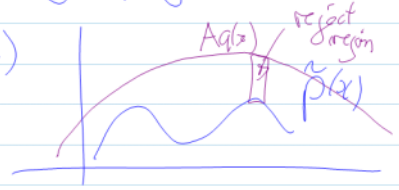
Lecture 20 - scribbles

Friday, November 11, 2016
13:37

rejection sampling:

say $p(x) = \frac{\tilde{p}(x)}{Z_p}$; say we can find $q(x)$ a distribution we can easily sample from

s.t. $Aq(x) \geq \tilde{p}(x)$



rule: · sample $X \sim q(x)$
· Accept with probability $\frac{\tilde{p}(x)}{Aq(x)} \in [0,1]$
(reject o.w.)

say you want to compute $p(x | \bar{x}_E)$

here $\tilde{p}(x) = p(x_{E^c}, \bar{x}_E)\, \delta(x_E, \bar{x}_E)$

here $Z_p = p(\bar{x}_E)$

if you sample from original joint using ancestral sampling (DGM)

$q(x) = p(x_{E^c}, x_E)$

here, we have $q(x) \geq \tilde{p}(x)$ $\forall x$ [take $A=1$]

acceptance prob. $= \frac{\tilde{p}(x)}{q(x)} = \begin{cases} 1 & \text{if } x_E = \bar{x}_E \\ 0 & \text{o.w.} \end{cases}$

[i.e. reject when $x_E \neq \bar{x}_E$]

$P\{\text{accept}\}$ marginally $= \frac{Z_p}{A} = p(\bar{x}_E)$

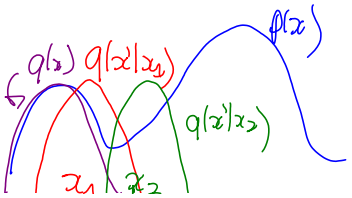
⊛ sidenote: when sample from joint, you are also sampling from marginal

i.e. $(X, Y) \sim p(x, y)$

then looking at X by itself, you have $X \sim p(x)$

MCMC

motivation:



$\| x_1 \| x_2 \|$

use instead adaptive proposal $q(x' | x)$

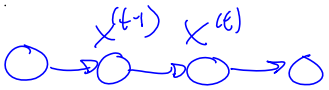
↳ defines a Markov chain

goal is that sample from MC "converges" to correct distribution

before: samples were $x^{(t)} \sim q$ (Markov transition probability)

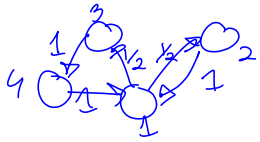
now $x^{(t)} | x^{(t-1)} \sim q(x' | x^{(t-1)})$

review of Markov chains [finite state MC; ie $|X| = K$]

as DGM 

there is also the transition prob view; say one node per states

[homogeneous M.C.]

(probabilistic FSA) 

4 states Markov chain

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

here A is a $K \times K$ matrix s.t. $\mathbb{1}^T A = \mathbb{1}$ (sum along a column $= 1$)

"left stochastic matrix"

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

then, new marginal is just $A\pi$

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

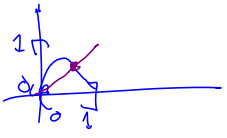
stationary dist π of A is π s.t.

$A\pi = \pi$ [right e-vector of A with e-value of 1]

[this implies that if $P(X_{t-1} = i\} = \pi_i$ then $P(X_t = i\} = \pi_i$]

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

(by Brouwer's fixed pt. thm.)



irreducible M.C. $\Leftrightarrow$ there is a positive probability "path" from every i to j

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

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

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

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

aka. regular MC
or ergodic M.C.

(note: a sufficient condition for being regular is $\exists i$ s.t. $A_{ii} > 0$] for an irreducible MC.)

thm: if a finite M.C. is ergodic (regular)
then $\exists$ a unique stationary dist. π
and for any starting dist. π_0, $A^t \pi_0 \to \pi$

the speed of convergence is related to the mixing time of the chain
which is $\frac{1}{1-|\lambda_2(A)|}$ ← 2nd biggest e-value of A

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

intuition for this (linear algebra)

suppose A is diagonalizable $A = U \Sigma U^{-1}$ with $\Sigma = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix}$

linearly independent e-vectors
$U = (u_1 \cdots u_k)$

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

take $u_1 = \pi$ [$A\pi = \pi$]

this is basis

let α_0 s.t. $\pi_0 = U\alpha_0$

$$A^t \pi_0 = (U\Sigma U^{-1})(U\Sigma U^{-1}) \cdots (U\Sigma U^{-1}) \overbrace{U\alpha_0}^{\pi_0}$$

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

$$= [(a_0)_1 \pi + \lambda_2^t (a_0)_2 u_2 + \dots + \lambda_K^t (a_0)_K u_K]$$

because $\mathbb{1}^T \pi_0 = 1$

$$\| A^t \pi_0 - \pi \|_1 \le C |\lambda_2|^t \qquad |\lambda_2| = 1 - \varepsilon_1 \qquad \varepsilon_1 = 1 - |\lambda_2|$$

$$|\lambda_2| \le \exp(-\varepsilon_1)$$
$$|\lambda_2|^t \le \exp(-\varepsilon_1 t) \to 1$$

$$\gamma = \frac{1}{1-|\lambda_2|}$$ mixing time

↳ can be exponentially big sometimes

(*) How do we design for π?

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

"detailed balance equation"

it means

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

$\Rightarrow$ that $A\pi = \pi$

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

note: detailed balance is a sufficient condition for stationarity but is not necessary

Metropolis-Hasting algorithm

→ construct a MC with stationary distribution $p(x)$ [our target] (assume $p(x) > 0$)

we consider proposal $q(x'|x)$

[ie. if in state x, we sample $x'|x \sim q(x'|x)$]

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

ratio of p ⇒ no need to normalize

acceptance ratio to satisfy detailed balance

reject ⇒ stay in same state x

[this is still a new sample.

[this is still a new sample.

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

alg.: start at $x^{(0)}$

for $t = 1, \ldots$

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

flip a biased coin with prob $a(x'^{(t)} | x)$

if accept

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

o.w.

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

end

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

↪ noisy hill-climbing alg.

let's verify detailed balance:

$A_{ii}\pi_i = A_{ii}\pi_i$ trivially

here $A_{ij} = q(i|j)\,a(i|j)$ for $i \neq j$

to have $A_{ij}\pi_j = A_{ji}\pi_i$ (target dist.)

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

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

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

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

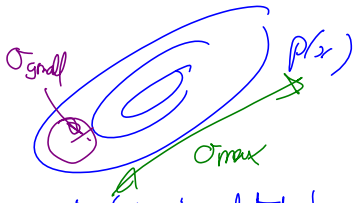
aperiodicity $q(x|x) > 0$ for some x

⊛ aside: it is still ok to change proposal with time (inhomogeneous MC) $q_t(x'|x)$ as long as choice of q_t does not depend on $x^{(t-1)}$

slow mixing example:

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

→ high prob of rejection



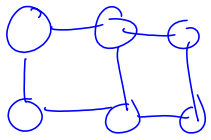
here the mixing time is related to ratio $\frac{\sigma_{max}}{\sigma_{small}}$

good book Casella & Berger Monte Carlo Statistical Methods

Gibbs sampling algorithm:

→ canonical choice of proposal $q_t(x'|x)$

UGM:



examples → UGM: $\tilde{p}(x) = \prod_c \psi_c(x_c)$

difficult conditional in DGM $p(x) = p(x, \bar{x}_E)$ & $p(x|\bar{x}_E)$

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

start at some $x^{(0)}$

for $t = 1, \dots$

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

true conditional as proposal

this is M-H:

with time varying proposal
suppose we pick i at time t

force rest to be constant

then $q_t(x'|x^{(t-1)}) = p(x_i'|x_{\neg i}^{(t-1)}) \delta(x_{\neg i}', x_{\neg i}^{(t-1)})$

acceptance ratio: $\frac{q_t(x^{(t-1)}|x')\,p(x')}{q_t(x'|x^{(t-1)})\,p(x^{(t-1)})}$

$p(x') = p(x_{\neg i}')\,p(x_i'|x_{\neg i}')$

$p(x^{(t-1)}) = p(x_{\neg i}^{(t-1)})\,p(x_i^{(t-1)}|x_{\neg i}^{(t-1)})$

here $x_{\neg i}^{(t-1)} = x_{\neg i}'$

$= 1$ always accepts? //

convergence:

let A be one full cycle (n steps)
→ homogeneous M.C.

A is irreducible & aperiodic since can get to any state with n flips

$\Rightarrow A^t \pi_0 \xrightarrow{t \to \infty} p$

☺ also works for random scan (pick $i \sim \text{Unif}(1:n)$ at each step)