Lecture9

Tuesday, October 9, 2018 9:25 AM

CONVERGENCE OF MCMC

LAST TIME:

1) Introduced MCMC: key: Metropolization

+ proposal:
  + $q(m_k, p)$
  + acceptance rate:

$$\alpha(m_k, p) = \min\left\{1, \frac{\pi(p)\, q(p, m_k)}{\pi(m_k)\, q(m_k, p)}\right\}$$

+ the transition kernel / probability induced by MCMC satisfies the reversibility $\Rightarrow$ $\pi(m)$ is a invariant density / measure of the Markov chain.

Is $\{m_i\}$ obtained from MCMC eventually distributed by $\pi(m)$?

- Invariant measures may not be unique, and it is possible that we cannot get from one state to another $\Rightarrow$ it is known as: the MC is reducible.

Irreducibility. If

a) the desired density $\pi(m)$ is finite everywhere

b) $q(\cdot,\cdot)$ is positive and continuous.

Remarks: + The first condition guarantees that we have non-zero acceptance rate

+ The second condition allows the Markov chain to explore everywhere in the parameter space. and avoids zero acceptance rate.

Aperiodicity. even the chain is irreducible it may still not converge to the target distribution $\pi(m)$. The reason is that the chain could oscillate between states. Fortunately the Metropolis-Hastings algorithm satisfies the aperiodicity condition (non-periodic) automatically when $m \in \mathbb{R}^n$

Thm: If the Markov chain (generated by

the above Metropolis-Hastings algorithm) has $\pi(m)$ as its stationary distribution and it is both irreducible and aperiodic. Then for almost everywhere $m \in \mathbb{R}^n$ the Markov chain is eventually distributed by $\pi(m)$. Specifically If we define the N-step transition probability:

$$P^N(m, A) := \mathbb{P}[m_N \in A \mid m_0 = m]$$

then:

$$\lim_{N \to \infty} P^N(m, A) = \mu(A) := \int_A \pi(m)\, dm$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \| $$

$$\int_A \pi(dm)$$

Furthermore: a LLN-type convergence holds

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} g(m_i) = \int_S g(m)\, d\pi(m)$$

where $\{m_i\}$ is the Markov chained obtained from M-H algorithm.

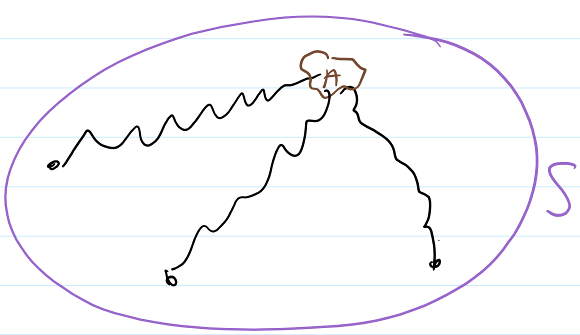
Harris recurrent:

- The above thm says that it is possible that the markov chain will not converge because the results only hold almost everywhere. A sufficient condition for the results to hold everywhere is known as Harris's recurrent!

Def: Harris recurrent:

- for all $A \subset S$: $\mu(A) > 0$,

If the markov chain starts at any $m \in S$, the chain eventually reaches A, then the Markov chain is called Harris's - recurrent



Sufficient condition for Harris's recurrent: If the proposal $q(m, \cdot)$ is absolutely continuous w.r.t the target density $\pi(\cdot)$, i.e.,

$$q(m, A) = \int c(m)\, d\pi(m) \quad \forall A \subset S$$

$$q(m, A) = \int_A c(m)\, d\pi(m) \quad \forall A \subset S$$

$$\Downarrow$$

$$\text{if } \mu_\pi(A) = 0 \Rightarrow \mu_q(A) = 0$$

$$\Downarrow$$

$$\mu_\pi(A_n) \to 0 \Rightarrow \mu_q(A_n) \to 0$$

Then the Markov chain is Harris-recurrent.

Thm: If a Markov chain is irreducible, aperiodic, and Harris-recurrent, then the chain is eventually distributed by $\pi(m)$ no matter where the chain starts.

Random-walk - Metropolis - Hastings

(RWMH)

- let's consider the following random walk

$$q(m, p) := \frac{1}{\sqrt{2\pi\gamma^2}} \exp\left(-\frac{1}{2\gamma^2} \|m-p\|^2\right)$$

(Gaussian centered at m, previous state)

then the acceptance rate can be simplified

$$\alpha(m_k, p) = \min\left\{1, \frac{\pi(p)\, q(p, m_k)}{\quad}\right\}$$

$$\alpha(m_k, p) = \min \left\{ 1, \frac{\pi(p)\, q(p, m_k)}{\pi(m_k)\, q(m_k, p)} \right\}$$

random walk $\;\Big\|\;$ $q(m_k, p) = q(p, m_k)$

$$= \min \left\{ 1, \frac{\pi(p)}{\pi(m_k)} \right\}$$

Remarks:

1) the RWMH suggests that we have high probability of accepting p if $\pi(p) \simeq \pi(m_k)$ of $\pi(p) > \pi(m_k)$ is ideal.

2) what remains is to determine the step length γ:

+ if γ is large $\Rightarrow$ good chance that $\pi(p) \ll \pi(m_k)$ $\Rightarrow$ rejection probability is high $\Rightarrow$ the Markov chain may not explore the state space well

+ if γ is small $\Rightarrow$ $\pi(p) \simeq \pi(m_k)$

$\Rightarrow$ accepts p most of the time $\Rightarrow$ the chain stays in some small region $\Rightarrow$ the chain may not explore well the state space either.

- It can be shown (for independent distribution in $\mathbb{R}^n$, $\pi(\vec{m}) \propto \pi(m^1) \times \pi(m^2) \times \dots \pi(m^n)$, where $\vec{m} = [m^1, \dots, m^n]^T$) that the compromised step length is given by

$$\gamma = O\left(\frac{1}{\sqrt{n}}\right)$$

then the acceptance rate is bound away from zero as $n \to \infty$

$$\Downarrow \text{ RWMH}$$

the optimal acceptance rate is 0.234

- Recall in RWMH, the proposal

$$q(m_k, p) \sim \exp\left(-\frac{1}{2\gamma^2}\|m_k - p\|^2\right)$$

$p \sim q(m_k, p)$ $\Downarrow$

$$p = m_k + \gamma w \quad \text{where } w \sim N(0,1)$$

in some sense a derivative free optimization technique for

$$J = -\log(\pi(m))$$

$\Rightarrow$ should be able to improve the sampling method by using derivative info.

$\Downarrow$ Langevin dynamics

$$p = m_k - \gamma^2 \nabla J(m_k) + \gamma w$$

$\Downarrow$ put it through M-H

Metropolis-Adjusted Langevin Algorithm (MALA)

$\Downarrow$

optimal scaling

$$\gamma \sim \frac{1}{\sqrt[4]{n}}$$

$\Downarrow$ $\frac{}{4\sqrt{n}}$

optimal acceptance rate ≈ 0.587 ??

$\boxed{\text{How about Newton ??}}$

Back to LLN for Markov chain

- For converged Markov chain (irreducibility, aperiodic, etc), then we know

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} m_i \xrightarrow{a.s} \mathbb{E}_{\pi}[m]$$

* For i.i.d samples

$$S_N := \frac{1}{N} \sum m_i \xrightarrow{MSE} \bar{m} \quad \text{with}$$

the rate $\frac{\sigma}{\sqrt{N}}$

* How about samples from a Markov chain?

for Markov chain the samples are not i.i.d. but correlated. What we can hope for is every k samples we have an independent samples $\Rightarrow$ effective sample size is N/k. In other words, what we hope is to have N/k "i.i.d. samples". Then the convergence rate

$$\| S_N - \mathbb{E}_\pi[m] \|_{L^2(\pi)} = O\left(\frac{\sigma}{\sqrt{N/k}} \right)$$

$$= O\left(\sqrt{k} \frac{\sigma}{\sqrt{N}} \right)$$

- This estimation suggests that we can improve the convergence of the Markov chain by at least two ways.

1) Reduce k: it turns out the exploring the structure of $J = -\log(\pi(m))$, via derivative information for example, reduces k

2) Improve the rate from $\frac{1}{\sqrt{N}}$ to $\frac{1}{N^\beta}$ where $\beta > 1/2$. One way to achieve

this is to use high-dimensional quadratures

$\Rightarrow$ this is the subject of quasi-Monte-Carlo methods.

HW (not submitted): read autocorrelation function as a means to estimate K.