MCMC sampling error

Qifan Song

Department of Statistics
Purdue University

Review of MCMC ground theory

- Make Bayesian Inference base on the posterior distribution;
- Non-normalized density $\pi(\theta)l(\theta; D_n)$;
- Monte Carlo sampling VS Riemann Integral;
- Sampling error between target π and actual sampled random variable X.

$$\|L_X - L_\pi\|_{TV} = \max_C |L_X(C) - \pi(C)| = (1/2) \int |f(x) - \pi(x)| dx$$

Review of MCMC ground theory

- Ergodicity of MCMC $\{X_i\}$:
  If a Markov chain is irreducible and aperiodic, and it admits a finite measure π as its invariant measure, then,

  $$\|\pi_0 P^n - \pi\|_{TV} \to 0,$$

- M-H algorithm: If π is bounded away from 0 and ∞ on any compact set of its support, and there exist δ and ϵ such that the proposal distribution satisfies:

  $$|x-y| \leq \delta \Rightarrow q(x,y) \geq \epsilon,$$

  the chain is ergodic and geometric ergodic.

Sampling error due to convergence rate

- Geometric ergodicity of MCMC $\{X_i\}$:
  If there exist a set C, constant $b < \infty$, $\beta > 0$, and function $V \geq 1$ finite at some one x_0, constant n_0 and a non-zero measure $\nu(A)$, satisfying

  $$E[V(X_1)|X_0 = x] - V(x) \leq -\beta V(x) + bI_C(x), \quad x \in \Omega,$$

  $$P^{n_0}(x, A) \geq \nu(A), \quad x \in C,$$

  then,

  $$\|\pi_0 P^n - \pi\|_{TV} \leq M\rho^C,$$

  for some $\rho \in (0, 1)$.
- Computable bound for M and ρ exist, in Meyn and Tweedie (1994), *Annals of App. Prob.*

Sampling error due to Computer Simulation

- There is no random sampling for computer simulation!
- There is no continuous variable for computer simulation!
- The precision of a double floating point 2^{-1023}.
- In MCMC, the error accumulates.
- My own experience: recursively algorithm for $(X_\xi^T X_\xi)^{-1}$.
- Robert, Rosenthal and Schwartz, *J. Appl. Prob* 1998; Chen, Dick and Owen, *Ann. Stat.* 2011

Today's topic

- "Theoretical guarantees for approximate sampling from smooth and log-concave densities" by Arnak Dalayan.
- Consider a smooth target distribution $\pi(x) \propto \exp\left(-f(x)\right)$
- For x, $x' \in \mathbb{R}^p$, f satisfied

$$f(x) - f(x') - \nabla f(x')(x - x')^T \geq m\|x - x'\|^2/2,$$

$$\|\nabla f(x) - \nabla f(x')\| \leq M\|x - x'\|.$$

- Unimodal, thin tailed distribution.
- Interested in Langevin Monte Carlo.

Langevin Diffusion

- Langevin diffusion

$$dL_t = \nabla \log \pi(L_t)dt + \sqrt{2}dW_t,$$

where W_t is p-dim Brownian Motion.

- If π is sufficiently smooth,

$$\|\nu P_L^t - \pi\|_{TV} \to 0,$$

where $P_L^t(x, A) = Pr(L_{t+s} \in A | L_s = x)$.

Langevin Diffusion

- Exponential convergence: $\|\nu P_L^t - \pi\|_{TV} < M\rho^t$.
- Roberts and Tweedie, *Bernoulli* 1996 studied the sufficient condition of exponential convergence of Langevin diffusion.
- No exponential convergence for $|f(x)| \to 0$.
- This work gives an non-asymptotic bound:

$$\|\nu P_L^t - \pi\|_{TV} < \frac{1}{2}\chi^2(\nu\|\pi)^{1/2}e^{tm/2}, \quad \text{(lemma 1)}$$

- $\chi^2(\nu\|\pi)$ measures the discrepancy between ν and π. (Defines in next slide)

Langevin Diffusion: sketch of the proof

$$\chi^2(\nu\|\pi) = \int \left(\frac{d\nu}{d\pi}(x) - 1\right)^2 \pi(dx)$$

$$\begin{aligned}
\|\nu P_L^t - \pi\|_{TV} &= \sup_A \left| \int P_L^t(x, A)\nu(x)dx - \pi(A) \right| \\
&= \sup_A \left| \int (P_L^t(x, A) - \pi(A))\nu(x)dx \right| \\
&= \sup_A \left| \int (P_L^t(x, A) - \pi(A))(\nu(x) - \pi(x))dx \right| \\
&\leq \sup_A \int \left| P_L^t(x, A) - \pi(A) \right| \left| \frac{\nu(x)}{\pi(x)} - 1 \right| \pi(x)dx
\end{aligned}$$

Sketch of the proof

$$
\begin{aligned}
\|\nu P_L^t - \pi\|_{TV} &\le \sup_A \int \left| P_L^t(x,A) - \pi(A) \right| \left| \frac{\nu(x)}{\pi(x)} - 1 \right| \pi(x)dx \\
&\le \sup_A \left(\int \left| P_L^t(x,A) - \pi(A) \right|^2 \pi(x)dx \right)^{1/2} \sqrt{\chi^2(\nu\|\pi)} \\
&\le \sup_A \left(e^{-tm}\pi(A)[1-\pi(A)] \right)^{1/2} \sqrt{\chi^2(\nu\|\pi)}.
\end{aligned}
$$

The last inequality follows Chen and Wang, *Trans. Amer. Math. Soc* 1997

Langevin Monte Carlo

- $dL_t = -\nabla f(L_t)dt + \sqrt{2}dW_t$, not able to sample from this SDE.
- Using discrete approximation $\{X^{(k,h)}\}$
- $\Delta X = X^{(k+1,h)} - X^{(k,h)} = -h\nabla f(X^{(k,h)}) + \sqrt{2h}\varepsilon^{(k)}$

- $X^{k,h}$ generally converge to a stationary distribution which is not π.

Metroplis-Adjusted Langevin Monte Carlo

- $X^{(k+0.5,h)} = X^{(k,h)} - h\nabla f(X^{(k,h)}) + \sqrt{2h}\varepsilon^{(k)}$
- $X^{(k+1,h)} = X^{(k+0.5,h)}$ or $X^{(k+1,h)} = X^{(k,h)}$ by metroplis ratio.

- Rondom Walk M-H algorithm converges to scaled Langevin diffusion $L_{\alpha(\sigma)t}$.
- Maximize $\alpha(\sigma)$ to get optimal scale, with optimal accept rate 0.234.

The discrepancy between Langevin Diffusion and LMC

- SDE for Langevin Diffusion L_t:

$$dL_t = -\nabla f(L_t)dt + \sqrt{2}dW_t.$$

- SDE for Langevin Monte Carlo D_t (a piecewise linear process based on $X^{(k,h)}$)

$$dD_t = b_t(D)dt + \sqrt{2}dW_t,$$

where $b_t(D) = -\sum_{k=0}^{\infty} \nabla f(D_{kh}) I\left(t \in [kh, kh+h)\right)$, i.e. the negative gradien at the nearest previous time knot.

The discrepancy between Langevin Diffusion and LMC

- Girsanov theorem describes the change of measure for stochastic process. provide close expression for Radon-Nikodym derivative $d\mu_1/d\mu_2$.
- KL divergence $\int \log[d\mu_1(x)/d\mu_2(x)]d\mu_2(x)$
- Total variation $\leq \sqrt{\text{KL divergence}/2}$
- $$KL(P_L^{x,Kh} \| P_D^{x,Kh}) = \frac{1}{4}\int_0^{Kh} E[\|\nabla f(D_t) + b_t(D)\|^2]dt$$

The discrepancy between Langevin Diffusion and LMC

$$
\begin{aligned}
KL(P_L^{x,Kh} \| P_D^{x,Kh}) &= \frac{1}{4}\int_0^{Kh} E[\|\nabla f(D_t) + b_t(D)\|^2]dt \\
\leq & \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{kh+h} E\|D_t - D_{kh}\|^2 dt \\
\leq & \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{kh+h} \left(E\|\nabla f(D_{kh})\|^2 (t-kh)^2 + 2p(t-kh)\right) dt \\
\leq & \frac{M^2h^3}{12}\sum_{k=0}^{K-1} E\|\nabla f(D_{kh})\|^2 + \frac{pKM^2h^2}{4}
\end{aligned}
$$

The discrepancy between Langevin Diffusion and LMC

- $Ef(D_{kh})$ is controlled by the nature of Langevin MC.
  $E[f(D_{kh}) - f^*) \leq (1-\rho)E[f(D_{(k-1)h}) - f^*) + Mhp$
- By strongly convex, so does $\|\nabla f(D_{kh})\|^2$.
- When $h \leq 1/(\alpha M)$ with $\alpha > 1$

$$KL(P_L^{x,Kh} \| P_D^{x,Kh}) \leq \frac{M^3 h^2 \alpha}{12(2\alpha - 1)}(\|x - x^*\|^2 + 2Khp) + \frac{pHM^2h^2}{4}$$

Error rate of LMC (main result)

If start with $\nu =$N$(x^*,M^{-1}I_p)$, $h \leq 1/(\alpha M)$ with $\alpha > 1$

$$\|\nu P^K - \pi\|_{TV} \leq \frac{1}{2}\exp\left\{\frac{p}{4}\log(M/m) - \frac{TM}{2}\right\} + \left\{\frac{pM^2Th\alpha}{4(2\alpha-1)}\right\}^{1/2}$$

Target error rate ϵ

Choose:

$$T = \frac{4\log(1/\epsilon) + p\log(M/m)}{2m}, \quad h = \frac{\epsilon^2(2\alpha - 1)}{M^2 T p \alpha}$$

with $\alpha = (1 + MpT\epsilon^{-2})/2$. The LMC achieve the require precision at (T/h)th step.