Variance bounds and concentration for Markov chains

Aldéric Joulin* and Yann Ollivier†
*Institut de Mathématiques de Toulouse and †Ecole Normale Supérieure de Lyon

Fifth meeting BOSANTOUVAL - 06/03/2009
Parc Ornithologique du Teich

Outline

Introduction

π probability measure on state space $(\mathcal{X}, d)$.
$f : \mathcal{X} \to \mathbb{R}$ function.

Aim: estimate $\pi(f) := \int_{\mathcal{X}} f d\pi$ when classical numerical methods fail (for instance when $\mathcal{X}$ is high-dimensional).

Classical Monte-Carlo method: simulate i.i.d. sequence $Z_1, Z_2, \ldots, Z_T \sim \pi$ and use Law of Large Numbers to estimate $\pi(f)$ by

$$\hat{\pi}(f) := \frac{1}{T} \sum_{k=1}^{T} f(Z_k).$$

Central Limit Theorem gives the shape and fluctuation of the error $\hat{\pi}(f) - \pi(f)$.

Introduction

Problem: if π is complicated, difficult to simulate Z_k.

Idea: the so-called Markov chain Monte Carlo method:

• Find an easy-to-simulate Markov chain $(X_N)_{N \in \mathbb{N}}$ on $\mathcal{X}$ with stationary distribution π (waiting for a time T_0 so that $\mathcal{L}(X_{T_0}) \approx \pi$).

• Estimate $\pi(f)$ by

$$\hat{\pi}(f) := \frac{1}{T} \sum_{k=T_0+1}^{T_0+T} f(X_k),$$

according to the ergodic theorem.

Various algorithms allows us to simulate a Markov chain for a given π (Hastings-Metropolis, Gibbs sampler, etc...).

Introduction

Burn-in period: T_0 generally chosen to be the mixing time

$$\tau := \inf\{N \in \mathbb{N} : \|P_x^N - \pi\|_{\mathrm{TV}} \leq 1/4\},$$

where the norm is classical total variation and P_x^N is the law of X_N starting from x.

Main problem: in essence, LLN and CLT are asymptotic results, whereas non-asymptotic estimates (in time) are required for simulation purposes, i.e. when one wants to estimate the minimum time to run the simulation algorithm in order to achieve a prescribed level of accuracy.

Introduction

Literature for non-asymptotic concentration results of type:

$$\mathbb{P}(|\hat{\pi}(f) - \pi(f)| > r), \quad r > 0.$$

• Lezaud (AAP 1998): discrete/continuous-time on $\mathcal{X}$ finite. Approach through spectral gap.

• Wu (AIHP 2000), Cattiaux-Guillin (ESAIM 2008), Guillin-Léonard-Wu-Yao (PTRF 2009): Markov processes, f Lipschitz, regularity of the initial distribution. Approach through functional inequalities (transportation, F-Sobolev, etc...) satisfied by π.

• Joulin (Bernoulli 2009): pure-jump Markov processes, f Lipschitz, initial measure: Dirac. Approach through curvature + tensorization of Laplace transform.

Introduction

Our objectives: to give non-asymptotic bounds on bias/variance and concentration of $\hat{\pi}(f) - \pi(f)$, f Lipschitz, that:

- are new.

- recover the existing ones.

- are easily applicable on several examples (discrete/continuous-time, finite or infinite space $\mathcal{X}$, jump or diffusion processes, etc...).

- need few informations on π.

- do not use reversibility.

Notation and results

$(X_N)_{N\in\mathbb{N}}$ Markov chain on $(\mathcal{X}, d)$ with stationary distribution π and transition probabilities $(P_x)_{x\in\mathcal{X}}$.

Wasserstein distance between $\mu, \nu \in \mathcal{P}_1(\mathcal{X})$:

$$\begin{aligned} W_1(\mu,\nu) &:= \inf_{\gamma\in\mathrm{Marg}(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)\gamma(dx,dy) \\ &= \sup\left\{ \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu : \|f\|_{\mathrm{Lip}} \leq 1 \right\}, \end{aligned}$$

by the Kantorovich-Rubinstein duality theorem.

Notation and results

Definition

$(X_N)_{n\in\mathbb{N}}$ *has Ricci curvature on* $(\mathcal{X}, d)$ *bounded below by* $\kappa \leq 1$ *if*

$$W_1(P_x, P_y) \leq (1-\kappa) d(x,y), \quad x, y \in \mathcal{X}.$$

• Link with geometry: Ollivier (JFA 2009) recovers the classical Ricci curvature on Riemannian manifolds.

• Discrete analogous of the Wasserstein curvature emphasized in Joulin (Bernoulli 2009).

• Classical Dobrushin coefficient in statistical mechanics.

Notation and results

Link with ergodicity: if $\kappa > 0$ then for any $x \in \mathcal{X}$,

$$W_1(P_x^N, \pi) \leq (1-\kappa)^N E(x) \underset{N \to \infty}{\longrightarrow} 0,$$

with the eccentricity

$$E(x) = \int_{\mathcal{X}} d(x,y)\pi(dy) \leq \begin{cases} \operatorname{diam} \mathcal{X}; \\ \frac{1}{\kappa} \int_{\mathcal{X}} d(x,y) P_x(dy). \end{cases}$$

Notation and results

Diffusion constant and granularity controlling jumps of $(X_N)_{N\in\mathbb{N}}$:

$$\frac{\sigma(x)^2}{n_x} := \sup\{\mathrm{Var}_x(f) : \|f\|_{\mathrm{Lip}} \le 1\};$$

$$\sigma_\infty := \frac{1}{2} \sup_{x\in\mathcal{X}} \mathrm{diam}\,\mathrm{Supp}\, P_x,$$

where the variance is

$$\mathrm{Var}_x(f) := \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{X}} (f(y) - f(z))^2 P_x(dy) P_x(dz).$$

Notation and results

Mean quadratic error:

$$\begin{aligned}
\mathbb{E}_x\left[|\hat{\pi}(f)-\pi(f)|^2\right] &= |\mathbb{E}_x[\hat{\pi}(f)]-\pi(f)|^2+\mathrm{Var}_x[\hat{\pi}(f)] \\
&= (\text{bias})^2+\text{variance} \\
&\leq \left(\frac{(1-\kappa)^{T_0+1}}{\kappa T}E(x)\|f\|_{\mathrm{Lip}}\right)^2+\text{variance}.
\end{aligned}$$

Hence a control of the variance is required to control the L^2-error.

Notation and results

Theorem

Assume that:

- *Ricci curvature* $\geq \kappa > 0$;
- *there exists a* C*-Lipschitz function* S *such that*

$$\frac{\sigma(x)^2}{n_x \kappa} \leq S(x), \quad x \in \mathcal{X}.$$

Then the variance of $\hat{\pi}(f) := \frac{1}{T} \sum_{k=T_0+1}^{T_0+T} f(X_k)$ *is bounded as:*

$$\mathrm{Var}_x \left[\hat{\pi}(f)\right] \leq \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa T} \left((1 + 1/\kappa T)\mathbb{E}_\pi[S] + \frac{2C(1-\kappa)^{T_0}}{\kappa T} E(x) \right).$$

Notation and results

Theorem

Under the following assumptions:

- *granularity* $\sigma_\infty < \infty$;
- *Ricci curvature* $\geq \kappa > 0$;
- *there exists a* C*-Lipschitz function* S *such that* $\frac{\sigma(x)^2}{n_x \kappa} \leq S(x)$,

we have the concentration result for any 1-Lipschitz function f *and any* $r > 0$:

$$\mathbb{P}_x \left(|\hat{\pi}(f) - \pi(f)| > r + \mathrm{bias} \right) \leq \begin{cases} 2\exp\left\{ -\frac{r^2}{16 V_{T,x}^2} \right\} & r \leq r_{\max} \\ 2\exp\left\{ -\frac{r\kappa T}{\max\{2C, 3\sigma_\infty\}} \right\} & r > r_{\max} \end{cases}$$

where Gaussian window is $r_{\max} := 4\kappa T V_{T,x}^2 / \max\{2C, 3\sigma_\infty\}$ *and* $V_{T,x}^2$ *is the latter variance bound.*

Notation and results

• Price to pay compared with the i.i.d. case: a κ-term at the denominator of the variance bound. It is expected since

$$\mathrm{Var}_x\left[\hat{\pi}(f)\right] \underset{T_0\to\infty}{\approx} \mathrm{Var}_\pi\left[\hat{\pi}(f)\right] \leq \frac{2}{\kappa T}\mathrm{Var}_\pi[f]$$

by the exponential decrease of correlations (at least in the reversible case by using spectral gap).

• Allow to have a Lipschitz diffusion constant (convenient for approximations), in contrast to the usual case where it is bounded.

• Estimate with $\delta_x \to \mu$ might be obtained. In this case, additional variance term of order $\frac{1}{(\kappa T)^2}\int_{\mathcal{X}}\int_{\mathcal{X}} d(y,z)^2\mu(dy)\mu(dz)$. In particular, no regularity assumption on the initial distribution.

• Gaussian behaviour in accordance with the CLT when rescaling ($r_{max} = \infty$ in this case).

Hypercube

Hypercube $\mathcal{X} = \{0,1\}^N$ equipped with Hamming metric:

$$d(x,y) = \sum_{k=1}^{N} |x_k - y_k| = \mathrm{Card}\,\{k \in \{1,\ldots,N\} : x_k \neq y_k\}.$$

Uniform probability on $\mathcal{X}$: $\pi(x) = 2^{-N}$, invariant measure of the lazy random walk:

$$P_x(y) = \begin{cases} 1/2 & \text{if} \quad y = x; \\ 1/2N & \text{if} \quad y \sim x. \end{cases}$$

Quantities of interest:

- Ricci curvature $\kappa = 1/N$.
- granularity $\sigma_\infty = 1$.
- diffusion constant $\sigma(x)^2/n_x \leq 1/2$.

Hypercube

Let f be 1-Lipschitz on the Hypercube.

- Bound on the bias:

$$|E_x[\hat{\pi}(f)] - \pi(f)| \leq \frac{N^2}{2T} \exp\{-T_0/N\}.$$

To ensure a small bias, choose $T_0 \asymp N\log(N)$ which is known to be the mixing time.

- Bound on the variance:

$$\mathrm{Var}_x[\hat{\pi}(f)] \leq \frac{N^2}{2T}(1 + N/T).$$

- Concentration: for $r = O(N)$ and $T_0 = 0$:

$$\mathbb{P}_x(|\hat{\pi}(f) - \pi(f)| > r + \mathrm{bias}) \leq 2\exp\left\{-\frac{Tr^2}{8N^2}\right\}.$$

$M/M/\infty$ queueing process

Markov process $(X_t)_{t\geq 0}$ on $\mathcal{X} = \mathbb{N}$ endowed with the classical metric $d(x,y) = |x-y|$. Transition probabilities:

$$P_x^t(y) = \begin{cases} \lambda t + o(t) & \text{if} \quad y = x+1; \\ xt + o(t) & \text{if} \quad y = x-1; \\ 1-(\lambda+x)t + o(t) & \text{if} \quad y = x. \end{cases}$$

Invariant distribution: $\pi \sim \mathrm{Poisson}(\lambda)$.

Using an approximation by Markov chain of binomial-type, one obtains for $\hat{\pi}(f) := \frac{1}{T}\int_0^T f(X_s)ds$, with $f : \mathbb{N} \to \mathbb{R}$ 1-Lipschitz:

$$\mathbb{P}_x\left(|\hat{\pi}(f) - \pi(f)| > r + \mathrm{bias}\right) \leq \begin{cases} 2\exp\left\{-\frac{Tr^2}{16(2\lambda+(\lambda+x)/T)}\right\} & r \leq r_{\max} \\ 2\exp\left\{-\frac{rT}{12}\right\} & r > r_{\max} \end{cases}$$

where Gaussian window is $r_{\max} := (8\lambda T + 4(\lambda + x))/3T$.

$M/M/\infty$ queueing process

• Difficulty of unbounded generator overcome by its Lipschitz property.

• Estimate comparable to that of Guillin-Léonard-Wu-Yao (PTRF 2009) obtained through large deviations combined with transportation-information inequalities, except that no regularity assumption required on the initial distribution.

Euler scheme for diffusions

Diffusion process on Euclidean space $\mathbb{R}^d$:

$$dX_t = b(X_t)dt + \sqrt{2}\rho(X_t)dW_t, \quad t > 0,$$

where $(W_t)_{t\geq 0}$ Brownian motion on $\mathbb{R}^d$, $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\rho : \mathbb{R}^d \to \mathcal{M}_{d\times d}(\mathbb{R})$.

Hilbert-Schmidt norm: $\|A\|_{\mathrm{HS}} = \sqrt{\mathrm{tr}(A^*A)} = \sqrt{\sum_{i,j} a_{ij}^2}$.

Operator (spectral) norm: $\|A\| = \sup_{v\neq 0} \frac{\|Av\|}{\|v\|} = \sqrt{\lambda_{\max}(A^*A)}$.

Assume the following: the functions b and ρ are Lipschitz and there exists $\alpha > 0$ such that

$$\text{(C)} \quad \|\rho(x)-\rho(y)\|_{\mathrm{HS}}^2+\langle x-y, b(x)-b(y)\rangle \leq -\alpha\|x-y\|^2, \quad x, y \in \mathbb{R}^d.$$

Euler scheme for diffusions

Approximation by the Euler scheme

$$X_{N+1}^{(\delta t)} = X_N^{(\delta t)} + b(X_N^{(\delta t)})\delta t + \sqrt{2\delta t}\rho(X_N^{(\delta t)})Y_N,$$

where $(Y_N)_{N\in\mathbb{N}}$ i.i.d. standard Gaussian.

Quantities of interest:

- Ricci curvature $\geq \kappa \geq \alpha\delta t + O(\delta t^2)$.
- granularity $\sigma_\infty \to 0$ as $\delta t \to 0$.
- diffusion constant bounded by $2\delta t\|\rho(x)\|^2$.

Euler scheme for diffusions

Corollary

Let $\hat{\pi}(f) := \frac{1}{T}\int_0^T f(X_s)ds$, *with* $f : \mathbb{R}^d \to \mathbb{R}$ *1-Lipschitz. Under the following assumptions:*

- *Condition (C) is fulfilled;*
- *there exists a C-Lipschitz function* $S : \mathbb{R}^d \to \mathbb{R}$ *such that* $S(x) \geq \frac{2}{\alpha}\|\rho(x)\|^2$. *Then*

$$\mathrm{Var}_x[\hat{\pi}(f)] \leq \frac{1}{\alpha T}\mathbb{E}_\pi[S] + \frac{CE(x)}{(\alpha T)^2} =: V_{T,x}^2,$$

$$\mathbb{P}_x\left(|\hat{\pi}(f) - \pi(f)| > r + \mathrm{bias}\right) \leq \begin{cases} 2\exp\left\{-\frac{r^2}{16V_{T,x}^2}\right\} & r \leq r_{\max} \\ 2\exp\left\{-\frac{r\alpha T}{8C}\right\} & r > r_{\max} \end{cases}$$

where Gaussian window is $r_{\max} := 2V_{T,x}^2\alpha T/C$.

Euler scheme for diffusions

• Basic example: OU-type process with $\rho = \mathrm{Id}$ and $b = -\nabla V$ (hence $\pi(dx) = e^{-V} dx$).

In this case, Assumption (C) is equivalent to Bakry-Emery criterion: $\mathrm{Hess}\, V \geq \alpha\, \mathrm{Id}$, and with $C = 0$ we obtain Gaussian control of concentration.

• Convenient for processes with volatility $\rho(x)$ growing like $\sqrt{x}$.

Some references

P. Cattiaux and A. Guillin.
Deviation bounds for additive functionals of Markov processes.
ESAIM Probab. Stat., 12:12–29, 2008.

A. Guillin, C. Léonard, L. Wu, and N. Yao.
Transportation-information inequalities for Markov processes.
To appear in Probab. Theory Relat. Fields, 2009.

A. Joulin.
A new Poisson-type deviation inequality for Markov jump processes with positive Wasserstein curvature.
To appear in Bernoulli, 2009.

P. Lezaud.
Chernoff-type bound for finite Markov chains.
Ann. Appl. Probab., 8(3):849–867, 1998.

Y. Ollivier.
Ricci curvature of Markov chains on metric spaces.
J. Funct. Anal., 256(3):810–864, 2009.

L. Wu.
A deviation inequality for non-reversible Markov processes.
Ann. Inst. H. Poincaré Probab. Statist., 36(4):435–445, 2000.