Approximate Inference: Sampling Methods

CMSC 678
UMBC
Outline

Recap

Monte Carlo methods

Sampling Techniques
  Uniform sampling
  Importance Sampling
  Rejection Sampling
  Metropolis-Hastings
  Gibbs sampling

Example: Collapsed Gibbs Sampler for Topic Models
Recap from last time...
Exponential Family Forms: Capture Common Distributions

Discrete *(Finite distributions)*

\[ p_\pi(x) = \prod \pi_i^{1[x=i]} \quad \theta = (\log \pi_i)_i, \quad f(x) = (1[x_j = i])_i \]

Dirichlet *(Distributions over (finite) distributions)*

\[ p_\alpha(x) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \prod x_i^{\alpha_i-1} \quad \theta = (\alpha_i - 1)_i, \quad f(x) = (\log x_i)_i \]

Gaussian

\[ p_{\mu,\sigma^2}(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad \theta = \left(\frac{\mu}{\sigma^2}\right), \quad f(x) = \left(\begin{array}{c} x \\ x^2 \end{array}\right) \]

Gamma, Exponential, Poisson, Negative-Binomial, Laplace, log-Normal,...
Exponential Family Forms: “Easy” Posterior Inference

$p$ is the conjugate prior for $q$

Posterior $p$ has same form as prior $p$

$$p(\theta | x) \propto q(x | \theta) p(\theta)$$

<table>
<thead>
<tr>
<th>Posterior</th>
<th>Likelihood</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirichlet (Beta)</td>
<td>Discrete (Bernoulli)</td>
<td>Dirichlet (Beta)</td>
</tr>
<tr>
<td>Normal</td>
<td>Normal (fixed var.)</td>
<td>Normal</td>
</tr>
<tr>
<td>Gamma</td>
<td>Exponential</td>
<td>Gamma</td>
</tr>
</tbody>
</table>
Variational Inference: A Gradient-Based Optimization Technique

Set $t = 0$
Pick a starting value $\lambda_t$
Until converged:
1. Get value $y_t = F(q(\cdot;\lambda_t))$
2. Get gradient $g_t = F'(q(\cdot;\lambda_t))$
3. Get scaling factor $\rho_t$
4. Set $\lambda_{t+1} = \lambda_t + \rho_t * g_t$
5. Set $t += 1$
Variational Inference: The Function to Optimize

KL-Divergence (expectation)

$$\min_q D_{KL}(q(\theta) \| p(\theta \mid x))$$

Find the best distribution

$$D_{KL}(q(\theta) \| p(\theta \mid x)) = \mathbb{E}_{q(\theta)} \left[ \log \frac{q(\theta)}{p(\theta \mid x)} \right]$$

Parameters for desired model

$$q(\theta \mid \lambda)$$

Variational parameters for $\theta$
Goal: Posterior Inference

Hyperparameters $\alpha$

Unknown parameters $\Theta$

Data:

Likelihood model:
Some Learning Techniques

MAP/MLE: Point estimation, basic EM

Variational Inference: Functional Optimization

Sampling/Monte Carlo
Outline

Recap

Monte Carlo methods

Sampling Techniques
  Uniform sampling
  Importance Sampling
  Rejection Sampling
  Metropolis-Hastings
  Gibbs sampling

Example: Collapsed Gibbs Sampler for Topic Models
Two Problems for Sampling Methods to Solve

Generate samples from \( p \)

\[
p(x) = \frac{u(x)}{Z}, \quad x \in \mathbb{R}^D
\]

\( x_1, x_2, ..., x_R \) samples

Q: Why is sampling from \( p(x) \) hard?
Two Problems for Sampling Methods to Solve

Generate samples from $p$

$$p(x) = \frac{u(x)}{Z}, x \in \mathbb{R}^D$$

$x_1, x_2, ..., x_R$ samples

**Q:** Why is sampling from $p(x)$ hard?

**A1:** Can we evaluate $Z$?

**A2:** Can we sample without enumerating? (Correct samples should be where $p$ is big)
Two Problems for Sampling Methods to Solve

Generate samples from $p$

$$p(x) = \frac{u(x)}{Z}, x \in \mathbb{R}^D$$

$x_1, x_2, ..., x_R$ samples

**Q:** Why is sampling from $p(x)$ hard?

**A1:** Can we evaluate $Z$?

**A2:** Can we sample without enumerating? (Correct samples should be where $p$ is big)

$$u(x) = \exp(0.4(x - 0.4)^2 - 0.08x^4)$$

(ITILA, Fig 29.1)
Two Problems for Sampling Methods to Solve

Generate samples from \( p \)

\[
p(x) = \frac{u(x)}{Z}, x \in \mathbb{R}^D
\]

\( x_1, x_2, ..., x_R \) samples

Estimate expectation of a function \( \phi \)

\[
\Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p} [\phi(x)] = \int p(x) \phi(x) dx
\]

Q: Why is sampling from \( p(x) \) hard?

A1: Can we evaluate \( Z \)?

A2: Can we sample without enumerating? (Correct samples should be where \( p \) is big)
Two Problems for Sampling Methods to Solve

Generate samples from $p$

$$p(x) = \frac{u(x)}{Z}, \; x \in \mathbb{R}^D$$

$x_1, x_2, \ldots, x_R$ samples

Estimate expectation of a function $\phi$

$$\Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] = \int p(x)\phi(x)dx$$

$$\hat{\Phi} = \frac{1}{R} \sum_r \phi(x_r)$$

Q: Why is sampling from $p(x)$ hard?

A1: Can we evaluate $Z$?

A2: Can we sample without enumerating? (Correct samples should be where $p$ is big)
Two Problems for Sampling Methods to Solve

Generate samples from $p$

$p(x) = \frac{u(x)}{Z}, x \in \mathbb{R}^D$

$x_1, x_2, ..., x_R$ samples

Q: Why is sampling from $p(x)$ hard?

A1: Can we evaluate $Z$?

A2: Can we sample without enumerating? (Correct samples should be where $p$ is big)

Estimate expectation of a function $\phi$

$\Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] = \int p(x)\phi(x)dx$

$\hat{\Phi} = \frac{1}{R} \sum_r \phi(x_r)$

If we could sample from $p$...

$\mathbb{E}[\hat{\Phi}] = \Phi$ consistent estimator
Outline

Recap

Monte Carlo methods

Sampling Techniques
  Uniform sampling
  Importance Sampling
  Rejection Sampling
  Metropolis-Hastings
  Gibbs sampling

Example: Collapsed Gibbs Sampler for Topic Models
Uniform Sampling

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

Sample uniformly:
\[ x_1, x_2, \ldots, x_R \]

\[ \hat{\Phi} = \sum_r \phi(x_r)p^*(x_r) \]
Uniform Sampling

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

sample uniformly:
\[ x_1, x_2, \ldots, x_R \]

\[ \hat{\Phi} = \sum_r \phi(x_r)p^*(x_r) \]

\[ p^*(x) = \frac{u(x)}{Z^*} \]

\[ Z^* = \sum_r u(x_r) \]
Uniform Sampling

\[ p^* x = u x \]

\[ Z^* \Phi = \Phi \]

\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

Goal:

\[ \Phi = \sum_r \phi(x_r) p^*(x_r) \]

\[ p^*(x) = \frac{u(x)}{Z^*} \]

\[ Z^* = \sum_r u(x_r) \]

sample uniformly:

\[ x_1, x_2, \ldots, x_R \]

this *might* work if \( R \) (the number of samples) sufficiently hits high probability regions
Uniform Sampling

Sample uniformly:

\[ x_1, x_2, \ldots, x_R \]

Goal:

\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

\[ \hat{\Phi} = \sum_r \phi(x_r)p^*(x_r) \]

\[ p^*(x) = \frac{u(x)}{Z^*} \]

\[ Z^* = \sum_r u(x_r) \]

*This might* work if \( R \) (the number of samples) sufficiently hits high probability regions

**Ising model example:**

- \( 2^H \) states of high probability
- \( 2^N \) states total
Uniform Sampling

\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

Goal:

\[ \hat{\Phi} = \sum_r \phi(x_r) p^*(x_r) \]

\[ p^*(x) = \frac{u(x)}{Z^*} \]

\[ Z^* = \sum_r u(x_r) \]

**Ising model example:**

- \( 2^H \) states of high probability
- \( 2^N \) states total

this *might* work if \( R \) (the number of samples) sufficiently hits high probability regions

chance of sample being in high prob. region: \( \frac{2^H}{2^N} \)

min. samples needed: \( \sim 2^{N-H} \)
Outline

Recap

Monte Carlo methods

Sampling Techniques
  Uniform sampling
  Importance Sampling
  Rejection Sampling
  Metropolis-Hastings
  Gibbs sampling

Example: Collapsed Gibbs Sampler for Topic Models
Importance Sampling

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

approximating distribution:
\[ Q(x) \propto u_q(x) \]

sample from \( Q \):
\[ x_1, x_2, \ldots, x_R \]
Importance Sampling

approximating distribution:
\( Q(x) \propto u_q(x) \)
/sample from \( Q \):
\( x_1, x_2, \ldots, x_R \)

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

\( x \) where \( Q(x) > p(x) \): over-represented
\( x \) where \( Q(x) < p(x) \): under-represented

ITILA, Fig 29.5
Importance Sampling

approximating distribution:
\[ Q(x) \propto u_q(x) \]

sample from \( Q \):
\[ x_1, x_2, \ldots, x_R \]

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = E_{x \sim p} [\phi(x)] \]

\[ \hat{\Phi} = \frac{\sum_r \phi(x_r)w(x_r)}{\sum_r w(x_r)} \]

\[ w(x_r) = \frac{u_p(x)}{u_q(x)} \]

\( x \) where \( Q(x) > p(x) \): over-represented
\( x \) where \( Q(x) < p(x) \): under-represented

ITILA, Fig 29.5
**Importance Sampling**

approximating distribution:
\[ Q(x) \propto u_q(x) \]

sample from \( Q \):
\[ x_1, x_2, \ldots, x_R \]

**Goal:**
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

\[ \hat{\Phi} = \frac{\sum_r \phi(x_r)w(x_r)}{\sum_r w(x_r)} \]

\[ w(x_r) = \frac{u_p(x)}{u_q(x)} \]

x where \( Q(x) > p(x) \): over-represented
x where \( Q(x) < p(x) \): under-represented

**Q:** How reliable will this estimator be?

ITILA, Fig 29.5
Importance Sampling

approximating distribution:
\[ Q(x) \propto u_q(x) \]

sample from \( Q \):
\[ x_1, x_2, \ldots, x_R \]

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

\[ \hat{\Phi} = \frac{\sum_r \phi(x_r)w(x_r)}{\sum_r w(x_r)} \]

\[ w(x_r) = \frac{u_p(x)}{u_q(x)} \]

Q: How reliable will this estimator be?

A: In practice, difficult to say. \( w(x_r) \) may not be a good indicator.
Importance Sampling

approximating
distribution:
\( Q(x) \propto u_q(x) \)

sample from \( Q \):
\( x_1, x_2, \ldots, x_R \)

\[
\Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)]
\]

\[
\hat{\Phi} = \frac{\sum_r \phi(x_r)w(x_r)}{\sum_r w(x_r)}
\]

\[
w(x_r) = \frac{u_p(x)}{u_q(x)}
\]

**Q:** How reliable will this estimator be?

**A:** In practice, difficult to say. \( w(x_r) \) may not be a good indicator.

**Q:** How do you choose a good approximating distribution?

ITILA, Fig 29.5
Importance Sampling

approximating distribution:
\[ Q(x) \propto u_q(x) \]

sample from \( Q \):
\[ x_1, x_2, \ldots, x_R \]

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p} \phi(x) \]

\[ \hat{\Phi} = \frac{\sum_r \phi(x_r)w(x_r)}{\sum_r w(x_r)} \]

\[ w(x_r) = \frac{u_p(x)}{u_q(x)} \]

- \( x \) where \( Q(x) > p(x) \): over-represented
- \( x \) where \( Q(x) < p(x) \): under-represented

\textbf{Q:} How reliable will this estimator be?

\textbf{A:} In practice, difficult to say. \( w(x_r) \) may not be a good indicator

\textbf{Q:} How do you choose a good approximating distribution?

\textbf{A:} Task/domain specific
Importance Sampling: Variance Estimator may vary

$q(x):$ Gaussian

$q(x):$ Cauchy distribution

ITILA, Fig 29.6
Outline

Recap

Monte Carlo methods

Sampling Techniques
  Uniform sampling
  Importance Sampling
  **Rejection Sampling**
  Metropolis-Hastings
  Gibbs sampling

Example: Collapsed Gibbs Sampler for Topic Models
Rejection Sampling

approximating distribution:

\[ Q(x) \propto u_q(x), \; c \cdot u_q > u_p \]

Goal:

\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]
Rejection Sampling

approximating distribution:
\( Q(x) \propto u_q(x), c \ast u_q > u_p \)

sample from \( Q \):
\( x_1, x_2, \ldots, x_{R^*} \)

sample uniformly:
\( z_k \sim \text{Unif}(0, c \ast u_q(x_k)) \)

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p} [\phi(x)] \]
Rejection Sampling

approximating distribution: $Q(x) \propto u_q(x), c \cdot u_q > u_p$

sample from $Q$:
$x_1, x_2, \ldots, x_{R^*}$
sample uniformly:
$z_k \sim \text{Unif}(0, c \cdot u_q(x_k))$

if $z_k \leq u_p(x_k)$: add $x_k$ to sampled R points
otherwise: reject it

Goal:
$\Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)]$
Rejection Sampling

approximating distribution: $Q(x) \propto u_q(x), c \cdot u_q > u_p$

sample from $Q$:
$x_1, x_2, \ldots, x_{R^*}$
sample uniformly:
$z_k \sim \text{Unif}(0, c \cdot u_q(x_k))$

Goal:
$\Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)]$

if $z_k \leq u_p(x_k)$: add $x_k$ to sampled $R$ points
otherwise: reject it

this produces samples from the $p$-distribution
Rejection Sampling

approximating distribution: 
\( Q(x) \propto u_q(x), c \cdot u_q > u_p \)

sample from \( Q \):
\( x_1, x_2, \ldots, x_{R^*} \)

sample uniformly:
\( z_k \sim \text{Unif}(0, c \cdot u_q(x_k)) \)

Goal:
\( \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \)

if \( z_k \leq u_p(x_k) \): add \( x_k \) to sampled \( R \) points
otherwise: reject it

\( \hat{\Phi} = \frac{1}{R} \sum_r \phi(x_r) \)
Rejection Sampling

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

approximating distribution:
\[ Q(x) \propto u_q(x), \ c \ast u_q > u_p \]

sample from \( Q \):
\[ x_1, x_2, \ldots, x_{R^*} \]
sample uniformly:
\[ z_k \sim \text{Unif}(0, c \ast u_q(x_k)) \]

if \( z_k \leq u_p(x_k) \): add \( x_k \) to sampled \( R \) points
otherwise: reject it

\[ \hat{\Phi} = \frac{1}{R} \sum_{r} \phi(x_r) \]

Q: How reliable will this estimator be?
Rejection Sampling

approximating distribution: $Q(x) \propto u_q(x), c * u_q > u_p$

sample from $Q$: $x_1, x_2, \ldots, x_{R^*}$

sample uniformly: $z_k \sim \text{Unif}(0, c * u_q(x_k))$

if $z_k \leq u_p(x_k)$: add $x_k$ to sampled $R$ points
otherwise: reject it

Goal:

$\Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)]$

$\hat{\Phi} = \frac{1}{R} \sum_r \phi(x_r)$

Q: How reliable will this estimator be?
A: How well does $Q$ approximate $P$?

Q: How do you choose a good approximating distribution?
Rejection Sampling

approximating distribution: 
\( Q(x) \propto u_q(x), \ c \ast u_q > u_p \)

sample from \( Q \): 
\( x_1, x_2, \ldots, x_{R^*} \)

sample uniformly:
\( z_k \sim \text{Unif}(0, c \ast u_q(x_k)) \)

Goal:
\[
\Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)]
\]

if \( z_k \leq u_p(x_k) \): add \( x_k \) to sampled \( R \) points
otherwise: reject it

\[
\hat{\Phi} = \frac{1}{R} \sum_r \phi(x_r)
\]

Q: How reliable will this estimator be?
A: How well does \( Q \) approximate \( P \)?

Q: How do you choose a good approximating distribution?
A: Task/domain specific
Rejection Sampling

approximating distribution: 
\( Q(x) \propto u_q(x), c \cdot u_q > u_p \)

sample \textbf{from} \( Q \):
\( x_1, x_2, \ldots, x_{R^*} \)
sample \textbf{uniformly}:
\( z_k \sim \text{Unif}(0, c \cdot u_q(x_k)) \)

\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

if \( z_k \leq u_p(x_k) \): add \( x_k \) to sampled \( R \) points
otherwise: reject it

\[ \hat{\Phi} = \frac{1}{R} \sum_r \phi(x_r) \]

\textbf{Q}: How reliable will this estimator be?
\textbf{A}: How well does \( Q \) approximate \( P \)?

\textbf{Q}: How do you choose a good approximating distribution?
\textbf{A}: Task/domain specific

rejection sampling can be difficult to use in high-dimensional spaces 😞
Outline

Recap

Monte Carlo methods

Sampling Techniques
  Uniform sampling
  Importance Sampling
  Rejection Sampling
  Metropolis-Hastings
  Gibbs sampling

Example: Collapsed Gibbs Sampler for Topic Models
Markov Chain Monte Carlo

\[ \theta^{(t)} \xrightarrow{\text{transition kernel}} \theta^{(t+1)} \]
Metropolis-Hastings

importance and rejection sampling:

a **single** proposal distribution

\[ Q(x) \propto u_q(x) \]

Metropolis-Hastings (and Gibbs):
create a proposal distribution based on **current state**

\[ Q(x|x^{(t)}) \propto u_q(x|x^{(t)}) \]

Goal:

\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]
Metropolis-Hastings

importance and rejection sampling: a **single** proposal distribution

\[ Q(x) \propto u_q(x) \]

Metropolis-Hastings (and Gibbs): create a proposal distribution based on **current state**

\[ Q(x|x^{(t)}) \propto u_q(x|x^{(t)}) \]

\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p} [\phi(x)] \]

Q does not need to look similar to P

ITILA, Fig 29.10
Metropolis-Hastings

transition kernel/distribution:
\[ Q(x|x^{(t)}) \propto u_q(x|x^{(t)}) \]

sample from \( Q(x|x^{(t)}) \):
\[ x_1, x_2, \ldots, x_{R^*} \]
sample uniformly:
\[ \alpha_k = \frac{u_p(x_k) u_q(x_k)}{u_p(x^{(t)}) u_q(x^{(t)})} \]

if \( \alpha_k \geq 1 \): add \( x_k \) to sampled \( R \) points
otherwise: accept with probability \( \alpha_k \)

\[ \hat{\Phi} = \frac{1}{R} \sum_r \phi(x_r) \]

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p} [\phi(x)] \]
Metropolis-Hastings

transition kernel/distribution:
\[ Q(x|x^{(t)}) \propto u_q(x|x^{(t)}) \]

sample from \( Q(x|x^{(t)}) \):
\[ x_1, x_2, ..., x_{R^*} \]
sample uniformly:
\[ \alpha_k = \frac{u_p(x_k) u_q(x_k)}{u_p(x^{(t)}) u_q(x^{(t)})} \]

if \( \alpha_k \geq 1 \): add \( x_k \) to sampled \( R \) points
otherwise: accept with probability \( \alpha_k \)

if accepted: \( x^{(t+1)} = x_k \)
otherwise: \( x^{(t+1)} = x^{(t)} \)

samples are not independent

\[ \Phi = \frac{1}{R} \sum_r \phi(x_r) \]

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p} [\phi(x)] \]
Goal: \[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

Metropolis-Hastings

transition kernel/distribution: \[ Q(x|x^{(t)}) \propto u_q(x|x^{(t)}) \]

sample from \( Q(x|x^{(t)}) \):
\[ x_1, x_2, ..., x_{R^*} \]
sample uniformly: \[ \alpha_k = \frac{u_p(x_k) u_q(x_k)}{u_p(x^{(t)}) u_q(x^{(t)})} \]

if \( \alpha_k \geq 1 \): add \( x_k \) to sampled \( R \) points
otherwise: accept with probability \( \alpha_k \)
\[
\text{if accepted: } x^{(t+1)} = x_k \\
\text{otherwise: } x^{(t+1)} = x^{(t)}
\]
samples are not independent

\[ \hat{\Phi} = \frac{1}{R} \sum_r \phi(x_r) \]

Metropolis-Hastings can be used effectively in high-dimensional spaces 😊
Outline

Recap

Monte Carlo methods

Sampling Techniques
  - Uniform sampling
  - Importance Sampling
  - Rejection Sampling
  - Metropolis-Hastings
  - Gibbs sampling

Example: Collapsed Gibbs Sampler for Topic Models
Gibbs Sampling

transition kernel/distribution:

\[ Q(x|x^{(t)}) = p(x | \text{all other variables}) \]

\[ x^{(t+1)}_{[i]} \sim p(\cdot | x^{(t+1)}_{[1]}, \ldots, x^{(t+1)}_{[i-1]}, x^{(t)}_{[i+1]}, \ldots, x^{(t)}_{[N]}) \]

Next sampled value of current variable

Values of all other variables, both new and old
Remember: Markov Blanket

the set of nodes needed to form the complete conditional for a variable $x_i$

$$p(x_i | x_j \neq i) = \frac{p(x_1, \ldots, x_N)}{\int p(x_1, \ldots, x_N) dx_i}$$

factorization of graph

$$= \frac{\prod_k p(x_k | \pi(x_k))}{\int \prod_k p(x_k | \pi(x_k)) \, dx_i}$$

factor out terms not dependent on $x_i$

$$= \frac{\prod_{k: k = i \text{ or } i \in \pi(x_k)} p(x_k | \pi(x_k))}{\int \prod_{k: k = i \text{ or } i \in \pi(x_k)} p(x_k | \pi(x_k)) \, dx_i}$$
Gibbs Sampling

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

transition kernel/distribution:
\[ Q(x|x^{(t)}) = p(x | \text{MB}(x^{(t)})) \]

sample (always accept) from
\[ Q(x|x^{(t)}): x_1, x_2, \ldots, x_{R^*} \]

samples are not independent

Gibbs Sampling can be used effectively in high-dimensional spaces 😊
Collapsed Gibbs Sampling

Goal:
\[ \Phi = \langle \phi(x) \rangle_p = \mathbb{E}_{x \sim p}[\phi(x)] \]

transition kernel/distribution:
\[ Q(x|x^{(t)}) = \int p(x | \text{MB}(x^{(t)})) \, dy = p(x | \text{MB}_{-y}(x^t)) \]

integrate out some of Markov blanket

sample (always accept) from
\[ Q(x|x^{(t)}): x_1, x_2, \ldots, x_{R^*} \]

\[ x^{(t+1)} = x_k \]

\[ \hat{\Phi} = \frac{1}{R} \sum_r \phi(x_r) \]

samples are not independent

Collapsed Gibbs can be used effectively in high-dimensional spaces 😊
Outline

Recap

Monte Carlo methods

Sampling Techniques
  Uniform sampling
  Importance Sampling
  Rejection Sampling
  Metropolis-Hastings
  Gibbs sampling

Example: Collapsed Gibbs Sampler for Topic Models
Latent Dirichlet Allocation
(Blei et al., 2003)

\[ w^{(d,n)} \sim \text{Discrete} (\phi_{z^{(d,n)}}) \]
\[ \phi_k \sim \text{Dirichlet} (\beta) \]
\[ z^{(d,n)} \sim \text{Discrete} (\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet} (\alpha) \]
Gibbs Sampler for LDirA

for each document d:
\[ \text{resample } \theta_d \mid z_{d,1}, \ldots, z_{d,N_d} \]
for each token i in d:
\[ \text{resample } z_{d,i} \mid w_{d,i}, \{ \psi_k \}, \theta_d \]

for each topic k:
\[ \text{resample } \psi_k \]
Latent Dirichlet Allocation
(Blei et al., 2003)

Per-document (unigram) word counts

Per-document (latent) topic usage

Per-topic word usage

\( w^{(d,n)} \sim \text{Discrete} (\phi_{\tilde{z}^{(d,n)}}) \)

\( \phi_k \sim \text{Dirichlet} (\beta) \)

\( \tilde{z}^{(d,n)} \sim \text{Discrete} (\theta^{(d)}) \)

\( \theta^{(d)} \sim \text{Dirichlet} (\alpha) \)

integrate these out
Collapsed Gibbs Sampler for LDirA

for each document \( d \):
    resample \( \theta_d \mid z_{d,1}, \ldots, z_{d,N_d} \)

for each token \( i \) in \( d \):
    resample \( z_{d,i} \mid w_{d,i}, \{\psi_k\}, \{z^*,-i\} \)

for each topic \( k \):
    resample \( \psi_k \)
Collapsed Gibbs Sampler for LDirA

for each document \(d\):

resample \(\theta_d \mid z_{d,1}, \ldots, z_{d,N_d}\)

for each token \(i\) in \(d\):

resample \(z_{d,i} \mid w_{d,i}, \{\psi_k\}, \{z_{*,-i}\}\)

for each topic \(k\):

resample \(\psi_k\)

\[
p(z_{di} \mid z_{*,-i}) = \frac{p(z_{*,*})}{p(z_{*,-i})}
\]
Sampling: Discrete Observations

\[ \theta \sim \text{Dirichlet}\ (\alpha) \]

\[ z_i \mid \theta \sim \text{Discrete}\ (\theta) \]
Sampling: Discrete Observations

\[ \theta \sim \text{Dirichlet} (\alpha) \]

\[ z_i \mid \theta \sim \text{Discrete} (\theta) \]

\[ p_\alpha(z) = \int_{\theta} p(z \mid \theta)p_\alpha(\theta)d\theta \]
Sampling: Discrete Observations

\[ \theta \sim \text{Dirichlet} \left( \alpha \right) \]

\[ z_i \mid \theta \sim \text{Discrete} \left( \theta \right) \]

\[ p_\alpha(z) = \int_\theta p(z | \theta) p_\alpha(\theta) d\theta \]

\[ = \frac{\Gamma \left( \sum_k \alpha_k \right)}{\Gamma \left( \sum_k (c(k) + \alpha_k) \right)} \prod_k \frac{\Gamma \left( c(k) + \alpha_k \right)}{\Gamma \left( \alpha_k \right)} \]

\[ = \text{DMC}_z \left( \alpha \right) \]

Griffiths and Stevers (PNAS, 2004)
Sampling: Discrete Observations

\[ \theta \sim \text{Dirichlet}(\alpha) \]

\[ z_i \mid \theta \sim \text{Discrete}(\theta) \]

\[
p_{\alpha}(z) = \int_{\theta} p(z \mid \theta) p_{\alpha}(\theta) d\theta
\]

\[
= \frac{\Gamma \left( \sum_k \alpha_k \right)}{\Gamma \left( \sum_k \left( c(k) + \alpha_k \right) \right)} \prod_k \frac{\Gamma \left( c(k) + \alpha_k \right)}{\Gamma \left( \alpha_k \right)}
\]

Gamma function fact: \( \Gamma(x + 1) = x\Gamma(x) \)

Griffiths and Stevers (PNAS, 2004)
Sampling: Discrete Observations

\[ p_\alpha(z) = \int_\theta p(z|\theta)p_\alpha(\theta) d\theta \]

\[ = \frac{\Gamma \left( \sum_k \alpha_k \right)}{\Gamma \left( \sum_k \left( c(k) + \alpha_k \right) \right)} \prod_k \frac{\Gamma \left( c(k) + \alpha_k \right)}{\Gamma \left( \alpha_k \right)} \]

Collapsed Gibbs Sampling goal:

\[ p(z_{di} | z_{*,-i}) = \frac{p(z_{*,*})}{p(z_{*,-i})} \]

Griffiths and Stevers (PNAS, 2004)
Sampling: Discrete Observations

\[ p_\alpha(z) = \int_\theta p(z|\theta) p_\alpha(\theta) d\theta \]

\[ = \frac{\Gamma \left( \sum_k \alpha_k \right)}{\Gamma \left( \sum_k (c(k) + \alpha_k) \right)} \prod_k \frac{\Gamma(c(k) + \alpha_k)}{\Gamma(\alpha_k)} \]

Gamma function fact:

\[ \Gamma(x + 1) = x\Gamma(x) \]

Collapsed Gibbs Sampling goal:

\[ p(z_{di} | z_{*,i}) = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k c(d, k) + \alpha_k)} \prod_k \frac{\Gamma(c(d, k) + \alpha_k)}{\Gamma(\alpha_k)} \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k c(d, k) - 1 + \alpha_k)} \prod_k \frac{\Gamma(c(d, k) - 1 + \alpha_k)}{\Gamma(\alpha_k)} \]
Sampling: Discrete Observations

\[ p_\alpha(z) = \int_\theta p(z|\theta) p_\alpha(\theta) d\theta \]

\[ = \frac{\Gamma \left( \sum_k \alpha_k \right)}{\Gamma \left( \sum_k (c(k) + \alpha_k) \right)} \prod_k \frac{\Gamma (c(k) + \alpha_k)}{\Gamma (\alpha_k)} \]

Gamma function fact:
\[ \Gamma(x + 1) = x\Gamma(x) \]

Collapsed Gibbs Sampling goal:

\[ p(z_{di} | z_{*,-i}) = \frac{\prod_k (c(d, k) - 1 + \alpha_k) \Gamma(c(d, k) - 1 + \alpha_k)}{(\sum_k c(d, k) - 1 + \alpha_k) \Gamma(\sum_k c(d, k) - 1 + \alpha_k)} \frac{\prod_k \Gamma(c(d, k) - 1 + \alpha_k)}{\Gamma(\sum_k c(d, k) - 1 + \alpha_k)} \]

Griffiths and Stevers (PNAS, 2004)
Sampling: Discrete Observations

\[ p_\alpha(z) = \int_\theta p(z|\theta)p_\alpha(\theta)d\theta \]

\[ = \frac{\Gamma \left( \sum_k \alpha_k \right)}{\Gamma \left( \sum_k \left( c(k) + \alpha_k \right) \right)} \prod_k \frac{\Gamma \left( c(k) + \alpha_k \right)}{\Gamma \left( \alpha_k \right)} \]

Gamma function fact:
\[ \Gamma(x + 1) = x\Gamma(x) \]

Collapsed Gibbs Sampling goal:
\[ p(z_{di} | z_{*,-i}) = \frac{\prod_k \left( c(d, k) - 1 + \alpha_k \right) \Gamma \left( c(d, k) - 1 + \alpha_k \right)}{\left( \sum_k c(d, k) - 1 + \alpha_k \right) \Gamma \left( \sum_k c(d, k) - 1 + \alpha_k \right)} \frac{\prod_k \Gamma \left( c(d, k) - 1 + \alpha_k \right)}{\Gamma \left( \sum_k c(d, k) - 1 + \alpha_k \right)} \]
Sampling: Discrete Observations

\[ p_{\alpha}(z) = \int_{\theta} p(z|\theta)p_{\alpha}(\theta)d\theta \]

\[ = \frac{\Gamma \left( \sum_k \alpha_k \right)}{\Gamma \left( \sum_k (c(k) + \alpha_k) \right)} \prod_k \frac{\Gamma \left( c(k) + \alpha_k \right)}{\Gamma (\alpha_k)} \]

Gamma function fact:
\[ \Gamma(x + 1) = x\Gamma(x) \]

Collapsed Gibbs Sampling goal:
\[ p(z_{di} = k \mid z_{*,-i}) \propto c(d, k) - 1 + \alpha_k \]

Griffiths and Stevers (PNAS, 2004)
Collapsed Gibbs Sampler for LDirA

for each document $d$:
  for each token $i$ in $d$:
    resample $z_{d,i} \mid w_{d,i}, \{\psi_k\}, \{z_{*,-i}\}$

$$p(z_{d,i} \mid z_{*,-i}) = \propto (c(d, k) - 1 + \alpha_k) \text{ *topic-word counts}$$
Collapsed Gibbs Sampler for LDirA

randomly assign $z_*, *$
maintain count tables:
  $c(d, k)$: document-topic counts
  $c(k, v)$: topic-word counts
for each document $d$:
  for each token $i$ in $d$:
    unassign topic: $z_{d,i}$
    resample $z_{d,i} | w_{d,i}, \{\psi_k\}, \{z_{*,-i}\}$
    reassign topic: $z_{d,i}$

$p(z_{di} | z_{*,-i}) = \propto (c(d, k) - 1 + \alpha_k) *$topic-word counts
Outline

Recap

Monte Carlo methods

Sampling Techniques
  Uniform sampling
  Importance Sampling
  Rejection Sampling
  Metropolis-Hastings
  Gibbs sampling

Example: Collapsed Gibbs Sampler for Topic Models