On the convergence of the Unadjusted and Metropolis Adjusted Langevin Algorithms

MCM 2023

Alain Durmus[1]
Joint work with: Valentin De Bortoli[2], Ana Fernandez-Vidal[3], Andreas Eberle[4], Éric Moulines[1],
Marcelo Pereyra[3],

[1]Ecole Polytechnique
[2]ENS Paris, CNRS
[3]Maxwell Institute for Mathematical Sciences, Heriot-Watt University
[4]Bonn university

Outline

1. The Langevin Monte Carlo algorithm / ULA and its convergence
   - Introduction and motivations
   - Convergence of discretization of diffusions
2. Convergence of the Metropolis Adjsuted Langevin Algorithm (MALA)
3. Stochastic optimization by Langevin Monte Carlo
   - Presentation of the methodology
   - Theoretical results
   - Numerical experiments

Bayesian setting

- Bayesian decision theory relies on computing expectations:

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\mathrm{d}\pi(x) = \int_{\mathbb{R}^d} f(x)\pi(x)\mathrm{d}x$$

  Generic problem: estimation of an integral $\pi(f)$, where
  - π is known up to a multiplicative factor ;
  - Sampling directly from π is not an option;
- A solution is to approximate $\pi(f)$ by $n^{-1}\sum_{i=1}^{n} f(X_i)$ where $(X_i)_{i\geq 0}$ is a Markov chain associated with a Markov kernel P with invariant distribution π.
- We assume that π is positive on $\mathbb{R}^d$,

$$\pi : x \mapsto \mathrm{e}^{-U(x)} \Big/ \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)}\mathrm{d}y \ ,$$

- U is referred to as the potential associated with π.

(Overdamped) Langevin diffusion

- Langevin SDE:

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ ,$$

  where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian Motion.
- Notation: $(P_t)_{t\geq 0}$ the Markov semigroup associated to the Langevin diffusion:

$$P_t(x, A) = \mathbb{P}(Y_t \in A | Y_0 = x) \ , \quad x \in \mathbb{R}^d, A \in \mathcal{B}(\mathbb{R}^d) \ .$$

- $\pi(x) \propto \exp(-U(x))$ is the unique invariant probability measure.

Discretized Langevin diffusion

- Idea: Sample the diffusion paths, using the Euler-Maruyama (EM) scheme:

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} G_{k+1}$$

where
  - $(G_k)_{k \geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$
  - $\gamma > 0$ is a stepsize
- This algorithm is referred to as the Unadjusted Langevin Algorithm (ULA) in Bayesian statistics or Langevin Monte Carlo (LMC).
- U is not necessarily convex here but still gradient Lipschitz.

Discretized Langevin diffusion: constant stepsize

- $(X_k)_{k\geq 1}$ is an homogeneous Markov chain with Markov kernel R_γ
- Under mild conditions, $\mathrm{R}_\gamma \rightsquigarrow$ unique invariant distribution π_γ
- π_γ which does not coincide with the target distribution π
- Questions:
  - For a given precision $\epsilon > 0$, how should I choose the stepsize $\gamma > 0$ and the number of iterations n so that : $\mathrm{d}(\delta_x \mathrm{R}_\gamma^n, \pi) \leq \epsilon$ where d is some distance [could be the TV or the Wasserstein distance]

Wasserstein metrics and total variation

- The set of all couplings of ξ and ξ' is denoted by $\Pi(\xi, \xi')$. $\zeta \in \Pi(\xi, \xi')$ if :

$$\zeta(A \times \mathbb{R}^d) = \xi(A) \text{ and } \zeta(\mathbb{R}^d \times A) = \xi'(A) \text{ for all } A \in \mathcal{B}(\mathbb{R}^d) .$$

- Let ξ, ξ' be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Define the Wasserstein or Kantorovich-Rubinstein distance of order p by

$$\mathbf{W}_p^p(\xi, \xi') = \inf_{\zeta \in \Pi(\xi, \xi')} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - x'\|^p \, \zeta(\mathrm{d}x\mathrm{d}x') .$$

- Let ξ, ξ' be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Define the total variation distance by

$$\|\xi - \xi'\|_{\mathrm{TV}} = \inf_{\zeta \in \Pi(\xi, \xi')} \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathbb{1}_{\Delta_{\mathbb{R}^d}^c}(x, x') \zeta(\mathrm{d}x\mathrm{d}x') , \quad \Delta_{\mathbb{R}^d} = \{(x, x) : x \in \mathbb{R}^d\} .$$

(Very incomplete) existing results for ULA

1. Weak errors estimates [TT90; LP03].
2. Explicit errors [Dal14; DM17].
3. These results are based on
   - the comparison between the discretization and the diffusion process
   - quantify how the error introduced by the discretization accumulate throughout the algorithm.
4. In the following, we consider a different approach.

Reminder: fixed step size ULA

- Consider $(X_k)_{k\in\mathbb{N}}$ a Markov chain associated with the Euler scheme

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} G_{k+1} \;, \tag{1}$$

  to sample from $\pi \propto \mathrm{e}^{-U}$.
- Recall that R_γ is the Markov kernel associated with (1).
- We assume in the sequel that for any $\gamma > 0$, R_γ has an invariant probability distribution $\pi_\gamma \neq \pi$.
- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a test function for which we want to compute $\pi(f)$ and consider the estimator:

$$\hat{\pi}_n(f) = n^{-1} \sum_{k=1}^{n} f(X_k) \;.$$

Bias-Variance decomposition

- Consider the Mean Square error for $\hat{\pi}_n(f)$:

$$\mathbb{E}\left[|\hat{\pi}_n(f) - \pi(f)|^2\right] = \mathrm{bias}^2_{n,\gamma}(f) + \mathrm{Var}\left\{n^{-1}\sum_{k=1}^{n} f(X_k)\right\} ,$$

$$\mathrm{bias}_{n,\gamma}(f) = \left|n^{-1}\sum_{k=1}^{n}\{\mathbb{E}\left[f(X_k)\right] - \pi(f)\}\right| = \left|n^{-1}\sum_{k=1}^{n}\{\mu_0 \mathrm{R}^k_\gamma f - \pi(f)\}\right| .$$

Bound for the bias

- The presented works and many papers on ULA/discretization establish that for some numerical sequence $(u_n)_{n\in\mathbb{N}^*}$:

$$\mathbf{W}_{\mathbf{c}}(\mu_0 \mathrm{R}_\gamma^n, \pi) \leq u_n(\mu_0, \pi, \gamma) \ , \text{ for some assumptions on } U \ ,$$

  where for $\mathbf{c} : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$, the Wasserstein metric/distance $\mathbf{W}_{\mathbf{c}}(\mu, \nu)$ between μ and ν by

$$\mathbf{W}_{\mathbf{c}}(\mu, \nu) = \inf_{\zeta \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \mathbf{c}(x, y) \mathrm{d}\zeta(x, y) \ . \tag{2}$$

- Implies that $\mathrm{bias}_{n,\gamma}(f)$ can be estimated if $f \in \mathcal{F}$.
- $\mathcal{F}$ a class of function related to $\mathbf{c}$.
- Examples:
  1. $\mathbf{W}_{\mathbf{c}} = \mathbf{W}_1$, $\mathbf{c}(x, y) = \|x - y\|$ and $\mathcal{F} = \{f \text{ Lipschitz}\}$;
  2. $\mathbf{W}_{\mathbf{c}} = \|\cdot\|_{\mathrm{TV}}$, $\mathbf{c}(x, y) = \mathbb{1}_{x\neq y}$ and $\mathcal{F} = \{f \text{ bounded}\}$.

Bound on the variance term

- Consider the Mean Square error for $\hat{\pi}_n(f)$:

$$\mathbb{E}\left[|\hat{\pi}_n(f) - \pi(f)|^2\right] = \mathrm{bias}^2_{n,\gamma}(f) + \mathrm{Var}\left\{n^{-1}\sum_{k=1}^{n} f(X_k)\right\} .$$

- Question: what about the variance term? Can we have also explicit bound on this term?
- Natural answer: we need to have quantitative bound for the convergence of R_γ to π_γ.
- The same problem appears when dealing with concentration inequalities:

$$\mathbb{P}\left(|\hat{\pi}_n(f) - \pi(f)| \geq t\right) \leq ?? \, , \quad t \geq 0 \, .$$

Going back to the bias

- Recall

$$\mathrm{bias}_{n,\gamma}(f) = \left| n^{-1} \sum_{k=1}^{n} \{\mathbb{E}\left[f(X_k)\right] - \pi(f)\} \right| = \left| n^{-1} \sum_{k=1}^{n} \{\mu_0 \mathrm{R}_\gamma^k f - \pi(f)\} \right| .$$

- If we can show and quantify convergence of $(X_k)_{k\in\mathbb{N}}$ to π_γ, we can think about considering the decomposition

$$\begin{aligned} \mathrm{bias}_{n,\gamma}(f) &\leq |\pi(f) - \pi_\gamma(f)| + \left| n^{-1} \sum_{k=1}^{n} \{\mathbb{E}\left[f(X_k)\right] - \pi_\gamma(f)\} \right| \\ &= |\pi(f) - \pi_\gamma(f)| + \left| n^{-1} \sum_{k=1}^{n} \{\mu_0 \mathrm{R}_\gamma^k f - \pi_\gamma(f)\} \right| . \end{aligned}$$

- It remains to bound $|\pi(f) - \pi_\gamma(f)|$: see Andreas' talk!

Convergence of Markov processes

- The study of the convergence of Markov processes is an active field.
- Pioneering results from [NT78; NT82; NT83].
- [Pop77; MT92] established (f, r)-ergodicity on general state space using Foster-Lyapunov drift conditions in combination of an appropriate minorization condition.
- Applied in numerous papers [Cha93; CT91; RP94].
- Later extended to continuous-time Markov processes in [Kha11; MT93]....

Convergence of Markov processes (II)

- Most of these results in total variation or in V-norm and are non-quantitative.
- Let ξ, ξ' be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Define the V-norm for $V : \mathbb{R}^d \to [1, +\infty)$ by

$$\|\xi - \xi'\|_V = \sup_{|f| \le V} \left| \int_{\mathbb{R}^d} f(x) \mathrm{d}\{\xi - \xi'\}(x) \right| .$$

- Explicit convergence bounds in the same metrics for Markov chains have been established in [Ros95; For01; DMR04; Ros02]...
- The techniques developed in these papers have not been adapted to continuous-time Markov processes, except in [RR96].
- Deriving quantitative minorization conditions for continuous-time process seems to be even more difficult than for their discrete counterpart.

Wasserstein vs total variation distance

- To avoid the use of minorization conditions, Wasserstein metrics have shown to be very interesting.
- Following [HM11], [HMS11] generalizes the Harris' theorem for V-norms to handle more general Wasserstein type metrics.
- Use of Wasserstein distance has been successively applied to the study of diffusion processes and MCMC algorithms [Ebe16; Cha+18; Bak+; HSV14].
- One key idea introduced in [HMS11] and [Ebe16] is the construction of an appropriate metric designed specifically for the Markov process under consideration.
- We can still wonder if "good minorization conditions" can be found to derive similar bounds using classical results cited above.
- In particular for R_γ, $\gamma > 0$ under some conditions on U.

Preliminary observations

- For any $\gamma > 0$ and $n \in \mathbb{N}$, R_γ^n is supposed to be an approximation of $P_{n\gamma}$.
- Recall that $(P_t)_{t \geq 0}$ is the Markov semigroup associated with

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ .$$

- So, the convergence of R_γ to π_γ should be $\approx$ the one of $(P_t)_{t \geq 0}$ to π.
- Using different techniques, it can be shown that for some semi-metric dist

$$\mathrm{dist}(\mu_0 P_t, \pi) \leq C(\mu_0)\rho^t \ , \quad C \geq 0 \ , \rho \in (0,1) \ .$$

- Therefore, it is expected that roughly

$$\mathrm{dist}(\mu_0 \mathrm{R}_\gamma^n, \pi_\gamma) \leq C(\mu_0)\rho^{n\gamma} \ .$$

- The rate of convergence should scale linearly with γ!

Strongly convex potential: Convergence in $\mathbf{W}_2$

- Assume that U is m-strongly convex and L-gradient Lipschitz.
- By an easy computation, using the synchronuous coupling,

$$\mathbf{W}_2^2(\mu_0 P_{n\gamma}, \pi) \leq C_{2,\mathrm{c}}(\mu_0)\rho_{2,\mathrm{c}}^{\gamma n}\,, \quad \rho_{2,\mathrm{c}} = \mathrm{e}^{-m}\,,$$
$$\mathbf{W}_2^2(\mu_0 \mathrm{R}_\gamma^n, \pi_\gamma) \leq C_{2,\mathrm{d}}(\mu_0)\rho_{2,\mathrm{d}}^{\gamma n}\,, \quad \rho_{2,\mathrm{d}} = \mathrm{e}^{-\varpi}\,, \quad \varpi = 2mL/(m+L)\,.$$

Strongly convex potential: Convergence in total variation

- Assume that U is m-strongly convex and L-gradient Lipschitz.
- For the total variation distance, using the reflexion coupling, we obtain the following result.

Theorem 1 (DM19)

- *For any* $\gamma > 0$ *and* $n \in \mathbb{N}^*$, $n \geq 2/\gamma$,

$$\|\mu_0 P_{n\gamma} - \pi\|_{\mathrm{TV}} \leq C_{\mathrm{TV,c}}(\mu_0)\rho_{\mathrm{TV,c}}^{\gamma n} , \quad \rho_{\mathrm{TV,c}} = \mathrm{e}^{-m} .$$

- *For any* $\gamma \in (0, 2/(m+L))$, *and* $n \in \mathbb{N}^*$, $n \geq 2/\gamma$,

$$\|\mu_0 \mathrm{R}_\gamma^n - \pi_\gamma\|_{\mathrm{TV}} \leq C_{\mathrm{TV,d}}(\mu_0)\rho_{\mathrm{TV,d}}^{\gamma n} , \quad \rho_{\mathrm{TV,d}} = \mathrm{e}^{-\varpi} .$$

- We get the same convergence rate for the total variation distance and the Wasserstein distance!

1. The Langevin Monte Carlo algorithm / ULA and its convergence
   - Introduction and motivations
   - Convergence of discretization of diffusions
2. Convergence of the Metropolis Adjsuted Langevin Algorithm (MALA)
3. Stochastic optimization by Langevin Monte Carlo
   - Presentation of the methodology
   - Theoretical results
   - Numerical experiments

Functional auto-regressive models

- In [DD19], we study the convergence in $\|\cdot\|_{\mathrm{TV}}$ of a class of Markov chains.
- For any $\gamma > 0$, R_γ belongs to this class.
- We study Markov chains $(\mathrm{X}_k)_{k\in\mathbb{N}}$ on $\mathbb{R}^d$ defined by the recursion:

$$\mathrm{X}_{k+1} = \mathcal{T}_\gamma(\mathrm{X}_k) + \sqrt{2\gamma} G_{k+1} \;,$$

where

- $(G_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. Gaussian random variable with zero mean and covariance identity.
- $\{\mathcal{T}_\gamma \,:\, \gamma \in (0, \bar{\gamma}]\}$ is a sequence of functions from $\mathbb{R}^d$ to $\mathbb{R}^d$.

- For the Euler-Maruyama discretization, $\mathcal{T}_\gamma \leftarrow \{x \mapsto x - \gamma \nabla U(x)\}$.

Assumption on functional auto-regressive models

- Consider the assumption: for any $\gamma \in (0, \bar{\gamma}]$ and $x, y \in \mathbb{R}^d$,

$$\|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\|^2 \leq (1 + \gamma\kappa(\gamma)) \|x - y\|^2 \ .$$

- For the EM discretization corresponds to a one-side Lipschitz condition on ∇U: there exists $\kappa \in \mathbb{R}$, for any $x, y \in \mathbb{R}^d$,

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq \kappa \|x - y\| \ .$$

  Indeed, it implies that

$$\|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\|^2 \leq (1 + \gamma\kappa(\gamma)) \|x - y\|^2 \ , \text{ with } \kappa(\gamma) = -2\kappa + L^2\gamma \ .$$

- κ measures in some sense the default of convexity of U:
  - if $\kappa > 0$, for γ small enough, $1 + \gamma\kappa(\gamma) \leq 1$;
  - if $\kappa < 0$, for any $\gamma > 0$, $1 + \gamma\kappa(\gamma) > 1$.

Minorization conditions for functional auto-regressive models

- Consider the assumption: for any $\gamma \in (0, \bar{\gamma}]$ and $x, y \in \mathbb{R}^d$,

$$\|\mathcal{T}_\gamma(x) - \mathcal{T}_\gamma(y)\|^2 \leq (1 + \gamma\kappa(\gamma)) \|x - y\|^2 \ .$$

- Then explicit minorization conditions [DD19] can be found for $(\mathrm{X}_k)_{k\in\mathbb{N}}$

$$\mathrm{X}_{k+1} = \mathcal{T}_\gamma(\mathrm{X}_k) + \sqrt{\gamma} G_{k+1} \ .$$

- The constants which appear do not depend on the dimension and are sharp!
- It remains to apply results developed in the litterature to obtain quantitative bounds for $(\mathrm{X}_k)_{k\in\mathbb{N}}$ if a Lyapunov condition holds!
- Or for some cases not...

Lyapunov conditions for functional auto-regressive models

- Denote by Q_γ the Markov kernel associated with $(\mathrm{X}_k)_{k\in\mathbb{N}}$:

$$\mathrm{X}_{k+1} = \mathcal{T}_\gamma(\mathrm{X}_k) + \sqrt{\gamma} G_{k+1} \ .$$

- Assume that Q_γ satisfies a Foster-Lyapunov condition for any $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$:

$$\mathrm{Q}_\gamma V(x) \leq \lambda_\gamma V(x) + b_\gamma \mathbb{1}_{\mathrm{D}}(x) \ , \text{ with } \mathrm{D} \subset \mathbb{R}^d \ , \lambda_\gamma \in (0,1) \text{ and } b_\gamma \geq 0 \ . \quad (3)$$

- It turns out that it is a really bad idea to apply as such existing results directly to Q_γ.
- We need to consider instead $\mathrm{Q}_\gamma^{\lceil 1/\gamma \rceil}$.
- Therefore, we need to consider a Lyapunov-drift condition for this Markov kernel!

Lyapunov conditions for functional auto-regressive models

- Assume that Q_γ satisfies a Foster-Lyapunov condition for any $\gamma \in (0, \bar{\gamma}]$ and $x \in \mathbb{R}^d$:

$$\mathrm{Q}_\gamma V(x) \leq \lambda_\gamma V(x) + b_\gamma \mathbb{1}_{\mathrm{D}}(x) \;, \text{ with } \mathrm{D} \subset \mathbb{R}^d \;, \lambda_\gamma \in (0,1) \text{ and } b_\gamma \geq 0 \;. \quad (4)$$

- If we iterate Equation (4), we end up with

$$\mathrm{Q}_\gamma^{\lceil 1/\gamma \rceil} V(x) \leq \tilde{\lambda} V(x) + \tilde{b} \;, \text{ with } \tilde{\lambda} \in (0,1) \text{ and } \tilde{b} \geq 0 \;.$$

- But we do not have the indicator function anymore which lead to non-sharp results.
- We adapt proofs of the results in [Ros92,DM17] to get sharp convergence bounds for Q_γ.

Application to potential strongly convex outside a ball

H1

- *U is L-gradient Lipschitz.*
- *There exists $m \in \mathbb{R}_-^*$ such that for any $x, y \in \mathbb{R}^d$,*

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m \|x - y\|^2 \ . \tag{5}$$

- *There exist $m^+ > 0$ and $R \geq 0$ such that for any $x, y \in \mathbb{R}^d$, $\|x - y\| \geq R$,*

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m^+ \|x - y\|^2 \ . \tag{6}$$

Convergence of the Euler discretization

Theorem 2

Assume **H**1. *Then there exist* $\bar{\gamma} > 0$, $D_{\bar{\gamma},1}, D_{\bar{\gamma},2}, E_{\bar{\gamma}} \geq 0$ *and* $\lambda_{\bar{\gamma}}, \rho_{\bar{\gamma}} \in [0,1)$ *with* $\lambda_{\bar{\gamma}} \leq \rho_{\bar{\gamma}}$, *which can be explicitly computed, such that for any* $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$

$$\mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq \lambda_{\bar{\gamma}}^{k\gamma/4} [D_{\bar{\gamma},1} \mathbf{c}(x,y) + D_{\bar{\gamma},2} \mathbb{1}_{x \neq y}] + E_{\bar{\gamma}} \rho_{\bar{\gamma}}^{k\gamma/4} \mathbb{1}_{x \neq y} \,, \tag{7}$$

where $\mathbf{c}(x,y) = \mathbb{1}_{x \neq y}(1 + \|x - y\| / R)$.

- It is sensible to obtain two different convergence rates $\lambda_{\bar{\gamma}}, \rho_{\bar{\gamma}}$ in Theorem 2.
- One characterizing the forgetting of the initial distance between the two starting points $x, y \in \mathbb{R}^d$, corresponding to a burn-in period.
- The other one characterizing the effective convergence.
- Note that $\lambda_{\bar{\gamma}} \ll \rho_{\bar{\gamma}}$.

Convergence of the Euler discretization (II)

Corollary 3

Assume **H***1 Then, there exist* $\bar{\gamma} > 0$, $E_{\bar{\gamma},1}, E_{\bar{\gamma},2} \geq 0$ *such that for any* $\gamma \in (0, \bar{\gamma}]$, $x, y \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$ *we have*

$$\|\delta_x \mathrm{R}_\gamma^k - \delta_y \mathrm{R}_\gamma^k\|_{\mathrm{TV}} \leq \mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq E_{\bar{\gamma},1} \rho_{\bar{\gamma}}^{k\gamma/4} \mathbf{c}(x, y) \;, \tag{8}$$

$$\mathbf{W}_1(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k) \leq E_{\bar{\gamma},2} \rho_{\bar{\gamma}}^{k\gamma/4} \|x - y\| \;, \tag{9}$$

In addition, the constants, $\{E_{\bar{\gamma},i} : i = 1, 2, 3\}$ *can be explicitly computed.*

- This result and the first one imply quantitative convergence for variance of additive functional and concentration inequality.
- No need of strict contraction!
- Bounds also for $\mathbf{W}_p(\delta_x \mathrm{R}_\gamma^k, \delta_y \mathrm{R}_\gamma^k)$.

Convergence of the diffusion

Theorem 4

Assume **H***1. Then there exist* $D_1, D_2, E \geq 0$ *and* $\lambda, \rho \in [0,1)$ *with* $\lambda \leq \rho$ *such that for any* $x, y \in \mathbb{R}^d$ *and* $t \geq 0$

$$\|\delta_x \mathrm{P}_t - \delta_y \mathrm{P}_t\|_{\mathrm{TV}} \leq \mathbf{W}_{\mathbf{c}}(\delta_x \mathrm{P}_t, \delta_y \mathrm{P}_t) \leq \lambda^{t/4}[D_1 \mathbf{c}(x,y) + D_2 \mathbb{1}_{x \neq y}] + E \rho^{t/4} \mathbb{1}_{x \neq y} \,, \tag{10}$$

where $\mathbf{c}(x,y) = \mathbb{1}_{x \neq y}(1 + \|x-y\|/R)$, $(\mathrm{P}_t)_{t \geq 0}$ *is the Markov semigroup associated with the Langevin semigroup and*

$$\begin{cases} D_1 = \lim_{\bar{\gamma} \to 0} D_{\bar{\gamma},1} \,, \quad D_2 = \lim_{\bar{\gamma} \to 0} D_{\bar{\gamma},2} \,, \quad E = \lim_{\bar{\gamma} \to 0} E_{\bar{\gamma}} \,, \\ \lambda = \lim_{\bar{\gamma} \to 0} \lambda_{\bar{\gamma}} \,, \quad \rho = \lim_{\bar{\gamma} \to 0} \rho_{\bar{\gamma}} \,, \end{cases} \tag{11}$$

and $D_{\bar{\gamma},1}, D_{\bar{\gamma},2}, E_{\bar{\gamma}}, \lambda_{\bar{\gamma}}, \rho_{\bar{\gamma}}$ *are given in Theorem 2.*

- Note that the constants D_1, D_2, E, λ and ρ have explicit expressions.

Comparison with existing results

- First, a major difference between our work and other ones on the same subject is that we use a completely different technique to establish our results.
- They show in general a strict contraction for $\mathbf{W}_{\mathbf{c}}$ for well-chosen $\mathbf{c}$.
- Therefore they do not dissociate the forgotten of initial conditions and the effective convergence rate.
- The convergence rate obtained in [EM18] [MMS18] are smaller than ours regarding the discretization.
- The convergence in the total variation we derive is new using a probabilistic strategy and improve [EGZ18] (dependence on the dimension).

Outline

The Metropolis Adjsuted Langevin Algorithm

- To circumvent the bias of ULA, [RT96b; RDF78; Nea93] suggest to use a Metropolis filter.
- This defines the Markov chain

$$X_{k+1} = Y_{k+1}\mathbb{1}\left\{\alpha(X_k, Y_{k+1}) \leq U_{k+1}\right\} + X_k\mathbb{1}\left\{\alpha(X_k, Y_{k+1}) > U_{k+1}\right\} , \quad (12)$$

where

$$Y_{k+1} = X_k - \gamma\nabla U(X_k) + \sqrt{2\gamma}G_{k+1} \quad (13)$$

$$\alpha(x, y) = 1 \wedge \left\{\frac{\pi(y)r_\gamma(y, x)}{\pi(x)r_\gamma(x, y)}\right\} . \quad (14)$$

- Denote by $\overline{\mathrm{R}}_\gamma$ the Markov kernel associated with MALA.

Convergence of MALA

- What about convergence of MALA?
- Under very mild assumption on U:

$$\lim_{k \to +\infty} \|\delta_x \bar{R}_\gamma - \pi\|_{\mathrm{TV}} = 0 \ , \ \text{for any } x \in \mathbb{R}^d \ . \tag{15}$$

- Question: can we quantify the convergence?
- Here we are particularly interested in geometric ergodicity.

Conditions on U

H2

- *U is L-gradient Lipschitz.*
- *$U = U_1 + U_2$ with $U_2, \nabla U_2$ bounded and*

$$\nabla^2 U_1(x) \succeq m \, \mathrm{Id} \ , \ \textit{for any } x \in \mathbb{R}^d, \ \|x\| \geq R \ . \tag{16}$$

- Condition (16) is stronger than condition **H**1–(6).

A first result

Theorem 5 ([DM23])

Assume **H**2. *Then, there exists* $\bar{\gamma} > 0$ *such that for any* $\gamma \in (0, \bar{\gamma}]$, *there exist* $C_\gamma \geq 0$, $\rho_\gamma \in [0, 1)$ *satisfying for any* x, k

$$\|\delta_x \bar{R}_\gamma^k - \pi\|_V \leq C_\gamma \rho_\gamma^k V(x) \,, \tag{17}$$

where

$$V(x) = \exp(m \|x\|^2 / 16) \,. \tag{18}$$

- $C_\gamma \geq 0$, $\rho_\gamma \in [0, 1)$ are non-quantitative.
- In particular, the dependence with respect to γ is unclear.

A quantitative result

H3

Assume that $\sup_x \left\| D^3 U(x) \right\| \leq M < +\infty$.

Theorem 6 ([DM23])

Assume **H2** *and* **H3**. *Then, there exists* $\bar{\gamma} > 0$, $C_{\bar{\gamma}} \geq 0$, $\rho_{\bar{\gamma}} \in [0, 1)$ *such that for any* $\gamma \in (0, \bar{\gamma}]$, x, k

$$\|\delta_x \bar{R}_\gamma^k - \pi\|_V \leq C_{\bar{\gamma}} \rho_{\bar{\gamma}}^{k\gamma} V(x) \,, \tag{19}$$

where $V(x) = \exp(m \|x\|^2 / 16)$.

- $C_{\bar{\gamma}} \geq 0$, $\rho_{\bar{\gamma}} \in [0, 1)$ are quantitative.
- We get back the linear dependence of the convergence rate with respect to γ:

$$\log(\rho_{\bar{\gamma}}^{\gamma}) = \gamma \log(\rho_{\bar{\gamma}}) \,. \tag{20}$$

Comparison with existing results I/III

- V-geometric ergodicity has been shown for MALA in [RT96a] under strong condition on U.

H4

Assume

- $$\lim_{\|x\| \to +\infty} \int \mathbb{1}_{A(x) \cap I(x)}(y) r_\gamma(x, y) \mathrm{d}y = 0 \,, \tag{21}$$

  where

  $$A(x) = \{y \,:\, \alpha(x, y) = 1\} \,, \quad I(x) = B(0, \|x\|) \,. \tag{22}$$

- $$\liminf_{\|x\| \to +\infty} \{\|x\| - \|x - \gamma \nabla U(x)\|\} \geq \eta \,. \tag{23}$$

- Condition very hard to verify!
- The Lyapunov function that [RT96a] depends on γ: $V(x) = \exp(a \|x\|)$, $a \leq \gamma\eta$.

Comparison with existing results II/III

- Analysis of MALA using conductance techniques [Dwi+18; Che+21]
- Assume U is strongly convex
- Require a proper initialization

Comparison with existing results III/III

H5

Assume that U is C^4 and

$$\nabla^2 U(x) \succeq m\,\mathrm{Id}\ , \quad \textit{for } \|x\| \geq R\ . \tag{24}$$

Theorem 7 (bourabee:hairer:2013)

Assume **H5**, *then there exists $\bar{\gamma} > 0$, $C_{\bar{\gamma}} \geq 0$, $\rho_{\bar{\gamma}} \in [0,1)$ such that for any $\gamma \in (0, \bar{\gamma}]$, $\|x\| \leq E_0$, k*

$$\|\delta_x \bar{R}_\gamma^k - \pi\|_{\mathrm{TV}} \leq C_{\bar{\gamma}}(E_0)(\rho_{\bar{\gamma}}^{k\gamma} V(x) + \phi(\gamma))\ . \tag{25}$$

Outline

1. The Langevin Monte Carlo algorithm / ULA and its convergence
   - Introduction and motivations
   - Convergence of discretization of diffusions
2. Convergence of the Metropolis Adjsuted Langevin Algorithm (MALA)
3. Stochastic optimization by Langevin Monte Carlo
   - Presentation of the methodology
   - Theoretical results
   - Numerical experiments

Stochastic optimization setting

- Let Θ be a convex closed set in $\mathbb{R}^{d_\Theta}$.
- Consider an objective function $f : \Theta \to \mathbb{R}$ which we want to minimize.
- Its gradient is given for any $\theta \in \Theta$ by

$$\nabla f(\theta) = \int_{\mathbb{R}^d} H_\theta(x)\pi_\theta(\mathrm{d}x) \,,$$

where

- $(\theta, x) \mapsto H_\theta(x) \in \mathrm{C}(\Theta \times \mathbb{R}^d, \mathbb{R})$;
- $(\pi_\theta)_{\theta\in\Theta}$ family of probability distributions over $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

Stochastic approximation (II)

- Consider an objective function $f : \Theta \to \mathbb{R}$ which we want to minimize
- Its gradient is given for any $\theta \in \Theta$ by

$$\nabla f(\theta) = \int_{\mathbb{R}^d} H_\theta(x)\pi_\theta(\mathrm{d}x) \; .$$

- To optimize f, consider the classical stochastic recursion [RM51]: starting from $\theta_0 \in \Theta$,

$$\theta_{n+1} = \Pi_\Theta \left[\theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} H_{\theta_n}(Y_k^n) \right] \; , \tag{26}$$

where

- $(\delta_n)_{n\in\mathbb{N}^*} \in (\mathbb{R}_+^*)^{\mathbb{N}^*}$, $(m_n)_{n\in\mathbb{N}} \in (\mathbb{N}^*)^{\mathbb{N}}$: stepsizes and batch sizes;
- Π_Θ orthogonal projection onto Θ;
- for any $n \in \mathbb{N}$, $(Y_k^n)_{k\in\{1,\dots,m_n\}}$ i.i.d. $\sim \pi_{\theta_n}$.

Stochastic approximation (III)

- Then, a sequence of approximate minimizers of f: $(\hat{\theta}_N)_{N\in\mathbb{N}^*}$ where for any $N \in \mathbb{N}^*$

$$\hat{\theta}_N = \left\{\sum_{n=1}^{N} \delta_n \theta_n\right\} \Big/ \left\{\sum_{n=1}^{N} \delta_n\right\} \ .$$

An illustrative example: empirical Bayes estimation

- Consider the hierarchical model based on the observation y:

$$\begin{aligned} & (y, x, \theta) \mapsto p(y|x, \theta) \\ \text{prior distributions} \quad & (x, \theta) \mapsto p(x|\theta) \text{ and } \theta \mapsto p(\theta) \ , \end{aligned}$$

  - $x \in \mathbb{R}^d$ is the parameter of interest;
  - $\theta \in \Theta \subset \mathbb{R}^{d_\Theta}$ is a hyperparameter.
- The the a posteriori distribution is given for any $x \in \mathbb{R}^d$ by

$$p(x|y) \propto \int_\Theta p(y|x, \theta) p(x|\theta) p(\theta) \mathrm{d}\theta \ .$$

- Question: how to sample from $p(x|y)$?
- One solution consists in sampling from $p(\theta|x, y)$ and $p(x|y, \theta)$ alternatively and performing inference using the marginal distribution along the variable x.

An illustrative example, the EB setting (II)

- The a posteriori distribution is given for any $x \in \mathbb{R}^d$ by

$$p(x|y) \propto \int_\Theta p(y|x,\theta)p(x|\theta)p(\theta)\mathrm{d}\theta \; .$$

- Another solution consists in approximating the a posteriori distribution of x given y by

$$p(y|x,\theta^*)p(x|\theta^*)p(\theta^*) \text{ up to a normalizing constant}$$

with $\theta^* \in \arg\max p(\theta|y)$

$$p(\theta|y) \propto \int_{\mathbb{R}^d} \frac{p(y|\theta,x)p(x|\theta)p(\theta)}{p(y)}\mathrm{d}x = \int_{\mathbb{R}^d} \frac{p(x,y,\theta)}{p(y)}\mathrm{d}x \; .$$

- It defines the empirical Bayes setting [CL00; Cas85; Rob85].
- Now how to estimate θ^*?

An illustrative example, the EB setting (III)

- $$\theta^* \in \arg\max p(\theta|y) \;, \quad p(\theta|y) \propto \int_{\mathbb{R}^d} p(y|\theta,x)p(x|\theta)p(\theta)\mathrm{d}x = \int_{\mathbb{R}^d} p(x,y,\theta)\mathrm{d}x \;.$$
- Now how to estimate θ^*?
- Solution: using stochastic approximation approach with

$$H_\theta(x) = \nabla_\theta p(x,y,\theta)/p(x,y,\theta) \;, \qquad \pi_\theta(x) = p(x|\theta,y) = \frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)} \;.$$

- Indeed, using that $\frac{p(y|x,\theta)p(x|\theta)}{p(y|\theta)} = \frac{p(x,y,\theta)}{p(\theta|y)}$, we get

$$\nabla_\theta \log p(\theta|y) = \int_{\mathbb{R}^d} \frac{\nabla_\theta p(x,y,\theta)}{p(\theta|y)}\mathrm{d}x = \int \frac{\nabla_\theta p(x,y,\theta)}{p(x,y,\theta)}\pi_\theta(x)\mathrm{d}x \;.$$

- However, in some case sampling from π_θ is not an option!
- Other applications where it is not possible: maximum marginal likelihood estimation, texture synthesis...

Stochastic optimization using MCMC

- Question: how to use stochastic approximation as sampling from π_θ is not an option.
- One solution: use MCMC methods.
- Specifically, the SA recursion is replaced by: starting from $\theta_0 \in \Theta$,

$$\theta_{n+1} = \Pi_\Theta \left[\theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} H_{\theta_n}(Y_k^n) \right] , \tag{27}$$

where

  - $(\delta_n)_{n \in \mathbb{N}^*} \in (\mathbb{R}_+^*)^{\mathbb{N}^*}$, $(m_n)_{n \in \mathbb{N}} \in (\mathbb{N}^*)^{\mathbb{N}}$: stepsizes and batch sizes;
  - Π_Θ orthogonal projection onto Θ;
  - for any $n \in \mathbb{N}$, $(Y_k^n)_{k \in \{1,\dots,m_n\}}$ is a Markov chain with invariant distribution π_{θ_n}.

- SA (with and without MCMC) was studied in numerous papers [BMP90; FM03; DHS11; AM06; Nem+08; AFM17]
- Question: can we use inexact MCMC methods as ULA instead of exact MCMC methods?

Stochastic optimization using inexact MCMC

- More precisely, consider a family of Markov kernels

$$\{K_{\gamma,\theta}, \gamma \in (0, \bar{\gamma}) \text{ and } \theta \in \Theta\} .$$

- Assume that for any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma})$, $K_{\gamma,\theta}$ admits an invariant probability distribution $\pi_{\gamma,\theta}$.
- Assume in addition that: the bias associated to the use of this family of Markov kernels can be controlled w.r.t. to γ uniformly in θ, *i.e.*there exists $C > 0$ such that for all $\gamma \in (0, \bar{\gamma})$ and $\theta \in \Theta$,

$$|\pi_{\gamma,\theta}(H_\theta) - \pi_\theta(H_\theta)| \leq C\gamma^\tau \text{ , for } \tau > 0 .$$

- It seems reasonable to use the recursion: starting from $\theta_0 \in \Theta$,

$$\theta_{n+1} = \Pi_\Theta \left[\theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} H_{\theta_n}(X_k^n) \right] , \tag{28}$$

where

  - for any $n \in \mathbb{N}$, $(X_k^n)_{k \in \{1,\ldots,m_n\}}$ is a Markov chain with Markov kernel K_{γ_n,θ_n}, where $(\gamma_n)_{n \in \mathbb{N}}$ is a sequence of step-size.

Stochastic approximation using inexact MCMC (II)

- starting from $\theta_0 \in \Theta$,

$$\theta_{n+1} = \Pi_\Theta \left[\theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} H_{\theta_n}(X_k^n) \right] , \tag{29}$$

where

  - for any $n \in \mathbb{N}$, $(X_k^n)_{k \in \{1,\dots,m_n\}}$ is a Markov chain with Markov kernel K_{γ_n,θ_n}, where $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of step-size.

- In our applications,
  - for any $\theta \in \Theta$,

$$\pi_\theta(x) \propto \exp\left(-U_\theta(x)\right) ;$$

  - $K_{\gamma,\theta}$ stands for $\mathrm{R}_{\gamma,\theta}$ for any $\gamma \in (0, \bar{\gamma})$, $\theta \in \Theta$ where $\mathrm{R}_{\gamma,\theta}$ is associated with

$$X_{k+1}^\theta = X_k^\theta - \nabla U_\theta(X_k^\theta) + \sqrt{2\gamma} G_{k+1} .$$

Stochastic approximation using inexact MCMC (III)

- starting from $\theta_0 \in \Theta$,

$$\theta_{n+1} = \Pi_\Theta \left[\theta_n - \frac{\delta_{n+1}}{m_n} \sum_{k=1}^{m_n} H_{\theta_n}(X_k^n) \right] , \tag{30}$$

where

  - for any $n \in \mathbb{N}$, $(X_k^n)_{k \in \{1,\dots,m_n\}}$ is a Markov chain with Markov kernel K_{γ_n,θ_n}, where $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of step-size.

- Question: convergence to minimizers/minimum of f of $(\hat{\theta}_N)_{N\in\mathbb{N}^*}/(f(\hat{\theta}_N))_{N\in\mathbb{N}^*}$ where for any $N \in \mathbb{N}^*$

$$\hat{\theta}_N = \left\{ \sum_{n=1}^{N} \delta_n \theta_n \right\} \Big/ \left\{ \sum_{n=1}^{N} \delta_n \right\} \ ?$$

1. The Langevin Monte Carlo algorithm / ULA and its convergence
   - Introduction and motivations
   - Convergence of discretization of diffusions
2. Convergence of the Metropolis Adjsuted Langevin Algorithm (MALA)
3. Stochastic optimization by Langevin Monte Carlo
   - Presentation of the methodology
   - Theoretical results
   - Numerical experiments

Assumptions on Θ and f

H6

Θ is a convex compact set and $\Theta \subset \overline{B}(0, M_\Theta)$ with $M_\Theta > 0$.

H7

There exist an open set $\mathsf{U} \subset \mathbb{R}^m$ and $L_f \geq 0$ such that $\Theta \subset \mathsf{U}$ and $f \in \mathrm{C}^1(\mathsf{U}, \mathbb{R})$ is convex and for any $\theta_1, \theta_2 \in \Theta$

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leq L_f \|\theta_1 - \theta_2\| \,. \tag{31}$$

Main results increasing batch size

Theorem 8 (DDPV 19)

Assume **H**6, **H**7 *and some conditions on* $\{K_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$. *Let* $(\gamma_n)_{n\in\mathbb{N}}$, $(\delta_n)_{n\in\mathbb{N}^*}$ *be sequences of non-increasing positive real numbers and* $(m_n)_{n\in\mathbb{N}}$ *be a sequence of positive integers satisfying* $\sup_{n\in\mathbb{N}} \delta_n < 1/L_f$, $\sup_{n\in\mathbb{N}} \gamma_n < \bar{\gamma}$ *and*

$$\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty \,, \qquad \sum_{n=0}^{+\infty} \delta_{n+1}\gamma_n^{\tau} < +\infty \,, \qquad \sum_{n=0}^{+\infty} \delta_{n+1}/(m_n\gamma_n) < +\infty \,.$$

Then, $(\hat{\theta}_n)_{n\in\mathbb{N}}$ *converges a.s to some* $\theta^* \in \arg\min_\Theta f$.

Fixed batch size setting

- The conditions:

$$\sum_{n=0}^{+\infty} \delta_{n+1} = +\infty \ , \qquad \sum_{n=0}^{+\infty} \delta_{n+1}/(m_n \gamma_n) < +\infty$$

  implies that $m_n \to +\infty$ as $n \to +\infty$.

- We also have the same kind of convergence results in the setting where $m_n = m \in \mathbb{N}^*$ for any $n \in \mathbb{N}^*$ but requires additional conditions which are satisfied for ULA.

Assumptions on $\{\mathrm{R}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$

- The condition on $\{\mathrm{R}_{\gamma,\theta} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ is satisfied under the following assumption on $(\pi_\theta)_{\theta \in \Theta}$.

H8

For any $\theta \in \Theta$, there exists $U_\theta : \mathbb{R}^d \to \mathbb{R}$ such that $\pi_\theta(x) \propto \exp(-U_\theta(x))$. In addition $(\theta, x) \mapsto U_\theta(x)$ is continuous, $x \mapsto U_\theta(x)$ is differentiable for all $\theta \in \Theta$ and there exists $\mathtt{L} \geq 0$ such that for any $x, y \in \mathbb{R}^d$,

$$\sup_{\theta \in \Theta} \|\nabla_x U_\theta(x) - \nabla_x U_\theta(y)\| \leq \mathtt{L} \|x - y\| \ ,$$

and $\{\|\nabla_x U_\theta(0)\| : \theta \in \Theta\}$ is bounded.

H9

There exist $\eta > 0$ and $\mathtt{m}_1, C, M_\eta \geq 0$ such that for any $\theta \in \Theta$ and $x \in \mathbb{R}^d$,

$$\langle \nabla_x U_\theta(x), x \rangle \geq \eta \|x\| \mathbb{1}_{\mathrm{B}(0, M_\eta)^c}(x) + \mathtt{m}_1 \|\nabla_x U_\theta(x)\|^2 - C \ .$$

1. The Langevin Monte Carlo algorithm / ULA and its convergence
   - Introduction and motivations
   - Convergence of discretization of diffusions
2. Convergence of the Metropolis Adjsuted Langevin Algorithm (MALA)
3. Stochastic optimization by Langevin Monte Carlo
   - Presentation of the methodology
   - Theoretical results
   - Numerical experiments

Numerical experiments: Statistical audio compression

- Consider an n-dimensional time-discrete signal $\mathbf{z} \in \mathbb{R}^n$.
- Assume it is sparse in some dictionary $\Psi \in \mathbb{R}^{n \times j}$, i.e, $\mathbf{z} = \Psi \mathbf{x}$ with $\mathbf{x} \in \mathbb{R}^j$ is sparse.
- We assume that the observation $\mathbf{y}$ is a noisy compressed version of $\mathbf{z}$:

$$\mathbf{y} = \mathbf{M}\mathbf{z} + \mathbf{w} \ ,$$

  where $\mathbf{w}$ is Gaussian and $\mathbf{M} \in \mathbb{R}^{p \times n}$ with $p < n$ is a measurement matrix.
- We consider the prior

$$p(\mathbf{x}|\theta) \propto \exp\left(-\theta \sum_{i=1}^{d} h_\lambda(\mathbf{x}_i)\right) \ ,$$

  where h_λ is the Huber function given for any $u \in \mathbb{R}$ by

$$h_\lambda(u) = \begin{cases} u^2/2 & \text{if } |u| \leq \lambda \ , \\ \lambda(|u| - \lambda/2) & \text{otherwise .} \end{cases}$$

Numerical experiments: Statistical audio compression (II)

- The a posteriori distribution is then given by:

$$p(\mathbf{x}|\mathbf{y},\theta) \propto \exp\left(-\frac{\|\mathbf{y}-\mathbf{M}\Psi\mathbf{x}\|_2^2}{2\sigma^2} - \theta\sum_{i=1}^{d} h_\lambda(\mathbf{x}_i)\right) .$$

- $\mathbf{z}$ is retrieved here by computing the maximum-a-posteriori (MAP) estimate $\hat{\mathbf{x}}_{MAP}$ that maximises $p(\mathbf{x}|\mathbf{y})$ and then setting $\hat{\mathbf{z}}_{MAP} = \Psi\hat{\mathbf{x}}_{MAP}$:

$$\hat{\mathbf{x}}_{MAP}(\theta) \in \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\|\mathbf{y}-\mathbf{M}\Psi\mathbf{x}\|_2^2/(2\sigma^2) + \theta\sum_{i=1}^{d} h_\lambda(\mathbf{x}_i)\right\} ,$$

- The problem we face is to select the value of the hyper-parameter $\theta > 0$.
- Here we consider the maximum marginal likelihood estimator

$$\hat{\theta}_{\mathrm{MMLE}}(\theta) = \underset{\theta\in\Theta}{\operatorname{argmax}}\ p(\mathbf{y}|\theta) , \qquad \Theta = \left[0.4444, 2.22\times 10^3\right) ,$$

computed using the SA approach below, since

$$p(\mathbf{y}|\theta) = \int_{\mathbb{R}^n} p(\mathbf{x},\mathbf{y}|\theta)\ \mathrm{d}\mathbf{x} .$$

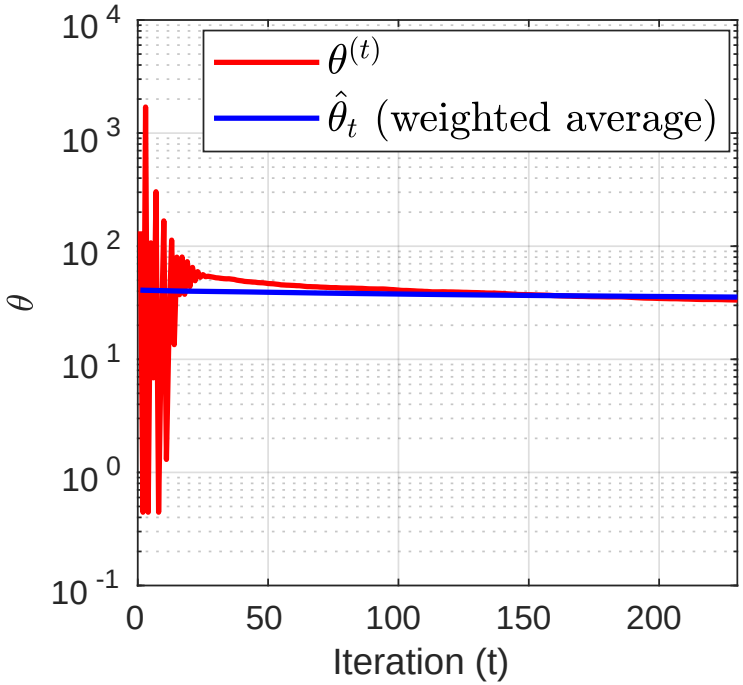
Numerical experiments: Statistical audio compression (III)

- We consider the audio compression experiment proposed in [BNE10] for the *"Mary had a little lamb"* song.
- The unknown parameter vector $\mathbf{x}$ is assumed to have dimension $d = 2900$.
- Ψ has row vectors which correspond to different piano notes.
- In the experiment proposed in [BNE10], θ is set to

$$\theta_{\text{cs}} = 0.1 \cdot \max(|(\mathbf{M}\Psi)^{\intercal} y|)/\sigma^2$$

.

- We use θ_{cs} as the initial value for θ in our algorithm.
- We use a fixed step size γ, mini-batch $m_n = 1$, and a decreasing sequence $(\delta_n)_{n\in\mathbb{N}} \propto n^{-0.8}$.



Numerical experiments: Statistical audio compression (IV)

- We compare the reconstruction for the MAP corresponding to our approximation of $\hat{\theta}_{\mathrm{MMLE}}$ and θ_{cs}.
- We consider the reconstruction mean squared error (MSE) $\|\mathbf{z} - \Psi\hat{\mathbf{x}}_{\mathrm{MAP}}\|_2$.
- θ_{MMLA} is close to the optimal value.

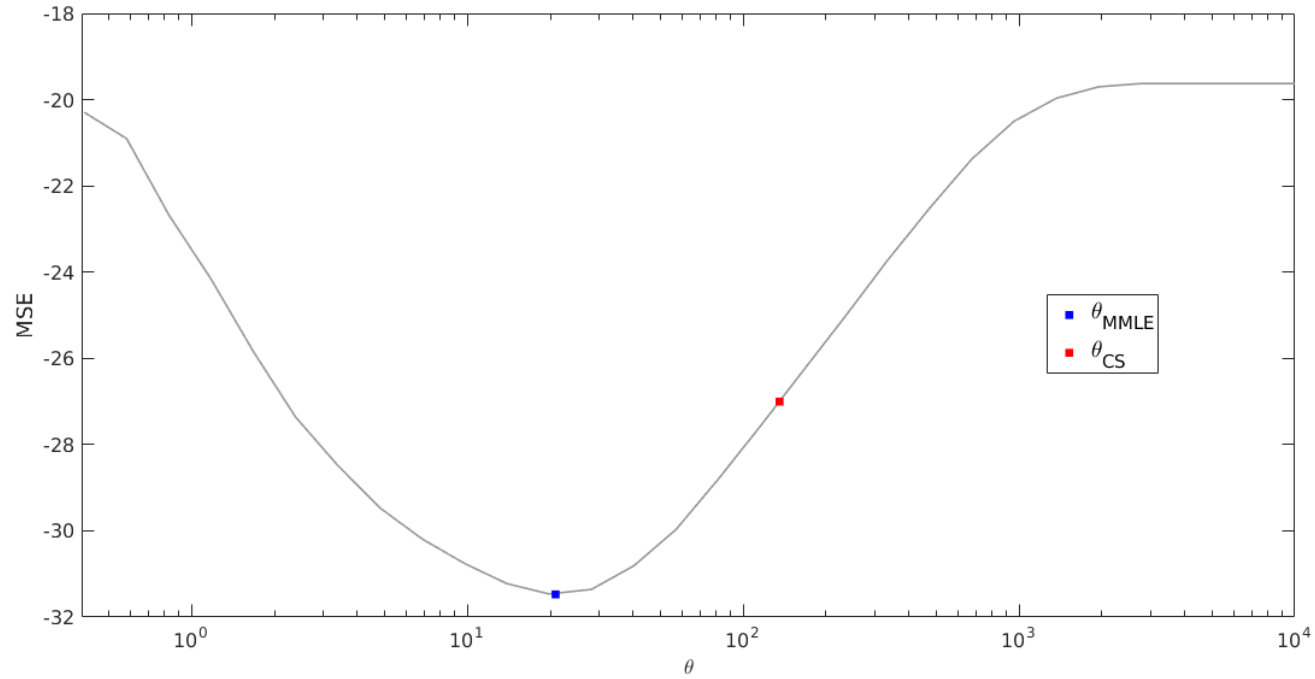


Figure: Statistical audio compression MSE for different values of the θ

Bibliography I

[AFM17] Y. F. Atchadé, G. Fort, and E. Moulines. "On perturbed proximal gradient algorithms". In: *J. Mach. Learn. Res* 18.1 (2017), pp. 310–342.

[AM06] C. Andrieu and É. Moulines. "On the ergodicity properties of some adaptive MCMC algorithms". In: *Ann. Appl. Probab.* 16.3 (2006), pp. 1462–1505. ISSN: 1050-5164.

[Bak+] Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. "Control variates for stochastic gradient MCMC". In: *Statistics and Computing* (), pp. 1–17.

[BMP90] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*. Vol. 22. Applications of Mathematics (New York). Translated from the French by Stephen S. Wilson. Springer-Verlag, Berlin, 1990, pp. xii+365. ISBN: 3-540-52894-6.

[BNE10] L. Balzano, R. Nowak, and J. Ellenberg. "Compressed sensing audio demonstration". In: (2010). URL:http://sunbeam.ece.wisc.edu/csaudio/.

[Cas85] George Casella. "An introduction to empirical Bayes data analysis". In: *Amer. Statist.* 39.2 (1985), pp. 83–87. ISSN: 0003-1305.

[Cha+18] Niladri Chatterji, Nicolas Flammarion, Yian Ma, Peter Bartlett, and Michael Jordan. "On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo". In: *International Conference on Machine Learning*. 2018, pp. 763–772.

[Cha93] K. S. Chan. "Asymptotic behavior of the Gibbs sampler". In: *J. Amer. Statist. Assoc.* 88.421 (1993), pp. 320–326. ISSN: 0162-1459.

Bibliography II

[Che+21] S. Chewi, C. Lu, K. Ahn, X. Cheng, T. Le Gouic, and P. Rigollet. "Optimal dimension dependence of the metropolis-adjusted langevin algorithm". In: *Conference on Learning Theory*. PMLR. 2021, pp. 1260–1300.

[CL00] Bradley P. Carlin and Thomas A. Louis. "Empirical Bayes: past, present and future". In: *J. Amer. Statist. Assoc.* 95.452 (2000), pp. 1286–1289. ISSN: 0162-1459.

[CT91] R. Chen and R. S. Tsay. "On the ergodicity of TAR(1) processes". In: *Ann. Appl. Probab.* 1.4 (1991), pp. 613–634. ISSN: 1050-5164.

[Dal14] A. Dalalyan. *Theoretical guarantees for approximate sampling from a smooth and log-concave density*. Submitted 1412.7392. arXiv, Dec. 2014, pp. 1–30.

[DE] Alain Durmus and Andreas Eberle. "Error bounds for inexact Markov chain Monte Carlo methods in high dimensions". In: *Arxiv* ().

[DHS11] J. Duchi, E. Hazan, and Y. Singer. "Adaptive subgradient methods for online learning and stochastic optimization". In: *J. Mach. Learn. Res.* 12 (2011), pp. 2121–2159. ISSN: 1532-4435.

[DM17] Alain Durmus and Éric Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". In: *Ann. Appl. Probab.* 27.3 (2017), pp. 1551–1587. ISSN: 1050-5164.

Bibliography III

[DM19] Alain Durmus and Éric Moulines. "High-dimensional Bayesian inference via the unadjusted Langevin algorithm". In: *Bernoulli* 25.4A (Nov. 2019), pp. 2854–2882.

[DM23] Alain Durmus and Eric Moulines. "Verifiable conditions for geometric ergodicity of MALA". In: *Accepted to Biometrika* (2023).

[DMR04] R. Douc, E. Moulines, and J. S. Rosenthal. "Quantitative bounds on convergence of time-inhomogeneous Markov chains". In: *Ann. Appl. Probab.* 14.4 (2004), pp. 1643–1665. ISSN: 1050-5164.

[Dwi+18] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. "Log-concave sampling: Metropolis-Hastings algorithms are fast!" In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 793–797.

[Ebe16] Andreas Eberle. "Reflection couplings and contraction rates for diffusions". In: *Probab. Theory Related Fields* 166.3-4 (2016), pp. 851–886. ISSN: 0178-8051.

[EGZ18] A. Eberle, A. Guillin, and R. Zimmer. "Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes". In: *Transactions of the American Mathematical Society* (2018).

[EM18] Andreas Eberle and Mateusz B Majka. "Quantitative contraction rates for Markov chains on general state spaces". In: *arXiv preprint arXiv:1808.07033* (2018).

Bibliography IV

[FM03] G. Fort and E. Moulines. "Polynomial ergodicity of Markov transition kernels". In: *Stochastic Process. Appl.* 103.1 (2003), pp. 57–99. ISSN: 0304-4149.

[For01] G. Fort. "Controle explicite d'ergodicite de chaines de Markov : Applications à l'analyse de convergence de l'algorithme Monte-Carlo EM". PhD thesis. Paris: Universite Pierre et Marie Curie, Paris, 2001.

[HM11] Martin Hairer and Jonathan C. Mattingly. "Yet another look at Harris' ergodic theorem for Markov chains". In: *Seminar on Stochastic Analysis, Random Fields and Applications VI*. Vol. 63. Progr. Probab. Birkhäuser/Springer Basel AG, Basel, 2011, pp. 109–117.

[HMS11] M. Hairer, J. C. Mattingly, and M. Scheutzow. "Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations". In: *Probab. Theory Related Fields* 149.1-2 (2011), pp. 223–259. ISSN: 0178-8051.

[HSV14] Martin Hairer, Andrew M. Stuart, and Sebastian J. Vollmer. "Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions". In: *Ann. Appl. Probab.* 24.6 (2014), pp. 2455–2490. ISSN: 1050-5164.

[Kha11] R. Khasminskii. *Stochastic stability of differential equations*. Vol. 66. Springer Science & Business Media, 2011.

[LP03] D. Lamberton and G. Pagès. "Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift". In: *Stoch. Dyn.* 3.4 (2003), pp. 435–451. ISSN: 0219-4937.

Bibliography V

[MMS18] Mateusz B Majka, Aleksandar Mijatović, and Lukasz Szpruch. "Non-asymptotic bounds for sampling algorithms without log-concavity". In: *arXiv preprint arXiv:1808.07105* (2018).

[MT92] Sean P. Meyn and R. L. Tweedie. "Stability of Markovian processes. I. Criteria for discrete-time chains". In: *Adv. in Appl. Probab.* 24.3 (1992), pp. 542–574. ISSN: 0001-8678.

[MT93] Sean P. Meyn and R. L. Tweedie. "Stability of Markovian processes. II. Continuous-time processes and sampled chains". In: *Adv. in Appl. Probab.* 25.3 (1993), pp. 487–517. ISSN: 0001-8678.

[Nea93] R. M. Neal. "Bayesian Learning via Stochastic Dynamics". In: *Advances in Neural Information Processing Systems 5, [NIPS Conference]*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, pp. 475–482. ISBN: 1-55860-274-7.

[Nem+08] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. "Robust stochastic approximation approach to stochastic programming". In: *SIAM J. Optim.* 19.4 (2008), pp. 1574–1609. ISSN: 1052-6234.

[NT78] E. Nummelin and R. L. Tweedie. "Geometric ergodicity and R-positivity for general Markov chains". In: *Ann. Probability* 6.3 (1978), pp. 404–420.

[NT82] Esa Nummelin and Pekka Tuominen. "Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory". In: *Stochastic Process. Appl.* 12.2 (1982), pp. 187–202. ISSN: 0304-4149.

Bibliography VI

[NT83] Esa Nummelin and Pekka Tuominen. "The rate of convergence in Orey's theorem for Harris recurrent Markov chains with applications to renewal theory". In: *Stochastic Process. Appl.* 15.3 (1983), pp. 295–311. ISSN: 0304-4149.

[Pop77] N. N. Popov. "Geometric ergodicity conditions for countable Markov chains". In: *Dokl. Akad. Nauk SSSR* 234.2 (1977), pp. 316–319. ISSN: 0002-3264.

[RDF78] P. J. Rossky, J. D. Doll, and H. L. Friedman. "Brownian dynamics as smart Monte Carlo simulation". In: *The Journal of Chemical Physics* 69.10 (1978), pp. 4628–4633.

[RM51] Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *Ann. Math. Statistics* 22 (1951), pp. 400–407. ISSN: 0003-4851.

[Rob85] H. Robbins. "An empirical Bayes approach to statistics". In: *Herbert Robbins Selected Papers*. Springer, 1985, pp. 41–47.

[Ros02] J. S. Rosenthal. "Quantitative convergence rates of Markov chains: a simple account". In: *Electron. Comm. Probab.* 7 (2002), pp. 123–128. ISSN: 1083-589X.

[Ros95] Jeffrey S. Rosenthal. "Minorization conditions and convergence rates for Markov chain Monte Carlo". In: *J. Amer. Statist. Assoc.* 90.430 (1995), pp. 558–566. ISSN: 0162-1459.

[RP94] G. O. Roberts and N. G. Polson. "On the Geometric Convergence of the Gibbs Sampler". In: *Journal of the Royal Statistical Society, Series B* 56 (1994), pp. 377–384.

Bibliography VII

[RR96] G. O. Roberts and J. S. Rosenthal. "Quantitative bounds for convergence rates of continuous time Markov processes". In: *Electron. J. Probab.* 1 (1996).

[RT96a] G. O. Roberts and R. L. Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* 2.4 (1996), pp. 341–363. ISSN: 1350-7265.

[RT96b] Gareth O. Roberts and Richard L. Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* 2.4 (Dec. 1996), pp. 341–363.

[TT90] D. Talay and L. Tubaro. "Expansion of the global error for numerical schemes solving stochastic differential equations". In: *Stochastic Anal. Appl.* 8.4 (1990), 483–509 (1991). ISSN: 0736-2994.