Markov Chain Monte Carlo methods

Tomas McKelvey and Lennart Svensson

Signal Processing Group
Department of Signals and Systems
Chalmers University of Technology, Sweden

November 26, 2012

Today's learning objectives

After this lecture you should be able to

- Explain what Monte Carlo methods are and why they are important.
- Describe how importance sampling works.
- Use the Metropolis-Hastings algorithm to sample from a general distribution.
- Explain what Gibbs sampling is and when it is useful.

Bayesian Inference

- Data: a realization x
- Parameters, latent variables: $\theta = [\theta_1, \theta_2, \ldots, \theta_p]$
- Likelihood: $L(x|\theta)$
- Inference based on the *joint posterior*

$$\begin{aligned}\pi(\theta|x) &= \frac{L(x|\theta)\pi_0(\theta)}{\int L(x|\theta)\pi_0(\theta)\, d\theta} \\ &\propto L(x|\theta)\pi_0(\theta) \\ \text{Posterior} &\propto \text{Likelihood} \times \text{Prior}\end{aligned}$$

Making decisions

Decisions are made by **minimizing posterior expected loss**.

Two examples:

1. Estimation:

$$\hat{\theta} = \mathbb{E}\{\theta|x\} = \int \theta\pi(\theta|x)\, d\theta$$

2. Model selection among two candidates m_1 and m_2:

$$\int f(x|\theta_1, m_1)\pi(\theta_1|m_1)\, d\theta_1 \underset{<}{\overset{>}{}} \int f(x|\theta_2, m_2)\pi(\theta_2|m_2)\, d\theta_2$$

m_1 above, m_2 below

Decisions by solving integrals

- In general, decisions are often made by computing integrals

$$I = \int g(\theta)\pi(\theta|x)\, d\theta$$

Today's problem formulation

Given: a distribution $p(\theta)$, known up to a normalization constant, and function, $g(\theta)$.
Objective: find

$$I = \int g(\theta)p(\theta)\, d\theta.$$

Analytical solutions

- If conjugate prior, posterior $p(\theta) = \pi(\theta|x)$ has nice analytical expression
  Examples: Beta-Binomial, Gauss-Gauss, Gamma-Poisson, etc.

- For some functions $g(\theta)$ the integral then has a closed form expression $\Rightarrow$ we are done!

- **Limitations:** rarely ever possible in problems of practical interest.
  More common: conjugate priors are used but solutions hard due to nuisance parameters

Deterministic approximations

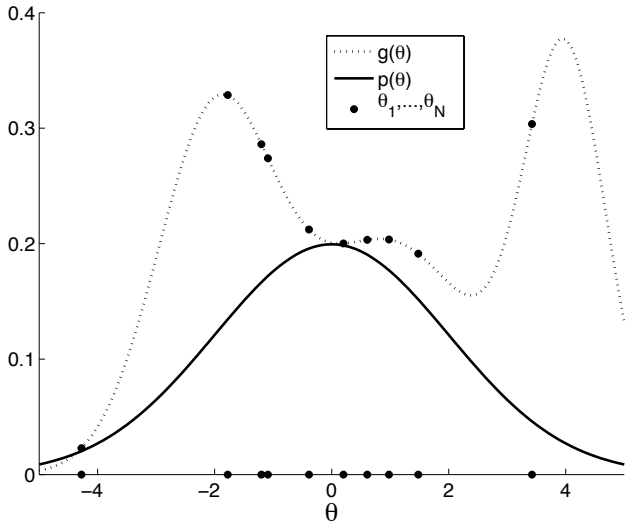
- There are many techniques to approximate such integrals (or to perform inference) deterministically.
- **Some examples:**
  - quadrature methods (related to unscented transform)
  - Laplace approximation (assumes unimodality)
  - message passing (uses conditional independence)
  - variational Bayes (approximate independence)
- Very important and useful techniques!
- **Weakness:** limited accuracy. In particular when approximations do not match model structure.

Monte Carlo methods

Method

- Generate independent and identically distributed (iid) samples $\theta_1, \ldots, \theta_N$ from $p(\theta)$. Approximate

$$I = \int g(\theta)p(\theta)\, d\theta \approx \frac{1}{N}\sum_{i=1}^{N} g(\theta_i).$$



- Law of large numbers: $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} g(\theta_i) = I$.

Monte Carlo methods – properties

- $\hat{I} = \frac{1}{N} \sum_{i=1}^{N} g(\theta_i)$ is an *unbiased estimate* of I

$$\mathbb{E}\left\{ \frac{1}{N} \sum_{i=1}^{N} g(\theta_i) \right\} = I$$

- Error covariance is

$$\mathrm{Cov}\{\hat{I}\} = \mathbb{E}\left\{ (\hat{I} - I)(\hat{I} - I)^T \right\} = \ldots$$

$$= ?$$

$\rightsquigarrow \ldots$, **independently of dimensionality** of θ!

- **Difficulty:** how can we generate $\theta_1, \ldots, \theta_N$?

Monte Carlo methods – properties

- $\hat{I} = \frac{1}{N}\sum_{i=1}^{N} g(\theta_i)$ is an *unbiased estimate* of I

$$\mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N} g(\theta_i)\right\} = I$$

- Error covariance is

$$\begin{aligned}\mathrm{Cov}\{\hat{I}\} &= \mathbb{E}\left\{(\hat{I} - I)(\hat{I} - I)^T\right\} = \ldots \\ &= \frac{1}{N}\mathrm{Cov}\{g(\theta)\}\end{aligned}$$

  $\rightsquigarrow$ vanishes as $1/N$, **independently of dimensionality** of θ!

- **Difficulty:** how can we generate $\theta_1, \ldots, \theta_N$?

Monte Carlo methods – feasible?

- Usually $p(\theta)$ is not a simple distribution, like Gaussian or Beta
  $\Rightarrow$ we cannot use built-in random generators directly!
- In fact, $p(\theta)$ is often the posterior

$$p(\theta) \propto \pi(\theta) f(x|\theta)$$

  with *unknown normalization*.
- Is it still possible to use the Monte Carlo idea?

Importance sampling – basic idea

Importance sampling

- Generate $\theta_1, \ldots, \theta_N$ from a proposal distribution $q(\theta)$ and use approximation

$$I = \int g(\theta) \frac{p(\theta)}{q(\theta)} q(\theta)\, d\theta \approx \frac{1}{N} \sum_{i=1}^{N} g(\theta_i) \frac{p(\theta_i)}{q(\theta_i)}$$

- Often, weights $\tilde{w}_i = \frac{p(\theta_i)}{q(\theta_i)N}$ contain unknown normalization, therefore replaced by

$$w_i = \frac{\tilde{w}_i}{\sum_{n=1}^{N} \tilde{w}_n}.$$

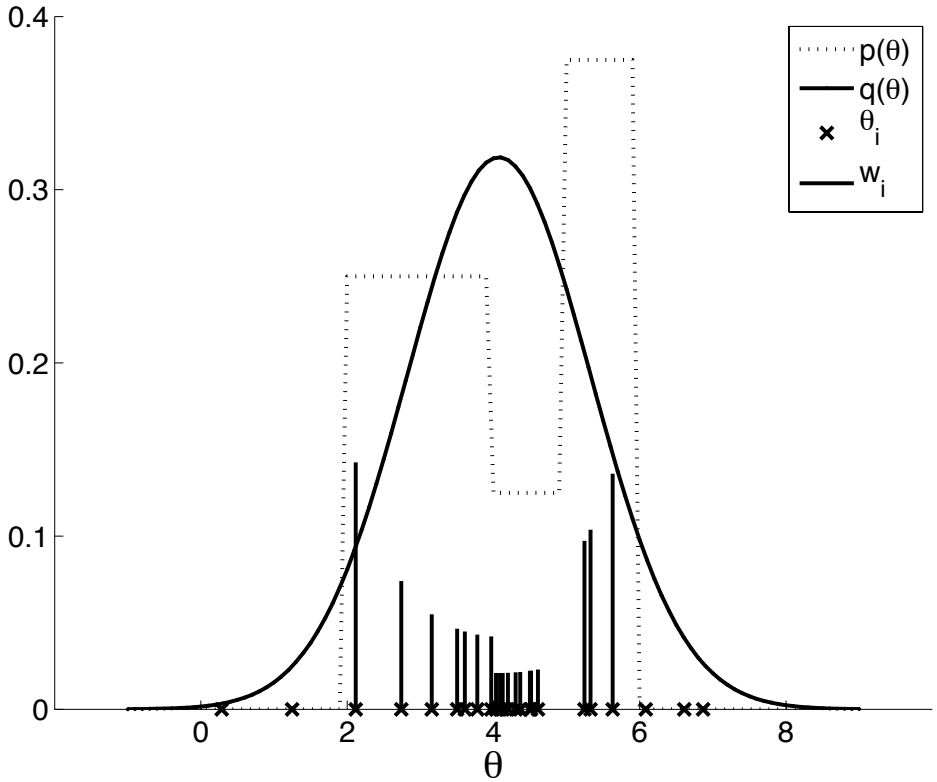
- Now

$$I \approx \sum_{i=1}^{N} w_i g(\theta_i)$$

with no need to know normalization of $p(\theta)$.

Importance sampling – an illustration

- An illustration of samples and weights



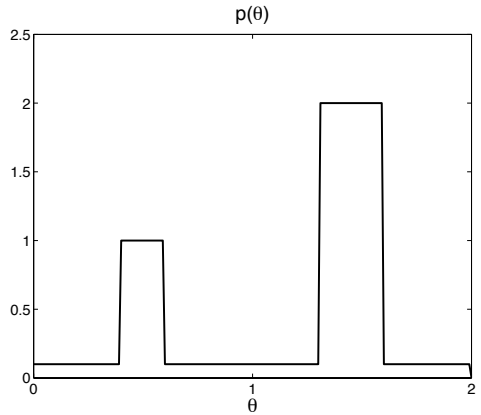
- Note that some samples are given zero weight.

Importance sampling – remarks

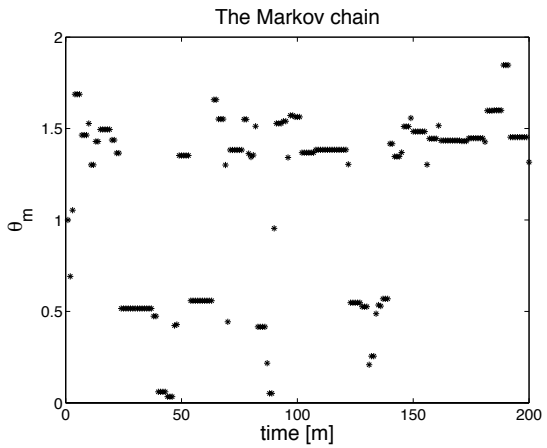
- Importance sampling "works" as long as support of $q(\theta)$ includes support of $p(\theta)$.
- Proposal should also be easy to generate samples from and similar to $p(\theta)$.
- Generates independent samples that can be used for Monte Carlo integration.
- Works **very well** if **proposal selected carefully**.
- Normally suffers from **curse of dimensionality**
  $\rightsquigarrow$ proposal mismatch grows quickly and makes most weights zero!

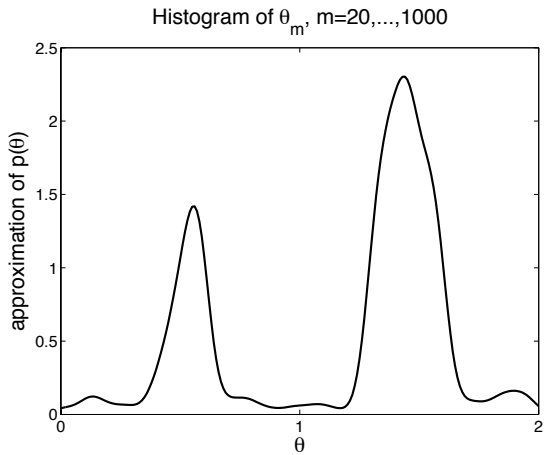
The MCMC concept

- Can Markov chains be used to perform Monte Carlo sampling?
- Suppose that we wish to generate samples from



- Samples can be obtained using a carefully designed Markov chain!





Markov chains and kernels

- A Markov chain is a sequence of random variables, $\theta_1, \theta_2, \ldots,$ that can be thought of as evolving over time.
- A key property is the conditional independence (Markov property):

$$f(\theta_m|\theta_{m-1}, \theta_{m-2}, \ldots, \theta_1) = f(\theta_m|\theta_{m-1})$$

- The distribution $f(\theta_m|\theta_{m-1})$ is called the **transition kernel**

$$K(\theta_m|\theta_{m-1}) = f(\theta_m|\theta_{m-1}).$$

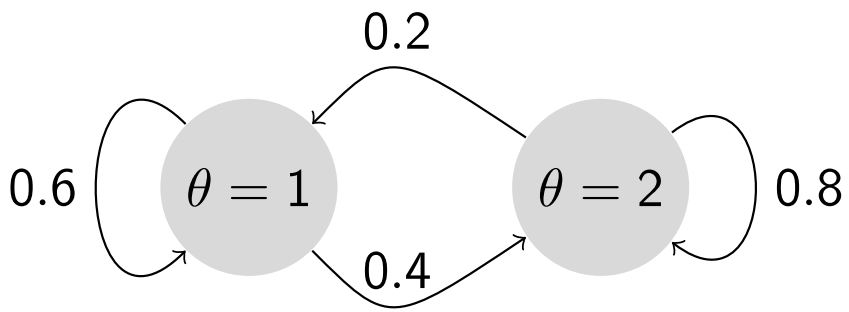
$$\Rightarrow \begin{cases} 1 = \int K(\theta_m|\theta_{m-1})\, d\theta_m \\ f(\theta_m) = \int K(\theta_m|\theta_{m-1}) f(\theta_{m-1})\, d\theta_{m-1} \end{cases}$$

Stationary distributions

- Many Markov chains converge to a **stationary distribution** over time
  $\rightsquigarrow \theta_{m-1}$ and θ_m (almost) identically distributed for large m.
- $\pi(\theta)$ is a stationary distribution if

$$\pi(\theta_m) = \int K(\theta_m | \theta_{m-1}) \pi(\theta_{m-1}) \, d\theta_{m-1}$$

- **Example:** compute $\pi_1 = \pi(\theta = 1)$ and $\pi_2 = \pi(\theta = 2)$.

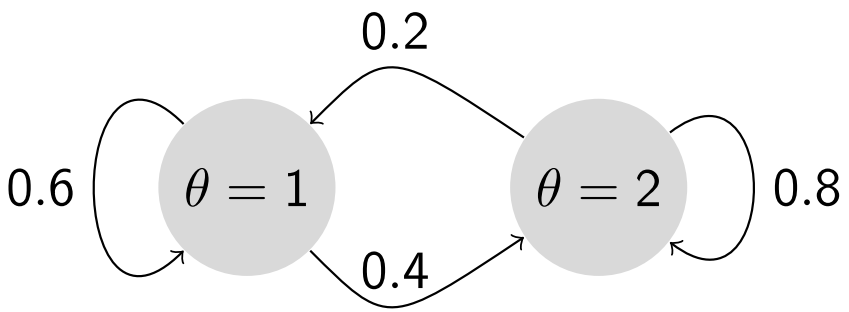


Stationary distributions

- Many Markov chains converge to a **stationary distribution** over time
  $\rightsquigarrow \theta_{m-1}$ and θ_m (almost) identically distributed for large m.
- $\pi(\theta)$ is a stationary distribution if

$$\pi(\theta_m) = \int K(\theta_m | \theta_{m-1}) \pi(\theta_{m-1}) \, d\theta_{m-1}$$

- **Example:** compute $\pi_1 = \pi(\theta = 1)$ and $\pi_2 = \pi(\theta = 2)$.



$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix}$$

Reversible chains

- A Markov chain is reversible if $\exists \pi(\theta)$:

$$\pi(\theta')K(\theta|\theta') = \pi(\theta)K(\theta'|\theta) \quad (1)$$

  $\rightsquigarrow$ probability of beeing at θ and passing to θ' and from beeing at θ' and passing to θ are the same.
- **Example:** verify that the above chain is reversible.

$$p = pP$$

- **Note:** (1) is called **detailed balance condition** and implies that $\pi(\theta)$ is a stationary distribution!

  Proof: integrate both sides with respect to θ

$$\int \pi(\theta')K(\theta|\theta')\, d\theta = \int \pi(\theta)K(\theta'|\theta)\, d\theta$$

  where lhs is simply $\pi(\theta')$. The condition for stationarity!

Reversible chains

- A Markov chain is reversible if $\exists \pi(\theta)$:

$$\pi(\theta')K(\theta|\theta') = \pi(\theta)K(\theta'|\theta) \quad (1)$$

  $\rightsquigarrow$ probability of beeing at θ and passing to θ' and from beeing at θ' and passing to θ are the same.
- **Example:** verify that the above chain is reversible.

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix} \qquad p = pP$$

- **Note:** (1) is called **detailed balance condition** and implies that $\pi(\theta)$ is a stationary distribution!

  Proof: integrate both sides with respect to θ

$$\int \pi(\theta')K(\theta|\theta')\, d\theta = \int \pi(\theta)K(\theta'|\theta)\, d\theta$$

  where lhs is simply $\pi(\theta')$. The condition for stationarity!

Reversible chains

- A Markov chain is reversible if $\exists \pi(\theta)$:

$$\pi(\theta')K(\theta|\theta') = \pi(\theta)K(\theta'|\theta) \quad (1)$$

  $\rightsquigarrow$ probability of beeing at θ and passing to θ' and from beeing at θ' and passing to θ are the same.
- **Example:** verify that the above chain is reversible.

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix} \quad P^{\infty} = \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{pmatrix} \quad p = pP$$

- Note: (1) is called **detailed balance condition** and implies that $\pi(\theta)$ is a stationary distribution!

  Proof: integrate both sides with respect to θ

$$\int \pi(\theta')K(\theta|\theta')\, d\theta = \int \pi(\theta)K(\theta'|\theta)\, d\theta$$

  where lhs is simply $\pi(\theta')$. The condition for stationarity!

Reversible chains

- A Markov chain is reversible if $\exists \pi(\theta)$:

$$\pi(\theta')K(\theta|\theta') = \pi(\theta)K(\theta'|\theta) \quad (1)$$

  $\rightsquigarrow$ probability of beeing at θ and passing to θ' and from beeing at θ' and passing to θ are the same.
- **Example:** verify that the above chain is reversible.

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{pmatrix} \quad P^{\infty} = \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{pmatrix} \quad p = pP$$

  $1/3 * 0.4 = 2/3 * 0.2$
- **Note:** (1) is called **detailed balance condition** and implies that $\pi(\theta)$ is a stationary distribution!

  Proof: integrate both sides with respect to θ

$$\int \pi(\theta')K(\theta|\theta')\, d\theta = \int \pi(\theta)K(\theta'|\theta)\, d\theta$$

  where lhs is simply $\pi(\theta')$. The condition for stationarity!

Metropolis-Hastings (MH) algorithm

- An MCMC algorithm which is flexible and very simple to use.
- A technique to design a **reversible Markov chain** with **stationary distribution** $\pi(\theta) = p(\theta)$.
- MH has two components
  1. **Proposal distribution** $q(\theta' \big| \theta)$: suggests a move from θ to θ'. Often simply $\theta' = \theta + n$ where $n \sim \mathcal{N}(0, I)$.
  2. **Acceptance probability** $a(\theta' \big| \theta)$: the proposed state is accepted with probability $a(\theta' \big| \theta)$
     $\rightsquigarrow$ $a(\theta' \big| \theta)$ is selected to ensure that the detailed balance conditioned is satisfied for $\pi(\theta) = p(\theta)$.

Metropolis-Hastings (MH) algorithm

Summary of the MH algorithm

- Start with an arbitrary initial value θ_0.
- Update from θ_m to θ_{m+1} $(m = 0, 1, \ldots, N)$ by
  1. Generate $\xi \sim q(\xi|\theta_m)$.
  2. Take

$$\theta_{m+1} = \begin{cases} \xi & \text{with probability } a(\xi|\theta_m) \\ \theta_m & \text{otherwise.} \end{cases}$$

- Given a realization of the chain we do Monte Carlo sampling:

$$\frac{1}{N - N_b + 1} \sum_{m=N_b}^{N} g(\theta_m)$$

where N_b is when we hope that the chain has reaches its stationary distribution (burn-in time)

The acceptance probability

- We wish to select $a(\theta' \big| \theta)$ such that

$$p(\theta')K(\theta \big| \theta') = p(\theta)K(\theta' \big| \theta)$$

  $\rightsquigarrow$ ensures that $p(\theta)$ is a stationary distribution.

- What is the transition kernel, $K(\theta' \big| \theta)$ (when $\theta' \neq \theta$)?

$$K(\theta' \big| \theta) = ?$$

- How can we select $a(\theta' \big| \theta)$ to satisfy the above detailed balance condition?

The acceptance probability

- We wish to select $a(\theta' \big| \theta)$ such that

$$p(\theta')K(\theta \big| \theta') = p(\theta)K(\theta' \big| \theta)$$

  $\rightsquigarrow$ ensures that $p(\theta)$ is a stationary distribution.

- What is the transition kernel, $K(\theta' \big| \theta)$ (when $\theta' \neq \theta$)?

$$K(\theta' \big| \theta) = q(\theta' \big| \theta)a(\theta' \big| \theta).$$

- How can we select $a(\theta' \big| \theta)$ to satisfy the above detailed balance condition?

The acceptance probability

- We wish to select $a(\theta' \big| \theta)$ such that

$$p(\theta')K(\theta \big| \theta') = p(\theta)K(\theta' \big| \theta)$$

  $\rightsquigarrow$ ensures that $p(\theta)$ is a stationary distribution.

- What is the transition kernel, $K(\theta' \big| \theta)$ (when $\theta' \neq \theta$)?

$$K(\theta' \big| \theta) = q(\theta' \big| \theta)a(\theta' \big| \theta).$$

- How can we select $a(\theta' \big| \theta)$ to satisfy the above detailed balance condition?

$$p(\theta')q(\theta \big| \theta')a(\theta \big| \theta') = p(\theta)q(\theta' \big| \theta)a(\theta' \big| \theta)$$
$$\Leftrightarrow \frac{a(\theta \big| \theta')}{a(\theta' \big| \theta)} = \frac{p(\theta)q(\theta' \big| \theta)}{p(\theta')q(\theta \big| \theta')}$$

Metropolis-Hastings (MH) algorithm

Summary of the MH algorithm

- Start with an arbitrary initial value θ_0.
- Update from θ_m to θ_{m+1} $(m = 0, 1, \ldots, N)$ by
  1. Generate $\xi \sim q(\xi|\theta_m)$.
  2. Take

$$\theta_{m+1} = \begin{cases} \xi & \text{with probability } \min\left\{1, \frac{p(\xi)q(\theta_m|\xi)}{p(\theta_m)q(\xi|\theta_m)}\right\} \\ \theta_m & \text{otherwise.} \end{cases}$$

- **Note:** algorithm involves point-wise evaluation of $p(\xi)/p(\theta_m)$
  $\rightsquigarrow$ possible even when the normalization constant is unknown!

The proposal distribution $q(\theta'|\theta)$

- We want to select $q(\theta|\theta')$ such that the chain *"explores the space quickly"*.
  If it does, it is said to "mix" well.
- The most common choice for q is the *random walk proposal*

$$q(\theta'|\theta) = f(\|\theta' - \theta\|)$$
$$\Leftrightarrow \xi = \theta_m + v,$$

  where v is a symmetric random variable.
- For this choice, the acceptance probability simplifies to

$$a(\xi|\theta_m) = \min\left\{1, \frac{p(\xi)}{p(\theta_m)}\right\}$$

  $\rightsquigarrow$ always accept proposals that move upwards.

Metropolis-Hastings – summary

- **Pros:**
  - The MH algorithm can be applied to virtually any problem.
  - We can collect all our samples from a single run of the Markov chain.
  - The freedom in the choice of proposal gives us the possibility to design a chain that mixes quickly.
- **Cons:**
  - We need to select the proposal density, $q(\theta|\theta')$.
  - A poor choice will give us samples that are highly correlated and do not represent the full distribution.
  - It is also difficult (impossible?) to know when the chain has reached the stationary distribution. Use convergence diagnostics!

Gibbs sampling – general algorithm

- Gibbs sampling is an MCMC algorithm that make use of the model structure.
- **Idea:** sample one variable (dimension) at a time conditioned on all the others.

The Gibbs sampler

- Start with arbitrary initial vector $\theta_0 = [\theta_0(1), \theta_0(2), \ldots, \theta_0(d)]^T$.
- For $m = 0, 1, \ldots, N$ generate

1) $\theta_{m+1}(1) \sim p(\theta(1) \big| \theta_m(2), \ldots, \theta_m(d))$
2) $\theta_{m+1}(2) \sim p(\theta(2) \big| \theta_{m+1}(1), \theta_m(3), \ldots, \theta_m(d))$

$\ldots$

d) $\theta_{m+1}(d) \sim p(\theta(d) \big| \theta_{m+1}(1), \ldots, \theta_{m+1}(d-1))$

Gibbs sampling – toy example

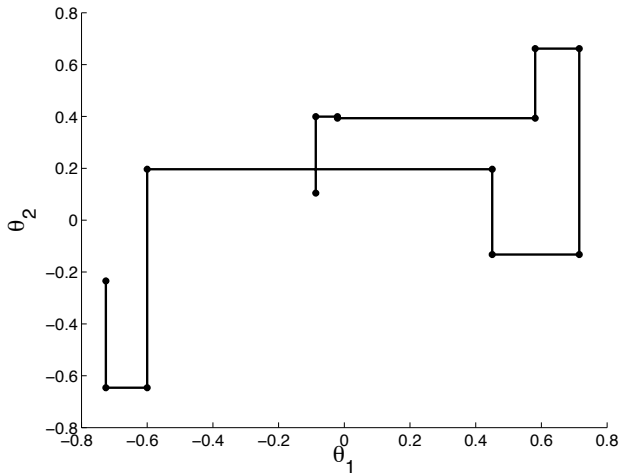
- The Gibbs sampler for the model

$$\begin{bmatrix}\theta(1)\\ \theta(2)\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$$

  is

  1) $\theta_{m+1}(1)\big|\theta_m(2) \sim \mathcal{N}(\rho\theta_m(2), 1-\rho^2)$
  2) $\theta_{m+1}(2)\big|\theta_{m+1}(1) \sim \mathcal{N}(\rho\theta_{m+1}(1), 1-\rho^2)$

– For $\rho = 0.7$ it can look like this:



Gibbs sampling vs Hierarchical models

- Gibbs sampling is particularly useful for Hierachical models.

- To sample $\theta \sim \pi(\theta|x)$, we can generate

$$(\theta, \phi) \sim \pi(\theta, \phi|x)$$

using Gibbs sampling:

$$\theta_{m+1} \sim \pi(\theta|\phi_m, x)$$
$$\phi_{m+1} \sim \pi(\phi|\theta_{m+1})$$

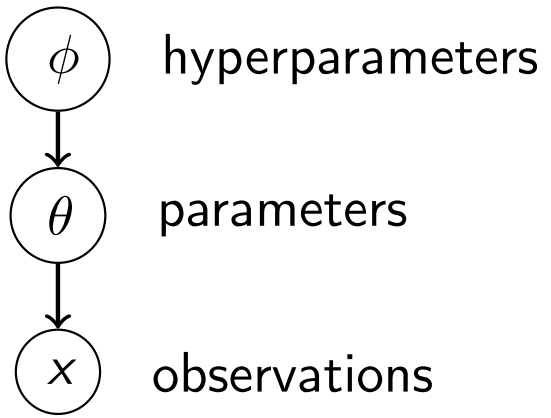


Figure: A very small hierachical model.

- While generating ϕ we can ignore x due to conditional independence.
  $\rightsquigarrow$ this benefit is substantially larger in bigger networks.

Gibbs sampling – summary

- **Pros:**
  - Few design choices makes it simple to use.
  - Utilizes the model structure.
  - Generates high dimensional variables using a sequence of low dimensional simulations.
- **Cons:**
  - Mixes poorly if variables are highly correlated.
  - Requires knowledge of $p(\theta)$.
  - Only applicable if conditionals are easy to simulate.

Other sampling methods

There are many other sampling methods that we could not cover today:

- Rejection sampling – a classical method to generate independent samples from a distribution.
- Slice sampler – a more general type of Gibbs sampler.
- Population Monte Carlo – a combination of MH and particle filters.
- Hamiltonian (or Hybrid) Monte Carlo – introduces additional latent variables that enables large steps in the state space.
- Adaptive MCMC – methods to adaptively improve the proposal distribution based on what we learn from the Markov chain.
- Reversible Jump MCMC – an extension to situations where the dimensionality of θ is unknown.

Today's learning objectives

After this lecture you should be able to

- Explain what Monte Carlo methods are and why they are important.
- Describe how importance sampling works.
- Use the Metropolis-Hastings algorithm to sample from a general distribution.
- Explain what Gibbs sampling is and when it is useful.

References

Some further readings:

- Gelfand, A.E. and Smith, A.F.M., "Sampling Based Approaches to Calculating Marginal Densities", *Journal of the American Statistical Association*, Vol. 85, pp. 398-409, 1990.
- Casella G. and George E.I. "Explaining the Gibbs Sampler", *The American Statistician*, Vol. 46 No. 3, pp. 167-174, 1992
- Hastings W.K. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika*, Vol. 57, No. 1, pp. 97-109, 1970
- Gilks W.R., Richardson S. and Spiegelhalter D.J., *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, 1996.