = \lim_{t\to\infty} t\mathrm{Var}\left(\frac{1}{t}\int_0^t \phi(X_t)dt\right)$ is smaller
- Large deviations rate function increases

Geometry-informed irreversible perturbations (GiIrr)

How to apply irreversibility to an already reversibly perturbed system?

Standard irreversibility applied to reversible perturbation (RMIrr)

$$\mathrm{d}X_t = [\beta \mathbf{B}(X_t)\nabla \log \pi(X_t) + \nabla \cdot \mathbf{B}(\theta_t) + \gamma(X_t)]\,\mathrm{d}t + \sqrt{2\beta \mathbf{B}(X_t)}\mathrm{d}W_t$$

$$\gamma(x) = \mathbf{D}\nabla \log \pi(x),\ \mathbf{D} = -\mathbf{D}^\top$$

Geometry-informed irreversible perturbations (GiIrr)

How to apply irreversibility to an already reversibly perturbed system?

Standard irreversibility applied to reversible perturbation (RMIrr)

$$\mathrm{d}X_t = [\beta \mathbf{B}(X_t)\nabla \log \pi(X_t) + \nabla \cdot \mathbf{B}(\theta_t) + \gamma(X_t)]\,\mathrm{d}t + \sqrt{2\beta \mathbf{B}(X_t)}\mathrm{d}W_t$$
$$\gamma(x) = \mathbf{D}\nabla \log \pi(x),\ \mathbf{D} = -\mathbf{D}^\top$$

Geometry-informed irreversibility (new!)

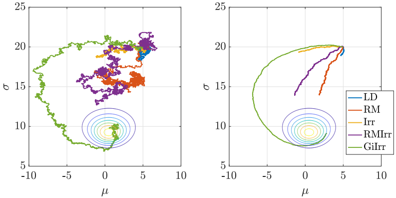
$$\mathrm{d}X_t = [\beta \mathbf{B}(X_t)\nabla \log \pi(X_t) + \nabla \cdot \mathbf{B}(X_t) + \gamma(X_t)]\,\mathrm{d}t + \sqrt{2\beta \mathbf{B}(X_t)}\mathrm{d}W_t$$
$$\gamma(x) = \mathbf{C}(x)\nabla \log \pi(x) + \nabla \cdot \mathbf{C}(x)$$
$$\mathbf{C}(x) = \frac{1}{2}\left[\mathbf{B}(x)\mathbf{D} + \mathbf{D}\mathbf{B}(x)\right],\ \text{note } \mathbf{C}(x) \text{ is still skew-symmetric!}$$

[Zhang, Marzouk, Spiliopoulos, Geometry-informed irreversible perturbations for accelerated convergence of Langevin dynamics, *Statistics and Computing*, 2022.]

Simple example: parameters of a normal distribution [Girolami 2011]

$$\log \pi(\mu, \sigma | \mathbf{X}) = \frac{N}{2} \log 2\pi - N \log \sigma - \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{2\sigma^2}$$

$$\mathbf{B}(\mu, \sigma) = \frac{\sigma^2}{N} \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix} \qquad \mathbf{D} = \delta \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$



	$\mathbb{E}[\mathrm{AVar}_\phi]$	$\mathrm{Std}[\mathrm{AVar}_\phi]$
LD	8332	4359
RM	4034	1378
Irr	2169	1072
RMIrr	1729	631.2
GiIrr	**479.4**	170.8

$\phi(\mu, \sigma) = \mu^2 + \sigma^2$

Independent component analysis [Amari 1996, Welling & Teh 2011]

$$\pi(\mathbf{W}|\mathbf{X}) = \det \mathbf{W} \prod_{i=1}^{m} p(\mathbf{w}_i^\top \mathbf{x}) \prod_{ij} \mathcal{N}(\mathbf{W}_{ij}; 0, \lambda^{-1})$$

- $p(y) = \frac{1}{4}\text{sech}^2(\frac{1}{2}y)$
- After vectorization, reversible perturbation that is also positive is

$$\mathbf{B}(\mathbf{W}) = \mathbf{W}^\top \mathbf{W} \otimes \mathbf{I}_d + I_{d^2}$$

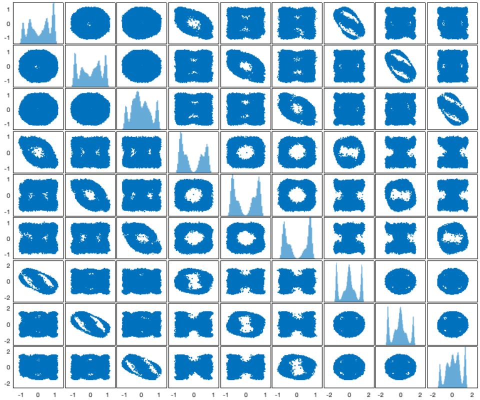
- Geometry-informed irreversible perturbation is

$$\gamma(\mathbf{W}) = \frac{1}{2}\left[\mathbf{D}\mathbf{B}(\mathbf{W}) + \mathbf{B}(\mathbf{W})\mathbf{D}\right]$$

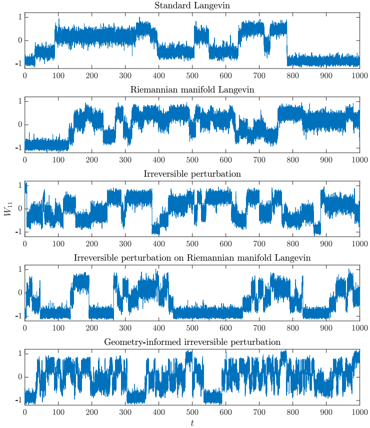
where $\mathbf{D}$ is any $m^2 \times m^2$ skew-symmetric matrix.

- In our experiments $\mathbf{D} = (\mathbf{I} \otimes \mathbf{C}_0 + \mathbf{C}_0 \otimes \mathbf{I})$, $\mathbf{C}_0 = -\mathbf{C}_0^\top$

Posterior distribution for an independent component analysis problem



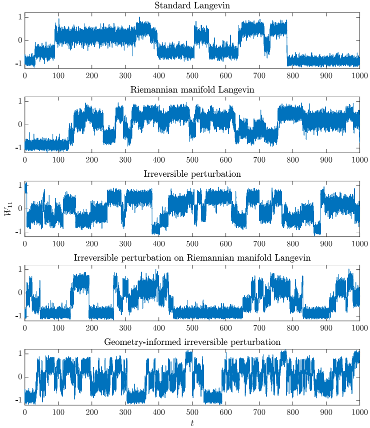
Geometry-informed irreversible perturbation mixes better



	$\mathbb{E}[\mathrm{AVar}_\phi]$	$\mathrm{Std}[\mathrm{AVar}_\phi]$
LD	50.17	17.92
RM	26.75	8.442
Irr	27.02	9.134
RMIrr	19.47	6.086
GiIrr	**6.381**	1.777

$$\phi(\mathbf{W}) = \left(\sum_{ij} \mathbf{W}_{ij}\right)^2$$

Geometry-informed irreversible perturbation mixes better

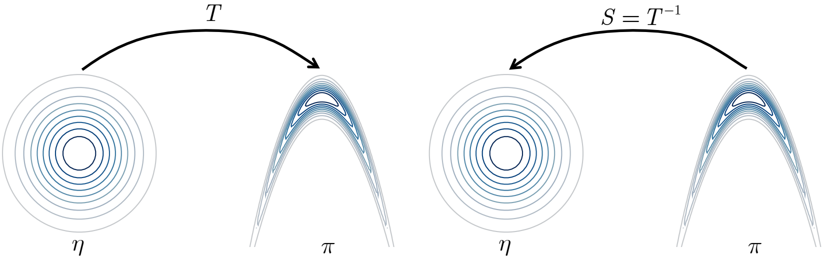


	$\mathbb{E}[\text{AVar}_\phi]$	$\text{Std}[\text{AVar}_\phi]$
LD	50.17	17.92
RM	26.75	8.442
Irr	27.02	9.134
RMIrr	19.47	6.086
GiIrr	**6.381**	1.777

$$\phi(\mathbf{W}) = \left(\sum_{ij} \mathbf{W}_{ij}\right)^2$$

- Open question: Are there guarantees for discretizations of perturbed LD?
- What about the reversible perturbation?
- **New/different perspective based on measure transport**

Transport maps are functional representations of random variables

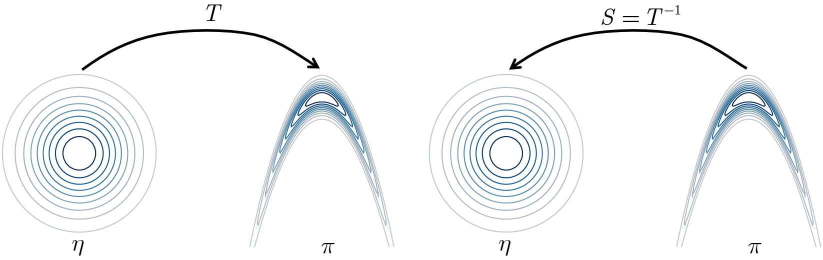


Transport maps

- Choose $X \sim \eta$ (e.g., standard Gaussian)
- Seek a *deterministic*, invertible map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\pi(y) = T_\sharp \eta(y) = \eta(S(y)) \det \mathbf{J}_S(y),\ \mathbf{J}_S(y) \text{ is the Jacobian of } S.$$

Transport maps are functional representations of random variables



Transport maps

- Choose $X \sim \eta$ (e.g., standard Gaussian)
- Seek a *deterministic*, invertible map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that

  $\pi(y) = T_\sharp \eta(y) = \eta(S(y)) \det \mathbf{J}_S(y)$, $\mathbf{J}_S(y)$ is the Jacobian of S.
  If $X \sim \eta$, then $Y = T(X) \sim \pi$.
- Many ways to find T: optimal transport, triangular transport, etc.

Transport maps define reversible and irreversible perturbations

- TM: Let $X \sim \eta$, $Y \sim \pi$. Transport maps $T_\sharp \eta = \pi$, $S_\sharp \pi = \eta$.

Proposition: `TM + LD = RMLD`

- LD on η: $dX_t = \nabla \log \eta(X_t)dt + \sqrt{2}dW_t$
- $Y_t = T(X_t)$ is an `RMLD` with $\mathbf{B}(Y_t) = (\mathbf{J}_S(Y_t)^* \mathbf{J}_S(Y_t))^{-1}$

$$dY_t = [\mathbf{B}(Y_t)\nabla \log \pi(Y_t) + \nabla \cdot \mathbf{B}(Y_t)]\,dt + \sqrt{2\mathbf{B}(Y_t)}dW_t$$

Transport maps define reversible and irreversible perturbations

Proposition: `TM + Irr = GiIrr`

- Irreversible LD on η with $\mathbf{D} = -\mathbf{D}^\top$:
  $dX_t = (\mathbf{I} + \mathbf{D})\nabla \log \eta(X_t)dt + \sqrt{2}dW_t$
- $Y_t = T(X_t)$ is a `GiIrr` with $\mathbf{B}(Y_t) = (\mathbf{J}_S(Y_t)^* \mathbf{J}_S(Y_t))^{-1}$,
  $\mathbf{C}(Y_t) = \mathbf{J}_S(Y_t)^{-1}\mathbf{D}\mathbf{J}_S^*(Y_t)^{-1}$

$$dY_t = [(\mathbf{B}(Y_t) + \mathbf{C}(Y_t))\nabla \log \pi(Y_t) + \nabla \cdot (\mathbf{B}(Y_t) + \mathbf{C}(Y_t))]\, dt + \sqrt{2\mathbf{B}(Y_t)}dW$$

Transport maps define reversible perturbations

Proposition: `TM + LD = RMLD`

- Langevin dynamics on $\eta = S_\sharp \pi$: $\mathrm{d}X_t = \nabla \log \eta(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$
- $Y_t = T(X_t)$ is an RMLD on π with $\mathbf{B}(Y_t) = (\mathbf{J}_S(Y_t)^* \mathbf{J}_S(Y_t))^{-1}$

$$\mathrm{d}Y_t = [\mathbf{B}(Y_t)\nabla \log \pi(Y_t) + \nabla \cdot \mathbf{B}(Y_t)]\,\mathrm{d}t + \sqrt{2\mathbf{B}(Y_t)}\mathrm{d}W_t$$

Insights and implications

- Transport maps *parameterize* reversible perturbations (or metrics)
- **Transport maps provide new way for discretizing RMLD**

Construction of triangular transports

$$S(y_1,\ldots,y_d)=\begin{bmatrix}S_1(y_1)\\ S_2(y_1,y_2)\\ \vdots\\ S_d(y_1,\ldots,y_d)\end{bmatrix} \implies \mathbf{J}_S(y)=\begin{bmatrix}\partial_{y_1}S_1 & & \\ \partial_{y_1}S_2 & \partial_{y_2}S_2 & \\ \vdots & \vdots & \ddots \\ \partial_{y_1}S_d & \cdots & \partial_{y_d}S_d\end{bmatrix}$$

$$\min_S D_{KL}(S_\sharp\pi\|\mathcal{N}(0,\mathbf{I}_d)) \implies \max_S \mathbb{E}_\pi\left[\log S^\sharp\mathcal{N}(0,\mathbf{I_d})\right]$$

- Each component map S_i is parametrized to be **monotone** in the leading variable

Construction of triangular transports

$$S(y_1,\ldots,y_d) = \begin{bmatrix} S_1(y_1) \\ S_2(y_1,y_2) \\ \vdots \\ S_d(y_1,\ldots,y_d) \end{bmatrix} \implies \mathbf{J}_S(y) = \begin{bmatrix} \partial_{y_1} S_1 & & \\ \partial_{y_1} S_2 & \partial_{y_2} S_2 & \\ \vdots & \vdots & \ddots \\ \partial_{y_1} S_d & \cdots & \partial_{y_d} S_d \end{bmatrix}$$

$$\min_S D_{KL}(S_\sharp \pi \| \mathcal{N}(0,\mathbf{I}_d)) \implies \max_S \mathbb{E}_\pi \left[\log S^\sharp \mathcal{N}(0,\mathbf{I_d})\right]$$

- Each component map S_i is parametrized to be **monotone** in the leading variable
- Monotone, **triangular** structure $\implies$ fast computation of S^{-1} and $\det \mathbf{J}_S$

Construction of triangular transports

$$S(y_1,\dots,y_d) = \begin{bmatrix} S_1(y_1) \\ S_2(y_1,y_2) \\ \vdots \\ S_d(y_1,\dots,y_d) \end{bmatrix} \implies \mathbf{J}_S(y) = \begin{bmatrix} \partial_{y_1} S_1 & & \\ \partial_{y_1} S_2 & \partial_{y_2} S_2 & \\ \vdots & \vdots & \ddots \\ \partial_{y_1} S_d & \cdots & \partial_{y_d} S_d \end{bmatrix}$$

$$\min_S D_{KL}(S_\sharp \pi \| \mathcal{N}(0,\mathbf{I}_d)) \implies \max_S \mathbb{E}_\pi \left[\log S^\sharp \mathcal{N}(0,\mathbf{I_d})\right]$$

- Each component map S_i is parametrized to be **monotone** in the leading variable
- Monotone, **triangular** structure $\implies$ fast computation of S^{-1} and $\det \mathbf{J}_S$
- Approximate map produces samples approximating π, but has **bias**

Construction of triangular transports

$$S(y_1,\dots,y_d) = \begin{bmatrix} S_1(y_1) \\ S_2(y_1,y_2) \\ \vdots \\ S_d(y_1,\dots,y_d) \end{bmatrix} \implies \mathbf{J}_S(y) = \begin{bmatrix} \partial_{y_1} S_1 & & \\ \partial_{y_1} S_2 & \partial_{y_2} S_2 & \\ \vdots & \vdots & \ddots \\ \partial_{y_1} S_d & \cdots & \partial_{y_d} S_d \end{bmatrix}$$

$$\min_S D_{KL}(S_\sharp \pi \| \mathcal{N}(0,\mathbf{I}_d)) \implies \max_S \mathbb{E}_\pi \left[\log S^\sharp \mathcal{N}(0,\mathbf{I_d})\right]$$

- Each component map S_i is parametrized to be **monotone** in the leading variable
- Monotone, **triangular** structure $\implies$ fast computation of S^{-1} and $\det \mathbf{J}_S$
- Approximate map produces samples approximating π, but has **bias**
- With a few samples from π, learn a *monotone triangular* map via `ATM`
  - On the representation and learning of monotone triangular transport maps [Baptista et al. 2020]

Transport map unadjusted Langevin algorithm

- Given target $\pi(y)$ and a triangular map $S(y) = T^{-1}(y)$
- *Define* reference $\eta(x) = (S_\sharp \pi)(x) = \pi(T(x)) \det \mathbf{J}_T(x)$, $\nabla \log \eta(x) = \nabla \log \pi(T(x)) + \nabla \log \det \mathbf{J}_T(x)$
- Construct Langevin dynamics on $\eta = S_\sharp \pi$, apply map $T = S^{-1}$ to trajectories on η

Transport map unadjusted Langevin algorithm

- Given target $\pi(y)$ and a triangular map $S(y) = T^{-1}(y)$
- *Define* reference $\eta(x) = (S_\sharp \pi)(x) = \pi(T(x)) \det \mathbf{J}_T(x)$,
  $\nabla \log \eta(x) = \nabla \log \pi(T(x)) + \nabla \log \det \mathbf{J}_T(x)$
- Construct Langevin dynamics on $\eta = S_\sharp \pi$, apply map $T = S^{-1}$ to trajectories on η

Transport map unadjusted Langevin algorithm (TMULA)

$$X_{k+1} = X_k + h \underbrace{\mathbf{J}_S^*(Y_k)^{-1} \left[\nabla \log \pi(Y_k) + \sum_{i=1}^{d} \left(\frac{\partial S_i}{\partial y_i}(Y_k) \right)^{-1} H_i(Y_k) \right]}_{\nabla \log \eta(X_k)} + \sqrt{2h}\xi_{k+1}$$

$$Y_{k+1} = T(X_{k+1})$$

where $H_i(Y_k) = \left[\frac{\partial^2 S_i}{\partial y_1 \partial y_i} \cdots \frac{\partial^2 S_i}{\partial y_d \partial y_i} \right]^\top$.

Other instances of transformed Langevin processes

- **Mirror Langevin** for sampling constrained distributions
  - Mirrored Langevin dynamics [Hsieh et al. 2018]
  - Wasserstein control of Mirror Langevin Monte Carlo [K Zhang et al. 2020]
  - Defines map as ∇h, where h is convex. Inverse is convex conjugate

Other instances of transformed Langevin processes

- **Mirror Langevin** for sampling constrained distributions
  - Mirrored Langevin dynamics [Hsieh et al. 2018]
  - Wasserstein control of Mirror Langevin Monte Carlo [K Zhang et al. 2020]
  - Defines map as ∇h, where h is convex. Inverse is convex conjugate
- **Transport map accelerated MCMC** (including TM-MALA) [Parno & Marzouk 2018]
  - Constructs triangular invertible transport for MCMC proposals
- **Adaptive Monte Carlo augmented with normalizing flows** [Gabrié et al. 2022]
  - Similar to TM-MCMC, where the maps are normalizing flows

Other instances of transformed Langevin processes

- **Mirror Langevin** for sampling constrained distributions
  - Mirrored Langevin dynamics [Hsieh et al. 2018]
  - Wasserstein control of Mirror Langevin Monte Carlo [K Zhang et al. 2020]
  - Defines map as ∇h, where h is convex. Inverse is convex conjugate
- **Transport map accelerated MCMC** (including TM-MALA) [Parno & Marzouk 2018]
  - Constructs triangular invertible transport for MCMC proposals
- **Adaptive Monte Carlo augmented with normalizing flows** [Gabrié et al. 2022]
  - Similar to TM-MCMC, where the maps are normalizing flows
- **Variable transformation to obtain geometric ergodicity** [Johnson & Geyer 2012]
  - Provides generic functions to transform tails of (sub)-exponentially light distributions for MCMC
- **Heavy-tailed sampling via transformed ULA** [He et al. 2022]
  - Provides generic functions to transform heavy-tailed distributions and applies ULA

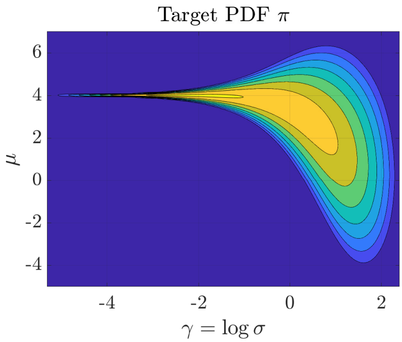
Numerical example: Funnel distribution

Bayesian inference problem

- Given data $\{X_i\}_{i=1}^N \sim \mathcal{N}(\mu, \sigma^2)$
- Infer $\mu, \gamma = \log \sigma \in \mathbb{R}$.
- Prior $\mu \sim \mathcal{N}(0, 3)$, $\gamma \sim \Gamma(2, 1)$

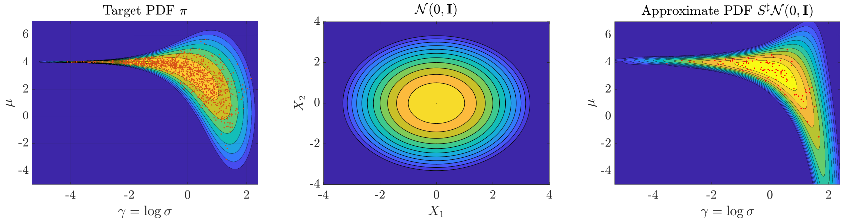
$\pi(\mu, \gamma | \mathbf{X}) \propto$

$$\exp\left((2-N)\gamma + 2\gamma - e^{\gamma} - \frac{\mu^2}{6} - \frac{1}{2}\sum_{i=1}^{N}(X_i - \mu)^2\right)$$

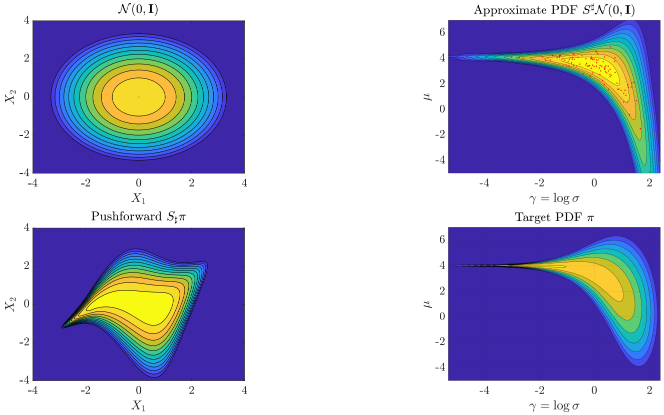


Numerical example: Funnel distribution

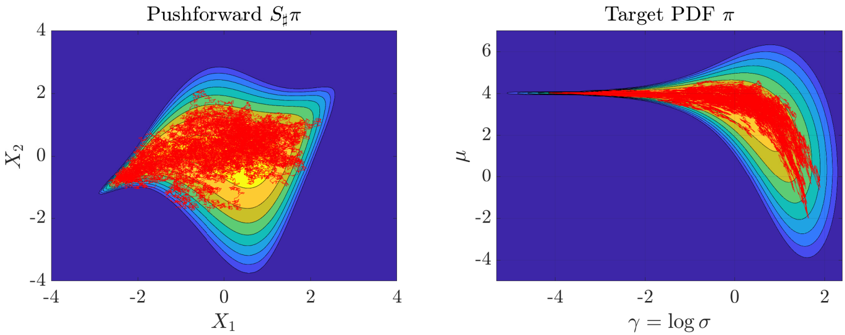
Learn a very approximate map via `ATM`



Numerical example: Funnel distribution

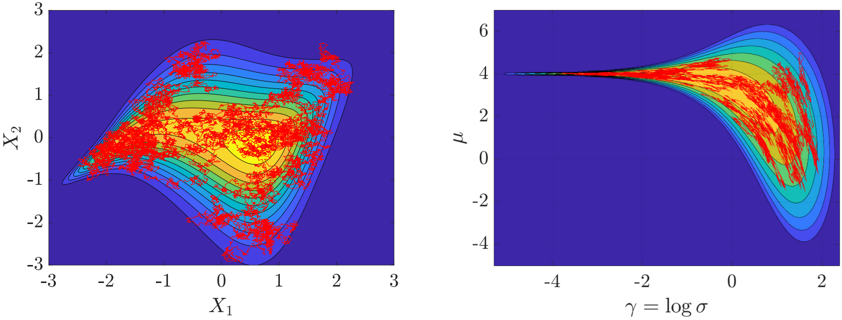


Numerical example: Funnel distribution



$$X_{k+1} = X_k + h\mathbf{J}_S^*(Y_k)^{-1}\left[\nabla \log \pi(Y_k) + \sum_{i=1}^d \left(\frac{\partial S_i}{\partial y_i}(Y_k)\right)^{-1} H_i(Y_k)\right] + \sqrt{2h}\xi_{k+1}$$
$$Y_{k+1} = T(X_{k+1})$$

Numerical example: `TM + Irr = GiIrr`



Numerical example: Funnel distribution

RMLD: $\mathbf{B}(\mu,\gamma) = \begin{bmatrix} \frac{1}{2Ne^{\gamma}} & 0 \\ 0 & \frac{1}{Ne^{-2\gamma}+1/3} \end{bmatrix}$ expected Fisher information plus negative Hessian of log prior.

TMRMLD: $\mathbf{B}(\mu,\gamma)^{-1} = \mathbf{J}_S^{\top}\mathbf{J}_S(\mu,\gamma)$.

Consider test functions $\phi_1(\mu,\gamma) = \exp(\gamma)$, $\phi_2(\mu,\gamma) = \gamma + \mu$, and $\phi_3(\mu,\gamma) = \gamma^2 + \mu^2$.

	$\mathbb{E}[\text{AVar}_{\phi_1}]$	$\text{Std}[\text{AVar}_{\phi_1}]$	$\mathbb{E}[\text{AVar}_{\phi_2}]$	$\text{Std}[\text{AVar}_{\phi_2}]$	$\mathbb{E}[\text{AVar}_{\phi_3}]$	$\text{Std}[\text{AVar}_{\phi_3}]$
`ULA`	8.759	1.797	1.957	0.4774	195.4	35.30
`RMLD`	25.46	7.550	28.82	2.860	1558	184.8
`TMRMLD`	1.344	0.2057	2.655	0.3705	108.7	15.48
`TMULA`	1.444	0.2061	2.480	0.3475	114.8	14.00
`TMULA + Irr`	**1.243**	**0.2131**	**1.961**	**0.2851**	**92.72**	**12.89**

Table 1: Asymptotic variance estimates for the funnel distribution.

Convergence guarantees

Proposition: guarantees revisited

Let $U(y) = -\log \eta = -\log S_\sharp \pi \in \mathcal{C}^2(\mathbb{R}^d)$, $Y_k \sim \pi^k$, $X_k \sim \eta^k$. If

- $m\mathbf{I} \preceq \nabla^2 U \preceq L\mathbf{I}$ (strong convexity and Lipschitz gradients)
- S is appropriately monotone $\|S(y) - S(y')\| \geq \rho \|y - y'\|$

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{Cr^k}{\rho^2} \mathcal{W}_2^2(\eta^0, \eta) + F(m, L, h)$$

where $r = 1 - \frac{mL}{(m+L)^2}$, $F(m, L, h)$ is the bias

- Does such a map S exist?

Convergence guarantees

Proposition: guarantees revisited

Let $U(y) = -\log \eta = -\log S_\sharp \pi \in \mathcal{C}^2(\mathbb{R}^d)$, $Y_k \sim \pi^k$, $X_k \sim \eta^k$. If

- $m\mathbf{I} \preceq \nabla^2 U \preceq L\mathbf{I}$ (strong convexity and Lipschitz gradients)
- S is appropriately monotone $\|S(y) - S(y')\| \geq \rho \|y - y'\|$

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{Cr^k}{\rho^2} \mathcal{W}_2^2(\eta^0, \eta) + F(m, L, h)$$

where $r = 1 - \frac{mL}{(m+L)^2}$, $F(m, L, h)$ is the bias

- Does such a map S exist? Yes: there exists (many) maps such that η is isotropic normal!

Convergence guarantees

Proposition: guarantees revisited

Let $U(y) = -\log \eta = -\log S_\sharp \pi \in \mathcal{C}^2(\mathbb{R}^d)$, $Y_k \sim \pi^k$, $X_k \sim \eta^k$. If

- $m\mathbf{I} \preceq \nabla^2 U \preceq L\mathbf{I}$ (strong convexity and Lipschitz gradients)
- S is appropriately monotone $\|S(y) - S(y')\| \geq \rho \|y - y'\|$

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{Cr^k}{\rho^2} \mathcal{W}_2^2(\eta^0, \eta) + F(m, L, h)$$

where $r = 1 - \frac{mL}{(m+L)^2}$, $F(m, L, h)$ is the bias

- Does such a map S exist? Yes: there exists (many) maps such that η is isotropic normal!
- Can the rate be optimized?

Convergence guarantees

Proposition: guarantees revisited

Let $U(y) = -\log \eta = -\log S_\sharp \pi \in \mathcal{C}^2(\mathbb{R}^d)$, $Y_k \sim \pi^k$, $X_k \sim \eta^k$. If

- $m\mathbf{I} \preceq \nabla^2 U \preceq L\mathbf{I}$ (strong convexity and Lipschitz gradients)
- S is appropriately monotone $\|S(y) - S(y')\| \geq \rho \|y - y'\|$

$$\mathcal{W}_2^2(\pi^k, \pi) \leq \frac{Cr^k}{\rho^2} \mathcal{W}_2^2(\eta^0, \eta) + F(m, L, h)$$

where $r = 1 - \frac{mL}{(m+L)^2}$, $F(m, L, h)$ is the bias

- Does such a map S exist? Yes: there exists (many) maps such that η is isotropic normal!
- Can the rate be optimized? Yes: optimal when $m = L \iff \eta$ is isotropic normal!

Transport map ULA is a different discretization of RMLD

RMLD

$$dY_t = [\mathbf{B}(Y_t)\nabla \log \pi(Y_t) + \nabla \cdot \mathbf{B}(Y_t)]\, dt + \sqrt{2\mathbf{B}(Y_t)} dW_t$$

$$\text{with } \mathbf{B}(Y_t) = (\mathbf{J}_S(Y_t)^* \mathbf{J}_S(Y_t))^{-1}$$

TMULA:

$$Y_{k+1} = T\left(S(Y_k) + h\mathbf{J}_S^*(Y_k)^{-1}\left[\nabla \log \pi(Y_k) + \sum_{i=1}^{d}\left(\frac{\partial S_i}{\partial y_i}(Y_k)\right)^{-1} H_i(Y_k)\right] + \sqrt{2h}\xi_{k+1}\right)$$

Euler-Maruyama applied to RMLD:

$$Y_{k+1} = Y_k + h(\mathbf{J}_S(Y_k)^* \mathbf{J}_S(Y_k))^{-1}\nabla \log \pi(Y_k) + \nabla \cdot (\mathbf{J}_S(Y_k)^* \mathbf{J}_S(Y_k))^{-1}) + \sqrt{2h}\xi_{k+1}$$

Transport map ULA is a different discretization of RMLD

TMULA:

$$Y_{k+1} = T\left(S(Y_k) + h\mathbf{J}_S^*(Y_k)^{-1}\left[\nabla \log \pi(Y_k) + \sum_{i=1}^{d}\left(\frac{\partial S_i}{\partial y_i}(Y_k)\right)^{-1} H_i(Y_k)\right] + \sqrt{2h}\xi_{k+1}\right)$$

EMRMLD:

$$Y_{k+1} = Y_k + h(\mathbf{J}_S(Y_k)^*\mathbf{J}_S(Y_k))^{-1}\nabla \log \pi(Y_k) + \nabla \cdot (\mathbf{J}_S(Y_k)^*\mathbf{J}_S(Y_k))^{-1}) + \sqrt{2h}\xi_{k+1}$$

Proposition: EMRMLD approximates TMULA

Let TMULA_m denote the mth component, $m = 1, \ldots d$. Then $\text{TMULA}_m = \text{EMRMLD}_m + h\left(\xi_{m+1}\nabla^2 T_m \xi_{m+1} - \sum_{i=1}^{d} \frac{\partial^2 T_m}{\partial x_i^2}\right) + \mathcal{O}(h^{3/2})$

Proposition: Regularity of T affects variance of error

$$\text{Var}\left(h\xi_{m+1}\nabla^2 T_m \xi_{m+1} - h\sum_{i=1}^{d}\frac{\partial^2 T_m}{\partial x_i^2}\right) = h^2\left(\sum_{ij}\frac{\partial^2 T_m}{\partial x_i \partial x_j} + 3\sum_{i=1}^{d}\left(\frac{\partial^2 T_m}{\partial x_i^2}\right)^2\right)$$

Convergence to the numerical invariant measure

Asymptotic bias of the ergodic estimator

$$e(\phi, h) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=0}^{K-1} \phi(Y_n) - \int \phi(Y)\pi(y)\mathrm{d}y$$

Rate of convergence of e [Abduelle et al. 2013]

With an integrator of local weak order p,

$$e(\phi, h) = -\lambda_p h^p + \mathcal{O}(h^{p+1})$$

with $\lambda_p = \int_0^\infty \int_{\mathbb{R}^d} \left(\frac{1}{(p+1)!}\mathcal{L}^{p+1} - A_p \right) u(y,t)\pi(y)\mathrm{d}y\mathrm{d}t.$

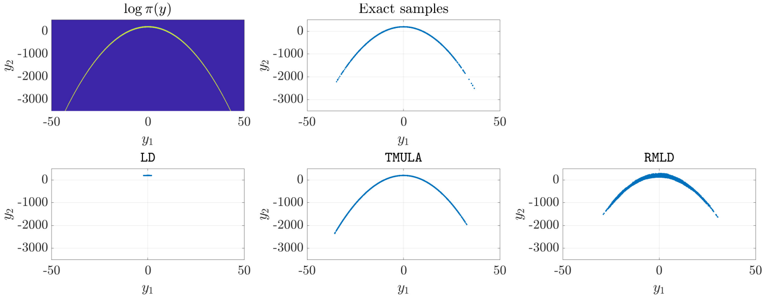
In very restrictive settings, can compute that λ_p for TMULA is smaller than for EMRMLD.

Numerical example: banana example

$$\log \pi(y) = -y_1^2/s^2 - (y_2 + by_1^2 - 100b)^2, \text{ with } s = 4, b = 0.01$$

$$S(y_1, y_2) = \begin{bmatrix} y_1/s \\ y_2 + by_1^2 - 100b \end{bmatrix} : \text{ this is pushing it to a Gaussian density}$$

$$\phi(y_1, y_2) = y_1^2 + y_1 + y_2^2 + y_2$$



Numerical example: banana example

Compute now the leading eigenvalue for the asymptotic bias:
$\lambda_1 \sim -e(\phi, h)/h$

$$\log \pi(y) = -y_1^2/s^2 - (y_2 + by_1^2 - 100b)^2, \text{ with } s = 4, b = 0.01$$
$$\phi(y_1, y_2) = y_1^2 + y_1 + y_2^2 + y_2$$

- We can calculate $\lambda_1^{\mathrm{TMULA}} = -0.62$ while $\lambda_1^{\mathrm{EMRMLD}} = 34.69$.
- Transport map accelerates convergence because it is a reversible perturbation.

Numerical example: Hybrid Rosenbrock [Pagani et al. 2022]

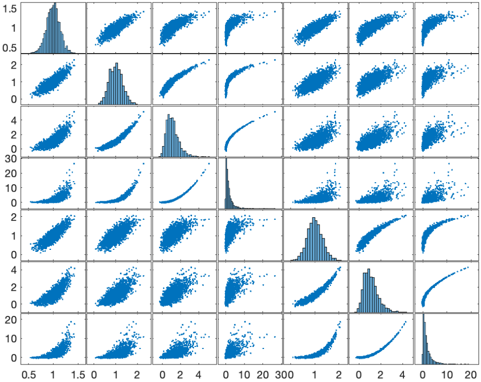


Figure 1: $\pi(\mathbf{y}) \propto \exp\left\{-a(y_1-\mu)^2 - \sum_{j=1}^{n_2}\sum_{i=2}^{n_1} b_{ji}(y_{j,i}-y_{j,i-1}^2)^2\right\}$

Numerical example: Hybrid Rosenbrock [Pagani et al. 2022]

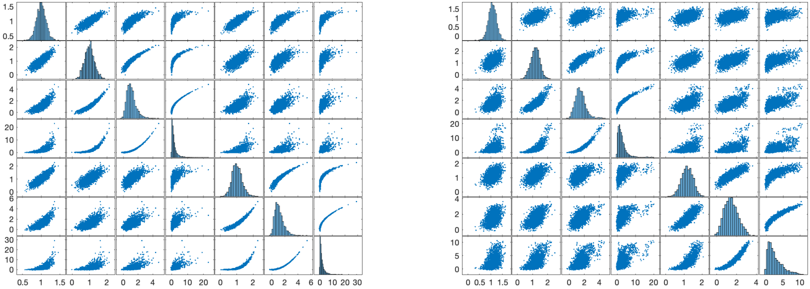


Figure 2: Left: TMULA, Right: ULA. Step size $h = 0.01$

Numerical example: Hybrid Rosenbrock [Pagani et al. 2022]

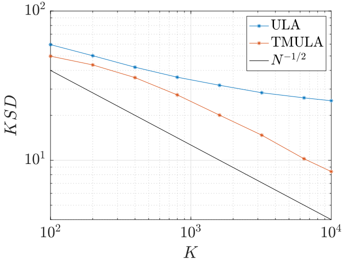
Measuring sample quality via kernelized Stein discrepancy [Gorham & Mackey 2017]

- Approximates integral probability metrics

$$d_{\mathcal{H}}(\hat{\pi}_K, \pi) = \sup_{\phi \in \mathcal{H}} |\mathbb{E}_{\hat{\pi}_K}[\phi(Z)] - \mathbb{E}_{\pi}[\phi(X)]|$$

- Only requires evaluations of $\nabla \log \pi(x)$ and a kernel function
- For $\mathcal{H}$ is large enough

$$d_{\mathcal{H}}(\hat{\pi}_K, \pi) \to 0 \iff \hat{\pi}_K \to \pi \text{ in distribution}$$



- Observation: TMULA allows taking larger step size. Here $h = 0.01$.

Numerical example: Hybrid Rosenbrock [Pagani et al. 2022]

Consider test functions $\phi_1(Y) = \sum_{i=1}^{7} Y^i$ and $\phi_s(Y) = \sum_{i=1}^{7} (Y^i)^2$.

	$\mathbb{E}[\mathrm{AVar}_{\phi_1}]$	$\mathrm{Std}[\mathrm{AVar}_{\phi_1}]$	$\mathbb{E}[\mathrm{AVar}_{\phi_2}]$	$\mathrm{Std}[\mathrm{AVar}_{\phi_2}]$
UILA	6762	2663	6.957×10^6	5.185×10^6
TMUILA	**65.03**	**28.54**	**6506**	**1284**

Table 2: Asymptotic variance estimates for the hybrid Rosenbrock distribution.

Some caveats

Discretizations of irreversibly perturbed systems

- Irreversible term increases stiffness
- May lead to *worse* performance due to extra bias

Lighter than Gaussian tails

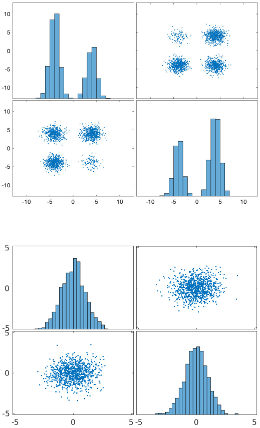
- Euler-Maruyama discretization may be transient (e.g., EM on $dX_t = -X_t^3 dt + \sqrt{2} dW_t$)!
- TMULA may blow up because it explores tails better (chain diverges there)

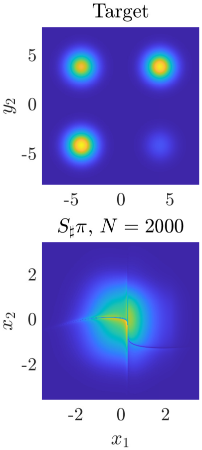
Implicit Euler-Maruyama schemes may be used

$$\begin{cases} S(Y^*) = S(Y^k) + h\mathbf{J}_S^\top(Y^*)^{-1}\left[\nabla_Y \log \pi(Y^*) - \sum_{i=1}^{d}\left(\frac{\partial S_i}{\partial y_i}(Y^*)\right)^{-1} H_i(Y^*)\right] \\ X_{k+1} = S(Y^*) + \sqrt{2h}\xi^{k+1} \\ Y_{k+1} = T(X_{k+1}), \end{cases} \tag{1}$$

where $H_i(Y^k) = \left[\frac{\partial^2 S_i}{\partial y_1 \partial y_i}, \cdots, \frac{\partial^2 S_i}{\partial y_d \partial y_i}\right]^\top$, where $\xi^{k+1} \sim \mathcal{N}(0, \mathbf{I})$.

Cautionary tale: Multimodal distributions





Conclusion

Algorithmic aspects

- **Improved sample quality** and **accelerated convergence** of LD
- Novel *geometry-informed irreversible* perturbations
- We considered *triangular transport*, but TMULA is agnostic

Theoretical aspects

- Transport map ULA can guarantee fast convergence for a larger class of distributions
- Transport map applied to Langevin dynamics is Riemannian manifold Langevin dynamics

Future directions

- Interacting particle systems formulation for learning maps
- Analyzing the TM-MALA (with Metropolis-Hastings correction)
- improve our theoretical understanding of how to characterize the transport map within a given approximate class that maximizes the efficiency of TMULA sampling
- Better approximation of transport maps in the presence of multimodality

References. I

1. K.A. Athreya, H. Doss and J. Sethuraman, On the convergence of the Markov chain simulation method, *Annals of Statistics*, Vol. 24, (1996), pp. 69-100.
2. C. Barbarosie, Representation of divergence-free vector fields, *Quarterly of Applied Mathematics*, Vol. 69, (2011), pp. 309–316.
3. M. Bedard and J.S. Rosenthal, Optimal Scaling of Metropolis Algorithms: Heading Towards General Target Distributions, *Canadian Jounral of Statistics*, Vol. 36, Issue 4, (2008), pp. 483-503.
4. P. Constantin, A. Kiselev, L. Ryshik and A. Zlatos, Diffusion and mixing in fluid flow *Annals of Mathematics*, Vo. 168 (2008), pp. 643-674

References. II

5. M.D. Donsker and S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large times, I, *Communications Pure in Applied Mathematics*, Vol. 28, (1975), pp. 1-47, II, *Communications on Pure in Applied Mathematics*, Vol. 28, (1975), pp. 279–301, and III, *Communications on Pure in Applied Mathematics*, Vol. 29, (1976), pp. 389-461.
6. P. Dupuis, Y. Liu, N. Plattner, and J. D. Doll, On the Infinite Swapping Limit for Parallel Tempering. *SIAM Multiscale Modeling and Simulation*, Vol. 10, Issue 3, (2012), pp. 986-1022.
7. M.I. Freidlin and A.D. Wentzell, Random Perturbations of Hamiltonian Systems, *Memoirs of the American Mathematical Society*, Vol. 109, No. 523, 1994.
8. B. Franke, C.-R. Hwang, H.-M. Pai, and S.-J. Sheu, The behavior of the spectral gap under growing drift, *Transactions of the American Mathematical Society*, Vol 362, No. 3 (2010), pp. 1325-1350.

References. III

9. A. Frigessi, C.R. Hwang and L. Younes, Optimal spectral structures of reversible stochatic matrices, Monte Carlo methods and the simulation of Markov random fields, *Annals of Applied Probability*, Vol. 2, (1992), pp. 610-628.
10. A. Frigessi, C.R. Hwang, S.J. Sheu and P. Di Stefano, Convergence rates of the Gibbs sampler, the Metropolis algorithm, and their single-site updating dynamics, *Journal of Royal Statistical Society Series B, Statistical Methodology*, Vol. 55, (1993), pp. 205-219.
11. A. Ichiki and M. Ohzeki, Violation of detailed balance condition accelerates relaxation, *Physical Review E*, Vol. 88, (2013), pp. 020101(R).
12. Jürgen Gärtner, On large deviations from the invariant measure, *Theory of probability and its applications*, Vol. XXII, No. 1, (1977), pp. 24-39.

References. IV

13. W.R. Gilks and G.O. Roberts, Strategies for improving MCMC, *Monte Carlo Markov Chain in practice*, Chapman and Hall, Boca Raton, FL, (1996), pp. 89-114.
14. C.R. Hwang, S.Y. Hwang-Ma and S.J. Sheu, Accelerating Gaussian diffusions. *The Annals of Applied Probability* Vol. 3, (1993) pp. 897-913.
15. C.R. Hwang, S.Y. Hwang-Ma and S.J. Sheu, Accelerating diffusions, *The Annals of Applied Probability*, Vol 15, No. 2, (2005), pp. 1433-1444.
16. T. Lelievre, F. Nier and G.A. Pavliotis, Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion, *Journal of Statistical Physics*, 152(2), 237-274, (2013)
17. K.L. Mergessen and R.L. Tweedie, Rates of convergence of the Hastings and Metropolis algorithms, *Annals of Statistics*, Vol. 24, (1996), pp. 101-121.

References. V

18. R. Pinsky, The I-function for diffusion processes with boundaries, *The Annals of Probability*, Vol. 13, No. 3, (1985), pp. 676-692.
19. G.O. Roberts and J.S. Rosenthal, General state space Markov Chain and MCMC algorithms, *Probability Surveys*, Vol. 1, (2004), pp. 20-71.
20. Luc Rey-Bellet and K. Spiliopoulos, Irreversible Langevin samplers and variance reduction: a large deviations approach, Nonlinearity , Vol. 28, (2015), pp. 2081-2103.
21. Luc Rey-Bellet and K. Spiliopoulos, Variance reduction for irreversible Langevin samplers and diffusion on graphs, Electronic Communications in Probability, Vol. 20, (2015), no. 15, pp. 1–16.
22. Luc Rey-Bellet and K. Spiliopoulos, Improving the convergence of reversible samplers, Journal of Statistical Physics, Vol. 164, Issue 3, (2016), pp. 472-494.

References. VI

23. Jianfeng Lu and K. Spiliopoulos, Analysis of multiscale integrators for multiple attractors and irreversible Langevin samplers, SIAM Multiscale Modeling and simulation , Vol. 16, Issue 4, (2018), pp. 1859–1883.
24. Michela Ottobre, Natesh S. Pillai and K. Spiliopoulos, Optimal scaling of the MALA algorithm with irreversible proposals for Gaussian targets, Stochastics and Partial Differential Equations: Analysis and Computations, Vol. 8, (2020), pp. 311–361.
25. Benjamin J. Zhang, Youssef M. Marzouk and K. Spiliopoulos, Geometry-informed irreversible perturbations for accelerated convergence of Langevin dynamics, Statistics and Computing, (2022), to appear.
26. Benjamin J. Zhang, Youssef M. Marzouk and K. Spiliopoulos, Transport map unadjusted Langevin algorithms, (2023), submitted.

Thank You!!!!!

Questions?