Markov chain Monte Carlo methods

- The approximations of expectation that we have looked at 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 often 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).
  - If there are faster solutions, you can incorporate that (and MCMC becomes a way to improve/select these good guesses).

Metropolis Example

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

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

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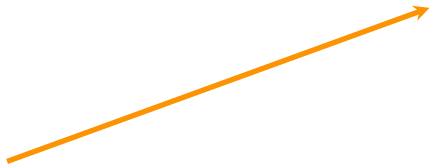
Sample $q\left(z \mid z^{(prev)}\right)$

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 $z^{(prev)}$.

}

Always emit one or the other (not needed if you are only searching for n good value).

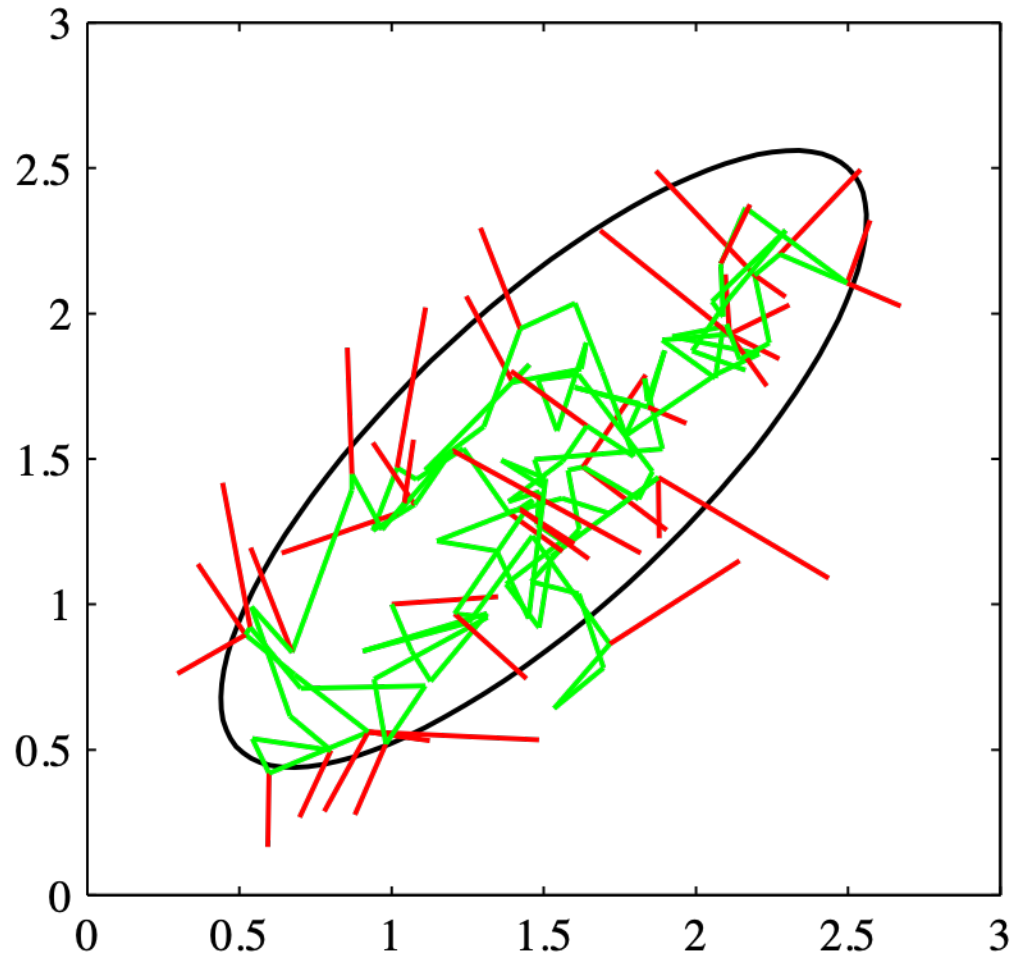If things get better, always accept. If they get worse, sometimes accept.

Metropolis Example

Note that

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

So 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 $p\left(z^{(1)}\right)$
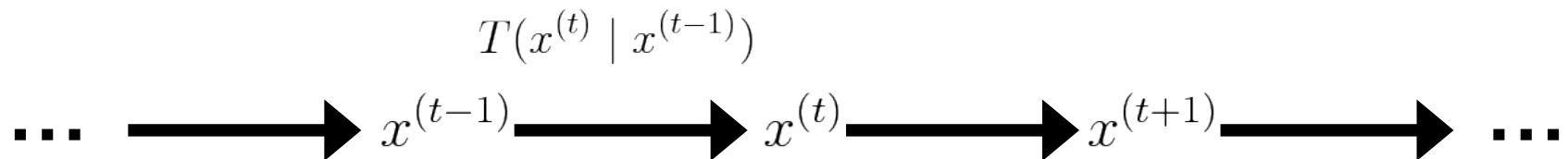
Define transition probabilities by:
$$T\left(z^{(prev)}, z\right) = p\left(z \mid z^{(prev)}\right) \qquad \text{(a probability distribution)}$$

T 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.

Markov Chain Monte Carlo (MCMC)

- Stochastic 1ˢᵗ order Markov process with transition kernel:

$$T(x^{(t)} \mid x^{(t-1)})$$

$$\cdots \longrightarrow x^{(t-1)} \longrightarrow x^{(t)} \longrightarrow x^{(t+1)} \longrightarrow \cdots$$

- Each $x^{(t)}$ full N-dimensional state vector
- MCMC samples $\ldots, x^{(t-1)}, x^{(t)}, x^{(t+1)}, \ldots$ **not independent**

Stationary Markov chains

- Recall that our goal is to have our Markov chain emit samples from our **target distribution** $p(z)$.

- This implies that the distribution being sampled at time $t+1$ would be the same as that of time t (**stationary**).

- If our stationary (target) distribution is $p()$, then if we 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 s1, otherwise, the percentage of states would not be stable.

  – Can we make this intuition more formal?

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 $p()$ to be a stationary distribution with respect to the positive T.

Detailed balance implies stationary

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

(because?)

Detailed balance implies stationary

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

$$= \sum_{z'} p^{(prev)}(z) T(z, z') \qquad \text{(because?)}$$

Detailed balance implies stationary

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

$$= \sum_{z'} p^{(prev)}(z) T(z, z') \qquad \text{(detailed balance)}$$

$$= p^{(prev)}(z) \sum_{z'} T(z, z') \qquad \textbf{(because?)}$$

Detailed balance implies stationary

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

$$= \sum_{z'} p^{(prev)}(z) T(z, z') \qquad \text{(detailed balance)}$$

$$= p^{(prev)}(z) \sum_{z'} T(z, z') \qquad \text{(moving constant out of sum)}$$

$$= p^{(prev)}(z) \sum_{z'} p(z' | z) \qquad \textbf{(because?)}$$

Detailed balance implies stationary

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

$$= \sum_{z'} p^{(prev)}(z) T(z, z') \qquad \text{(detailed balance)}$$

$$= p^{(prev)}(z) \sum_{z'} T(z, z') \qquad \text{(moving constant out of sum)}$$

$$= p^{(prev)}(z) \sum_{z'} p(z' | z) \qquad \text{(definition of T)}$$

$$= p^{(prev)}(z) \sum_{z'} \frac{p(z', z)}{p(z)} \qquad \textcolor{red}{\textbf{(because?)}}$$

Detailed balance implies stationary

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

$$= \sum_{z'} p^{(prev)}(z) T(z, z') \qquad \text{(detailed balance)}$$

$$= p^{(prev)}(z) \sum_{z'} T(z, z') \qquad \text{(moving constant out of sum)}$$

$$= p^{(prev)}(z) \sum_{z'} p(z' | z) \qquad \text{(definition of T)}$$

$$= p^{(prev)}(z) \sum_{z'} \frac{p(z', z)}{p(z)} \qquad \text{(definition of "|")}$$

$$= p^{(prev)}(z) \frac{p(z)}{p(z)} \qquad \textbf{(because?)}$$

Detailed balance implies stationary

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

$$= \sum_{z'} p^{(prev)}(z) T(z, z') \qquad \text{(detailed balance)}$$

$$= p^{(prev)}(z) \sum_{z'} T(z, z') \qquad \text{(moving constant out of sum)}$$

$$= p^{(prev)}(z) \sum_{z'} p(z' | z) \qquad \text{(definition of T)}$$

$$= p^{(prev)}(z) \sum_{z'} \frac{p(z', z)}{p(z)} \qquad \text{(definition of "|")}$$

$$= p^{(prev)}(z) \frac{p(z)}{p(z)} \qquad \text{(marginalization)}$$

$$= p^{(prev)}(z) \qquad \textbf{(because?)}$$

Detailed balance implies stationary

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

$$= \sum_{z'} p^{(prev)}(z) T(z, z') \qquad \text{(detailed balance)}$$

$$= p^{(prev)}(z) \sum_{z'} T(z, z') \qquad \text{(moving constant out of sum)}$$

$$= p^{(prev)}(z) \sum_{z'} p(z' | z) \qquad \text{(definition of T)}$$

$$= p^{(prev)}(z) \sum_{z'} \frac{p(z', z)}{p(z)} \qquad \text{(definition of "|")}$$

$$= p^{(prev)}(z) \frac{p(z)}{p(z)} \qquad \text{(marginalization)}$$

$$= p^{(prev)}(z) \qquad \text{(canceling)}$$

Detailed balance (continued)

- 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 soon.
  - For MCMC algorithms like Metropolis, it is important that the stationary state is the distribution **we want** (most Markov chains converge to *something*),

- Does the Metropolis algorithm have detailed balance?

Metropolis has detailed balance

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

For detailed balance, we need to show (in general)

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

Metropolis has detailed balance

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

For detailed balance, we need to show (in general)

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

In Metropolis this is

$$\textcolor{blue}{p(z')q(z|z')A(z,z')} = \textcolor{red}{p(z)q(z'|z)A(z',z)}$$

Probability of transition from *z* to *z'* is the probability that *z'* is proposed, **and** it is accepted.

Metropolis has detailed balance

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

$$p\left(z'\right)q\left(z\middle|z'\right)A\left(z,z'\right) = q\left(z\middle|z'\right)\min\left(p\left(z'\right), p\left(z\right)\right) \qquad \textbf{(because?)}$$

Metropolis has detailed balance

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

$p(z')q(z|z')A(z,z') = q(z|z')\min(p(z'),p(z))$ (bring $p(z')$ into A)

$= q(z'|z)\min(p(z'),p(z))$ **(because?)**

Metropolis has detailed balance

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

$$p\left(z'\right)q\left(z\mid z'\right)A\left(z,z'\right)=q\left(z\mid z'\right)\min\left(p\left(z'\right),p\left(z\right)\right) \qquad \text{(bring } p\left(z'\right) \text{ into } A\text{)}$$

$$=q\left(z'\mid z\right)\min\left(p\left(z'\right),p\left(z\right)\right) \qquad \text{q() is symmetric}$$

$$=p\left(z\right)q\left(z'\mid z\right)\min\left(\dfrac{p\left(z'\right)}{p\left(z\right)},1\right) \qquad \textbf{(because?)}$$

Metropolis has detailed balance

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

$$p(z')q(z|z')A(z,z') = q(z|z')\min(p(z'),p(z)) \qquad \text{(bring } p(z') \text{ into } A\text{)}$$

$$= q(z'|z)\min(p(z'),p(z)) \qquad \text{q() is symmetric}$$

$$= p(z)q(z'|z)\min\left(\frac{p(z')}{p(z)},1\right) \qquad \text{(divide the min()}$$
$$\text{by } p(z))$$

$$= p(z)q(z'|z)\min\left(1,\frac{p(z')}{p(z)}\right) \qquad \text{(switch order}$$
$$\text{in min())}$$

$$= p(z)q(z'|z)A(z',z) \qquad \textbf{(because?)}$$

Metropolis has detailed balance

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

$$p(z')q(z|z')A(z,z') = q(z|z')\min(p(z'), p(z)) \qquad \text{(bring } p(z') \text{ into } A)$$

$$= q(z'|z)\min(p(z'), p(z)) \qquad \text{q() is symmetric}$$

$$= p(z)q(z'|z)\min\left(\dfrac{p(z')}{p(z)}, 1\right) \qquad \text{(divide the min()}$$
$$\text{by } p(z))$$

$$= p(z)q(z'|z)\min\left(1, \dfrac{p(z')}{p(z)}\right) \qquad \text{(switch order}$$
$$\text{in min())}$$

$$= p(z)q(z'|z)A(z',z) \qquad \text{(definition of } A(z',z))$$

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 (there exists a stationary state) with T>0 (you can get there).
    - This is sufficient, but not necessary for convergence.

*Infinite or uncountable state spaces introduces additional complexities, but the main thrust is similar.

Evolution of ergodic chains

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

Let $\pi(z)$ be the stationary distribution

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

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

Evolution of ergodic chains

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

Let $\pi(z)$ be the stationary distribution

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

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

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

$$= \pi(z) - \Delta^{(t+1)}(z)$$

Evolution of ergodic chains

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

Evolution of ergodic chains

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

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

$$= \pi(z) - \Delta^{(t+1)}(z)$$

Claim that $\left| \Delta^{(t)}(z) \right| < (1-v)^t$

where $\quad v = \min_{z} \, \min_{z':\pi(z')>0} \dfrac{T(z,z')}{\pi(z)}$

and we have $\quad 0 < v \le 1$

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:

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

What does this equation look like?

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, $\mathbf{p}^* = \mathbf{T}^\infty \mathbf{p}^{(0)}$ For any $\mathbf{p}^{(0)}$!

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

- In our problem, T is a left (column) stochastic matrix.
    - If you want to be right handed, take the transpose

- The column vector, **p**, also has non-negative elements, that sum to one (stochastic vector).

Aside on stochastic matrices

- In our problem, T is a left (column) stochastic matrix.
  - If you want to be right handed, take the transpose

- The column vector, **p**, also has non-negative elements, that sum to one (stochastic vector).

- Fun facts
  - 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, $T = E\Lambda E^{-1}$

$$T^N = ?$$

Aside on (stochastic) matrix powers

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

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

Aside on (stochastic) matrix powers

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

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

T^N cannot grow without bound,   Why not?

Aside on (stochastic) matrix powers

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

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

T^N cannot grow without bound,

because it is a stochastic matrix.

Aside on (stochastic) matrix powers

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

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

T^N cannot grow without bound,

because it is a stochastic matrix.

Logic:
- Product of stochastic matrix is a stochastic matrix
- Colums of (left) stochastic matrix sum to 1
- Power is a bunch of products

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 eigenvalue magnitudes (remember they can be complex) are inside [0,1].

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 eigenvalue magnitudes (remember they can be complex) are inside [0,1].

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 (otherwise T will go to zero).

Aside on (stochastic) matrix powers

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

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

Aside on (stochastic) matrix powers

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

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

Aside on (stochastic) matrix powers

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

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

$$\Lambda^\infty E^{-1} = \begin{pmatrix} & \\ & ? \\ & \end{pmatrix}$$

Aside on (stochastic) matrix powers

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

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

$$\Lambda^\infty E^{-1} = \begin{pmatrix} E^{-1}(1,:) \\ \mathbf{0}^T \\ ... \\ \mathbf{0}^T \end{pmatrix}$$

Aside on (stochastic) matrix powers

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

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

$$E^{-1}(1,:) \bullet \mathbf{p} = \sum_i a_i E^{-1}(1,:) \bullet \mathbf{e}_i = a_1$$

Aside on (stochastic) matrix powers

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

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

$$E^{-1}\left(1,:\right) \bullet \mathbf{p} = \sum_i a_i E^{-1}\left(1,:\right) \bullet \mathbf{e}_i = a_1$$

$$\left(\begin{array}{l} E^{-1} \bullet E = I \\ \text{And the columns of E are } \mathbf{e}_i \\ \text{So, } E^{-1}(1,:) \bullet E = \left(1,0,0,\cdots.0\right) \\ \text{(first row of the inverse), and} \\ \text{so } E^{-1}(1,:) \bullet \mathbf{e}_1 = 1 \\ \text{and } E^{-1}(1,:) \bullet \mathbf{e}_{i \neq 1} = 0 \end{array}\right)$$

Aside on (stochastic) matrix powers

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

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

$$E^{-1}\left(1,:\right) \bullet \mathbf{p} = \sum_i a_i E^{-1}\left(1,:\right) \bullet \mathbf{e}_i = a_1$$

and, $\quad \Lambda^\infty E^{-1} \mathbf{p} = \begin{pmatrix} E^{-1}\left(1,:\right) \bullet \mathbf{p} \\ 0 \\ \ldots \\ 0 \end{pmatrix} = \begin{pmatrix} a_1 \\ 0 \\ \ldots \\ 0 \end{pmatrix}$

Aside on (stochastic) matrix powers

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

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

So, $\qquad\qquad \mathrm{E}\,\Lambda^\infty E^{-1}\mathbf{p} = \ ?$

Aside on (stochastic) matrix powers

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

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

So, $\qquad \mathrm{E} \, \Lambda^{\infty} E^{-1} \mathbf{p} = a_1 \mathbf{e_1}$

Aside on (stochastic) matrix powers

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

$\mathrm{E}\Lambda^{\infty}E^{-1}\mathbf{p} = \mathbf{p}^*$ no mater what the initial point $\mathbf{p}$ is.

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

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. $\lim\limits_{k \to \infty} A^k / r^k = vw^T$, where the left and right eigenvectors for A are normalized so that $w^T v = 1$. Moreover, the matrix $v w^T$ 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 $x_i \ne 0$. 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 for positive square matrices

- The maximal eigenvalue is strictly maximal and real valued (item 1).

- Its eigenvector (as computed by software*) has all positive (or negative) real components (item 3).

- The maximal eigenvalue of a stochastic matrix has absolute value 1 (item 8 applied to stochastic matrix).

*P-F says that the positive version exists, but software might hand you the negative of that, but you can negate it to be consistent with P-F.

Summary on matrix version of stationarity

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

We have written it as $\mathbf{p}^* \parallel \mathbf{e}^1$ because $\mathbf{e}^1$ is the eigenvector normalized to norm 1 (not stochastic).

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

Algebraic proof

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

MCMC so far

- Under reasonable (easily checked and/or arranged) conditions, ensembles of chains over discretized states converge to an equilibrium state.

- Easiest way to prove (or check) that this is the case is to show detailed balance and use T>0.

- Nice analogy with powers of stochastic matrices, which converge to an operator based on the largest magnitude eigenvector (not covered in F18)

- In theory, to use MCMC for sampling a distribution, we simply need to ensure that our target distribution is the equilibrium state.

- In practice we do not know even know if we have visited the best place yet. (The ensemble metaphor runs into trouble if you have a small number of chains compared to the number of states).

MCMC in theory

- 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?

  - In theory we can take every N^{th} sample (some theories about how long to wait exist, but it depends on the algorithm and distribution).

MCMC for ML in practice

- We use MCMC for machine learning problems with very complex distributions over high dimensional spaces.

- Variables can be either discrete or continuous (often both)

- Despite the gloomy worst case scenario, MCMC is often a good way to find good solutions (either by MAP or integration).
  - Key reason is that there is generally structure in our distributions.
  - We need to exploit this knowledge in our proposal distributions.
  - Instead of getting hung up about whether you actually have convergence
    - Enjoy that fact that what you are doing is principled and can improve any answer (with respect to your model) that you can get by other means
      - Your model should be able to tell you which proposed solution are good.

Metropolis-Hastings MCMC method

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

- Metropolis is a special case of MH.

Metropolis-Hastings MCMC method

While not_bored

{

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

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

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

}

Does Metropolis-Hastings converge to the target distribution?

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

- If Metropolis-Hastings has detailed balance, then it converges to the target distribution under weak conditions.

  – The converse is not true, but generally samplers of interest will have detailed balance

Does Metropolis-Hastings have detailed balance?

To show detailed balance we need to show

$$p(z')q(z|z')A(z,z') = p(z)q(z'|z)A(z',z)$$

$$p(z')q(z|z')A(z,z') = \min\left(p(z')q(z|z'), p(z)q(z'|z)\right)$$

Does Metropolis-Hastings have detailed balance?

To show detailed balance we need to show

$$p(z')q(z|z')A(z,z') = p(z)q(z'|z)A(z',z)$$

$$p(z')q(z|z')A(z,z') = \min\left(p(z')q(z|z'), p(z)q(z'|z)\right)$$

$$\left\{ \text{Recall that} \quad A(z,z') = \min\left(1, \frac{p(z)q(z'|z)}{p(z')q(z|z')}\right) \right\}$$

Does Metropolis-Hastings have detailed balance?

To show detailed balance we need to show

$$p(z')q(z|z')A(z,z') = p(z)q(z'|z)A(z',z)$$

$$p(z')q(z|z')A(z,z') = \min\left(p(z')q(z|z'),\ p(z)q(z'|z)\right)$$

$$= p(z)q(z'|z)\min\left(\frac{p(z')}{p(z)}\frac{q(z|z')}{q(z'|z)}, 1\right)$$

Does Metropolis-Hastings have detailed balance?

To show detailed balance we need to show

$$p(z')q(z|z')A(z,z') = p(z)q(z'|z)A(z',z)$$

$$p(z')q(z|z')A(z,z') = \min\big(p(z')q(z|z'),\, p(z)q(z'|z)\big)$$

$$= p(z)q(z'|z)\min\left(\frac{p(z')}{p(z)}\frac{q(z|z')}{q(z'|z)},1\right)$$

$$= p(z)q(z'|z)\min\left(1,\frac{p(z')}{p(z)}\frac{q(z|z')}{q(z'|z)}\right)$$

Does Metropolis-Hastings have detailed balance?

To show detailed balance we need to show

$$p(z')q(z|z')A(z,z') = p(z)q(z'|z)A(z',z)$$

$$p(z')q(z|z')A(z,z') = \min\left(p(z')q(z|z'),\, p(z)q(z'|z)\right)$$

$$= p(z)q(z'|z)\min\left(\frac{p(z')}{p(z)}\frac{q(z|z')}{q(z'|z)},1\right)$$

$$= p(z)q(z'|z)\min\left(1,\frac{p(z')}{p(z)}\frac{q(z|z')}{q(z'|z)}\right)$$

$$= p(z)q(z'|z)A(z',z)$$

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.

- q() can be anything, but you need to specify the reverse move (often tricky)
    - If you are using MH for optimization (not integration), then getting this only approximately right might be OK.

Metropolis-Hastings

Transition kernel with target distribution:

$$p(x) = 1/Z \, f(x)$$

1. Sample proposal: $x' \mid x^{(t-1)} \sim q(\cdot)$
2. Accept with probability:

$$\min\{1, a\} \quad \text{where} \quad a = \frac{f(x')}{f(x^{(t-1)})} \frac{q(x^{(t-1)} \mid x')}{q(x' \mid x^{(t-1)})}$$

Example Gaussian proposal: $q(x^{(t)} \mid x^{(t-1)}) = \mathcal{N}(x^{(t-1)}, \epsilon^2)$

- Acceptance ratio simplifies to: $a = f(x')/f(x^{(t-1)})$
- True for any symmetric proposal: $q(x^{(t)} \mid x^{(t-1)}) = q(x^{(t-1)} \mid x^{(t)})$
- Known as Metropolis algorithm in this case
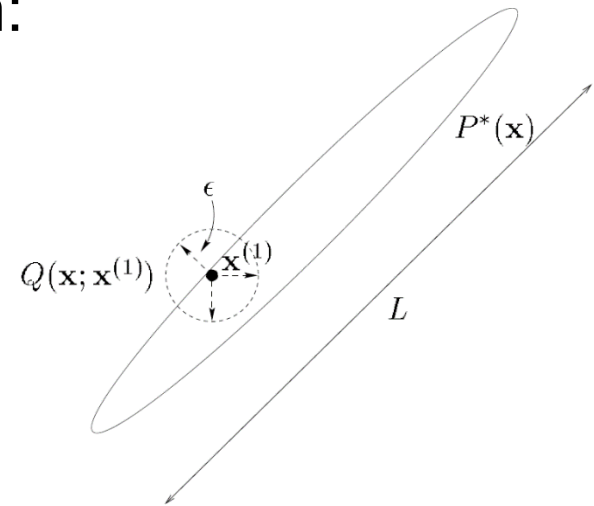
Independent Samples

Q How many M-H samples are required for an independent sample?

A Consider Gaussian proposal:

$$q(x^{(t)} \mid x^{(t-1)}) = \mathcal{N}(x^{(t-1)}, \epsilon^2)$$

- Typically $\epsilon \ll L$ for adequate acceptance rate
- Leads to random walk dynamics, which can be slow to converge
- <u>Rule of Thumb</u>: If average acceptance is $f \in (0, 1)$ need to run for roughly $T \approx (L/\epsilon)^2/f$ iterations for an independent sample

This is only a lower bound (and potentially very loose)
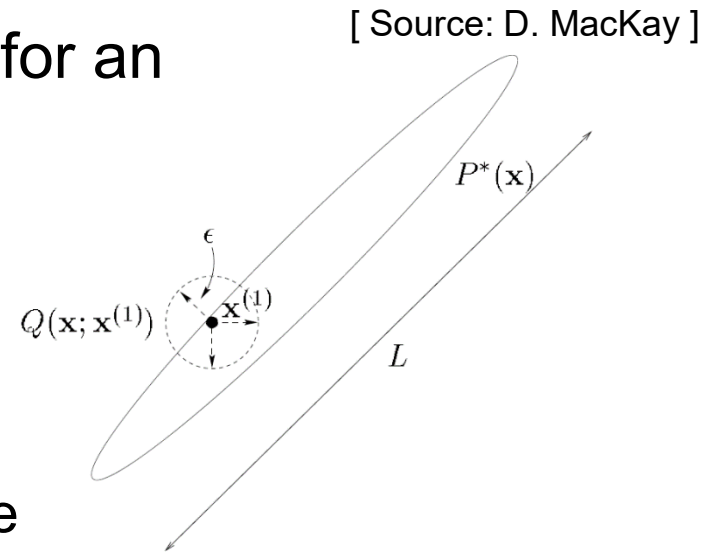
Example: Independent Samples

Metropolis   **Independent**   [Source: D. MacKay]
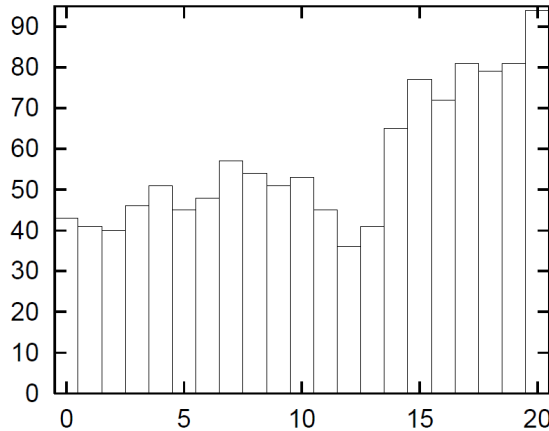
1200 iterations   1200 iterations
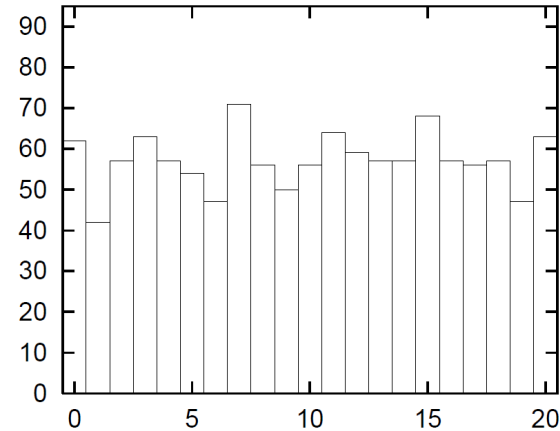
Target: $p(x) = \begin{cases} \frac{1}{21} & x \in \{0, \ldots, 20\} \\ 0 & \text{otherwise} \end{cases}$

Proposal: $q(x' \mid x) = \begin{cases} \frac{1}{2} & x' = x \pm 1 \\ 0 & \text{otherwise} \end{cases}$

From $x_0 = 10$ need ~400 steps to reach both end states (0 and 20). So, ~400 steps to generate 1 independent sample!

Very important to avoid random walk dynamics

Administrivia

- Homework 5 out, due Monday, Dec. 7 (2 weeks)
  - Particle filtering
  - Gibbs sampling

- We **do** have class this Wednesday

Gibbs Sampling

Suppose target distribution is:

$$p(x) = \prod_{s \in \mathcal{V}} p(x_s \mid \mathrm{Pa}(s))$$

where Pa(s) are parents of node *s.*

Metropolis-Hastings Proposal:

For system with K variables,

$$x_1^{(t+1)} \sim P(x_1 | x_2^{(t)}, x_3^{(t)}, \ldots x_K^{(t)})$$
$$x_2^{(t+1)} \sim P(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots x_K^{(t)})$$
$$x_3^{(t+1)} \sim P(x_3 | x_1^{(t+1)}, x_2^{(t+1)}, \ldots x_K^{(t)}), \mathrm{etc.}$$
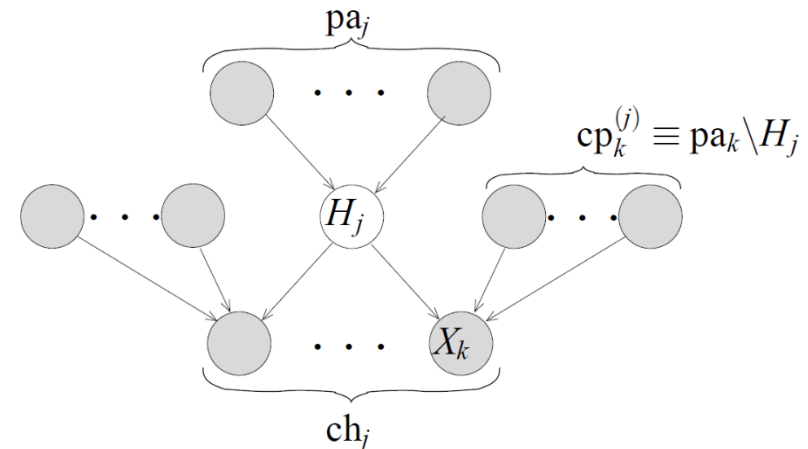
By conditional independence, Gibbs samples drawn from Markov blanket

Recall for undirected MRFs the Markov Blanket are immediate neighbors

Gibbs sampling

- Gibbs sampling is special case of MH.

- The proposal distribution will be cycle over $p\left(z_n \middle| \left\{z_{i \neq n}\right\}\right)$

- We will always accept the proposal.

- 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, z_1, z_1, ..., z_N. Then Gibbs says

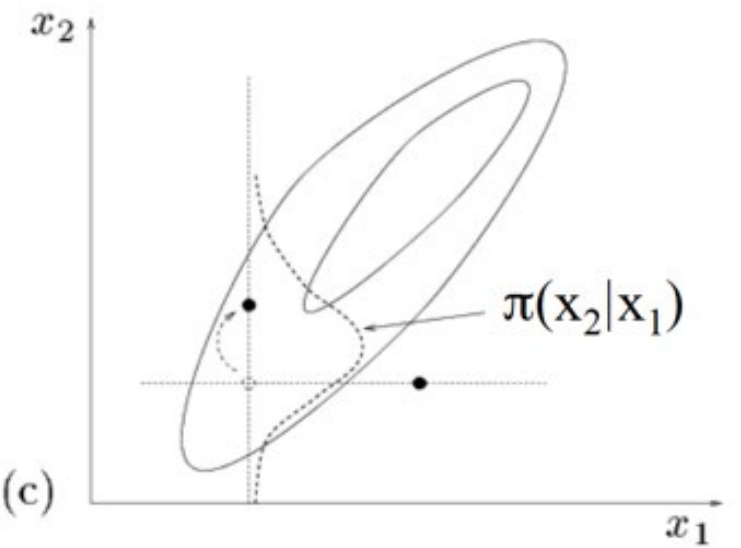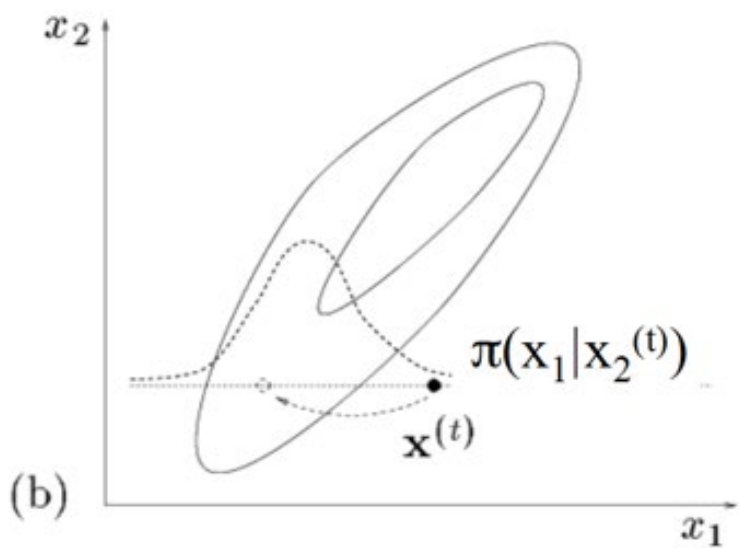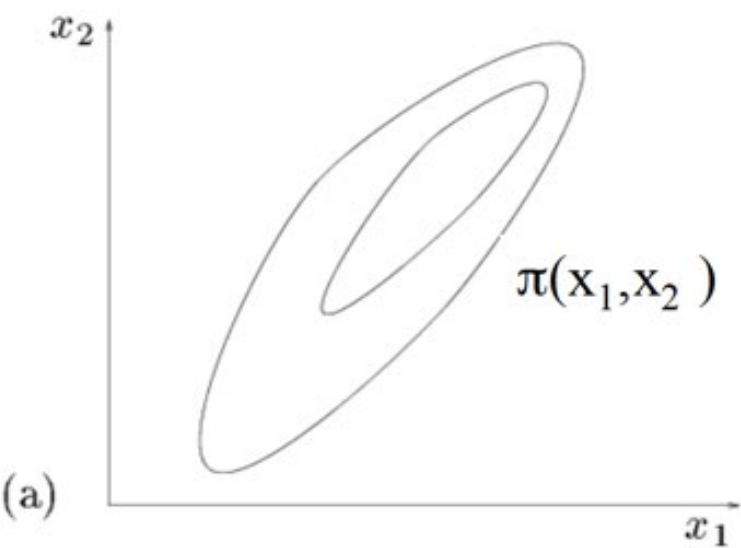Initialize $\left\{ z_i^{(0)} : i = 1,...,N \right\}$

While not_bored

{

    For i=1 to N

    {

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

    }

}

(a) $\pi(x_1, x_2)$

(Source: D. MacKay)

(a) $\pi(x_1, x_2)$

(b) $\pi(x_1 | x_2^{(t)})$ $\mathbf{x}^{(t)}$

(Source: D. MacKay)

(a) $\pi(x_1, x_2)$

(b) $\pi(x_1 | x_2^{(t)})$ $\mathbf{x}^{(t)}$

(c) $\pi(x_2 | x_1)$

(Source: D. MacKay)

(a) $\pi(x_1, x_2)$

(b) $\pi(x_1 | x_2^{(t)})$, $\mathbf{x}^{(t)}$

(c) $\pi(x_2 | x_1)$

(d) $\mathbf{x}^{(t+2)}$, $\mathbf{x}^{(t+1)}$, $\mathbf{x}^{(t)}$

(Source: D. MacKay)
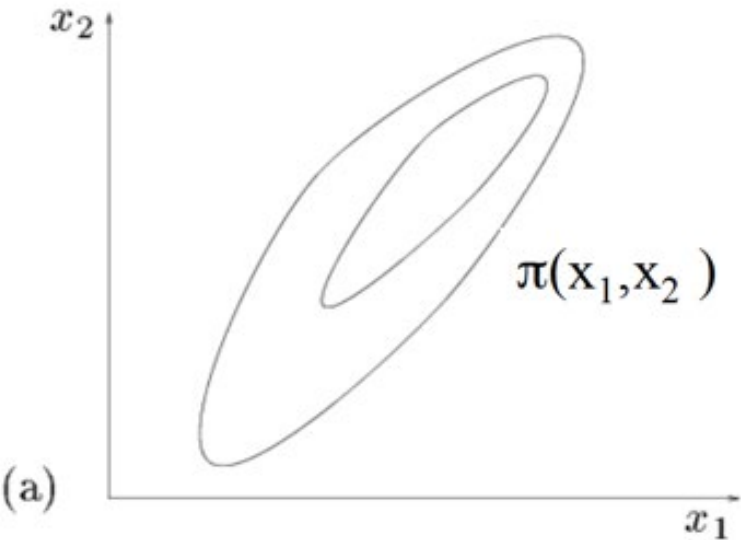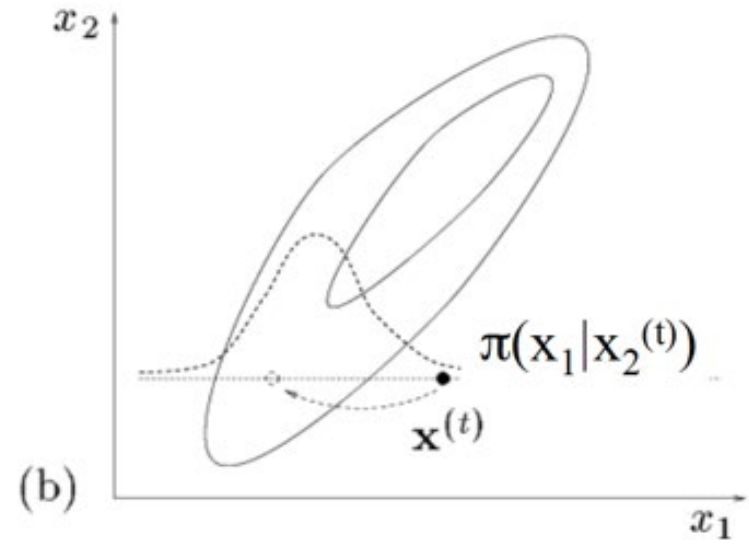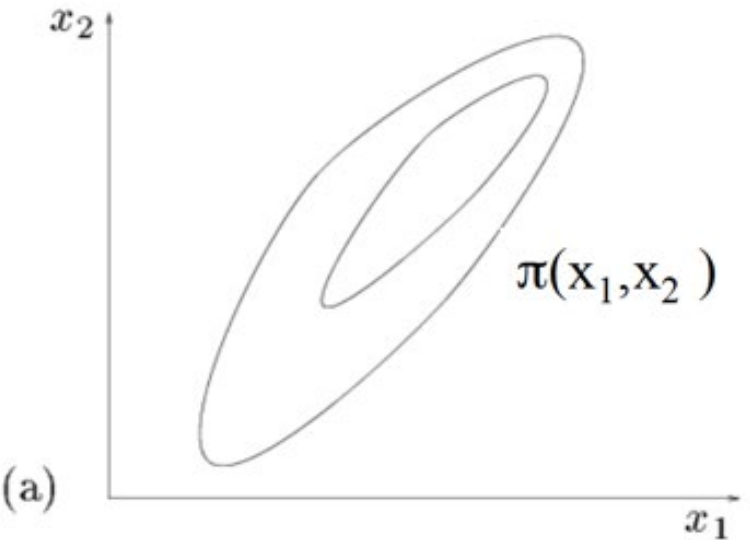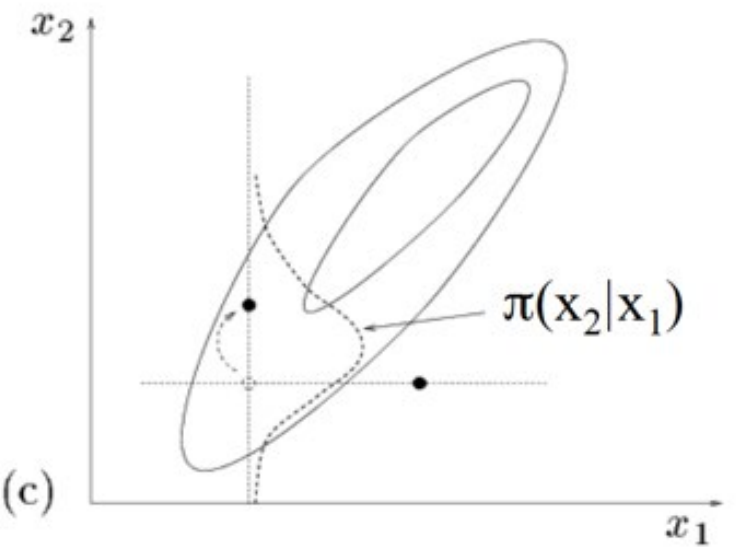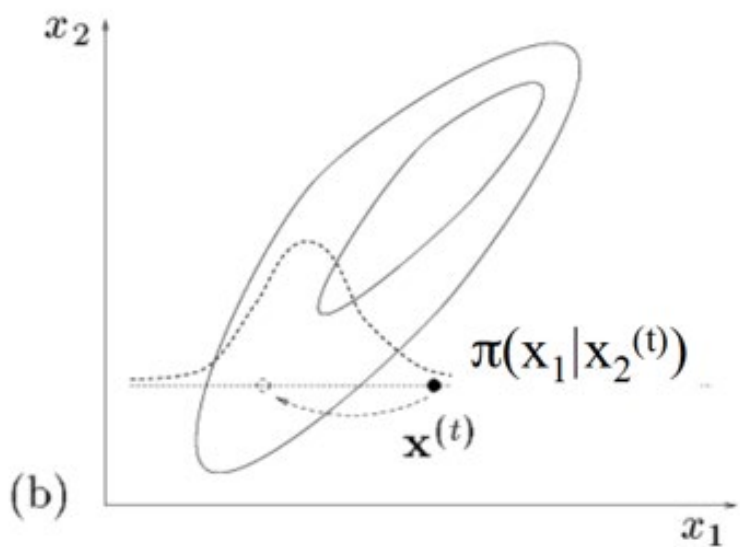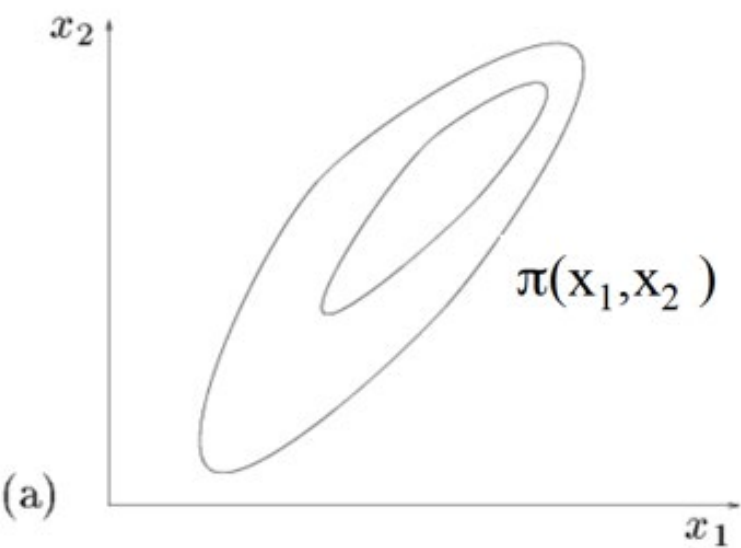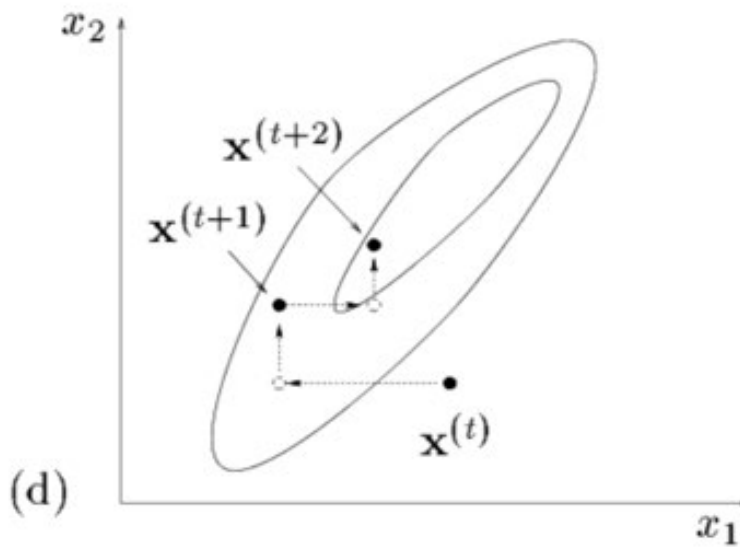
Examples of Gibbs

- If one can specify the conditional distributions in a way that they can be sampled, Gibbs can be a very good method.

- Typical examples include symmetric systems like the Markov random field grids we had for images.
  - With a Markov property, the conditional probability can be quite simple.

Examples of Gibbs

MRF: $P(x|\sim x) = P(x|N)$

50%

50%

Examples of Gibbs

Weak Affinity to Neighbors   Strong Affinity to Neighbors
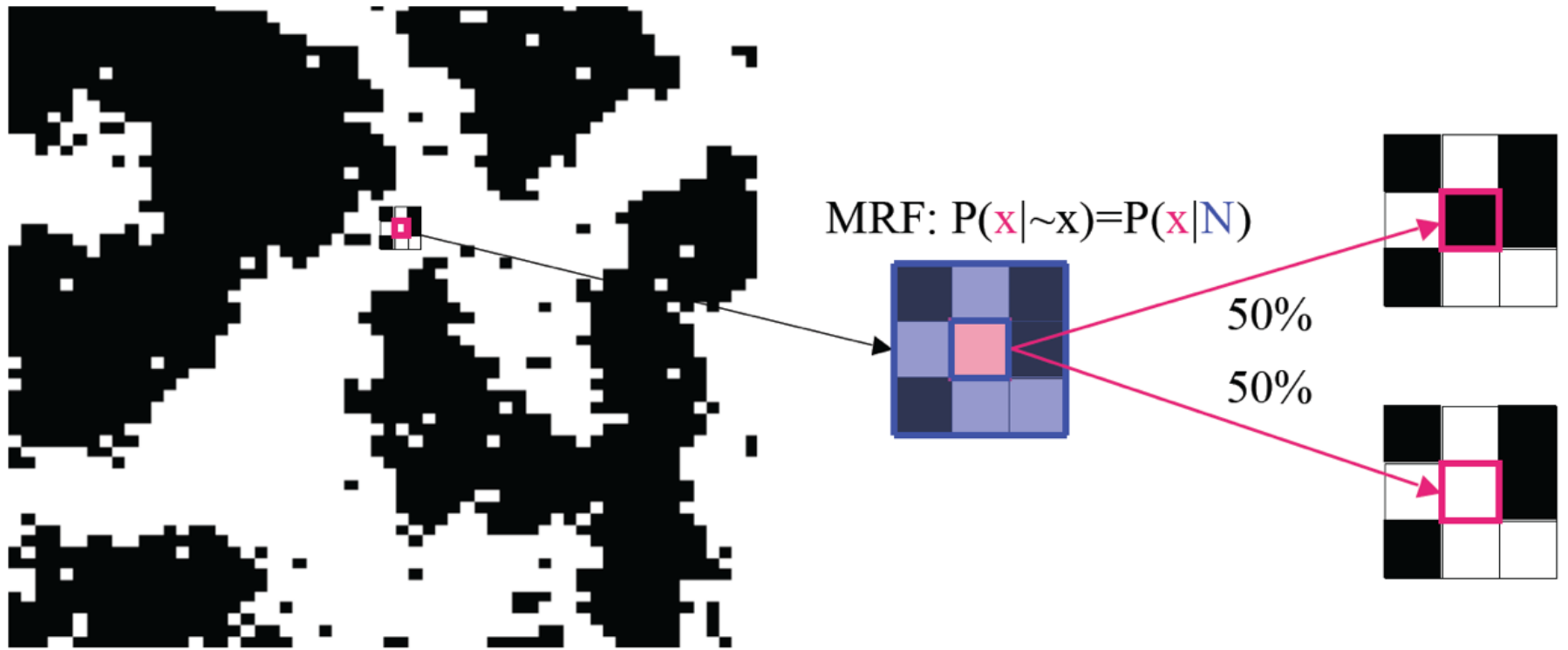
(From Dellaert and Zhu tutorial)

Gibbs as Metropolis Hastings (M-H)

To see Gibbs as MH, and to understand why we always accept, consider that if it were MH, then our proposal distribution, $q_i()$, for a given variable, i, would be

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

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

The "*" here means next state, NOT stationary state.

Gibbs as M-H

$$A\left(\mathbf{z}^{*}, \mathbf{z}\right) = \min\left(1, \frac{p\left(\mathbf{z}^{*}\right)q_{i}\left(\mathbf{z}\middle|\mathbf{z}^{*}\right)}{p\left(\mathbf{z}\right)q_{i}\left(\mathbf{z}^{*}\middle|\mathbf{z}\right)}\right)$$

(def'n of A())

Gibbs as M-H

$$A\left(\mathbf{z}^*, \mathbf{z}\right) = \min\left(1, \frac{p\left(\mathbf{z}^*\right) q_i\left(\mathbf{z} \big| \mathbf{z}^*\right)}{p\left(\mathbf{z}\right) q_i\left(\mathbf{z}^* \big| \mathbf{z}\right)}\right)$$

(def'n of A())

$$= \min\left(1, \frac{p\left(\mathbf{z}^*_{\backslash i}\right) p\left(z^*_i \big| \mathbf{z}^*_{\backslash i}\right) q_i\left(\mathbf{z} \big| \mathbf{z}^*\right)}{p\left(\mathbf{z}_{\backslash i}\right) p\left(z_i \big| \mathbf{z}_{\backslash i}\right) q_i\left(\mathbf{z}^* \big| \mathbf{z}\right)}\right)$$

(because?)

Gibbs as M-H

$$A\left(\mathbf{z}^{*}, \mathbf{z}\right) = \min\left(1, \frac{p\left(\mathbf{z}^{*}\right) q_i\left(\mathbf{z} \middle| \mathbf{z}^{*}\right)}{p\left(\mathbf{z}\right) q_i\left(\mathbf{z}^{*} \middle| \mathbf{z}\right)}\right)$$ (def'n of A())

$$= \min\left(1, \frac{p\left(\mathbf{z}_{\backslash i}^{*}\right) p\left(z_i^{*} \middle| \mathbf{z}_{\backslash i}^{*}\right) q_i\left(\mathbf{z} \middle| \mathbf{z}^{*}\right)}{p\left(\mathbf{z}_{\backslash i}\right) p\left(z_i \middle| \mathbf{z}_{\backslash i}\right) q_i\left(\mathbf{z}^{*} \middle| \mathbf{z}\right)}\right)$$ (def'n of "bar")

$$= \min\left(1, \frac{p\left(\mathbf{z}_{\backslash i}^{*}\right) p\left(z_i^{*} \middle| \mathbf{z}_{\backslash i}^{*}\right) p\left(z_i \middle| \mathbf{z}_{\backslash i}^{*}\right)}{p\left(\mathbf{z}_{\backslash i}\right) p\left(z_i \middle| \mathbf{z}_{\backslash i}\right) p\left(z_i^{*} \middle| \mathbf{z}_{\backslash i}\right)}\right)$$ **(because?)**

Gibbs as M-H

$$A(\mathbf{z}^*, \mathbf{z}) = \min\left(1, \frac{p(\mathbf{z}^*)q_i(\mathbf{z}|\mathbf{z}^*)}{p(\mathbf{z})q_i(\mathbf{z}^*|\mathbf{z})}\right) \quad \text{(def'n of A())}$$

$$= \min\left(1, \frac{p(\mathbf{z}^*_{\backslash i})p(z_i^*|\mathbf{z}^*_{\backslash i})q_i(\mathbf{z}|\mathbf{z}^*)}{p(\mathbf{z}_{\backslash i})p(z_i|\mathbf{z}_{\backslash i})q_i(\mathbf{z}^*|\mathbf{z})}\right) \quad \text{(def'n of "bar")}$$

$$= \min\left(1, \frac{{\color{red}p(\mathbf{z}^*_{\backslash i})}{\color{green}p(z_i^*|\mathbf{z}^*_{\backslash i})}{\color{blue}p(z_i|\mathbf{z}^*_{\backslash i})}}{{\color{red}p(\mathbf{z}_{\backslash i})}{\color{blue}p(z_i|\mathbf{z}_{\backslash i})}{\color{green}p(z_i^*|\mathbf{z}_{\backslash i})}}\right) \quad \text{(Gibbs, coloring)}$$

$$= \min(1,1) \qquad \text{(cancel colors using } \mathbf{z}^*_{\backslash i} = \mathbf{z}_{\backslash i} \text{ , as only } z_i \text{ changes)}$$

$$= 1$$

Exploring the space

- Algorithms like Metropolis-Hastings exhibit "random walk behavior" if the step size (proposal variance) is small
  - Random walk dynamics is practical limitation of MCMC
  - Leads to long mixing times (e.g. long burn-in time)

- If the step size is too big, then you get rejected too often

- Adaptive methods exist (see slice sampling in Bishop)

Gibbs Sampling Extensions

Standard Gibbs suffers same random walk behavior as M-H (but no adjustable parameters, so that's a plus…)

Block Gibbs Jointly sample subset $S \subset \mathcal{V}$ from $p(x_S \mid x_{\neg S})$
- Reduces random walk caused by highly correlated variables
- Requires that conditional $p(x_S \mid x_{\neg S})$ can be sampled efficiently

Collapsed Gibbs Marginalize some variables out of joint:

$$p(x_{\mathcal{V} \setminus S}) = \int p(x) dx_S$$

- Reduces dimensionality of space to be sampled
- Requires that marginals are computable in closed-form

Combined samplers

Different samplers fail in different ways, so combine them…

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.

…can also combine with Gibbs proposals

Mixing MCMC Kernels

Consider a set of MCMC kernels $T_1, T_2, \ldots, T_K$ all having target distribution p(x) then the mixture:

$$T = \sum_{k=1}^{K} \pi_k T_k$$

→ **Mixing weights**

Is a valid MCMC kernel with target distribution p(x)

Mixture MCMC Transition kernel given by:
1. Sample $k \sim \pi$
2. Sample $x^{(t+1)} \sim T_k(x \mid x^{(t)})$

MCMC Summary

- Markov chain induced by MCMC transition kernel T(z,z')

- Converges to stationary distribution iff chain is **ergodic**
  - Chain is ergodic if it is **irreducible** (can get from any z to any z') and **aperiodic** (doesn't get trapped in cycles)

- Easier to prove **detailed balance,** which implies ergodicity

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

- Metropolis algorithm samples from symmetric proposal q(z'|z) and accepts sample z' with probability,

$$A = \min\left(1, \frac{\widetilde{p}(z')}{\widetilde{p}(z)}\right)$$

- Metropolis-Hastings allows non-symmetric proposal q(z'|z) and accepts sample z' with probability,

$$A = \min\left(1, \frac{\widetilde{p}(z')}{\widetilde{p}(z)} \frac{q(z \mid z')}{q(z' \mid z)}\right)$$

- Gibbs sampler on random vector $z = (z_1, \dots, z_d)^T$ successively samples from *complete conditionals*,

$$z_1^{\mathrm{new}} \sim p(z_1 \mid z_2^{\mathrm{old}}, \dots, z_d^{\mathrm{old}})$$

$$z_2^{\mathrm{new}} \sim p(z_2 \mid z_1^{\mathrm{new}}, z_3^{\mathrm{old}}, \dots, z_d^{\mathrm{old}})$$

$$\dots$$

$$z_d^{\mathrm{new}} \sim p(z_d \mid z_1^{\mathrm{new}}, \dots, z_{d-1}^{\mathrm{new}})$$

- Gibbs is instance of M-H which *always accepts*

Inference (and related) Tasks

- Simulation: $x \sim p(x) = \dfrac{1}{Z} f(x)$

- Compute expectations: $\mathbb{E}[\phi(x)] = \displaystyle\int p(x) \phi(x) \, dx$

- Optimization: $x^* = \arg\max_x f(x)$

- Compute normalizer: $Z = \displaystyle\int f(x) \, dx$

Inference (and related) Tasks

- Simulation: $x \sim p(x) = \dfrac{1}{Z} f(x)$

- Compute expectations: $\mathbb{E}[\phi(x)] = \displaystyle\int p(x)\phi(x)\,dx$

- Optimization: $x^* = \arg\max_x f(x)$

- Compute normalizer: $Z = \displaystyle\int f(x)\,dx$

Simulated 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.

Simulated 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).

- The terminology makes sense because the number of states accessible to a physical system decreases with temperature.

Simulated 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 $u < \mathcal{A}(x^{(i)}, x^\star) = \min\left\{1, \dfrac{p^{\frac{1}{T_i}}(x^\star) q(x^{(i)} | x^\star)}{p^{\frac{1}{T_i}}(x^{(i)}) q(x^\star | x^{(i)})}\right\}$

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

     else

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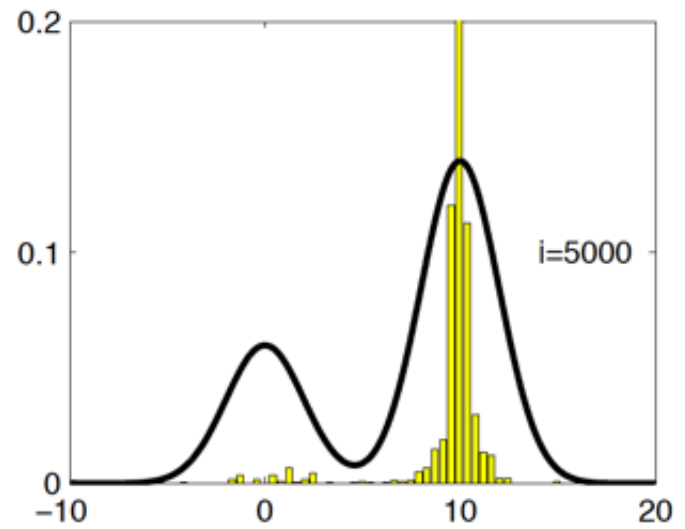
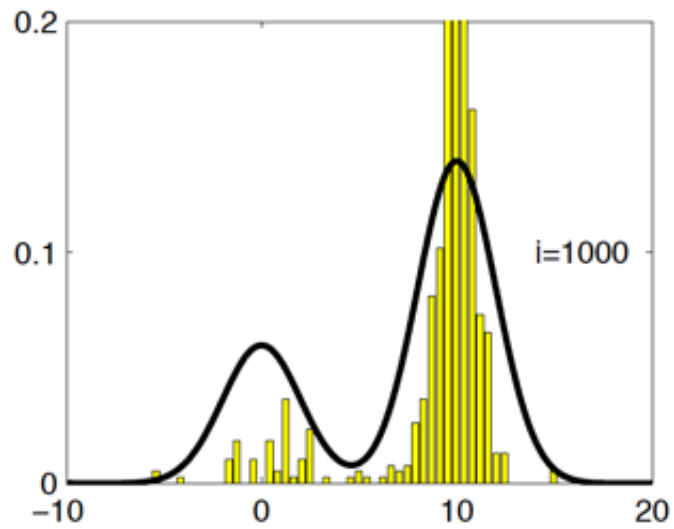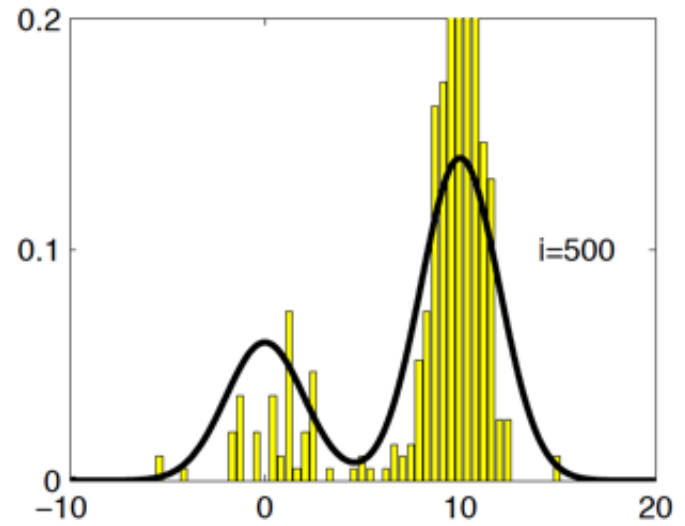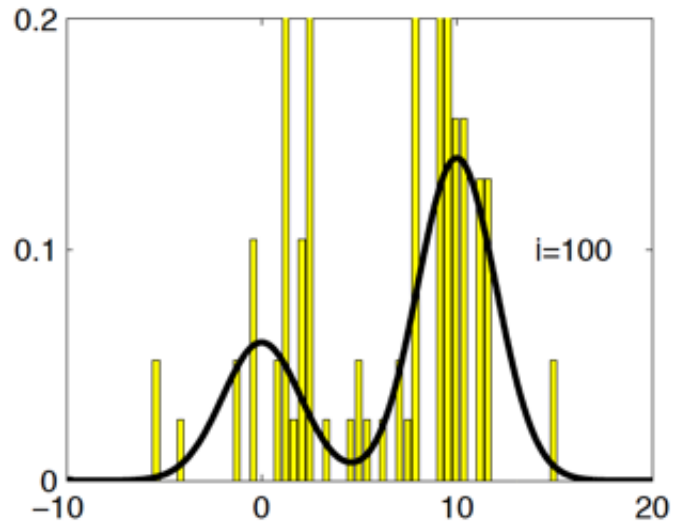   - Set T_{i+1} according to a chosen cooling schedule.

(From Andrieu et al)

i=100
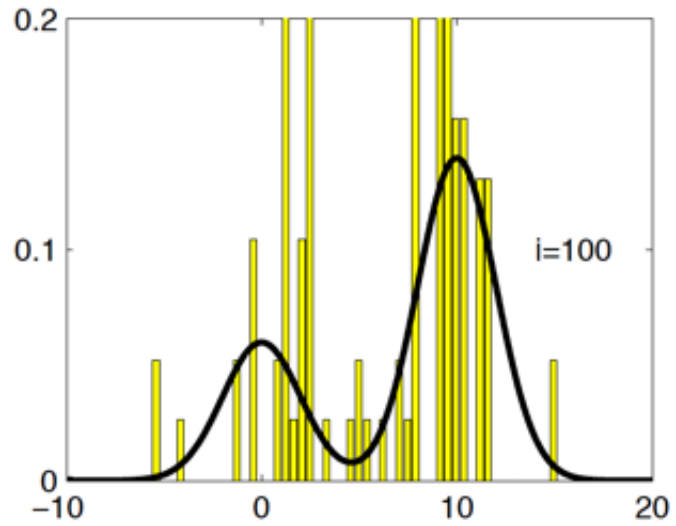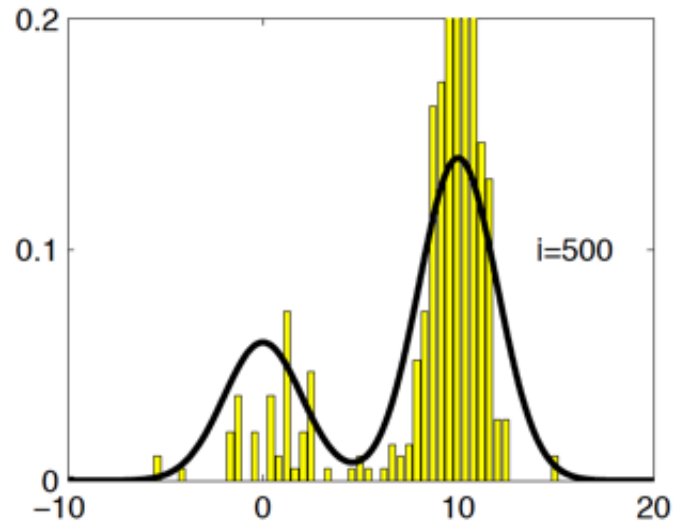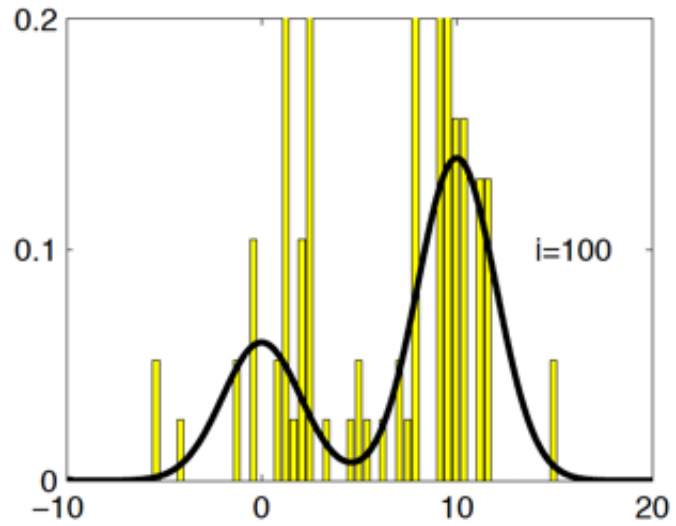
(From Andrieu et al)
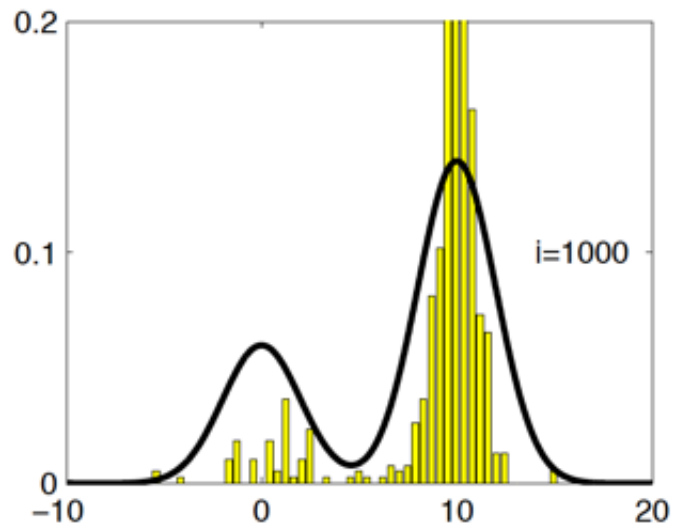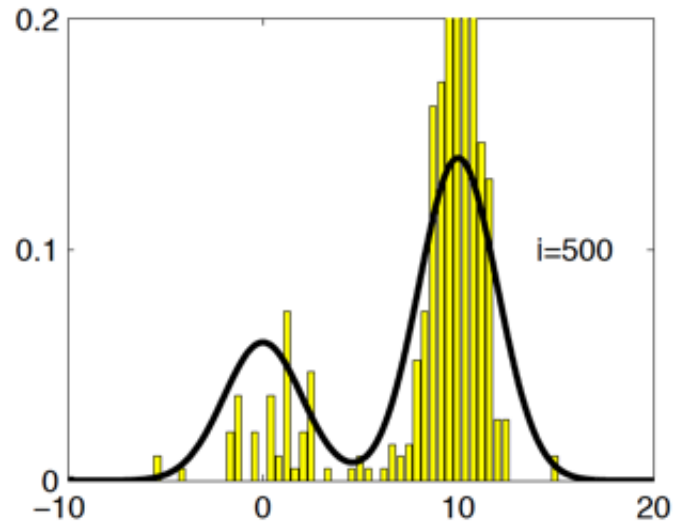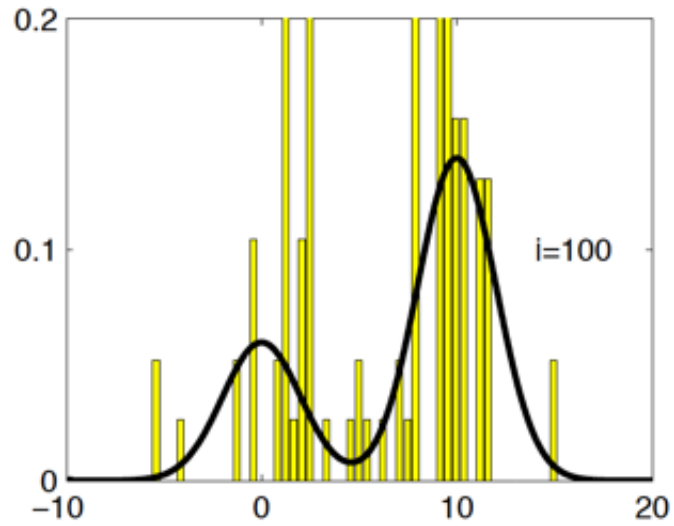
(From Andrieu et al)

(From Andrieu et al)

(From Andrieu et al)

Simulated Annealing

Let *annealing distribution* at temp τ be given by:

$$p_\tau(x) \propto (f(x))^{1/\tau}$$

As $\tau \to 0$ we have:

$$\lim_{\tau \to 0} p_\tau(x) = \delta(x^*) \qquad \text{where} \qquad x^* = \arg\max_x f(x)$$

SA for Global Optimization:
Annealing schedule $\tau_0 \geq \ldots \geq \tau_t \geq \ldots \geq 0$
1. Sample $x^{(t)}$ from MCMC kernel T_t with target $p_{\tau_t}(x)$
2. Set τ_{t+1} according to annealing schedule

SA for Convergence: $\tau_0 \geq \ldots \geq 1$ Final temperature = 1