Planning for the last mile

- A5 is due Friday (Monday is OK)
- Take home M2 to be posted early next week.
- Last day of classes is May 4.
  - Five left including this one
- Exam is May 6!
  - Do we want an in class exam?
  - Assignment instead?

Markov chain Monte Carlo methods

- Samples are conditioned on the previous one (this is the Markov chain).
- We have given up the very natural preference for independent samples.
- Basic intuition why this might be a good idea
  - If you have **finally** found a region of high probability, stick around for a bit, enjoy yourself, grab some more samples.
- MCMC is generally a good hammer for complex, high dimensional, problems.

Recall terminology and chain evolution

Denote an initial probability distribution by $p\left(z^{(1)}\right)$

Define transition probabilities by:

$$T\left(z^{(prev)}, z\right) = p\left(z \middle| z^{(prev)}\right)$$

$T = T_m(\ )$ can change over time, but for now, assume that it it is always the same (homogeneous chain)

Chains (think ensemble) evolve according to:

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

Stationary Markov chains

- Our goal is to have our Markov chain emit samples from our target distribution.
- This implies that the distribution being sampled at time *t+1* is the same as that of time *t* (stationary).
- If our stationary (target) distribution is *p()*, then if imagine an ensemble of chains, the are in each state with (long-run) probability *p()*.
  - On average, a switch from s1 to s2 happens as often as going from s2 to s, otherwise, the percentage of states would not be stable

Detailed balance

- Detailed balance is defined by:

$$p(z)T(z,z') = p(z')T(z',z)$$

(We assume that $T(\bullet)>0$)

- Detailed balance is a sufficient condition for a stationary distribution.

- Detailed balance is also referred to as reversibility.

Detailed balance implies stationary

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

If we have detailed balance, then

$$p(z)T(z,z') = p^{(prev)}(z')T(z',z)$$

So,

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

Hence, detailed balance implies the distribution is stationary.

Detailed balance (cont)

- Detailed balance (for p()) means that *if* our chain was generating samples from *p()*, it would continue to due so.
  - We will address how it gets there shortly

- Does the Metropolis algorithm have detailed balance?

Metropolis Example

While not_bored

{

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

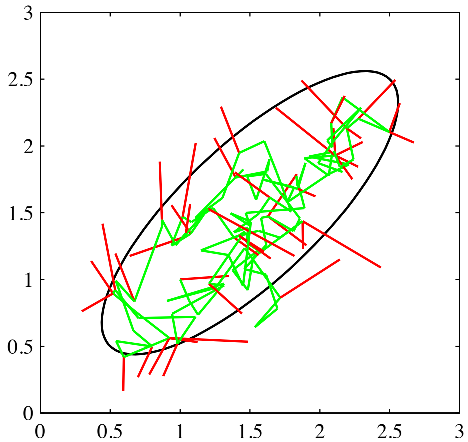
Accept with probability $A\left(z, z^{(prev)}\right) = \min\left(1, \dfrac{\tilde{p}(z)}{\tilde{p}\left(z^{(prev)}\right)}\right)$

If accept, emit z, otherwise, emit $z^{(prev)}$.

}

Same as $\dfrac{p(z)}{p\left(z^{(prev)}\right)}$

Metropolis Example



Green follows accepted proposals
Red are rejected moves.

Metropolis Example

Recall that in Metropolis, $A(z,z') = \min\left(1, \frac{p(z)}{p(z')}\right)$

$$\begin{aligned} p(z')q(z|z')A(z,z') &= q(z|z')\min(p(z'),p(z)) \\ &= q(z'|z)\min(p(z'),p(z)) \qquad (q() \text{ is symmetric}) \\ &= p(z)q(z'|z)\min\left(\frac{p(z')}{p(z)},1\right) \\ &= p(z)q(z'|z)\min\left(1,\frac{p(z')}{p(z)}\right) \\ &= p(z)q(z'|z)A(z',z) \end{aligned}$$

Ergodic chains

- Different starting probabilities will give different chains
- We want our chains to converge (in the limit) to the same stationary state, regardless of starting distribution.
- Such chains are called ergodic, and the common stationary state is called the equilibrium state.
- Ergodic chains have a unique equilibrium.

When do our chains converge?

- Important theorem tells us that our chains converge to equilibrium under two relatively weak conditions.
- (1) Irreducible
  - We can get from any state to any other state
- (2) Aperiodic
  - The chain does not get trapped in cycles
- These are true for detailed balance which is sufficient, but not necessary for convergence.

Intuition behind ergodic chains

Let $p^{(t)}(z)$ be the distribution at some time (e.g., initial distribution)

Let $p^{*}(z)$ be the stationary distribution

Let $p^{(t)}(z) = p^{*}(z) - q^{(t)}(z)$

Note that the elements of $p^{(t+1)}(z)$ and $p^{*}(z)$ sum to one, and thus the elements of $q(z)$ sum to zero.

Intuition behind ergodic chains

Let $p^{(t)}(z)$ be the distribution at some time (e.g., initial distribution)

Let $p^{*}(z)$ be the stationary distribution

Let $p^{(t)}(z) = p^{*}(z) - q^{(t)}(z)$

What is $p^{(t+1)}(z)$ in terms of $p^{*}(z)$?

Intuition behind ergodic chains

Let $p^{(t)}(z)$ be the distribution at some time (e.g., initial distribution)

Let $p^{*}(z)$ be the stationary distribution

Let $p^{(t)}(z) = p^{*}(z) - q^{(t)}(z)$

$$\begin{aligned}
p^{(t+1)}(z) &= \sum_{z'} p^{(t)}(z')\, T(z,z') \\
&= \sum_{z'} p^{*}(z')\, T(z,z') - \sum_{z'} q^{(t)}(z')\, T(z,z') \\
&= p^{*}(z) - q^{(t+1)}(z)
\end{aligned}$$

Intuition behind ergodic chains

$$\begin{aligned}
p^{(t+1)}(z) &= \sum_{z'} p^{(t)}(z')\, T(z,z') \\
&= \sum_{z'} p^{*}(z')\, T(z,z') - \sum_{z'} q^{(t)}(z')\, T(z,z') \\
&= p^{*}(z) - q^{(t+1)}(z)
\end{aligned}$$

Claim that $\left|q^{(t+1)}(z)\right| < \left|q^{(t)}(z)\right|$

Matrix representation

A single transition is given by

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

First note what happens for stationary state:

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

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

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).

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 should do 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, $T = E\Lambda E^{-1}$

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

Since T^N cannot grow without bound, the eigenvalues are inside $[-1,1]$. Also, "aperiodic" $\Leftrightarrow$ only one e.v. equal to 1.

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

Aside on (stochastic) Matrix powers

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

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

$$\Lambda^\infty E^{-1} = \begin{pmatrix} \mathbf{e}_1^T \\ 0 \\ \ldots \\ 0 \end{pmatrix} \quad \text{and} \quad E\Lambda^\infty E^{-1}\mathbf{p} \parallel \mathbf{p}^* \quad (\text{where } \mathbf{p}^* = \mathbf{e}_1)$$

Aside on (stochastic) Matrix powers

Summary

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

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