# Bayesian Theory and Computation

# Lecture 4: Markov Chain Monte Carlo I


**Cheng Zhang**

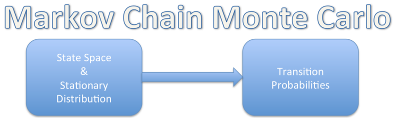
School of Mathematical Sciences, Peking University


March 04, 2024

# Markov chain Monte Carlo

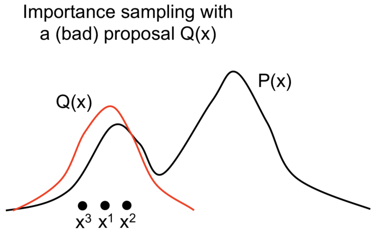


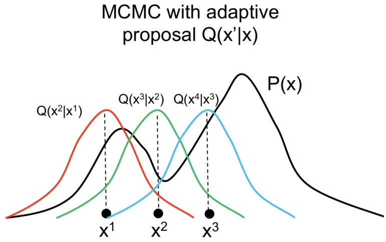
- Now suppose we are interested in sampling from a distribution $\pi$ (e.g., the unnormalized posterior)
- Markov chain Monte Carlo (MCMC) is a method that samples from a Markov chain whose stationary distribution is the target distribution $\pi$. It does this by constructing an appropriate transition probability for $\pi$
- MCMC, therefore, can be viewed as an inverse process of Markov chains

# Markov chain Monte Carlo

- The transition probability in MCMC resembles the proposal distribution we used in previous Monte Carlo methods.
- Instead of using a fixed proposal (as in importance sampling and rejection sampling), MCMC algorithms feature adaptive proposals




Figures adapted from Eric Xing (CMU)

# The Metropolis Algorithm



- Suppose that we are interested in sampling from a distribution $\pi$, whose density we know up to a constant $P(x) \propto \pi(x)$
- We can construct a Markov chain with a transition probability (i.e., proposal distribution) $Q(x'|x)$ which is symmetric; that is, $Q(x'|x) = Q(x|x')$
- Example. A normal distribution with the mean at the current state and fixed variance $\sigma^2$ is symmetric since

$$\exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) = \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)$$

# The Metropolis Algorithm



In each iteration we do the following

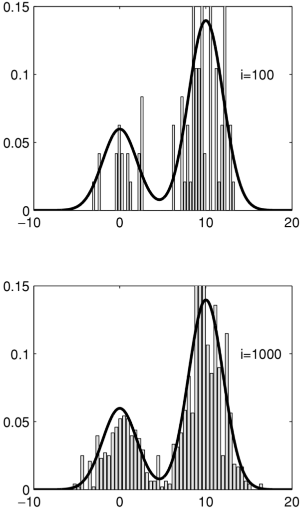
- Draws a sample $x'$ from $Q(x'|x)$, where $x$ is the previous sample
- Calculated the acceptance probability

$$a(x'|x) = \min\left(1, \frac{P(x')}{P(x)}\right)$$

  Note that we only need to compute $\frac{P(x')}{P(x)}$, the unknown constant cancels out
- Accept the new sample with probability $a(x'|x)$ or remain at state $x$. The acceptance probability ensures that, after sufficient many draws, our samples will come from the true distribution $\pi(x)$

# Example: Gaussian Mixture Model






Adapted from Andrieu, Freitas, Doucet, Jordan, 2003

- How do we know that the chain is going to converge to $\pi$?
- Suppose the support of the proposal distribution is $\mathcal{X}$ (e.g., Gaussian distribution), then the Markov chain is irreducible and aperiodic.
- We only need to verify the detailed balance condition

$$\begin{aligned}
\pi(dx)p(x, dx') &= \pi(x)dx \cdot Q(x'|x)a(x'|x)dx' \\
&= \pi(x)Q(x'|x) \min\left(1, \frac{\pi(x')}{\pi(x)}\right) dxdx' \\
&= Q(x'|x) \min(\pi(x), \pi(x'))dxdx' \\
&= Q(x|x') \min(\pi(x'), \pi(x))dxdx' \\
&= \pi(x')dx' \cdot Q(x|x') \min\left(1, \frac{\pi(x)}{\pi(x')}\right) dx \\
&= \pi(dx')p(x', dx)
\end{aligned}$$

- It turned out that symmetric proposal distribution is not necessary. Hastings (1970) later on generalized the above algorithm using the following acceptance probability for general $Q(x'|x)$

$$a(x'|x) = \min\left(1, \frac{P(x')Q(x|x')}{P(x)Q(x'|x)}\right)$$

- Similarly, we can show that detailed balanced condition is preserved

# Proposal Distribution



- Under mild assumptions on the proposal distribution $Q$, the algorithm is ergodic
- However, the choice of $Q$ is important since it determines the speed of convergence to $\pi$ and the efficiency of sampling
- Usually, the proposal distribution depend on the current state. But it can be independent of current state, which leads to an independent MCMC sampler that is somewhat like a rejection/importance sampling method
- Some examples of commonly used proposal distributions
  - $Q(x'|x) \sim \mathcal{N}(x, \sigma^2)$
  - $Q(x'|x) \sim \text{Uniform}(x - \delta, x + \delta)$
- Finding a good proposal distribution is hard in general

- Recall the univariate Gaussian model with known variance

$$y_i \sim \mathcal{N}(\theta, \sigma^2)$$
$$p(y|\theta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - \theta)^2}{2\sigma^2}\right)$$

- Note that there is a conjugate $\mathcal{N}(\mu_0, \tau_0^2)$ prior for $\theta$, and the posterior has a close form normal distribution
- Now let's pretend that we don't know this exact posterior distribution and use a Markov chain to sample from it.

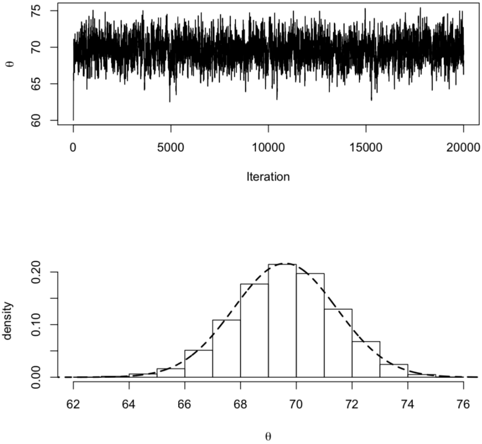
# Examples: Gaussian Model with Known Variance

- We can of course write the posterior distribution up to a constant

$$p(\theta|y) \propto \exp\left(\frac{(\theta-\mu_0)^2}{2\tau_0^2}\right)\prod_{i=1}^{n}\exp\left(-\frac{(y_i-\theta)^2}{2\sigma^2}\right) = P(\theta)$$

- We use $\mathcal{N}(\theta^{(i)}, 1)$, a normal distribution around our current state, to propose the next step
- Starting from an initial point $\theta^{(0)}$ and propose the next step $\theta' \sim \mathcal{N}(\theta^{(0)}, 1)$, we either accept this value with probability $a(\theta'|\theta^{(0)})$ or reject and stay where we are
- We continue these steps for many iterations

- As we can see, the posterior distribution we obtained using the Metropolis algorithm is very similar to the exact posterior



# Examples: Binomial Model with Beta Prior



- Recall the binomial model:

$$p(y|n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

- Assuming the conjugate prior $\text{Beta}(\alpha, \beta)$ for $\theta$, we saw that the posterior is $\text{Beta}(\alpha + y, \beta + n - y)$.
- For the election example, we mentioned that out of 100 people surveyed, 39 said they are going to vote for $A$. We used a conjugate $\text{Beta}(1, 1)$ prior and obtained $\text{Beta}(40, 62)$ as the posterior distribution for $\theta$.
- Now let's not use the closed form of the posterior distribution and use the Metropolis algorithm instead.

- We first need to find the posterior distribution (up to a constant).
- The prior distribution is of course uniform: $p(\theta) = 1$.
- The likelihood is (ignore the irrelevant constant)

$$p(y|\theta) \propto \theta^y (1-\theta)^{n-y}$$

  where $n = 100$ and $y = 39$.
- Therefore, using the Bayes' theorem, the posterior is

$$p(\theta|y) \propto p(\theta)p(y|\theta) \propto \theta^{39}(1-\theta)^{61} = P(\theta)$$

- Next, we need to choose a transition (i.e., proposal) distribution.
- Let's use Uniform$(0, 1)$. This is of course symmetric.
- Now we start from $x_0 = 0.5$ and repeat the following steps
  - sample $\theta'$ from Uniform$(0, 1)$
  - calculate the acceptance probability

$$a(\theta'|\theta^{(i)}) = \min\left(1, \frac{(\theta')^{39}(1-\theta')^{61}}{(\theta^{(i)})^{39}(1-\theta^{(i)})^{61}}\right)$$

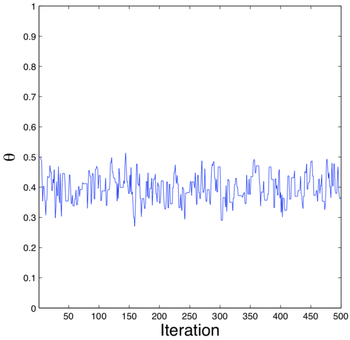
  - Accept the proposed value with probability $a(\theta'|\theta)$. For this, we can sample $u \sim$ Uniform$(0, 1)$ and set

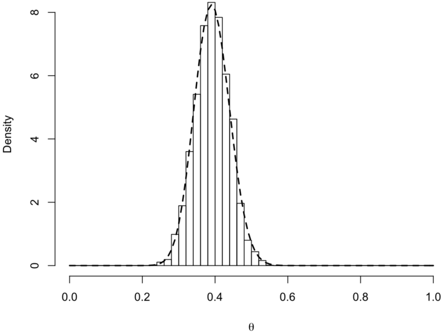
$$\theta^{(i+1)} = \begin{cases} \theta' & u < a(\theta'|\theta) \\ \theta^{(i)} & \text{otherwise} \end{cases}$$

# Examples: Binomial Model with Beta Prior



Trace plot and posterior estimation



- Recall the Beckham's example. We modeled the number of goals $y_i$ he scores in a game using a Poisson model

$$y_i \sim \text{Poisson}(\theta)$$

- He scored 0 and 1 goals in the first two games respectively
- We used $\text{Gamma}(1.4, 10)$ prior for $\theta$, and because of conjugacy, the posterior distribution also had a Gamma distribution

$$\theta|y \sim \text{Gamma}(2.4, 12)$$

- Again, let's ignore the closed form posterior and use MCMC for sampling the posterior distribution

# Examples: Poisson Model with Gamma Prior

- The prior is
$$p(\theta) \propto \theta^{0.4} \exp(-10\theta)$$
- The likelihood is
$$p(y|\theta) \propto \theta^{y_1+y_2} \exp(-2\theta)$$
where $y_1 = 0$ and $y_2 = 1$
- Therefore, the posterior is proportional to
$$p(\theta|y) \propto \theta^{0.4} \exp(-10\theta) \cdot \theta^{y_1+y_2} \exp(-2\theta) = P(\theta)$$

- Symmetric proposal distributions such as

$$\text{Uniform}(\theta^{(i)} - \delta, \theta^{(i)} + \delta) \text{ or } \mathcal{N}(\theta^{(i)}, \sigma^2)$$

  might not be efficient since they do not take the non-negative support of the posterior into account.
- Here, we use a non-symmetric proposal distribution such as $\text{Uniform}(0, \theta^{(i)} + \delta)$ and use the Metropolis-Hastings (MH) algorithm instead
- We set $\delta = 1$

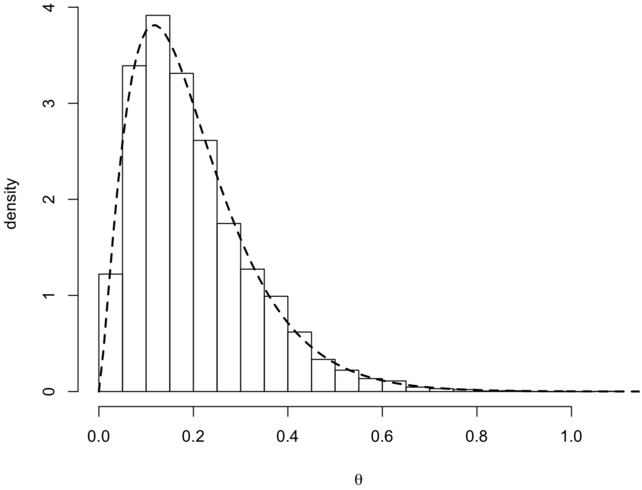
We start from $\theta_0 = 1$ and follow these steps in each iteration

- Sample $\theta'$ from $\mathcal{U}(0, \theta^{(i)} + 1)$
- Calculate the acceptance probability

$$a(\theta'|\theta^{(i)}) = \min\left(1, \frac{P(\theta')\text{Uniform}(\theta^{(i)}|0, \theta' + 1)}{P(\theta^{(i)})\text{Uniform}(\theta'|0, \theta^{(i)} + 1)}\right)$$

- Sample $u \sim \mathcal{U}(0, 1)$ and set

$$\theta^{(i+1)} = \begin{cases} \theta' & u < a(\theta'|\theta^{(i)}) \\ \theta^{(i)} & \text{otherwise} \end{cases}$$


density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
θ

- What if the distribution is multidimensional, *i.e.*, $x = (x_1, x_2, \ldots, x_d)$
- We can still use the Metropolis algorithm (or MH), with a multivariate proposal distribution, *i.e.*, we now propose $x' = (x'_1, x'_2, \ldots, x'_d)$
- For example, we can use a multivariate normal $\mathcal{N}_d(x, \sigma^2 I)$, or a $d$-dimensional uniform distribution around the current state

# Examples: Banana Shape Distribution

- Here we construct a banana-shaped posterior distribution as follows

$$y|\theta \sim \mathcal{N}(\theta_1 + \theta_2^2, \sigma_y^2), \quad \sigma_y = 2$$

We generate data $y_i \sim \mathcal{N}(1, \sigma_y^2)$

- We use a bivariate normal prior for $\theta$

$$\theta = (\theta_1, \theta_2) \sim \mathcal{N}(0, I)$$

- The posterior is

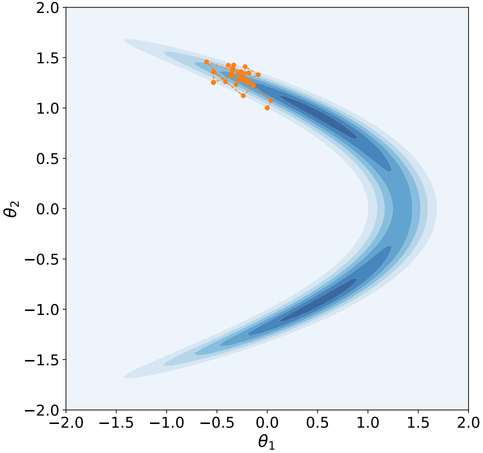
$$p(\theta|y) \propto \exp\left(-\frac{\theta_1^2 + \theta_2^2}{2}\right) \cdot \exp\left(-\frac{\sum_i (y_i - \theta_1 - \theta_2^2)^2}{2\sigma_y^2}\right)$$

- We use the Metropolis algorithm to sample from posterior, with a bivariate normal proposal distribution such as $\mathcal{N}(\theta^{(i)}, (0.15)^2 I)$

# Examples: Banana Shape Distribution



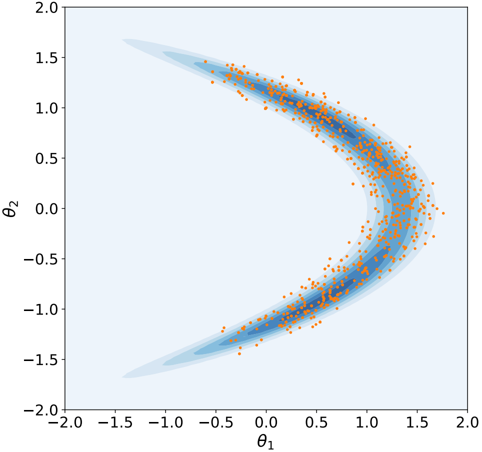
The first few samples from the posterior distribution of $\theta = (\theta_1, \theta_2)$, using a bivariate normal proposal



# Examples: Banana Shape Distribution

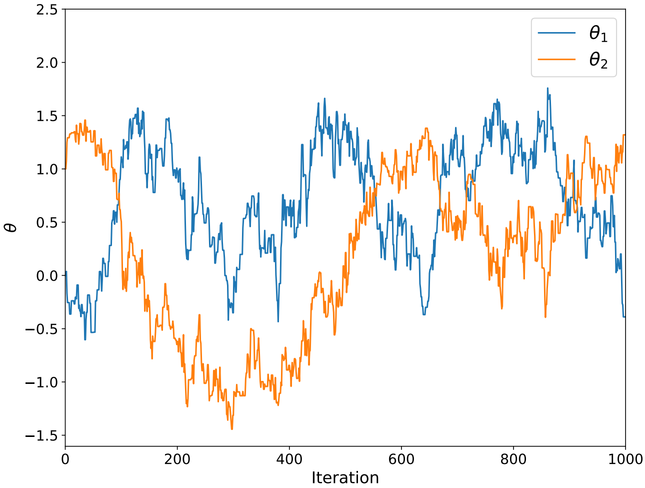


Posterior samples for $\theta = (\theta_1, \theta_2)$



# Examples: Banana Shape Distribution

Trace plot of posterior samples for $\theta = (\theta_1, \theta_2)$



- Sometimes, it is easier to decompose the parameter space into several components, and use the Metropolis (or MH) algorithm for one component at a time
- At iteration $i$, given the current state $(x_1^{(i)}, \ldots, x_d^{(i)})$, we do the following for all components $k = 1, 2, \ldots, d$
  - Sample $x'_k$ from the univariate proposal distribution $Q(x'_k | \ldots, x_{k-1}^{(i+1)}, x_k^{(i)}, \ldots)$
  - Accept this new value and set $x_k^{(i+1)} = x'_k$ with probability

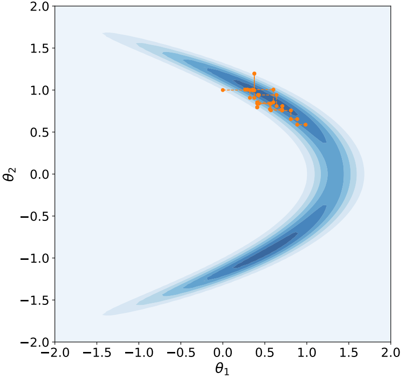
$$a(x'_k | \ldots, x_{k-1}^{(i+1)}, x_k^{(i)}, \ldots)) = \min\left(1, \frac{P(\ldots, x_{k-1}^{(i+1)}, x'_k, \ldots)}{P(\ldots, x_{k-1}^{(i+1)}, x_k^{(i)}, \ldots)}\right)$$

or reject it and set $x_k^{(i+1)} = x_k^{(i)}$

# Decomposing the Parameter Space

- Note that in general, we can decompose the space of random variable into blocks of components
- Also, we can update the components sequentially or randomly
- As long as each transition probability individually leaves the target distribution invariant, their sequence would leave the target distribution invariant
- In Bayesian models, this is especially useful if it is easier and computationally less intensive to evaluate the posterior distribution when one subset of parameters change at a time

- In the example of banana-shaped distribution, we can sample $\theta_1$ and $\theta_2$ one at a time
- The first few samples from the posterior distribution of $\theta = (\theta_1, \theta_2)$, using a univariate normal proposal sequentially



- As the dimensionality of the parameter space increases, it becomes difficult to find an appropriate proposal distributions (e.g., with appropriate step size) for the Metropolis (or MH) algorithm
- If we are lucky (in some situations we are!), the conditional distribution of one component, $x_j$, given all other components, $x_{-j}$ is tractable and has a close form so that we can sample from it directly
- If that's the case, we can sample from each component one at a time using their corresponding conditional distributions $P(x_j|x_{-j})$

# The Gibbs Sampler



- This is known as the Gibbs sampler (GS) or "heat bath" (Geman and Geman, 1984)
- Note that in Bayesian analysis, we are mainly interested in sampling from $p(\theta|y)$
- Therefore, we use the Gibbs sampler when $P(\theta_j|y, \theta_{-j})$ has a closed form, e.g., there is a conditional conjugacy
- One example is the univariate normal model. As we will see later, given $\sigma$, the posterior $P(\mu|y, \sigma^2)$ has a closed form, and given $\mu$, the posterior distribution of $P(\sigma^2|\mu, y)$ also has a closed form

- The Gibbs sampler works as follows
- Initialize starting value for $x_1, x_2, \ldots, x_d$
- At each iteration, pick an ordering of the $d$ variables (can be sequential or random)
    1. Sample $x \sim P(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$, *i.e.*, the conditional distribution of $x_i$ given the current values of all other variables
    2. Update $x_i \leftarrow x$
- When we update $x_i$, we immediately use it new value for sampling other variables $x_j$

# GS is A Special Case of MH

- Note that in GS, we are not proposing anymore, we are directly sampling, which can be viewed as a proposal that will always be accepted
- This way, the Gibbs sampler can be viewed as a special case of MH, whose proposal is

$$Q(x_i', x_{-i}|x_i, x_{-i}) = P(x_i'|x_{-i})$$

- Applying MH with this proposal, we obtain

$$a(x_i', x_{-i}|x_i, x_{-i}) = \min\left(1, \frac{P(x_i', x_{-i})Q(x_i, x_{-i}|x_i', x_{-i})}{P(x_i, x_{-i})Q(x_i', x_{-i}|x_i, x_{-i})}\right)$$

$$= \min\left(1, \frac{P(x_i', x_{-i})P(x_i|x_{-i})}{P(x_i, x_{-i})P(x_i'|x_{-i})}\right) = \min\left(1, \frac{P(x_i', x_{-i})P(x_i, x_{-i})}{P(x_i, x_{-i})P(x_i', x_{-i})}\right)$$

$$= 1$$

- We can now use the Gibbs sampler to simulate samples from the posterior distribution of the parameters of a univariate normal $y \sim \mathcal{N}(\mu, \sigma^2)$ model, with prior

$$\mu \sim \mathcal{N}(\mu_0, \tau_0^2), \quad \sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

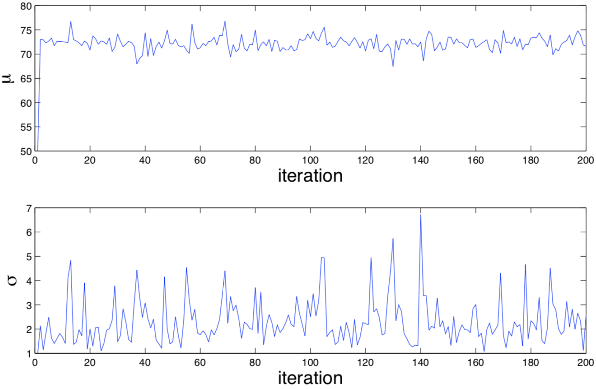
- Given $(\sigma^{(i)})^2$ at the $i^{\text{th}}$ iteration, we sample $\mu^{(i+1)}$ from

$$\mu^{(i+1)} \sim \mathcal{N}\left(\frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{(\sigma^{(i)})^2}}{\frac{1}{\tau_0^2} + \frac{n}{(\sigma^{(i)})^2}}, \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{(\sigma^{(i)})^2}}\right)$$

- Given $\mu^{(i+1)}$, we sample a new $\sigma^2$ from

$$(\sigma^{(i+1)})^2 \sim \text{Inv-}\chi^2(\nu_0 + n, \frac{\nu_0\sigma_0^2 + \nu n}{\nu_0 + n}), \quad \nu = \frac{1}{n}\sum_{j=1}^{n}(y_j - \mu^{(i+1)})^2$$

- The following graphs show the trace plots of the posterior samples (for both $\mu$ and $\sigma$)



Gibbs sampling algorithms have been widely used in **probabilistic graphical models**

- Conditional distributions are fairly easy to derive for many graphical models (e.g., mixture models, Latent Dirichlet allocation)
- Have reasonable computation and memory requirements, only needs to sample one random variable at a time
- Can be Rao-Blackwellized (integrate out some random variable) to decrease the sampling variance. This is called *collapsed Gibbs sampling*
- We will see examples later.

# Combining Metropolis with Gibbs



- For more complex models, we might only have conditional conjugacy for one part of the parameters
- In such situations, we can combine the Gibbs sampler with the Metropolis method
- That is, we update the components with conditional conjugacy using Gibbs sampler and for the rest parameters, we use the Metropolis (or MH)

# References




- Metropolis, N. (1953). Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21, 1087–1092.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.
- Andrieu, C., De Freitas, N., Doucet, A. and Jordan, M. I. (2003). An introduction to MCMC for machine learning. Machine learning 50, 5–43.