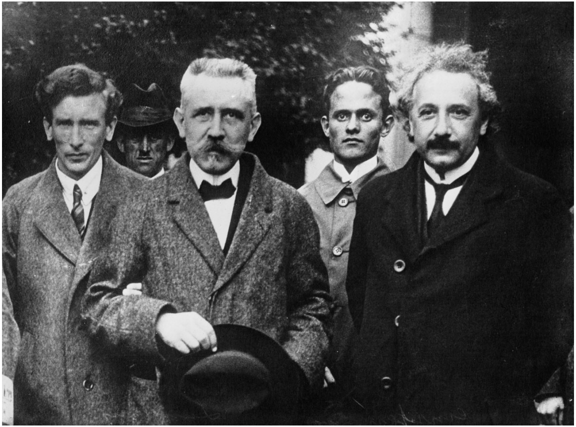
Discretized Langevin algorithms for non-strongly log-concave targets

Arnak S. Dalalyan

CREST/ ENSAE Paris / IP Paris



Paul Langevin and Albert Einstein 1923 (from `wikimedia`)

1. Introduction

Sampling from a density

Problem: Given a probability density function $\pi : \mathbb{R}^p \to \mathbb{R}$, generate a random vector $\boldsymbol{X}$ such that

$$\boldsymbol{X} \sim \pi,$$

that is $\mathbf{P}(\boldsymbol{X} \in A) = \int_A \pi(\boldsymbol{x})\, d\boldsymbol{x}$.

Warm-up: rejection sampling 1/2

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

Rejection method

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

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

Warm-up: rejection sampling 1/2

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

Rejection method

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

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

- Let K be the number of rounds required to sample $\boldsymbol{X}$.
  - the random variable $K \sim$ Geom(p)
  - with $p = \mathbf{P}(U \leq \pi(\boldsymbol{Y})/\nu(\boldsymbol{Y})) = 1/M$
  - the average number of rounds: $\mathbf{E}[K] = 1/p = M$.

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 $\pi(\boldsymbol{x}) \propto \mathbb{1}(\boldsymbol{x} \in \mathcal{C})$ with $\mathcal{C} \subset [0,1]^p$ 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 $\nu(\boldsymbol{x}) = \mathbb{1}(\boldsymbol{x} \in [0,1]^p)$.
- Then the almost only possible choice for M is $M = 1/\text{Vol}(B_r^p)$.

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 $\pi(\boldsymbol{x}) \propto \mathbb{1}(\boldsymbol{x} \in \mathcal{C})$ with $\mathcal{C} \subset [0,1]^p$ 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 $\nu(\boldsymbol{x}) = \mathbb{1}(\boldsymbol{x} \in [0,1]^p)$.
- Then the almost only possible choice for M is $M = 1/\mathrm{Vol}(B_r^p)$.

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

$$f(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta}).$$

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

$$f(\boldsymbol{\theta}) - f(\bar{\boldsymbol{\theta}}) - \nabla f(\bar{\boldsymbol{\theta}})^\top(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \geq \frac{m}{2}\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2^2, \qquad \textbf{(C1)}$$

$$\|\nabla f(\boldsymbol{\theta}) - \nabla f(\bar{\boldsymbol{\theta}})\|_2 \leq M\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2, \qquad \textbf{(C2)}$$

for all $\boldsymbol{\theta}, \bar{\boldsymbol{\theta}} \in \mathbb{R}^p$.

Goal: find nonasymptotic guarantees for approximately sampling from π. More precisely, for every $\epsilon > 0$ find a density μ such that one can efficiently sample from μ and

$$\big\|\mu - \pi\big\|_{\rm TV} \leq \epsilon \quad \text{or} \quad W_2(\mu, \pi) \leq \epsilon.$$

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

$$\|\mathbf{P}^k(\boldsymbol{x},\cdot) - \pi\|_{\rm TV} \le C_1(E_0)\big(\rho^k + e^{-C_2/h^{1/4}}\big)$$

holds for all k. Here $\mathbf{P}^k$ is the transition probability of a k-step MCMC.

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

$$\|\mathbf{P}^k(\boldsymbol{x},\cdot) - \pi\|_{\mathrm{TV}} \le C_1(E_0)\big(\rho^k + e^{-C_2/h^{1/4}}\big)$$

holds for all k. Here $\mathbf{P}^k$ is the transition probability of a k-step MCMC.

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

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

holds for all k. Here $\mathbf{P}^k$ 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 $U(\boldsymbol{x}) \geq 1$ and, for any $C > 0$ there exists an $E > 0$ such that*

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

for all $U(\boldsymbol{x}) > E$.

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

$$\Delta U(\boldsymbol{x}) \leq c|\nabla U(\boldsymbol{x})|^2 - dU(\boldsymbol{x}), \tag{2.4}$$

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

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

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

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

D) *There exists a constant $C > 0$ such that the first four derivatives of*

Optimization versus integration

Guarantees for sampling II

Fast Algorithms for Logconcave Functions: Sampling, Rounding, Integration and Optimization

László Lovász
Microsoft Research

Santosh Vempala *
Georgia Tech and MIT

Corollary 1.2 *Let f be a logconcave function in $\mathbb{R}^n$, given in the sense of (LS1), (LS2) and (LS3). Then for*

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

the total variation distance of σ^m and π_f is less than ε.

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

$$\|\mathbf{P}^k(\boldsymbol{x},\cdot) - \pi\|_{\mathrm{TV}} \leq \epsilon.$$

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 $\boldsymbol{\vartheta}^{(0)} \in \mathbb{R}^p$ and use the update rule

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

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

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

$$\boldsymbol{y}^{(k+1)} = \bar{\boldsymbol{\vartheta}}^{(k)} - h\nabla f(\bar{\boldsymbol{\vartheta}}^{(k)}) + \sqrt{2h}\,\boldsymbol{\xi}^{(k+1)},$$

$$\bar{\boldsymbol{\vartheta}}^{(k+1)} = \begin{cases} \boldsymbol{y}^{(k+1)}, & \text{with prob. } \alpha_k, \\ \bar{\boldsymbol{\vartheta}}^{(k)}, & \text{with prob. } 1 - \alpha_k \end{cases}$$

for a properly chosen acceptance rate $\alpha_k = \alpha(\bar{\boldsymbol{\vartheta}}^{(k)}, \boldsymbol{y}^{(k+1)})$.

Background on the Langevin algorithm
Langevin diffusion

- $\boldsymbol{\vartheta}^{(k)}$ is the Euler discretisation of the Langevin diffusion $\boldsymbol{L}_t$,
- the Langevin diffusion is defined by the SDE

$$d\boldsymbol{L}_t = -\nabla f(\boldsymbol{L}_t)\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \qquad t \geq 0.$$

- 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 $\mathbf{P}^t_{\boldsymbol{L}}(\boldsymbol{x}, \cdot\,)$, that is $\mathbf{P}^t_{\boldsymbol{L}}(\boldsymbol{x}, A) = \mathbf{P}(\boldsymbol{L}_t \in A | \boldsymbol{L}_0 = \boldsymbol{x})$.
- (C1-C2) yield the spectral gap property of the semigroup $\{\mathbf{P}^t_{\boldsymbol{L}} : t \in \mathbb{R}_+\}$. For any probability density ν,

$$\|\nu\mathbf{P}^t_{\boldsymbol{L}} - \pi\|_{\mathrm{TV}} \leq \frac{1}{2} D_{\mathrm{KL}}(\nu\|\pi)^{1/2} e^{-tm/2}, \qquad \forall t \geq 0.$$

Illustration of the link between Langevin diffusion and sampling

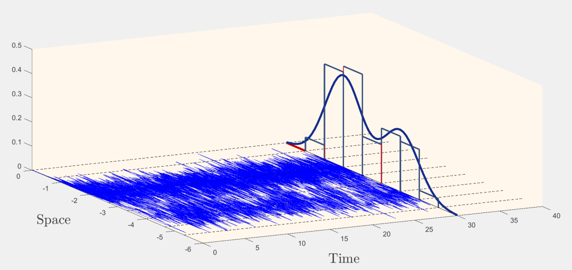


Figure: Illustration of Langevin dynamics. The blue lines represent different paths of a Langevin process. We see that the histogram of the state at time $t = 30$ is close to the target density (the dark blue line).

Background on the Langevin algorithm
Euler discretization

- the Langevin diffusion is defined by the SDE

$$d\boldsymbol{L}_t = -\nabla f(\boldsymbol{L}_t)\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \qquad t \geq 0.$$

- $\boldsymbol{\vartheta}^{(k)}$ is the Euler discretisation of the Langevin diffusion $\boldsymbol{L}_t$: $\boldsymbol{\vartheta}^{(k)} \approx \boldsymbol{L}_{kh}$.
- To be more precise, we introduce a diffusion-type continuous-time process $\boldsymbol{D}$ obeying the following SDE:

$$d\boldsymbol{D}_t = b_t(\boldsymbol{D})\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \qquad t \geq 0,$$

  with the drift $b_t(\boldsymbol{D}) = -\nabla f(\boldsymbol{D}_{kh})$ if $t \in [kh, (k+1)h[$.
- For this process, we have

$$(\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(k)}) \stackrel{\mathscr{D}}{=} (\boldsymbol{D}_h, \ldots, \boldsymbol{D}_{kh}).$$

Optimization versus sampling

Optimization

- **Problem:** compute

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

Sampling

- **Problem:** Sample $\boldsymbol{\vartheta}$ from the pdf

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

Optimization versus sampling

Optimization

- **Problem:** compute

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

- **Method:** gradient descent

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

Sampling

- **Problem:** Sample $\boldsymbol{\vartheta}$ from the pdf

$$\pi(\boldsymbol{\theta}) = \tfrac{1}{C} e^{-f(\boldsymbol{\theta})}, \quad C = \int_{\mathbb{R}^p} e^{-f}$$

- **Method:** Langevin Monte Carlo

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

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

Optimization versus sampling

Optimization

- **Problem:** compute

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

- **Method:** gradient descent

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

Sampling

- **Problem:** Sample $\boldsymbol{\vartheta}$ from the pdf

$$\pi(\boldsymbol{\theta}) = \tfrac{1}{C} e^{-f(\boldsymbol{\theta})}, \quad C = \int_{\mathbb{R}^p} e^{-f}$$

- **Method:** Langevin Monte Carlo

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

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

What about theoretical guarantees?

Optimization versus sampling
Theoretical guarantees

- We assume that for some $m, M > 0$

$$\begin{cases} f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') - \nabla f(\boldsymbol{\theta}')^\top(\boldsymbol{\theta} - \boldsymbol{\theta}') \geq ({}^m\!/{}_2)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2, \\ \|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \leq M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \end{cases} \qquad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p,$$

- **Theorem 0 (optim.):** If $h \leq 2/(m+M)$, then

$$\|\boldsymbol{\theta}^K - \boldsymbol{\theta}^*\|_2 \leq (1 - mh)^K \|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^*\|_2.$$

Optimization versus sampling
Theoretical guarantees

- We assume that for some $m, M > 0$

$$\begin{cases} f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') - \nabla f(\boldsymbol{\theta}')^\top(\boldsymbol{\theta} - \boldsymbol{\theta}') \geq (m/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2, \\ \|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \leq M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \end{cases} \qquad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p,$$

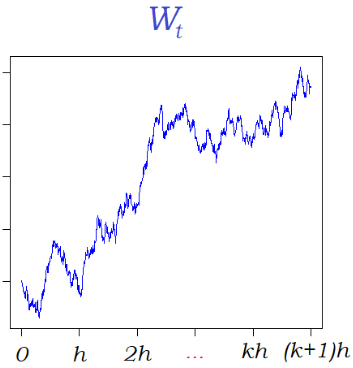
- **Theorem 0 (optim.):** If $h \leq 2/(m+M)$, then

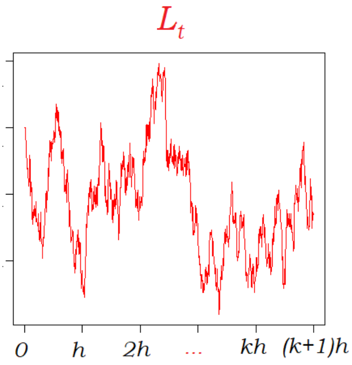
$$\|\boldsymbol{\theta}^K - \boldsymbol{\theta}^*\|_2 \leq (1 - mh)^K \|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^*\|_2.$$

- **Theorem 1(sampling):** If $h \leq 2/(m+M)$,

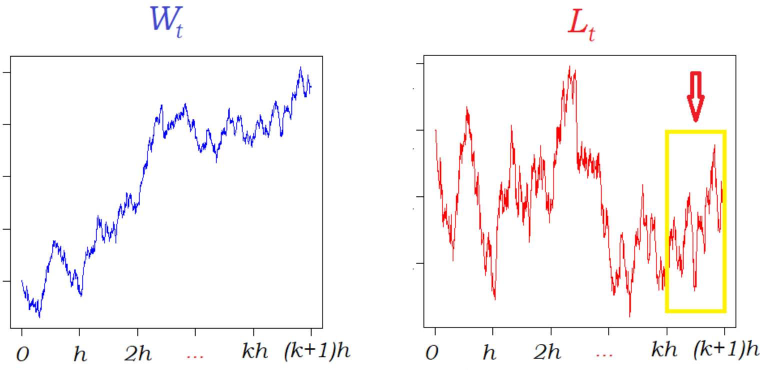
$$W_2(\nu_K, \pi) \leq (1 - mh)^K W_2(\nu_0, \pi) + \frac{2M}{m}(hp)^{1/2}.$$

(Durmus and Moulines, 2019; Dalalyan, 2017b)

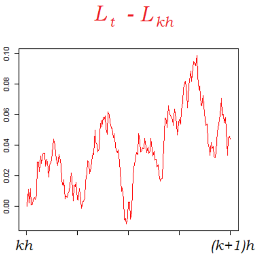




$$\boldsymbol{L}_t = \boldsymbol{L}_0 - \int_0^t \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\, \boldsymbol{W}_t$$

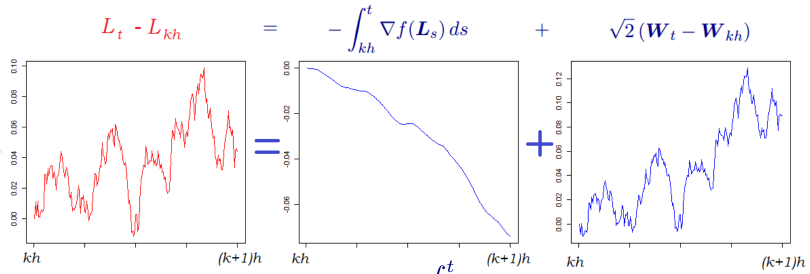


$$\boldsymbol{L}_t = \boldsymbol{L}_0 - \int_0^t \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\, \boldsymbol{W}_t$$



$$\boldsymbol{L}_t - \boldsymbol{L}_{kh} = -\int_{kh}^{t} \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\,(\boldsymbol{W}_t - \boldsymbol{W}_{kh})$$

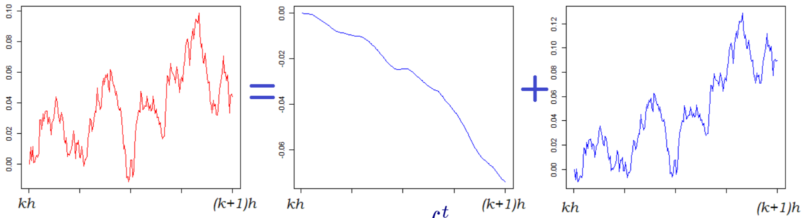
$$\boldsymbol{D}_t - \boldsymbol{D}_{kh} = -(t - kh)\nabla f(\boldsymbol{D}_{kh}) + \sqrt{2}\,(\boldsymbol{W}_t - \boldsymbol{W}_{kh})$$



$$\boldsymbol{L}_t - \boldsymbol{L}_{kh} = -\int_{kh}^{t} \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\,(\boldsymbol{W}_t - \boldsymbol{W}_{kh})$$

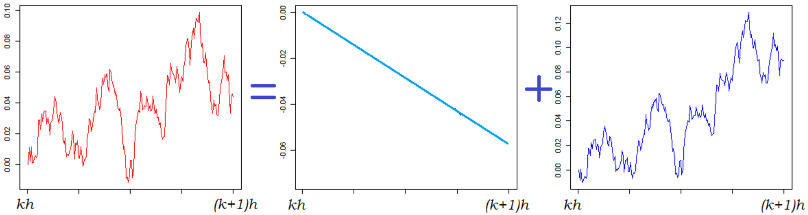
$$\boldsymbol{D}_t - \boldsymbol{D}_{kh} = -(t - kh)\nabla f(\boldsymbol{D}_{kh}) + \sqrt{2}\,(\boldsymbol{W}_t - \boldsymbol{W}_{kh})$$

$L_t - L_{kh} = -\int_{kh}^{t} \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\,(\boldsymbol{W}_t - \boldsymbol{W}_{kh})$



$$\boldsymbol{L}_t - \boldsymbol{L}_{kh} = -\int_{kh}^{t} \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\,(\boldsymbol{W}_t - \boldsymbol{W}_{kh})$$

$\boldsymbol{D}_t - \boldsymbol{D}_{kh} = -(t-kh)\nabla f(\boldsymbol{D}_{kh}) + \sqrt{2}\,(\boldsymbol{W}_t - \boldsymbol{W}_{kh})$



$$\boldsymbol{D}_t - \boldsymbol{D}_{kh} = -(t-kh)\nabla f(\boldsymbol{D}_{kh}) + \sqrt{2}\,(\boldsymbol{W}_t - \boldsymbol{W}_{kh})$$

Sketch of the proof/2

- This readily yields
$$\boldsymbol{L}_{(k+1)h} - \boldsymbol{D}_{(k+1)h} = \boldsymbol{L}_{kh} - \boldsymbol{D}_{kh} - h\big(\nabla f(\boldsymbol{L}_{kh}) - \nabla f(\boldsymbol{D}_{kh})\big) + \int_0^h (\nabla f(\boldsymbol{L}_{kh+s}) - \nabla f(\boldsymbol{L}_{kh}))\, ds.$$
Moreover, $\mathbf{I} - h\nabla f$ is a contraction.
- We then check that with $\rho = 1 - mh$,
$$\|\boldsymbol{L}_{(k+1)h} - \boldsymbol{D}_{(k+1)h}\|_{L_2} \leq \varrho\, \|\boldsymbol{L}_{kh} - \boldsymbol{D}_{kh}\|_{L_2} + 2M(h^3 p)^{1/2}.$$
- Using this inequality repeatedly for $k+1, k, \ldots, 1$, we get
$$\begin{aligned} \|\boldsymbol{L}_{(k+1)h} - \boldsymbol{D}_{(k+1)h}\|_{L_2} & \\ &\leq \varrho^{k+1}\, \|\boldsymbol{L}_0 - \boldsymbol{D}_0\|_{L_2} + 2M(h^3 p)^{1/2}(1 + \varrho + \ldots \varrho^k) \\ &\leq \varrho^{k+1}\, W_2(\nu_0, \pi) + 2M(h^3 p)^{1/2}(1 - \varrho)^{-1}. \end{aligned}$$

Improved result with variable step-size

Theorem 2 (Dalalyan and Karagulyan, 2017)

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

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

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

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

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 π (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 $\leq \varepsilon$.

Langevin as gradient flow in the space of measures

(Durmus et al., 2019)

The distribution $\boldsymbol{\nu}_t$ of the Langevin difusion $\boldsymbol{L}_t$ is the solution of

$$\dot{\boldsymbol{\nu}}_t = -\nabla \mathscr{F}(\boldsymbol{\nu}_t), \qquad t \geq 0,$$

where

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

and the time-derivative of the mapping $t \mapsto \boldsymbol{\nu}_t$ 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 ∇f might be costly or even impossible.
- But one might have access to a noisy version of it:

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

where $\{\zeta^{(k)}\}$ satisfy
  - (bounded bias) $\mathbf{E}\big[\big\|\mathbf{E}(\boldsymbol{\zeta}^k|\boldsymbol{\vartheta}^k)\big\|_2^2\big] \leq \delta^2 p$,
  - (bounded variance) $\mathbf{E}[\|\boldsymbol{\zeta}^k - \mathbf{E}(\boldsymbol{\zeta}^k|\boldsymbol{\vartheta}^k)\|_2^2] \leq \sigma^2 p$,
  - (ind. of updates) $\boldsymbol{\xi}^{(k+1)}$ is independent of $(\boldsymbol{\zeta}^0, \ldots, \boldsymbol{\zeta}^k)$.
- The noisy LMC (nLMC) algorithm is then

$$\boldsymbol{\vartheta}^{(k+1,h)} = \boldsymbol{\vartheta}^{(k,h)} - h\boldsymbol{Y}^{(k,h)} + \sqrt{2h}\,\boldsymbol{\xi}^{(k+1)}.$$

The case of noisy gradient
Error estimate

- One has access to a noisy version of the gradient:

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

  where $\{\boldsymbol{\zeta}^{(k)}\}$ satisfy
  - $\mathbf{E}\big[\big\|\mathbf{E}(\boldsymbol{\zeta}^k|\boldsymbol{\vartheta}^k)\big\|_2^2\big] \le \delta^2 p$ and $\mathbf{E}[\|\boldsymbol{\zeta}^k - \mathbf{E}(\boldsymbol{\zeta}^k|\boldsymbol{\vartheta}^k)\|_2^2] \le \sigma^2 p$,
  - (ind. of updates) $\boldsymbol{\xi}^{(k+1)}$ is independent of $(\boldsymbol{\zeta}^0, \ldots, \boldsymbol{\zeta}^k)$.
- The noisy LMC (nLMC) algorithm is then

$$\boldsymbol{\vartheta}^{(k+1,h)} = \boldsymbol{\vartheta}^{(k,h)} - h\boldsymbol{Y}^{(k,h)} + \sqrt{2h}\,\boldsymbol{\xi}^{(k+1)}.$$

Theorem 3

Let $\boldsymbol{\vartheta}^{(K,h)}$ be the K-th iterate of the nLMC and ν_K be its distribution. If $h \le {}^2/_{M+m}$ then we have

$$W_2(\nu_K, \pi) \le (1 - mh)^K W_2(\nu_0, \pi) + \frac{2M}{m}(hp)^{1/2} + \frac{\delta\sqrt{p}}{m} + \sigma(hp/m)^{1/2}.$$

Guarantees under additional smoothness

CONDITION F: $f \in C^2$ and for some m, M, $M_2 > 0$,

- (strong convexity) $\nabla^2 f(\boldsymbol{\theta}) \succeq m\mathbf{I}_p$, for every $\boldsymbol{\theta} \in \mathbb{R}^p$,
- (bounded second derivative) $\nabla^2 f(\boldsymbol{\theta}) \preceq M\mathbf{I}_p$, for every $\boldsymbol{\theta} \in \mathbb{R}^p$,
- (further smoothness) $\|\nabla^2 f(\boldsymbol{\theta}) - \nabla^2 f(\boldsymbol{\theta}')\| \le M_2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$.

Theorem 4

Let $\boldsymbol{\vartheta}_{K,h}$ be the K-th iterate of the LMC and ν_K be its distribution. Then, for every $h \le 2/(m+M)$

$$W_2(\nu_K, \pi) \le (1 - mh)^K W_2(\nu_0, \pi) + \frac{M_2 hp}{2m} + \frac{11Mh\sqrt{Mp}}{5m}, \quad (1)$$

$$W_2(\nu_K^{\text{LMCO}}, \pi) \le (1 - 0.25mh)^k W_2(\nu_0, \pi) + \frac{11.5M_2 h(p+1)}{m}. \quad (2)$$

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

$$d\begin{bmatrix}\boldsymbol{V}_t\\ \boldsymbol{L}_t\end{bmatrix} = \begin{bmatrix}-(\gamma\boldsymbol{V}_t + u\nabla f(\boldsymbol{L}_t))\\ \boldsymbol{V}_t\end{bmatrix} dt + \sqrt{2\gamma u}\begin{bmatrix}\mathbf{I}_p\\ \mathbf{0}_{p\times p}\end{bmatrix} d\boldsymbol{W}_t, \qquad (3)$$

  where $\gamma > 0$ is the friction coeff. and $u > 0$ is the inverse mass.
- The Langevin diffusion is obtained as a limit of $\boldsymbol{L}_{\gamma t}$, where $\boldsymbol{L}$ is defined as in (3) with $u = 1$, when γ tends to infinity.
- The continuous-time Markov process $(\boldsymbol{L}_t, \boldsymbol{V}_t)$ is positive recurrent. The corresponding invariant density is given by

$$p_*(\boldsymbol{\theta}, \boldsymbol{v}) \propto \exp\Big\{-f(\boldsymbol{\theta}) - \frac{1}{2u}\|\boldsymbol{v}\|_2^2\Big\}, \qquad \boldsymbol{\theta} \in \mathbb{R}^p,\ \boldsymbol{v} \in \mathbb{R}^p. \qquad (4)$$

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

Kinetic Langevin diffusion

- One can discretize this process to sample from p_* (hence from π).
- 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 $\{\boldsymbol{L}\}$ are smooth of order $\approx 3/2$ since

$$\boldsymbol{L}_t = \boldsymbol{L}_0 + \int_0^t \boldsymbol{V}_s \, ds.$$

- 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 $K = O^*((p/\varepsilon^2)^{1/2})$.

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 γ?

Q2. Is it possible to improve the rate of convergence of the KLMC by optimizing it over the choice of u, γ 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?

	gradient-Lipschitz	Hessian-Lipschitz
LMC	p/ε^2	p/ε
KLMC	$\sqrt{p/\varepsilon^2}$	???

Mixing rate for any (γ, u)

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

  Lemma The modified process $(\bar{\boldsymbol{V}}_t, \bar{\boldsymbol{L}}_t) = (u^{-1/2}\boldsymbol{V}_{t/\sqrt{u}}, \boldsymbol{L}_{t/\sqrt{u}})$ 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

$$W_2(\mu\mathbf{P}_t^{\boldsymbol{L}}, \mu'\mathbf{P}_t^{\boldsymbol{L}}) \leq (\sqrt{2}/\gamma)e^{-\beta\, t}W_2(\mu, \mu'). \tag{5}$$

- Slightly better β is

$\gamma^2 \in$	$]0, M]$	$]M, m+M]$	$[m+M, 3m+M[$	$[3m+M, +\infty[$
β	NA	$\dfrac{\gamma^2 - M}{\gamma}$	$\dfrac{\gamma}{2} - \dfrac{M-m}{2\sqrt{2(m+M)-\gamma^2}}$	$\dfrac{\gamma - \sqrt{\gamma^2 - 4m}}{2}$

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:

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\vartheta}_k) \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\vartheta}_k) \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \boldsymbol{\xi}_{k+1} \\ \boldsymbol{\xi}'_{k+1} \end{bmatrix}, \quad (6)$$

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

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

- This recursion is obtained by replacing $\nabla f(\boldsymbol{L}_t)$ by $\nabla f(\boldsymbol{L}_{kh})$, on $t \in [kh, (k+1)h]$, by renaming $(\boldsymbol{V}_{kh}, \boldsymbol{L}_{kh})$ into $(\boldsymbol{v}_k, \boldsymbol{\vartheta}_k)$ 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 ν_k of the kth iterate $\boldsymbol{\vartheta}_k$ of the KLMC algorithm (6) satisfies

$$W_2(\nu_k, \pi) \leq \sqrt{2}\Big(1 - \frac{0.75mh}{\gamma}\Big)^k W_2(\nu_0, \pi) + \frac{Mh\sqrt{2p}}{m}. \qquad (7)$$

- The second term in the upper bound scales linearly as a function of the condition number $\varkappa \triangleq M/m$, whereas the corresponding term in (Cheng et al., 2018) scales as $\varkappa^{3/2}$.
- If we denote by K the number of iterations sufficient for the error to be smaller than ε, 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 $\mathbf{H}_k = \nabla^2 f(\boldsymbol{\vartheta}_k)$ and

$$\begin{bmatrix} \boldsymbol{v}_{k+1} \\ \boldsymbol{\vartheta}_{k+1} \end{bmatrix} = \begin{bmatrix} \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\vartheta}_k) \\ \boldsymbol{\vartheta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\vartheta}_k) \end{bmatrix} + \sqrt{2\gamma} \begin{bmatrix} \boldsymbol{\xi}_{k+1}^{(1)} \\ \boldsymbol{\xi}_{k+1}^{(2)} \end{bmatrix}$$
$$- \begin{bmatrix} \varphi_2(h)\mathbf{H}_k\boldsymbol{v}_k \\ \varphi_3(h)\mathbf{H}_k\boldsymbol{v}_k \end{bmatrix} - \sqrt{2\gamma} \begin{bmatrix} \mathbf{H}_k\boldsymbol{\xi}_{k+1}^{(3)} \\ \mathbf{H}_k\boldsymbol{\xi}_{k+1}^{(4)} \end{bmatrix},$$

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

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

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

Guarantees for the second-order KLMC

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 ν_k of the kth iterate of the second-order KLMC algorithm satisfies

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

May be compared to the analogous bound for the KLMC:

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

Concluding remarks

- As soon as $\gamma^2 > M$, the KL process mixes exponentially fast with a rate at least equal to $\{m \wedge (\gamma^2 - M)\}/\gamma$. 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 γ 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 $\sqrt{p}/\varepsilon$).


рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
mersi
kia ora
barka
welalin
tack
spas
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
vinaka
спасибі
blagodaram
kiitos
dankie
nanni
nandri
dhanyavad
bedankt
bayarlalaa
gràçie
hvala
maururu
köszönöm
enkosi
thank you
akun dankon aĉiŭ
gracias
хвала
asante
manana
obrigada
tenki
murakoze
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
dĕkuji
obrigado
sagolun
sukriya
kop khun krap
taiku
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
terima kasih
rahmet
kam sah hamnida
rahmat
diolch
dhanyavadagalu
shukriya
mercé
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci

References I

N. Brosse, A. Durmus, É. Moulines, and M. Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In Proceedings of COLT, pages 319–342, 07–10 Jul 2017.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. Discrete & Computational Geometry, 59(4):757–783, Jun 2018.

Y. Chen, R. Dwivedi, M. Wainwright, and B. Yu. Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients. J. Mach. Learn. Res., 21:92:1–92:72, 2020.

X. Cheng and P. Bartlett. Convergence of Langevin MCMC in KL-divergence. ArXiv e-prints, May 2017.

X. Cheng, N. Chatterji, P. Bartlett, and M. Jordan. Underdamped langevin MCMC: A non-asymptotic analysis. In Conference On Learning Theory, COLT 2018, pages 300–323, 2018.

A. Dalalyan. Theoretical guarantees for approximate sampling from a smooth and log-concave density. J. R. Stat. Soc. B, 79:651–676, 2017a.

A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In Proceedings of COLT, pages 678–689, 07–10 Jul 2017b.

References II

A. Dalalyan and A. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. ArXiv e-prints, dec 2017.

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

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. Ann. Appl. Probab., 27(3):1551–1587, 06 2017.

A. Durmus and É. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. Bernoulli, 25(4A):2854 – 2882, 2019.

A. Durmus, S. Majewski, and B. Miasojedow. Analysis of langevin monte carlo via convex optimization. Journal of Machine Learning Research, 20(73): 1–46, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In Conference On Learning Theory, COLT 2018, pages 793–797, 2018.

R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.