Discretized Langevin algorithms for non-strongly log-concave targets

Arnak S. Dalalyan
CREST/ ENSAE Paris / IP Paris
Paul Langevin and Albert Einstein 1923 (from wikimedia)
1. Introduction
Problem: Given a probability density function $\pi : \mathbb{R}^p \to \mathbb{R}$, generate a random vector $X$ such that

$$X \sim \pi,$$

that is $P(X \in A) = \int_A \pi(x) \, dx$. 

Sampling from a density

Dalalyan, A.S.
Warm-up: rejection sampling 1/2

- Let $\nu : \mathbb{R}^p \rightarrow \mathbb{R}$ be an auxiliary, easily samplable, density.
- Assume for a known $M > 0$, we have $\pi(x) \leq M \nu(x), \forall x$.

Rejection method

Step 1 sample independently $Y \sim \nu$ and $U \sim \text{Unif}([0, M])$

Step 2 if $U \leq \pi(Y)/\nu(Y)$, then set $X = Y$,
else reject $Y$ and return to Step 1.
Warm-up: rejection sampling 1/2

Let $\nu : \mathbb{R}^p \rightarrow \mathbb{R}$ be an auxiliary, easily samplable, density.

Assume for a known $M > 0$, we have $\pi(x) \leq M \nu(x)$, $\forall x$.

Rejection method

Step 1 sample independently $Y \sim \nu$ and $U \sim \text{Unif}([0, M])$

Step 2 if $U \leq \frac{\pi(Y)}{\nu(Y)}$, then set $X = Y$,
else reject $Y$ and return to Step 1.

Let $K$ be the number of rounds required to sample $X$.

the random variable $K \sim \text{Geom}(p)$

with $p = \mathbb{P}(U \leq \frac{\pi(Y)}{\nu(Y)}) = 1/M$

the average number of rounds: $\mathbb{E}[K] = 1/p = M$. 
Warm-up: rejection sampling 2/2
Uniform distribution on a compact set

**Drawback of rejection sampling**: in most cases $M$ grows exponentially fast in dimension $p$.

- Consider the particular case $\pi(x) \propto 1(x \in C)$ with $C \subset [0, 1]^p$ compact.
- We do not know the volume $V_C$ of the set $C$ but we know that $C$ contains a ball of radius $r > 0$.
- We naturally choose $\nu(x) = 1(x \in [0, 1]^p)$.
- Then the almost only possible choice for $M$ is $M = 1/V\text{ol}(B_r^p)$. 
Drawback of rejection sampling: in most cases $M$ grows exponentially fast in dimension $p$.

- Consider the particular case $\pi(x) \propto \mathbb{1}(x \in C)$ with $C \subset [0, 1]^p$ compact.
- We do not know the volume $V_C$ of the set $C$ but we know that $C$ contains a ball of radius $r > 0$.
- We naturally choose $\nu(x) = \mathbb{1}(x \in [0, 1]^p)$.
- Then the almost only possible choice for $M$ is $M = 1/\text{Vol}(B_r^p)$.

Most Markov Chain Monte Carlo algorithms suffer from the same drawback.
Precise setting
Sampling from a log-concave density

We define the (log-posterior) function

\[ f(\theta) = -\log \pi(\theta). \]

and assume that it satisfies the smoothness and the strong convexity assumptions: there exist \( m > 0 \) and \( M < \infty \) such that

\[
\begin{align*}
    f(\theta) - f(\bar{\theta}) - \nabla f(\bar{\theta})^\top (\theta - \bar{\theta}) & \geq \frac{m}{2} \|\theta - \bar{\theta}\|_2^2, \quad (C1) \\
    \|\nabla f(\theta) - \nabla f(\bar{\theta})\|_2 & \leq M \|\theta - \bar{\theta}\|_2, \quad (C2)
\end{align*}
\]

for all \( \theta, \bar{\theta} \in \mathbb{R}^p \).

**Goal:** find nonasymptotic guarantees for approximately sampling from \( \pi \). More precisely, for every \( \epsilon > 0 \) find a density \( \mu \) such that one can efficiently sample from \( \mu \) and

\[
\|\mu - \pi\|_{TV} \leq \epsilon \quad \text{or} \quad W_2(\mu, \pi) \leq \epsilon.
\]
Optimization versus integration
Guarantees for sampling I

IMA Journal of Numerical Analysis (2013) 33, 80–110
Advance Access publication on March 19, 2012

Nonasymptotic mixing of the MALA algorithm

N. Bou-Rabee* AND M. Hairer

Theorem

Under natural assumptions on the target distribution \( \pi(x) \propto e^{-f(x)} \) for \( h \) small enough and for \( x \in \mathbb{R}^p \) satisfying \( f(x) < E_0 \), there exist positive constants \( \rho \in (0, 1) \), \( C_1(E_0) \) and \( C_2 \) independent of \( h \) such that the bound

\[
\| P^k(x, \cdot) - \pi \|_{TV} \leq C_1(E_0)(\rho^k + e^{-C_2/h^{1/4}})
\]

holds for all \( k \). Here \( P^k \) is the transition probability of a \( k \)-step MCMC.
Under natural assumptions on the target distribution $\pi(x) \propto e^{-f(x)}$ for $h$ small enough and for $x \in \mathbb{R}^p$ satisfying $f(x) < E_0$, there exist positive constants $\rho \in (0, 1)$, $C_1(E_0)$ and $C_2$ independent of $h$ such that the bound

$$\|P^k(x, \cdot) - \pi\|_{TV} \leq C_1(E_0) \left( \rho^k + e^{-C_2/h^{1/4}} \right)$$

holds for all $k$. Here $P^k$ is the transition probability of a $k$-step MCMC.
Theorem

Under natural assumptions on the target distribution \(\pi(x) \propto e^{-f(x)}\) for \(h\) small enough and for \(x \in \mathbb{R}^p\) satisfying \(f(x) < E_0\), there exist positive constants \(\rho \in (0, 1), C_1(E_0)\) and \(C_2\) independent of \(h\) such that the bound

\[
\|P^k(x, \cdot) - \pi\|_{TV} \leq C_1(E_0) \left( \rho^k + e^{-C_2/h^{1/4}} \right)
\]

holds for all \(k\). Here \(P^k\) is the transition probability of a \(k\)-step MCMC.
Assumption 2.1. The potential energy $U \in C^4(\mathbb{R}^n, \mathbb{R})$ satisfies the following.

A) One has $U(x) \geq 1$ and, for any $C > 0$ there exists an $E > 0$ such that

$$U(x) \geq C(1 + |x|^2),$$

for all $U(x) > E$.

B) There exist constants $c \in (0, \beta), d > 0$ and $E > 0$ such that

$$\Delta U(x) \leq c|\nabla U(x)|^2 - dU(x), \quad (2.4)$$

for all $x \in \mathbb{R}^n$ satisfying $U(x) > E$.

C) The Hessian of $U$ is bounded from below in the sense that there exists $C \geq 0$ such that

$$D^2U(x)(\eta, \eta) \geq -C|\eta|^2,$$

uniformly for all $x, \eta \in \mathbb{R}^n$.

D) There exists a constant $C > 0$ such that the first four derivatives of
Corollary 1.2 Let $f$ be a logconcave function in $\mathbb{R}^n$, given in the sense of (LS1), (LS2) and (LS3). Then for

$$m > 10^{31} p^4 \frac{n^3 R^2}{r^2} \ln^5 \frac{n R^2}{\varepsilon r d / \beta},$$

the total variation distance of $\sigma^m$ and $\pi_f$ is less than $\varepsilon$.

Our notation: $k > 10^{31} p^4 (M/m)^2 \log^5 (\square p / \varepsilon)$ implies that

$$\| P_k(x, \cdot) - \pi \|_{TV} \leq \varepsilon.$$
2. Sampling using the Langevin diffusion
Langevin based algorithms

To sample from $\pi \propto e^{-f}$, one can consider two versions of the Langevin Monte Carlo (LMC) algorithm.

**LMC** (aka ULA) Start from $\vartheta^{(0)} \in \mathbb{R}^p$ and use the update rule

$$
\vartheta^{(k+1)} = \vartheta^{(k)} - h \nabla f(\vartheta^{(k)}) + \sqrt{2h} \xi^{(k+1)};
$$

where $h > 0$ is the step-size, and $\xi^{(1)}, \ldots, \xi^{(k)}, \ldots$ are iid standard Gaussian and independent of $\vartheta^{(0)}$.

**MALA** (Metropolis adjusted Langevin algorithm) Start from $\bar{\vartheta}^{(0)} \in \mathbb{R}^p$ and use the update rule

$$
\begin{align*}
\bar{\vartheta}^{(k+1)} &= \bar{\vartheta}^{(k)} - h \nabla f(\bar{\vartheta}^{(k)}) + \sqrt{2h} \xi^{(k+1)}, \\
\bar{\vartheta}^{(k+1)} &= \begin{cases} 
\bar{y}^{(k+1)}, & \text{with prob. } \alpha_k, \\
\bar{\vartheta}^{(k)}, & \text{with prob. } 1 - \alpha_k
\end{cases}
\end{align*}
$$

for a properly chosen acceptance rate

$$
\alpha_k = \alpha(\bar{\vartheta}^{(k)}, \bar{y}^{(k+1)}).
$$

Dalalyan, A.S.
Background on the Langevin algorithm

Langevin diffusion

- $\mathcal{O}^{(k)}$ is the Euler discretisation of the Langevin diffusion $L_t$,
- the Langevin diffusion is defined by the SDE

$$dL_t = -\nabla f (L_t) \, dt + \sqrt{2} \, dW_t, \quad t \geq 0.$$ 

- Under (C1-C2), the SDE has a unique strong solution which is a Markov process. It is ergodic with stationary density $\pi \propto e^{-f}$.
- The transition kernel of this process is denoted by $P_L^t (x, \cdot)$, that is $P_L^t (x, A) = P (L_t \in A | L_0 = x)$.
- (C1-C2) yield the spectral gap property of the semigroup \{ $P_L^t : t \in \mathbb{R}_+$ \}. For any probability density $\nu$,

$$\| \nu P_L^t - \pi \|_{TV} \leq \frac{1}{2} D_{KL} (\nu \| \pi)^{1/2} e^{-tm/2}, \quad \forall t \geq 0.$$ 

Dalalyan, A.S.
Illustration of the link between Langevin diffusion and sampling

**Figure:** Illustration of Langevin dynamics. The blue lines represent different paths of a Langevin process. We see that the histogram of the state at time $t = 30$ is close to the target density (the dark blue line).
Background on the Langevin algorithm

Euler discretization

- the Langevin diffusion is defined by the SDE

$$dL_t = -\nabla f(L_t) \, dt + \sqrt{2} \, dW_t, \quad t \geq 0.$$ 

- $\vartheta^{(k)}$ is the Euler discretisation of the Langevin diffusion $L_t$:

$\vartheta^{(k)} \approx L_{kh}$.

- To be more precise, we introduce a diffusion-type continuous-time process $D$ obeying the following SDE:

$$dD_t = b_t(D) \, dt + \sqrt{2} \, dW_t, \quad t \geq 0,$$

with the drift $b_t(D) = -\nabla f(D_{kh})$ if $t \in [kh, (k + 1)h[$.

- For this process, we have

$$(\vartheta^{(1)}, \ldots, \vartheta^{(k)}) \overset{\mathcal{D}}{=} (D_h, \ldots, D_{kh}).$$
Optimization versus sampling

**Optimization**

- **Problem:** compute

\[ \theta^* \in \arg \min_{\theta \in \mathbb{R}^p} f(\theta). \]

**Sampling**

- **Problem:** Sample \( \vartheta \) from the pdf

\[ \pi(\theta) = \frac{1}{C} e^{-f(\theta)}, \quad C = \int_{\mathbb{R}^p} e^{-f} \]
Optimization versus sampling

**Optimization**

**Problem:** compute

\[ \theta^* \in \arg \min_{\theta \in \mathbb{R}^p} f(\theta). \]

**Method:** gradient descent

\[ \theta^{k+1} = \theta^k - h \nabla f(\theta^k). \]

**Sampling**

**Problem:** Sample \( \vartheta \) from the pdf

\[ \pi(\theta) = \frac{1}{C} e^{-f(\theta)}, \quad C = \int_{\mathbb{R}^p} e^{-f} \]

**Method:** Langevin Monte Carlo

\[ \vartheta^{k+1} = \vartheta^k - h \nabla f(\vartheta^k) + \sqrt{2h} \xi^k. \]

with \( \xi^k \) iid \( \mathcal{N}(0, I) \).
Optimization versus sampling

**Optimization**

- **Problem:** compute

\[ \theta^* \in \arg \min_{\theta \in \mathbb{R}^p} f(\theta). \]

- **Method:** gradient descent

\[ \theta^{k+1} = \theta^k - h \nabla f(\theta^k). \]

**Sampling**

- **Problem:** Sample \( \vartheta \) from the pdf

\[ \pi(\theta) = \frac{1}{C} e^{-f(\theta)}, \quad C = \int_{\mathbb{R}^p} e^{-f}. \]

- **Method:** Langevin Monte Carlo

\[ \vartheta^{k+1} = \vartheta^k - h \nabla f(\vartheta^k) + \sqrt{2h} \xi^k. \]

with \( \xi^k \) iid \( \mathcal{N}(0, I) \).

**What about theoretical guarantees?**
Optimization versus sampling
Theoretical guarantees

- We assume that for some $m, M > 0$
  \[
  \begin{align*}
  f(\theta) - f(\theta') - \nabla f(\theta')^\top (\theta - \theta') & \geq \frac{m}{2}\|\theta - \theta'\|_2^2, \\
  \|\nabla f(\theta) - \nabla f(\theta')\|_2 & \leq M\|\theta - \theta'\|_2,
  \end{align*}
  \forall \theta, \theta' \in \mathbb{R}^p,
  \]

- Theorem 0 (optim.): If $h \leq 2/(m + M)$, then
  \[
  \|\theta^K - \theta^*\|_2 \leq (1 - mh)^K \|\theta^0 - \theta^*\|_2.
  \]
Optimization versus sampling
Theoretical guarantees

- We assume that for some $m, M > 0$
\[
\begin{align*}
  f(\theta) - f(\theta') - \nabla f(\theta')^\top (\theta - \theta') & \geq (m/2)\|\theta - \theta'\|^2_2, \\
  \|\nabla f(\theta) - \nabla f(\theta')\|_2 & \leq M\|\theta - \theta'\|_2,
\end{align*}
\]
\hspace{5mm} $\forall \theta, \theta' \in \mathbb{R}^p$,

- Theorem 0 (optim.): If $h \leq 2/(m + M)$, then
\[
\|\theta^K - \theta^*\|_2 \leq (1 - mh)^K \|\theta^0 - \theta^*\|_2.
\]

- Theorem 1(sampling): If $h \leq 2/(m + M)$,
\[
W_2(\nu_K, \pi) \leq (1 - mh)^K W_2(\nu_0, \pi) + \frac{2M}{m} (hp)^{1/2}.
\]

(Durmus and Moulines, 2019; Dalalyan, 2017b)
\[ L_t = L_0 - \int_0^t \nabla f(L_s) \, ds + \sqrt{2} W_t \]
\[ L_t = L_0 - \int_0^t \nabla f(L_s) \, ds + \sqrt{2} W_t \]
\[ L_t - L_{kh} = - \int_{k}^{t} \nabla f(L_s) \, ds + \sqrt{2} (W_t - W_{kh}) \]

\[ D_t - D_{kh} = -(t - kh) \nabla f(D_{kh}) + \sqrt{2} (W_t - W_{kh}) \]
\[ L_t - L_{kh} = - \int_{kh}^{t} \nabla f(L_s) \, ds + \sqrt{2} (W_t - W_{kh}) \]

\[ D_t - D_{kh} = -(t - kh) \nabla f(D_{kh}) + \sqrt{2} (W_t - W_{kh}) \]
\[ L_t - L_{kh} = -\int_{kh}^{t} \nabla f(L_s) \, ds + \sqrt{2} (W_t - W_{kh}) \]

\[ D_t - D_{kh} = -(t - kh) \nabla f(D_{kh}) + \sqrt{2} (W_t - W_{kh}) \]
Sketch of the proof/2

This readily yields

\[ L_{(k+1)h} - D_{(k+1)h} = L_{kh} - D_{kh} - h(\nabla f(L_{kh}) - \nabla f(D_{kh})) \]

\[ + \int_0^h (\nabla f(L_{kh+s}) - \nabla f(L_{kh})) \, ds. \]

Moreover, \( I - h\nabla f \) is a contraction.

We then check that with \( \rho = 1 - mh \),

\[ \| L_{(k+1)h} - D_{(k+1)h} \|_{L_2} \leq \rho \| L_{kh} - D_{kh} \|_{L_2} + 2M(h^3p)^{1/2}. \]

Using this inequality repeatedly for \( k+1, k, \ldots, 1 \), we get

\[ \| L_{(k+1)h} - D_{(k+1)h} \|_{L_2} \leq \rho^{k+1} \| L_0 - D_0 \|_{L_2} + 2M(h^3p)^{1/2}(1 + \rho + \ldots \rho^k) \]

\[ \leq \rho^{k+1} W_2(\nu_0, \pi) + 2M(h^3p)^{1/2}(1 - \rho)^{-1}. \]
Improved result with variable step-size

**Theorem 2** *(Dalalyan and Karagulyan, 2017)*

Consider the LMC with varying step-size $h_{k+1}$ defined by

$$h_{k+1} = \frac{2}{M + m + (2/3)m(k - K_1)_+}, \quad k = 1, 2, \ldots$$

where $K_1 \geq 0$ is the smallest integer satisfying

$$K_1 \geq \frac{\ln \left( W_2(\nu_0, \pi)/\sqrt{p} \right) + \ln(m/M) + (1/2) \ln(M + m)}{\ln(1 + 2m/M - m)}.$$

For every positive integer $k \geq K_1$, we have

$$W_2(\nu_k, \pi) \leq \frac{3.5M\sqrt{p}}{m\sqrt{M + m + (2/3)m(k - K_1)}}.$$
Remarks

1. **Theorem 3** implies that $O(p/\varepsilon^2 \log p/\varepsilon^2)$ gradient evaluations are enough for getting precision $\leq \varepsilon$.

2. **Theorem 2** implies that $O(p/\varepsilon^2)$ gradient evaluations are enough for getting precision $\leq \varepsilon$.

3. Similar result holds true for
   - the TV-distance (Dalalyan, 2017a), (Durmus and Moulines, 2017),
   - the KL-divergence (Cheng and Bartlett, 2017),
   - compact support $\pi$ (Bubeck et al., 2018), (Brosse et al., 2017).

4. **Further smoothness:** if $f$ is Hessian-Lipschitz, then $O(p/\varepsilon \log p/\varepsilon^2)$ gradient evaluations are enough for getting precision $\leq \varepsilon$ by the LMC. (Durmus and Moulines, 2019)

5. (Dwivedi et al., 2018; Chen et al., 2020) proved that for MALA, $O^*(p)$ gradient evaluations are enough for getting precision $\leq \varepsilon$. 

Dalalyan, A.S.
The distribution $\nu_t$ of the Langevin diffusion $L_t$ is the solution of

$$\dot{\nu}_t = -\nabla \mathcal{F}(\nu_t), \quad t \geq 0,$$

where

$$\mathcal{F}(\nu) = \int_{\mathbb{R}^p} f(\theta) \nu(\theta) d\theta + \int_{\mathbb{R}^p} \nu(\theta) \log \nu(\theta) d\theta.$$

and the time-derivative of the mapping $t \mapsto \nu_t$ should be understood in the sense of the Wasserstein-2 distance.

**Theorem 1 bis (sampling, improved):** If $h \leq 1/M$,

$$W_2(\nu_K, \pi) \leq (1 - mh)^{K/2} W_2(\nu_0, \pi) + (2Mhp/m)^{1/2}.$$

The difference with Theorem 1 is that the condition number $(M/m) > 1$ is now within the square root.
The case of noisy gradient

The setting

- The computation of $\nabla f$ might be costly or even impossible.
- But one might have access to a noisy version of it:

$$Y^k = \nabla f(\vartheta^k) + \zeta^k,$$

where $\{\zeta^{(k)}\}$ satisfy

- (bounded bias) $\mathbb{E}[\|\mathbb{E}(\zeta^k | \vartheta^k)\|_2^2] \leq \delta^2 p$,
- (bounded variance) $\mathbb{E}[\|\zeta^k - \mathbb{E}(\zeta^k | \vartheta^k)\|_2^2] \leq \sigma^2 p$,
- (ind. of updates) $\xi^{(k+1)}$ is independent of $(\zeta^0, \ldots, \zeta^k)$.

- The noisy LMC (nLMC) algorithm is then

$$\vartheta^{(k+1, h)} = \vartheta^{(k, h)} - h Y^{(k, h)} + \sqrt{2h} \xi^{(k+1)}.$$
The case of noisy gradient

Error estimate

- One has access to a noisy version of the gradient:
  \[ Y^k = \nabla f(\vartheta^k) + \zeta^k, \]

  where \( \{\zeta^{(k)}\} \) satisfy
  - \( \mathbb{E}\left[\|\mathbb{E}(\zeta^k | \vartheta^k)\|^2\right] \leq \delta^2p \) and \( \mathbb{E}\left[\|\zeta^k - \mathbb{E}(\zeta^k | \vartheta^k)\|^2\right] \leq \sigma^2p, \)
  - (ind. of updates) \( \xi^{(k+1)} \) is independent of \( (\zeta^0, \ldots, \zeta^k) \).

- The noisy LMC (nLMC) algorithm is then
  \[ \vartheta^{(k+1),h} = \vartheta^{(k),h} - hY^{(k),h} + \sqrt{2h} \xi^{(k+1)}. \]

**Theorem 3**

Let \( \vartheta^{(K),h} \) be the \( K \)-th iterate of the nLMC and \( \nu_K \) be its distribution. If \( h \leq \frac{2}{M+m} \) then we have

\[ W_2(\nu_K, \pi) \leq (1 - mh)^K W_2(\nu_0, \pi) + \frac{2M}{m}(hp)^{1/2} + \frac{\delta \sqrt{p}}{m} + \sigma(hp/m)^{1/2}. \]
Guarantees under additional smoothness

**CONDITION F:** $f \in C^2$ and for some $m, M, M_2 > 0$,
- (strong convexity) $\nabla^2 f(\theta) \succeq mI_p$, for every $\theta \in \mathbb{R}^p$,
- (bounded second derivative) $\nabla^2 f(\theta) \preceq MI_p$, for every $\theta \in \mathbb{R}^p$,
- (further smoothness) $\|\nabla^2 f(\theta) - \nabla^2 f(\theta')\| \leq M_2\|\theta - \theta'\|_2$.

**Theorem 4**

Let $\theta_{K,h}$ be the $K$-th iterate of the LMC and $\nu_K$ be its distribution. Then, for every $h \leq 2/(m+M)$

$$W_2(\nu_K, \pi) \leq (1 - mh)^K W_2(\nu_0, \pi) + \frac{M_2hp}{2m} + \frac{11Mh\sqrt{Mp}}{5m}, \quad (1)$$

$$W_2(\nu_{K,\text{LMCO}}, \pi) \leq (1 - 0.25mh)^K W_2(\nu_0, \pi) + \frac{11.5M_2h(p+1)}{m}. \quad (2)$$
3. Sampling using the kinetic Langevin diffusion
Kinetic Langevin diffusion

Under the same assumptions on the log-target $f$, one can consider the kinetic Langevin diffusion

$$
\begin{align*}
\frac{d}{dt} \begin{bmatrix} V_t \\ L_t \end{bmatrix} &= \begin{bmatrix} -(\gamma V_t + u \nabla f(L_t)) \\ V_t \end{bmatrix} dt + \sqrt{2\gamma u} \begin{bmatrix} I_p \\ 0_{p \times p} \end{bmatrix} dW_t,
\end{align*}
$$

(3)

where $\gamma > 0$ is the friction coeff. and $u > 0$ is the inverse mass.

The Langevin diffusion is obtained as a limit of $L_{\gamma t}$, where $L$ is defined as in (3) with $u = 1$, when $\gamma$ tends to infinity.

The continuous-time Markov process $(L_t, V_t)$ is positive recurrent. The corresponding invariant density is given by

$$
\begin{align*}
p_\star(\theta, v) &\propto \exp \left\{ -f(\theta) - \frac{1}{2u} \|v\|^2 \right\}, \\
&\quad \theta \in \mathbb{R}^p, \ v \in \mathbb{R}^p.
\end{align*}
$$

(4)

So under the invariant distribution, $L$ and $V$ are independent, $L \sim \pi$ and $V \sim \mathcal{N}(0, u)$. 

Dalalyan, A.S.
Kinetic Langevin diffusion

- One can discretize this process to sample from $p_*$ (hence from $\pi$).
- The quality of the resulting sampler will depend on two key properties of the process: rate of mixing and smoothness of sample paths.
- (Cheng et al., 2018) establishes that for $(\gamma, u) = (2, 1/M)$, the mixing rate in the Wasserstein distance is $e^{-(m/2M)t}$.
- On the other hand, sample paths of $\{L\}$ are smooth of order $\approx 3/2$ since
  \[ L_t = L_0 + \int_0^t V_s \, ds. \]
- Combining these two properties, (Cheng et al., 2018) prove that a suitable discretization of (3) leads to a sampler that achieves an error $\leq \varepsilon$ after $K$ iterations with $K = O^*(\varepsilon^2)^{1/2}$.
Main questions answered in our work

**Q1.** What is the rate of mixing of the continuous-time kinetic Langevin diffusion for general values of the parameters \( u \) and \( \gamma \)?

**Q2.** Is it possible to improve the rate of convergence of the KLMC by optimizing it over the choice of \( u, \gamma \) and the step-size?

**Q3.** If the function \( f \) happens to have a Lipschitz-continuous Hessian, is it possible to devise a discretization that takes advantage of this additional smoothness and leads to improved rates of convergence?

<table>
<thead>
<tr>
<th></th>
<th>gradient-Lipschitz</th>
<th>Hessian-Lipschitz</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMC</td>
<td>( p/\varepsilon^2 )</td>
<td>( p/\varepsilon )</td>
</tr>
<tr>
<td>KLMC</td>
<td>( \sqrt{p/\varepsilon^2} )</td>
<td>???</td>
</tr>
</tbody>
</table>
Mixing rate for any \((\gamma, u)\)

- A first observation is that, without loss of generality, we can focus our attention to the case \(u = 1\).

**Lemma** The modified process \((\bar{V}_t, \bar{L}_t) = (u^{-1/2}V_t/\sqrt{u}, L_t/\sqrt{u})\) is an kinetic Langevin diffusion with parameters \(\bar{\gamma} = \gamma/\sqrt{u}\) and \(\bar{u} = 1\).

**Theorem 1** For every \(\gamma, t > 0\), there exists \(\beta \geq \{m \wedge (\gamma^2 - M)\}/\gamma\) such that

\[
W_2(\mu P^L_t, \mu' P^L_t) \leq (\sqrt{2}/\gamma)e^{-\beta t}W_2(\mu, \mu').
\] (5)

- Slightly better \(\beta\) is

<table>
<thead>
<tr>
<th>(\gamma^2 \in ]0, M[</th>
<th>]M, m + M[</th>
<th>[m + M, 3m + M[</th>
<th>[3m + M, +\infty[</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta)</td>
<td>NA</td>
<td>(\frac{\gamma^2 - M}{\gamma})</td>
<td>(\frac{\gamma}{2} - \frac{M - m}{2\sqrt{2(m + M) - \gamma^2}})</td>
</tr>
</tbody>
</table>
The KLMC algorithm

- Set \( \psi_0(t) = e^{-\gamma t} \) and \( \psi_{k+1}(t) = \int_0^t \psi_k(s) \, ds \).
- The discretization is defined by the recursion:

\[
\begin{bmatrix}
\mathbf{v}_{k+1} \\
\mathbf{\vartheta}_{k+1}
\end{bmatrix} = \begin{bmatrix}
\psi_0(h)\mathbf{v}_k - \psi_1(h) \nabla f(\mathbf{\vartheta}_k) \\
\mathbf{\vartheta}_k + \psi_1(h)\mathbf{v}_k - \psi_2(h) \nabla f(\mathbf{\vartheta}_k)
\end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix}
\xi_{k+1} \\
\xi'_{k+1}
\end{bmatrix},
\]

where \((\xi_{k+1}, \xi'_{k+1})\) is a centered Gaussian satisfying s.t.
- \((\xi_j, \xi'_j)\)'s are iid,
- for any \(j\), the vectors \(((\xi_j)_1, (\xi'_j)_1), ((\xi_j)_2, (\xi'_j)_2), \ldots, ((\xi_j)_p, (\xi'_j)_p)\) are iid with the covariance matrix

\[
\mathbf{C} = \int_0^h \begin{bmatrix}
\psi_0(t) & \psi_1(t)
\end{bmatrix}^\top \begin{bmatrix}
\psi_0(t) & \psi_1(t)
\end{bmatrix} \, dt.
\]

- This recursion is obtained by replacing \( \nabla f(L_t) \) by \( \nabla f(L_{kh}) \), on \( t \in [kh, (k+1)h] \), by renaming \((V_{kh}, L_{kh})\) into \((\mathbf{v}_k, \mathbf{\vartheta}_k)\) and by explicitly solving the obtained linear SDE.

- This algorithm, that we will refer to as KLMC, has been first analyzed by Cheng et al. (2018).
Guarantees for the KLMC algorithm

**Theorem 5** (Dalalyan and Riou-Durand, 2020)

For every \( \gamma \geq \sqrt{m + M} \) and \( h \leq m/(4\gamma M) \), the distribution \( \nu_k \) of the \( k \)th iterate \( \vartheta_k \) of the KLMC algorithm (6) satisfies

\[
W_2(\nu_k, \pi) \leq \sqrt{2} \left(1 - \frac{0.75mh}{\gamma}\right)^k W_2(\nu_0, \pi) + \frac{Mh\sqrt{2\pi}}{m}.
\]

(7)

- The second term in the upper bound scales linearly as a function of the condition number \( \kappa \triangleq M/m \), whereas the corresponding term in (Cheng et al., 2018) scales as \( \kappa^{3/2} \).

- If we denote by \( K \) the number of iterations sufficient for the error to be smaller than \( \varepsilon \), our result leads to an expression of \( K \) in which \( W_2(\nu_0, \pi) \) is within a logarithm. The expression of \( K \) in (Cheng et al., 2018, Theorem 1) scales linearly in \( W_2(\nu_0, \pi) \).
Second-order KLMC

For $k \in \mathbb{N}$, we define $H_k = \nabla^2 f(\vartheta_k)$ and

$$
\begin{bmatrix}
v_{k+1} \\
\vartheta_{k+1}
\end{bmatrix} = 
\begin{bmatrix}
\psi_0(h)v_k - \psi_1(h)\nabla f(\vartheta_k) \\
\vartheta_k + \psi_1(h)v_k - \psi_2(h)\nabla f(\vartheta_k)
\end{bmatrix} + \sqrt{2\gamma}
\begin{bmatrix}
\xi^{(1)}_{k+1} \\
\xi^{(2)}_{k+1}
\end{bmatrix}
\begin{bmatrix}
\varphi_2(h)H_kv_k \\
\varphi_3(h)H_kv_k
\end{bmatrix} - \sqrt{2\gamma}
\begin{bmatrix}
H_k\xi^{(3)}_{k+1} \\
H_k\xi^{(4)}_{k+1}
\end{bmatrix},
\end{bmatrix}
$$

where $\varphi_{k+1}(t) = \int_0^t e^{-\gamma(t-s)}\psi_k(s)\,ds$ and

- the $p \times 4$-matrices $\Xi_{k+1} := (\xi^{(1)}_{k+1}, \xi^{(2)}_{k+1}, \xi^{(3)}_{k+1}, \xi^{(4)}_{k+1})$ are iid,
- the $p$ rows of $\Xi_{k+1}$ are iid centered Gaussian with the covariance matrix

$$
\tilde{C} = \int_0^h [\psi_0(t); \psi_1(t); \varphi_2(t); \varphi_3(t)]^\top [\psi_0(t); \psi_1(t); \varphi_2(t); \varphi_3(t)]\,dt.
$$

Dalalyan, A.S.
Theorem 6 (Dalalyan and Riou-Durand, 2020)

Assume that $f$ is $m$-strongly convex, its gradient is $M$-Lipschitz, and its Hessian is $M_2$-Lipschitz for the spectral norm. For every $\gamma \geq \sqrt{m + M}$ and $h \leq m / (5\gamma M)$, the distribution $\nu_k$ of the $k$th iterate of the second-order KLMC algorithm satisfies

$$W_2(\nu_k, \pi) \leq 7 \left(1 - \frac{mh}{4\gamma}\right)^{2k} W_2(\nu_0, \pi) + \frac{33h^2 M_2 M p}{m^2} + \frac{2h^2 M \sqrt{Mp}}{m}.$$ 

May be compared to the analogous bound for the KLMC:

$$W_2(\nu_k, \pi) \leq \sqrt{2} \left(1 - \frac{0.75mh}{\gamma}\right)^k W_2(\nu_0, \pi) + \frac{M h \sqrt{2p}}{m}.$$ 

Guarantees for the second-order KLMC
Concluding remarks

- As soon as $\gamma^2 > M$, the KL process mixes exponentially fast with a rate at least equal to $\left\{ m \wedge (\gamma^2 - M) \right\} / \gamma$. Therefore, for fixed values of $m$ and $M$, the nearly fastest rate of mixing is obtained for $\gamma^2 = m + M$ and is equal to $m/\sqrt{m + M}$.

- Optimization with respect to $\gamma$ and $u$ leads to improved constants but does not improve the rate as compared to the values $\gamma = 2$ and $u = 1/M$ used in (Cheng et al., 2018).

- Leveraging second-order information may help to reduce the number of steps of the algorithm by a factor proportional to $1/\sqrt{\varepsilon}$ ($\sqrt{p/\varepsilon}$ versus $\sqrt{p/\varepsilon}$).

- Better discretization error obtained by the randomized mid-point method (Shen and Lee, 2019) ($p^{1/3}/\varepsilon^{2/3}$ versus $\sqrt{p}/\varepsilon$).
thank you
References I


References II


