Sampling for Inferencing

Tractability of Bayesian Inferencing

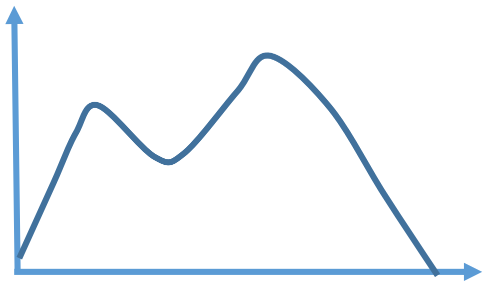
- $p(\theta|x) = \underbrace{p(x|\theta)}_{\text{data}}\ \underbrace{p(\theta)}_{\text{belief}} / \underbrace{p(x)}_{\text{evidence}} = p(x|\theta)\ p(\theta) / \int p(x|\theta)\ p(\theta) d\theta$

- The (normalizing) denominator is difficult to compute in closed
  - We only know the posterior distribution up to the normalizing factor
  - Lets call it $p^{\text{tilde}}(\theta|x) = p(x|\theta)\ p(\theta)$

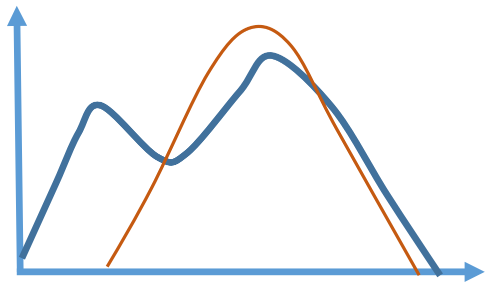
- Need for Approximate Bayesian methods

Sampling Approximation of Distributions

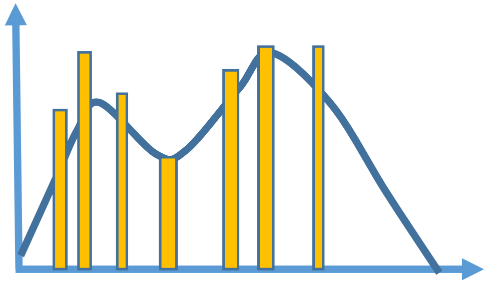
- Approximate the posterior distribution using a sampled data set



Original Distribution



Mean Field VI Approximation



Sampling Approximation
(Union of delta functions
at the sample points)

If we can draw large number of samples from the target distribution we get a good approximation

Difficulties in Sampling

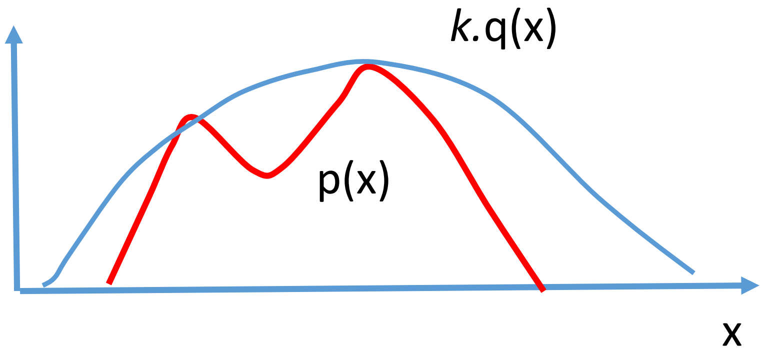
- The target distribution is unknown/intractable
  - Or known only up to the normalizing constant

- Can we use another known/standard distribution to draw a sample from the target distribution

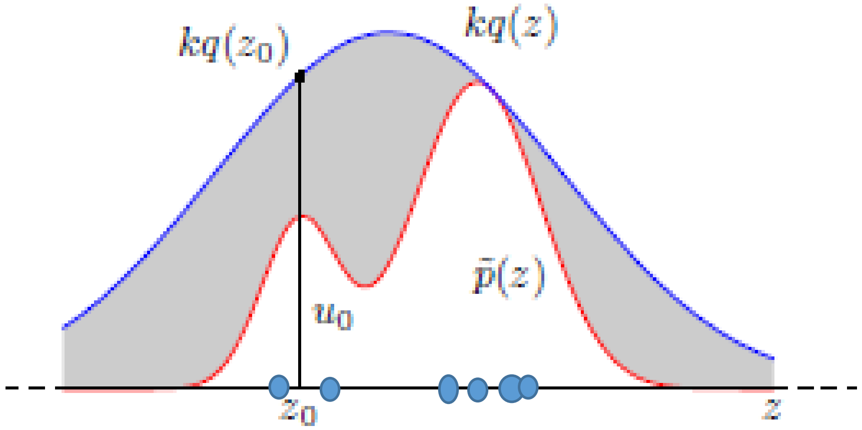
- Often you have access to a uniform random number generator, or even a randn() function to generate a random variable distributed according to standard normal distribution

Univariate Sampling

- Target distribution – p(x) (intractable)
- Let us upper bound p(x) by a known distribution q(x) times a const. *k.*
  - Constant *k* is a must for upper bounding
  - q(x) is a standard distribution, say Normal with variance unity
- If the upper bound is tight every where, samples drawn from *k*.q(x) will approximate the target distribution p(x)



Rejection Sampling



Steps:

1. Generate a sample z_0, using q(z)
2. Accept it with probability $u_0/kq(z_0)$, $u_0 = p^{tilde}(z_0)$

The grey part is the rejection region.

More points generated near the maxima of the target distribution.

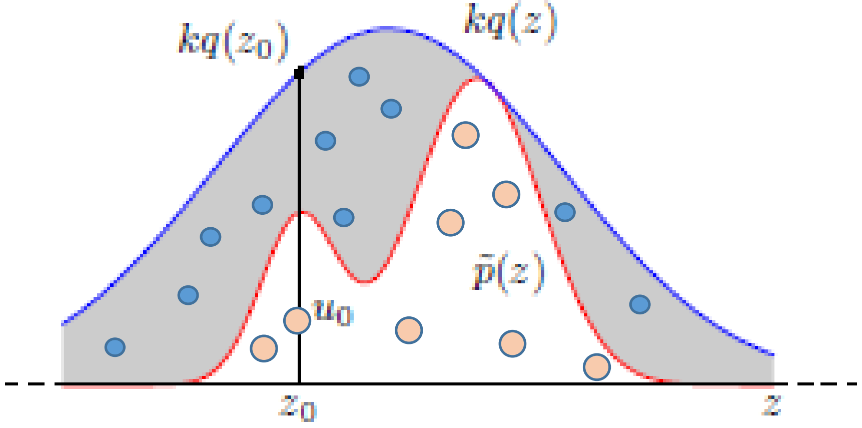
How many samples get rejected?

- Fraction of samples accepted is proportional to the ratio of the (white area)/(grey + white area) = *k*

(Since, area under a distribution function is 1)

p^{tilde}(z) is same as p(z) up to normalizing constant

We can absorb the normalizing constant within *k*, and use p^{tilde}(z) in rejection sampling



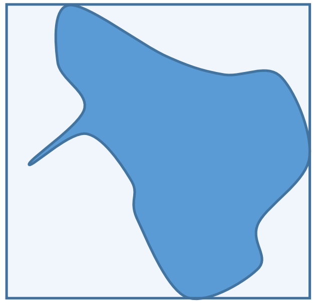
Monte Carlo Approximation

- Simulation method to approximately compute area of complex regions

Probabilistic model: $p(\boldsymbol{x}) = \frac{1}{Z}\tilde{p}(\boldsymbol{x})$, $Z = \int \tilde{p}(\boldsymbol{x}) d\boldsymbol{x}$.

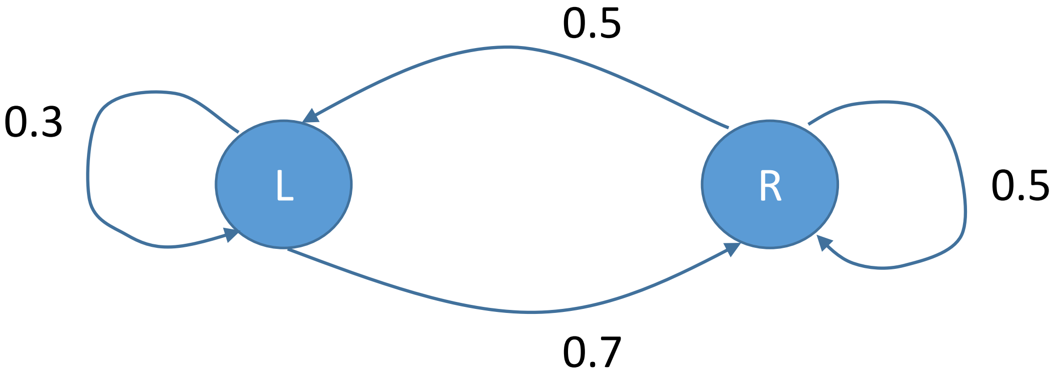
construct samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ from $p(\boldsymbol{x})$ using only $\tilde{p}(\boldsymbol{x})$.

Use the samples to approximately compute expectations.



Usage: $\mathbb{E}_{p(\boldsymbol{x})} f(\boldsymbol{x}) = \int f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \approx \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{x}_i)$.

Markov Chain



T(L → R) = 0.7
T(R → L) = 0.5
T(L → L) = 0.3
T(R → R) = 0.5

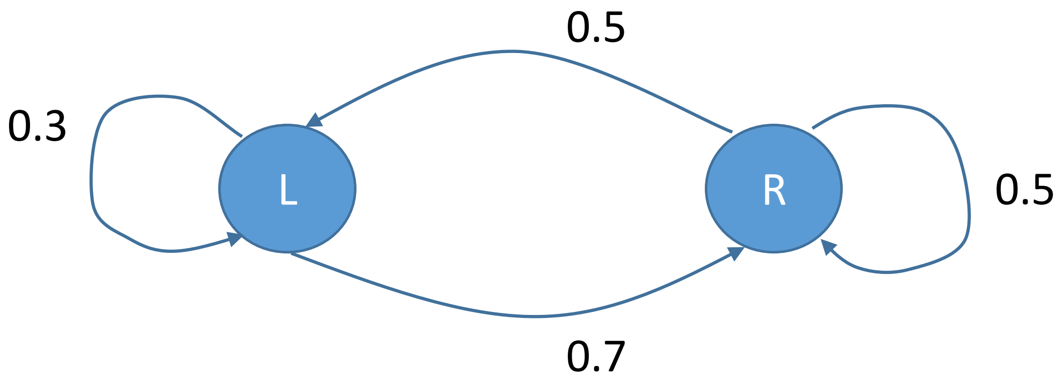
Transition depend on current state only.

Simulation of a Markov chain with time steps:

LLRLRRRLLLL

What is the probability that the system is in state L after t time steps?

Markov Chains



Time Step	p(x=L)	p(x=R)
x^1	1	0
x^2	0.3	0.7
x^3	0.3x0.3+0.7x0.5	…
x^{1M}	0.42	0.58
x^{2M}	0.42	0.58

← (converging probabilities, defines a distribution over states)

$p(x^3) = p(x^3 | x^2 = L)p(x^2 = L) + p(x^3 | x^2 = R)p(x^2 = R)$ [Marginalization over previous state values]

Different ways it can reach L/R in the third step

Using Markov Chains to Generate Samples

Simulate the Markov chain a large number of times:

LRRRLLRL…LRR
LRLRLRRL…LRR
LRRRLLRL…LRL
LRLRLRLL…LRL
LRRRRLRL…LRR

p(L) ~ 2/5 = 0.42
p(R) ~ 3/5 = 0.58

If we look at the final value of the state (L/R) at the end of each simulation we get a sample distributed according to a Bernoulli distribution with probabilities 0.42 and 1 - 0.42.

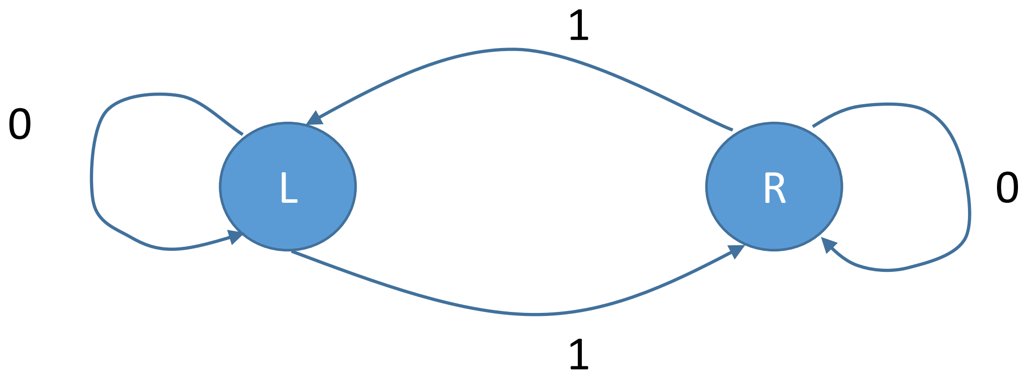
In practice, we need not look at just the last value, but may throw away first 1000 values in each chain and use the rest of the values as samples.

The principle can be extended to discrete distributions with many possible value,
or to continuous distributions where the marginalization is an integration.

Markov Chain Sampling

- We want to sample from p(x)
- Build a Markov chain that converges to p(x)
- Start from any state x^0
- Generate $x^{k+1} \sim T(x^{k+1} \rightarrow x^k)$
- Eventually x^k will look like samples from p(x)

Does a Markov Chain always converge?



It oscillates between L and R

Stationary Distribution

- p(x) is a stationary distribution for a Markov chain iff

$$p(x') = \sum_x T(x'|x)p(x)$$

If we start from the distribution p(x) over states, and transition one time step we get the same distribution p(x)

Once we encounter the stationary distribution during the sampling process, we stay at it henceforth

Theorem

- If $T(x'|x) > 0$ for all x’, x

(any state can be reached from any other state)

- Then there exists an unique stationary distribution for the Markov chain

- And the Markov chain converges to the stationary distribution from any starting point (ergodic)

Markov Chain Monte Carlo

- Given a target distribution p(x)

- How do you design a Markov chain such that simulating it generates samples from p(x)? - stationary distribution
  - Combine with Monte Carlo sampling

- Markov Chain Monte Carlo (MCMC)

Gibbs Sampling

Goal: generate samples from $p(\boldsymbol{x}) = \frac{1}{Z}\tilde{p}(\boldsymbol{x})$, where $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k]$.

Suppose $\boldsymbol{x} \sim p(\boldsymbol{x})$. Then the next point $\boldsymbol{x}^{new}$ is generated as follows: (Markov chain transition)

- $\boldsymbol{x}_1^{new} \sim p(\boldsymbol{x}_1 | \boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_k)$; (1D distributions are easy to sample from)
- $\boldsymbol{x}_2^{new} \sim p(\boldsymbol{x}_2 | \boldsymbol{x}_1^{new}, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_k)$;
- $\ldots$;
- $\boldsymbol{x}_k^{new} \sim p(\boldsymbol{x}_k | \boldsymbol{x}_1^{new}, \boldsymbol{x}_3^{new}, \ldots, \boldsymbol{x}_{k-1}^{new})$;
- $\boldsymbol{x}^{new} = [\boldsymbol{x}_1^{new}, \boldsymbol{x}_2^{new}, \ldots, \boldsymbol{x}_k^{new}]$.

It is easy to show than $p(\boldsymbol{x})$ is invariant under such transition probability. If all conditionals $p(\boldsymbol{x}_i | \boldsymbol{x}_{\setminus i}) > 0$, then the corresponding Markov chain would be ergodic.

Gibbs Sampling: Proof of Correctness

- p(x) is a stationary distribution if: $p(X') = \sum_{X} T(X \rightarrow X')p(X)$
- Let us consider a 3D state X=(x,y,z) with transition: T(x,y,z → x’,y’,z’)

$$\sum_{x,y,z} T(x,y,z \rightarrow x',y',z')p(x,y,z)$$

$$= \sum_{x,y,z} p(x'|y=y,z=z)p(y'|x=x',z=z)p(z'|x=x',y=y')p(x,y,z)$$

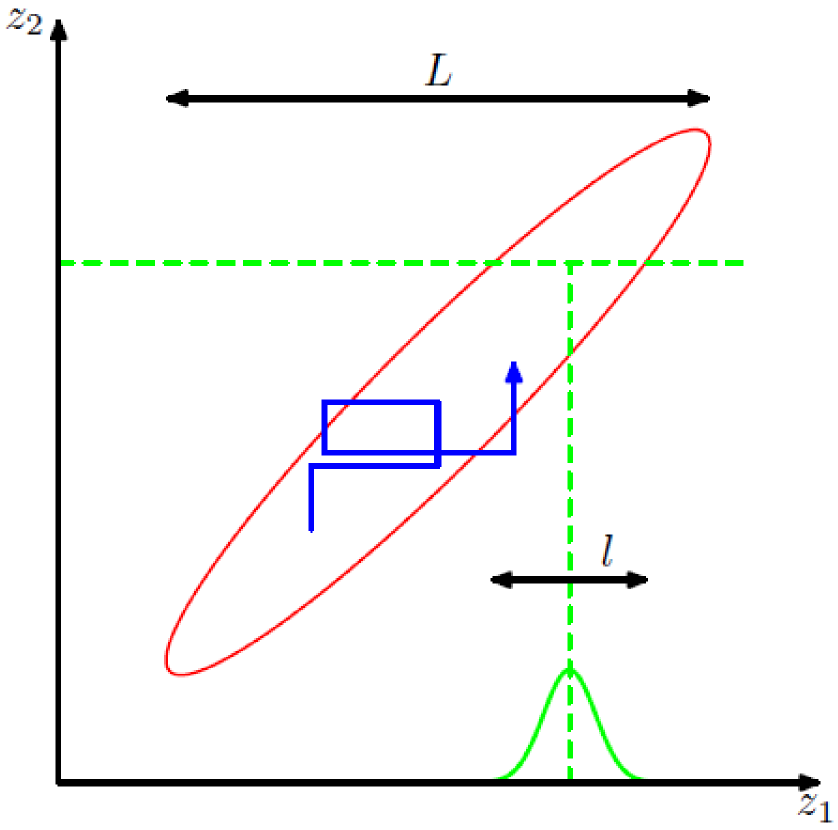
$$= p(z'|x',y')\sum_{y,z}(p(x'|y,z)p(y'|x',z)\sum_{x} p(x,y,z))$$

$$= p(z'|x',y')\sum_{z}(p(y'|x',z)\sum_{y} p(x',y,z))$$

$$= p(z'|x',y')\sum_{z} p(y',x',z) = p(z'|x',y')p(y',x') = p(x',y',z')$$

Example: Bivariate Gaussian

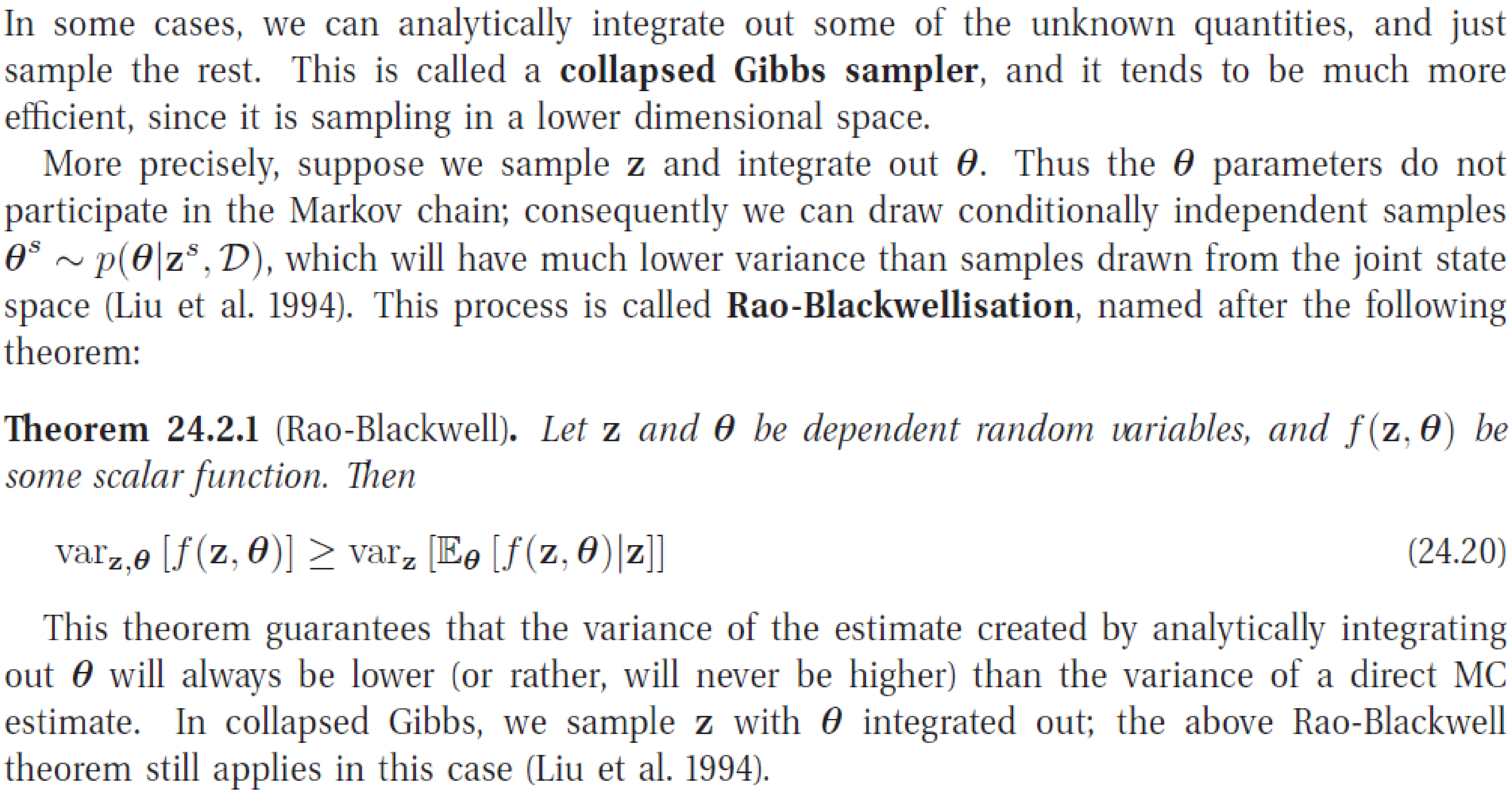
Illustration of Gibbs sampling by alternate updates of two variables whose distribution is a correlated Gaussian. The step size is governed by the standard deviation of the conditional distribution (green curve), and is $O(l)$, leading to slow progress in the direction of elongation of the joint distribution (red ellipse). The number of steps needed to obtain an independent sample from the distribution is $O((L/l)^2)$.



Summary of Gibbs Sampling

- Long burn-in phase
- Small steps if variables are correlated
- Can not be parallelized
- Large number of steps for high dimension

Collapsed Gibbs Sampling



In some cases, we can analytically integrate out some of the unknown quantities, and just sample the rest. This is called a **collapsed Gibbs sampler**, and it tends to be much more efficient, since it is sampling in a lower dimensional space.

More precisely, suppose we sample $\mathbf{z}$ and integrate out $\boldsymbol{\theta}$. Thus the $\boldsymbol{\theta}$ parameters do not participate in the Markov chain; consequently we can draw conditionally independent samples $\boldsymbol{\theta}^s \sim p(\boldsymbol{\theta}|\mathbf{z}^s, \mathcal{D})$, which will have much lower variance than samples drawn from the joint state space (Liu et al. 1994). This process is called **Rao-Blackwellisation**, named after the following theorem:

Theorem 24.2.1 (Rao-Blackwell)**.** *Let* $\mathbf{z}$ *and* $\boldsymbol{\theta}$ *be dependent random variables, and* $f(\mathbf{z}, \boldsymbol{\theta})$ *be some scalar function. Then*

$$\text{var}_{\mathbf{z},\boldsymbol{\theta}}\left[f(\mathbf{z}, \boldsymbol{\theta})\right] \geq \text{var}_{\mathbf{z}}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[f(\mathbf{z}, \boldsymbol{\theta})|\mathbf{z}\right]\right] \tag{24.20}$$

This theorem guarantees that the variance of the estimate created by analytically integrating out $\boldsymbol{\theta}$ will always be lower (or rather, will never be higher) than the variance of a direct MC estimate. In collapsed Gibbs, we sample $\mathbf{z}$ with $\boldsymbol{\theta}$ integrated out; the above Rao-Blackwell theorem still applies in this case (Liu et al. 1994).

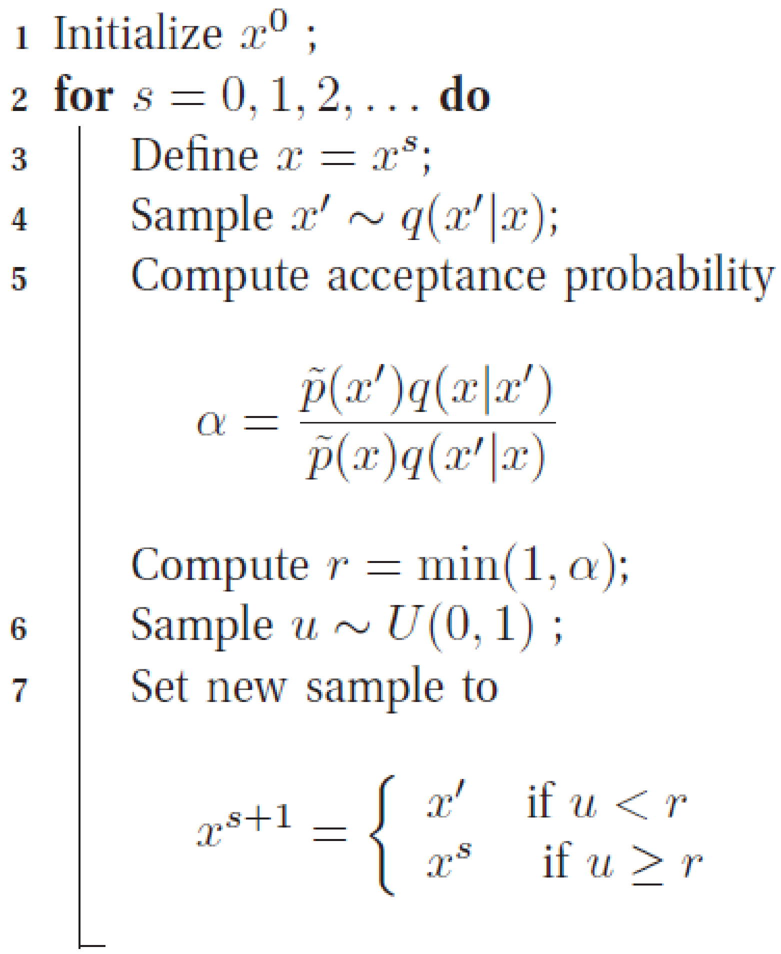
Metropolis-Hastings Sampling

- Propose larger transitions – reject if they are not good
- Proposal – critic framework

- Monte Carlo sampling over Markov chains

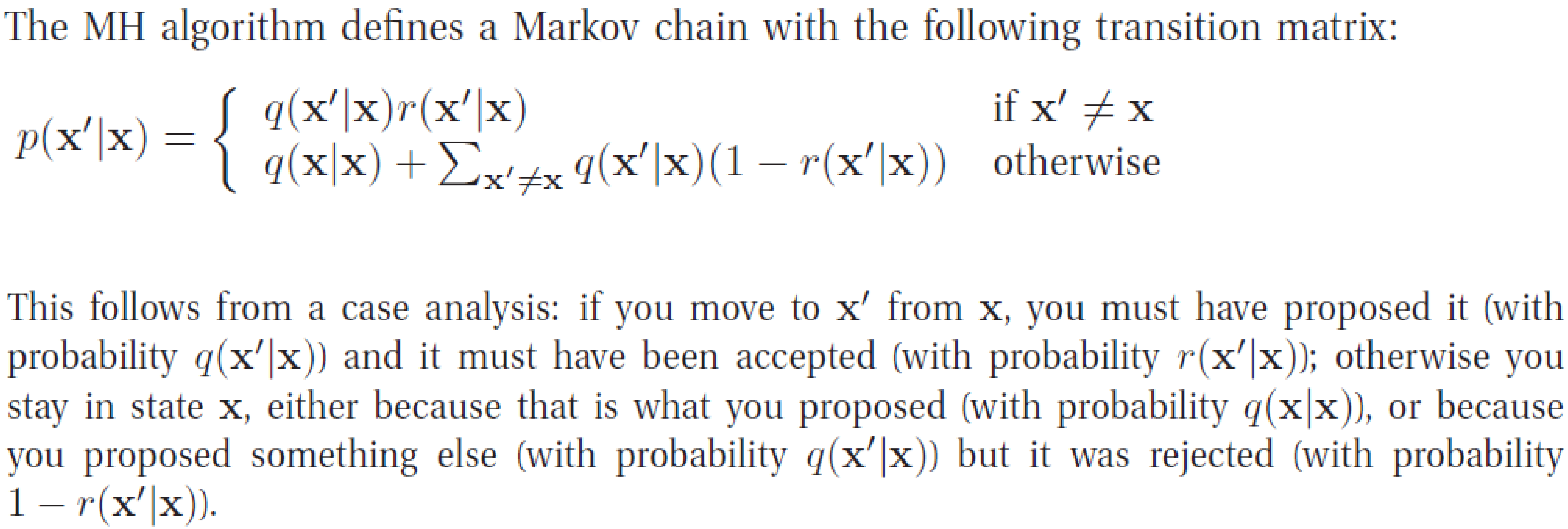
- Markov Chain Monte Carlo

Metropolis-Hastings (MH) Algorithm



```
Initialize $x^0$ ;
for $s = 0, 1, 2, \ldots$ do
    Define $x = x^s$;
    Sample $x' \sim q(x'|x)$;
    Compute acceptance probability
        $\alpha = \frac{\tilde{p}(x')q(x|x')}{\tilde{p}(x)q(x'|x)}$
    Compute $r = \min(1, \alpha)$;
    Sample $u \sim U(0, 1)$ ;
    Set new sample to
        $x^{s+1} = \begin{cases} x' & \text{if } u < r \\ x^s & \text{if } u \geq r \end{cases}$
```

MH as a Markov Chain



The MH algorithm defines a Markov chain with the following transition matrix:

$$p(\mathbf{x}'|\mathbf{x}) = \begin{cases} q(\mathbf{x}'|\mathbf{x})r(\mathbf{x}'|\mathbf{x}) & \text{if } \mathbf{x}' \neq \mathbf{x} \\ q(\mathbf{x}|\mathbf{x}) + \sum_{\mathbf{x}' \neq \mathbf{x}} q(\mathbf{x}'|\mathbf{x})(1 - r(\mathbf{x}'|\mathbf{x})) & \text{otherwise} \end{cases}$$

This follows from a case analysis: if you move to $\mathbf{x}'$ from $\mathbf{x}$, you must have proposed it (with probability $q(\mathbf{x}'|\mathbf{x})$) and it must have been accepted (with probability $r(\mathbf{x}'|\mathbf{x})$); otherwise you stay in state $\mathbf{x}$, either because that is what you proposed (with probability $q(\mathbf{x}|\mathbf{x})$), or because you proposed something else (with probability $q(\mathbf{x}'|\mathbf{x})$) but it was rejected (with probability $1 - r(\mathbf{x}'|\mathbf{x})$).

Detailed Balance

How do we construct a transition $q(x'|x)$ with given $p(x)$ as its stationary distribution? This problem can be simplified if we consider special transitions that satisfy the *detailed balance* condition. If we are given the marginal distribution $p(x)$, the detailed balance condition for a transition q is

$$\frac{q(x'|x)}{q(x|x')} = \frac{p(x')}{p(x)}, \qquad \forall x, x'$$

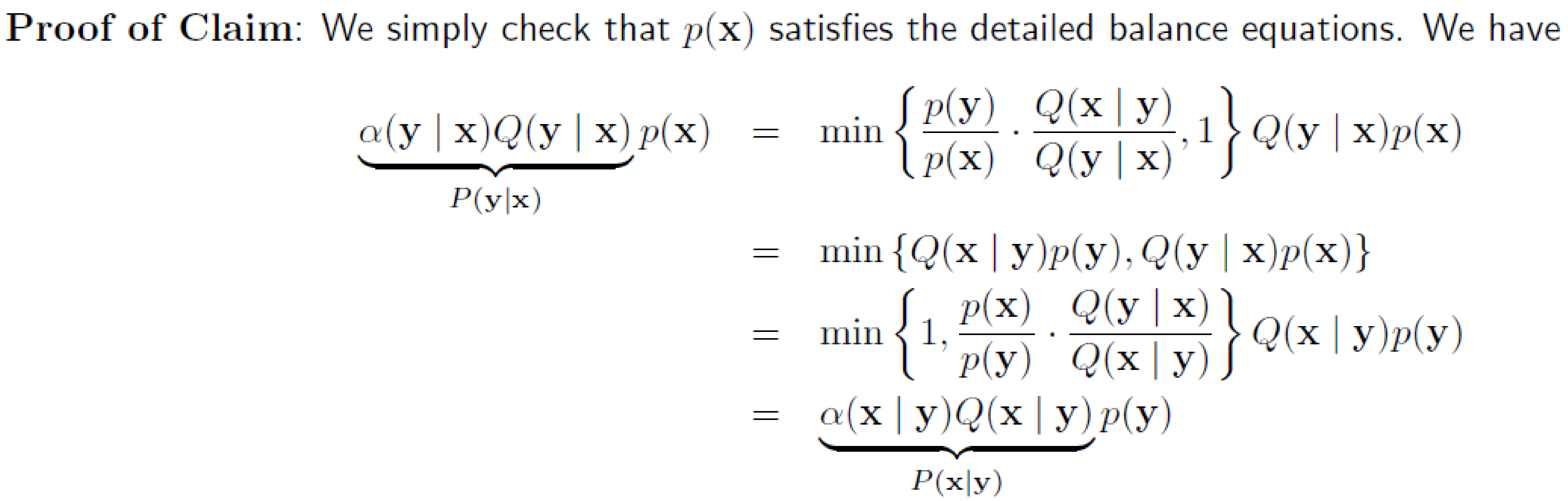
Only ratios of distributions are considered.
No need of the normalizing constant.

Since:

$$\int_x q(x'|x)p(x) = \int_x q(x|x')p(x') = p(x')$$

so that $p(x)$ is the stationary distribution of $q(x'|x)$.

Detailed Balance Holds for MH

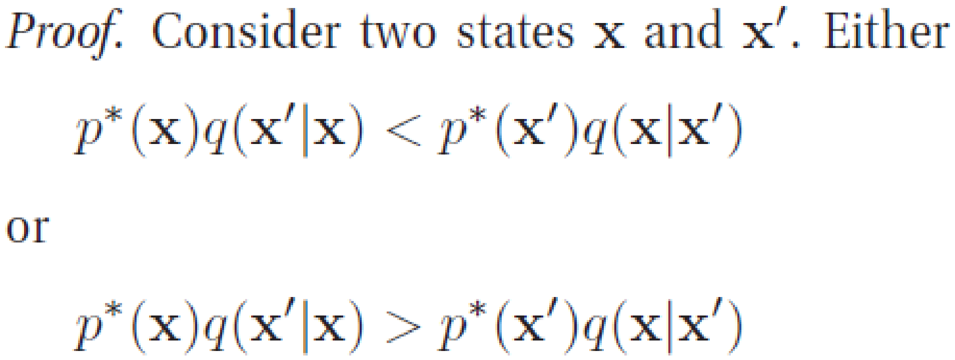


Proof of Claim: We simply check that $p(\mathbf{x})$ satisfies the detailed balance equations. We have

$$\begin{aligned}
\underbrace{\alpha(\mathbf{y} \mid \mathbf{x}) Q(\mathbf{y} \mid \mathbf{x})}_{P(\mathbf{y}|\mathbf{x})} p(\mathbf{x}) &= \min \left\{ \frac{p(\mathbf{y})}{p(\mathbf{x})} \cdot \frac{Q(\mathbf{x} \mid \mathbf{y})}{Q(\mathbf{y} \mid \mathbf{x})}, 1 \right\} Q(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) \\
&= \min \{ Q(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y}), Q(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) \} \\
&= \min \left\{ 1, \frac{p(\mathbf{x})}{p(\mathbf{y})} \cdot \frac{Q(\mathbf{y} \mid \mathbf{x})}{Q(\mathbf{x} \mid \mathbf{y})} \right\} Q(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y}) \\
&= \underbrace{\alpha(\mathbf{x} \mid \mathbf{y}) Q(\mathbf{x} \mid \mathbf{y})}_{P(\mathbf{x}|\mathbf{y})} p(\mathbf{y})
\end{aligned}$$

Proof of Correctness of MH

Theorem: The transition matrix of the MH algorithm has p*(x) as its stationary distribution



Proof. Consider two states $\mathbf{x}$ and $\mathbf{x}'$. Either

$$p^*(\mathbf{x})q(\mathbf{x}'|\mathbf{x}) < p^*(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')$$

or

$$p^*(\mathbf{x})q(\mathbf{x}'|\mathbf{x}) > p^*(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')$$

Forward Case

$$\alpha(\mathbf{x}'|\mathbf{x}) = \frac{p^*(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p^*(\mathbf{x})q(\mathbf{x}'|\mathbf{x})} < 1$$

assume that $p^*(\mathbf{x})q(\mathbf{x}'|\mathbf{x}) > p^*(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')$.

Hence we have $r(\mathbf{x}'|\mathbf{x}) = \alpha(\mathbf{x}'|\mathbf{x})$ and $r(\mathbf{x}|\mathbf{x}') = 1$.

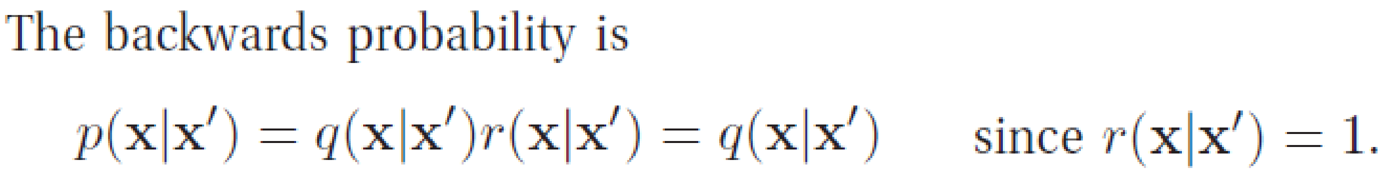
Now to move from $\mathbf{x}$ to $\mathbf{x}'$ we must first propose $\mathbf{x}'$ and then accept it. Hence

$$p(\mathbf{x}'|\mathbf{x}) = q(\mathbf{x}'|\mathbf{x})r(\mathbf{x}'|\mathbf{x}) = q(\mathbf{x}'|\mathbf{x})\frac{p^*(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p^*(\mathbf{x})q(\mathbf{x}'|\mathbf{x})} = \frac{p^*(\mathbf{x}')}{p^*(\mathbf{x})}q(\mathbf{x}|\mathbf{x}')$$

Hence

$$p^*(\mathbf{x})p(\mathbf{x}'|\mathbf{x}) = p^*(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')$$

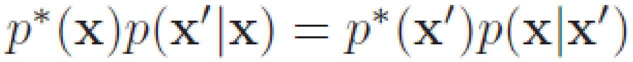
Backward Case



The backwards probability is

$$p(\mathbf{x}|\mathbf{x}') = q(\mathbf{x}|\mathbf{x}')r(\mathbf{x}|\mathbf{x}') = q(\mathbf{x}|\mathbf{x}') \qquad \text{since } r(\mathbf{x}|\mathbf{x}') = 1.$$

Here also details balance holds.

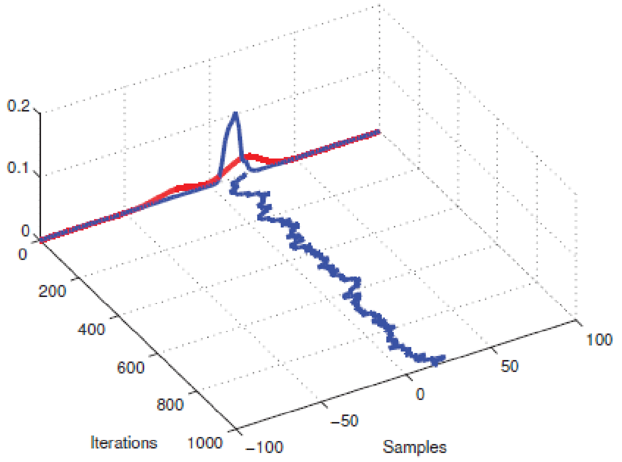


$$p^*(\mathbf{x})p(\mathbf{x}'|\mathbf{x}) = p^*(\mathbf{x}')p(\mathbf{x}|\mathbf{x}')$$

Hence proved the target distribution is the stationary distribution for MH.

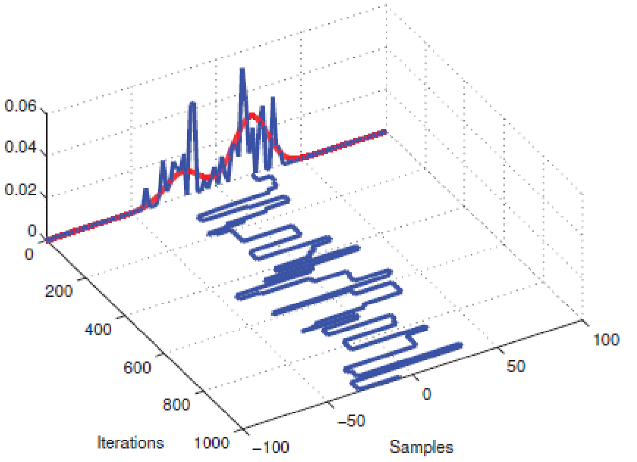
MH: Gaussian Proposal

MH with N(0,1.000^2) proposal



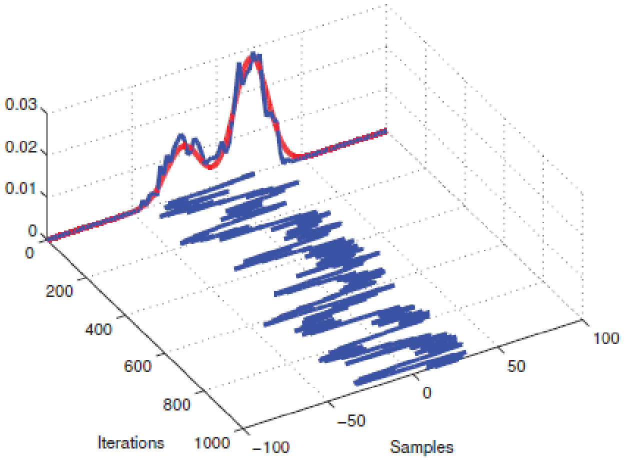
(a)

MH with N(0,500.000^2) proposal



(b)

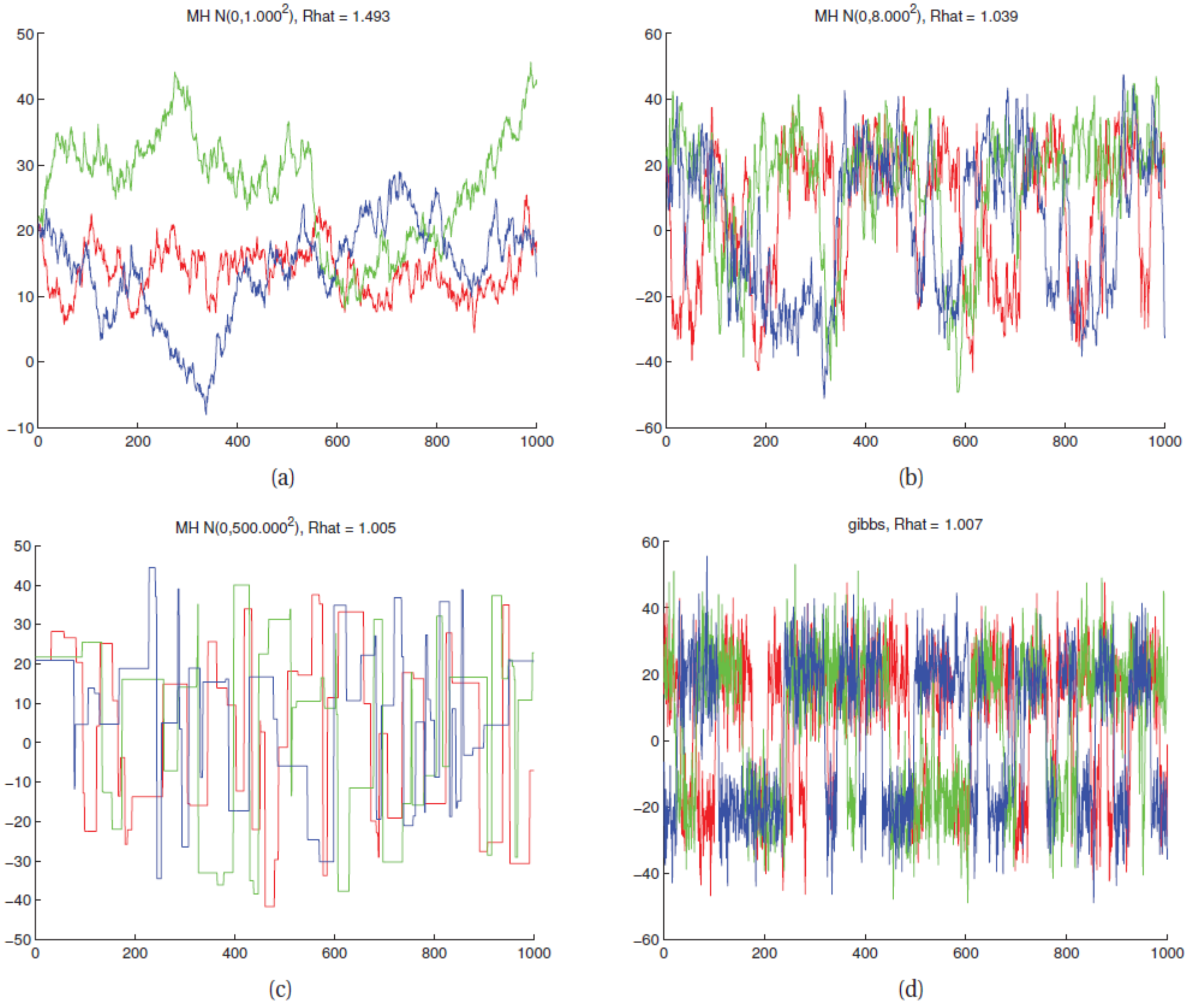
MH with N(0,8.000^2) proposal



(c)

Proposal distribution: $\mathscr{N}$(x,1)

Trace Plots



(a)

(b)

(c)

(d)

Summary of MCMC

- Unbiased
  - If we simulate MCMC for large enough number of steps we can get arbitrarily accurate approximation of the target distribution

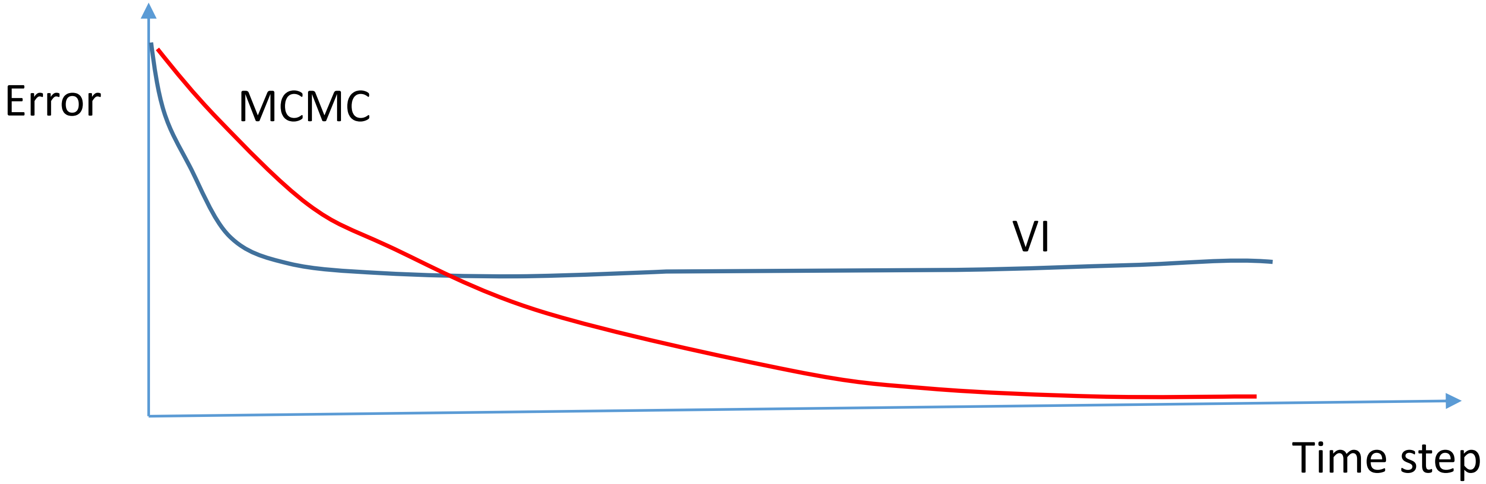
- May be slow
  - Long burn-in
  - Slow exploration

- Parameter choice
  - Gibbs: No parameter tuning
  - Metropolis-Hastings: Proposal distribution has to be chosen

- Sensitive to initialization

Comparison between VI, MCMC

Variational Inference	MCMC
Biased	Unbiased
Fast	Slow
Choice of approximations (Mean Field)	Choice of proposal distributions


Error
MCMC
VI
Time step