Lecture 22 - MCMC

Friday, November 20, 2020 13:26

today : MCMC
- review Markov chains
- M-H algo

MCMC - Markov chain Monte-Carlo

idea : is to relax indep. assumption between samples

to allow adaptive proposal dist.

i.e. we'll run a chain $X_t \mid X_{t-1}$ s.t. $X_t \xrightarrow{t\to\infty}$ in dist. to target dist. p
"stationary dist. of chain"

then we can approximate

$$\mathbb{E}_p[f(x)] \text{ as } \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} f(x_t)$$

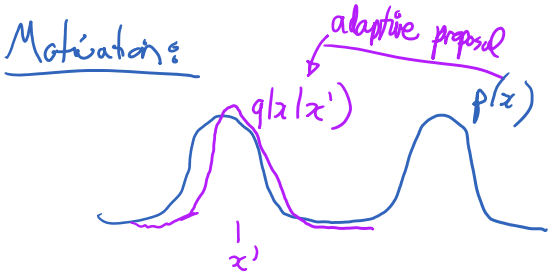
T_0 is called "burn-in period" ~> depends on "mixing time" of Markov chain

* no need to thin the samples [i.e. use Δt between samples to get more independence]

as this yield higher variance

→ better to use all samples after T_0 to estimate μ (unless it is too expensive)

Motivation:



before : samples were $X^{(t)} \overset{iid}{\sim} q$

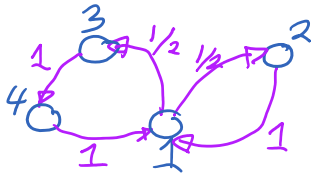
MCMC $X^{(t)} \mid X^{(t-1)} \sim q(\cdot \mid X^{(t-1)})$

Markov transition prob.

Review of (finite state space) Markov chains $[|X| = k]$

- as a DGM, 

- there is also transition prob. pt. of view : use one node per state (probabilistic FSA)

eg. $k=4$ 

[homogeneous M.C.]

↳ i.e. $P\{X_t = i \mid X_{t-1} = j\} = A_{ij}$ (no time dep.)

A is a $k \times k$ matrix s.t. $\mathbb{1}_k^T A = \mathbb{1}_k^T$

"left-stochastic matrix" — $\mathbb{1}_k$: vector of ones of size k

⊛ (as in HMM) suppose $P\{X_{t-1} = j\} = (\pi_{t-1})_j$

$$P\{X_t = i\} = \sum_j \underbrace{P\{X_t = i \mid X_{t-1} = j\}}_{A_{ij}} \underbrace{P\{X_{t-1} = j\}}_{(\pi_{t-1})_j}$$

$$\pi_{t+1} = A\pi_t$$

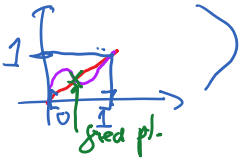
$$\Rightarrow \boxed{\pi_t = A^t \pi_0}$$

stationary dist π of A is a dist. π s.t. $A\pi = \pi$

[note that π is a right e-vector of A with e-value 1]

fact: every stochastic matrix has at least 1 stat. dist.

(by Brouwer's fixed pt. thm.)



def: irreducible Markov chain ⟺ there exists a positive prob. "path" from any i to any j (states)

$\forall (i,j)$, $\exists$ an integer m_{ij} s.t. $(A^{m_{ij}})_{ij} > 0$

(by Perron-Frobenius thm.) ⇒ irreducible M.C. has a unique stat. dist.

[multiplicity of e-value 1 is 1]

⊛ in order to converge to it, we need aperiodicity as well

irreducible and aperiodic M.C. ⟺ $\exists$ an integer m s.t. $A^m > 0$ ($(A^m)_{ij} > 0 \ \forall ij$)

aka regular M.C. (finite state space)

or ergodic M.C.

⊛ [note: a sufficient condition for an irreducible M.C.

to be aperiodic $\Leftrightarrow \exists i$ s.t. $A_{ii} > 0$ ~

example of a regular M.C. on k states

$$A = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix} = \frac{1}{k-1}\left(\mathbb{1}\mathbb{1}^T - I\right)$$

$$A^2 = \frac{1}{(k-1)^2}\left(\mathbb{1}\mathbb{1}^T\mathbb{1}\mathbb{1}^T - 2\mathbb{1}\mathbb{1}^T + I\right)$$

$$= \frac{1}{(k-1)^2}\left((k-2)\mathbb{1}\mathbb{1}^T + I\right) \quad \text{for } k \geq 3, \text{ this} > 0$$

[but, for $k=2$, it is not aperiodic : $A^2 = I$]

thm.: if a finite M.C. is ergodic (regular)

then $\exists$ a unique stationary dist. π

and for any starting dist. π_0, $\lim_{t\to\infty} A^t \pi_0 = \pi$

the speed of convergence is related to the mixing time τ of the chain

$$\tau \triangleq \frac{1}{1 - |\lambda_2(A)|}$$

(λ_2 : 2nd biggest eigenvalue of A)

$$\| A^t \pi_0 - \pi \|_1 \leq C \exp(-t/\tau)$$ (not 100% sure)

14h32

after τ steps, error decreases $\frac{1}{e}$

⊛ intuition (from linear algebra) [informal argument]

simpler case, suppose A is diagonalizable with orthogonal matrix U (here A symmetric)

$$A = U \Sigma U^T \quad \text{with} \quad \Sigma = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix}$$

$U \rightarrow$ basis of e-vectors

$U = (u_1, \dots u_k)$

[by Perron-Frobenius thm., $\lambda_1 = 1 > |\lambda_2| \geq \dots \geq |\lambda_k|$

$u_1 = \frac{\pi}{\|\pi\|_2}$ $A\pi = \pi$

$U^T U = U U^T = I$

let α_0 be coordinate of π_0 in U basis

ie. $\pi_0 = U \alpha_0$ $\quad (\alpha_0 = U^T \pi_0)$

ie. $\pi_0 = U\alpha_0$ $\quad (\alpha_0 = U^T\pi_0)$

$A^t\pi_0 = (U\Sigma U^T)(U\Sigma U^T)\cdots(U\Sigma U^T)(U\pi_0)$ $\quad (\alpha_0)_1 = \langle \pi_0, \frac{\pi}{\|\pi\|_2}\rangle$

$= U\Sigma^t\alpha_0$ $\quad \Sigma^t = \begin{pmatrix} \lambda_1^t & & 0 \\ & \ddots & \\ 0 & & \lambda_K^t \end{pmatrix}$

$A^t\pi_0 = (\alpha_0)_1 \lambda_1^t u_1 + (\alpha_0)_2 \lambda_2^t u_2 + \dots + (\alpha_0)_K \lambda_K^t u_K$

$u_1 = \frac{\pi}{\|\pi\|_2}$, $\lambda_2^t \to 0$, $\lambda_K^t \to 0$

(if $\frac{(\alpha_0)_1}{\|\pi\|_2} = 1$) [sloppy]

$\|A^t\pi_0 - \pi\|_2 = \|(\alpha_0)_2\lambda_2^t u_2 + \dots\| \le C|\lambda_2|^t$ $\quad$ first e-gap

$|\lambda_2| = 1 - \varepsilon_1$ $\quad \varepsilon_1 \triangleq 1 - |\lambda_2|$

$|\lambda_2| \le \exp(-\varepsilon_1)$ $\quad 1 - x \le \exp(-x)\ \forall x$

$|\lambda_2|^t \le \exp(-t\varepsilon_1)$

$\frac{1}{\gamma} \Rightarrow \gamma = \frac{1}{1-|\lambda_2|}$ //

(*) mixing time is often (usually) exponentially big ?

(*) How we design A s.t. $A^t\pi_0 \to \pi$?

one "easy way"

reversible M.-C. $\Leftrightarrow$ iff $\exists$ dist. π s.t. $A_{ij}\pi_j = A_{ji}\pi_i$ $\forall i,j$

"detailed balance equation"

This is sufficient condition to get $A\pi = \pi$ (but not necessary)

it means when $P\{X_{t-1} = i\} = \pi_i$ then $P\{X_t = i, X_{t-1} = j\} = P\{X_t = j, X_{t-1} = i\}$

proof: $(A\pi)_i = \sum_j A_{ij}\pi_j \overset{\text{detailed balance}}{=} \sum_j A_{ji}\pi_i = \pi_i \underbrace{\left(\sum_j A_{ji}\right)}_{1}$ //

Metropolis-Hastings alg.:

goal $\rightsquigarrow$ construct a M.C. with stat dist. $p(x)$ [our target]

[assume $p(x) > 0\ \forall x$]

use some proposal $q(x'|x)$ $\quad$ does not depend on normalizing cst of p

use same proposal $q(x'|x)$

accept new state x' with prob. $a(x'|x) \triangleq \min\left\{1, \frac{q(x|x')p(x')}{q(x'|x)p(x)}\right\}$ (dependence on normalizing const. of p)

if reject it → stay in same state

[this is still new sample vs. rejection sampling where only "accepted states" are new samples]

acceptance ratio to satisfy detailed balance

M.-H. alg.

start at $x^{(0)}$

for $t=1,\dots,$ (important design choice)

- propose $x'^{(t)} \sim q(x'|x^{(t-1)})$
- flip a biased coin with prob. $a(x'^{(t)}|x^{(t-1)})$ to be 1
- if accept (coin=1) let $x^{(t)} = x'^{(t)}$

  o.w. $x^{(t)} = x^{(t-1)}$

end for

note: for symmetric proposal $q(x'|x) = q(x|x')$, always accept if $p(x') \geq p(x)$

↝ like a noisy hill-climbing alg.

[Metropolis alg.]

[verify as exercise that it satisfies detailed balance with $\pi = p$ (target)]

⊛ for convergence is: if M.H. chain is ergodic, then we converge to unique stat dis p

sufficient conditions — for irreducibility $q(x'|x) > 0 \ \forall x, x' \in \mathcal{X}$

for aperiodicity either $q(x|x) > 0$ for some $x \in \mathcal{X}$

$\Downarrow$ $A_{ii} > 0$ for some i

or $a(x'|x) < 1$ for some $x_i \in \mathcal{X}$ x' s.t. $q(x'|x) > 0$

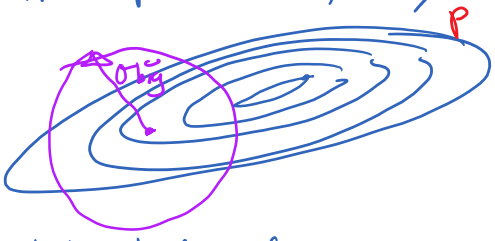
aside: it is still ok. to change proposal with time (inhomogeneous M.C.) $q_t(x'|x)$

as long as choice of q_t does not depend on $x^{(t-1)}$

then convergence theory above will go through

slow mixing example

Suppose p is a $N(\mu, \Sigma)$ & $q(x'|x) = N(x'|x, \sigma^2 I)$



*high prob of rejection

"small moves" ⇒ many steps needed

here the best mixing time is related to ratio $\frac{\sigma_{max}}{\sigma_{min}}$

reference for mixing times:

Markov Chains and Mixing Times
David A. Levin, Yuval Peres, Elizabeth L. Wilmer
https://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf