Lecture 21 - MCMC, Markov chains

November 19, 2021 3:02 PM

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

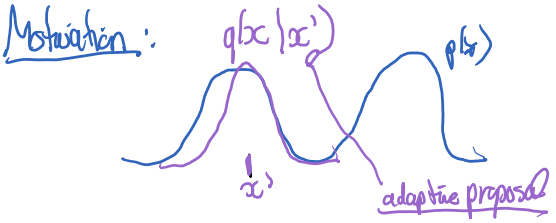
MCMC - Markov chain Monte Carlo

idea: is to relax indep assumption samples
to allow adaptive proposal dist.
ie. 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
$\mathbb{E}_p[f(x)]$ 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 [ie. 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:

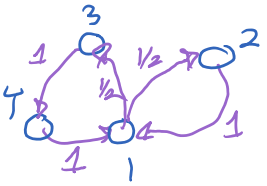


before: samples were $X^{(t)} \overset{iid}{\sim} q$
MCMC $X^{(t)} \mid X^{(t-1)} \sim q(\cdot \mid X^{(t-1)})$
Markov transition proba

Review of (finite state space) Markov chains [$|\mathcal{X}| = K$]

• as a DGM, $x^{(0)} \to x^{(1)} \to x^{(2)} \to \cdots \to x^{(t-1)} \to x^{(t)}$

• there is also transition proba pf. of vars: use one per state
(probabilistic FSA) e.g. $K=4$



[homogeneous M.C.]
↳ ie. $P\{X_t = i \mid X_{t-1} = j\} = A_{ij}$ (no time dep.)
A is a $K \times K$ matrix s.t. $\mathbb{1}_K^T A = \mathbb{1}_K^T$
"left-stochastic matrix"

A is a $k \times k$ matrix s.t. $\mathbb{1}_k' A = \mathbb{1}_k'$

"left-stochastic matrix" — vector of ones of size k

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

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

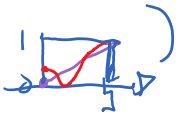
$$\pi_{t+1} = A\pi_t$$

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

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

[note that π is a right-eigenvector 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 $\Leftrightarrow$ there exists a positive-proba "path" from any i to any j (states)

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

(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 (positive) integer m s.t. $(A^m) > 0$ $\left((A^m)_{ij} > 0 \ \forall i,j\right)$

aka regular M.C. (finite state space)

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

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

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

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

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

[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, $\lim_{t\to\infty} A^t \pi_0 = \pi$

16h03

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

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

λ_2 is 2nd biggest e-value of A

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

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

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

simpler case, suppose A is symmetric $\geq$ psd, by spectral theorem, A is diagonalizable with orthogonal matrix U

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

$U \to$ basis of e-vectors (orthonormal basis)

$U U^T = U^T U = I \quad U = (u_1, \dots u_k)$

[by Perron-Frobenius thm

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

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

let α_0 be coordinates of π_0 in U basis

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

$$A^t \pi_0 = (U\Sigma U^T)(U\Sigma U^T)\cdots(U\Sigma U^T)(U\alpha_0)$$

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

$$A^t \pi_0 = (\alpha_0)_1 1^t u_1 + (\alpha_0)_2 \lambda_2^t u_2 + \dots + (\alpha_0)_k \lambda_k^t u_k$$

$u_1 = \frac{\pi}{\|\pi\|_2}$; $\lambda_2^t \to 0$; $\lambda_k^t \to 0$

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

first e-gap

$\|A^t\pi - \pi\|_1 = \|(\alpha_0)_2 \lambda_2^t u_2 + \dots\| \leq C|\lambda_2|^t$ $\quad |\lambda_2| = 1 - \underbrace{\varepsilon_1}_{\triangleq 1-|\lambda_2|}$

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

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

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

(*) 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. $\Leftrightarrow \exists$ distr. π s.t. $\boxed{A_{ij}\pi_j = A_{ji}\pi_i}\ \forall (i,j)$

"detailed balance equation"

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

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 → construct a M.C. with stat. dist $p(x)$ [can target]

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

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

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

if reject it → stay in same state

no dependance on normalization of p

acceptance ratio to satisfy detailed balance

[this is still new sample]

Vs. rejection sampling where only "accepted states" are new samples.

M.H. alg

start at $x^{(0)}$

for $t=1\dots$

important design

start at $x^{(0)}$

for $t=1, \ldots,$

- propose $x'^{(t)} \sim q(x' | x^{(t-1)})$ (important design choice)
- 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.
  $x^{(t)} = x^{(t-1)}$

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 M.H. satisfies the detailed balance eq. with $\pi = p_{\theta}$ target]

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

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

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

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

$\Downarrow$

$A_{ii} > 0$ for some i

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

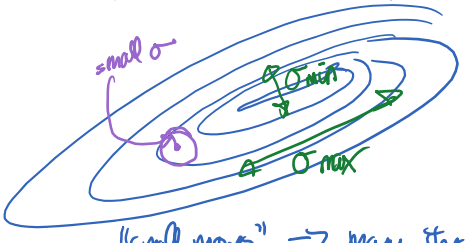
as long as choice of q_t does not depend on $x^{(t-1)}$

then convergence theory (above) will go through

slow mixing example:

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

* 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