High-order Langevin Diffusion Yields an Accelerated MCMC Algorithm

Wenlong Mou, Yi-An Ma, Martin Wainwright, Peter Bartlett, Michael Jordan

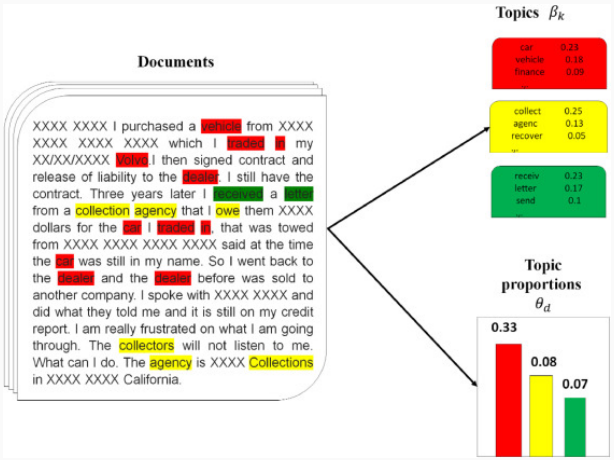
Preliminaries: from Langevin diffusion to high-order variants

Problem setup

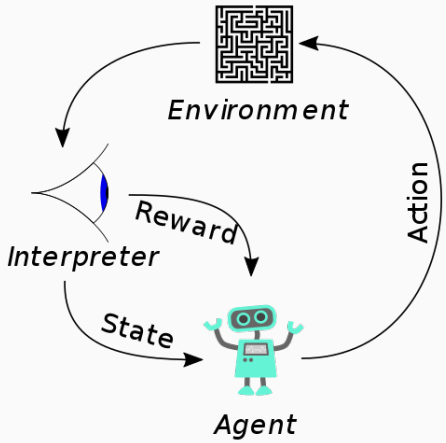
Goal: sample from target density $\pi \propto e^{-f}$, where f satisfies:

- β-smooth: $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2$.
- μ-strongly convex: $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu}{2} \|x - y\|_2^2$.

Example of applications:



Bayesian inference



Sequential decision making

Key algorithmic challenge: high dimensions.

Continuous-time Langevin diffusion

(Overdamped) Langevin diffusion:

$$d\theta_t = -\nabla f(\theta_t)dt + \sqrt{2}dB_t.$$

Unique stationary measure $\pi \propto e^{-f}$

Synchronous coupling:

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left[\left\|\theta_t^{(1)} - \theta_t^{(2)}\right\|_2^2\right] &= -\mathbb{E}\left[\langle \nabla f(\theta_t^{(1)}) - \theta_t^{(2)}, \theta_t^{(1)} - \theta_t^{(2)}\rangle\right] \\ &\leq -\frac{\mu}{2}\mathbb{E}\left[\left\|\theta_t^{(1)} - \theta_t^{(2)}\right\|_2^2\right].\end{aligned}$$

Forward Euler method for Langevin diffusion

(Overdamped) Langevin diffusion:

$$d\theta_t = -\nabla f(\theta_t)dt + \sqrt{2}dB_t.$$

Forward Euler method:

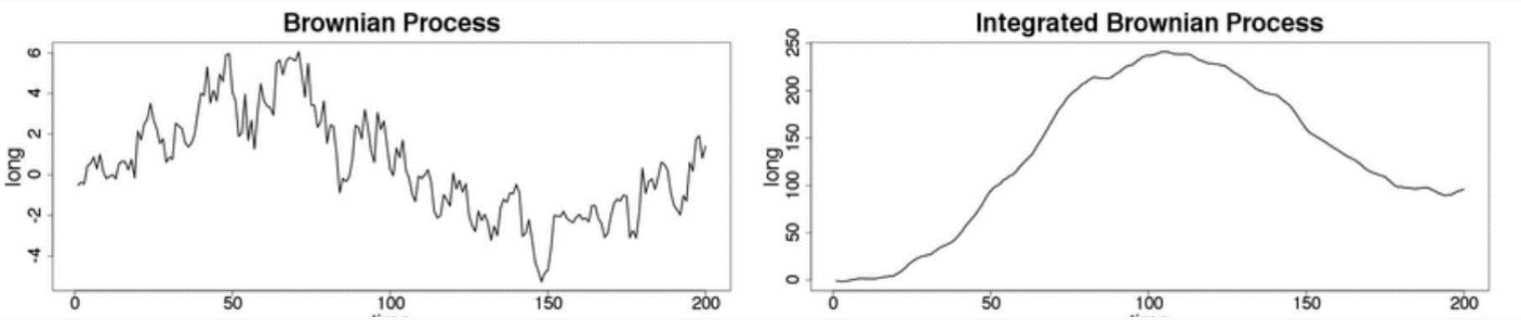
$$d\widehat{\theta}_t = -\nabla f\left(\widehat{\theta}_{k\eta}\right)dt + \sqrt{2}dB_t, \quad \text{for } t \in [k\eta, (k+1)\eta).$$

Discrete-time analysis:

- By fluctuation of Brownian motion: $\left\|\nabla f(\widehat{\theta}_t) - \nabla f(\widehat{\theta}_{k\eta})\right\|_2 \lesssim \sqrt{\eta d}$, leading to $O(\sqrt{\eta d})$ discretization error in total [Dal17, DM17].
- Improved analysis when $\nabla^2 f$ is also Lipschitz (in operator norm): $O(\eta d)$ discretization error [DM19, MFWB19].

Taking a smoother path: underdamped Langevin diffusion

Discretization error of ULA comes mainly from $O(\sqrt{\eta d})$ oscillation of Brownian motion.



Underdamped Langevin diffusion:

$$d\theta_t = r_t dt$$

$$dr_t = -\frac{1}{\beta}\nabla f(\theta_t) - \xi r_t dt + \sqrt{2\xi/\beta}\, dB_t.$$

Proof via synchronous coupling with Lyapunov function:

$$\Phi_t = \begin{bmatrix} \theta_t^{(1)} - \theta_t^{(2)} \\ r_t^{(1)} - r_t^{(2)} \end{bmatrix}^\top \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \theta_t^{(1)} - \theta_t^{(2)} \\ r_t^{(1)} - r_t^{(2)} \end{bmatrix}.$$

Discretization of underdamped Langevin diffusion

For the time interval $t \in [k\eta, (k+1)\eta)$:

$$d\widehat{\theta}_t = \widehat{r}_t dt$$
$$d\widehat{r}_t = -\frac{1}{\beta}\nabla f(\widehat{\theta}_{k\eta}) - \xi \widehat{r}_t dt + \sqrt{2\xi/\beta} dB_t.$$

Ornstein-Uhlenbeck process, implementable with explicit solution.

An intuitive analysis: note that $\|\widehat{r}_t\|_2 \sim O(\sqrt{d})$.

$$\left\|\widehat{\theta}_t - \widehat{\theta}_{k\eta}\right\|_2 \le O(\eta\sqrt{d}) \Rightarrow \left\|\widehat{r}_{(k+1)\eta} - r_{(k+1)\eta}\right\|_2 \le \|\widehat{r}_{k\eta} - r_{k\eta}\|_2 + O(\eta^2\sqrt{d})$$
$$\Rightarrow \|\widehat{r}_t - r_t\|_2 \le O(\eta\sqrt{d}) \Rightarrow \left\|\widehat{\theta}_t - \theta_t\right\|_2 \le O(\eta\sqrt{d}) \Rightarrow O(\sqrt{d}/\varepsilon) \text{ mixing time.}$$

Avoid fluctuation of BM to enter the gradient term.

Can we do even better?

Algorithm class	Mixing time bounds (in TV or Wasserstein)
Langevin diffusion	[Dal17, DM17]: $O(d/\varepsilon^2)$; [DM19, MFWB19]: $O(d/\varepsilon)$. No convexity needed.
Underdamped Langevin	[CCBJ18]: $O(\sqrt{d}/\varepsilon)$; [SL19]: $O(d^{1/3}/\varepsilon^{2/3})$ using randomized midpoint.
HMC	[MS17, MS19]: $O(\sqrt{d}/\varepsilon)$; [MV18, LSV18]: $\mathrm{polylog}(d, \varepsilon^{-1})$ under incoherence assumptions for generalized linear model.
High-order Langevin	This work: $O(d^{1/4}/\varepsilon^{1/2})$ under an integration oracle (e.g. for ridge-separable functions).

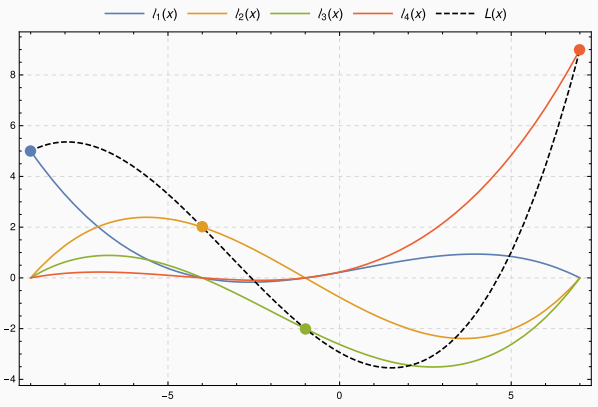
Difficulties with discretization and the integration oracle

Classical discretization vs. high-dimensional discretization

Consider computing an integral $\mathbb{R}^d$:

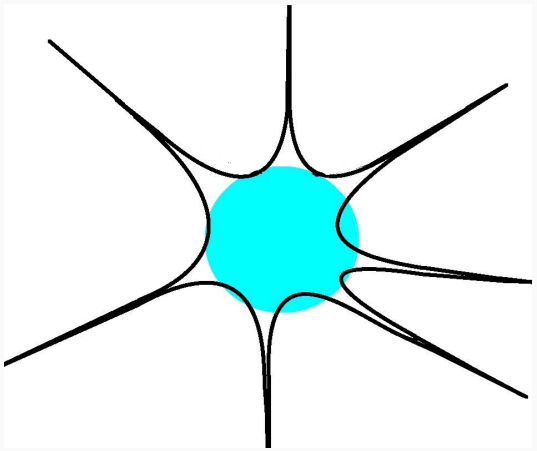
$$\int_0^\eta b(X_s)ds \approx \sum_{i=1}^{k} \alpha_i b(X_{s_i})(s_i - s_{i-1}).$$

Low-dimensional discretization:



$O(\eta^k)$ discretization error for k interpolation points with stepsize η. (e.g. LMM, Runge-Kutta, collocation).

Applied to high-dimensional problems

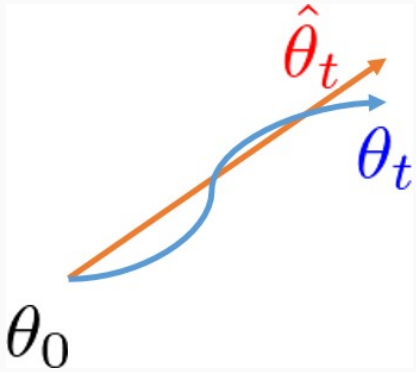


Dimension dependency grows with order of smoothness, leading to $O\left((\eta\sqrt{d})^k\right)$ discretization error.

Idea 1: computing the integral directly

Recall underdamped Langevin diffusion:

$$r_t - r_{k\eta} = \int_{k\eta}^{t} \left(-\frac{1}{\beta} \nabla f(\theta_s) - \xi r_s \right) ds + \int_{k\eta}^{t} \sqrt{2\xi/\beta} dB_s.$$



First approximate $(\theta_s)_{k\eta \le s \le (k+1)\eta}$, and then use the approximate path to compute the integral.

Questions:

- How to compute the integral $\int_{k\eta}^{t} \nabla f(\widehat{\theta}_s) ds$?
- How to incorporate the integral to approximate r_t.

The integration oracle

Given f and a line segment $\{x + tz : t \in [0, 1]\}$, exactly compute the vector-valued integral:

$$J_f(x; z) := \int_0^1 \nabla f(x + tz) dt.$$

e.g. ridge separable function and Newton-Leibniz formula.

$$f(\theta) = \sum_{i=1}^{n} u_i\left(a_i^\top \theta\right),$$

$$\int_0^1 \nabla f(\theta + tp) dt = \sum_i \left(u_i\left(a_i^\top(\theta + p)\right) - u_i\left(a_i^\top \theta\right)\right) \frac{a_i}{a_i^\top p}.$$

- Covers most models for machine learning and Bayesian inference.
- Requires NO quantitative assumption on individual data; only assuming the function f to be strongly convex and smooth.

Problem with the second-order Langevin diffusion

Want to compute the stochastic integral:

$$\widetilde{r}_t - \widetilde{r}_{k\eta} = \underbrace{\int_{k\eta}^{t} \left(-\frac{1}{\beta} \nabla f(\widehat{\theta}_s) \right) ds}_{\text{Integration oracle}} - \underbrace{\int_{k\eta}^{t} \xi \widetilde{r}_s + \int_{k\eta}^{t} \sqrt{2\xi/\beta} dB_s}_{\text{Ornstein-Uhlenbeck}} .$$

Not compatible! The integral affects $\widetilde{r}$ in the path, and enters the OU term. Nonlinear effects exist and cannot be exactly computed.

Solution: separate the Brownian motion and integral oracle part.

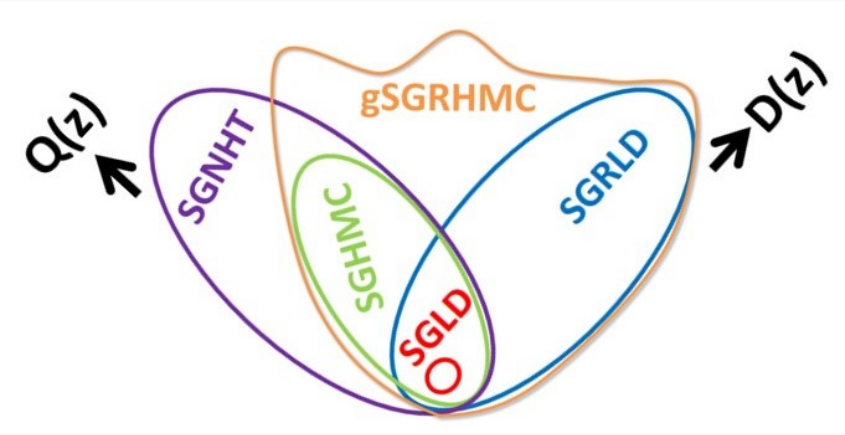
Construction of the third-order Langevin diffusion

Further dimension lifting: the general framework

[MCF15, MFCW18]: a class of Langevin algorithms with the correct stationary distribution $\propto e^{-H}$.

$$dZ_t = -(D+Q)\nabla H(Z_t)dt + \sqrt{2D}dB_t,$$

with $D \succeq 0$, $Q = -Q^\top$ and $D+Q$ of full rank.



Design principle:

- Matrix D with less non-zero entries (smoother curve)
- Matrix Q separates the Brownian motion and path integral.

Third-order Langevin diffusion

Continuous dynamics with noise injected only in one of three variables:

$$\begin{cases} d\theta_t = p_t dt \\ dp_t = -\frac{1}{\beta}\nabla f(\theta_t)dt + \gamma r_t dt \\ dr_t = -\gamma p_t dt - \xi r_t dt + \sqrt{2\xi/\beta}dB_t^r \end{cases}.$$

Continuous-time convergence: Lyapunov function $\Phi_t := \mathbb{E}\left[(Z_t^{(1)} - Z_t^{(2)})^\top S (Z_t^{(1)} - Z_t^{(2)})\right]$, with S given by

$$S := \begin{pmatrix} \frac{\kappa^7+3\kappa^4+5\kappa^3+\kappa+1}{4\kappa^5}\mathrm{I} & \frac{\kappa}{2}\mathrm{I} & \frac{1}{4}\left(1-\frac{1}{\kappa^3}-\frac{1}{\kappa^4}\right)\kappa\mathrm{I} \\ \frac{\kappa}{2}\mathrm{I} & \frac{4\kappa^4+6\kappa^3+\kappa+1}{4\kappa^4}\mathrm{I} & \frac{\kappa+1}{2\kappa}\mathrm{I} \\ \frac{1}{4}\left(1-\frac{1}{\kappa^3}-\frac{1}{\kappa^4}\right)\kappa\mathrm{I} & \frac{\kappa+1}{2\kappa}\mathrm{I} & \frac{\kappa+2}{4\kappa}\mathrm{I} \end{pmatrix}.$$

We have the bound $\frac{d}{dt}\Phi_t \leq -\frac{1}{5\kappa^2+10}\Phi_t$.

Construction idea: Lyapunov stability criteria, $SL^\top + LS \prec 0$ and $S \succ 0$.

Improved discretization scheme

Main results

Theorem (informal)
Given f satisfying the $\mu I_d \preceq \nabla^2 f \preceq \beta I_d$ with $\kappa := \beta/\mu = O(1)$. There exists a discretization scheme of the third-order Langevin diffusion, such that for any $\varepsilon > 0$, by taking stepsize:

$$\eta \asymp \varepsilon^{-1/2} d^{-1/4},$$

the algorithm output $\theta^{(k)} \sim \pi^{(k)}$ with $\mathcal{W}_2(\pi^{(k)}, \pi^*) \leq \varepsilon$, with iterations:

$$k = O\left(d^{1/4}\varepsilon^{-1/2} \log \varepsilon^{-1}\right).$$

Each iteration invokes an integral oracle and additional computation feasible within $O(d)$ time (same as the cost of gradient oracle).

Stage I: constructing coarse estimates for the path

Recall the third-order dynamics:

$$\begin{cases} d\theta_t = p_t dt \\ dp_t = -\frac{1}{\beta}\nabla f(\theta_t)dt + \gamma r_t dt \\ dr_t = -\gamma p_t dt - \xi r_t dt + \sqrt{2\xi/\beta}dB_t^r. \end{cases}$$

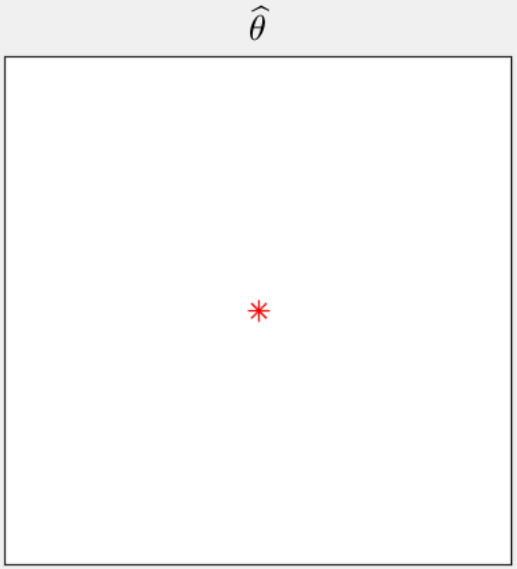
Integration oracle for θ available on straight line segment.

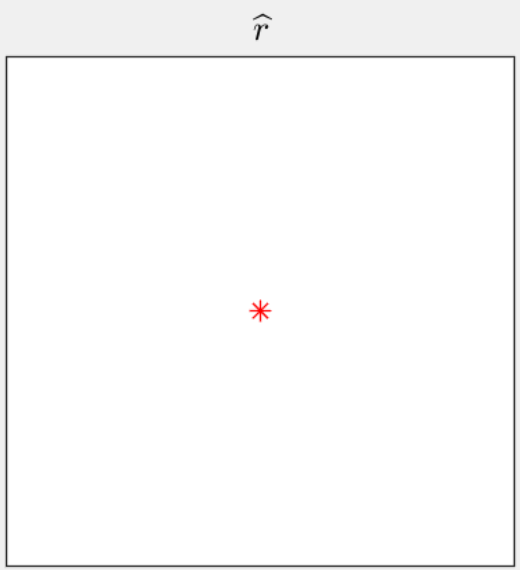
$$\begin{cases} \mathrm{d}\widehat{\theta}_t = \widetilde{p}_{k\eta}\mathrm{d}t, \\ \mathrm{d}\widehat{r}_t = -\gamma\widetilde{p}_{k\eta}\mathrm{d}t - \xi\widehat{r}_t\mathrm{d}t + \sqrt{2\xi/L}\mathrm{d}B_t^r, \end{cases} \quad \text{for all } t \in [k\eta, (k+1)\eta].$$

Prepare the coarse path for computing $\widetilde{p}_t$ using integration oracle.

$$\left\|\widetilde{\theta}_{k\eta} - \widehat{\theta}_t\right\|_2 \le O(\eta\sqrt{d}), \quad \|\widetilde{r}_{k\eta} - \widehat{r}_t\|_2 \le O(\eta\sqrt{d}).$$

Stage I: constructing coarse estimates for the path





Stage II: Using the coarse path and integration oracle

$$\begin{cases} \mathrm{d}\widetilde{p}_t = -\frac{1}{\beta}\nabla f\left(\widehat{\theta}_t\right)\mathrm{d}t + \gamma\widehat{r}_t dt, \\ \mathrm{d}\widehat{p}_t = -\frac{1}{\beta\eta}\left(\int_{k\eta}^{(k+1)\eta} \nabla f\left(\widehat{\theta}_t\right)\mathrm{d}s\right)\mathrm{d}t + \gamma\widehat{r}_t dt, \end{cases} \qquad \forall t \in [k\eta, (k+1)\eta].$$

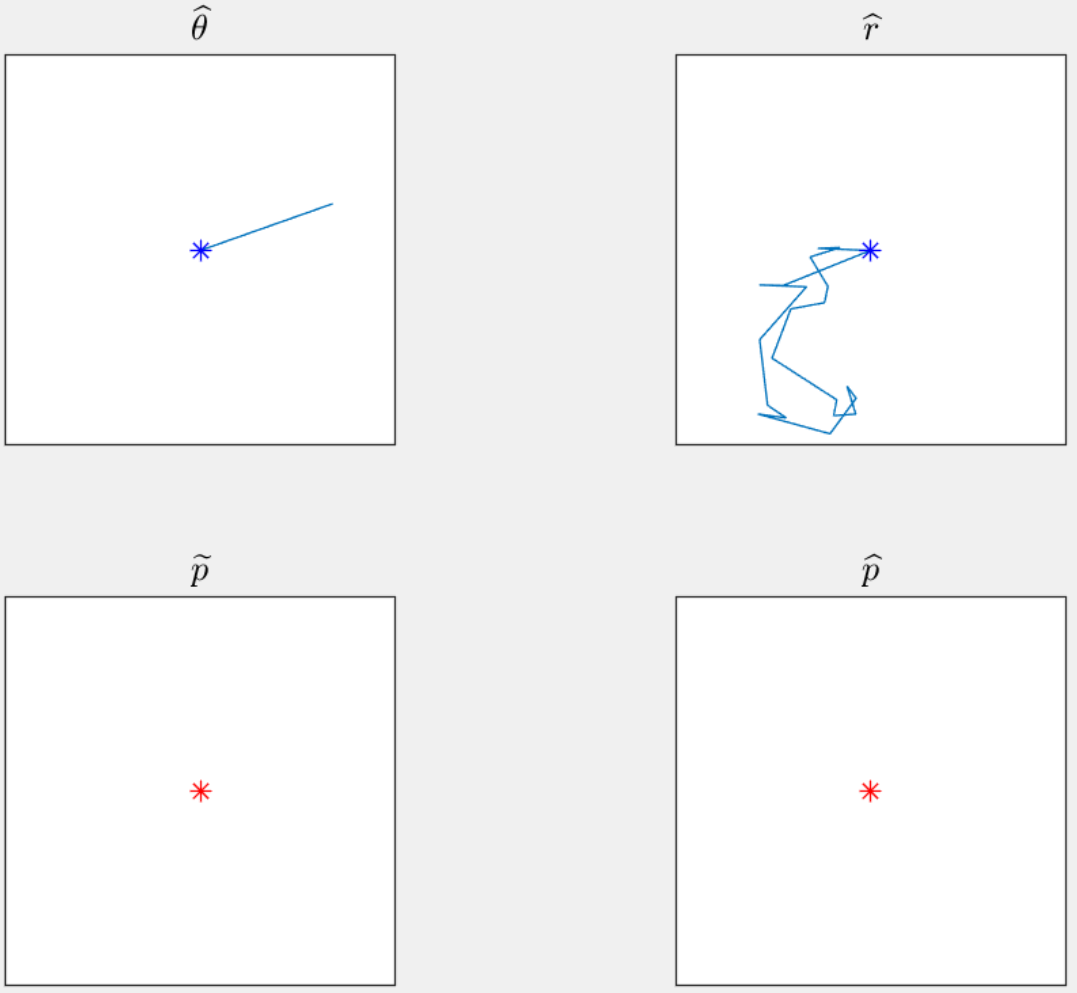
Only the endpoint $\widetilde{p}_{(k+1)\eta}$ used (for next step's stage 1). Computable through integration oracle:

$$\widetilde{p}_{(k+1)\eta} = \widetilde{p}_{k\eta} - \underbrace{\frac{1}{\beta}\int_0^\eta \nabla f\left(\widehat{\theta}_{k\eta} + t\widetilde{p}_{k\eta}\right)\mathrm{d}t}_{\text{integration oracle}} + \underbrace{\int_0^\eta \gamma\widehat{r}_t dt}_{\text{additive Gaussian}}$$

Another coarse approximation $\widehat{p}$ used in the third stage.

$$\widetilde{p}_{(k+1)\eta} = \widehat{p}_{(k+1)\eta}, \quad \|\widetilde{p}_t - \widehat{p}_t\|_2 \le O(\eta^2\sqrt{d}).$$

Stage II: Using the coarse path and integration oracle



Stage III: using $\widehat{p}$ to compute next-step $(\widetilde{\theta}, \widetilde{r})$

Substituting the refined path $\widehat{p}$ back to the processes $\widetilde{\theta}$ and $\widetilde{r}$ to correct the path

$$\begin{cases} \mathrm{d}\widetilde{\theta}_t = \widehat{p}_t \mathrm{d}t \\ \mathrm{d}\widetilde{r}_t = -\gamma \widehat{p}_t \mathrm{d}t - \xi \widetilde{r}_t \mathrm{d}t + \sqrt{2\xi/\beta}\, \mathrm{d}B_t^r, \end{cases} \qquad t \in [k\eta, (k+1)\eta].$$

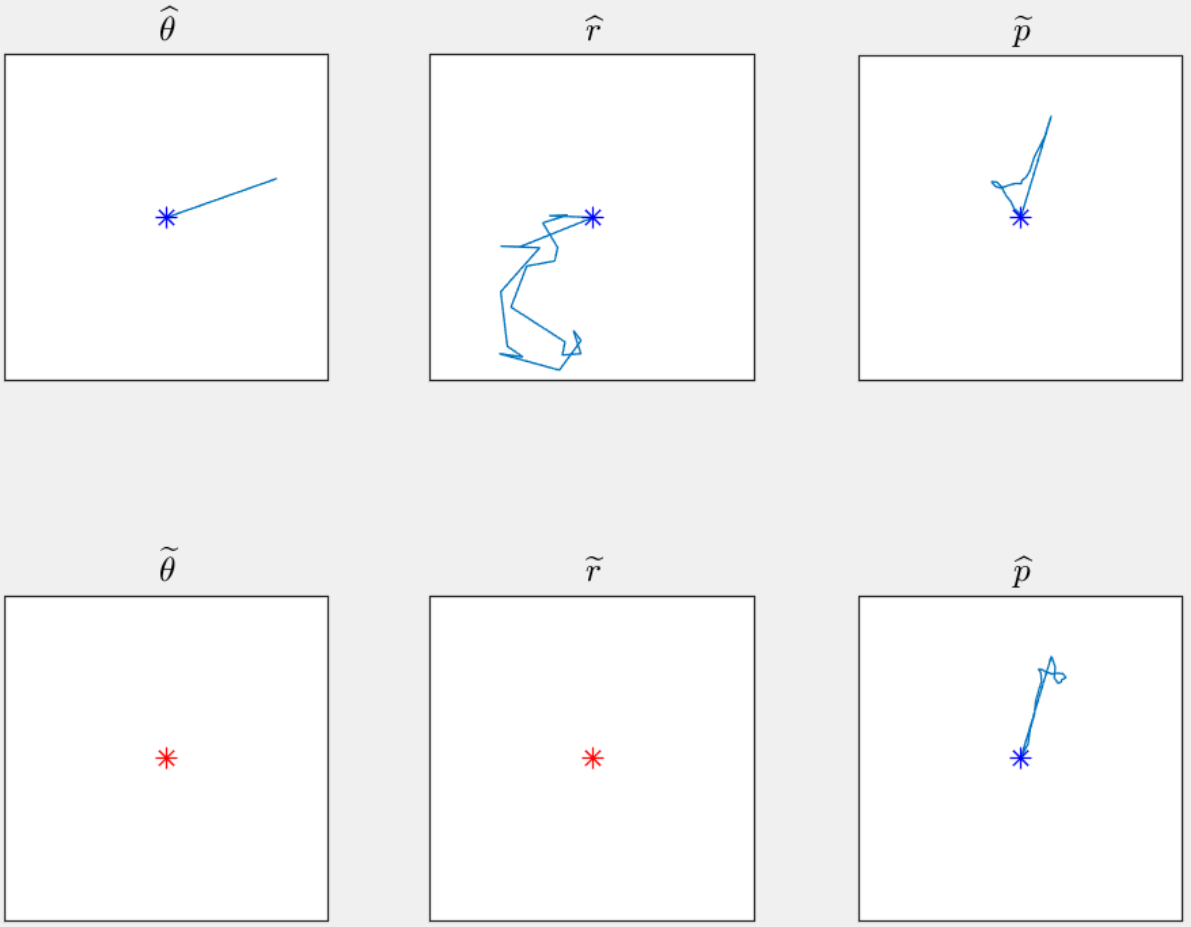
The $5d$-dimensional process $(\widetilde{\theta}, \widehat{\theta}, \widehat{p}, \widetilde{r}, \widehat{r})$ is jointly OU, using pre-computed integral oracle (a deterministic quantity given the filtration up to time $k\eta$). Implementation can be explicit!

Key properties used in the analysis:

$$\left\| \widehat{\theta}_t - \widetilde{\theta}_t \right\|_2 \le O(\eta^2 \sqrt{d}), \quad \|\widehat{r}_t - \widetilde{r}_t\|_2 \le O(\eta^2 \sqrt{d}).$$

Substituting the contraction of the Lyapunov function, leading to $O(\eta^2 \sqrt{d})$ final discretization error bound.

Stage II: Using the coarse path and integration oracle



Discussion and open problems

- High-order Langevin diffusion + special structure allowing integration $\Rightarrow$ faster high-dimensional sampling algorithm.
- $O(d^{1/4}\varepsilon^{-1/2})$ mixing time, best known dimension dependency for such class, easy-to-implement explicit scheme.

Open questions:

- Integration oracle on a curve – even better dimension dependence via further lifted dynamics.
- Use of integration oracle in other fields of computing (numerical analysis, optimization)
- Improved dimension dependence without the integration oracle.

Thank you!