Approximate Inference: Sampling Methods (2)

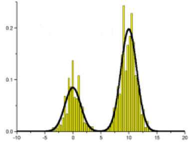
Piyush Rai

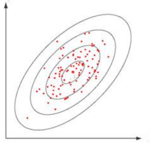
Probabilistic Machine Learning (CS772A)

Oct 3, 2017

Sampling Methods: Recap

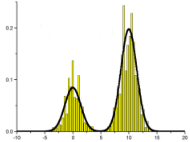
- Any probability distribution $p(\boldsymbol{z})$ can be (approximately) represented using a set of samples

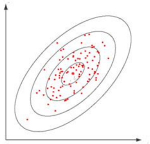




Sampling Methods: Recap

- Any probability distribution $p(\boldsymbol{z})$ can be (approximately) represented using a set of samples

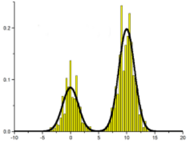


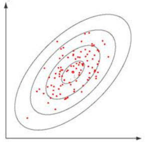


- Samples can come from $p(\boldsymbol{z})$ or some "proposal distribution" if $p(\boldsymbol{z})$ is a "difficult" distribution

Sampling Methods: Recap

- Any probability distribution $p(\boldsymbol{z})$ can be (approximately) represented using a set of samples





- Samples can come from $p(\boldsymbol{z})$ or some "proposal distribution" if $p(\boldsymbol{z})$ is a "difficult" distribution
- Given a set of samples $\{\boldsymbol{z}^{(\ell)}\}_{\ell=1}^{L}$, the sample-based approximation of $p(\boldsymbol{z})$ can be written as

$$p(\boldsymbol{z}) \approx \frac{1}{L}\sum_{\ell=1}^{L}\delta(\boldsymbol{z}=\boldsymbol{z}^{(\ell)}) \quad \text{or} \quad p(\boldsymbol{z}) \approx \frac{1}{L}\sum_{\ell=1}^{L}\delta_{\boldsymbol{z}^{(\ell)}}(\boldsymbol{z})$$

Sampling Methods: Recap

- Looked at some basic methods for generating samples from a probability distribution
- Also, using the samples to approximate difficult to compute expectations

Sampling Methods: Recap

- Looked at some basic methods for generating samples from a probability distribution
  - Transformation based methods
  - Rejection sampling
- Also, using the samples to approximate difficult to compute expectations

Sampling Methods: Recap

- Looked at some basic methods for generating samples from a probability distribution
  - Transformation based methods
  - Rejection sampling
- Also, using the samples to approximate difficult to compute expectations

Monte Carlo Sampling: $$\mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz \approx \frac{1}{L}\sum_{\ell=1}^{L} f(z^{(\ell)}) \quad \text{where} \quad \{z^{(\ell)}\}_{\ell=1}^{L} \sim p(z)$$

Sampling Methods: Recap

- Looked at some basic methods for generating samples from a probability distribution
  - Transformation based methods
  - Rejection sampling
- Also, using the samples to approximate difficult to compute expectations

$$\text{Monte Carlo Sampling:} \quad \mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz \approx \frac{1}{L}\sum_{\ell=1}^{L} f(z^{(\ell)}) \quad \text{where} \quad \{z^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

$$\text{Important Sampling (1):} \quad \mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz \approx \frac{1}{L}\sum_{\ell=1}^{L} f(z^{(\ell)})\frac{p(\boldsymbol{z}^{(\ell)})}{q(\boldsymbol{z}^{(\ell)})} \quad \text{where} \quad \{z^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$$

Sampling Methods: Recap

- Looked at some basic methods for generating samples from a probability distribution
  - Transformation based methods
  - Rejection sampling
- Also, using the samples to approximate difficult to compute expectations

$$\text{Monte Carlo Sampling:} \quad \mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz \approx \frac{1}{L}\sum_{\ell=1}^{L} f(z^{(\ell)}) \quad \text{where} \quad \{z^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

$$\text{Important Sampling (1):} \quad \mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz \approx \frac{1}{L}\sum_{\ell=1}^{L} f(z^{(\ell)})\frac{p(\boldsymbol{z}^{(\ell)})}{q(\boldsymbol{z}^{(\ell)})} \quad \text{where} \quad \{z^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$$

$$\text{Important Sampling (2):} \quad \mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})} \quad \text{where} \quad \{z^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$$

Sampling Methods: Recap

- Looked at some basic methods for generating samples from a probability distribution
  - Transformation based methods
  - Rejection sampling
- Also, using the samples to approximate difficult to compute expectations

$$\text{Monte Carlo Sampling:} \quad \mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz \approx \frac{1}{L}\sum_{\ell=1}^{L} f(z^{(\ell)}) \quad \text{where} \quad \{z^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{z})$$

$$\text{Important Sampling (1):} \quad \mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz \approx \frac{1}{L}\sum_{\ell=1}^{L} f(z^{(\ell)})\frac{p(\boldsymbol{z}^{(\ell)})}{q(\boldsymbol{z}^{(\ell)})} \quad \text{where} \quad \{z^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$$

$$\text{Important Sampling (2):} \quad \mathbb{E}_{p(z)}[f(z)] = \int f(z)p(z)dz \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{\ell=1}^{L} f(\boldsymbol{z}^{(\ell)})\frac{\tilde{p}(\boldsymbol{z}^{(\ell)})}{\tilde{q}(\boldsymbol{z}^{(\ell)})} \quad \text{where} \quad \{z^{(\ell)}\}_{\ell=1}^{L} \sim q(\boldsymbol{z})$$

[Note: I.S. (1) assumes $p(\boldsymbol{z})$ can be <u>evaluated</u> at any $\boldsymbol{z}$, I.S. (2) assumes $p(\boldsymbol{z}) = \frac{\tilde{p}(z)}{Z_p}$ can only be evaluated up to a prop. constant]

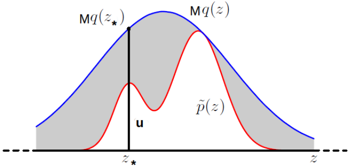
Limitations of Basic Sampling Methods

Limitations of Basic Sampling Methods

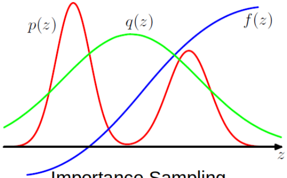
- Transformation based methods: Usually limited to drawing from standard distributions

Limitations of Basic Sampling Methods

- Transformation based methods: Usually limited to drawing from standard distributions
- Rejection Sampling and Importance Sampling: Require good proposal distributions



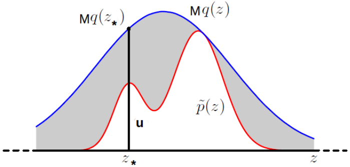
Rejection Sampling



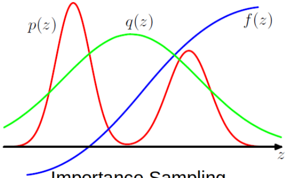
Importance Sampling

Limitations of Basic Sampling Methods

- Transformation based methods: Usually limited to drawing from standard distributions
- Rejection Sampling and Importance Sampling: Require good proposal distributions



Rejection Sampling

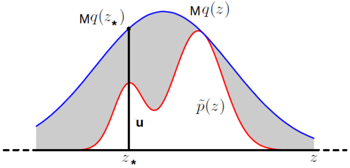


Importance Sampling

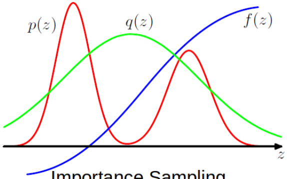
- Difficult to find good prop. distr. especially when $\boldsymbol{z}$ is high-dim. (e.g., models with many params)

Limitations of Basic Sampling Methods

- Transformation based methods: Usually limited to drawing from standard distributions
- Rejection Sampling and Importance Sampling: Require good proposal distributions



Rejection Sampling

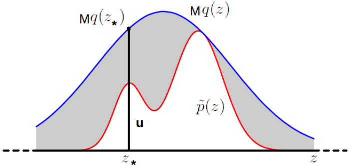


Importance Sampling

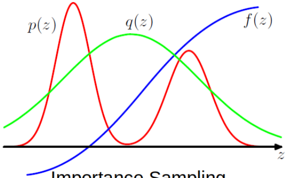
- Difficult to find good prop. distr. especially when $\boldsymbol{z}$ is high-dim. (e.g., models with many params)
  - In high dimensions, most of the mass of $p(\boldsymbol{z})$ is concentrated in a tiny region of the $\boldsymbol{z}$ space

Limitations of Basic Sampling Methods

- Transformation based methods: Usually limited to drawing from standard distributions
- Rejection Sampling and Importance Sampling: Require good proposal distributions



Rejection Sampling



Importance Sampling

- Difficult to find good prop. distr. especially when $\boldsymbol{z}$ is high-dim. (e.g., models with many params)
  - In high dimensions, most of the mass of $p(\boldsymbol{z})$ is concentrated in a tiny region of the $\boldsymbol{z}$ space
  - Difficult to *a priori* know what those regions are, thus difficult to come up with good proposal dist.

Markov Chain Monte Carlo (MCMC) Methods

Markov Chain Monte Carlo (MCMC)

- Goal: Generate samples from some target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$, where $\boldsymbol{z}$ is high-dimensional

Markov Chain Monte Carlo (MCMC)

- Goal: Generate samples from some target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$, where $\boldsymbol{z}$ is high-dimensional
- Will again assume that we can evaluate $p(\boldsymbol{z})$ at least up to a proportionality constant

Markov Chain Monte Carlo (MCMC)

- Goal: Generate samples from some target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$, where $\boldsymbol{z}$ is high-dimensional
- Will again assume that we can evaluate $p(\boldsymbol{z})$ at least up to a proportionality constant
- Basic idea in MCMC: Use a Markov Chain to generate samples from $p(\boldsymbol{z})$

$$\boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \ldots \rightarrow \boldsymbol{z}^{(L)}$$

Markov Chain Monte Carlo (MCMC)

- Goal: Generate samples from some target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$, where $\boldsymbol{z}$ is high-dimensional
- Will again assume that we can evaluate $p(\boldsymbol{z})$ at least up to a proportionality constant
- Basic idea in MCMC: Use a Markov Chain to generate samples from $p(\boldsymbol{z})$

$$\boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \ldots \rightarrow \boldsymbol{z}^{(L)}$$

- How: Given a current sample $\boldsymbol{z}^{(\ell)}$ from the chain, generate the next sample $\boldsymbol{z}^{(\ell+1)}$ as

Markov Chain Monte Carlo (MCMC)

- Goal: Generate samples from some target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$, where $\boldsymbol{z}$ is high-dimensional
- Will again assume that we can evaluate $p(\boldsymbol{z})$ at least up to a proportionality constant
- Basic idea in MCMC: Use a Markov Chain to generate samples from $p(\boldsymbol{z})$

$$\boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \ldots \rightarrow \boldsymbol{z}^{(L)}$$

- How: Given a current sample $\boldsymbol{z}^{(\ell)}$ from the chain, generate the next sample $\boldsymbol{z}^{(\ell+1)}$ as
  - Use a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\ell)})$ to generate a candidate sample $\boldsymbol{z}^*$

Markov Chain Monte Carlo (MCMC)

- Goal: Generate samples from some target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$, where $\boldsymbol{z}$ is high-dimensional
- Will again assume that we can evaluate $p(\boldsymbol{z})$ at least up to a proportionality constant
- Basic idea in MCMC: Use a Markov Chain to generate samples from $p(\boldsymbol{z})$

$$\boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \ldots \rightarrow \boldsymbol{z}^{(L)}$$

- How: Given a current sample $\boldsymbol{z}^{(\ell)}$ from the chain, generate the next sample $\boldsymbol{z}^{(\ell+1)}$ as
  - Use a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\ell)})$ to generate a candidate sample $\boldsymbol{z}^*$
  - Accept/reject $\boldsymbol{z}^*$ as the next sample based on an acceptance criterion (will see later)

Markov Chain Monte Carlo (MCMC)

- Goal: Generate samples from some target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$, where $\boldsymbol{z}$ is high-dimensional
- Will again assume that we can evaluate $p(\boldsymbol{z})$ at least up to a proportionality constant
- Basic idea in MCMC: Use a Markov Chain to generate samples from $p(\boldsymbol{z})$

$$\boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \ldots \rightarrow \boldsymbol{z}^{(L)}$$

- How: Given a current sample $\boldsymbol{z}^{(\ell)}$ from the chain, generate the next sample $\boldsymbol{z}^{(\ell+1)}$ as
  - Use a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\ell)})$ to generate a candidate sample $\boldsymbol{z}^*$
  - Accept/reject $\boldsymbol{z}^*$ as the next sample based on an acceptance criterion (will see later)
  - If accepted, $\boldsymbol{z}^{(\ell+1)} = \boldsymbol{z}^*$. If rejected, $\boldsymbol{z}^{(\ell+1)} = \boldsymbol{z}^{(\ell)}$

Markov Chain Monte Carlo (MCMC)

- Goal: Generate samples from some target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$, where $\boldsymbol{z}$ is high-dimensional
- Will again assume that we can evaluate $p(\boldsymbol{z})$ at least up to a proportionality constant
- Basic idea in MCMC: Use a Markov Chain to generate samples from $p(\boldsymbol{z})$

$$\boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \ldots \rightarrow \boldsymbol{z}^{(L)}$$

- How: Given a current sample $\boldsymbol{z}^{(\ell)}$ from the chain, generate the next sample $\boldsymbol{z}^{(\ell+1)}$ as
  - Use a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\ell)})$ to generate a candidate sample $\boldsymbol{z}^*$
  - Accept/reject $\boldsymbol{z}^*$ as the next sample based on an acceptance criterion (will see later)
  - If accepted, $\boldsymbol{z}^{(\ell+1)} = \boldsymbol{z}^*$. If rejected, $\boldsymbol{z}^{(\ell+1)} = \boldsymbol{z}^{(\ell)}$
- If $q(\boldsymbol{z}|\boldsymbol{z}^{(\ell)})$ has certain properties, the Markov chain's stationary distribution will be $p(\boldsymbol{z})$

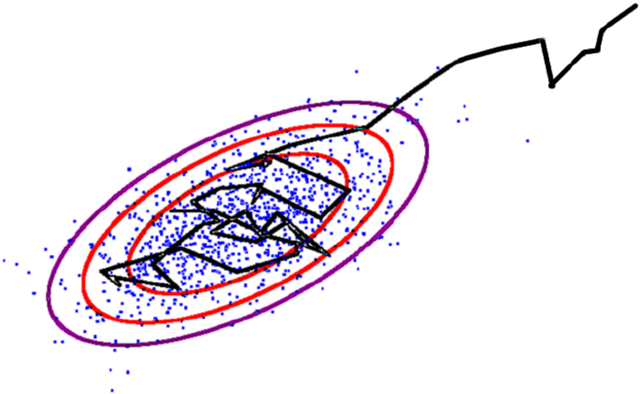
Markov Chain Monte Carlo (MCMC)

- Goal: Generate samples from some target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$, where $\boldsymbol{z}$ is high-dimensional
- Will again assume that we can evaluate $p(\boldsymbol{z})$ at least up to a proportionality constant
- Basic idea in MCMC: Use a Markov Chain to generate samples from $p(\boldsymbol{z})$

$$\boldsymbol{z}^{(1)} \rightarrow \boldsymbol{z}^{(2)} \rightarrow \ldots \rightarrow \boldsymbol{z}^{(L)}$$

- How: Given a current sample $\boldsymbol{z}^{(\ell)}$ from the chain, generate the next sample $\boldsymbol{z}^{(\ell+1)}$ as
  - Use a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\ell)})$ to generate a candidate sample $\boldsymbol{z}^*$
  - Accept/reject $\boldsymbol{z}^*$ as the next sample based on an acceptance criterion (will see later)
  - If accepted, $\boldsymbol{z}^{(\ell+1)} = \boldsymbol{z}^*$. If rejected, $\boldsymbol{z}^{(\ell+1)} = \boldsymbol{z}^{(\ell)}$
- If $q(\boldsymbol{z}|\boldsymbol{z}^{(\ell)})$ has certain properties, the Markov chain's stationary distribution will be $p(\boldsymbol{z})$
  - Informally, stationary distribution means where the chain will eventually "reach"

Markov Chain Monte Carlo (MCMC)



Markov Chain

- Consider a sequence of random variables $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$

Markov Chain

- Consider a sequence of random variables $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$
- A first-order Markov Chain assumes

$$p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(\ell)}) = p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)}) \qquad \forall \ell$$

Markov Chain

- Consider a sequence of random variables $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$
- A first-order Markov Chain assumes

$$p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(\ell)}) = p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)}) \qquad \forall \ell$$

- A first order Markov chain can be defined by the following

Markov Chain

- Consider a sequence of random variables $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$
- A first-order Markov Chain assumes

$$p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(\ell)}) = p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)}) \qquad \forall \ell$$

- A first order Markov chain can be defined by the following
  - An initial state distribution $p(\boldsymbol{z}^{(0)})$

Markov Chain

- Consider a sequence of random variables $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$
- A first-order Markov Chain assumes

$$p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(\ell)}) = p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)}) \qquad \forall \ell$$

- A first order Markov chain can be defined by the following
  - An initial state distribution $p(\boldsymbol{z}^{(0)})$
  - Transition probabilities $T_\ell(\boldsymbol{z}^{(\ell)}, \boldsymbol{z}^{(\ell+1)}) = p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$

Markov Chain

- Consider a sequence of random variables $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$
- A first-order Markov Chain assumes

$$p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(\ell)}) = p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)}) \qquad \forall \ell$$

- A first order Markov chain can be defined by the following
  - An initial state distribution $p(\boldsymbol{z}^{(0)})$
  - Transition probabilities $T_\ell(\boldsymbol{z}^{(\ell)}, \boldsymbol{z}^{(\ell+1)}) = p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$: Distribution over the possible values of $\boldsymbol{z}^{(\ell+1)}$

Markov Chain

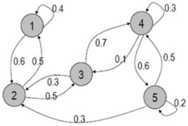
- Consider a sequence of random variables $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$
- A first-order Markov Chain assumes

$$p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(\ell)}) = p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)}) \qquad \forall \ell$$

- A first order Markov chain can be defined by the following
  - An initial state distribution $p(\boldsymbol{z}^{(0)})$
  - Transition probabilities $T_\ell(\boldsymbol{z}^{(\ell)}, \boldsymbol{z}^{(\ell+1)}) = p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$: Distribution over the possible values of $\boldsymbol{z}^{(\ell+1)}$

Transition probabilities can be defined using a *KxK* table if **z** is a discrete r.v. with *K* possible values

$$\begin{array}{c c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0.4 & 0.6 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.7 & 0.0 \\ 0.0 & 0.0 & 0.1 & 0.3 & 0.6 \\ 0.0 & 0.3 & 0.0 & 0.5 & 0.2 \end{bmatrix} \end{array}$$



Markov Chain

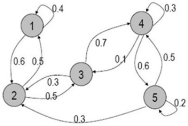
- Consider a sequence of random variables $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(L)}$
- A first-order Markov Chain assumes

$$p(\boldsymbol{z}^{(\ell+1)} | \boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(\ell)}) = p(\boldsymbol{z}^{(\ell+1)} | \boldsymbol{z}^{(\ell)}) \qquad \forall \ell$$

- A first order Markov chain can be defined by the following
  - An initial state distribution $p(\boldsymbol{z}^{(0)})$
  - Transition probabilities $T_\ell(\boldsymbol{z}^{(\ell)}, \boldsymbol{z}^{(\ell+1)}) = p(\boldsymbol{z}^{(\ell+1)} | \boldsymbol{z}^{(\ell)})$: Distribution over the possible values of $\boldsymbol{z}^{(\ell+1)}$

Transition probabilities can be defined using a *KxK* table if **z** is a discrete r.v. with *K* possible values

$$\begin{array}{c|c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0.4 & 0.6 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.7 & 0.0 \\ 0.0 & 0.0 & 0.1 & 0.3 & 0.6 \\ 0.0 & 0.3 & 0.0 & 0.5 & 0.2 \end{bmatrix} \end{array}$$



- Homogeneous Markov Chain: Transition probabilities $T_\ell = T$ (same everywhere along the chain)

Some Properties

- Consider the discrete case when $\boldsymbol{z}$ has K possible states (and T is a matrix of size $K \times K$)

Some Properties

- Consider the discrete case when $\boldsymbol{z}$ has K possible states (and T is a matrix of size $K \times K$)
- Assume the graph representing the possible state transitions is irreducible and aperiodic

Some Properties

- Consider the discrete case when $\boldsymbol{z}$ has K possible states (and T is a matrix of size $K \times K$)
- Assume the graph representing the possible state transitions is irreducible and aperiodic
- Ergodic Property: Under the above assumption, for any choice of an initial probability vector $\boldsymbol{v}$

$$\boldsymbol{v}T^m = \boldsymbol{p}^* \qquad \text{as} \quad m \to \infty$$

where the probability vector $\boldsymbol{p}^*$ represents the invariant or stationary distribution of the chain

Some Properties

- Consider the discrete case when $\boldsymbol{z}$ has K possible states (and T is a matrix of size $K \times K$)
- Assume the graph representing the possible state transitions is irreducible and aperiodic
- Ergodic Property: Under the above assumption, for any choice of an initial probability vector $\boldsymbol{v}$

$$\boldsymbol{v}T^m = \boldsymbol{p}^* \qquad \text{as} \quad m \to \infty$$

  where the probability vector $\boldsymbol{p}^*$ represents the invariant or stationary distribution of the chain
- Why do we need the graph to be irreducible and aperiodic?

Some Properties

- Consider the discrete case when $\boldsymbol{z}$ has K possible states (and T is a matrix of size $K \times K$)
- Assume the graph representing the possible state transitions is irreducible and aperiodic
- Ergodic Property: Under the above assumption, for any choice of an initial probability vector $\boldsymbol{v}$

$$\boldsymbol{v} T^m = \boldsymbol{p}^* \qquad \text{as} \quad m \to \infty$$

  where the probability vector $\boldsymbol{p}^*$ represents the invariant or stationary distribution of the chain
- Why do we need the graph to be irreducible and aperiodic?
  - Irreducible: No disjoint sets of nodes. Can reach from any state to any state

Some Properties

- Consider the discrete case when $\boldsymbol{z}$ has K possible states (and T is a matrix of size $K \times K$)
- Assume the graph representing the possible state transitions is irreducible and aperiodic
- Ergodic Property: Under the above assumption, for any choice of an initial probability vector $\boldsymbol{v}$

$$\boldsymbol{v}T^m = \boldsymbol{p}^* \qquad \text{as} \quad m \to \infty$$

  where the probability vector $\boldsymbol{p}^*$ represents the invariant or stationary distribution of the chain
- Why do we need the graph to be irreducible and aperiodic?
  - Irreducible: No disjoint sets of nodes. Can reach from any state to any state
  - Aperiodic: No cycles in the graph (otherwise would oscillate forever). Consider this example

$$\boldsymbol{v} = [1/5, 4/5] \qquad T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

  .. multiplying $\boldsymbol{v}$ by T repeatedly leads to oscillating values without convergence

Some Properties

- Note that, for Ergodic case, $\boldsymbol{p}^* T = \boldsymbol{p}^*$

Some Properties

- Note that, for Ergodic case, $\boldsymbol{p}^* T = \boldsymbol{p}^*$. Therefore $\boldsymbol{\pi}^*$ is the left eigenvector of T with eigenvalue 1

Some Properties

- Note that, for Ergodic case, $\boldsymbol{p}^* T = \boldsymbol{p}^*$. Therefore $\boldsymbol{\pi}^*$ is the left eigenvector of T with eigenvalue 1
- For the discrete-valued $\boldsymbol{z}$ with $K = 5$ possible states, $\boldsymbol{p}^* = [p_1^*, \ldots, p_5^*]$, and we can write

$$\sum_{i=1}^{5} p_i^* T_{ij} = p_j^*$$

Some Properties

- Note that, for Ergodic case, $\boldsymbol{p}^* T = \boldsymbol{p}^*$. Therefore $\boldsymbol{\pi}^*$ is the left eigenvector of T with eigenvalue 1
- For the discrete-valued $\boldsymbol{z}$ with $K = 5$ possible states, $\boldsymbol{p}^* = [p_1^*, \ldots, p_5^*]$, and we can write

$$\sum_{i=1}^{5} p_i^* T_{ij} = p_j^*$$

- For the continuous $\boldsymbol{z}$ case, we can equivalently write (for any two state values $\boldsymbol{z}$ and $\boldsymbol{z}'$)

$$\int p^*(\boldsymbol{z}') T(\boldsymbol{z}', \boldsymbol{z}) d\boldsymbol{z}' = p^*(\boldsymbol{z})$$

Some Properties

- Note that, for Ergodic case, $\boldsymbol{p}^* T = \boldsymbol{p}^*$. Therefore $\boldsymbol{\pi}^*$ is the left eigenvector of T with eigenvalue 1
- For the discrete-valued $\boldsymbol{z}$ with $K = 5$ possible states, $\boldsymbol{p}^* = [p_1^*, \ldots, p_5^*]$, and we can write

$$\sum_{i=1}^{5} p_i^* T_{ij} = p_j^*$$

- For the continuous $\boldsymbol{z}$ case, we can equivalently write (for any two state values $\boldsymbol{z}$ and $\boldsymbol{z}'$)

$$\int p^*(\boldsymbol{z}') T(\boldsymbol{z}', \boldsymbol{z}) d\boldsymbol{z}' = p^*(\boldsymbol{z})$$

- Suppose a Markov chain with transition probabilities T satisfies

$$p^*(\boldsymbol{z}) T(\boldsymbol{z}, \boldsymbol{z}') = p^*(\boldsymbol{z}') T(\boldsymbol{z}', \boldsymbol{z}) \quad \text{(Detailed Balance)}$$

Some Properties

- Note that, for Ergodic case, $\boldsymbol{p}^* T = \boldsymbol{p}^*$. Therefore $\boldsymbol{\pi}^*$ is the left eigenvector of T with eigenvalue 1
- For the discrete-valued $\boldsymbol{z}$ with $K = 5$ possible states, $\boldsymbol{p}^* = [p_1^*, \ldots, p_5^*]$, and we can write

$$\sum_{i=1}^{5} p_i^* T_{ij} = p_j^*$$

- For the continuous $\boldsymbol{z}$ case, we can equivalently write (for any two state values $\boldsymbol{z}$ and $\boldsymbol{z}'$)

$$\int p^*(\boldsymbol{z}')T(\boldsymbol{z}', \boldsymbol{z})d\boldsymbol{z}' = p^*(\boldsymbol{z})$$

- Suppose a Markov chain with transition probabilities T satisfies

$$p^*(\boldsymbol{z})T(\boldsymbol{z}, \boldsymbol{z}') = p^*(\boldsymbol{z}')T(\boldsymbol{z}', \boldsymbol{z}) \quad \text{(Detailed Balance)}$$

- Integrating both sides w.r.t. $\boldsymbol{z}'$ gives $\int p^*(\boldsymbol{z}')T(\boldsymbol{z}', \boldsymbol{z})d\boldsymbol{z}' = p^*(\boldsymbol{z})$ (i.e., the Ergodic property)
  - Thus a Markov chain with detailed balance will always converge to a stationary distribution $p^*(\boldsymbol{z})$

Some Properties

- Note that, for Ergodic case, $\boldsymbol{p}^* T = \boldsymbol{p}^*$. Therefore $\boldsymbol{\pi}^*$ is the left eigenvector of T with eigenvalue 1
- For the discrete-valued $\boldsymbol{z}$ with $K = 5$ possible states, $\boldsymbol{p}^* = [p_1^*, \ldots, p_5^*]$, and we can write

$$\sum_{i=1}^{5} p_i^* T_{ij} = p_j^*$$

- For the continuous $\boldsymbol{z}$ case, we can equivalently write (for any two state values $\boldsymbol{z}$ and $\boldsymbol{z}'$)

$$\int p^*(\boldsymbol{z}')T(\boldsymbol{z}', \boldsymbol{z})d\boldsymbol{z}' = p^*(\boldsymbol{z})$$

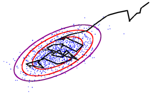
- Suppose a Markov chain with transition probabilities T satisfies

$$p^*(\boldsymbol{z})T(\boldsymbol{z}, \boldsymbol{z}') = p^*(\boldsymbol{z}')T(\boldsymbol{z}', \boldsymbol{z}) \quad \text{(Detailed Balance)}$$

- Integrating both sides w.r.t. $\boldsymbol{z}'$ gives $\int p^*(\boldsymbol{z}')T(\boldsymbol{z}', \boldsymbol{z})d\boldsymbol{z}' = p^*(\boldsymbol{z})$ (i.e., the Ergodic property)
  - Thus a Markov chain with detailed balance will always converge to a stationary distribution $p^*(\boldsymbol{z})$
  - Homogeneous Markov Chains satisfy detailed balance/ergodic property under mild conditions

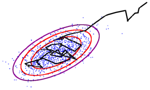
MCMC: The Basic Scheme

- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



MCMC: The Basic Scheme

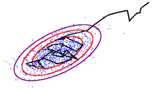
- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



- But we usually require several samples to approximate the distribution. How do we get those?

MCMC: The Basic Scheme

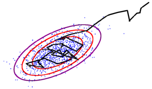
- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



- But we usually require several samples to approximate the distribution. How do we get those?
  - Start at an initial $\boldsymbol{z}^{(0)}$. Using a prop. dist. $p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$, run the chain long enough, say T_1 steps

MCMC: The Basic Scheme

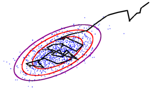
- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



- But we usually require several samples to approximate the distribution. How do we get those?
  - Start at an initial $\boldsymbol{z}^{(0)}$. Using a prop. dist. $p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$, run the chain long enough, say T_1 steps
  - Discard the first $(T_1 - 1)$ samples (called "burn-in" samples) and take the last sample $\boldsymbol{z}^{(T_1)}$

MCMC: The Basic Scheme

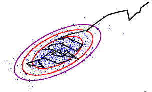
- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



- But we usually require several samples to approximate the distribution. How do we get those?
    - Start at an initial $\boldsymbol{z}^{(0)}$. Using a prop. dist. $p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$, run the chain long enough, say T_1 steps
    - Discard the first $(T_1 - 1)$ samples (called "burn-in" samples) and take the last sample $\boldsymbol{z}^{(T_1)}$
    - Continue from $\boldsymbol{z}^{(T_1)}$ up to T_2 steps, discard intermediate samples, take the last sample $\boldsymbol{z}^{(T_2)}$

MCMC: The Basic Scheme

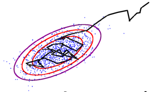
- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



- But we usually require several samples to approximate the distribution. How do we get those?
  - Start at an initial $\boldsymbol{z}^{(0)}$. Using a prop. dist. $p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$, run the chain long enough, say T_1 steps
  - Discard the first $(T_1 - 1)$ samples (called "burn-in" samples) and take the last sample $\boldsymbol{z}^{(T_1)}$
  - Continue from $\boldsymbol{z}^{(T_1)}$ up to T_2 steps, discard intermediate samples, take the last sample $\boldsymbol{z}^{(T_2)}$
    - This helps ensure that $\boldsymbol{z}^{(T_1)}$ and $\boldsymbol{z}^{(T_2)}$ are uncorrelated

MCMC: The Basic Scheme

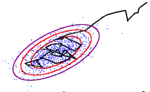
- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



- But we usually require several samples to approximate the distribution. How do we get those?
  - Start at an initial $\boldsymbol{z}^{(0)}$. Using a prop. dist. $p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$, run the chain long enough, say T_1 steps
  - Discard the first $(T_1 - 1)$ samples (called "burn-in" samples) and take the last sample $\boldsymbol{z}^{(T_1)}$
  - Continue from $\boldsymbol{z}^{(T_1)}$ up to T_2 steps, discard intermediate samples, take the last sample $\boldsymbol{z}^{(T_2)}$
    - This helps ensure that $\boldsymbol{z}^{(T_1)}$ and $\boldsymbol{z}^{(T_2)}$ are uncorrelated
  - Repeat the same for a total of S times

MCMC: The Basic Scheme

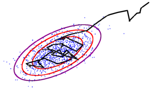
- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



- But we usually require several samples to approximate the distribution. How do we get those?
  - Start at an initial $\boldsymbol{z}^{(0)}$. Using a prop. dist. $p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$, run the chain long enough, say T_1 steps
  - Discard the first $(T_1 - 1)$ samples (called "burn-in" samples) and take the last sample $\boldsymbol{z}^{(T_1)}$
  - Continue from $\boldsymbol{z}^{(T_1)}$ up to T_2 steps, discard intermediate samples, take the last sample $\boldsymbol{z}^{(T_2)}$
    - This helps ensure that $\boldsymbol{z}^{(T_1)}$ and $\boldsymbol{z}^{(T_2)}$ are uncorrelated
  - Repeat the same for a total of S times
  - In the end, we have S i.i.d. samples from $p(\boldsymbol{z})$, i.e., $\boldsymbol{z}^{(T_1)}, \boldsymbol{z}^{(T_2)}, \ldots, \boldsymbol{z}^{(T_S)} \sim p(\boldsymbol{z})$

MCMC: The Basic Scheme

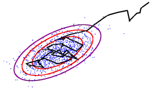
- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



- But we usually require several samples to approximate the distribution. How do we get those?
  - Start at an initial $\boldsymbol{z}^{(0)}$. Using a prop. dist. $p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$, run the chain long enough, say T_1 steps
  - Discard the first $(T_1 - 1)$ samples (called "burn-in" samples) and take the last sample $\boldsymbol{z}^{(T_1)}$
  - Continue from $\boldsymbol{z}^{(T_1)}$ up to T_2 steps, discard intermediate samples, take the last sample $\boldsymbol{z}^{(T_2)}$
    - This helps ensure that $\boldsymbol{z}^{(T_1)}$ and $\boldsymbol{z}^{(T_2)}$ are uncorrelated
  - Repeat the same for a total of S times
  - In the end, we have S i.i.d. samples from $p(\boldsymbol{z})$, i.e., $\boldsymbol{z}^{(T_1)}, \boldsymbol{z}^{(T_2)}, \ldots, \boldsymbol{z}^{(T_S)} \sim p(\boldsymbol{z})$
  - Note: Good choices for T_1 and $T_i - T_{i-1}$ are usually based on heuristics

MCMC: The Basic Scheme

- Running the MCMC chain infinitely long gives us ONE sample from the target distribution



- But we usually require several samples to approximate the distribution. How do we get those?
  - Start at an initial $\boldsymbol{z}^{(0)}$. Using a prop. dist. $p(\boldsymbol{z}^{(\ell+1)}|\boldsymbol{z}^{(\ell)})$, run the chain long enough, say T_1 steps
  - Discard the first $(T_1 - 1)$ samples (called "burn-in" samples) and take the last sample $\boldsymbol{z}^{(T_1)}$
  - Continue from $\boldsymbol{z}^{(T_1)}$ up to T_2 steps, discard intermediate samples, take the last sample $\boldsymbol{z}^{(T_2)}$
    - This helps ensure that $\boldsymbol{z}^{(T_1)}$ and $\boldsymbol{z}^{(T_2)}$ are uncorrelated
  - Repeat the same for a total of S times
  - In the end, we have S i.i.d. samples from $p(\boldsymbol{z})$, i.e., $\boldsymbol{z}^{(T_1)}, \boldsymbol{z}^{(T_2)}, \ldots, \boldsymbol{z}^{(T_S)} \sim p(\boldsymbol{z})$
  - Note: Good choices for T_1 and $T_i - T_{i-1}$ are usually based on heuristics
  - Note: MCMC is an approximate method because we don't usually know what T_1 is "long enough"

Some MCMC Sampling Algorithms

Metropolis-Hastings (MH) Sampling

- Assume a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, e.g., $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2\mathbf{I}_D)$

Metropolis-Hastings (MH) Sampling

- Assume a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, e.g., $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2 \mathbf{I}_D)$
- In each step, draw $\boldsymbol{z}^* \sim q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$ and accept the sample $\boldsymbol{z}^*$ with probability

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right)$$

Metropolis-Hastings (MH) Sampling

- Assume a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, e.g., $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2 \mathbf{I}_D)$
- In each step, draw $\boldsymbol{z}^* \sim q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$ and accept the sample $\boldsymbol{z}^*$ with probability

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right)$$

- The acceptance probability makes intuitive sense. Note the kind of $\boldsymbol{z}^*$ would it favor/unfavor:

Metropolis-Hastings (MH) Sampling

- Assume a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, e.g., $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2 \mathbf{I}_D)$
- In each step, draw $\boldsymbol{z}^* \sim q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$ and accept the sample $\boldsymbol{z}^*$ with probability

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right)$$

- The acceptance probability makes intuitive sense. Note the kind of $\boldsymbol{z}^*$ would it favor/unfavor:
  - It favors accepting $\boldsymbol{z}^*$ if $\tilde{p}(\boldsymbol{z}^*)$ has a higher value than $\tilde{p}(\boldsymbol{z}^{(\tau)})$

Metropolis-Hastings (MH) Sampling

- Assume a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, e.g., $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2 \mathbf{I}_D)$
- In each step, draw $\boldsymbol{z}^* \sim q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$ and accept the sample $\boldsymbol{z}^*$ with probability

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right)$$

- The acceptance probability makes intuitive sense. Note the kind of $\boldsymbol{z}^*$ would it favor/unfavor:
  - It favors accepting $\boldsymbol{z}^*$ if $\tilde{p}(\boldsymbol{z}^*)$ has a higher value than $\tilde{p}(\boldsymbol{z}^{(\tau)})$
  - Unfavors $\boldsymbol{z}^*$ if the proposal distribution q unduly favors it (i.e., if $q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})$ is large)

Metropolis-Hastings (MH) Sampling

- Assume a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, e.g., $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2 \mathbf{I}_D)$
- In each step, draw $\boldsymbol{z}^* \sim q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$ and accept the sample $\boldsymbol{z}^*$ with probability

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right)$$

- The acceptance probability makes intuitive sense. Note the kind of $\boldsymbol{z}^*$ would it favor/unfavor:
  - It favors accepting $\boldsymbol{z}^*$ if $\tilde{p}(\boldsymbol{z}^*)$ has a higher value than $\tilde{p}(\boldsymbol{z}^{(\tau)})$
  - Unfavors $\boldsymbol{z}^*$ if the proposal distribution q unduly favors it (i.e., if $q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})$ is large)
  - Favors $\boldsymbol{z}^*$ if we can "reverse" to $\boldsymbol{z}^{(\tau)}$ from $\boldsymbol{z}^*$ (i.e., if $q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)$ is large). Needed for good "mixing"

Metropolis-Hastings (MH) Sampling

- Assume a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, e.g., $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2 \mathbf{I}_D)$
- In each step, draw $\boldsymbol{z}^* \sim q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$ and accept the sample $\boldsymbol{z}^*$ with probability

$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min \left(1, \frac{\tilde{p}(\boldsymbol{z}^*) q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)}) q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right)$$

- The acceptance probability makes intuitive sense. Note the kind of $\boldsymbol{z}^*$ would it favor/unfavor:
  - It favors accepting $\boldsymbol{z}^*$ if $\tilde{p}(\boldsymbol{z}^*)$ has a higher value than $\tilde{p}(\boldsymbol{z}^{(\tau)})$
  - Unfavors $\boldsymbol{z}^*$ if the proposal distribution q unduly favors it (i.e., if $q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})$ is large)
  - Favors $\boldsymbol{z}^*$ if we can "reverse" to $\boldsymbol{z}^{(\tau)}$ from $\boldsymbol{z}^*$ (i.e., if $q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)$ is large). Needed for good "mixing"
- Transition probability of the Markov chain in MH sampling: $T(\boldsymbol{z}^{(\tau)}, \boldsymbol{z}^*) = A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})$

Metropolis-Hastings (MH) Sampling

- Assume a proposal distribution $q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$, e.g., $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2 \mathbf{I}_D)$
- In each step, draw $\boldsymbol{z}^* \sim q(\boldsymbol{z}|\boldsymbol{z}^{(\tau)})$ and accept the sample $\boldsymbol{z}^*$ with probability

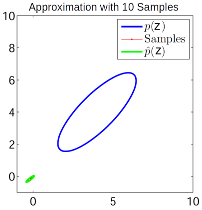
$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})}\right)$$

- The acceptance probability makes intuitive sense. Note the kind of $\boldsymbol{z}^*$ would it favor/unfavor:
  - It favors accepting $\boldsymbol{z}^*$ if $\tilde{p}(\boldsymbol{z}^*)$ has a higher value than $\tilde{p}(\boldsymbol{z}^{(\tau)})$
  - Unfavors $\boldsymbol{z}^*$ if the proposal distribution q unduly favors it (i.e., if $q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})$ is large)
  - Favors $\boldsymbol{z}^*$ if we can "reverse" to $\boldsymbol{z}^{(\tau)}$ from $\boldsymbol{z}^*$ (i.e., if $q(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^*)$ is large). Needed for good "mixing"
- Transition probability of the Markov chain in MH sampling: $T(\boldsymbol{z}^{(\tau)}, \boldsymbol{z}^*) = A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)})q(\boldsymbol{z}^*|\boldsymbol{z}^{(\tau)})$
- Exercise: Show that $T(\boldsymbol{z}, \boldsymbol{z}^{(\tau)})$ satisfies the detailed balance property

$$T(\boldsymbol{z}, \boldsymbol{z}^{(\tau)})p(\boldsymbol{z}) = T(\boldsymbol{z}^{(\tau)}, \boldsymbol{z})p(\boldsymbol{z}^{(\tau)})$$

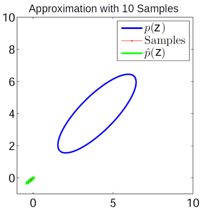
MH Sampling in Action: A Toy Example..

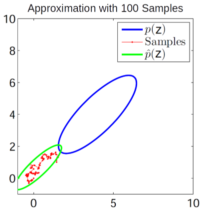
Target $p(\boldsymbol{z}) = \mathcal{N}\left(\begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right)$, Proposal $q(\boldsymbol{z}^{(t)}|\boldsymbol{z}^{(t-1)}) = \mathcal{N}\left(\boldsymbol{z}^{(t-1)}, \begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$



MH Sampling in Action: A Toy Example..

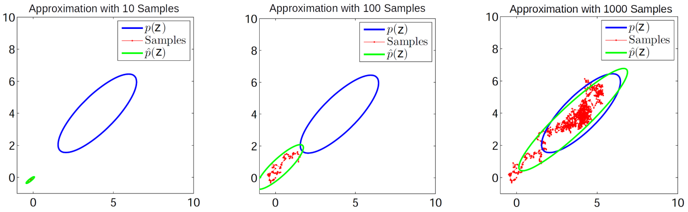
Target $p(\boldsymbol{z}) = \mathcal{N}\left(\begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right)$, Proposal $q(\boldsymbol{z}^{(t)}|\boldsymbol{z}^{(t-1)}) = \mathcal{N}\left(\boldsymbol{z}^{(t-1)}, \begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$





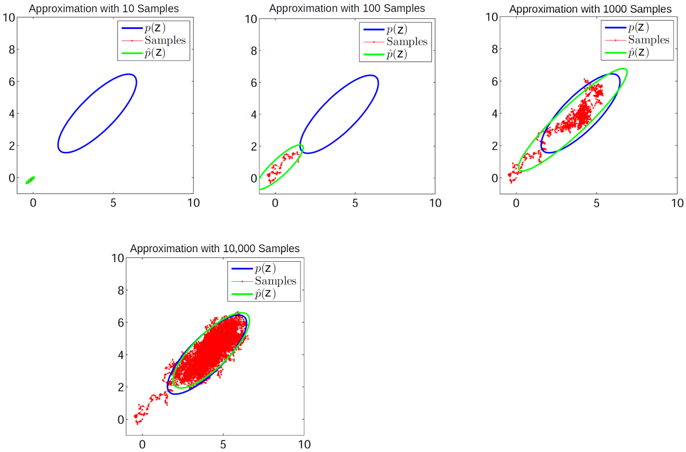
MH Sampling in Action: A Toy Example..

Target $p(\boldsymbol{z}) = \mathcal{N}\left(\begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right)$, Proposal $q(\boldsymbol{z}^{(t)}|\boldsymbol{z}^{(t-1)}) = \mathcal{N}\left(\boldsymbol{z}^{(t-1)}, \begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$



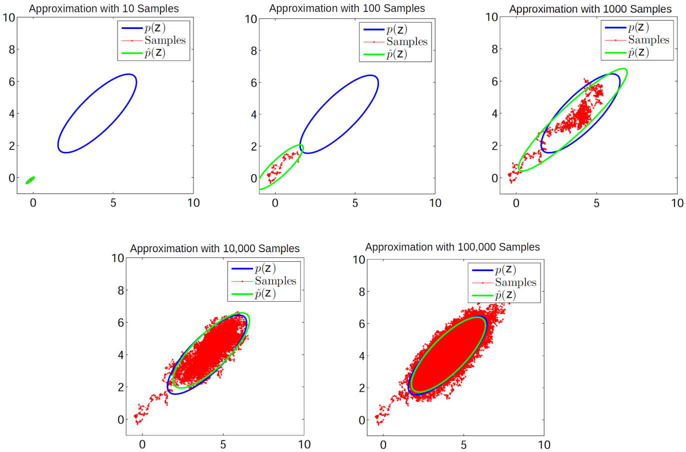
MH Sampling in Action: A Toy Example..

Target $p(\boldsymbol{z}) = \mathcal{N}\left(\begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right)$, Proposal $q(\boldsymbol{z}^{(t)}|\boldsymbol{z}^{(t-1)}) = \mathcal{N}\left(\boldsymbol{z}^{(t-1)}, \begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$



MH Sampling in Action: A Toy Example..

Target $p(\mathbf{z}) = \mathcal{N}\left(\begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right)$, Proposal $q(\mathbf{z}^{(t)}|\mathbf{z}^{(t-1)}) = \mathcal{N}\left(\mathbf{z}^{(t-1)}, \begin{bmatrix}0.01 & 0\\0 & 0.01\end{bmatrix}\right)$



MH Sampling: Some Comments

- Special Case: If proposal distrib. is symmetric, we get Metropolis Sampling algorithm with

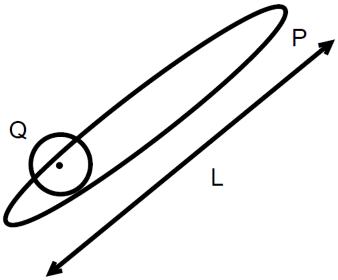
$$A(\boldsymbol{z}^*, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^*)}{\tilde{p}(\boldsymbol{z}^{(\tau)})}\right)$$

MH Sampling: Some Comments

- Special Case: If proposal distrib. is symmetric, we get Metropolis Sampling algorithm with

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

- Limitation: MH can have a very slow convergence



Generic proposals use $Q(x'; x) = \mathcal{N}(x, \sigma^2)$

σ **large** $\rightarrow$ **many rejections**

σ **small** $\rightarrow$ **slow diffusion:** $\sim (L/\sigma)^2$ iterations required

Gibbs Sampling (Geman & Geman, 1984)

- Suppose we wish to sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$

Gibbs Sampling (Geman & Geman, 1984)

- Suppose we wish to sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$
- However, suppose we can't sample from $p(\boldsymbol{z})$ but can sample from each conditional $p(z_i|\boldsymbol{z}_{-i})$

Gibbs Sampling (Geman & Geman, 1984)

- Suppose we wish to sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$
- However, suppose we can't sample from $p(\boldsymbol{z})$ but can sample from each conditional $p(z_i|\boldsymbol{z}_{-i})$
    - Can we done easily if we have a locally conjugate model (e.g., Gaussian matrix factorization)

Gibbs Sampling (Geman & Geman, 1984)

- Suppose we wish to sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$
- However, suppose we can't sample from $p(\boldsymbol{z})$ but can sample from each conditional $p(z_i|\boldsymbol{z}_{-i})$
  - Can we done easily if we have a locally conjugate model (e.g., Gaussian matrix factorization)
- Gibbs sampling uses the conditionals $p(z_i|\boldsymbol{z}_{-i})$ as the proposal distribution

Gibbs Sampling (Geman & Geman, 1984)

- Suppose we wish to sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$
- However, suppose we can't sample from $p(\boldsymbol{z})$ but can sample from each conditional $p(z_i|\boldsymbol{z}_{-i})$
    - Can we done easily if we have a locally conjugate model (e.g., Gaussian matrix factorization)
- Gibbs sampling uses the conditionals $p(z_i|\boldsymbol{z}_{-i})$ as the proposal distribution
- Gibbs sampling samples from these conditionals in a cyclic order

Gibbs Sampling (Geman & Geman, 1984)

- Suppose we wish to sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$
- However, suppose we can't sample from $p(\boldsymbol{z})$ but can sample from each conditional $p(z_i|\boldsymbol{z}_{-i})$
  - Can we done easily if we have a locally conjugate model (e.g., Gaussian matrix factorization)
- Gibbs sampling uses the conditionals $p(z_i|\boldsymbol{z}_{-i})$ as the proposal distribution
- Gibbs sampling samples from these conditionals in a cyclic order
- Gibbs sampling is equivalent to Metropolis Hastings sampling with acceptance prob. $= 1$

Gibbs Sampling (Geman & Geman, 1984)

- Suppose we wish to sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$
- However, suppose we can't sample from $p(\boldsymbol{z})$ but can sample from each conditional $p(z_i|\boldsymbol{z}_{-i})$
  - Can we done easily if we have a locally conjugate model (e.g., Gaussian matrix factorization)
- Gibbs sampling uses the conditionals $p(z_i|\boldsymbol{z}_{-i})$ as the proposal distribution
- Gibbs sampling samples from these conditionals in a cyclic order
- Gibbs sampling is equivalent to Metropolis Hastings sampling with acceptance prob. $= 1$

$$A(\boldsymbol{z}^*, \boldsymbol{z}) = \frac{p(\boldsymbol{z}^*)q(\boldsymbol{z}|\boldsymbol{z}^*)}{p(\boldsymbol{z})q(\boldsymbol{z}^*|\boldsymbol{z})}$$

Gibbs Sampling (Geman & Geman, 1984)

- Suppose we wish to sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$
- However, suppose we can't sample from $p(\boldsymbol{z})$ but can sample from each conditional $p(z_i|\boldsymbol{z}_{-i})$
  - Can we done easily if we have a locally conjugate model (e.g., Gaussian matrix factorization)
- Gibbs sampling uses the conditionals $p(z_i|\boldsymbol{z}_{-i})$ as the proposal distribution
- Gibbs sampling samples from these conditionals in a cyclic order
- Gibbs sampling is equivalent to Metropolis Hastings sampling with acceptance prob. $= 1$

$$A(\boldsymbol{z}^*, \boldsymbol{z}) = \frac{p(\boldsymbol{z}^*)q(\boldsymbol{z}|\boldsymbol{z}^*)}{p(\boldsymbol{z})q(\boldsymbol{z}^*|\boldsymbol{z})} = \frac{p(z_i^*|\boldsymbol{z}_{-i}^*)p(\boldsymbol{z}_{-i}^*)p(z_i|\boldsymbol{z}_{-i}^*)}{p(z_i|\boldsymbol{z}_{-i})p(\boldsymbol{z}_{-i})p(z_i^*|\boldsymbol{z}_{-i})}$$

Gibbs Sampling (Geman & Geman, 1984)

- Suppose we wish to sample from a joint distribution $p(\boldsymbol{z})$ where $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$
- However, suppose we can't sample from $p(\boldsymbol{z})$ but can sample from each conditional $p(z_i|\boldsymbol{z}_{-i})$
  - Can we done easily if we have a locally conjugate model (e.g., Gaussian matrix factorization)
- Gibbs sampling uses the conditionals $p(z_i|\boldsymbol{z}_{-i})$ as the proposal distribution
- Gibbs sampling samples from these conditionals in a cyclic order
- Gibbs sampling is equivalent to Metropolis Hastings sampling with acceptance prob. $= 1$

$$A(\boldsymbol{z}^*, \boldsymbol{z}) = \frac{p(\boldsymbol{z}^*)q(\boldsymbol{z}|\boldsymbol{z}^*)}{p(\boldsymbol{z})q(\boldsymbol{z}^*|\boldsymbol{z})} = \frac{p(z_i^*|\boldsymbol{z}^*_{-i})p(\boldsymbol{z}^*_{-i})p(z_i|\boldsymbol{z}^*_{-i})}{p(z_i|\boldsymbol{z}_{-i})p(\boldsymbol{z}_{-i})p(z_i^*|\boldsymbol{z}_{-i})} = 1$$

where we use the fact that $\boldsymbol{z}^*_{-i} = \boldsymbol{z}_{-i}$

Gibbs Sampling: Sketch of the Algorithm

M: Total number of variables, T: number of Gibbs sampling steps

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

Gibbs Sampling: Sketch of the Algorithm

M: Total number of variables, T: number of Gibbs sampling steps

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

Note: When sampling each variable from its conditional posterior, we use the most recent values of all other variables (this is akin to a co-ordinate ascent like procedure)

Gibbs Sampling: Sketch of the Algorithm

M: Total number of variables, T: number of Gibbs sampling steps

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.

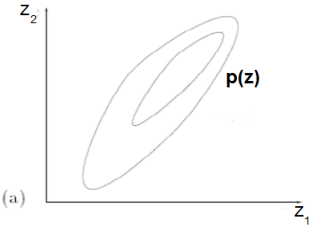
     $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

Note: When sampling each variable from its conditional posterior, we use the most recent values of all other variables (this is akin to a co-ordinate ascent like procedure)

Note: Order of updating the variables *usually* doesn't matter (but see "Scan Order in Gibbs Sampling: Models in Which it Matters and Bounds on How Much" from NIPS 2016)

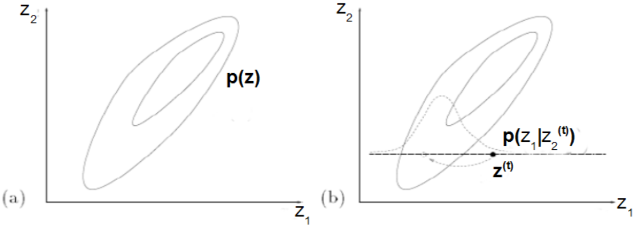
Gibbs Sampling: A Simple Example

Can sample from a 2-D Gaussian using 1-D Gaussians (recall that if the joint distribution is a 2-D Gaussian, conditionals will simply be 1-D Gaussians)



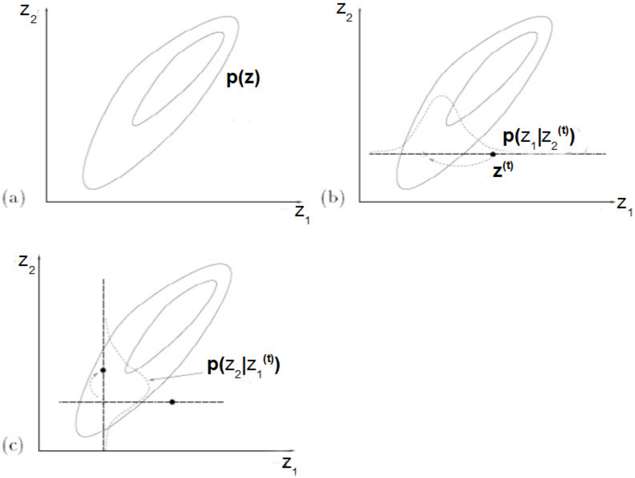
Gibbs Sampling: A Simple Example

Can sample from a 2-D Gaussian using 1-D Gaussians (recall that if the joint distribution is a 2-D Gaussian, conditionals will simply be 1-D Gaussians)



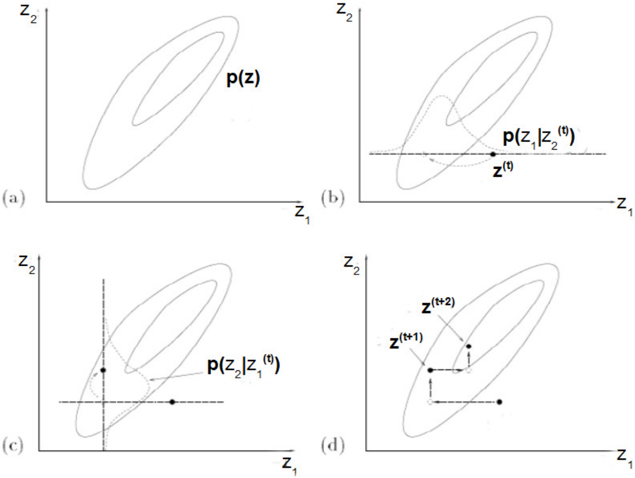
Gibbs Sampling: A Simple Example

Can sample from a 2-D Gaussian using 1-D Gaussians (recall that if the joint distribution is a 2-D Gaussian, conditionals will simply be 1-D Gaussians)



Gibbs Sampling: A Simple Example

Can sample from a 2-D Gaussian using 1-D Gaussians (recall that if the joint distribution is a 2-D Gaussian, conditionals will simply be 1-D Gaussians)



Next Class..

- More examples of Gibbs sampling
- Random-walk avoiding MCMC methods
- “Using” MCMC. Pros and Cons.
- Some recent advances in MCMC