Lecture 20 - scribbles

Friday, November 22, 2019 14:04

today: MCMC
review Markov chains
Metropolis-Hastings

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 $X_t | X_{t-1}$ s.t. $X_t \xrightarrow{t\to\infty}$ in dist. to target dist. p
"stationary dist. of chain"

then we can approximate

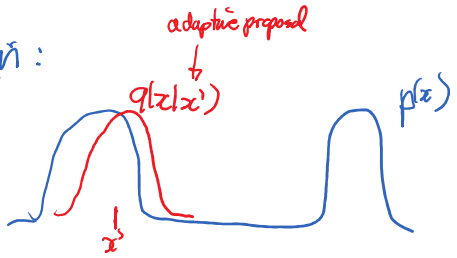
$$\mathbb{E}_p[f(x)] \text{ as } \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} f(x_t)$$

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

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

as this yields higher variance
→ better to use all samples after T_0 (unless it is too expensive)

Motivation:



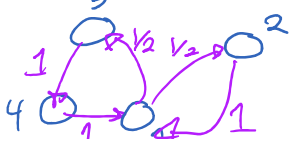
before: samples were $x^{(t)} \overset{iid}{\sim} q$

MCMC $x^{(t)} | x^{(t-1)} \sim q(\cdot | x^{(t-1)})$
↑ Markov transition prob.

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

- as a DGM, 

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

e.g. 4 states 

[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. $\mathbb{1}_k^T A = \mathbb{1}_k$

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

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

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

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

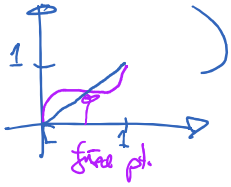
$$\Rightarrow \boxed{\pi_t = A^t \pi_0}$$

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

[note that π is a right 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** $\Leftrightarrow$ there exists a positive prob "path" from every i to j

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

(by Perron-Frobenius thm) $\Rightarrow$ irreducible M.C. has a unique stationary dist.

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 MC** or **ergodic M.C.** (finite state) (i.e. $(A^m)_{ij} > 0 \ \forall i,j$)

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

eg. of a regular M.C. $A = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix} = \frac{1}{2}(\mathbb{1}\mathbb{1}^T - I)$

$A^2 = \frac{1}{4}(\underbrace{\mathbb{1}\mathbb{1}^T\mathbb{1}\mathbb{1}^T}_{K} - 2\mathbb{1}\mathbb{1}^T + I)$

$= \frac{1}{4}((K-2)\mathbb{1}\mathbb{1}^T + I)$ for $k \geq 3$, this > 0

[but for $k=2$, it is not aperiodic] (fix this for $k \neq 3$ (2))

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

$\tau \triangleq \frac{1}{1 - |\lambda_2(A)|}$ ← 2nd biggest evalue of A

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

⊛ intuition (from linear algebra) [informal argument]

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

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

$U \to$ basis of evectors

$U = (u_1 \dots u_K)$

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

key Perron-Frobenius thm, $\lambda_1 = 1 > |\lambda_2| \geq \dots \geq |\lambda_K|$

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

Let α_0 be coordinate of π_0 in U basis i.e. $\pi_0 = U\alpha_0$ $(\alpha_0 = U^T \pi_0)$

$(\alpha_0)_1 = \langle \pi_0, \frac{\pi}{\|\pi\|_2} \rangle$

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

$= U \Sigma^t \alpha_0$

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

$A^t = (\quad)^t \dots (\quad)^t$

$= U \Sigma^t \alpha_0$

$A^t \pi_0 = (\alpha_0)_1 \lambda_1^t u_1 + (\alpha_0)_2 \lambda_2^t u_2 + \dots + (\alpha_0)_k \lambda_k^t u_k$ $\quad \left(\begin{pmatrix} \lambda_2^t & & \\ & \ddots & \\ & & \lambda_k^t \end{pmatrix}\right)$

$\lambda_2 \to 0$, $\lambda_k \to 0$

$A^t \pi_0 \xrightarrow{t \to \infty} (\alpha_0)_1 u_1 = \frac{\pi}{\|\pi\|_1}$ $\quad \|A^t \pi_0\|_1 = 1 \ \forall t \Rightarrow (\alpha_0)_1 = \frac{1}{\|\pi\|_1}$

$\|A^t \pi_0 - \pi\|_1 = \|(\alpha_0)_2 \lambda_2^t u_2 + \dots\| \le C |\lambda_2|^t$ $\quad |\lambda_2| = 1 - \varepsilon_1$ (first e-gap) $\quad \varepsilon_1 \triangleq 1 - |\lambda_2|$

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

$|\lambda_2|^t \le \exp(-t\varepsilon_1)$ $\Rightarrow \tau = \frac{1}{1 - |\lambda_2|}$ (mixing time $\approx \frac{1}{\varepsilon_1}$) //

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

15h16

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

one "easy way"

reversible M.C. $\Leftrightarrow$ iff $\exists$ dist. π $\boxed{A_{ij} \pi_j = A_{ji} \pi_i}$ $\forall (i,j)$

"detailed balance equation"

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

(when $P\{X_{t-1} = i\} = \pi_i$)

this is

↳ sufficient condition (but not necessary)

$A\pi = \pi$

proof: $(A\pi)_i = \sum_j A_{ij} \pi_j = \sum_j A_{ji} \pi_i = \pi_i \left(\sum_j A_{ji}\right) = \pi_i \cdot 1$ //

(detailed balance)

Metropolis-Hastings alg.

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

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

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

note $p(x')/p(x)$ does not depend on normalization of p

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

- accept new state x' with prob
- if reject it → stay in same state x

[this still a raw sample]

vs. rejection sampling where only "accepted states" are samples

$$a(x'|x) \triangleq \min\left\{1, \frac{q(x|x')p(x')}{q(x'|x)p(x)}\right\}$$

does not depend on normalization of p

acceptance ratio to satisfy detailed balance

M.H. alg.:

start at $x^{(0)}$

for $t=1, \ldots,$

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

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

end for

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

↝ like a noisy hill-climbing alg

[Metropolis alg.]

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

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

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 and x'

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

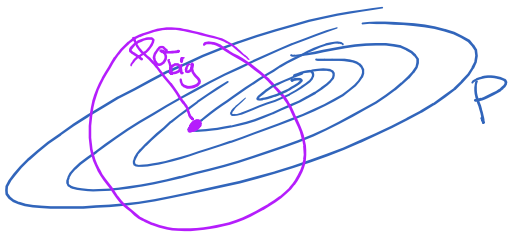
as long as choice of q_t does not depend on $x^{(t-1)}$

then convergence theory above will go through

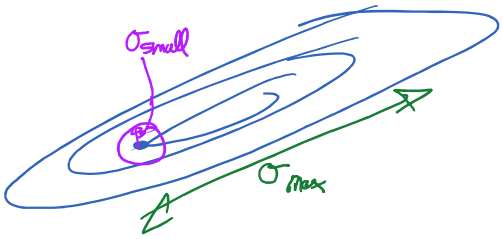
[i.e. detailed balance, etc. will give right stat. distribution]

slow mixing example

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



* high prob of rejection



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

good book: Casella & Berger
"Monte Carlo Statistical methods"