Approximate Inference: Sampling Methods

CMSC 691
UMBC
(Some) Learning Techniques

MAP/MLE: Point estimation, basic EM

Variational Inference: Functional Optimization

Sampling/Monte Carlo
Outline

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 $p$

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

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

Q: Why might sampling from $p(x)$ be hard?
Two Problems for Sampling Methods to Solve

Generate samples from $p$

$$p(x) = \frac{u(x)}{Z}, x \in \mathbb{R}^D$$
$$x_1, x_2, ..., x_R \text{ samples}$$

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

A1: Can we evaluate $Z$?

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

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

**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 might sampling from \( p(x) \) be 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 might sampling from \( p(x) \) be 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

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

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

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

**Goal:**

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

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

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

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

\[ p^* x = u x \]
\[ Z^* \Phi = \sum r \Phi(x_r) p^*(x_r) \]
\[ p^*(x) = \frac{u(x)}{Z^*} \]
\[ Z^* = \sum_r u(x_r) \]

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

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

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

sample uniformly: $x_1, x_2, \ldots, x_R$

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

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

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

Goal:

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

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

Monte Carlo methods

Sampling Techniques

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

Example: Collapsed Gibbs Sampler for Topic Models
Importance Sampling

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

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

where \( Q(x) > p(x) \): over-represented
\( x \) where \( Q(x) < p(x) \): under-represented
**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 \( Q \):
\[ x_1, x_2, \ldots, x_R \]

\[ \phi(x) \]

\[ p(x) \]

\[ Q^*(x) \]

\[ x \text{ where } Q(x) > p(x): \text{ over-represented} \]
\[ x \text{ where } Q(x) < p(x): \text{ under-represented} \]

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

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

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

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

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

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

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

Q: How reliable will this estimator be?

ITILA, Fig 29.5
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.
Importance Sampling

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

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

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

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

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

Goal:

**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?
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

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

ITILA, Fig 29.5
Importance Sampling:
Variance Estimator may vary

$q(x)$: Gaussian

$q(x)$: Cauchy distribution

ITILA, Fig 29.6
Outline

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

ITILA, Fig 29.8
Rejection Sampling

approximating distribution:

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

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
Rejection Sampling

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

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

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

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

select tuples

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

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

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

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

ITILA, Fig 29.8
Rejection Sampling

approximating distribution:

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

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

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

Q: How reliable will this estimator be?
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  
\[ \hat{\Phi} = \frac{1}{R} \sum_{r} \phi(x_r) \]

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

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

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

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

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

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

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

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

\[ \theta(t) \xrightarrow{\text{transition kernel}} \theta(t+1) \]
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:
\[ \alpha_k = \frac{u_p(x_k)}{u_p(x^{(t)})} \frac{u_q(x_k)}{u_q(x^{(t)})} \]

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

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

\[ \widehat{\Phi} = \frac{1}{R} \sum_r \phi(x_r) \]
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:
\[ \alpha_k = \frac{u_p(x_k) \cdot u_q(x_k)}{u_p(x^{(t)}) \cdot u_q(x^{(t)})} \]

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

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

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

samples are not independent

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

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

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

Monte Carlo methods

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

Example: Collapsed Gibbs Sampler for Topic Models
**Gibbs Sampling**

Goal:

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

transition kernel/distribution:

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

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 $x_i$

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

Markov blanket of a node $x$ is its parents, children, and children's parents
**Gibbs Sampling**

**Goal:**

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

**Transition kernel/distribution:**

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

\[ x^{(t+1)}_{[5]} \sim p(\cdot|x^{[2]}_{(t-1)}, x^{(t)}_{[3]}, x^{[4]}_{(t-1)}, x^{[6]}_{(t-1)}, x^{(t)}_{[7]}, x^{[8]}_{(t-1)}) \]

Values of just the Markov blanket variables, both old and new.
Gibbs Sampling

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

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

Markov blanket

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

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

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

samples are not independent

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

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

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

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 😊
Collapsed Gibbs Sampling

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

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

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

Warning: collapsing changes the Markov blanket

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

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

Let’s integrate out \( x_4 \)

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

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

Let’s integrate out \( x_4 \)
Collapsed Gibbs Sampling

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

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

Let’s integrate out \( x_4 \)

Next sampled value of current variable

Values of some of the Markov blanket variables, both old and old
### What Are the Trade-offs of CGS?

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Drawbacks</th>
</tr>
</thead>
</table>
| • Collapsing variables removes variables from the model via *integration/marginalization*  
  – Depending on which variables are marginalized out, this could be a drastic reduction  
  – The priors/hyperparams of the collapsed variables still impact the result  
| • The “steps” are less incremental |
What Are the Trade-offs of CGS?

Benefits

• Collapsing variables removes variables from the model via integration/marginalization
  – Depending on which variables are marginalized out, this could be a drastic reduction
  – The priors/hyperparams of the collapsed variables still impact the result
• The “steps” are less incremental

Drawbacks

• Collapsing removes conditional independences in the model
• Math may not be easy
• You may be restricted via conjugacy/other statistical properties
• You still have the drawbacks of sampling
Outline

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)

\[ w^{(d,n)} \sim \text{Discrete} \left( \phi_{z^{(d,n)}} \right) \]
\[ \phi_k \sim \text{Dirichlet} \left( \beta \right) \]
\[ z^{(d,n)} \sim \text{Discrete} \left( \theta^{(d)} \right) \]
\[ \theta^{(d)} \sim \text{Dirichlet} \left( \alpha \right) \]
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 $z_{d,i} \mid w_{d,i}, \{\psi_k\}, \theta_d$

for each topic $k$:
    resample $\psi_k$
Latent Dirichlet Allocation (Blei et al., 2003)

Per-document (unigram) word counts

Per-document (latent) topic usage

Per-topic word usage

\[ w^{(d, n)} \sim \text{Discrete} \left( \phi_{z^{(d, n)}} \right) \]

\[ \phi_k \sim \text{Dirichlet} \left( \beta \right) \]

\[ z^{(d, n)} \sim \text{Discrete} \left( \theta^{(d)} \right) \]

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

integrate these out
Collapsed 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:
    \[ \text{resample } z_{d,i} \mid w_{d,i}, \{ \psi_k \}, \{ z_{*,-i} \} \]

for each topic k:
    \[ \text{resample } \psi_k \]
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

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

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

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

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

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

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

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

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

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

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

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

Collapsed Gibbs Sampling goal:

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

maintain count tables

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

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

$$p(z_{d,i} \mid z_{*,-i}) = \alpha (c(d,k) - 1 + \alpha_k) \text{ *topic-word counts}$$
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 $z_{d,i} \mid w_{d,i}, \{\psi_k\}, \{z_{*,-i}\}$
    reassign topic: $z_{d,i}$

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

Monte Carlo methods

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

Example: Collapsed Gibbs Sampler for Topic Models