Convergence Rates for Discretized (Stochastic Gradient) Kinetic Langevin Dynamics

Peter Whalley[1]
Collaborators: B. Leimkuhler[1] and D. Paulin[1]

[1]University of Edinburgh

Monte Carlo Methods and Applications, Sorbonne University, June 2023

Outline

- Introduce kinetic Langevin dynamics, some discretisations and their convergence guarantees;
- A method to prove convergence with weak stepsize assumptions in the strongly log-concave setting;
- The effect of using a stochastic gradient approximation on convergence;
- An application to Bayesian Logistic regression;

Underdamped Langevin Dynamics

A popular MCMC method is based on the underdamped Langevin dynamics SDE:

$$dV_t = -\nabla U(X_t)dt - \gamma V_t dt + \sqrt{2\gamma}dW_t$$
$$dX_t = V_t dt,$$

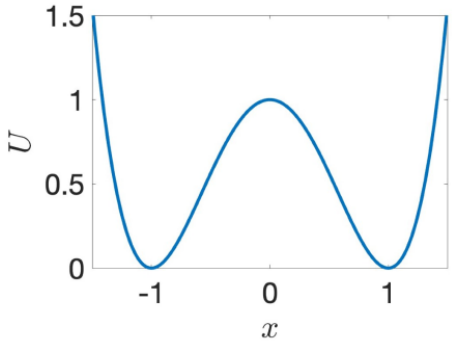
where γ is a friction parameter. This has been studied by physicists ([Einstein, 1905]) and mathematicians and has invariant measure $\pi \propto \exp\left(-U(x) - \frac{1}{2}\|v\|^2\right)$.

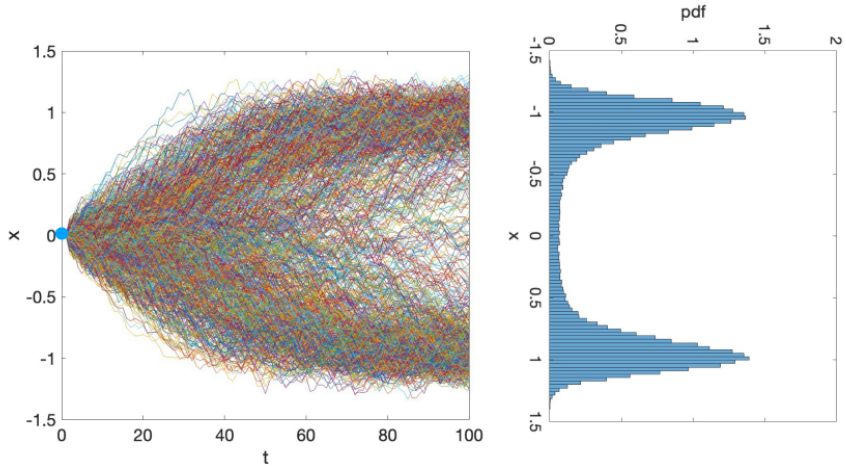
Underdamped Langevin Dynamics

A popular MCMC method is based on the underdamped Langevin dynamics SDE:

$$dV_t = -\nabla U(X_t)dt - \gamma V_t dt + \sqrt{2\gamma}dW_t$$
$$dX_t = V_t dt,$$

where γ is a friction parameter.
In practice this SDE is discretised and the individual timesteps generated by integration are viewed as approximate draws from the target distribution.





Popular Discretisations

The second order dynamics have additional complexity compared to the overdamped dynamics and there are many more possible discretisations. These include

- Euler-Maruyama (EM);
- BAOAB, OBABO, OABAO splitting methods [Bussi and Parrinello, 2007, Leimkuhler and Matthews, 2013];
- Stochastic Euler Scheme (SES) [Ermak and Buckholz, 1980];
- Stochastic Position Verlet (SPV), Stochastic Velocity Verlet (SVV) [Melchionna, 2007];
- BBK scheme [Brünger et al., 1984];
- ...

A metric of error: Non-asymptotic guarantees

A metric of error that they use is

$$\mathcal{W}_2(\mu_0 P_h^n, \mu) \leq \underbrace{\mathcal{W}_2(\mu_0 P_h^n, \mu_h)}_{\text{Convergence Rate}} + \underbrace{\mathcal{W}_2(\mu_h, \mu)}_{\text{Bias}}$$

for some initial measure μ_0 and target measure μ and P_h^n is the transition kernel of the discretisation with step-size h and invariant measure μ_h.

The aim is to optimally tune parameters to minimise the number of steps for

$$\mathcal{W}_2(\mu_0 P_h^n, \mu) < \epsilon$$

for some error ϵ.

A metric of error: Non-asymptotic guarantees

A metric of error that they use is

$$\mathcal{W}_2(\mu_0 P_h^n, \mu) \leq \underbrace{\mathcal{W}_2(\mu_0 P_h^n, \mu_h)}_{\text{Convergence Rate}} + \underbrace{\mathcal{W}_2(\mu_h, \mu)}_{\text{Bias}}$$

for some initial measure μ_0 and target measure μ and P_h^n is the transition kernel of the discretisation with step-size h and invariant measure μ_h.

The aim is to optimally tune parameters to minimise the number of steps for

$$\mathcal{W}_2(\mu_0 P_h^n, \mu) < \epsilon$$

for some error ϵ.

This talk: Convergence rate to the invariant measure for many different discretisations, trying to get results that hold for a large range of stepsizes.

Splitting Methods

One can split up the dynamics into parts which can be integrated exactly, see [Bussi and Parrinello, 2007].

$$dV_t = -\nabla U(X_t)dt - \gamma V_t dt + \sqrt{2\gamma}dW_t$$
$$dX_t = V_t dt,$$

Then you can integrate each part exactly

$$\mathcal{B} : v \to v - h\nabla U(x),$$
$$\mathcal{A} : x \to x + hv,$$
$$\mathcal{O} : v \to \eta v + \sqrt{1-\eta^2}\xi,$$

where $\eta := \exp(-\gamma h)$. For example the second order method $\mathcal{BAOAB}$.

Splitting Methods

One can split up the dynamics into parts which can be integrated exactly, see [Bussi and Parrinello, 2007].

$$dV_t = -\nabla U(X_t)dt - \gamma V_t dt + \sqrt{2\gamma}dW_t$$
$$dX_t = V_t dt,$$

Then you can integrate each part exactly

$$\mathcal{B} : v \to v - h\nabla U(x),$$
$$\mathcal{A} : x \to x + hv,$$
$$\mathcal{O} : v \to \eta v + \sqrt{1-\eta^2}\xi,$$

where $\eta := \exp(-\gamma h)$. For example the second order method $\mathcal{BAOAB}$.

Remark

One can create a kinetic Langevin integrator by considering a Hamiltonian integrator between two $\mathcal{O}$ steps. For example the randomised midpoint integrator of [Bou-Rabee and Marsden, 2022], we will refer to this integrator as rOABAO.

Stochastic Euler Scheme

A popular method in machine learning literature (the Stochastic Euler Scheme) is based on fixing the force over an interval and integrate $\mathcal{A} + \mathcal{B} + \mathcal{O}$ exactly.

$$X_{k+1} = X_k + \frac{1-\eta}{\gamma} V_k - \frac{\gamma h + \eta - 1}{\gamma^2} \nabla U(X_k) + \zeta_{k+1},$$
$$V_{k+1} = \eta V_k - \frac{1-\eta}{\gamma} \nabla U(X_k) + \omega_{k+1},$$

[Cheng et al., 2018, Dalalyan and Riou-Durand, 2020, Sanz-Serna and Zygalakis, 2021]

Assumptions

We assume that the target measure takes the form

$$\mu(dx) \propto \exp\left(-U(x)\right)dx,$$

for a potential U.

Assumption (M-∇Lipschitz)

There exists a $M > 0$ *such that for all* $x, y \in \mathbb{R}^d$

$$\left|\nabla U\left(x\right) - \nabla U\left(y\right)\right| \leq M\left|x - y\right|.$$

Assumption (m-convexity)

There exists a $m > 0$ *such that for all* $x, y \in \mathbb{R}^d$

$$\left\langle \nabla U(x) - \nabla U(y), x - y \right\rangle \geq m\left|x - y\right|^2.$$

There are results in the non-convex setting see for example [Eberle et al., 2019].

Twisted norm and Wasserstein Distance

- One cannot get Wasserstein convergence with respect to the standard Euclidean norm, but one can get Wasserstein convergence with respect to a "twisted Euclidean norm". [1]

$$|| (x, v) ||^2_{a,b} = ||x||^2 + 2b\langle x, v\rangle + a||v||^2,$$

for $a, b > 0$ such that $b^2 < a$.

[1] [Cheng et al., 2018, Dalalyan and Riou-Durand, 2020, Monmarché, 2021, Gouraud et al., 2022, Sanz-Serna and Zygalakis, 2021]

Twisted norm and Wasserstein Distance

- One cannot get Wasserstein convergence with respect to the standard Euclidean norm, but one can get Wasserstein convergence with respect to a "twisted Euclidean norm". [1]

$$||(x,v)||_{a,b}^2 = ||x||^2 + 2b\langle x, v\rangle + a||v||^2,$$

for $a, b > 0$ such that $b^2 < a$.

We define the p-Wasserstein between two probability measures μ and ν with respect to the norm $||\cdot||_{a,b}$ to be

$$\mathcal{W}_{p,a,b}(\nu,\mu) = \left(\inf_{\xi\in\Gamma(\nu,\mu)} \int_{\mathbb{R}^{2d}} ||z_1 - z_2||_{a,b}^p d\xi(z_1, z_2)\right)^{1/p},$$

where $\Gamma(\mu,\nu)$ is the set of measures with marginals μ and ν (the set of all couplings between μ and ν).

[1] [Cheng et al., 2018, Dalalyan and Riou-Durand, 2020, Monmarché, 2021, Gouraud et al., 2022, Sanz-Serna and Zygalakis, 2021]

Wasserstein Convergence

Let $z_k = (x_k, v_k)$ and $\tilde{z}_k = (\tilde{x}_k, \tilde{v}_k)$ be two **synchronously** coupled trajectories of a numerical scheme for kinetic Langevin. Then if they have the contraction property

$$||\tilde{z}_{k+1} - z_{k+1}||_{a,b}^2 \leq (1 - c(h)) ||\tilde{z}_k - z_k||_{a,b}^2$$

for $a, b > 0$ such that $b^2 > a$. Then we have that

$$\mathcal{W}_p^2 (\nu P_h^n, \mu P_h^n) \leq C (1 - c(h))^n \mathcal{W}_p^2 (\nu, \mu).$$

for all $1 \leq p \leq \infty$.
Our aim is to find a and b to provide explicit assumptions on the stepsize h and friction parameter γ.

Convergence rates?

Let $\overline{z}_j = \tilde{z}_j - z_j$ for $j \in \mathbb{N}$, then

$$||\tilde{z}_{k+1} - z_{k+1}||_{a,b}^2 \leq (1 - c(h)) ||\tilde{z}_k - z_k||_{a,b}^2, \tag{1}$$

is equivalent to showing that

$$\overline{z}_k^T \left((1 - c(h)) N - P^T N P \right) \overline{z}_k \geq 0, \quad \text{where} \quad N = \begin{pmatrix} 1 & b \\ b & a \end{pmatrix},$$

and $\overline{z}_{k+1} = P\overline{z}_k$.

Convergence rates?

Let $\overline{z}_j = \tilde{z}_j - z_j$ for $j \in \mathbb{N}$, then

$$||\tilde{z}_{k+1} - z_{k+1}||_{a,b}^2 \leq (1 - c(h)) ||\tilde{z}_k - z_k||_{a,b}^2, \tag{1}$$

is equivalent to showing that

$$\overline{z}_k^T \left((1 - c(h)) N - P^T N P \right) \overline{z}_k \geq 0, \quad \text{where} \quad N = \begin{pmatrix} 1 & b \\ b & a \end{pmatrix},$$

and $\overline{z}_{k+1} = P \overline{z}_k$.

Proving contraction is equivalent to showing that the matrix $\mathcal{H} := (1 - c(h)) N - P^T N P \succ 0$ is positive definite.

Example

As an example we have for the Euler-Maruyama scheme the update rule for $\overline{z}_k$

$$\overline{x}_{k+1} = \overline{x}_k + h\overline{v}_k, \qquad \overline{v}_{k+1} = \overline{v}_k - \gamma h\overline{v}_k - hQ\overline{x}_k,$$

where by mean value theorem we can define $Q = \int_{t=0}^{1} \nabla^2 U(\tilde{x}_k + t(x_k - \tilde{x}_k))dt$, then $\nabla U(\tilde{x}_k) - \nabla U(x_k) = Q\overline{x}$. One can show that

$$P = \begin{pmatrix} I & hI \\ -hQ & (1-\gamma h)\, I \end{pmatrix}.$$

Convergence rates?

The matrix $\mathcal{H} := (1 - c(h))\, N - P^T NP \succ 0$ is symmetric and hence of the form

$$\mathcal{H} = \begin{pmatrix} A & B \\ B & C \end{pmatrix}, \tag{2}$$

we can show that $\mathcal{H}$ is positive definite.

Proposition

Let $\mathcal{H}$ be a symmetric matrix of the form (2), then $\mathcal{H}$ is positive definite if and only if $A \succ 0$ and $C - BA^{-1}B \succ 0$. Further if A, B and C commute then $\mathcal{H}$ is positive definite if and only if $A \succ 0$ and $AC - B^2 \succ 0$.

Results

If $\gamma^2 \geq \mathcal{O}(M)$, for the choice of $a = \frac{1}{M}$, $\eta = \exp\{-\gamma h\}$, we have

Algorithm	b	$c(h)$	step-size restriction
EM	$1/\gamma$	$\mathcal{O}(mh/\gamma)$	$\mathcal{O}(1/\gamma)$
BBK	$h/2 + 1/\gamma$	$\mathcal{O}(mh/\gamma)$	$\mathcal{O}(1/\gamma)$
SPV and SVV	$h/(1-\eta)$	$\mathcal{O}(mh/\gamma)$	$\mathcal{O}(1/\gamma)$
BAOAB	$h/(1-\eta)$	$\mathcal{O}(mh^2/(1-\eta))$	$\mathcal{O}(1/\sqrt{M})$
OBABO	$h/(1-\eta)$	$\mathcal{O}(mh^2/(1-\eta))$	$\mathcal{O}(1/\sqrt{M})$
rOABAO	$h/(1-\eta)$	$\mathcal{O}(mh^2/(1-\eta))$	$\mathcal{O}(1/\sqrt{M})$
SES/EB	$1/\gamma$	$\mathcal{O}(mh/\gamma)$	$\mathcal{O}(1/\gamma)$

Remark

The convergence rate of the continuous dynamics on this class of functions is known to be $\mathcal{O}(m/\gamma)$.

High friction limit

If you take the limit as $\gamma \to \infty$ for BAOAB we obtain

$$x_{k+1} = x_k - \frac{h^2}{2}\nabla U(x_k) + \frac{h}{2}\left(\xi_k + \xi_{k+1}\right),$$

which is simply the [Leimkuhler and Matthews, 2013] (LM) scheme with stepsize $h^2/2$ and $\lim_{\gamma\to\infty} c\left(h\right) = \frac{h^2 m}{4}$.

High friction limit

If you take the limit as $\gamma \to \infty$ for BAOAB we obtain

$$x_{k+1} = x_k - \frac{h^2}{2}\nabla U(x_k) + \frac{h}{2}\left(\xi_k + \xi_{k+1}\right),$$

which is simply the [Leimkuhler and Matthews, 2013] (LM) scheme with stepsize $h^2/2$ and $\lim_{\gamma\to\infty} c\left(h\right) = \frac{h^2 m}{4}$. Now we take the limit as $\gamma \to \infty$ for OBABO we obtain

$$x_{k+1} = x_k - \frac{h^2}{2}\nabla U(x_k) + h\xi_{k+1},$$

which is the Euler-Maruyama scheme for overdamped Langevin with stepsize $h^2/2$, which has convergence rate $\mathcal{O}\left(h^2 m\right)$.

High friction limit

If you take the limit as $\gamma \to \infty$ for BAOAB we obtain

$$x_{k+1} = x_k - \frac{h^2}{2}\nabla U(x_k) + \frac{h}{2}\left(\xi_k + \xi_{k+1}\right),$$

which is simply the [Leimkuhler and Matthews, 2013] (LM) scheme with stepsize $h^2/2$ and $\lim_{\gamma\to\infty} c\left(h\right) = \frac{h^2 m}{4}$. Now we take the limit as $\gamma \to \infty$ for OBABO we obtain

$$x_{k+1} = x_k - \frac{h^2}{2}\nabla U(x_k) + h\xi_{k+1},$$

which is the Euler-Maruyama scheme for overdamped Langevin with stepsize $h^2/2$, which has convergence rate $\mathcal{O}\left(h^2 m\right)$.

- Euler-Maruyama for Kinetic Langevin Dynamics (no well-defined limit).
- Stochastic Euler Scheme: we obtain the update rule $x_{k+1} = x_k$ in the limit.

Convergence Rates Plots

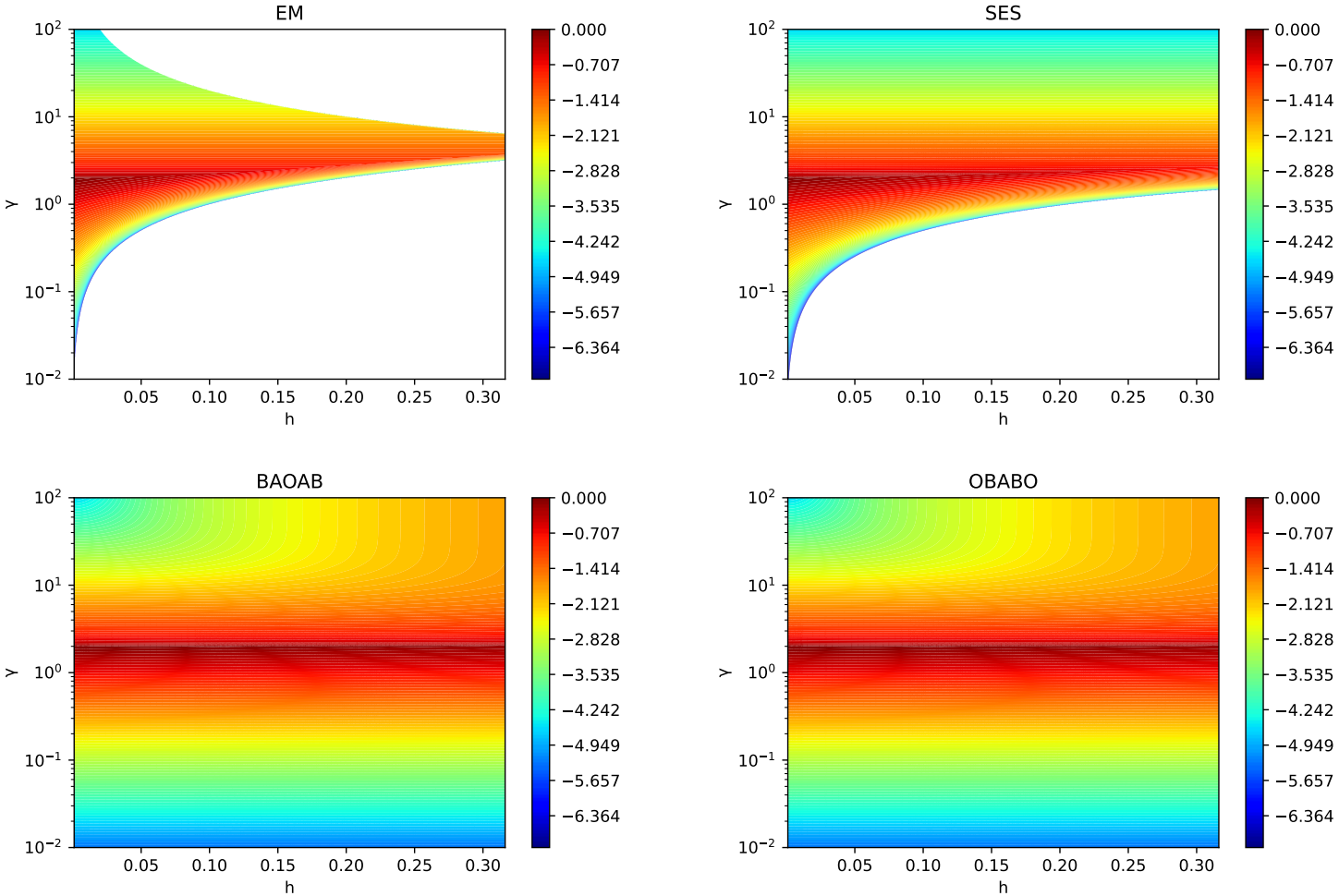


Figure: Contour plots of $\ln\left(\frac{1-c(h)}{h}\right)$ for an anisotropic Gaussian with $M = 10$ and $m = 1$.

Connection with HMC

[Gouraud et al., 2022] can explain the behaviour of OBABO and BAOAB through its relation to another sampling method Hamiltonian Monte Carlo. OBABO can be written as the velocity verlet integrator

$$v \to v - \frac{h}{2}\nabla U(x)$$
$$x \to x + hv$$
$$v \to v - \frac{h}{2}\nabla U(x)$$

with auto-regressive velocity refreshments given by

$$v \mapsto \eta v + \sqrt{1-\eta^2}\xi,$$

where $\xi \sim \mathcal{N}(0,1)$. This is precisely HMC with partial velocity refreshments.

Tightness of restrictions

- For this method of proof we require $\gamma^2 \geq \mathcal{O}(M)$ [Monmarché, 2020].
- Stability threshold for Euler-Maruyama, BAOAB, OBABO are the same as the step-size restriction for Gaussian targets.
- For other schemes the stability threshold for Gaussian target is not the same as the step-size restriction, but we do not expect the schemes to have reasonable bias outside this regime.

Stochastic gradients

Definition

A *stochastic gradient approximation* of a potential U is defined by a function $\mathcal{G} : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$ and a probability distribution ρ on a Polish space Ω, satisfying that $\mathcal{G}$ is measurable on $(\Omega, \mathcal{F})$, and that for every $x \in \mathbb{R}^n$, for $W \sim \rho$,

$$\mathbb{E}(\mathcal{G}(x, W)) = \nabla U(x).$$

Assumption (Variance of Jacobian)

We assume that the Jacobian of the stochastic gradient $\mathcal{G}$, $D_x\mathcal{G}(x, W)$ exists and it is measurable on $(\Omega, \mathcal{F})$. We also assume there exists $C_G > 0$ such that for $W \sim \rho$,

$$\sup_{x \in \mathbb{R}^n} \mathbb{E}\|D_x\mathcal{G}(x, W) - \nabla^2 U(x)\|^2 \leq C_G.$$

Convergence with Stochastic gradients

Let $\bar{z}_k = \tilde{z}_k - z_k$, if we synchronously couple the stochastic gradients we are able to get the expected contraction result ($Q = \mathbb{E}(\tilde{Q})$ and $\tilde{Q}$ is defined by MVT for $\mathcal{G}$):

$$\mathbb{E}\|z_{k+1}\|_{a,b}^2 \leq (1 - c_{\text{der}}(h)) \|z_k\|_{a,b}^2 + z_k^T \begin{pmatrix} a_2\mathbb{E}(\tilde{Q} - Q)^2 & b_2\mathbb{E}(\tilde{Q} - Q)^2 \\ b_2\mathbb{E}(\tilde{Q} - Q)^2 & c_2\mathbb{E}(\tilde{Q} - Q)^2 \end{pmatrix} z_k$$

Algorithm	c(h)
EM	$\mathcal{O}\left(mh/\gamma - h^2 C_G/M\right)$
BBK	$\mathcal{O}\left(mh/\gamma - h^2 C_G/M\right)$
SPV	$\mathcal{O}\left(mh/\gamma - h^2 C_G/M\right)$
SVV	$\mathcal{O}\left(mh/\gamma - h^2 C_G/M\right)$
BAOAB	$\mathcal{O}\left(h^2 m/(1-\eta) - h^2 C_G\left(\eta/M + h^2\right)\right)$
OBABO	$\mathcal{O}\left(h^2 m/(1-\eta) - h^2 C_G/M\right)$
rOABAO	$\mathcal{O}\left(h^2 m/(1-\eta) - h^2 C_G\left(\eta/M + h^2\right)\right)$
SES/EB	$\mathcal{O}\left(mh/\gamma - h^2 C_G/M\right)$

Table: Contraction rates $c(h)$ in stochastic gradient setting

Simulation example - MNIST classification

- MNIST data set [LeCun et al., 2010] has $60,000$ training data points and $10,000$ test data points.
- The images are of size 28 by 28 pixels and hence can be represented in $\mathbb{R}^{784}$.
- However, we will consider the problem of classification between the 3 and the 5 digits by Bayesian logistic regression.

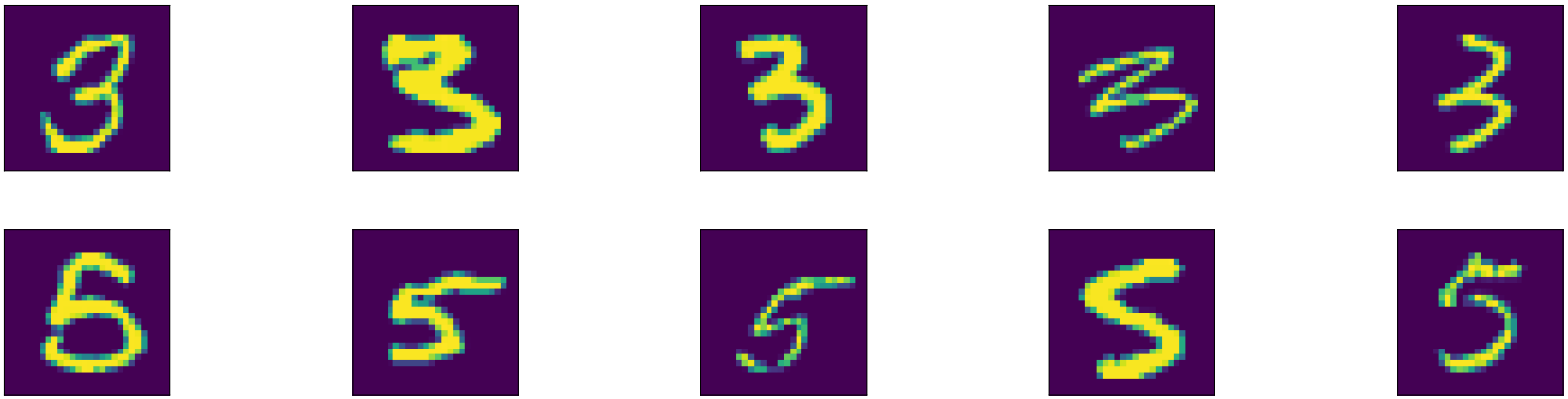


Figure: MNIST 3 and 5 digits.

Simulation example - MNIST classification

We use a i.i.d. Gaussian prior p_0 with mean 0 and variance $\sigma^2 = 0.001$. A more accurate estimator is the *variance reduced stochastic gradient* ([Johnson and Zhang, 2013]), also called control variate method in the context of MCMC (see [Quiroz et al., 2018], [Baker et al., 2019]).

- We compare our discretization schemes on this MNIST example.
- Both bias and effective sample sizes are evaluated, test function is chosen as the potential U.
- Ground truth is established via HMC with 40 million gradient evaluations, each method is evaluated using 8 million steps (80 runs with 100000 steps each).

Bias, $\gamma = \sqrt{M}$

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{M}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{M}$
EM	$4.2(\pm 0.089)$	$1.5(\pm 0.13)$	$0.79(\pm 0.18)$	$0.28(\pm 0.23)$
BBK	$2.7(\pm 0.061)$	$0.67(\pm 0.099)$	$0.016(\pm 0.14)$	$-0.18(\pm 0.2)$
SPV	$123(\pm 0.079)$	$32.1(\pm 0.091)$	$8.19(\pm 0.13)$	$2.07(\pm 0.18)$
SVV	$126(\pm 0.097)$	$32.8(\pm 0.091)$	$8.17(\pm 0.13)$	$2.03(\pm 0.17)$
BAOAB	$-0.043(\pm 0.049)$	$-0.002(\pm 0.058)$	$0.13(\pm 0.086)$	$-0.055(\pm 0.12)$
BAOAB VRSG	$0.47(\pm 0.043)$	$0.23(\pm 0.066)$	$0.035(\pm 0.087)$	$0.036(\pm 0.12)$
OBABO	$2.7(\pm 0.056)$	$0.67(\pm 0.076)$	$0.22(\pm 0.13)$	$0.17(\pm 0.19)$
rOABAO	$-2.6(\pm 0.062)$	$-0.61(\pm 0.094)$	$0.025(\pm 0.13)$	$-0.16(\pm 0.19)$
SES/EB	$2.6(\pm 0.072)$	$1.2(\pm 0.094)$	$0.71(\pm 0.11)$	$0.2(\pm 0.18)$

Table: Bias for potential function, $\gamma = \sqrt{M}$

Gradient evaluations/effective sample, $\gamma = \sqrt{M}$

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{M}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{M}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{M}$
EM	146($\pm$0.7)	221($\pm$0.998)	282($\pm$0.822)	327($\pm$0.581)
BBK	85($\pm$0.535)	148($\pm$0.726)	221($\pm$0.969)	285($\pm$0.933)
SPV	86.7($\pm$0.554)	148($\pm$0.775)	221($\pm$0.887)	284($\pm$0.992)
SVV	86.5($\pm$0.645)	147($\pm$0.801)	222($\pm$0.916)	283($\pm$0.825)
BAOAB	44.3($\pm$0.304)	88.7($\pm$0.585)	152($\pm$0.812)	228($\pm$0.822)
BAOAB VRSG	44.6($\pm$0.332)	86.8($\pm$0.578)	152($\pm$0.915)	226($\pm$0.934)
OBABO	68.6($\pm$0.491)	140($\pm$0.84)	218($\pm$0.942)	282($\pm$0.809)
rOABAO	68.5($\pm$0.507)	140($\pm$0.692)	219($\pm$0.781)	283($\pm$0.862)
SES/EB	87.4($\pm$0.593)	149($\pm$0.663)	220($\pm$0.831)	284($\pm$0.809)

Table: Gradient evaluations / ESS (potential function), $\gamma = \sqrt{M}$

Bias, $\gamma = \sqrt{m}$

- $\gamma = \mathcal{O}(\sqrt{m})$ is the best choice of friction in the continuous setting [Cao et al., 2019].

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{m}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{m}$
EM	$6.4 \cdot 10^4(\pm 0.82)$	$1.5 \cdot 10^4(\pm 0.72)$	$1.1 \cdot 10^3(\pm 0.73)$	$4.9(\pm 0.11)$
BBK	$2.8(\pm 0.034)$	$0.68(\pm 0.041)$	$0.1(\pm 0.05)$	$0.0038(\pm 0.066)$
SPV	$0.72(\pm 0.036)$	$0.14(\pm 0.043)$	$0.06(\pm 0.054)$	$-0.014(\pm 0.073)$
SVV	$3.5(\pm 0.036)$	$0.81(\pm 0.043)$	$0.26(\pm 0.061)$	$0.05(\pm 0.089)$
BAOAB	$0.03(\pm 0.038)$	$-0.011(\pm 0.049)$	$-0.046(\pm 0.062)$	$0.043(\pm 0.074)$
BAOAB VRSG	$6.4(\pm 0.04)$	$2.4(\pm 0.051)$	$1.1(\pm 0.063)$	$0.55(\pm 0.075)$
OBABO	$2.7(\pm 0.032)$	$0.65(\pm 0.041)$	$0.22(\pm 0.052)$	$0.11(\pm 0.071)$
rOABAO	$-1.7(\pm 0.041)$	$-0.55(\pm 0.041)$	$-0.2(\pm 0.054)$	$-0.033(\pm 0.081)$
SES/EB	$6.0 \cdot 10^4(\pm 0.61)$	$1.5 \cdot 10^4(\pm 0.48)$	$1.1 \cdot 10^3(\pm 0.59)$	$4.7(\pm 0.068)$

Table: Bias for potential function, $\gamma = \sqrt{m}$

Gradient evaluations/effective sample, $\gamma = \sqrt{m}$

Algorithm	$h = 2/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/\sqrt{M}$ $\gamma = \sqrt{m}$	$h = 1/(2\sqrt{M})$ $\gamma = \sqrt{m}$	$h = 1/(4\sqrt{M})$ $\gamma = \sqrt{m}$
EM	N.A.	N.A.	N.A.	189($\pm$0.955)
BBK	15($\pm$0.124)	30.1($\pm$0.233)	57.5($\pm$0.352)	108($\pm$0.717)
SPV	15.1($\pm$0.106)	29.7($\pm$0.209)	57.4($\pm$0.408)	109($\pm$0.725)
SVV	15($\pm$0.121)	29.9($\pm$0.222)	57.5($\pm$0.341)	108($\pm$0.628)
BAOAB	18.8($\pm$0.128)	36.4($\pm$0.288)	66.4($\pm$0.461)	116($\pm$0.849)
BAOAB VRSG	19.7($\pm$0.169)	36.4($\pm$0.242)	67.8($\pm$0.447)	114($\pm$0.662)
OBABO	15($\pm$0.118)	30($\pm$0.204)	57.5($\pm$0.471)	108($\pm$0.711)
rOABAO	16.5($\pm$0.236)	29.7($\pm$0.218)	58.2($\pm$0.356)	109($\pm$0.669)
SES/EB	N.A.	N.A.	N.A.	108($\pm$0.652)

Table: Gradient evaluations / ESS (potential function), $\gamma = \sqrt{m}$. N.A. indicates that the method did not converge for the given stepsize.

Future Work/Open questions

- Wasserstein bias estimates for BAOAB and other schemes.
- Can we get similar step-size restrictions with a more sophisticated metric and coupling to deal with the non-convex case?

References I


Baker, J., Fearnhead, P., Fox, E. B., and Nemeth, C. (2019).
Control variates for stochastic gradient mcmc.
Statistics and Computing, 29:599–615.


Bou-Rabee, N. and Marsden, M. (2022).
Unadjusted hamiltonian mcmc with stratified monte carlo time integration.
arXiv preprint arXiv:2211.11003.


Brünger, A., Brooks III, C. L., and Karplus, M. (1984).
Stochastic boundary conditions for molecular dynamics simulations of st2 water.
Chemical physics letters, 105(5):495–500.


Bussi, G. and Parrinello, M. (2007).
Accurate sampling using langevin dynamics.
Phys. Rev. E, 75:056707.

References II


Cao, Y., Lu, J., and Wang, L. (2019).
On explicit l^2-convergence rate estimate for underdamped langevin dynamics.
arXiv preprint arXiv:1908.04746.


Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. (2018).
Underdamped langevin mcmc: A non-asymptotic analysis.
In *Conference on learning theory*, pages 300–323. PMLR.


Dalalyan, A. S. and Riou-Durand, L. (2020).
On sampling from a log-concave density using kinetic langevin diffusions.
Bernoulli, 26(3):1956–1988.


Eberle, A., Guillin, A., and Zimmer, R. (2019).
Couplings and quantitative contraction rates for langevin dynamics.
The Annals of Probability, 47(4):1982–2010.

References III


Einstein, A. (1905).
Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen.
Annalen der physik, 4.


Ermak, D. L. and Buckholz, H. (1980).
Numerical integration of the langevin equation: Monte carlo simulation.
Journal of Computational Physics, 35(2):169–182.


Gouraud, N., Le Bris, P., Majka, A., and Monmarché, P. (2022).
Hmc and underdamped langevin united in the unadjusted convex smooth case.
arXiv e-prints, pages arXiv–2202.

References IV


Johnson, R. and Zhang, T. (2013).
Accelerating stochastic gradient descent using predictive variance reduction.
Advances in neural information processing systems, 26.


LeCun, Y., Cortes, C., Burges, C., et al. (2010).
Mnist handwritten digit database.


Leimkuhler, B. and Matthews, C. (2013).
Rational construction of stochastic numerical methods for molecular sampling.
Applied Mathematics Research eXpress, 2013(1):34–56.


Melchionna, S. (2007).
Design of quasisymplectic propagators for langevin dynamics.
The Journal of chemical physics, 127(4):044108.

References V


Monmarché, P. (2020).
Almost sure contraction for diffusions on rd. application to generalised langevin diffusions.
arXiv preprint arXiv:2009.10828.


Monmarché, P. (2021).
High-dimensional mcmc with a standard splitting scheme for the underdamped langevin diffusion.
Electronic Journal of Statistics, 15(2):4117–4166.


Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2018).
Speeding up mcmc by efficient data subsampling.
Journal of the American Statistical Association.

References VI


Sanz-Serna, J. M. and Zygalakis, K. C. (2021).
Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations.
J. Mach. Learn. Res., 22:242–1.