Lecture 20 - MCMC, Markov chains

Thursday, November 16, 2023 2:30 PM

today: • MCMC
• review Markov chains
• M-H. alg.

MCMC - Markov chain Monte-Carlo

idea: is to relax indep assumption 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 a chain"

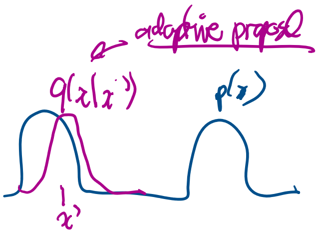
then we can approximate

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

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 $X^{(t)} \overset{\text{iid}}{\sim} q$

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

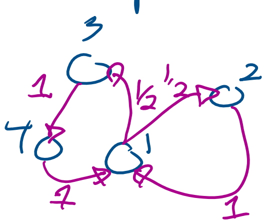
Markov transition prob

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

• as DGM $X^{(0)} \to X^{(1)} \to \dots \to X^{(t-1)} \to X^{(t)}$

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

e.g. $k=4$



[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. $\underbrace{\mathbb{1}_K^T}_{\text{vector of ones of size } K} A = \mathbb{1}_K^T$

"left-stochastic matrix"

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

$$P\{X_t = i\} = \sum_j \underbrace{P\{X_t = i \mid X_{t-1} = j\}}_{A_{ij}} \underbrace{P\{X_{t-1} = j\}}_{\pi_j}$$

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

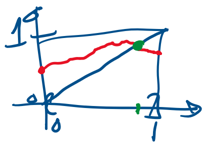
$$\Rightarrow \boxed{\pi_t = A^t \pi_0}$$

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

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

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

(by Brower's fixed pl. thm.



def: **irreducible Markov chain** $\Leftrightarrow$ there exists a positive prob "path" from any i to any j (states)

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

(by Perron-Frobenius thm. $\Rightarrow$ 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. $\Leftrightarrow$ $\exists$ an integer m s.t. $(A^m) > 0$

aka **regular M.C.** (finite state space) $\qquad ((A^m)_{ij} > 0 \ \forall i,j)$

or **ergodic M.C.**

(*) [note: a sufficient condition for an irreducible M.C. to] be aperiodic is $\exists\, i$ s.t. $A_{ii} > 0$

example of a regular M.C. on k states ($k \geq 3$)

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

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

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

$m = 2$

[but, for $k=2$ it is not aperiodic $A^2 = I$ $\quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$]

> **thm.** if a finite M.C. is ergodic (regular)
> then $\exists$ a unique stat. 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

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

$\uparrow$ 2nd largest e-value of A

$$\boxed{\|A^t \pi_0 - \pi\|_1 \leq C \exp(-t/\tau)}$$

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

15h24

(*) intuition (from linear algebra) [informal argument]

simpler case, suppose A is symmetric & psd

by spectral thm, A is diagonalizable with orthogonal U

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

[by Perron-Frobenius thm

$\lambda_1 = 1 > |\lambda_2| \geq \dots \geq \lambda_K$

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

$U \rightarrow$ orthonormal basis of e-vectors of A

$UU^T = U^TU = I \qquad U = (u_1 \cdots u_K)$

let α_0 be coordinates of π_0 in U basis let α_0 be coordinates of π_0 in U basis

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

$$A^t\pi_0 = (U\Sigma \underbrace{U^T)(U}_{I}\Sigma \underbrace{U^T)}_{I}\cdots(U\Sigma \underbrace{U^T)(U}_{I}\alpha_0)$$

$$= U\Sigma^t\alpha_0 \qquad \Sigma^t = \begin{pmatrix} 1 & & & 0 \\ & \lambda_2^t & & \\ & & \lambda_3^t & \\ 0 & & & \ddots \; \lambda_K^t \end{pmatrix}$$

$$= (\alpha_0)_1 1^t \underbrace{u_1}_{\frac{\pi}{\|\pi\|_2}} + (\alpha_0)_2 \lambda_2^t u_2 + \dots + (\alpha_0)_K \lambda_K^t u_K$$

$\lambda_2^t \to 0$, $\lambda_K^t \to 0$

$\|A^t\pi - \pi\|_2 =$ (if $\frac{(\alpha_0)_1}{\|\pi\|_2} = 1$ not true in general)

$\|(\alpha_0)_2 \lambda_2^t u_2 + \dots \| \leq C|\lambda_2|^t$

first e-gap

$|\lambda_2| = 1 - \varepsilon_1$

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

$1+x \leq \exp(x) \; \forall x \qquad |\lambda_2| \leq \exp(-\varepsilon_1)$

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

$\varepsilon_1 = \frac{1}{\tau}$ $\Rightarrow \tau = \frac{1}{1-|\lambda_2|}$ //

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

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

one "easy way" reversible M.C. $\Leftrightarrow$ $\exists$ dist. π s.t. $\boxed{A_{ij}\pi_j = A_{ji}\pi_i \; \forall (i,j)}$

"detailed balance equation"

this is sufficient condition to get $A\pi = \pi$

It means if $P\{X_t = i\} = \pi_i$

then $P\{X_t = i, X_{t-1} = j\}$
$= \mathbb{P}\{X_t = j, X_{t-1} = i\}$

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

Metropolis-Hastings alg.:

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

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

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

no dependence on normalization of p

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

accept new state x' with prob. $\boxed{a(x'|x) \triangleq \min\{1, \frac{q(x|x')\,p(x')}{q(x'|x)\,p(x)}\}}$

if reject → stay in same state x

[this is still new sample]

acceptance ratio to satisfy detailed balance

vs.

rejection sampling where only "accepted states" are samples

M.H. alg.:

start at $x^{(0)}$

important design choice

for $t=1,\ldots,$

- propose $x'^{(t)} \sim q(x' | x^{(t-1)})$

  flip a coin, with prob. $a(x'^{(t)} | x^{(t-1)})$ to be 1

- if accept (coin=1)

  let $x^{(t)} = x'^{(t)}$

  o.w.

  let $x^{(t)} = x^{(t-1)}$

end for

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

[Metropolis alg.] $\rightsquigarrow$ like noisy hill-climbing

[verify as exercise that M.H. satisfies detailed balance eq. with $\pi = p^{\text{target}}$]

(*) for convergence: if M.H chain is ergodic, then converge to unique stat. dist. p

sufficient conditions → for irreducible $q(x'|x) > 0 \;\; \forall x' \neq x \in \mathcal{X}$

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

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

⇓

$A_{ii} > 0$ for some i

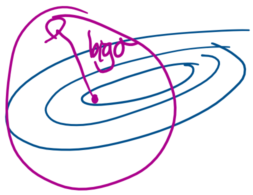
(*) aside: it is still ok change proposal with time

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

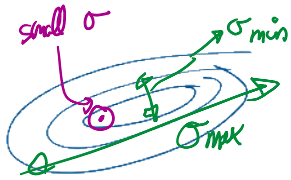
then convergence theory above will go through

Slow mixing example

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



* high prob rejection



"small moves" ⇒ many steps needed

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

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