Top ten algorithms for scientific computing

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler
- QR Algorithm for Computing Eigenvalues
- Quicksort Algorithm for Sorting
- Fast Fourier Transform
- Integer Relation Detection
- Fast Multipole Method

Metropolis-Hastings MCMC method

```
While not_bored
{
    Sample $q\left(z \middle| z^{(prev)}\right)$
    Accept with probability $A\left(z, z^{(prev)}\right) = \min\left(1, \dfrac{\tilde{p}(z)\, q\left(z^{(prev)} \middle| z\right)}{\tilde{p}\left(z^{(prev)}\right) q\left(z \middle| z^{(prev)}\right)}\right)$
    If accept, emit $z$, otherwise, emit $z^{(prev)}$.
}
```

Does Metropolis-Hastings 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

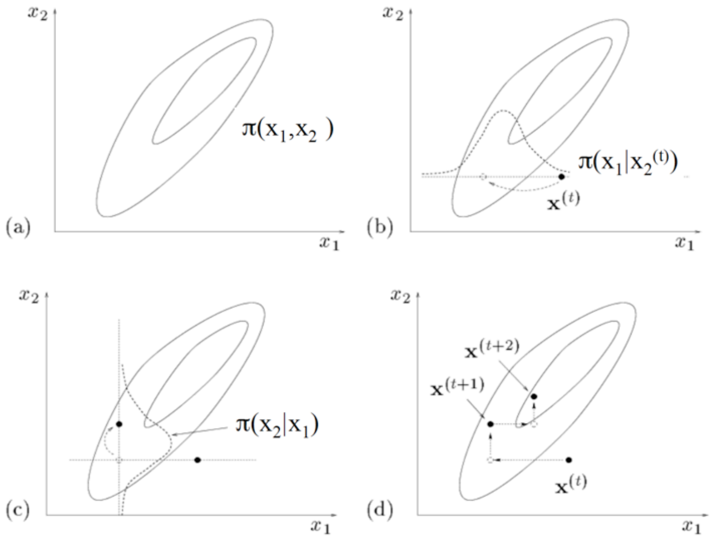
- Gibbs sampling is another special case of MH.

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



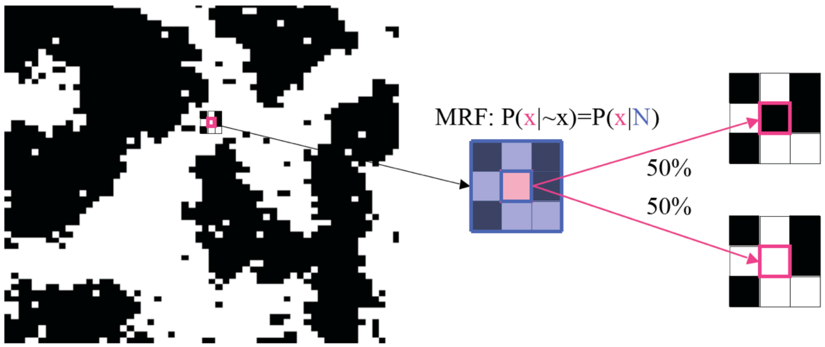
(From Dellaert and Zhu tutorial)

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



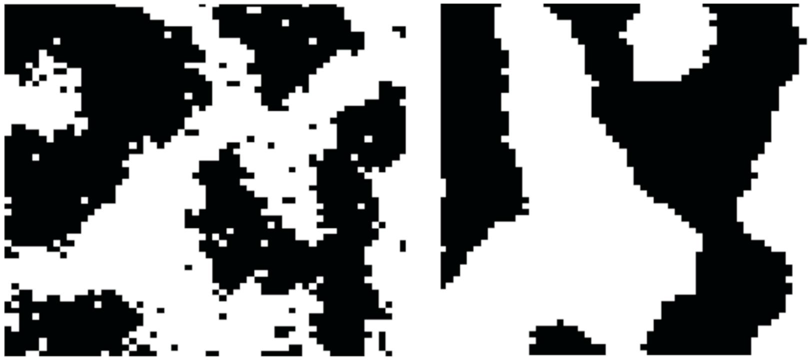
MRF: P(x|~x)=P(x|N)

50%

50%

(From Dellaert and Zhu tutorial)

Examples of Gibbs



Weak Affinity to Neighbors

Strong Affinity to Neighbors

(From Dellaert and Zhu tutorial)

Gibbs as MH

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

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

Gibbs as MH

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

$$
\begin{aligned}
& q_i\left(\mathbf{z}\middle|\mathbf{z}^{*}\right) = p\left(\mathbf{z}_i\middle|\mathbf{z}_{\backslash \mathrm{i}}^{*}\right) \\
\text{and} \quad & q_i\left(\mathbf{z}^{*}\middle|\mathbf{z}\right) = p\left(\mathbf{z}_i^{*}\middle|\mathbf{z}_{\backslash \mathrm{i}}\right) \\
\text{and} \quad & \mathbf{z}_{\backslash \mathrm{i}} = \mathbf{z}_{\backslash \mathrm{i}}^{*}
\end{aligned}
$$

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

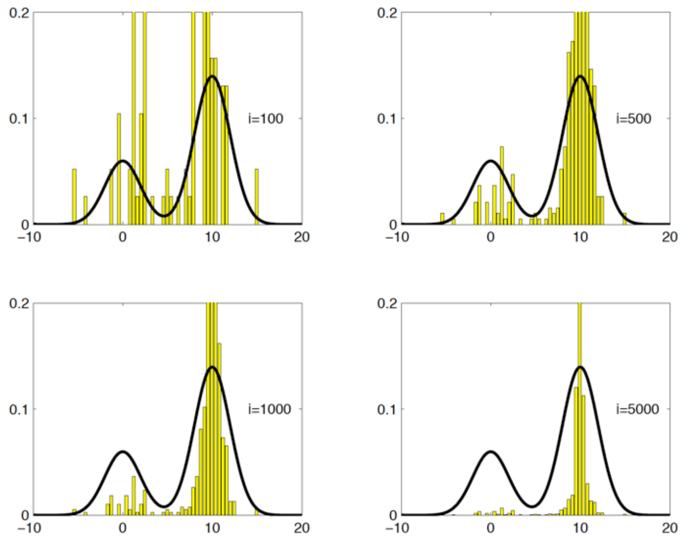
- Define a temperature T, and a cooling schedule (black magic part)

- 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

- 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