The Langevin MCMC: Theory and Methods

Alain Durmus, Eric Moulines

ENS Paris-Saclay and Ecole Polytechnique

August 8, 2019

1 Motivation

2 Metropolis-Hastings algorithm : a crash course

3 The Langevin Diffusion, ULA, MALA, SGLD

Introduction

- Sampling distribution over high-dimensional state-space has recently attracted a lot of research efforts in computational statistics and machine learning community...
- Applications (non-exhaustive)
    1. Bayesian inference for high-dimensional models,
    2. Bayesian inverse problems (e.g., image restoration and deblurring),
    3. Aggregation of estimators and experts,
    4. Bayesian non-parametrics.
- Most of the sampling techniques known so far do not scale to high-dimension and become impractical when the number of data samples is huge (big data)... Challenges are numerous in this area...

Logistic and probit regression

- Likelihood: Binary regression set-up in which the binary observations (responses) $\{Y_i\}_{i=1}^n$ are conditionally independent Bernoulli random variables with success probability $\{F(\boldsymbol{\beta}^T X_i)\}_{i=1}^n$, where
  1. X_i is a d dimensional vector of known covariates,
  2. $\boldsymbol{\beta}$ is a d dimensional vector of unknown regression coefficient
  3. F is the link function.
- Two important special cases:
  1. probit regression: F is the standard normal cumulative distribution function,
  2. logistic regression: F is the standard logistic cumulative distribution function:

$$F(t) = \mathrm{e}^t/(1+\mathrm{e}^t)$$

Bayes 101

- Bayesian analysis requires a prior distribution for the unknown regression parameter

$$\pi(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right) \quad \text{or} \quad \pi(\boldsymbol{\beta}) = \exp\left(-\sum_{i=1}^{d} \alpha_i|\beta_i|\right)$$

.

- The posterior of $\boldsymbol{\beta}$ is up to a proportionality constant given by

$$\pi(\boldsymbol{\beta}|(Y,X)) \propto \prod_{i=1}^{n} F^{Y_i}(\beta' X_i)(1 - F(\beta' X_i))^{1-Y_i}\pi(\boldsymbol{\beta})$$

New challenges

- Problem the number of predictor variables d is large (10^4 and up) the number of observations is $10^5 - 10^6$.
- Examples
  - text categorization,
  - matrix factorization,
  - genomics and proteomics (gene expression analysis),
  - other data mining tasks (recommendations, longitudinal clinical trials, ..).

A daunting problem ?

- For Gaussian prior (ridge regression), the potential U is smooth strongly convex.
- For Laplace prior (Lasso our fused Lasso) regression, the potential U is non-smooth but still convex...
- A wealth of efficient optimisation algorithms are now available to solve this problem in very high-dimension...
- (long term) Objective:
  - Contribute to fill the gap between optimization and simulation. Good optimization methods are in general a good source of inspiration to design efficient sampler.
  - Develop algorithms converging to the target distribution polynomially with the dimension (more precise statements below)

1 Motivation

2 Metropolis-Hastings algorithm : a crash course

3 The Langevin Diffusion, ULA, MALA, SGLD

The Metropolis-Hastings algorithm (I)

The Metropolis-Hastings algorithm gives a generic method to build Markov kernels P reversible w.r.t. the stationary distribution π.

Assumption: π has a density w.r.t. to some σ-finite measure ν. The density denoted h_π is everywhere positive (sometimes, we will use π for the density).

Transition density $q(x, y)$ w.r.t. ν :

1. $(x, y) \mapsto q(x, y)$ is measurable,
2. For all $x \in \mathbb{R}^d$, $y \mapsto q(x, y)$ is a probability density function.

The Metropolis-Hastings algorithm

Given X_k,

1. Sample a candidate $Y_{k+1} \sim q(X_k, \cdot)$.
2. Accept or reject his candidate:

$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with probability } \alpha(X_k, Y_{k+1})\,, \\ X_k & \text{with probability } 1 - \alpha(X_k, Y_{k+1})\,. \end{cases}$$

where

$$\boxed{\alpha(x,y) = 1 \wedge \frac{h_\pi(y)}{h_\pi(x)} \frac{q(y,x)}{q(x,y)}}\,.$$

Reversibility

Definition

Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. A σ-finite measure ξ on $\mathcal{X}$ is said to be reversible with respect to P if the measure $\xi \otimes P$ on $\mathcal{X} \otimes \mathcal{X}$ is symmetric, *i.e.* for all $(A, B) \in \mathcal{X} \times \mathcal{X}$

$$\xi \otimes P(A \times B) = \xi \otimes P(B \times A) ,$$

where for each $C \in \mathcal{X} \times \mathcal{X}$,

$$\xi \otimes P(C) = \iint \xi(\mathrm{d}x) P(x, \mathrm{d}x') \mathbb{1}_C(x, x')$$

.

Reversibility

- Reversibility means that for all bounded measurable functions f defined on $(\mathsf{X} \times \mathsf{X}, \mathcal{X} \otimes \mathcal{X})$,

$$\iint_{\mathsf{X}\times\mathsf{X}} \xi(\mathrm{d}x) P(x, \mathrm{d}x') f(x, x') = \iint_{\mathsf{X}\times\mathsf{X}} \xi(\mathrm{d}x) P(x, \mathrm{d}x') f(x', x) .$$

- If X is a countable state space, a (finite or σ-finite) measure ξ is reversible w.r.t. P if and only if, for all $(x, x') \in \mathsf{X} \times \mathsf{X}$,

$$\xi(x) P(x, x') = \xi(x') P(x', x) ,$$

a condition often referred to as the detailed balance condition.

Reversibility

If $\{X_k,\ k \in \mathbb{N}\}$ is a Markov chain with kernel P and initial distribution ξ, the reversibility condition precisely means that (X_0, X_1) and (X_1, X_0) have the same distribution, *i.e.* for all $f \in \mathbb{F}_b(\mathsf{X} \times \mathsf{X}, \mathcal{X} \otimes \mathcal{X})$,

$$\begin{aligned}\mathbb{E}_\xi[f(X_0, X_1)] &= \iint \xi(\mathrm{d}x_0) P(x_0, \mathrm{d}x_1) f(x_0, x_1) \\ &= \iint \xi(\mathrm{d}x_0) P(x_0, \mathrm{d}x_1) f(x_1, x_0) = \mathbb{E}_\xi[f(X_1, X_0)] \,.\end{aligned}$$

- The distribution of X_1 is the same as that of X_0: ξ is P-invariant.
- For all $n \in \mathbb{N}$,

$$\mathbb{E}_\xi[f(X_0, \ldots, X_n)] = \mathbb{E}_\xi[f(X_n, \ldots, X_0)].$$

Reversibility: recap

Theorem

Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$ and $\xi \in \mathbb{M}_1(\mathcal{X})$. If ξ is reversible w.r.t. P, then

- **(i)** *ξ is P-invariant*
- **(ii)** *the homogeneous Markov chain $\{X_k,\ k \in \mathbb{N}\}$ with Markov kernel P and initial distribution ξ is reversible, i.e. for any $n \in \mathbb{N}$, $(X_0, \ldots, X_n)$ and $(X_n, \ldots, X_0)$ have the same distribution.*

Reversibility of the Metropolis-Hastings kernel

The Metropolis-Hastings algorithm produces a Markov chain, $\{X_k,\ k \in \mathbb{N}\}$, with Markov kernel P given by

$$P(x, A) = \int_A \alpha(x, y) q(x, y) \nu(\mathrm{d}y) + \bar{\alpha}(x) \delta_x(A) ,$$

with

$$\bar{\alpha}(x) = \int_{\mathsf{X}} \{1 - \alpha(x, y)\} q(x, y) \nu(\mathrm{d}y) .$$

The quantity $\bar{\alpha}(x)$ is the probability of remaining at the same point.

Theorem

The distribution π is reversible w.r.t. the Metropolis-Hastings kernel P.

Reversibility of the MH-Proof 1

Note first that for every $x, y \in \mathsf{X}$, it holds that

$$\begin{aligned} h_\pi(x)\alpha(x,y)q(x,y) &= \{h_\pi(x)q(x,y)\} \wedge \{h_\pi(y)q(y,x)\} \\ &= h_\pi(y)\alpha(y,x)q(y,x) \; . \end{aligned}$$

Thus for $C \in \mathcal{X} \times \mathcal{X}$,

$$\iint h_\pi(x)\alpha(x,y)q(x,y)\mathbb{1}_C(x,y)\nu(\mathrm{d}x)\nu(\mathrm{d}y) \\ = \iint h_\pi(y)\alpha(y,x)q(y,x)\mathbb{1}_C(x,y)\nu(\mathrm{d}x)\nu(\mathrm{d}y) \; .$$

Reversibility of the MH-Proof 2

1. On the other hand,

$$\iint h_\pi(x)\delta_x(\mathrm{d}y)\bar{\alpha}(x)\mathbb{1}_C(x,y)\nu(\mathrm{d}x)$$
$$= \int h_\pi(x)\bar{\alpha}(x)\mathbb{1}_C(x,x)\nu(\mathrm{d}x) = \int h_\pi(y)\bar{\alpha}(y)\mathbb{1}_C(y,y)\nu(\mathrm{d}y)$$
$$= \iint h_\pi(y)\delta_y(\mathrm{d}x)\bar{\alpha}(y)\mathbb{1}_C(x,y)\nu(\mathrm{d}y)\;.$$

2. Hence, summing up the two terms

$$\iint h_\pi(x)P(x,\mathrm{d}y)\nu(\mathrm{d}x)\mathbb{1}_C(x,y) = \iint h_\pi(y)P(y,\mathrm{d}x)\mathbb{1}_C(x,y)\nu(\mathrm{d}y)\;.$$

3. Conclusion: π is reversible w.r.t. P.

Symmetric Random walk Metropolis-Hastings (I)

- The idea in the RWM is to propose local moves around the current states.
- The proposal mechanism is given by

  $$Y_{k+1} = X_k + Z_{k+1} \ ,$$

  where Z_{k+1} is independent of X_k and is distributed according to a probability measure with a symmetric probability density function $\tilde{q}$.
- The proposal distribution is of the form $q(x, y) = \tilde{q}(y - x)$.

Symmetric Random walk Metropolis-Hastings (II)

1. Generate Z_{k+1} from $\tilde{q}$ and set $Y_{k+1} = X_k + Z_{k+1}$.
2. Set
$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with probability } \alpha(X_k, Y_{k+1}) \,, \\ X_k & \text{with probability } 1 - \alpha(X_k, Y_{k+1}) \,. \end{cases}$$

   where
$$\boxed{\alpha(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)}} \,.$$

MH: properties

Simple condition to apply the Ergodic theorem:

- q and π are continuous.
- For all x, y such that $\pi(y) > 0$, $q(x, y) > 0$.

- P is irreducible and therefore admits a unique invariant distribution
- The Markov kernel P is ergodic: for any $f \in \mathrm{L}^1(\pi)$,

$$\hat{\pi}_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \underset{a.s.}{\longrightarrow} \int f(x) \pi(\mathrm{d}x) \, .$$

- If $f \in \mathrm{L}^2(\pi)$, the $n^{-1/2}(\hat{\pi}_n(f) - \pi(f))$ satisfies a Central Limit Theorem provided that the series $\sum_{k=0}^{\infty} \mathrm{Cov}_\pi(f, P^k f)$ is converging.

Total Variation distance

Definition (Total variation distance)

Let ξ be a finite signed measure on $(\mathsf{X}, \mathcal{X})$ with Jordan decomposition (ξ^+, ξ^-). The total variation norm of ξ is defined by

$$\|\xi\|_{\mathrm{TV}} = |\xi|(\mathsf{X}) \; .$$

The total variation distance between two probability measures $\xi, \xi' \in \mathbb{M}_1(\mathsf{X})$ is defined by

$$d_{\mathrm{TV}}(\xi, \xi') = \frac{1}{2}\|\xi - \xi'\|_{\mathrm{TV}} \; .$$

Some properties of the total variation distance

Theorem

For $\xi \in \mathbb{M}_{\mathrm{s}}(\mathcal{X})$*,*

$$\|\xi\|_{\mathrm{TV}} = \sup\left\{\xi(f) \,:\, f \in \mathbb{F}_b(\mathsf{X}), |f|_\infty \leq 1\right\} \;.$$

If ξ *is a finite signed measure such that* $\xi(\mathsf{X}) = 0$*, then,*

$$\|\xi\|_{\mathrm{TV}} = 2\sup\left\{\xi(f) \,:\, f \in \mathbb{F}_b(\mathsf{X}),\ \mathrm{osc}\,(f) \leq 1\right\} \;.$$

where $\mathrm{osc}\,(f) = \sup_{(x,x') \in \mathsf{X}\times\mathsf{X}} |f(x) - f(x')|$

we consider a function $f \in \mathbb{F}(\mathsf{X})$ taking values in $[1, \infty]$. We denote $D_f = \{x \in \mathsf{X} : f(x) < \infty\}$.

Definition (f-norm)

The space of finite signed measures ξ such that $|\xi|(f) < \infty$ is denoted by $\mathbb{M}_f(\mathcal{X})$.

(i) The f-norm of a measure $\xi \in \mathbb{M}_f(\mathcal{X})$ is $\|\xi\|_f = |\xi|(f)$.

(ii) The f-norm of a function $h \in \mathbb{F}(\mathsf{X})$ is

$$|h|_f = \sup_{x \in D_f} \frac{|h(x)|}{f(x)} .$$

(iii) The f-oscillation of a function $h \in \mathbb{F}(\mathsf{X})$ is

$$\operatorname{osc}_f(h) \overset{\text{def}}{=} \sup_{(x,x') \in D_f \times D_f} \frac{|h(x) - h(x')|}{f(x) + f(x')} .$$

f-norm

Theorem

For $\xi \in \mathbb{M}_f(\mathcal{X})$,

$$\|\xi\|_f = \sup\left\{\xi(h) \, : \, h \in \mathbb{F}_b(\mathsf{X}), |h|_f \leq 1\right\} \; .$$

Let $\xi \in \mathbb{M}_0(\mathcal{X}) \cap \mathbb{M}_f(\mathcal{X})$. *Then,*

$$\|\xi\|_f = \sup\left\{\xi(h) \, : \, \operatorname{osc}_f(h) \leq 1\right\} \; .$$

Definition (f-geometric ergodicity)

Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$ and $f : \mathsf{X} \to [1, \infty)$ be a measurable function. The Markov kernel P is said to be f-geometrically ergodic if P is irreducible, positive with invariant probability π and if there exist

(i) a measurable function $M : \mathsf{X} \to [0, \infty]$ such that $\pi(\{M < \infty\}) = 1$

(ii) a measurable function $\beta : \mathsf{X} \to [1, \infty)$ such that $\pi(\{\beta > 1\}) = 1$

satisfying for all $n \in \mathbb{N}$ and $x \in \mathsf{X}$,

$$\beta^n(x) \left\| P^n(x, \cdot) - \pi \right\|_f \leq M(x) \ .$$

If $f \equiv 1$, we say that P is geometrically ergodic instead of 1-geometrically ergodic.

f-Geometric recurrence

Definition

Let $f : \mathsf{X} \to [1, \infty)$ be a measurable function and $\delta > 1$. A set $C \in \mathcal{X}$ is said to be (f, δ)-geometrically recurrent if

$$\sup_{x \in C} \mathbb{E}_x \left[\sum_{k=0}^{\sigma_C - 1} \delta^k f(X_k) \right] < \infty .$$

- The set C is said to be f-geometrically recurrent if it is (f, δ)-geometrically recurrent for some $\delta > 1$.
- The set C is said to be geometrically recurrent if it is f-geometrically recurrent for some $f \geq 1$.

f-geometrically recurrence

Theorem

Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$, $C \in \mathcal{X}$, $\delta > 1$ and $f : \mathsf{X} \to [1, \infty)$ be a measurable function. The following conditions are equivalent.

(i) *The set C is (f, δ)-geometrically recurrent.*

(ii) *There exists a measurable function $V : \mathsf{X} \to [0, \infty]$ and $b \in [0, \infty)$ such that*

$$PV + f \leq \delta^{-1} V + b \mathbb{1}_C \ ,$$

and $\sup_{x \in C} V(x) < \infty$.

Small sets

Definition (Small Set)

Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$. A set $C \in \mathcal{X}$ is called a small set if there exist $m \in \mathbb{N}^*$ and a non-zero measure $\mu \in \mathbb{M}_+(\mathcal{X})$ such that for all $x \in C$ and $A \in \mathcal{X}$,

$$P^m(x, A) \geq \mu(A) \; .$$

The set C is then said to be an (m, μ)-small set.

The definition entails that μ is a finite measure and $0 < \mu(\mathsf{X}) \leq 1$. Hence it can be written $\mu = \epsilon\nu$ with $\epsilon = \mu(\mathsf{X})$ and ν is a probability measure.

Small sets

Lemma

$$P(x,A) \geq \int_A t(x,y) \mathrm{Leb}_d(\mathrm{d}y) \;, \qquad A \in \mathcal{B}\left(\mathbb{R}^d\right) \;,$$

where t is a positive l.s.c function on $\mathbb{R}^d \times \mathbb{R}^d$. Then every compact set C with positive Lebesgue measure is small.

f-geometric ergodicity

Theorem

Let P be an irreducible aperiodic Markov kernel on $\mathsf{X} \times \mathcal{X}$ and $f : \mathsf{X} \to [1, \infty)$ be a measurable function. Assume that there exist a function $V : \mathsf{X} \to [0, \infty]$ such that $\{V < \infty\} \neq \emptyset$, a non empty small set C, $\lambda \in [0, 1)$ and $b < \infty$ such that

$$PV + f \leq \lambda V + b\mathbb{1}_C \;;$$

Then, denoting by π the invariant probability measure, there exist a set $S \in \mathcal{X}$ such that $\pi(S) = 1$, $\{V < \infty\} \subset S$, and $\beta > 1$ such that for all $x \in S$,

$$\sum_{n=0}^{\infty} \beta^n \left\| P^n(x, \ \cdot) - \pi \right\|_f \leq \kappa\{V(x) + 1\} \,.$$

Ergodicity of RWM

- Under mild assumptions the RWM is geometrically ergodic with respect to π (see Jarner and Hansen, 2000).
- However, the constants κ and β are very conservative.
- In practice, converges very slowly when the dimension becomes large.

Data Augmentation-I

- Assume that $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ are Polish spaces equipped with their Borel σ-fields. Again, we wish to simulate from a probability measure π defined on $(\mathsf{X}, \mathcal{X})$ using a sequence $\{X_k,\ k \in \mathbb{N}\}$ of X-valued random variables.
- Data augmentation algorithms consist in writing the target distribution π as the marginal of the distribution π^* on the product space $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ defined by $\pi^* = \pi \otimes R$ where R is a kernel on $\mathsf{X} \times \mathcal{Y}$.

Data Augmentation-II

- There exists also a kernel S on $\mathsf{Y} \times \mathcal{X}$ and a probability measure $\tilde{\pi}$ on $(\mathsf{Y}, \mathcal{Y})$ such that $\pi^*(C) = \iint \mathbb{1}_C(x, y)\tilde{\pi}(\mathrm{d}y)S(y, \mathrm{d}x)$ for $C \in \mathcal{X} \otimes \mathcal{Y}$.
- In other words, if (X, Y) is a pair of random variables with distribution π^* then $R(x, \cdot)$ is the distribution of Y conditionally on $X = x$ and $S(y, \cdot)$ is the distribution of X conditionally on $Y = y$.
- The bivariate distribution π^* can then be expressed as follows

$$\pi^*(\mathrm{d}x\mathrm{d}y) = \pi(\mathrm{d}x)R(x, \mathrm{d}y) = S(y, \mathrm{d}x)\tilde{\pi}(\mathrm{d}y) \; .$$

Data Augmentation-III

- A data augmentation algorithm consists in running a Markov Chain $\{(X_k, Y_k),\ k \in \mathbb{N}\}$ with invariant probability π^* and to use $n^{-1} \sum_{k=0}^{n-1} f(X_k)$ as an approximation of $\pi(f)$.
- The transition from (X_k, Y_k) to (X_{k+1}, Y_{k+1}) is decomposed into two successive steps: Y_{k+1} is first drawn given (X_k, Y_k) and then X_{k+1} is drawn given (X_k, Y_{k+1}).
- Intuitively, Y_{k+1} can be used as an auxiliary variable, which directs the moves of X_k toward interesting regions with respect to the target distribution.

Data Augmentation

When sampling from R and S is feasible, a classical choice consists in following the two successive steps: given (X_k, Y_k),

(i) sample Y_{k+1} from $R(X_k, \cdot)$,

(ii) sample X_{k+1} from $S(Y_{k+1}, \cdot)$.

$$\begin{tikzcd}
X_k \arrow[dr, "R"] & X_{k+1} \\
Y_k \arrow[u, "S"] & Y_{k+1} \arrow[u, "S"']
\end{tikzcd}$$

Figure: In this example, sampling from R and S is feasible.

Theorem

$\{X_k,\ k \in \mathbb{N}\}$ is a Markov chain with Markov kernel RS and π is reversible w.r.t. RS.

Data Augmentation: proof

Proof.

For $A, B \in \mathcal{X}$,

$$
\begin{aligned}
&\pi \otimes RS(A \times B) \\
&\quad = \int_{\mathsf{X}\times\mathsf{Y}} \pi(\mathrm{d}x) R(x, \mathrm{d}y) \mathbb{1}_A(x) S(y, B) = \int_{\mathsf{X}\times\mathsf{Y}} \mathbb{1}_A(x) S(y, B) \pi^*(\mathrm{d}x\mathrm{d}y) \\
&\quad = \int_{\mathsf{X}\times\mathsf{Y}} \mathbb{1}_A(x) S(y, B) S(y, \mathrm{d}x) \tilde{\pi}(\mathrm{d}y) = \int_{\mathsf{Y}} S(y, A) S(y, B) \tilde{\pi}(\mathrm{d}y) \;.
\end{aligned}
$$

□

State of the art

The most popular algorithms for Bayesian inference in binary regression models are based on data augmentation

- Instead on sampling $\pi(\boldsymbol{\beta}|(X,Y))$ sample $\pi(\boldsymbol{\beta},W|(X,Y))$ probability measure on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and take the marginal w.r.t. $\boldsymbol{\beta}$.
- Typical application of the Gibbs sampler: sample in turn $\pi(\boldsymbol{\beta}|(X,Y,W))$ and $\pi(W|(X,Y,\boldsymbol{\beta}))$.
- The choice of the DA should make these two steps reasonably easy...
  - probit link: Albert and Chib (1993).
  - logistic link: Polya-Gamma sampler, Polsson and Scott (2012)... !

State of the art: shortcomings

- The state of the art samplers have been shown to be uniformly geometrically ergodic (more on this later): there exists a constant C such that

  $$\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq C\rho^n \quad \text{for any } x \in \mathsf{X}$$

  BUT $1-\rho$ is exponentially small with the dimension d
- The algorithms are very demanding in terms of computational resources...
  - applicable only when is d small 10 to moderate 100 but certainly not when d is large (10^4 or more).
  - convergence rate is extremely slow for $d \geq 10^2$.

1 Motivation

2 Metropolis-Hastings algorithm : a crash course

3 The Langevin Diffusion, ULA, MALA, SGLD

Framework

- Denote by π a target density w.r.t. the Lebesgue measure on $\mathbb{R}^d$, known up to a normalisation factor

$$x \mapsto \pi(x) \overset{\text{def}}{=} \mathrm{e}^{-U(x)} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y \ ,$$

  Implicitly, $d \gg 1$.

- Assumption: U is L-smooth : twice continuously differentiable and there exists a constant L such that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla U(x) - \nabla U(y)\| \leq L \|x - y\| \ .$$

(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:

$$P_t(x,A) = \mathbb{P}(X_t \in A | X_0 = x) \ , \quad x \in \mathbb{R}^d, A \in \mathcal{B}(\mathbb{R}^d) \ .$$

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

Ergodicity

- Key property 1: For all $x \in \mathbb{R}^d$,

$$\lim_{t \to +\infty} \|\delta_x P_t - \pi\|_{\mathrm{TV}} = 0 \ .$$

- Key property 2: for "nice" functions

$$\frac{1}{T} \int_0^T f(X_t) \mathrm{d}t \xrightarrow{\mathbb{P}_x - \text{a.s.}} \pi(f) = \int \pi(\mathrm{d}x) f(x)$$
$$\frac{1}{\sqrt{T}} \int_0^T \{f(X_t) - \pi(f)\} \mathrm{d}t \overset{\mathbb{P}_x}{\Longrightarrow} \mathcal{N}(0, \sigma^2(\pi, f)) \ .$$

- The Langevin diffusion provides a mean to sample any smooth distribution... Of course, this is a highly theoretical solution...

Discretized Langevin diffusion

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

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

  where

  - $(Z_k)_{k \geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$
  - $(\gamma_k)_{k \geq 1}$ is a sequence of stepsizes, which can either be held constant or be chosen to decrease to 0 at a certain rate.

- Closely related to the (stochastic) gradient descent algorithm.

Discretized Langevin diffusion: constant stepsize

- When the stepsize is held constant, *i.e.* $\gamma_k = \gamma$, then $(X_k)_{k \geq 1}$ is an homogeneous Markov chain with Markov kernel R_γ
- Under some appropriate conditions, this Markov chain is irreducible, positive recurrent $\rightsquigarrow$ unique invariant distribution 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 : $\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \epsilon$
  - Is there a way to choose the starting point x cleverly ?
  - Auxiliary question: quantify the distance between π_γ and π.

Discretized Langevin diffusion: decreasing stepsize

- When $(\gamma_k)_{k\geq 1}$ is nonincreasing and non constant, $(X_k)_{k\geq 1}$ is an inhomogeneous Markov chain associated with the kernels $(R_{\gamma_k})_{k\geq 1}$.
- Notation: Q_γ^p is the composition of Markov kernels

$$Q_\gamma^p = R_{\gamma_1} R_{\gamma_2} \dots R_{\gamma_p}$$

With this notation, $\mathbb{E}_x[f(X_p)] = \delta_x Q_\gamma^p f$.

- Questions:
  - Convergence : is there a way to choose the step sizes so that $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \to 0$ and if yes, what is the optimal way of choosing the stepsizes ?...
  - Optimal choice of simulation parameters : What is the number of iterations required to reach a neighborhood of the target: $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \epsilon$ starting from a given point x
  - Should we use fixed or decreasing step sizes ?

Metropolis-Adjusted Langevin Algorithm

- To correct the bias in the stationary distribution induced by the discretization (the discretization targets π_γ instead of π), a Metropolis-Hastings step can be included $\rightsquigarrow$ Metropolis Adjusted Langevin Agorithm (MALA).
  - Key reference: Roberts and Tweedie, 1996; several improvements have been considered since then.
- Algorithm:
  1. Propose $Y_{k+1} \sim X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1}$, $Z_{k+1} \sim \mathcal{N}(0, \mathrm{I}_d)$
  2. Compute the acceptance ratio $\alpha_\gamma(X_k, Y_{k+1})$:

$$\alpha_\gamma(x,y) = 1 \wedge \frac{\pi(y) r_\gamma(y,x)}{\pi(x) r_\gamma(x,y)} \; , r_\gamma(x,y) \propto \mathrm{e}^{-\|y-x-\gamma\nabla U(x)\|^2/(4\gamma)}$$

  3. Accept / Reject the proposal.

MALA: pros and cons

- Require to compute one gradient at each iteration and to evaluate one time the objective function
- Geometric convergence is established under the condition that in the tail the acceptance region is inwards in q,

$$\lim_{\|x\|\to\infty} \int_{\mathcal{A}_\gamma(x)\Delta\mathcal{I}(x)} r_\gamma(x,y)\mathrm{d}y = 0 \; .$$

where $\mathcal{I}(x) = \{y, \|y\| \leq \|x\|\}$ and $A_\gamma(x)$ is the acceptance region

$$\mathcal{A}_\gamma(x) = \{y, \pi(x) r_\gamma(x,y) \leq \pi(y) r_\gamma(y,x)\}$$

- Even in simple settings, checking this assumption is non-trivial (see **?**)

Stochastic Gradient Langevin Algorithm

- We shall consider the situation where the traget density π is the density of the posterior distribution under a Bayesian model where there are $N \gg 1$ i.i.d. observations, the so-called Big Data regime

$$\pi(x) \propto \mathrm{p}_0(x) \prod_{i=1}^{N} \mathrm{p}_i(y_i|x)$$

- Here, both computing the gradient term $\nabla \log \pi(x)$ and evaluating the Metropolis-Hastings ratio require a computational budget that scales unfeasibly as $O(N)$.
- Approach #1: use a standard MH and and try to approximate the Metropolis-Hastings accept-reject ratio using only a subset of data (see Bardenet et al, 2014-2017)... not very convincing.

Stochastic Gradient Langevin Algorithm

- The stochastic gradient Langevin dynamics (SGLD) is an alternative approach proposed by Welling and Teh (2011).
- This follows the opposite route and chooses to completely avoid the computation of the Metropolis-Hastings ratio.
- By choosing a discretization of the Langevin diffusion with a sufficiently small step-size, because the Langevin diffusion is ergodic with respect to π, the hope is that even if the Metropolis-Hastings accept-reject mechanism is completely avoided, the resulting Markov chain still has an invariant distribution that is close to the target posterior
- To further make this approach viable in large N settings, the gradient term $\nabla \log \pi$ can be further approximated using a subsampling strategy.

Stochastic Gradient Langevin Algorithm

- For an integer $n \in \{1, \ldots, N\}$ and a random subset $\tau = (\tau_1, \ldots, \tau_n)$ of $[N] = \{1, \ldots, N\}$ generated by sampling with or without replacement from $[N]$, the quantity

$$\nabla \log \mathrm{p}_0(x) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log \mathrm{p}_i(y_i|x)$$

  is an unbiased estimator of the gradient
- Most importantly, this stochastic estimate can be computed with a computational budget that scales as $O(n)$ with n potentially much smaller that N.
- Stochastic gradient method have a long history in optimisation and machine learning and are especially relevant in the large dataset regime.

A generalized SGLD

- Observations: At each iteration, we observe a noisy value of the gradient

$$Y_k = \nabla U(X_k) + \zeta_k$$

where $\{\zeta_k,\ k \in \mathbb{N}\}$ is an adapted sequence

- Noisy ULA:

$$X_{k+1} = X_k - \gamma_{k+1} Y_k + \sqrt{2\gamma_{k+1}} Z_{k+1}$$

- Typical assumptions:
  1. Independence Z_{k+1} is independent of $\sigma((\zeta_\ell, X_\ell), \ell \leq k)$
  2. Bounded variance $\mathbb{E}[\|\zeta_k - \mathbb{E}^{X_k}[\zeta_k]\|^2] \leq \sigma^2 d$
  3. Unbiasedness

Some (very incomplete) references: Early references

(i) **Statistical physics**: Parisi, 1981, *Correlation function and Computer Simulations*, Nuclear Physics.

(ii **Bayesian statistics**: Grenander and Miller (in discussion Besag, *Representation of knowledge in Complex Systems*, JRSS B). First theoretical results given by Roberts and Tweedie, 1996, *Exponential Convergence of Langevin Distributions and Their Discrete Approximations*, Bernoulli, Stramer and Tweedie, *Langevin-type models. I. Diffusions with given stationary distributions and their discretizations.*, MCAP, 1999

(ii Most of these results are qualitative (e.g. conditions upon which the sampler is geometrically ergodic).

Some (very incomplete) references: Euler discretisation

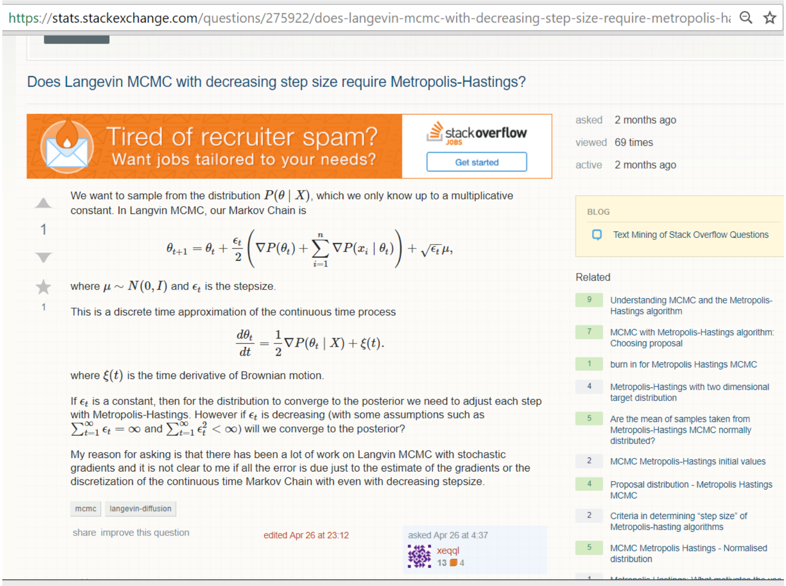
(i) Lamberton, D.; Pagès, G. Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift. Stoch. Dyn. 3 (2003), no. 4, 435–451.

(ii Lemaire, Vincent Behavior of the Euler scheme with decreasing step in a degenerate situation. ESAIM Probab. Stat. 11 (2007), 236–247.

(ii Lemaire, V.; Menozzi, S. On some non asymptotic bounds for the Euler scheme. Electron. J. Probab. 15 (2010), no. 53, 1645–1681.

Some (very incomplete) references: recent attempts in Machine Learning

1. Welling, M. and Teh, Y.W., *Bayesian learning via stochastic gradient Langevin dynamic*, ICML, 2011
2. Vollmer, S.; Zygalakis, K.; Teh, Y. W. *Exploration of the (non-)asymptotic bias and variance of stochastic gradient Langevin dynamics*, J. Mach. Learn. Res. 17 (2016),
3. Teh, Y.W.; Thiery, A.; Vollmer, S., *Consistency and fluctuations for stochastic gradient Langevin dynamics* J. Mach. Learn. Res. 17 (2016)
4. Dalalyan, A., *Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent*, Proc. of the 2017 Conference on Learning Theory, volume 65



https://stats.stackexchange.com/questions/275922/does-langevin-mcmc-with-decreasing-step-size-require-metropolis-ha

Does Langevin MCMC with decreasing step size require Metropolis-Hastings?

asked	2 months ago
viewed	69 times
active	2 months ago

1

1

We want to sample from the distribution $P(\theta \mid X)$, which we only know up to a multiplicative constant. In Langvin MCMC, our Markov Chain is

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2}\left(\nabla P(\theta_t) + \sum_{i=1}^{n} \nabla P(x_i \mid \theta_t)\right) + \sqrt{\epsilon_t}\mu,$$

where $\mu \sim N(0, I)$ and ϵ_t is the stepsize.

This is a discrete time approximation of the continuous time process

$$\frac{d\theta_t}{dt} = \frac{1}{2}\nabla P(\theta_t \mid X) + \xi(t).$$

where $\xi(t)$ is the time derivative of Brownian motion.

If ϵ_t is a constant, then for the distribution to converge to the posterior we need to adjust each step with Metropolis-Hastings. However if ϵ_t is decreasing (with some assumptions such as $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\sum_{t=1}^{\infty} \epsilon_t^2 < \infty$) will we converge to the posterior?

My reason for asking is that there has been a lot of work on Langvin MCMC with stochastic gradients and it is not clear to me if all the error is due just to the estimate of the gradients or the discretization of the continuous time Markov Chain with even with decreasing stepsize.

edited Apr 26 at 23:12

asked Apr 26 at 4:37
xeqql
13 ● 4

BLOG

Text Mining of Stack Overflow Questions

Related

9	Understanding MCMC and the Metropolis-Hastings algorithm
7	MCMC with Metropolis-Hastings algorithm: Choosing proposal
1	burn in for Metropolis Hastings MCMC
4	Metropolis-Hastings with two dimensional target distribution
5	Are the mean of samples taken from Metropolis-Hastings MCMC normally distributed?
2	MCMC Metropolis-Hastings initial values
4	Proposal distribution - Metropolis Hastings MCMC
2	Criteria in determining “step size” of Metropolis-hasting algorithms
5	MCMC Metropolis Hastings - Normalised distribution