Sampling based inference

- Resources.
  - Bishop, chapter 11
  - Koller and Friedman, chapter 12
  - Andrieu et al. (linked to on lecture page).
- Koller and Friedman uses "particles" terminology instead of "samples".

Sampling based inference

- We have studied two themes in inference.
  - Marginalization / expectation / summing out or integration
  - Optimization
- Two flavors of activities
  - Fitting (inference using a model)
  - Learning (inference to find a model)
- These activities are basically the same in the generative modeling approach.

Motivation for sampling methods

- Real problems are typically complex and high dimensional.
- Example, images as evidence for stuff in the world

Motivation for sampling methods

- Real problems are typically complex and high dimensional.
- Suppose that we *could* generate samples from a distribution that is proportional to one we are interested in.

Typical case we are often interested in is $p(\theta|D)$

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

Consider $\tilde{p}(z) = p(\theta)p(D|\theta)$

Motivation for sampling methods

- Generally, θ lives in a very high dimensional space.
- Generally, regions of high $\tilde{p}(z)$ is very little of that space.
- IE, the probability mass is very localized.
- Watching samples from $\tilde{p}(z)$ should provide a good maximum (one of our inference problems)

Motivation for sampling methods (II)

- Now consider computing the expectation of a function $f(z)$ over $p(z)$.
- Recall that this looks like $E_{p(z)}[f] = \int_z f(z)p(z)dz$
- How can we approximate or estimate E?

Motivation for sampling methods (II)

- Now consider computing the expectation of a function $f(z)$ over $p(z)$.
- Recall that this looks like $E_{p(z)}[f] = \int_z f(z)p(z)dz$
- A bad plan for computing E:

  Discretize the space where z lives into L blocks

  Then compute $E_{p(z)}[f] \cong \frac{1}{L}\sum_{l=1}^{L} p(z)f(z)$

Motivation for sampling methods (II)

- Now consider computing the expectation of a function $f(z)$ over $p(z)$.
- Recall that this looks like $E_{p(z)}[f] = \int_z f(z)p(z)dz$
- A better plan, assuming we can sample $\tilde{p}(z)$

  Given independant samples $z^{(l)}$ from $\tilde{p}(z)$

  Estimate $E_{p(z)}[f] \cong \frac{1}{L}\sum_{l=1}^{L} f(z)$

Challenges for sampling

In real problems sampling $p(z)$ is very difficult.

We typically do not know the normalization constant, Z. (So we need to use $\tilde{p}(z)$).

Even if we can draw samples, it is hard to know if (when) they are good, and if we have enough of them.

Evaluating $\tilde{p}(z)$ is generally much easier (although, it can also be quite involved).

Sampling framework

We assume that sampling from $\tilde{p}(z)$ is hard, but that evaluating $\tilde{p}(z)$ is relatively easy.

We also assume that the dimension of z is high, and that $\tilde{p}(z)$ may not have closed from (but we can evaluate it).

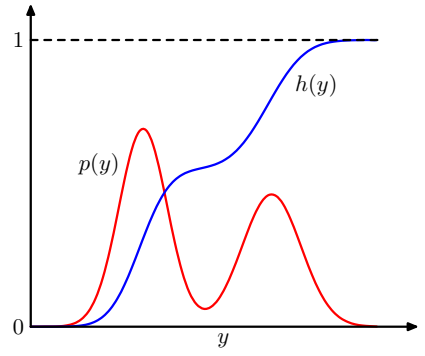
We will develop the material in the context of computing expections, but sampling also supports picking a good answer, such as a MAP estimate of parameters.

Basic Sampling (so far)

- Uniform sampling (everything builds on this)
- Sampling from a multinomial
- Sampling for selected other distributions (e.g., Gaussian)
  - At least, Matlab knows how to do it.
- Sampling univariate distributions using the inverse of the cumulative distribution (recall from HW 2).

Basic Sampling (so far)

- Sampling univariate distributions using the inverse of the cumulative distribution.



Basic Sampling (so far)

- Sampling directed graphical models using ancestral sampling.

Rejection Sampling

Assume that we have an easy to sample function, $q(z)$,
and a constant, k, where we know that $p(z) \leq k \bullet q(z)$.

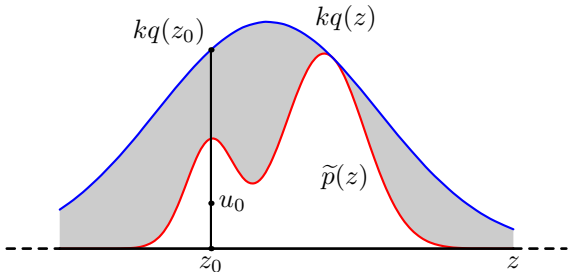
1) Sample $q(z)$

2) Keep samples in proportion to $\frac{p(z)}{k \bullet q(z)}$ and reject the rest.

Rejection Sampling

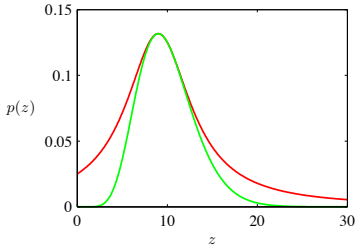
1) Sample $q(z)$

2) Keep samples in proportion to $\frac{p(z)}{k \bullet q(z)}$ and reject the rest.



Rejection Sampling

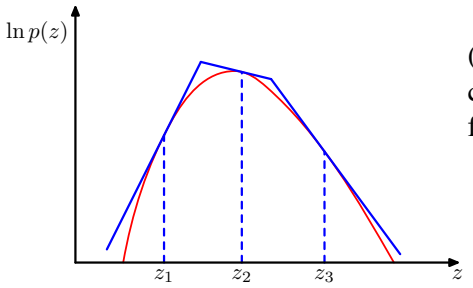
- Rejection sampling is hopeless in high dimensions, but is useful for sampling low dimensional "building block" functions.
- E.G., the Box-Muller method for generating samples from a Gaussian uses rejection sampling.



A second example where a gamma distribution is approximated by a Cauchy proposal distribution.

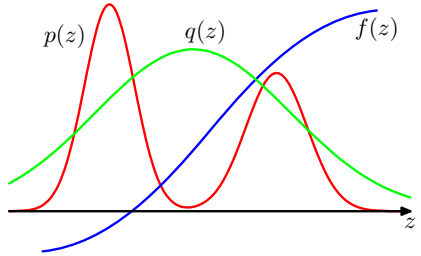
Rejection Sampling

- For complex functions, a good *q()* and *k* may not be available.
- One attempt to adaptively find a good *q()* (see Bishop 11.1.3)



(For log concave functions)

Importance Sampling



Rewrite $E_{p(z)}[f] = \int f(z)\,p(z)\,dz$

$$= \int f(z)\frac{p(z)}{q(z)}q(z)\,dz$$

$$\cong \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^{(l)})}{q(z^{(l)})}f(z^{(l)})$$ where samples come from $q(z)$

Importance Sampling (unnormalized)

$$p(z) = \frac{\tilde{p}(z)}{Z_p} \quad \text{and} \quad q(z) = \frac{\tilde{q}(z)}{Z_q}$$

$$E_{p(z)}[f] \cong \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^{(l)})}{q(z^{(l)})}f(z^{(l)})$$ (samples from $q(z^{(l)})$, equivalently, $\tilde{q}(z^{(l)})$)

$$\cong \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\frac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})}f(z^{(l)})$$

$$= \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l f(z^{(l)})$$ (introducing $\tilde{r}_l = \frac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})}$)

$$Z_p = \int \tilde{p}(z)\,dz$$

$$\frac{Z_p}{Z_q} = \int \frac{\tilde{p}(z)}{Z}\,dz = \int \frac{\tilde{p}(z)}{\tilde{q}(z)}q(z)\,dz$$ (because $Z_q = \frac{\tilde{q}(z)}{q(z)}$)

$$\cong \frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l$$ (samples coming from $\tilde{q}(z^{(l)})$)

Importance Sampling (unnormalized)

$$E_{p(z)}[f] \cong \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l f(z^{(l)})$$ (samples coming from $\tilde{q}(z^{(l)})$)

and $$\frac{Z_p}{Z_q} \cong \frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l q(z^{(l)})$$ (samples coming from $\tilde{q}(z^{(l)})$)

so $$E_{p(z)}[f] \cong \frac{\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l f(z^{(l)})}{\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l q(z^{(l)})}$$ (samples coming from $\tilde{q}(z^{(l)})$)

(from Koller and Friedman)

Importance sampling for graphical models

We know how to sample from directed graphical models where no variables are observed or conditioned on.

Suppose we want to use sampling to compute $p(Y = y)$.

$$p(Y = y) \cong \frac{1}{L}\sum_{l} I\left(y^{(l)}, y\right) \quad \text{(samples from } p(y)\text{)}$$

$$\text{where } I\left(y^{(l)}, y\right) = \begin{cases} 1 & \text{if } y^{(l)} = y \\ 0 & \text{otherwise} \end{cases}$$

(from Koller and Friedman)

Importance sampling for graphical models

We know how to sample from directed graphical models where no variables are observed or conditioned on.

What about the case of a particular value of a subset of the variables.

EG, we might want to sample: $p\left(Y \mid E = e\right)$

or, we might want to evaluate: $p\left(y = Y \mid E = e\right)$

(from Koller and Friedman)

Importance sampling for graphical models

EG, we might want to sample: $p\left(Y \mid E = e\right)$

or, we might want to evaluate: $p\left(y = Y \mid E = e\right)$

A fool-proof plan is to sample $p(y, e)$, and reject e ≠ E

(Potentially very expensive!)

(from Koller and Friedman)

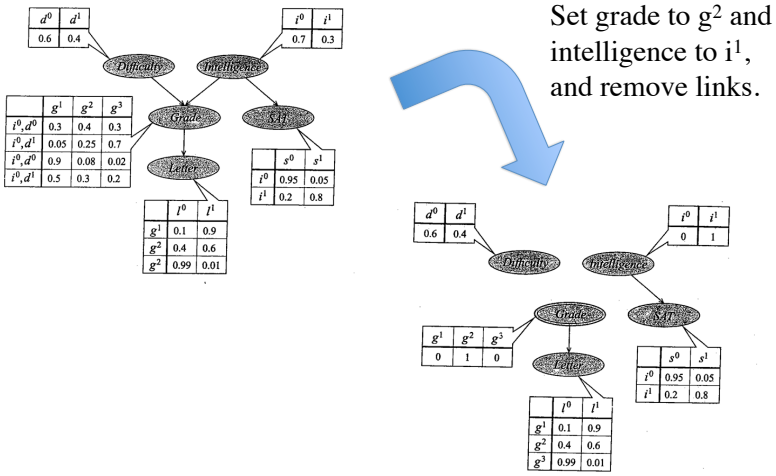
Importance sampling for graphical models

A natural idea is to use ancestral sampling on the graph, where we set E=e.

Kollar and Friedman develop this as sampling from the "mutilated" Bayesian network.

(from Koller and Friedman)

Mutilating graphical models



Set grade to g^2 and intelligence to i^1, and remove links.

(from Kollar and Friedman)

Importance sampling for graphical models

A natural idea is to use ancestral sampling on the graph, where we set E=e.

However, when E=e, this can influence the correct sampling of Y, and we have ignored this!

Instead, we use samples from the mutilated network for the proposal distribution in importance sampling .

(from Koller and Friedman)

Importance sampling for graphical models

$$\frac{p(y|e)}{q(y|e)} = \frac{P_{BN}(y|e)}{P_{MBN}(y|e)} = \frac{P_{BN}(y,e)}{P_{MBN}(y,e)}$$

$$p(y|e) \cong \frac{1}{L}\sum_{l} \frac{P_{BN}(y,e)}{P_{MBN}(y,e)} I(Y=y) \quad \text{(samples from } P_{MBN}(Y,e)\text{)}$$

Markov chain Monte Carlo methods

- The approximations of expectation so far have assumed that the samples are independent draws.
- This sounds good, but in high dimensions, we do not know how to get **good** independent samples from the distribution.
- MCMC methods drop this requirement.
- Basic intuition
  - If you have **finally** found a region of high probability, stick around for a bit, enjoy yourself, grab some more samples.

Markov chain Monte Carlo methods

- Samples are conditioned on the previous one (this is the Markov chain).
- MCMC is generally a good hammer for complex, high dimensional, problems.
- Main downside is that it is not "plug-and-play"
  - Doing well requires taking advantage to the structure of your problem
  - MCMC tends to be expensive (but take heart---there may not be any other solution, and at least your problem is being solved).

Metropolis Example

We want samples $z^{(1)}, z^{(2)}, \ldots$.

Again, write $p(z) = \tilde{p}(z)/Z$

Assume that $q\left(z \middle| z^{(prev)}\right)$ can be sampled easily

Also assume that $q(\ )$ is symmetric, i.e., $q\left(z_A \middle| z_B\right) = q\left(z_B \middle| z_A\right)$

For example, $q\left(z \middle| z^{(prev)}\right) \sim \mathbb{N}\left(z; z^{(prev)}, \sigma^2\right)$

Metropolis Example

While not_bored

{

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

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

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

}

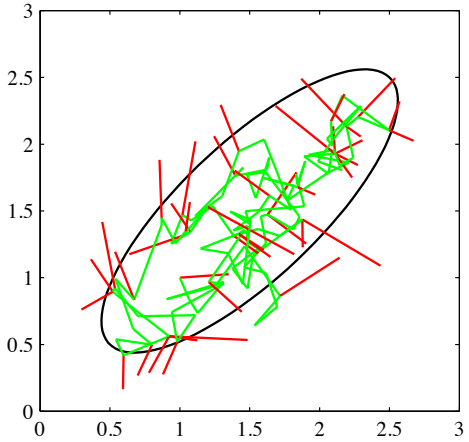
Metropolis Example

Note that

$$A\left(z, z^{(prev)}\right) = \min\left(1, \frac{\tilde{p}(z)}{\tilde{p}\left(z^{(prev)}\right)}\right) = \min\left(1, \frac{p(z)}{p\left(z^{(prev)}\right)}\right)$$

We do not need to normalize $p(z)$

Metropolis Example



Green follows accepted proposals
Red are rejected moves.

Markov chain view

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

Define transition probabilities by:

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

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

A given chain evolves from a sample of $p\left(z^{(1)}\right)$, and is an instance from an essemble of chains.

Stationary Markov chains

- Recall that 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, they 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
- If our stationary (target) distribution is *p()*, what do the transition probabilities look like?

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.

Review in 2011

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.

Review in 2011

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:

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

Review in 2011

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, they 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

Review in 2011

- 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

Review in 2011

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

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

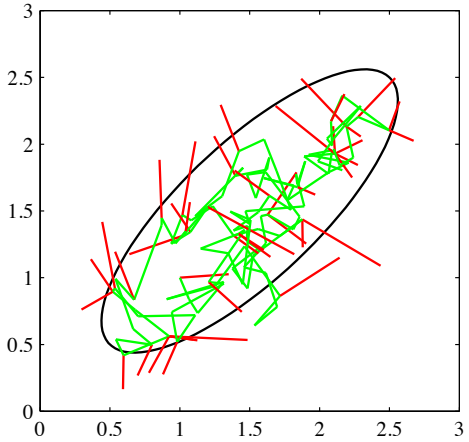
Accept with probability $A\left(z, z^{(prev)}\right) = \min\left(1, \frac{\tilde{p}(z)}{\tilde{p}\left(z^{(prev)}\right)}\right)$

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

}

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

Metropolis Example



Green follows accepted proposals
Red are rejected moves.

Metropolis Example

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

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

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 (for finite state spaces*) 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.

*Infinite or uncountable state spaces introduces additional complexities.

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

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

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

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-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 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 $[-1,1]$.

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

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

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

Aside on (stochastic) Matrix powers

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

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

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

Aside on (stochastic) Matrix powers

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

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

And, $E\Lambda^\infty E^{-1}\mathbf{p} = ?$

Aside on (stochastic) Matrix powers

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

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

In summary, $\mathbf{p}^* \parallel \mathbf{e}_1$ and $\mathbf{p}^*$ 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, $\mathbf{p}^* \| \mathbf{e}^{(1)}$

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: $A v = r v$, $v_i > 0$ 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 $w^T v = 1$. 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 $\mathbf{p}^* \| \mathbf{e}^{\mathbf{1}}$ because $\mathbf{e}^{\mathbf{1}}$ 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 might study this further).

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)

Metropolis-Hastings MCMC method

- Like Metropolis, but now q() is not symmetric.

Metropolis-Hastings MCMC method

While not_bored

{

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

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

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

}

Does Metropolis-Hastings have detailed balance?

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

Metropolis-Hastings comments

- Again it does not matter if we use unnormalized probabilities.
- It should be clear that the previous version, where q() is symmetric, is a special case.

Reversible Jump MH

- Suppose the dimension of your problem is not known (e.g., you want to estimate the number of clusters).
- Sampling now includes "jumping" changes probability space
- Requires a modification to Metropolis Hastings
  - Reversible jump MCMC, Green 95, 03
- RJMCMC is only about sampling. It does not tell you the number of dimensions
  - This must come from either the prior or the likelihood.

Gibbs sampling

- Gibbs sampling is another special case of MH.
- You might notice that the transition function, T(), varies (cycles) over time.
  - This is a relaxation of our assumption used to provide intuition about convergence
  - However, it still OK because the concatenation of the T() for a cycle converge

Consider a set of N variables, x_1, x_1, ...,x_N, Gibbs says

Initialize $\left\{z_i^{(0)} : i = 1,...,M\right\}$

While not_bored

{

For i=1 to M

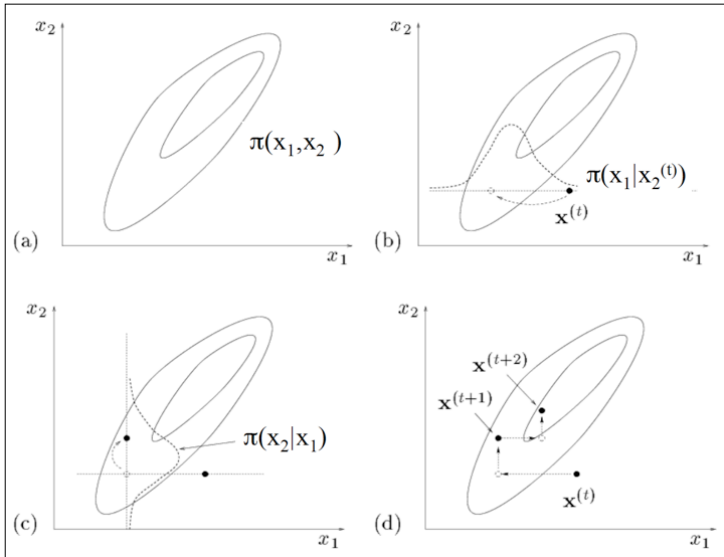
{

Sample $z_i^{(\tau+1)} \sim p\left(z_i \middle| z_1^{(\tau+1)},...,z_{i-1}^{(\tau+1)},z_{i+1}^{(\tau)},...,z_M^{(\tau)}\right)$

Always accept (emit $z = z_1^{(\tau+1)},...,z_{i-1}^{(\tau+1)},z_i^{(\tau+1)},z_{i+1}^{(\tau)},...,z_M^{(\tau)}$)

}

}

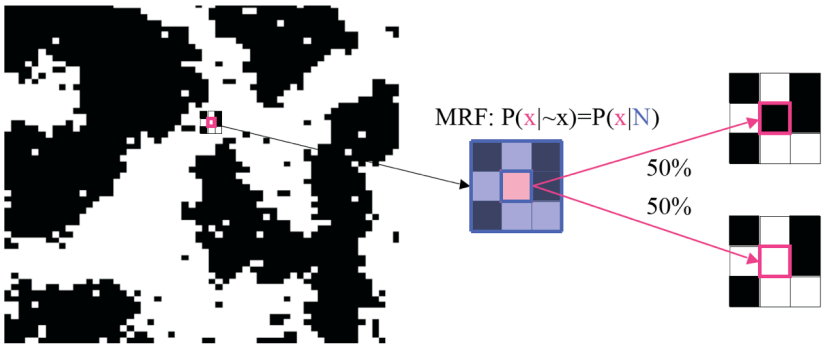


(From Dellaert and Zhu tutorial)

Examples of Gibbs

- If one can specify the conditional distributions so that they can be sampled, Gibbs is often a very good method.
- Typical examples include symmetric systems like the Markov random fields we had for images.
  - With a Markov property, the conditional probability can be quite simple.

Examples of Gibbs



(From Dellaert and Zhu tutorial)

Examples of Gibbs

Weak Affinity to Neighbors Strong Affinity to Neighbors

(From Dellaert and Zhu tutorial)

Gibbs as MH

$$q_i\left(\mathbf{z}\middle|\mathbf{z}^*\right)=p\left(z_i\middle|\mathbf{z}^*_{\backslash i}\right) \quad \text{and} \quad q_i\left(\mathbf{z}^*\middle|z\right)=p\left(z_i^*\middle|\mathbf{z}_{\backslash i}\right)$$

And we have $\mathbf{z}_{\backslash i}=\mathbf{z}^*_{\backslash i}$ because only i changes.

Gibbs as MH

$$\begin{aligned}
\mathrm{A}\left(\mathbf{z}^*,\mathbf{z}\right) &= \frac{p\left(\mathbf{z}^*\right)q_i\left(\mathbf{z}\middle|\mathbf{z}^*\right)}{p\left(\mathbf{z}\right)q_i\left(\mathbf{z}^*\middle|\mathbf{z}\right)} \\
&= \frac{p\left(\mathbf{z}^*_{\backslash i}\right)p\left(z_i^*\middle|\mathbf{z}^*_{\backslash i}\right)q_i\left(\mathbf{z}\middle|\mathbf{z}^*\right)}{p\left(\mathbf{z}_{\backslash i}\right)p\left(z_i\middle|\mathbf{z}_{\backslash i}\right)q_i\left(\mathbf{z}^*\middle|\mathbf{z}\right)} \\
&= \frac{p\left(\mathbf{z}^*_{\backslash i}\right)p\left(z_i^*\middle|\mathbf{z}^*_{\backslash i}\right)p\left(z_i\middle|\mathbf{z}^*_{\backslash i}\right)}{p\left(\mathbf{z}_{\backslash i}\right)p\left(z_i\middle|\mathbf{z}_{\backslash i}\right)p\left(z_i^*\middle|\mathbf{z}_{\backslash i}\right)} \\
&= 1
\end{aligned}$$

$$q_i\left(\mathbf{z}\middle|\mathbf{z}^*\right)=p\left(\mathbf{z}_i\middle|\mathbf{z}^*_{\backslash i}\right)$$

and $q_i\left(\mathbf{z}^*\middle|\mathbf{z}\right)=p\left(\mathbf{z}_i^*\middle|\mathbf{z}_{\backslash i}\right)$

and $\mathbf{z}_{\backslash i}=\mathbf{z}^*_{\backslash i}$

Exploring the space

- Algorithms like Metropolis-Hastings exhibit "random walk behavior" if the step size (proposal variance) is small
- If the step size is too big, then you get rejected too often
- Adaptive methods exist (see slice sampling in Bishop)
- Another approach is to combine samplers with different properties

Combined samplers

1. Initialise $x^{(0)}$.
2. For $i = 0$ to $N - 1$
   - Sample $u \sim \mathcal{U}_{[0,1]}$.
   - If $u < \nu$

     Apply the MH algorithm with a global proposal.
   - else

     Apply the MH algorithm with a random walk proposal.

Annealing

- Analogy with physical systems
- Relevant for optimization (not integration)
- Powers of probability distributions emphasize the peaks
- If we are looking for a maximum within a lot of distracting peaks, this can help.

Annealing

- Define a temperature T, and a cooling schedule (black magic part)
- Lower temperatures correspond to emphasized maximal peaks.
  - Hence we exponentiate by (1/T).

Annealing

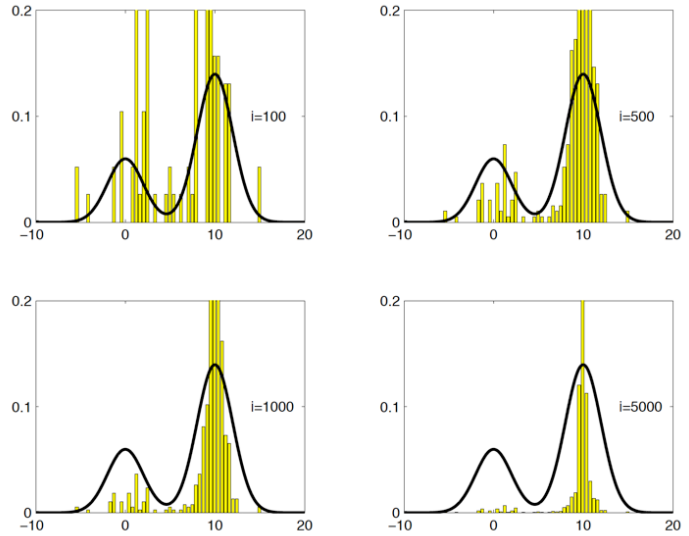
1. Initialise $x^{(0)}$ and set $T_0 = 1$.
2. For $i = 0$ to $N - 1$
   - Sample $u \sim \mathcal{U}_{[0,1]}$.
   - Sample $x^\star \sim q(x^\star | x^{(i)})$.
   - If $u < \mathcal{A}(x^{(i)}, x^\star) = \min\left\{1, \frac{p^{\frac{1}{T_i}}(x^\star) q(x^{(i)}|x^\star)}{p^{\frac{1}{T_i}}(x^{(i)}) q(x^\star|x^{(i)})}\right\}$

     $$x^{(i+1)} = x^\star$$

     else

     $$x^{(i+1)} = x^{(i)}$$
   - Set T_{i+1} according to a chosen cooling schedule.

(From Andrieu et al)



(From Andrieu et al)

Continuous versus discrete variables

- Derivatives of continuous distributions can tell you about the structure of your problem.
  - Opportunities for going much faster
- Naive approach is gradient ascent with added stochastic properties
  - Take a step, then perturb the result.
- Typical approach is to link the probability distribution to a potential energy function
  - Follow the system to find low energy (high probability)
  - Stochastic sampling via random momentum
  - An effective example method is Hybrid Monte Carlo

Hybrid Monte Carlo

- A more effective example method is Hybrid Monte Carlo
- Link the probability distribution to a potential energy function
  - Alternate stochastic sampling with "dynamics".
  - The dynamics follow the system to find low energy (high probability)
- HMC is an "auxiliary variable sampler"
  - Important trick
  - To sample p(z) we sample p(z,r) or p(z,r_1,r_2,...)
  - Ignore the auxiliary variables when we use the samples.

Hamiltonian Dynamics

$$p(\mathbf{z}) = \frac{1}{Z_p}\exp\left(-E(\mathbf{z})\right)$$

We equate $\mathbf{z}$ with position, so $E(\mathbf{z})$ is the potential energy.

High probability $\Leftrightarrow$ Low energy

$$E(\mathbf{z}) = -\log\left(Z_p\right) - \log\left(p(\mathbf{z})\right).$$

Hamiltonian Dynamics

Recall that the gradient, ∇, is the vector of partial derivatives.

Recall from physics that force is the negative gradient of energy

From before $E(\mathbf{z}) = -\log\left(Z_p\right) - \log\left(p(\mathbf{z})\right)$

So $\nabla E(\mathbf{z}) = \nabla\left(-\log\left(p(\mathbf{z})\right)\right)$

Or, in terms of log probabilities, we define

$\Delta(\mathbf{z}) = \nabla\left(\log\left(p(\mathbf{z})\right)\right) = -\nabla E(\mathbf{z})$ (This is the force)

Hamiltonian Dynamics

$$p(\mathbf{z}) = \frac{1}{Z_p}\exp\left(-E(\mathbf{z})\right) \quad \text{and} \quad \nabla E(\mathbf{z}) = \nabla\left(-\log\left(p(\mathbf{z})\right)\right)$$

Let $\mathbf{r}$ be the momentum vector for the system. Denote the kinetic energy by $K(\mathbf{r})$.

$$K(\mathbf{r}) = \frac{1}{2}\|\mathbf{r}\|^2 = \frac{1}{2}\sum_i r_i^2$$ (We assume that mass is one).

Hamiltonian Dynamics

$H(\mathbf{z},\mathbf{r}) = E(\mathbf{z}) + K(\mathbf{r})$ (conserved)

Our distribution with auxilary variables is

$$p(\mathbf{z},\mathbf{r}) = \frac{1}{Z}\exp\left(-H(\mathbf{z},\mathbf{r})\right)$$

Hamiltonian Dynamics

$H(\mathbf{z},\mathbf{r}) = E(\mathbf{z}) + K(\mathbf{r})$ (conserved)

We follow $\mathbf{z}$ according to H with a random $\mathbf{r}$

This can rapidly transport us towards (but not to) a local minimun thus avoiding random walk.

To follow H, we observe that $\mathbf{z}$ changes proportional to $\mathbf{r}$, and $\mathbf{r}$ changes proportion to force $(-\nabla E)$.

Again $-\nabla E = \nabla\left(\log\left(p(\mathbf{z})\right)\right) \equiv \Delta p(\mathbf{z})$

Following Dynamics

In HMC we follow the dynamics for L time steps of size τ (tunable parameters).

In the "leap frog" method for each τ.

1. Take 1/2 step in $\mathbf{r}$.
2. Take a full step in $\mathbf{z}$.
3. Take 1/2 step in $\mathbf{r}$.

Following Dynamics

For L leap frog steps we have.

1. Take 1/2 step in $\mathbf{r}$.
2. $(L-1)$ times take a full steps in $\mathbf{z}$, then $\mathbf{r}$.
3. Take a full step in $\mathbf{z}$.
4. Take 1/2 step in $\mathbf{r}$

Following Dynamics

To take a full step in $\mathbf{z}$.

$\mathbf{z}(\tau+1) = \mathbf{z}(\tau) + \varepsilon \cdot \Delta\left(\mathbf{r}(\tau)\right)$

(ε is the step size).

Following Dynamics

To take 1/2 step in $\mathbf{r}$.

$$\mathbf{r}\left(\tau+\frac{1}{2}\right)=\mathbf{r}(\tau)+\frac{1}{2}\varepsilon\cdot\Delta(\mathbf{z}(\tau))$$

Where $\Delta(\mathbf{z}(\tau))=\nabla\log\left(p(\mathbf{z}(\tau))\right)=-\nabla E(\mathbf{z})$ (force)

Following Dynamics

- After L steps of size t, we are at a new point with some bias of being at a lower potential energy (higher probability) and higher momentum.
- Momentum allows us to jump out of wells.

HMC dynamics step acceptance

- If our integration is perfect (i.e., in the limit as t --> 0) then energy is conserved.
  - Thus the value of distribution p($\mathbf{z}$,$\mathbf{r}$) is the same after the dynamics.
- If we assume no integration errors, we simply accept this step
- If we want to account for error accumulation, we accept the result according to:

$$\min\left(1,\frac{p(\mathbf{z}^*,\mathbf{r}^*)}{p(\mathbf{z},\mathbf{r})}\right)=\min\left(1,\exp\left(H(\mathbf{z},\mathbf{r})-H(\mathbf{z}^*,\mathbf{r}^*)\right)\right)$$

HMC stochastic step

- Typical instantiations sample the momentum variable
- Two common strategies
  - Sample the $\mathbf{r}$ independently from a Gaussian
  - Sample $\mathbf{r}$ from a Gaussian using Gibbs
- Note that in both of these cases the proposals are always accepted.

Putting it all together (A typical vision lab sampler)

- Discrete variables are sampled using (reversible jump) Metropolis Hastings.
- Continuous variables are sampled using stochastic dynamics (essentially hybrid Monte Carlo).
- Discrete variables typically control topology or components
  - The number of components and their type (block, cylinder)
  - How components are connected (branches from a stem)

A typical vision lab sampler

- Randomly proposing structure is too expensive because of the high rejection rate.
- Solution (part one) is to use data driven sampling
  - Proposals are conditioned on distributions computed before we begin using the data
  - For example, the probability of a corner being present in each point in the image.
- Solution (part two) is to delay acceptance
  - Adjust continuous parameters using stochastic dynamics so that the proposed structure is a good fit to the data.

A typical vision lab sampler

- We thus alternate between
  - (1) data driven proposals for new structure (or to switch or kill existing structure)
  - (2) exploring the continuous parameters of the structure
- Additional gains in optimization through having multiple samplers running in parallel exchange information