Borne de convergence explicite pour l'algorithme Langevin Monte Carlo cinétique.

Aurélien Enfroy[1]
Joint work with: Alain Durmus[2], Eric Moulines[2],

[1]Télécom SudParis

[2]ENS Paris-Saclay

[3]Ecole Polytechnique

Outline

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... Challenges are numerous in this area...

Bayesian setting

- A Bayesian model is specified by
  1. the sampling distribution of the observed data conditional on its parameters, often termed likelihood: $Y \sim \mathsf{L}(\cdot|\theta)$
  2. a prior distribution π_0 on the parameter space $\theta \in \mathbb{R}^d$
- The inference is based on the posterior distribution:

$$\boxed{\pi(\mathrm{d}\theta) = \frac{\pi_0(\mathrm{d}\theta)\mathsf{L}(Y|\theta)}{\int \mathsf{L}(Y|u)\pi_0(\mathrm{d}u)}\,.}$$

- In most cases the normalizing constant is not tractable:

$$\pi(\mathrm{d}\theta) \propto \pi_0(\mathrm{d}\theta)\mathsf{L}(Y|\theta)\,.$$

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.

Markov chain theory

Let $(X_k)_{k\geq 0}$ be a Markov chain on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

- P Markov kernel associated with $(X_k)_{k\geq 0}$ if
  - for any ν, νP is the distribution of X_1 starting from $X_0 \sim \nu$
  - νP^k the distribution of X_k for $k \geq 0$
- Invariant probability measure: π is said to be an invariant probability measure for the Markov kernel P if

$$X_0 \sim \pi \text{ then } X_1 \sim \pi \qquad \text{equivalent to } \pi P = \pi$$

- Ergodic Theorem (Meyn and Tweedie, 2003): If π is invariant, With some conditions on P, we have for any $f \in \mathrm{L}^1(\pi)$,

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

Convergence of Markov chains

- A measure of efficiency of MCMC to target π associated to a Markov kernel P:

$$\|P^k(x,\cdot) - \pi\|_{\mathrm{TV}} \leq C(x)v(k) ,$$

1. The total variation distance defined for μ, ν two probability measures on $\mathbb{R}^d$ by

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{|f|\leq 1} |\mu(f) - \nu(f)| .$$

2. $C(x) \geq 0$: dependence on the initial condition.
3. Ideally $\lim_{k\to+\infty} v(k) = 0$ (or close to 0) with the better possible rate.

We answer to the following questions:

- For a target precision $\varepsilon > 0$, we can find $N \geq 0$ such that

$$\|\delta_x P^n - \pi\|_{\mathrm{TV}} \leq \varepsilon \text{ for all } n \geq N .$$

- In general N is not explicit.

Outline

Framework

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

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

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

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

Kinetic Langevin diffusion

- Kinetic Langevin SDE:

$$\mathrm{d}\mathbf{X}_t = \mathbf{V}_t \mathrm{d}t \; ,$$
$$\mathrm{d}\mathbf{V}_t = -(\kappa_1 \mathbf{V}_t + \kappa_2 \nabla U(\mathbf{X}_t))\mathrm{d}t + \sqrt{2\kappa_1\kappa_2}\mathrm{d}B_t \; ,$$

  where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian Motion and κ_1, κ_2 are positive constants.

- Notation: $(P_t)_{t\geq 0}$ the Markov semigroup associated to the Kinetic Langevin diffusion:

$$P_t(z, A) = \mathbb{P}((\mathbf{X}_t, \mathbf{V}_t) \in A | (\mathbf{X}_0, \mathbf{V}_0) = z) \; , \quad x \in \mathbb{R}^{2d}, A \in \mathcal{B}(\mathbb{R}^{2d}) \; .$$

- $\mu(x, v) \propto \exp(-U(x)) \exp(-\|v\|^2/2) \propto \pi(x) \exp(-\|v\|^2/2)$ is the unique invariant probability measure.

Discretized Kinetic Langevin diffusion

- Idea: Sample the diffusion paths:

$$\mathrm{d}\tilde{\mathbf{X}}_t = \tilde{\mathbf{V}}_t \mathrm{d}t \ ,$$
$$\mathrm{d}\tilde{\mathbf{V}}_t = -(\kappa_1 \tilde{\mathbf{V}}_t + \kappa_2 \nabla U(\tilde{\mathbf{X}}_{\Gamma_k}))\mathrm{d}t + \sqrt{2\kappa_1\kappa_2}\mathrm{d}B_t \ ,$$

  where

  - $(\gamma_k)_{k\geq 1}$ is a sequence of stepsizes, which can either be held constant or be chosen to decrease to 0 at a certain rate
  - For any $k \in \mathbb{N}$, $\Gamma_k = \sum_{i=1}^{k} \gamma_i$
- Noation: For any $k \in \mathbb{N}$, $(X_k, V_k) = (\mathbf{X}_{\Gamma_k}, \mathbf{V}_{\Gamma_k})$
- This algorithm is referred to as the Kinetic Langevin Algorithm.

Explicit form of the Kinetic Langevin Algorithm

- Explicit form:

$$X_{k+1} = X_k - \frac{1}{\kappa_1}(\mathrm{e}^{-\kappa_1\gamma_{k+1}} - 1)V_k - \frac{\kappa_2}{\kappa_1}\left(\gamma_{k+1} + \frac{\mathrm{e}^{-\kappa_1\gamma_{k+1}} - 1}{\kappa_1}\right)\nabla U(X_k)$$
$$- p_{\mathbb{R}^d \times \{0\}}\left(\Sigma_{\gamma_{k+1}}^{1/2} G_{k+1}\right) ,$$
$$V_{k+1} = \mathrm{e}^{-\kappa_1\gamma_{k+1}} V_k + \frac{\kappa_2}{\kappa_1}(\mathrm{e}^{-\kappa_1\gamma_{k+1}} - 1)\nabla U(X_k) + p_{\{0\}\times\mathbb{R}^d}\left(\Sigma_{\gamma_{k+1}}^{1/2} G_{k+1}\right) ,$$

where

- $(G_k)_{k\geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_{2d})$
- $(\gamma_k)_{k\geq 1}$ is a sequence of stepsizes, which can either be held constant or be chosen to decrease to 0 at a certain rate
-

$$\Sigma_\gamma = 2\kappa_1^{-1}\kappa_2 \begin{pmatrix} \frac{\kappa_1\gamma - 2^{-1}(1-\mathrm{e}^{-\kappa_1\gamma})^2 - (1-\mathrm{e}^{-\kappa_1\gamma})}{\kappa_1}\mathrm{I}_d & -2^{-1}(1-\mathrm{e}^{-\kappa_1\gamma})^2\mathrm{I}_d \\ -2^{-1}(1-\mathrm{e}^{-\kappa_1\gamma})^2\mathrm{I}_d & 2^{-1}\kappa_1(1-\mathrm{e}^{-2\kappa_1\gamma})\mathrm{I}_d \end{pmatrix} .$$

Discretized Kinetic Langevin diffusion: constant stepsize

- When the stepsize is held constant, *i.e.* $\gamma_k = \gamma$, then $(X_k, V_k)_{k \geq 1}$ is an homogeneous Markov chain with Markov kernel R_γ
- Under some appropriate conditions,

$$R_\gamma$$

is irreducible, positive recurrent $\rightsquigarrow$ unique invariant distribution μ_γ which does not coincide with the target distribution μ.
- Questions:
  - For a given precision $\epsilon > 0$, how should I choose the stepsize $\gamma > 0$ and the number of iterations n so that : $\|\delta_x R_\gamma^n - \mu\|_{\mathrm{TV}} \leq \epsilon$
  - quantify the distance between μ_γ and μ.

Discretized Kinetic Langevin diffusion: decreasing stepsize

- When $(\gamma_k)_{k\geq 1}$ is nonincreasing and non constant, $(X_k, V_k)_{k\geq 1}$ 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, $\mathbb{E}_x[f(X_p, V_p)] = \delta_x Q_\gamma^p f$.

- Questions:
  - Convergence : is there a way to choose the step sizes so that $\|\delta_x Q_\gamma^p - \mu\|_{\mathrm{TV}} \to 0$?

Outline

Kullback-Leibler divergence

- For $f:\mathbb{R}^d \to \mathbb{R}$, such that for any $x \in \mathbb{R}^d$, $f(x) > 0$, $\int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x) < +\infty$, define the entropy of f with respect to μ by

$$\mathrm{Ent}_\mu(f) = \int_{\mathbb{R}^d} f(x)\log(f(x))\mathrm{d}\mu(x) - \log\left(\int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x)\right)\int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x)\,.$$

- The Kullback-Leibler divergence between μ and ν is defined by

$$\mathrm{KL}(\nu,\mu) = \mathrm{Ent}_\mu(\mathrm{d}\nu/\mathrm{d}\mu)\ ,$$

if $\nu \ll \mu$ and $\mathrm{KL}(\mu,\nu) = +\infty$ otherwise.

Main result

Theorem 1

For any $k \in \mathbb{N}^*$ and $z \in \mathbb{R}^{2d}$,

$$\mathrm{KL}\left(\delta_z P_{\Gamma_k} \,|\, \delta_z Q_\gamma^k\right) \leq (L^2 \kappa_2/(2\kappa_1)) \sum_{j=0}^{k-1} \gamma_{j+1}^3 D(\gamma_{j+1}, \delta_z Q_\gamma^j) ,$$

where for any distribution ν on $\mathcal{B}(\mathbb{R}^d)$ and $\gamma \in \mathbb{R}_+^*$,

$$D(\gamma, \nu) = \int_{\mathbb{R}^{2d}} \{\mathrm{A}(\gamma) \left\|x'\right\|^2 + \mathrm{B}(\gamma) \left\|v'\right\|^2\} \mathrm{d}\nu(\mathrm{d}x' \mathrm{d}v') + d\kappa_1\kappa_2\gamma/4 ,$$

$$\mathrm{A}(\gamma) = L^2 \left(\frac{\kappa_2^2 \gamma^2}{15} + \frac{\kappa_2 \gamma}{8}\right) , \quad \mathrm{B}(\gamma) = \frac{1}{3} + \frac{\kappa_2 \gamma}{8} .$$

Main result

Corollary 2

For any $z \in \mathbb{R}^{2d}$, $k \in \mathbb{N}^*$, $p \in \mathbb{N}$, $p < k$,

$$\left\|\delta_z Q_\gamma^k - \mu\right\|_{TV} \leq L\sqrt{\frac{\kappa_2}{\kappa_1}\left(\sum_{j=p}^{k-1} \gamma_{j+1}^3 D(\gamma_{j+1}, \delta_z Q_\gamma^j)\right)} + \left\|\delta_z Q_\gamma^p P_{\Gamma_{p,k}} - \mu\right\|_{TV} .$$

Outline

Poincaré

- Objective compute bound for $\left\|\delta_z Q_\gamma^p P_{\Gamma_{p,k}} - \mu\right\|_{TV}$
- Assumption: π satisfies a Poincaré inequality i.e there exists C_P such that, for any $f \in C^2(\mathbb{R}^d)$ satisfying $\int_{\mathbb{R}^d} f \mathrm{d}\pi = 0$, $\int_{\mathbb{R}^d} (\nabla f)^2 \mathrm{d}\pi \geq C_P \int_{\mathbb{R}^d} f^2 \mathrm{d}\pi$.
- Remark: If U is such that $\langle \nabla U(x), x\rangle \geq L\chi_2 \|x\| - \tau_2$ then π satisfies a Poincaré inequality.

Theorem 3

For any $t \in \mathbb{R}_+$ and initial distribution ν_0 such that $\nu_0 \ll \mu$,

$$\|\nu_0 P_t - \mu\|_{TV} \leq A\mathrm{e}^{-\alpha t} Var_\mu(\mathrm{d}\nu_0/\mathrm{d}\mu)$$

where A and α are explicit constants.

Log-Sobolev

- Assumption:
  - π satisfies a log-Sobolev inequality *i.e.* there exists $C_{LS} > 0$ such that, for any continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfying $f(x) > 0$ for any $x \in \mathbb{R}^d$, $\int_{\mathbb{R}^d} f(x)\mathrm{d}x < +\infty$

$$\mathrm{Ent}_\pi(f) \leq C_{LS} \int_{\mathbb{R}^d} \{\|\nabla f(x)\|^2 / f(x)\} \mathrm{d}\pi(x) .$$

  - The potential U is infinitely continuously differentiable on $\mathbb{R}^d$ and for any $\alpha \in \mathbb{N}^d$ multi-index, $\sup_{x \in \mathbb{R}^d} |\partial_x^\alpha U(x)| < +\infty$.
- Remark: If U is strongly convex then π satisfies a log-Sobolev inequality.

Log-Sobolev

Theorem 4

For any distribution ν_0 *such that* $\mathrm{d}\nu_0/\mathrm{d}\mu \in \mathrm{C}_b^{2,+,*}(\mathbb{R}^{2d})$ *and* $t \in \mathbb{R}_+$ *,*

$$\mathsf{KL}(\nu_0 P_t, \mu) \le \exp(-\alpha t)\left(\mathsf{KL}(\nu_0, \mu) + \frac{1}{\beta}\int \Phi_2\left(\frac{\mathrm{d}\nu_0}{\mathrm{d}\mu}\right)\mathrm{d}\mu\right) ,$$

where α *is an explicit constant and* Φ_2*,* β *are defined by,*

$$\Phi_2(f) = \left(\|(\nabla_x + \kappa_1\nabla_v)f\|^2 + \|\nabla_v f\|^2\right)/f , \qquad \beta = 2(1+\kappa_2^2 L^2 + \kappa_1^2/2)/(\kappa_1^2\kappa_2) .$$

Outline

Convergence

- By the previous section there exist $\varrho \in (0,1)$, such that for any initial distribution ν_0, there exists $C(\nu_0) < +\infty$ such that,

$$\left\| \nu_0 P_t - \mu \right\|_{TV} \leq C(\nu_0)\varrho^t$$

Theorem 5

Assume that $\lim_{k\to+\infty} \gamma_k = 0$, $\lim_{k\to+\infty} \Gamma_k = +\infty$ *and some technical conditions. Then for any* $z \in \mathbb{R}^{2d}$

$$\lim_{k\to+\infty} \left\| \delta_z Q_\gamma^k - \mu \right\|_{TV} = 0$$

Algorithm complexity

Theorem 6

For all $\varepsilon > 0$, *we get* $\left\| \delta_z R_\gamma^k - \mu \right\|_{TV} \leq \varepsilon$ *if*

$$k > T\gamma^{-1} + 1 \qquad \textit{and} \qquad \gamma \leq \sqrt{\frac{\kappa_1 \varepsilon^2}{4\kappa_2 \overline{D}(z) L^2 (T + \overline{\gamma})}} \wedge \overline{\gamma} \,,$$

where

$$T = (\log(\overline{C}(z)) - \log(\varepsilon/2))/(-\log(\varrho)) \,,$$

for some explicit constants $\overline{C}(z)$ *and* $\overline{D}(z)$.

Distance between μ_γ and μ

- Assumtion: There exist $C < +\infty$, $v > 0$ such that, for any initial distribution ν_0, μ_0,

$$\|\nu_0 P_t - \mu_0 P_t\|_{TV} \leq C\mathrm{e}^{-vt} \|\nu_0 - \mu_0\|_{TV} .$$

Theorem 7

For any $\gamma < \overline{\gamma}$,

$$\|\mu - \mu_\gamma\|_{TV} \leq \gamma E$$

for some constant E .

Thank you for your attention.