Today: MCMC

Markov Chain Monte Carlo:

Idea: is to relax independence assumption between samples to allow adaptive proposal distributions.

i.e. we'll run a chain \( X_t / X_{t-1} \) s.t. \( X_t \sim \text{target distribution } p \)

Then, we can approximate

\[
\text{Ep}[f(X)] \approx \frac{1}{T-t_0} \sum_{t=t_0}^{T} f(X_t)
\]

To is called "burn-in" period \( \Rightarrow \) depends on "mixing time" of Markov chain.

\( \Rightarrow \) no need to thin the samples \( \Rightarrow \) i.e. use \( X_t \) between samples to get more independence.

Motivation: \( q(x) \) \( \rightarrow \) \( \pi(x|x_{-1}) \)

Adaptive proposal...
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

- as a DOM $\cdots \rightarrow \cdots \rightarrow \cdots$

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

(Petrodolic finite state automaton)

Homogeneous M.C.

I.e., $P_x X_t = i \mid X_{t-1} = i \sim A_{i,j}$ (no time dependence)

$A$ is a $k \times k$ matrix, s.t. $\sum_k A_{i,k} = 1$

"Left stochastic matrix"

($\pi$ as in HMM), suppose $P_x X_{t-1} = i \sim \pi(i)$
\[ P(X_t = i | X_0 = j) = \sum_j P(X_{t+1} = i | X_t = j) P(X_t = j) \]

\[ \pi_{t+1} = A \pi_t \]

\[ \Rightarrow \pi_t = A^T \pi_0 \]

**Stationary dist.** \( \pi \) of \( A \) is a dist. \( \pi \) s.t. \( A \pi = \pi \)

[Note that \( \pi \) is a right eigenvector of \( A \) with \( \pi \) value of 1.]

**Fact:** Every stochastic matrix has at least 1 stationary distribution

(key Brownian's fixed pt. thm.)

**Def.:**

**Irreducible MC** \( \iff \) there is a positive probability path from every \( i \) to every \( j \)

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

(key 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.** \( \iff \exists \text{ an integer } m \text{ s.t. } A^m > 0 \)

(i.e. \( (A^m)_{ij} > 0 \))

**aka.** **Regular MC.**
Theorem: If a finite M.C. is ergodic (regular)
then \( \exists \) a unique stationary 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 \( T \) of the chain:

\[
\gamma \leq \frac{1}{2 - \lambda_2 \|A\|}
\]

2nd largest e-value of A

\[
\|A^t \pi_0 - \pi\|_1 \leq C \exp(-t \gamma)
\]

* Intuition (from linear algebra)
  suppose \( A \) is diagonalizable \( A = U\Sigma U^T \) with \( \Sigma = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_k \end{pmatrix} \)

  - linear basis of e-vectors
    \( U = (u_1, \ldots, u_k) \)
  - take \( u_1 = \pi \)
  - let \( \alpha_0 = \pi \cdot u_0 \)

  by Perron-Frobenius thm.
  can show that \( \lambda_1 > |\lambda_2| > \ldots > |\lambda_k| \)
\[ A^{t} \pi_{0} = (U \Sigma U^{*} )(U \Sigma U^{*})^{T} \delta_{0} \]
\[ = U \Sigma \pi_{0} \]
\[ \delta_{t} = \begin{pmatrix} \lambda_{1}^{t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \]
\[ A^{t} \pi_{0} - \pi_{0} \subseteq \left[ \sum (\lambda_{0}) \lambda_{2}^{t} + (\lambda_{0}) \lambda_{2}^{t} u_{2} + \ldots + (\lambda_{0}) \lambda_{2}^{t} u_{k} \right] \]

because \( A^{t} \pi_{0} = \pi_{0} \) for all \( t \) \( \Rightarrow \) \( (\lambda_{0}) = 1 \) \( \delta \) is exponential

\[ ||A^{t} \pi_{0} - \pi||_{1} \leq C ||\lambda_{0}||^{t} \]

\[ ||\lambda_{0}|| = 4 \epsilon \]

\[ ||\lambda_{0}|| \leq \exp(-C_{A} t) \]

\[ ||\lambda_{0}|| \leq \exp(-\gamma t) \]

\[ \Rightarrow \gamma = \frac{4}{1-1/4} \]

missing line can be exponentially big sometimes

\( \bigcirc \) how do we design A s.t. \( A^{t} \pi_{0} \rightarrow \pi \)?

one "easy way"

\( \text{Reversible M.C. if } \exists \text{ dist. } \pi \text{ s.t. } \exists A_{i} \pi_{i} = A_{i} \pi_{i} \forall (i,j) \)
sufficient condition (not necessarily)

\[ \text{Algorithm: Metropolis-Hastings} \]

1. \( a \in \mathbb{R} \), \( x \in A \) 
2. \( r \) is a random draw from \( p(x) \) 
3. \( \alpha_{x \to y} = \min\left\{ 1, \frac{p(y)}{p(x)} \right\} \) 
4. \( \text{accept} \) if \( r \leq \alpha_{x \to y} \) or \( \text{stay in state} \) with prob. \( \alpha_{x \to x} \)
5. \( y \) is state in state space
6. \( \text{accept} \) if \( r \leq \alpha_{x \to y} \) or \( \text{stay in state} \) with prob. \( \alpha_{x \to x} \)

(assumed \( p(x) > 0 \))

\[ \sum_{y \in A} \alpha_{x \to y} = 1 \]
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 $\text{accept}$ ($\text{coin} = 1$)
      - Let $x^{(t)} = x^{(t)}$
    - Else
      - $x^{(t)} = x^{(t-1)}$

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

Let's verify detailed balance

- $A_{ij} = A_{ji}$
- $A_{ii} = A_{ii}$, trivially
- To have $A_{ij} := A_{ji}$

Need: $q(i | j) a(i | j) p_j = q(j | i) a(j | i) p_i$

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

To finish: use the max-$\exists j$ to look at cases

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

Sufficient conditions: irreducibility $q(x' | x) > 0 \ \forall x', x \in X$
\[ \text{apenodicly } q(x|x) > 0 \text{ for some } x \in \mathcal{X} \]

\[ \bigcirc \text{ aside: it is still ok to change proposal with } \]
\[ \text{(inhomogeneous MCMC) } q_\theta(x' \mid x) \]
\[ \text{ as long as } \text{choice of } q_\theta \text{ does not depend on } x \text{ (\textit{e.g.})} \]
\[ \text{then convergence theorem will go through} \]
\[ \text{(i.e. detailed balance etc... will give right stationary dist.)} \]

\[ \text{slow mixing example:} \]
\[ \text{suppose } p \text{ is multivariate normal } \]
\[ q(x' \mid x) = N(x' \mid x, \sigma^2 I) \]

\[ \text{high prob of rejection} \]
\[ \text{here the mixing time is} \]
\[ \text{related to} \]
\[ \text{O} \text{ \textit{max \text{O} \text{ \textit{small}}} \}

\[ \text{good book: } \]
\[ \text{Caffo, } \text{Berger } \text{Monte Carlo Statistical Methods} \]

\[ \text{Gibbs sampling algorithm} \]
\[ MH \text{ with clever choice of proposal } q_\theta(x' \mid x) \]
\[ \text{examples of applications. } \]
\[ \text{\textit{e.g.}} \]
\[ \text{A.M: } p(x) = \int p_0(x) \]

Teaching Page 8
cyclic Gibbs sampling alg:

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

for $t = 1, \ldots, n$

1. pick $i = (t \mod n) + 1$
2. sample $x_i^{(t)} \sim p(x_i = \cdot | x_{-i}^{(t-1)} = x_{-i})$
3. set $x_j^{(t)} = x_j^{(t-1)}$ for $j \neq i$

end