# Lecture 20 - MCMC, Markov chains

Friday, November 18, 2022 3:00 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

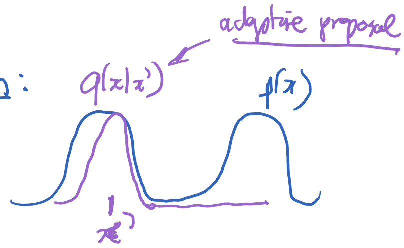
(*) no need to thin the samples [i.e. use $\Delta t$ between samples to get more independence]

as this yield higher variance

→ better to use all samples after $T_0$ to estimate $\mu$

[unless it is too expensive]

motivation:



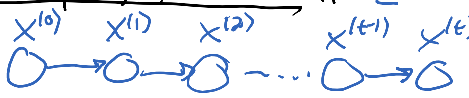

before: samples were $X^{(t)} \overset{iid}{\sim} q$

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

Markov transition prob

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

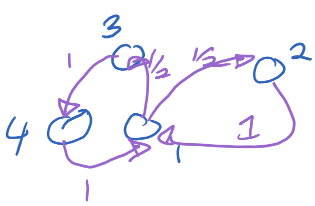
• as DGM,




• there is also transition prob. pt. of view: use one node 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. $\underbrace{\mathbb{1}_k^T}_{\text{vector of ones of size } k} A = \mathbb{1}_k^T$

"left-stochastic matrix"

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

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

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

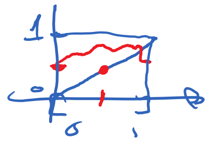
$$\Rightarrow \boxed{\pi_t = A^t \pi_0}$$

$\boxed{\text{stationary dist. } \pi \text{ of } A \text{ is a dist. } \pi \text{ s.t. } A\pi = \pi}$

[note that this $\pi$ 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 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 (positive) integer $m$ s.t. $(A^m) > 0$

$((A^m)_{ij} > 0 \ \forall i,j)$

a.k.a. regular M.C. (finite state space)

or ergodic M.C.

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

be aperiodic ie. $\exists i$ s.t. $A_{ii} > 0$

example of a regular M.C on k states

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

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

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

[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. $\pi$
and for any starting dist. $\pi_0$, $\lim_{t\to\infty} A^t \pi_0 = \pi$

the speed of convergence is related to the mixing time $\tau$ of the chain

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

$\lambda_2$: 2nd biggest e-value of A

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

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

16h05

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

simpler case, suppose A is symmetric & psd., by spectral thm., A is diagonalizable with orthogonal matrix 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.

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

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

U → orthonormal basis of e-vectors

$UU^T = U^TU = I$ $\quad U = (u_1 \cdots u_k)$

let $\alpha_0$ be coordinates of $\pi_0$ in U basis

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

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

$$= U\Sigma^t \alpha_0$$

$$\Sigma^t = \begin{pmatrix} 1 & & & & 0 \\ & \lambda_2^t & & & \\ & & \lambda_3^t & & \\ & & & \ddots & \\ 0 & & & & \lambda_k^t \end{pmatrix}$$

$$= (\alpha_0)_1 1^t \underbrace{u_1}_{\frac{\pi}{\|\pi\|_2}} + (\alpha_0)_2 \lambda_2^t u_2 + \ldots + (\alpha_0)_k \lambda_k^t u_k$$

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

$\left(\text{if } \frac{(\alpha_0)_1}{\|\pi\|_2} = 1\right)$ [fishy; not true in general]

$$\|A^t\pi - \pi\|_1 = \|(\alpha_0)_2 \lambda_2^t u_2 + \ldots\| \leq C|\lambda_2|^t$$

first e-gap

$|\lambda_2| = 1 - \varepsilon_1$

$\varepsilon_1 = 1 - |\lambda_2|$

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

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

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

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

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

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

"detailed balance equation"

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

it means $P\{X_{t-1} = i\} = \pi_i$

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

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

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

## 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 $q(x'|x)$

- 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 $\to$ stay in same state $x$

no dependence on normalization of $p$

acceptance ratio to satisfy detailed balance

[this is still new sample

vs. rejection sampling where only "accepted states" are new sam.]

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

## M.H. alg.

start at $x^{(0)}$

for $t = 1, \dots,$

- propose $x'^{(t)} \sim q(x' \mid x^{(t-1)})$ — important design choice

  flip a coin, with prob. $a(x'^{(t)} \mid x^{(t-1)})$ 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 $q(x' \mid x) = q(x \mid x')$, always accept if $p(x') \geq p(x)$ → like noisy hill-climbing alg.

[Metropolis alg.]

[verify as exercise that M.H. satisfies the detailed eq. with $\pi = p$ (target)]

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

sufficient conditions
- for irreducibility: $q(x' \mid x) > 0 \ \forall x' \neq x \in \mathcal{X}$
- for aperiodicity ($\Downarrow$ $A_{ii} > 0$ for some $i$)
  - either $q(x \mid x) > 0$ for some $x \in \mathcal{X}$
  - or $a(x' \mid x) < 1$ for some $x \in \mathcal{X}$ and $x'$ s.t. $q(x' \mid x) > 0$

(*) aside: it is still ok to change proposal with time $q_t(x' \mid x)$

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 $N(\mu, \Sigma)$ & $q(x' \mid x) = N(x' \mid x, \sigma^2 I)$


"small moves" → many steps needed

* high prob. of rejection

"small moves" $\Rightarrow$ many steps needed

here best mixing time 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