Lecture 19 - scribbles

Tuesday, November 14, 2017
14:29

today: MCMC

Markov chain Monte-Carlo:

idea: is to relax independence assumption between samples
to allow adaptature proposal distributions
ie. we'll run a chain $X_t \mid X_{t-1}$ s.t. $X_t \xrightarrow{t\to\infty}$ target distribution p
"stationary dist. of chain"

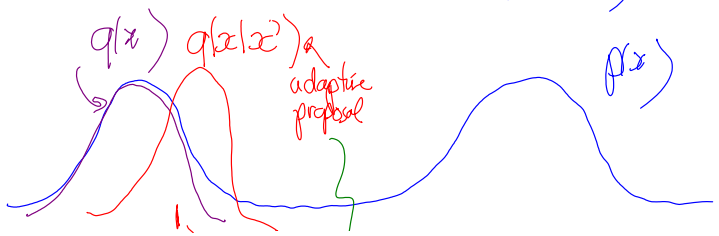
then, we can approximate

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

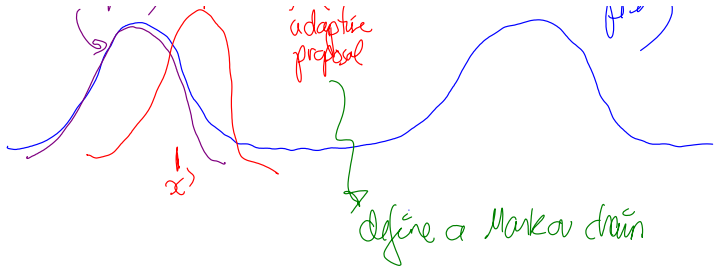
T_0 is called "burn-in" period → depends on "mixing time" of Markov chain

(*) no need to thin the samples [ie. use Δt between samples to get more independence]
as this yields higher variance
→ better to use all samples after T_0 (unless it is too expansive)

Motivation:



Motivation:



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

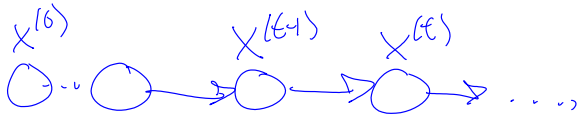
MCMC: $X^{(t)} \mid X^{(t-1)} \sim q(\cdot \mid x^{(t-1)})$

↑ Markov transition probability

review of (finite state space) Markov chain

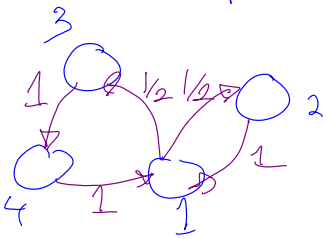
[finite state space $|X| = K$]

- as a DGM



- there is also the transition prob. point of view: use one node per state

(probabilistic finite state automaton (FSA))



4 states example

[homogeneous M.C.]

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

A is a $k \times k$ matrix s.t. $\mathbb{1}_k^T A = \mathbb{1}_k^T$ (sum along a column is equal to 1)

↑ vector of k ones

"left-stochastic matrix"

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

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

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

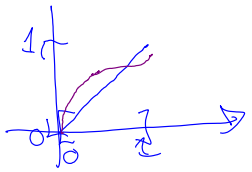
$$\Rightarrow \pi_t = A^t \pi_0$$

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 of 1]

fact: every stochastic matrix has at least 1 stationary distribution

(by Brouwer's fixed pt. thm.)



def:

irreducible MC $\Leftrightarrow$ there is a positive probability "path" from every $i \leftrightarrow j$

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

(by Perron-Frobenius thm.) $\Rightarrow$ unique stationary dist. for irreducible MC.

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

irreducible and aperiodic MC $\Leftrightarrow$ $\exists$ an integer m s.t. $A^m > 0$

aka. regular MC. (i.e. $(A^m)_{ij} > 0$)

ergodic MC.

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

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 e-value of A

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

* intuition (from linear algebra)
suppose A is diagonalizable $A = U \Sigma U^{-1}$ with $\Sigma = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix}$

by Perron-Frobenius thm.
can show that $\lambda_1 = 1 > |\lambda_2| \geq \dots \geq |\lambda_k|$

linear basis of e-vectors
$U = (u_1 \dots u_k)$

take $u_1 = \pi$ [$A\pi = \pi$] (e-value = 1)

let α_0 s.t.
$\pi_0 = U \alpha_0$

$$A^t \pi_0 = (U \Sigma U^{-1})(U \Sigma U^{-1}) \cdots (U \Sigma U^{-1}) \pi_0$$

$$= U \Sigma^t \alpha_0 \qquad (U^{-1}\pi_0 = \alpha_0)$$

$$\Sigma^t = \begin{pmatrix} 1 & & 0 \\ & \lambda_2^t & \\ 0 & & \ddots \lambda_K^t \end{pmatrix}$$

$$A^t \pi_0 = U \Sigma^t \alpha_0 = \left[(\alpha_0)_1 u_1 + (\alpha_0)_2 \lambda_2^t u_2 + \cdots + (\alpha_0)_K \lambda_K^t u_K \right]$$

because $\mathbb{1}^T A^t \pi_0 = 1$ for all $t \Rightarrow (\alpha_0)_1 = 1$

$$\| A^t \pi_0 - \pi \|_1 \leq C |\lambda_2|^t$$

first e-gap

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

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

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

$\varepsilon_1 \to \frac{1}{\tau}$

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

//

mixing time can be exponentially big sometimes

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

one "easy way"

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

"detailed balance equation"

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

[when $P\{X_{t-1} = i\} = \pi_i$]

sufficient condition (but not necessary)

for $A\pi = \pi$

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

Metropolis-Hastings algorithm

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

(assume $p(x) > 0$)

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

[i.e. if in state x, we same $x'|x \sim q(x'|x)$]

- accept new state x' with prob. $a(x'|x) \triangleq \min\left\{1, \frac{q(x|x')p(x')}{q(x'|x)p(x)}\right\}$
- if reject → stay in same state x

depends only on $p(x')/p(x)$ → no need to normalize

acceptance ratio to satisfy detailed balance

[this is still a new sample]

(vs. rejection sampling where only "accepted" states are sample)

alg.: s. 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

if accept (coin = 1)

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

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

end

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

→ like noisy hill-climbing alg.

Let's verify detailed balance $\qquad A_{ij}\pi_j = A_{ji}\pi_i$

- $A_{ii}\pi_i = A_{ii}\pi_i$ trivially
- to have $A_{ij}\pi_j = A_{ji}\pi_i$

need. $q(i|j)\,a(i|j)\,p_j = q(j|i)\,a(j|i)\,p_i$ (target dist.)

$\Rightarrow$ want $\boxed{\dfrac{a(i|j)}{a(j|i)} = \dfrac{q(j|i)p_i}{q(i|j)p_j}}$

to finish: use the $\max\{1, \_\}$ to look at cases

For convergence: if MH chain is ergodic, then we converge to correct unique stationary dist. p

sufficient conditions ← irreducibility $q(x'|x) > 0 \quad \forall x' \neq x \in X$

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

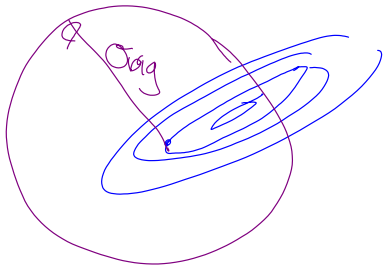
(*) aside: it is still ok to change proposal with time (inhomogeneous M.C.) $q_t(x'|x)$

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

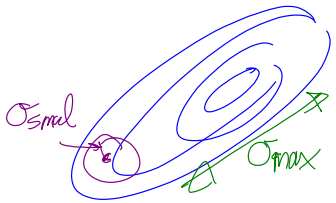
then convergence theory above will go through (i.e. detailed balance etc... will give right stationary dist.)

slow mixing example:

suppose p is multivariate normal & $q(x'|x) = N(x'|x, \sigma^2 I)$



* high prob. of rejection



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

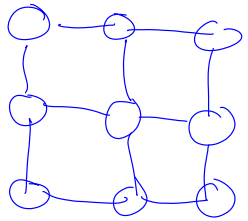
good book: Casella & Berger Monte Carlo Statistical Methods

Gibbs sampling algorithm

MH with clever choice of proposal $q_t(x'|x)$

examples of applications. — UGM: $p(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$

UGM's



difficult conditional in DGM

$\tilde{p}(x) = p(x, \bar{x}_E)$

$\propto p(x \mid \bar{x}_E)$

cyclic Gibbs sampling alg.: nodes $i = 1, \dots, n$

start at some $x^{(t)}$

for $t = 1, \dots,$

- pick $i = (t \bmod n) + 1$
- sample $x_i^{(t)} \sim p(X_i = \cdot \mid X_{\neg i} = x_{\neg i}^{(t-1)})$ ($1:n \setminus \{i\}$; call other nodes except i)
- set $x_j^{(t)} = x_j^{(t-1)}$ for $j \neq i$ (true conditional as proposal)

end