High order numerical simulation of the underdamped Langevin diffusion

James Foster

University of Bath
(joint with Terry Lyons and Harald Oberhauser)

Outline

What is the underdamped Langevin diffusion?

The underdamped Langevin diffusion (ULD) is a model for molecular dynamics and is given by the stochastic differential equation (SDE):

$$\begin{aligned} dx_t &= v_t\, dt, \\ dv_t &= -\gamma v_t\, dt - u\nabla f(x_t)\, dt + \sqrt{2\gamma u}\, dW_t, \end{aligned} \tag{1}$$

where

- $x, v \in \mathbb{R}^d$ will represent the *position* and *momentum* of a particle
- $f : \mathbb{R}^d \to \mathbb{R}$ is a *scalar potential* that the particle moves around in
- $\gamma > 0$ is the *friction* coefficient
- $u > 0$ is the *gradient* coefficient (often just set to $u = 1$)
- $W = \{W_t\}_{t\geq 0}$ is a standard d-dimensional Brownian motion

$$dW_t \sim \mathcal{N}(0, I_d\, dt)$$

Applications of the underdamped Langevin diffusion

"ULD = Newton's second law + frictional forces + stochastic forces"

$$
\begin{aligned}
dx_t &= v_t\, dt, \\
dv_t &= \underbrace{-\gamma v_t\, dt}_{\text{friction}} \;-\; \underbrace{u\nabla f(x_t)\, dt}_{\substack{\text{gradient of the} \\ \text{potential / target}}} \;+\; \underbrace{\sigma\, dW_t}_{\text{noise}},
\end{aligned}
$$

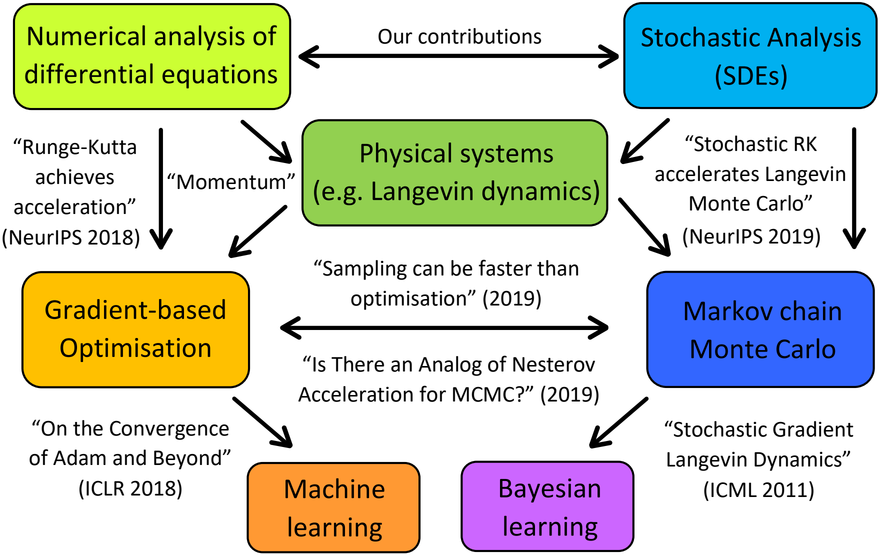
Under mild assumptions on f, the SDE admits a unique strong solution that is ergodic with stationary distribution $\pi(x,v) \propto e^{-f(x)+\frac{1}{2u}\|v\|^2}$ [1].

In addition to being a fundamental model in statistical mechanics [2], ULD has recently been applied to sampling problems in data science as simulating ULD allows one to generate samples from $\pi(x) \propto e^{-f(x)}$.

(technically, samples are "close" to π in an optimal transport sense [3])

In practice, (1) cannot be solved exactly, so we must approximate ULD.

Langevin MCMC as part of a larger ecosystem



Outline

A high order ODE-based approximation of ULD

One strategy for discretizing the underdamped Langevin diffusion is

1. Replace the Brownian motion W by a piecewise linear path $\widehat{W}$.

2. Along each piece of $\widehat{W}$, we approximate the SDE (1) using the ODE:

$$d\widehat{x}_t = \widehat{v}_t \, dt, \tag{2}$$

$$d\widehat{v}_t = -\gamma \widehat{v}_t \, dt - u \nabla f(\widehat{x}_t) \, dt + \sigma \, d\widehat{W}_t, \tag{3}$$

where $\sigma := \sqrt{2\gamma u}$.

3. In each step, we discretize (2) and (3) using a suitable ODE solver.

Note that we will use the notation $W_{s,t} := W_t - W_s$ and $\widehat{W}_{s,t} := \widehat{W}_t - \widehat{W}_s$.

The stochastic Taylor expansion of ULD

When f is three times differentiable, ULD admits the Taylor expansion:

$$\begin{aligned}\begin{pmatrix} x_t \\ v_t \end{pmatrix} = \begin{pmatrix} x_s \\ v_s \end{pmatrix} + \begin{pmatrix} v_s \\ -\gamma v_s - u\nabla f(x_s) \end{pmatrix}(t-s) + \begin{pmatrix} 0 \\ \sigma \end{pmatrix} W_{s,t} \\ + (\cdots)\int_s^t W_{s,r}\,dr + (\cdots)(t-s)^2 + (\cdots)(t-s)^3 \\ + (\cdots)\int_s^t\int_s^r W_{s,v}\,dv\,dr + (\cdots)\int_s^t\int_s^r\int_s^v W_{s,w}\,dw\,dv\,dr \\ + (\cdots)\int_s^t\int_s^u\int_s^v (r-s)\,dW_r\,dv\,du + R_{s,t}, \end{aligned} \tag{4}$$

where $(\cdots)$ are terms involving the vector fields and their derivatives.

Theorem (Stochastic Taylor expansion, Theorem 5.5.1 of [10])

If $\mathbb{E}\big[\|\nabla f(x_s)\|_2^4\big] < \infty$, $\mathbb{E}\big[\|v_s\|_2^8\big] < \infty$ *and* $\nabla^k f$ *is Lipschitz continuous for* $k = 1, 2, 3$ *then* $\mathbb{E}\big[\|R_{s,t}\|_2^2\big]^{\frac{1}{2}} \sim O\big((t-s)^4\big)$.

The Taylor expansion of the ODE approximation

When f is three times differentiable, (2, 3) admits the Taylor expansion:

$$\begin{aligned}\begin{pmatrix}\widehat{x}_t\\ \widehat{v}_t\end{pmatrix} &= \begin{pmatrix}\widehat{x}_s\\ \widehat{v}_s\end{pmatrix} + \begin{pmatrix}\widehat{v}_s\\ -\gamma\widehat{v}_s - u\nabla f(\widehat{x}_s)\end{pmatrix}(t-s) + \begin{pmatrix}0\\ \sigma\end{pmatrix}\widehat{W}_{s,t} \\ &\quad + (\cdots)\int_s^t \widehat{W}_{s,r}\,dr + (\cdots)(t-s)^2 + (\cdots)(t-s)^3 \\ &\quad + (\cdots)\int_s^t\int_s^r \widehat{W}_{s,v}\,dv\,dr + (\cdots)\int_s^t\int_s^r\int_s^v \widehat{W}_{s,w}\,dw\,dv\,dr \\ &\quad + (\cdots)\int_s^t\int_s^u\int_s^v (r-s)\,d\widehat{W}_r\,dv\,du + \widehat{R}_{s,t},\end{aligned} \tag{5}$$

where $(\cdots)$ are the same terms appearing in the expansion (4) of ULD.

Remark

We want to construct $\widehat{W}$ to match certain iterated time integrals of W.

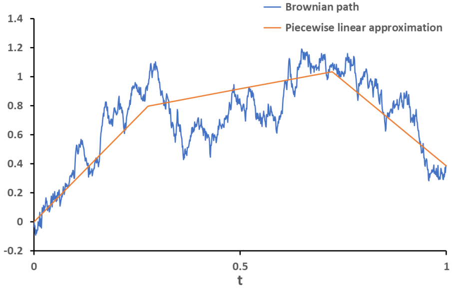
Piecewise linear discretization of Brownian motion

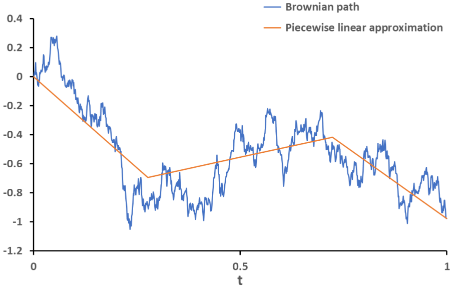
Based on these expansions, we want a piecewise linear path $\widehat{W}$ so that

$$\widehat{W}_{s,t} = W_{s,t}, \tag{6}$$

$$\int_s^t \widehat{W}_{s,r}\, dr = \int_s^t W_{s,r}\, dr, \tag{7}$$

$$\int_s^t \int_s^r \widehat{W}_{s,v}\, dv\, dr = \int_s^t \int_s^r W_{s,v}\, dv\, dr. \tag{8}$$

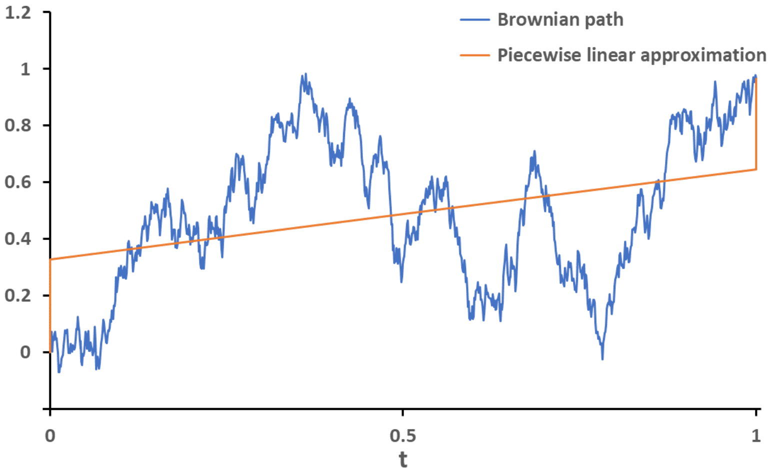




Piecewise linear discretization of Brownian motion

To reduce computational cost, we construct $\widehat{W}$ using "vertical pieces"!

Along vertical pieces, $dt = 0$ so the ODE becomes $d\widehat{x}_t = 0, d\widehat{v}_t = \sigma\, d\widehat{W}_t$.



Generating iterated time integrals of Brownian motion

Let $\{t_n\}_{n\geq 0}$ be a sequence of times with $t_0 = 0$ and $t_{n+1} > t_n$. We define

$$h_n := t_{n+1} - t_n\,,$$

$$W_n := W_{t_{n+1}} - W_{t_n}\,,$$

$$H_n := \frac{1}{h_n}\int_{t_n}^{t_{n+1}}\bigg(\big(W_t - W_{t_n}\big) - \frac{t - t_n}{h_n}W_n\bigg)\,dt,$$

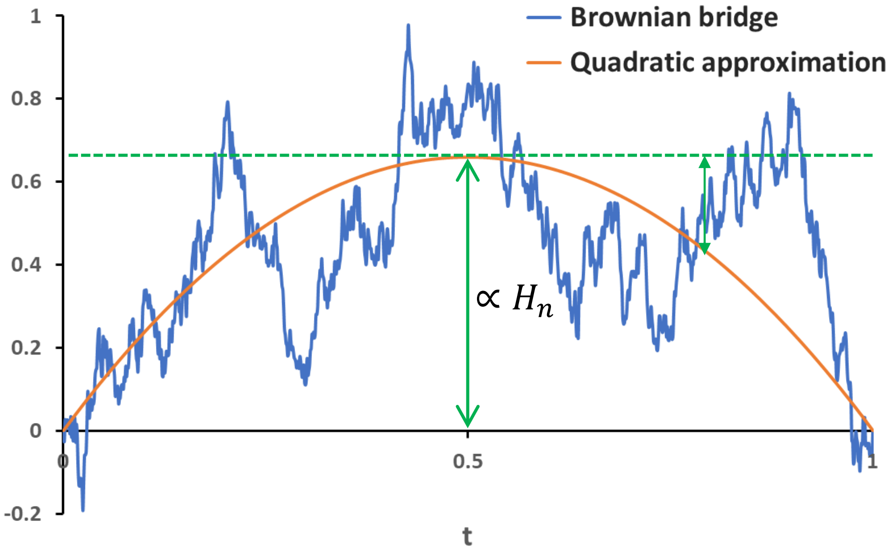
$$K_n := \frac{1}{h_n^2}\int_{t_n}^{t_{n+1}}\bigg(\frac{1}{2}h_n - (t - t_n)\bigg)\bigg(\big(W_t - W_{t_n}\big) - \frac{t - t_n}{h_n}W_n\bigg)\,dt.$$

Lemma (Direct consequence of Theorem 2.2 in [11])

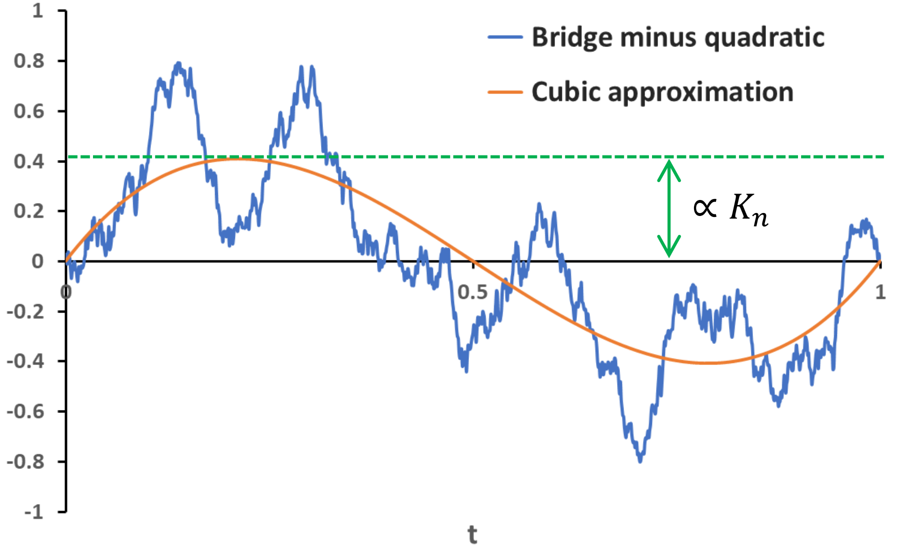
The random vectors W_n, H_n, K_n are independent and distributed as

$$W_n \sim \mathcal{N}\big(0, I_d\, h_n\big), \qquad H_n \sim \mathcal{N}\Big(0, \frac{1}{12} I_d\, h_n\Big), \qquad K_n \sim \mathcal{N}\Big(0, \frac{1}{720} I_d\, h_n\Big).$$

Generating iterated time integrals of Brownian motion



Generating iterated time integrals of Brownian motion



Generating a piecewise linear approximation of W

It is straightforward to obtain the iterated integrals from (W_n, H_n, K_n).

Lemma (Iterated integrals and polynomial coefficients of W)

$$\int_{t_n}^{t_{n+1}} W_{t_n,t}\, dt = \frac{1}{2}h_n W_n + h_n H_n\,,$$

$$\int_{t_n}^{t_{n+1}} \int_{t_n}^{t} W_{t_n,s}\, ds\, dt = \frac{1}{6}h_n^2 W_n + \frac{1}{2}h_n^2 H_n + h_n^2 K_n\,.$$

Definition (Piecewise linear discretization of Brownian motion)

We define $\widehat{W}$ on each $[t_n, t_{n+1}]$ as the piecewise linear path connecting

$$\big(t_n, W_{t_n}\big),\ \big(t_n, W_{t_n} + H_n + 6K_n\big),\ \big(t_{n+1}, W_{t_{n+1}} + H_n - 6K_n\big),\ \big(t_{n+1}, W_{t_{n+1}}\big)$$

in said order.

From the lemma, we can check that $\widehat{W}$ satisfies properties (6), (7), (8).

An ODE method for underdamped Langevin dynamics

Definition (Shifted ODE method for ULD [12])

We define a numerical solution $\{(\widetilde{x}_n, \widetilde{v}_n)\}$ by setting $(\widetilde{x}_0, \widetilde{v}_0) := (x_0, v_0)$ and for each $n \geq 0$, defining $(\widetilde{x}_{n+1}, \widetilde{v}_{n+1})$ as

$$\begin{pmatrix} \widetilde{x}_{n+1} \\ \widetilde{v}_{n+1} \end{pmatrix} := \begin{pmatrix} \overline{x}_1^n \\ \overline{v}_1^n \end{pmatrix} - (H_n - 6K_n)\begin{pmatrix} 0 \\ \sigma \end{pmatrix}$$

where $\{(\overline{x}_t^n, \overline{v}_t^n)\}_{t\in[0,1]}$ solves the following Langevin-type ODE,

$$\frac{d}{dt}\begin{pmatrix} \overline{x}^n \\ \overline{v}^n \end{pmatrix} = \begin{pmatrix} \overline{v}^n \\ -\gamma\overline{v}^n - u\nabla f(\overline{x}^n) \end{pmatrix} h_n + (W_n - 12K_n)\begin{pmatrix} 0 \\ \sigma \end{pmatrix}, \qquad (9)$$

with initial condition

$$\begin{pmatrix} \overline{x}_0^n \\ \overline{v}_0^n \end{pmatrix} := \begin{pmatrix} \widetilde{x}_n \\ \widetilde{v}_n \end{pmatrix} + (H_n + 6K_n)\begin{pmatrix} 0 \\ \sigma \end{pmatrix}.$$

Outline

Error analysis of the ODE approximation

Theorem (Convergence of shifted ODE with fixed step sizes [12])

Suppose that the function $f : \mathbb{R}^d \to \mathbb{R}$ is m-strongly convex,

$$f(y) \geq f(x) + \big\langle \nabla f(x), y - x \big\rangle + \frac{1}{2} m \|x - y\|_2^2 \,, \tag{10}$$

and twice continuously differentiable with an M-Lipschitz continuous gradient ∇f,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq M \|x - y\|_2 \,, \tag{11}$$

for all $x, y \in \mathbb{R}^d$.

Let $\{(x_t, v_t)\}$ and $\{(\widetilde{x}_n, \widetilde{v}_n)\}$ be defined from the same Brownian motion, $(x_0, v_0) \sim \pi$ and $\widetilde{v}_0 = v_0 \sim \mathcal{N}(0, uI_d)$. Then there exists $c_0, c_1 > 0$ so that

$$\| \widetilde{x}_n - x_{t_n} \|_{L^2(\mathbb{P})} \leq c_0 \, e^{-n\alpha h} \| \widetilde{x}_0 - x_0 \|_{L^2(\mathbb{P})} + c_1 \sqrt{d} \, h^{\frac{3}{2}} \,, \tag{12}$$

Error analysis of the ODE approximation

Theorem (Convergence of the shifted ODE method (continued))

where for a random vector X and $p \geq 1$, the norm $\|X\|_{L^p(\mathbb{P})}$ is defined as

$$\|X\|_{L^p(\mathbb{P})} := \mathbb{E}\big[\|X\|_2^p\big]^{\frac{1}{p}},$$

and the rate of contraction α is given by

$$\alpha = \frac{(\gamma^2 - uM) \vee um}{\gamma}.$$

In addition, if $\nabla^2 f$ is Lipschitz continuous then there exists $c_2 > 0$ so that

$$\|\widetilde{x}_n - x_{t_n}\|_{L^2(\mathbb{P})} \leq c_0\, e^{-n\alpha h}\|\widetilde{x}_0 - x_0\|_{L^2(\mathbb{P})} + c_2\, d\, h^{\frac{5}{2}}. \tag{13}$$

If $\nabla^2 f$ and $\nabla^3 f$ are Lipschitz continuous then there exists $c_3 > 0$ so that

$$\|\widetilde{x}_n - x_{t_n}\|_{L^2(\mathbb{P})} \leq c_0\, e^{-\frac{1}{2}n\alpha h}\|\widetilde{x}_0 - x_0\|_{L^2(\mathbb{P})} + c_3\, d^{\frac{3}{2}}\, h^3. \tag{14}$$

Error analysis of the ODE approximation

Sketch Proof.

Firstly, we use a change of variable to rewrite the ODE approximation as

$$\begin{pmatrix} \widetilde{x}_{n+1} \\ \widetilde{v}_{n+1} \end{pmatrix} := \begin{pmatrix} \widehat{x}^n_{t_{n+1}} \\ \widehat{v}^n_{t_{n+1}} \end{pmatrix} + 12K_n \begin{pmatrix} 0 \\ \sigma \end{pmatrix},$$

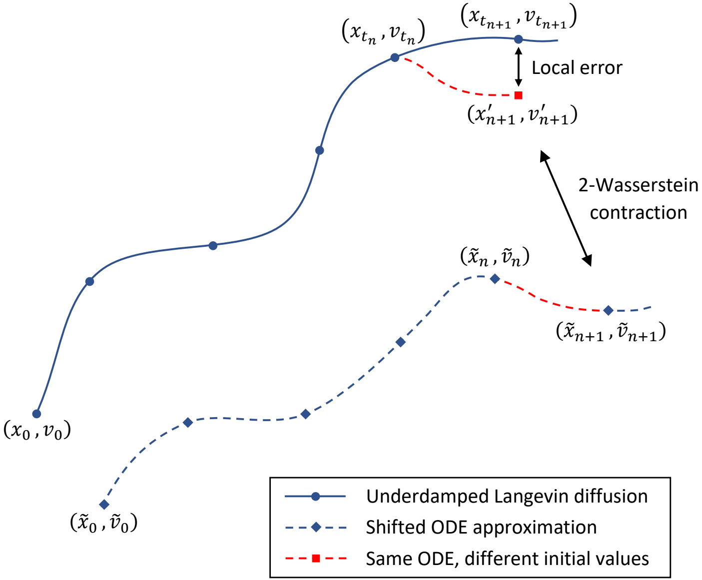
where $\big\{\big(\widehat{x}^n_t, \widehat{v}^n_t\big)\big\}_{t\in[t_n, t_{n+1}]}$ solves the following Langevin-type ODE,

$$\frac{d}{dt}\begin{pmatrix} \widehat{x}^n \\ \widehat{v}^n \end{pmatrix} = \begin{pmatrix} \widehat{v}^n + \sigma\big(H_n + 6K_n\big) \\ -\gamma\big(\widehat{v}^n + \sigma\big(H_n + 6K_n\big)\big) - u\nabla f\big(\widehat{x}^n\big) \end{pmatrix} h_n + \frac{W_n - 12K_n}{h_n}\begin{pmatrix} 0 \\ \sigma \end{pmatrix},$$

with initial condition $\big(\widehat{x}^n_{t_n}, \widehat{v}^n_{t_n}\big) := \big(\widetilde{x}_n, \widetilde{v}_n\big)$.

This will help us establish local error estimates at times within $[t_n, t_{n+1}]$.

Error analysis of the ODE approximation



Error analysis of the ODE approximation

Sketch Proof. (Global bounds on the diffusion process)

- Since $(x_0, v_0) \sim \pi$ we have $(x_t, v_t) \sim \pi$ for all $t \geq 0$. In particular,

$$v_t \sim \mathcal{N}\big(0, uI_d\big), \tag{15}$$

and

$$\mathbb{E}\big[\|\nabla f(x_t)\|_2^2\big] \leq Md, \tag{16}$$

$$\mathbb{E}\big[\|\nabla f(x_t)\|_2^4\big] \leq 3M^2d^2, \tag{17}$$

for all $t \geq 0$ (see Lemma 2 in [13] and Theorem C.11 in [12]).

- Since we start the ODE approximation from $\big(x_{t_n}, v_{t_n}\big)$, we can use the above to estimate errors without imposing boundedness on ∇f.

- We note that the strong convexity assumption is not required here.

Error analysis of the ODE approximation

Sketch Proof. (Crude local error estimates)

- We also define

$$\begin{pmatrix} x'_{n+1} \\ v'_{n+1} \end{pmatrix} := \begin{pmatrix} \widehat{x}^n_{t_{n+1}} \\ \widehat{v}^n_{t_{n+1}} \end{pmatrix} + 12K_n \begin{pmatrix} 0 \\ \sigma \end{pmatrix},$$

where $\big\{\big(\widehat{x}^n_t, \widehat{v}^n_t\big)\big\}_{t\in[t_n,t_{n+1}]}$ is the solution to the Langevin-type ODE used previously, but with initial condition $\big(\widehat{x}^n_{t_n}, \widehat{v}^n_{t_n}\big) := (x_{t_n}, v_{t_n})$.

- Using just the Lipschitz regularity of ∇f, we can obtain the estimates

$$\|x_t - \widehat{x}^n_t\|_{L^p(\mathbb{P})} \leq C_1(p)\sqrt{d}\,(h_n)^{\frac{1}{2}}(t - t_n), \tag{18}$$

$$\|v_t - \widehat{v}^n_t\|_{L^p(\mathbb{P})} \leq C_2(p)\sqrt{d}\,(t - t_n)^{\frac{1}{2}}, \tag{19}$$

for $p \in \{2, 4, 8\}$ and $t \in [t_n, t_{n+1}]$.

Error analysis of the ODE approximation

Sketch Proof. (Local error estimates)

- Now that we have $L^p(\mathbb{P})$ bounds for the SDE and ODE solutions, we can Taylor expand $\left\{\left(x_t^n, v_t^n\right)\right\}_{t\in[t_n,t_{n+1}]}$ and $\left\{\left(\widehat{x}_t^n, \widehat{v}_t^n\right)\right\}_{t\in[t_n,t_{n+1}]}$ and estimate the remainder terms.

- Using the Lipschitz regularity of ∇f, we can show, for $p \in \{2, 4, 8\}$,

$$\|x_{t_{n+1}} - x_{n+1}'\|_{L^p(\mathbb{P})} \leq C_3(p)\sqrt{d}\,(h_n)^{\frac{7}{2}}, \tag{20}$$

$$\|v_{t_{n+1}} - v_{n+1}'\|_{L^p(\mathbb{P})} \leq C_4(p)\sqrt{d}\,(h_n)^{\frac{5}{2}}. \tag{21}$$

- In addition, from the Lipschitz regularity of ∇f and $\nabla^2 f$, we have

$$\|v_{t_{n+1}} - v_{n+1}'\|_{L^2(\mathbb{P})} \leq C_5\, d\,(h_n)^{\frac{7}{2}}. \tag{22}$$

- Again, we note that the strong convexity of f is not required here.

Error analysis of the ODE approximation

Theorem (Exponential contractivity of the ODE approximation)

Suppose that f is m-strongly convex and ∇f is M-Lipschitz continuous. Let $\lambda \in [0, \frac{1}{2}\gamma)$ and define $\eta := \gamma - \lambda$. Then for $n \geq 0$, we have

$$\left\| \begin{pmatrix} (\lambda \widetilde{x}_{n+1} + \widetilde{v}_{n+1}) - (\lambda x'_{n+1} + v'_{n+1}) \\ (\eta \widetilde{x}_{n+1} + \widetilde{v}_{n+1}) - (\eta x'_{n+1} + v'_{n+1}) \end{pmatrix} \right\|_2 \leq e^{-\alpha h_n} \left\| \begin{pmatrix} (\lambda \widetilde{x}_n + \widetilde{v}_n) - (\lambda x_{t_n} + v_{t_n}) \\ (\eta \widetilde{x}_n + \widetilde{v}_n) - (\eta x_{t_n} + v_{t_n}) \end{pmatrix} \right\|_2,$$

almost surely, where

$$\alpha = \frac{(\eta^2 - uM) \vee (um - \lambda^2)}{\gamma - 2\lambda}.$$

Proof

Follows by essentially the same argument applied to ULD in [14]. □

Convergence in the 2-Wasserstein metric

Numerical method	Assumptions on the strongly convex f	Number of steps to achieve an error of $W_2(\widetilde{x}_n, e^{-f}) \leq \varepsilon$
Shifted ODE [12]	Lipschitz gradient + Lipschitz $\nabla^2 f$ + Lipschitz $\nabla^3 f$	$\mathcal{O}\big(\sqrt[3]{d}/\varepsilon^{\frac{2}{3}}\big)$ $\mathcal{O}\big(d^{\frac{2}{5}}/\varepsilon^{\frac{2}{5}}\big)$ $\mathcal{O}\big(\sqrt{d}/\varepsilon^{\frac{1}{3}}\big)$
OBABO splitting [15, 16]	Lipschitz gradient + Lipschitz $\nabla^2 f$	$\mathcal{O}\big(\sqrt{d}/\varepsilon\big)$ $\mathcal{O}\big(\sqrt{d}/\sqrt{\varepsilon}\big)$
Randomized midpoint [17, 18]	Lipschitz gradient	$\mathcal{O}\big(\sqrt[3]{d}/\varepsilon^{\frac{2}{3}}\big)$
Left-point method [3]	Lipschitz gradient	$\mathcal{O}\big(\sqrt{d}/\varepsilon\big)$

Discretization of the shifted ODE (SORT method)

Definition (Shifted ODE with Runge-Kutta Three)

$$
\begin{aligned}
V_n^{(1)} &:= V_n + \sigma\big(H_n + 6K_n\big),\\
X_n^{(1)} &:= X_n + \left(\frac{1 - e^{-\frac{1}{2}\gamma h_n}}{\gamma}\right)V_n^{(1)} - \left(\frac{e^{-\frac{1}{2}\gamma h_n} + \frac{1}{2}\gamma h_n - 1}{\gamma^2}\right)u\nabla f(X_n)\\
&\qquad + \sigma\left(\frac{e^{-\frac{1}{2}\gamma h_n} + \frac{1}{2}\gamma h_n - 1}{\gamma^2 h_n}\right)\big(W_n - 12K_n\big),\\
X_{n+1} &:= X_n + \left(\frac{1 - e^{-\gamma h_n}}{\gamma}\right)V_n^{(1)} + \sigma\left(\frac{e^{-\gamma h_n} + \gamma h_n - 1}{\gamma^2 h_n}\right)\big(W_n - 12K_n\big)\\
&\qquad - \left(\frac{e^{-\gamma h_n} + \gamma h_n - 1}{\gamma^2}\right)\left(\frac{1}{3}u\nabla f(X_n) + \frac{2}{3}u\nabla f(X_n^{(1)})\right),\\
V_{n+1} &:= e^{-\gamma h_n}V_n^{(1)} + \sigma\left(\frac{1 - e^{-\gamma h_n}}{\gamma h_n}\right)\big(W_n - 12K_n\big) - \sigma\big(H_n - 6K_n\big)\\
&\qquad - uh_n\left(\frac{1}{6}e^{-\gamma h_n}\nabla f(X_n) + \frac{2}{3}e^{-\frac{1}{2}\gamma h_n}\nabla f(X_n^{(1)}) + \frac{1}{6}\nabla f(X_{n+1})\right).
\end{aligned}
$$

Outline

A second order SDE approximation of ULD

1. Evaluate ∇f at X_n

2. Compute X_{n+1} by solving the following SDE on $[t_n\,, t_{n+1}]$:

$$
\begin{aligned}
dx_t &= v_t\,dt,\\
dv_t &= -\lambda v_t\,dt - u\nabla f(X_n)\,dt + \sigma\,dW_t,
\end{aligned}
$$

with initial value $(X_n\,, V_n)$. Computing V_{n+1} gives the left-point method.

3. Evaluate ∇f at X_{n+1}

4. Compute V_{n+1} by solving the following SDE on $[t_n\,, t_{n+1}]$:

$$
dv_t = -\lambda v_t\,dt - u\Big(\nabla f(X_n) + \frac{t - t_n}{h_n}\big(f(X_{n+1}) - f(X_n)\big)\Big)dt + \sigma\,dW_t.
$$

with initial value V_n.

A second order SDE approximation of ULD

Definition (ULD discretized by linearly interpolating gradients)

$$X_{n+1} := X_n + \Big(\frac{1 - e^{-\gamma h_n}}{\gamma}\Big)V_n - \Big(\frac{e^{-\gamma h_n} + \gamma h_n - 1}{\gamma^2}\Big)u\nabla f(X_n)$$
$$+ \sigma\int_{t_n}^{t_{n+1}}\int_{t_n}^{t} e^{-\gamma(t-s)}\,dW_s\,dt,$$

$$V_{n+1} := e^{-\gamma h_n}V_n - \Big(\frac{1 - (1+\gamma h_n)e^{-\gamma h_n}}{\gamma^2 h_n}\Big)u\nabla f(X_n)$$
$$- \Big(\frac{e^{-\gamma h_n} + \gamma h_n - 1}{\gamma^2 h_n}\Big)u\nabla f(X_{n+1})$$
$$+ \sigma\int_{t_n}^{t_{n+1}} e^{-\gamma(t_{n+1}-t)}\,dW_t.$$

Although we only expect a $O(h^2)$ convergence rate, this method has the advantage that it only uses one additional gradient evaluation per step.

Numerical experiment (logistic regression)

- The dataset is m pairs of labels $y_i \in \{-1, 1\}$ and features $x_i \in \mathbb{R}^d$.
- Target density $\pi(\theta) \propto \exp(-f(\theta))$ comes from a logistic regression:

$$f(\theta) = \frac{\delta}{2}\|\theta\|_2^2 + \sum_{i=1}^{m} \log\left(1 + \exp(-y_i x_i^\top \theta)\right),$$

where δ is a regularization parameter which we will set to $\delta = 0.1$.

- German credit data from UCI repository [19] $(m = 1000, d = 49)$.
- Estimate $L^2(\mathbb{P})$ error by simulating chains with step sizes h and $\frac{1}{2}h$:

$$S_{N,n} := \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\| \overrightarrow{\theta}_{N,i}^{\,h} - \overrightarrow{\theta}_{N,i}^{\,\frac{1}{2}h} \right\|_2^2}\,,$$

where we use a fixed time horizon $T = 1000$ with step size $h = \frac{T}{N}$.

Numerical experiment (logistic regression)

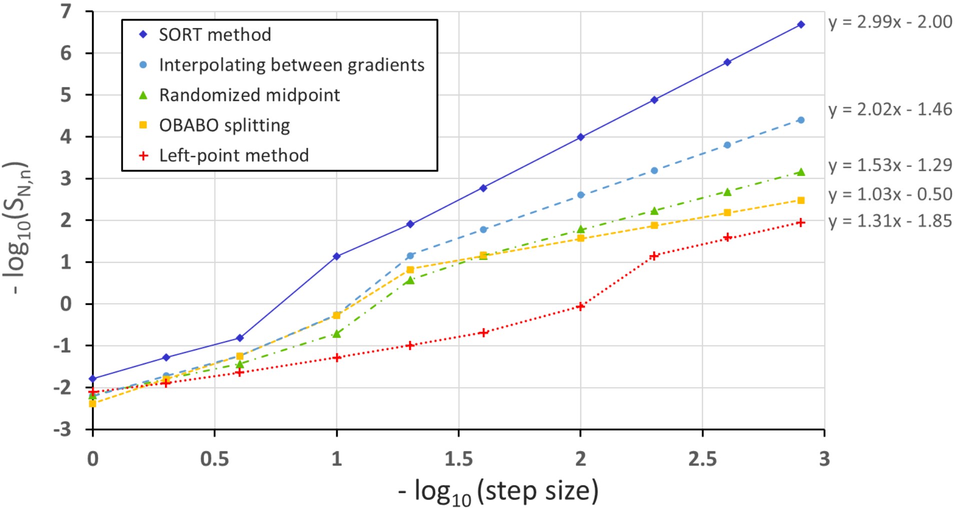


Figure: $S_{N,n}$ computed with $n = 100$ sample paths using a fixed step size.

Outline

Conclusion

- Shifted ODE method
  - First to achieve third order convergence without derivatives of ∇f
  - Scales sublinearly with the dimension d
  - Allows one to use modern ODE solvers
  - No problem with adaptive step sizes
- SORT method (Shifted ODE with Runga-Kutta Three)
  - Practical (two additional gradient evaluations per step)
  - Can empirically demonstrate third order convergence
  - Difficult to analyse!
- Interpolating between gradients
  - Very practical (one additional gradient evaluation per step)
  - Should be possible to establish second order convergence
  - Natural candidate for noisy gradients

Thank you
for your attention!

and our preprint can be found at:

J. Foster, T. Lyons and H. Oberhauser, *The shifted ODE method for underdamped Langevin MCMC*, arxiv.org/abs/2101.03446, 2021.

Outline

References I

G. A. Pavliotis. *Stochastic Processes and Applications*, Springer, New York, 2014.

W. T. Coffey, Y. P. Kalmykov and J. T. Waldron, *The Langevin Equation: With Applications to Stochastic Problems in Physics, Chemistry and Electrical Engineering*, World Scientifc, 2012.

X. Cheng, N. S. Chatterji, P. L. Bartlett and M. I. Jordan. *Underdamped Langevin MCMC: A non-asymptotic analysis*, Proceedings of Machine Learning Research, vol. 75, 2018.

J. Zhang, A. Mokhtari, S. Sra and A. Jadbabaie. *Direct Runge-Kutta Discretization Achieves Acceleration*, Advances in Neural Information Processing Systems, 2018.

References II

X. Li, D. Wu, L. Mackey and M. A. Erdogdu. *Stochastic Runge-Kutta Accelerates Langevin Monte Carlo and Beyond*, Advances in Neural Information Processing Systems, 2019.

Y. Ma, Y. Chen, C. Jin, N. Flammarion and M. I. Jordan. *Sampling can be faster than optimization*, Proceedings of the National Academy of Sciences of the USA, vol. 116, no. 42, 2019.

Y. Ma, , N. Chatterji, X. Cheng, N. Flammarion, P. Bartlett and M. I. Jordan. *Is There an Analog of Nesterov Acceleration for MCMC?*, https://arxiv.org/abs/1902.00996, 2019.

S. J. Reddi, S. Kale and S. Kumar. *On the Convergence of Adam and Beyond* Proceedings of the International Conference on Learning Representations (ICLR), 2018.

References III

- M. Welling and Y. W. Teh. *Bayesian Learning via Stochastic Gradient Langevin Dynamics*, Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.
- P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, 1992.
- J. Foster, T. Lyons and H. Oberhauser, *An optimal polynomial approximation of Brownian motion*, SIAM Journal on Numerical Analysis, vol. 58, no. 3, pp. 1393–1421, 2020.
- J. Foster, T. Lyons and H. Oberhauser, *The shifted ODE method for underdamped Langevin MCMC*, https://arxiv.org/abs/2101.03446, 2021.

References IV

- A. S. Dalalyan, *Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent*, Proceedings of the 2017 Conference on Learning Theory, 2017.
- A. S. Dalalyan and L. Riou-Durand, *On sampling from a log-concave density using kinetic Langevin diffusions*, Bernoulli, vol. 26, no. 3, pp 1956-1988, 2020.
- Z. Song and Z. Tan. *Hamiltonian Assisted Metropolis Sampling*, https://arxiv.org/pdf/2005.08159, 2020.
- P. Monmarché. *High-dimensional MCMC with a standard splitting scheme for the underdamped Langevin diffusion*, https://arxiv.org/pdf/2007.05455, 2020.

References V

R. Shen and Y. T. Lee. *The Randomized Midpoint Method for Log-Concave Sampling*, Advances in Neural Information Processing Systems, 2019.

Y. He, K. Balasubramanian and M. A. Erdogdu. *On the Ergodicity, Bias and Asymptotic Normality of Randomized Midpoint Sampling Method*, Advances in Neural Information Processing Systems, 2020.

M. Lichman. *UCI machine learning repository*, https://archive.ics.uci.edu/ml, 2013.