Markov chain Monte Carlo methods

Review

- Sampling distributions (e.g., a posterior) supports estimating maximums and expectations without attempting "exact inference"
- Different from sampling methods discussed previously, MCMC relaxes having independent draws.
  - Independent draws would be preferred, but for complex distributions in high dimensions, we typically do not know how to get **good** independent samples from the distribution.
- Samples are conditioned on previous one(s)
  - Explores promising parts of the space before moving on
  - Associate "states" with emission of a particular sample

Metropolis Example

Review

Get an initial value (state) $z^{(0)}$

While not_bored

{

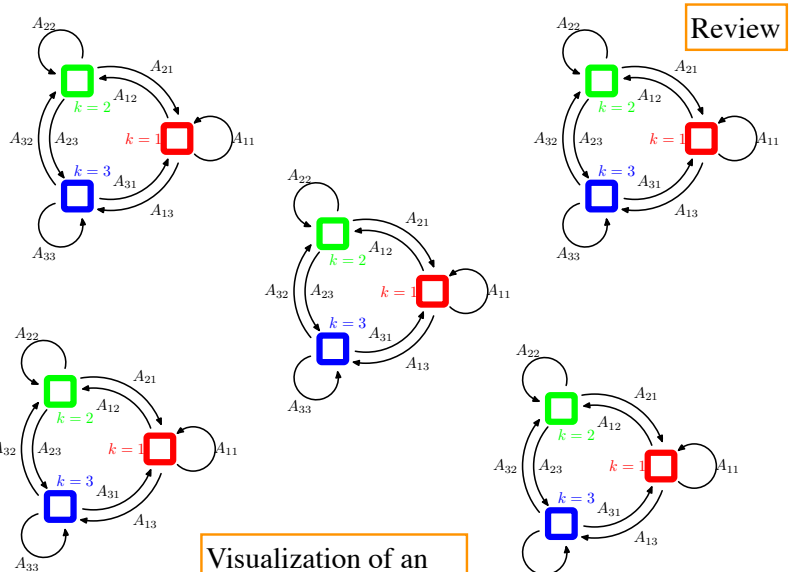
Sample $q\left(z\middle|z^{(t)}\right)$

Accept with probability $A\left(z,z^{(t)}\right)=\min\left(1,\frac{\tilde{p}(z)}{\tilde{p}\left(z^{(t)}\right)}\right)$

If accept, emit $z^{(t+1)}=z$, otherwise, emit $z^{(t+1)}=z^{(t)}$.

}

If things get better, always accept. If they get worse, sometimes accept.



Review

Visualization of an ensemble of samplers

Evolution of ergodic chains

Review

Let $p^{(t)}(z)=\pi(z)-\Delta^{(t)}(z)$

$$p^{(t+1)}(z)=\sum_{z'}p^{(t)}(z')\,T(z,z')$$
$$=\sum_{z'}\pi(z')\,T(z,z')-\sum_{z'}\Delta^{(t)}(z')\,T(z,z')$$
$$=\pi(z)-\Delta^{(t+1)}(z)$$

Cannot die!

Dies out

Matrix-vector representation

Chains (think ensemble) evolve according to:

$$p(z) = \sum_{z'} p(z')\, T(z', z)$$

Matrix vector representation:

$$\mathbf{p} = \mathbf{T}\mathbf{p}'$$

And, after n iterations after a starting point:

$$\mathbf{p}^{(n)} = \mathbf{T}^N \mathbf{p}^{(0)}$$

Matrix representation

A single transition is given by

$$\mathbf{p} = \mathbf{T}\mathbf{p}'$$

Note what happens for stationary state:

$$\mathbf{p}^* = \mathbf{T}\mathbf{p}^*$$

So, $\mathbf{p}^*$ is an eigenvector with eigenvalue one.

And, intutively, if things converge, $\mathbf{p}^* = \mathbf{T}^\infty \mathbf{p}^{(0)}$

Aside on stochastic Matrices

- A right (row) stochastic matrix has non-negative entries, and its rows sum to one.
- A left (column) stochastic matrix has non-negative entries, and its columns sum to one.
- A doubly stochastic matrix has both properties.

Aside on stochastic Matrices

- T is a left (column) stochastic matrix.
  - If you are right handed, take the transpose
- The column vector, **p**, also has non-negative elements, that sum to one (sometimes this is called a stochastic vector).
- Fun facts that we did on the board
  - The product of a stochastic matrix and vector is a stochastic vector.
  - The product of two stochastic matrices is a stochastic matrix.

Aside on (stochastic) Matrix powers

Consider the eigenvalue decomposition of T, $T = E\Lambda E^{-1}$

$$\mathrm{T}^N = E\Lambda^N E^{-1}$$

Since T^N cannot grow without bound, the eigenvalues are inside $[-1,1]$.

In fact, for our situation, the second biggest absolute value of the eigenvalues is less than one (not so easy to prove), which also means the biggest one is 1.

Aside on (stochastic) Matrix powers

We have $\mathrm{T}^N = E\Lambda^N E^{-1}$

$$\Lambda = \begin{pmatrix} 1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_K \end{pmatrix} \quad \text{and} \quad \Lambda^\infty = \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \cdots & \\ & & & 0 \end{pmatrix}$$

$$\Lambda^\infty E^{-1} = \begin{pmatrix} \mathbf{e}_1^T \\ 0 \\ \ldots \\ 0 \end{pmatrix}$$

Aside on (stochastic) Matrix powers

Write $\mathbf{p}$ in terms of the eigen basis

$$\mathbf{p} = \sum_i a_i \mathbf{e}_i$$

$$\mathbf{e}_1^T \mathbf{p} = \sum_i a_i \mathbf{e}_1^T \mathbf{e}_i = a_1$$

and, $$\Lambda^\infty E^{-1}\mathbf{p} = \begin{pmatrix} \mathbf{e}_1^T \cdot \mathbf{p} \\ 0 \\ \ldots \\ 0 \end{pmatrix} = \begin{pmatrix} a_1 \\ 0 \\ \ldots \\ 0 \end{pmatrix}$$

Aside on (stochastic) Matrix powers

Recall that we are studying $\mathrm{E}\,\Lambda^\infty E^{-1}\mathbf{p}$

$$\Lambda^\infty E^{-1}\mathbf{p} = \begin{pmatrix} a_1 \\ 0 \\ \ldots \\ 0 \end{pmatrix}$$

So, $\mathrm{E}\,\Lambda^\infty E^{-1}\mathbf{p} = a_1\mathbf{e}_1$

Aside on (stochastic) Matrix powers

So, $E\Lambda^{\infty}E^{-1}\mathbf{p} = \mathbf{e}_1\underbrace{\left(\mathbf{e}_1^T\cdot\mathbf{p}\right)}_{a_1} \parallel \mathbf{e}_1 \parallel \mathbf{p}^*$

In summary, $\mathbf{p}^* \parallel \mathbf{e}_1$ together with $\mathbf{p}^*$ stochastic means that $E\Lambda^{\infty}E^{-1}\mathbf{p} = \mathbf{p}^*$

This is true, no mater what the initial point $\mathbf{p}$ is.

So, glossing over details, we have convergence to equilibrium.

Demo

- According to the previous, if T is a stochastic matrix, then:

$\mathbf{p}^* \cong \mathbf{T}^N\mathbf{p}$

(No matter what p! They all will give the same answer).

Also, $\mathbf{p}^* \parallel \mathbf{e}^{(1)}$

No demo, this was bonus homework.

Justification relies on Perron Frobenius theorem

Let $A = (a_{ij})$ be an $n \times n$ positive matrix: $a_{ij} > 0$ for $1 \le i, j \le n$. Then the following statements hold.

1. There is a positive real number r, called the **Perron root** or the **Perron–Frobenius eigenvalue**, such that r is an eigenvalue of A and any other eigenvalue λ (possibly, complex) is strictly smaller than r in absolute value, $|\lambda| < r$. Thus, the spectral radius $\rho(A)$ is equal to r.
2. The Perron–Frobenius eigenvalue is simple: r is a simple root of the characteristic polynomial of A. Consequently, the eigenspace associated to r is one-dimensional. (The same is true for the left eigenspace, i.e., the eigenspace for A^T.)
3. There exists an eigenvector $v = (v_1,\ldots,v_n)$ of A with eigenvalue r such that all components of v are positive: $Av = rv$, $v_i > 0$ for $1 \le i \le n$. (Respectively, there exists a positive left eigenvector w: $w^T A = r w^T$, $w_i > 0$.)
4. There are no other positive (moreover non-negative) eigenvectors except v (respectively, left eigenvectors except w), i.e. all other eigenvectors must have at least one negative or non-real component.
5. $\lim_{k\to\infty} A^k/r^k = vw^T$, where the left and right eigenvectors for A are normalized so that $w^Tv = 1$. Moreover, the matrix vw^T is the projection onto the eigenspace corresponding to r. This projection is called the **Perron projection**.
6. **Collatz–Wielandt formula**: for all non-negative non-zero vectors x, let $f(x)$ be the minimum value of $[Ax]_i / x_i$ taken over all those i such that $x_i \neq 0$. Then f is a real valued function whose maximum is the Perron–Frobenius eigenvalue.
7. A "Min-max" Collatz–Wielandt formula takes a form similar to the one above: for all strictly positive vectors x, let $g(x)$ be the maximum value of $[Ax]_i / x_i$ taken over i. Then g is a real valued function whose minimum is the Perron–Frobenius eigenvalue.
8. The Perron–Frobenius eigenvalue satisfies the inequalities

$$\min_i \sum_j a_{ij} \le r \le \max_i \sum_j a_{ij}.$$

From Wikipedia

Main points about P-F

- The maximal eigenvalue is strictly maximal (item 1).
- The corresponding eigenvector is “simple” (item 2)
- It has all positive (or negative) components (item 3).
- There is no other eigenvector that can be made non-negative.
- The maximal eigenvalue of a stochastic matrix has absolute value 1 (item 8 applied to stochastic matrix).

Aside on (stochastic) Matrix powers

Summary

$\mathbf{p}^* = \mathbf{T}\mathbf{p}^*$ is an eigenvector with eigenvalue one.

We have written it as $\mathbf{p}^* \| \mathbf{e}^1$ because $\mathbf{e}^1$ is the eigenvector normalized to norm 1 (standard form).

Intuitively (perhaps), T will reduce any component of p orthogonal to p^*, and T^N will kill off such components as $N \to \infty$.

Algebraic proof

Neal '93 provides an algebraic proof which does not rely on spectral theory.

(A question on the final studies this further for those that are interested).

Summary so far

- Under reasonable (easily checked and/or arranged) conditions, our chains converge to an equilibrium state.
- Easiest way to prove (or check) that this is the case is to show detailed balance.
- To use MCMC for sampling a distribution, we simply ensure that our target distribution is the equilibrium state.
- Variations on MCMC are mostly about improving the speed of convergence for particular situations.

Summary so far

- The time it takes to get reasonably close to equilibrium (where samples come from the target distribution) is called "burn in" time.
  - I.E., how long does it take to forget the starting state.
  - There is no general way to know when this has occurred.
- The average time it takes to visit a state is called "hit time".
- What if we really want independent samples?
  - We can take every N^{th} sample (some theories about how long to wait exist, but it depends on the algorithm and distribution)