CPSC 540: Machine Learning

Conjugate Priors and Monte Carlo Methods

Mark Schmidt

University of British Columbia

Winter 2016

Admin

- Nothing exciting?

Last Time: Bayesian Statistics

- In Bayesian statistics we work with posterior over parameters,

$$p(\theta|x,\alpha,\beta) = \frac{p(x|\theta)p(\theta|\alpha,\beta)}{p(x|\alpha,\beta)}.$$

Last Time: Bayesian Statistics

- In Bayesian statistics we work with posterior over parameters,

$$p(\theta|x,\alpha,\beta) = \frac{p(x|\theta)p(\theta|\alpha,\beta)}{p(x|\alpha,\beta)}.$$

- We discussed empirical Bayes, where you optimize prior using marginal likelihood,

$$\underset{\alpha,\beta}{\text{argmax}}\, p(x|\alpha,\beta) = \underset{\alpha,\beta}{\text{argmax}} \int_\theta p(x|\theta)p(\theta|\alpha,\beta)d\theta.$$

  - Can be used to optimize λ_j, polynomial degree, RBF σ_i, polynomial vs. RBF, etc.

Last Time: Bayesian Statistics

- In Bayesian statistics we work with posterior over parameters,

$$p(\theta|x,\alpha,\beta) = \frac{p(x|\theta)p(\theta|\alpha,\beta)}{p(x|\alpha,\beta)}.$$

- We discussed empirical Bayes, where you optimize prior using marginal likelihood,

$$\underset{\alpha,\beta}{\text{argmax}}\, p(x|\alpha,\beta) = \underset{\alpha,\beta}{\text{argmax}} \int_\theta p(x|\theta)p(\theta|\alpha,\beta)d\theta.$$

  - Can be used to optimize λ_j, polynomial degree, RBF σ_i, polynomial vs. RBF, etc.
- We also considered hierarchical Bayes, where you put a prior on the prior,

$$p(\alpha,\beta|x,\gamma) = \frac{p(x|\alpha,\beta)p(\alpha,\beta|\gamma)}{p(x|\gamma)}.$$

  - But is the hyper-prior really needed?

Hierarchical Bayes as Graphical Model

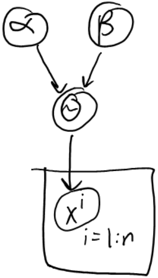
- Let x^i be a binary variable, representing if treatment works on patient i,

$$x^i \sim \mathsf{Ber}(\theta).$$

- As before, let's assume that θ comes from a beta distribution,

$$\theta \sim \mathcal{B}(\alpha, \beta).$$

- We can visualize this as a graphical model:

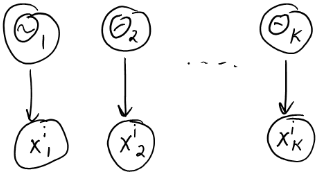


Hierarchical Bayes for Non-IID Data

- Now let x^i represent if treatment works on patient i in hospital j.
- Let's assume that treatment depends on hospital,

$$x_j^i \sim \text{Ber}(\theta_j).$$

- The x_j^i are IID given the hospital.



Hierarchical Bayes for Non-IID Data

- Now let x^i represent if treatment works on patient i in hospital j.
- Let's assume that treatment depends on hospital,

$$x_j^i \sim \mathrm{Ber}(\theta_j).$$

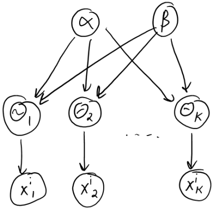
- The x_j^i are IID given the hospital.
- But we may have more data for some hospitals than others:
  - Can we use data from one hospital to learn about others?
  - Can we say anything about a hospital with no data?

Hierarchical Bayes for Non-IID Data

- Common approach: assume θ_j drawn from common prior,

$$\theta_j \sim \mathcal{B}(\alpha, \beta).$$

- This ties the parameters from the different hospitals together:

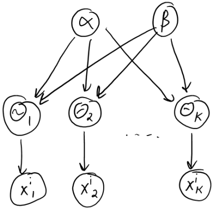


Hierarchical Bayes for Non-IID Data

- Common approach: assume θ_j drawn from common prior,

$$\theta_j \sim \mathcal{B}(\alpha, \beta).$$

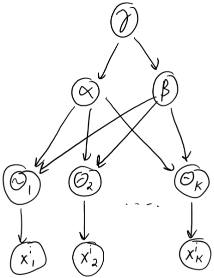
- This ties the parameters from the different hospitals together:



- But, if you fix α and β then you can't learn across hospitals:
  - The θ_j and d-separated given α and β.

Hierarchical Bayes for Non-IID Data

- If α and β are random variables and you use a hyperprior:

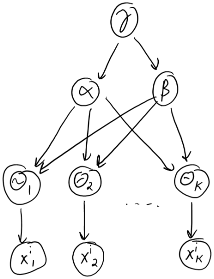


  - You can now consider posterior over both types of variables given data and γ:

$$p(\theta, \alpha, \beta | x, \gamma).$$

Hierarchical Bayes for Non-IID Data

- If α and β are random variables and you use a hyperprior:



  - You can now consider posterior over both types of variables given data and γ:

$$p(\theta, \alpha, \beta | x, \gamma).$$

  - Now there is a dependency between the different θ_j.
  - You combine the non-IID data across different hospitals.
  - Data-rich hospitals inform posterior for data-poor hospitals.
  - You even consider the posterior for new hospitals.

Outline

Conjugate Priors

- Bayesian framework gives simple solutions to several difficult problems:
  - Estimating uncertainty, choosing many hyper-parameters, handling non-IID data, etc.

Conjugate Priors

- Bayesian framework gives simple solutions to several difficult problems:
  - Estimating uncertainty, choosing many hyper-parameters, handling non-IID data, etc.
- But it often requires:
  1. Representing high-dimensional distributions.
  2. Solving high-dimensional integration problems.

Conjugate Priors

- Bayesian framework gives simple solutions to several difficult problems:
  - Estimating uncertainty, choosing many hyper-parameters, handling non-IID data, etc.
- But it often requires:
  1. Representing high-dimensional distributions.
  2. Solving high-dimensional integration problems.
- We've seen this is possible in some special cases:
  - Bernoulli likelihood with discrete prior gives discrete posterior ($\theta = 0.5$ or $\theta = 1$).
  - Bernoulli likelihood with beta prior gives beta posterior.
  - Gaussian likelihood with Gaussian prior gives Gaussian posterior (linear regression).

Conjugate Priors

- Bayesian framework gives simple solutions to several difficult problems:
  - Estimating uncertainty, choosing many hyper-parameters, handling non-IID data, etc.
- But it often requires:
  1. Representing high-dimensional distributions.
  2. Solving high-dimensional integration problems.
- We've seen this is possible in some special cases:
  - Bernoulli likelihood with discrete prior gives discrete posterior ($\theta = 0.5$ or $\theta = 1$).
  - Bernoulli likelihood with beta prior gives beta posterior.
  - Gaussian likelihood with Gaussian prior gives Gaussian posterior (linear regression).
- These are easy because the posterior is in the same 'family' as the prior:
  - This is called a conjugate prior to the likelihood.

Conjugate Priors

- Basic idea of conjugate priors:

$$x \sim D(\theta), \quad \theta \sim P(\lambda) \quad \Rightarrow \quad \theta \mid x \sim P(\lambda').$$

Conjugate Priors

- Basic idea of conjugate priors:

$$x \sim D(\theta), \quad \theta \sim P(\lambda) \quad \Rightarrow \quad \theta \mid x \sim P(\lambda').$$

- Beta-bernoulli example:

$$x \sim \text{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta), \quad \Rightarrow \quad \theta \mid x \sim \mathcal{B}(\alpha', \beta'),$$

and in particular if we see h heads and t tails then the posterior is $\mathcal{B}(h + \alpha, t + \beta)$.

Conjugate Priors

- Basic idea of conjugate priors:

$$x \sim D(\theta), \quad \theta \sim P(\lambda) \quad \Rightarrow \quad \theta \mid x \sim P(\lambda').$$

- Beta-bernoulli example:

$$x \sim \text{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta), \quad \Rightarrow \quad \theta \mid x \sim \mathcal{B}(\alpha', \beta'),$$

  and in particular if we see h heads and t tails then the posterior is $\mathcal{B}(h + \alpha, t + \beta)$.
- Gaussian-Gaussian example:

$$x \sim \mathcal{N}(\mu, \Sigma), \quad \mu \sim \mathcal{N}(\mu_0, \Sigma_0), \quad \Rightarrow \quad \mu \mid x \sim \mathcal{N}(\mu', \Sigma'),$$

  and posterior predictive is also a Gaussian.

Conjugate Priors

- Basic idea of conjugate priors:

$$x \sim D(\theta), \quad \theta \sim P(\lambda) \quad \Rightarrow \quad \theta \mid x \sim P(\lambda').$$

- Beta-bernoulli example:

$$x \sim \text{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta), \quad \Rightarrow \quad \theta \mid x \sim \mathcal{B}(\alpha', \beta'),$$

  and in particular if we see h heads and t tails then the posterior is $\mathcal{B}(h + \alpha, t + \beta)$.
- Gaussian-Gaussian example:

$$x \sim \mathcal{N}(\mu, \Sigma), \quad \mu \sim \mathcal{N}(\mu_0, \Sigma_0), \quad \Rightarrow \quad \mu \mid x \sim \mathcal{N}(\mu', \Sigma'),$$

  and posterior predictive is also a Gaussian.
- If Σ is also a random variable:
  - Conjugate prior is normal-inverse-Wishart.
  - Posterior predictive is a student t.

Conjugate Priors

- Basic idea of conjugate priors:

$$x \sim D(\theta), \quad \theta \sim P(\lambda) \quad \Rightarrow \quad \theta \mid x \sim P(\lambda').$$

- Beta-bernoulli example:

$$x \sim \text{Ber}(\theta), \quad \theta \sim \mathcal{B}(\alpha, \beta), \quad \Rightarrow \quad \theta \mid x \sim \mathcal{B}(\alpha', \beta'),$$

  and in particular if we see h heads and t tails then the posterior is $\mathcal{B}(h + \alpha, t + \beta)$.
- Gaussian-Gaussian example:

$$x \sim \mathcal{N}(\mu, \Sigma), \quad \mu \sim \mathcal{N}(\mu_0, \Sigma_0), \quad \Rightarrow \quad \mu \mid x \sim \mathcal{N}(\mu', \Sigma'),$$

  and posterior predictive is also a Gaussian.
- If Σ is also a random variable:
  - Conjugate prior is normal-inverse-Wishart.
  - Posterior predictive is a student t.
- For the conjugate priors of many standard distributions, see:

  `https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions`

Existence of Conjugate Priors

- Conjugate priors make Bayesian inference easier:
  - Posterior involves updating parameters of prior.
  - Marginal likelihood has closed form as ratio of normalizing constants.
  - In many cases posterior predictive also has a nice form.

Existence of Conjugate Priors

- Conjugate priors make Bayesian inference easier:
  - Posterior involves updating parameters of prior.
  - Marginal likelihood has closed form as ratio of normalizing constants.
  - In many cases posterior predictive also has a nice form.
- Do conjugate priors always exist?
  - No, only exist for exponential family likelihoods.
  - If you aren't in the exponential family (e.g., student t), Bayesian inference gets ugly.

Exponential Family

- Exponential family distributions can be written in the form

$$p(x|\theta) \propto h(x)\exp(\theta^T\phi(x)).$$

- We often have $h(x) = 1$, and $\phi(x)$ are called the sufficient statistics.
  - If you have $\phi(x)$ for a dataset x, you don't need data $x^1, x^2, \dots, x^n$.

Exponential Family

- Exponential family distributions can be written in the form

$$p(x|\theta) \propto h(x)\exp(\theta^T\phi(x)).$$

- We often have $h(x) = 1$, and $\phi(x)$ are called the sufficient statistics.
  - If you have $\phi(x)$ for a dataset x, you don't need data $x^1, x^2, \ldots, x^n$.
- If $\phi(x) = x$, we say that the θ are the cannonical parameters.
  - For Bernoulli, write it as

$$\begin{aligned}
p(x|\pi) = \pi^x(1-\pi)^{1-x} &= \exp(\log(\pi^x(1-\pi)^{1-x})) \\
&= \exp(x\log\pi + (1-x)\log(1-\pi)) \\
&= \exp\left(x\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right) \\
&\propto \exp\left(x\log\left(\frac{\pi}{1-\pi}\right)\right),
\end{aligned}$$

Exponential Family

- Exponential family distributions can be written in the form

$$p(x|\theta) \propto h(x)\exp(\theta^T\phi(x)).$$

- We often have $h(x) = 1$, and $\phi(x)$ are called the sufficient statistics.
  - If you have $\phi(x)$ for a dataset x, you don't need data $x^1, x^2, \ldots, x^n$.
- If $\phi(x) = x$, we say that the θ are the cannonical parameters.
  - For Bernoulli, write it as

$$\begin{aligned} p(x|\pi) &= \pi^x(1-\pi)^{1-x} = \exp(\log(\pi^x(1-\pi)^{1-x})) \\ &= \exp(x\log\pi + (1-x)\log(1-\pi)) \\ &= \exp\left(x\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right) \\ &\propto \exp\left(x\log\left(\frac{\pi}{1-\pi}\right)\right), \end{aligned}$$

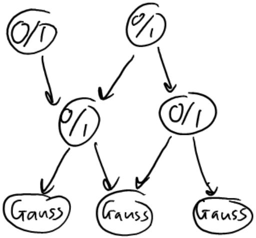
  and parameterize in terms of log-odds, $\theta = \log(\pi/(1-\pi))$.
  (solve for π using sigmoid function, $\pi = 1/(1+\exp(-\theta))$)

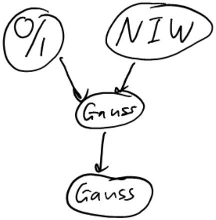
Conjugate Graphical Models

- Discrete priors are "conjugate" to all likelihoods:
  - Posterior will be discrete, although it still might be NP-hard to use.

Conjugate Graphical Models

- Discrete priors are “conjugate” to all likelihoods:
  - Posterior will be discrete, although it still might be NP-hard to use.
- Conjugacy also helps in more complex situations.
- Consider DAGs where marginal of parent is conjugate prior for child:
  - Unconditional inference and sampling will be easy.
- Examples:
  - Gaussian graphical models.
  - Discrete graphical models.
  - Hybrid Gaussian/discrete, where discrete nodes can’t have Gaussian parents.
  - Gaussian graphical model with normal-inverse-Wishart parents.





Outline

Need for Approximate Integration

- Posterior often doesn't have a closed-form expression.
  - We don't just want to flip coins and multiply Gaussians.
  - You can use mixtures of conjugate priors, but we'll consider a different approach.
- Can we approximate the posterior with a simpler closed-form distribution?

Need for Approximate Integration

- Posterior often doesn't have a closed-form expression.
  - We don't just want to flip coins and multiply Gaussians.
  - You can use mixtures of conjugate priors, but we'll consider a different approach.
- Can we approximate the posterior with a simpler closed-form distribution?
- Two main strategies:
  1. Variational methods.
  2. Monte Carlo methods.
- Both are classic ideas from statistical physics, but in 90s revolutionized Bayesian stats/ML.
- Also used extensively in graphical models and deep learning.

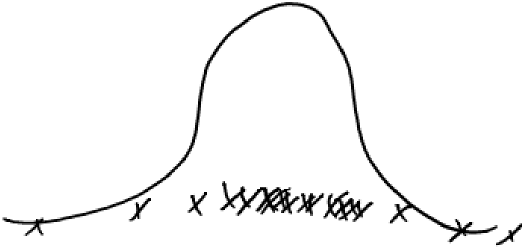
Monte Carlo Methods

- Our goal is to approximate a probability distribution $p(x)$ with a simpler distribution.
    - This could be a posterior distribution or a graphical model or a deep belief network.

Monte Carlo Methods

- Our goal is to approximate a probability distribution $p(x)$ with a simpler distribution.
  - This could be a posterior distribution or a graphical model or a deep belief network.
- Basic idea between Monte Carlo methods:
  1. Generate n samples proportional to $p(x)$,

$$x^i \sim p$$



Monte Carlo Methods

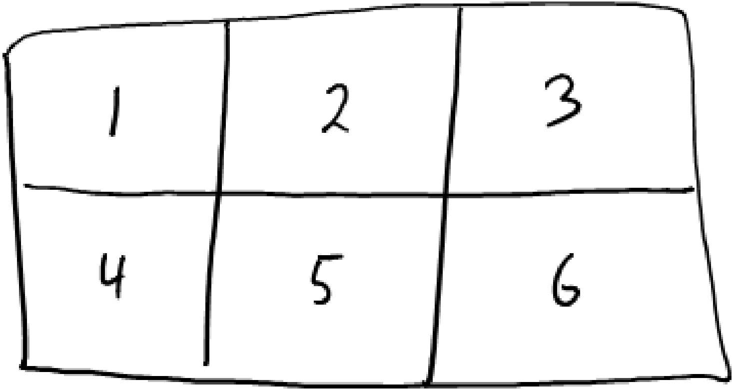
- Our goal is to approximate a probability distribution $p(x)$ with a simpler distribution.
  - This could be a posterior distribution or a graphical model or a deep belief network.
- Basic idea between Monte Carlo methods:
  1. Generate n samples proportional to $p(x)$,
  $$x^i \sim p$$
  2. Use these samples as an approximation of the distribution.
  $$p(x) \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[x = x^i].$$

  (Area where you have more samples means higher probaiblity.)

Monte Carlo Methods

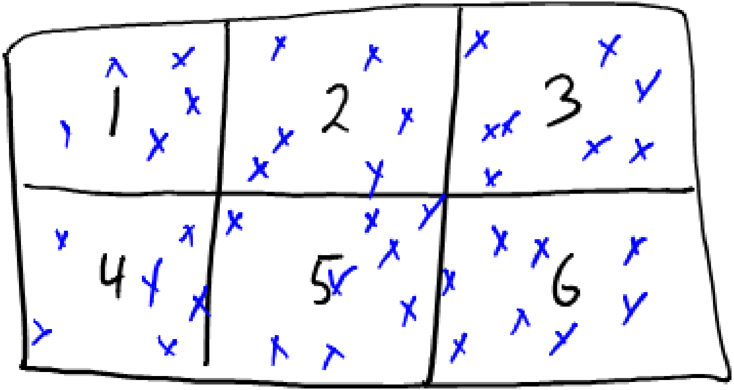
- Our goal is to approximate a probability distribution $p(x)$ with a simpler distribution.
  - This could be a posterior distribution or a graphical model or a deep belief network.
- Basic idea between Monte Carlo methods:
  1. Generate n samples proportional to $p(x)$,
  $$x^i \sim p$$
  2. Use these samples as an approximation of the distribution.
  $$p(x) \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[x = x^i].$$
  (Area where you have more samples means higher probaiblity.)
- As $n \to \infty$, "converges" to the true distribution.
- We can use this "empirical measure" to approximate the original probability.
  - E.g., if you want $\mathbb{E}[f(x)]$, compute $\frac{1}{n}\sum_{i=1}^{n} f(x)$.
  - Converges to expectation as $n \to \infty$ by law of large numbers.

Monte Carlo Methods Example: Rolling di



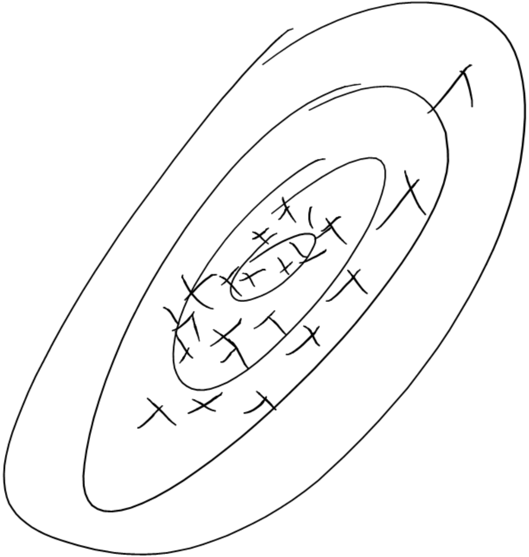
Probability of event: (number of samples consistent with event)/(number of sampes)

Monte Carlo Methods Example: Rolling di



Probability of event: (number of samples consistent with event)/(number of sampes)

Monte Carlo Methods Example: Gaussian distribution



Overview of Monte Carlo Methods

- We'll assume you have a way to sample uniformly over $[0, 1]$.
    - Usually, a "pseudo-random" number generator is good enough.
    - E.g., Matlab's *rand* function.

Overview of Monte Carlo Methods

- We'll assume you have a way to sample uniformly over $[0, 1]$.
  - Usually, a "pseudo-random" number generator is good enough.
  - E.g., Matlab's *rand* function.
- First class of Monte Carlo method generate independent samples:
  1. Inverse transform and ancestral sampling.
  2. Rejection and importance sampling.

Overview of Monte Carlo Methods

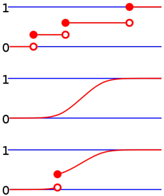
- We'll assume you have a way to sample uniformly over $[0, 1]$.
  - Usually, a "pseudo-random" number generator is good enough.
  - E.g., Matlab's *rand* function.
- First class of Monte Carlo method generate independent samples:
  1. Inverse transform and ancestral sampling.
  2. Rejection and importance sampling.
- Second class of Monte Carlo methods generate dependent samples:
  1. Markov chain Monte Carlo.
     - Gibbs sampling, Metropolis-Hastings.
  2. Sequential Monte Carlo.
     - AKA sequential importance sampling or particle filtering.

Inverse Transform Method (Exact 1D Sampling)

- Recall that we're using $p(x)$ as a short way to write $p(X = x)$:
  - Probability that random variable X has the value x.

Inverse Transform Method (Exact 1D Sampling)

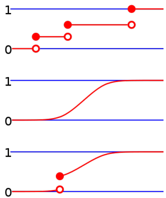
- Recall that we're using $p(x)$ as a short way to write $p(X = x)$:
  - Probability that random variable X has the value x.
- The cumulative distribution function (CDF) F is $p(X \leq x)$.
  - $F(x)$ is between 0 and 1 a gives proportion of times X is below x.



https://en.wikipedia.org/wiki/Cumulative_distribution_function

Inverse Transform Method (Exact 1D Sampling)

- Recall that we're using $p(x)$ as a short way to write $p(X = x)$:
  - Probability that random variable X has the value x.
- The cumulative distribution function (CDF) F is $p(X \leq x)$.
  - $F(x)$ is between 0 and 1 a gives proportion of times X is below x.

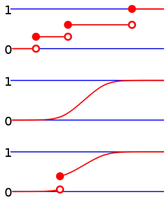


https://en.wikipedia.org/wiki/Cumulative_distribution_function

- The inverse CDF (or quantile function) F^{-1} is its inverse:
  - Given a number u between 0 and 1, gives x such that $p(X \leq x) = u$.

Inverse Transform Method (Exact 1D Sampling)

- Recall that we're using $p(x)$ as a short way to write $p(X = x)$:
  - Probability that random variable X has the value x.
- The cumulative distribution function (CDF) F is $p(X \leq x)$.
  - $F(x)$ is between 0 and 1 a gives proportion of times X is below x.



https://en.wikipedia.org/wiki/Cumulative_distribution_function

- The inverse CDF (or quantile function) F^{-1} is its inverse:
  - Given a number u between 0 and 1, gives x such that $p(X \leq x) = u$.
- Inverse transfrom method for exact sampling in 1D:
  1. Sample $u \sim \mathcal{U}(0, 1)$.
  2. Compute $x = F^{-1}(u)$.

Inverse Transform Method (Exact 1D Sampling)

- Consider a discrete distribution:

$$p(X = 1) = 0.4, \quad p(X = 2) = 0.1, \quad p(X = 3) = 0.2, \quad p(X = 4) = 0.3.$$

Inverse Transform Method (Exact 1D Sampling)

- Consider a discrete distribution:

$$p(X=1)=0.4, \quad p(X=2)=0.1, \quad p(X=3)=0.2, \quad p(X=4)=0.3.$$

- Inverse transform method:
  1. Generate $u \sim \mathcal{U}(0,1)$.
  2. If $u \leq p(X=1)$, output 1.
  3. If $u \leq p(X=1)+p(X=2)$, output 2.
  4. If $u \leq p(X=1)+p(X=2)+p(X=3)$, output 3.
  5. Otherwise, output 4.

Inverse Transform Method (Exact 1D Sampling)

- Consider a discrete distribution:

$$p(X=1)=0.4, \quad p(X=2)=0.1, \quad p(X=3)=0.2, \quad p(X=4)=0.3.$$

- Inverse transform method:
  1. Generate $u \sim \mathcal{U}(0,1)$.
  2. If $u \leq p(X=1)$, output 1.
  3. If $u \leq p(X=1)+p(X=2)$, output 2.
  4. If $u \leq p(X=1)+p(X=2)+p(X=3)$, output 3.
  5. Otherwise, output 4.
- With k states, cost to generate a sample is $O(k)$.

Inverse Transform Method (Exact 1D Sampling)

- Consider a discrete distribution:

$$p(X = 1) = 0.4, \quad p(X = 2) = 0.1, \quad p(X = 3) = 0.2, \quad p(X = 4) = 0.3.$$

- Inverse transform method:
  1. Generate $u \sim \mathcal{U}(0, 1)$.
  2. If $u \leq p(X = 1)$, output 1.
  3. If $u \leq p(X = 1) + p(X = 2)$, output 2.
  4. If $u \leq p(X = 1) + p(X = 2) + p(X = 3)$, output 3.
  5. Otherwise, output 4.
- With k states, cost to generate a sample is $O(k)$.
- If you are generating multiple samples, store the sums and do binary search:
  - $O(k)$ pre-processing cost, then $O(\log k)$ cost per sample.

Inverse Transform Method (Exact 1D Sampling)

- Consider a Gaussian distribution,

$$x \sim \mathcal{N}(\mu, \sigma^2).$$

- CDF has the form

$$F(x) = p(X \leq x) = \frac{1}{2}\left[1 + \mathsf{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right],$$

where erf the CDF of $\mathcal{N}(0, 1)$.

Inverse Transform Method (Exact 1D Sampling)

- Consider a Gaussian distribution,

$$x \sim \mathcal{N}(\mu, \sigma^2).$$

- CDF has the form

$$F(x) = p(X \leq x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right],$$

  where erf the CDF of $\mathcal{N}(0, 1)$.
- Inverse CDF has the form

$$F^{-1}(u) = \mu + \sigma\sqrt{2}\text{erf}^{-1}(2u - 1).$$

- To sample from a Gaussian:
  1. Generate $u \sim \mathcal{U}(0, 1)$.
  2. Compute $F^{-1}(u)$.

Ancestral Sampling (Exact Multidimensional Sampling)

- We've seen already for DAG models.
- If you want to sample from $p(x_1, x_2, x_3)$,
  - Sample x_1 from $p(x_1)$.
  - Using x_1, sample x_2 from $p(x_2|x_1)$.
  - Using x_1 and x_2, sample x_3 from $p(x_3|x_1, x_2)$.

Ancestral Sampling (Exact Multidimensional Sampling)

- We've seen already for DAG models.
- If you want to sample from $p(x_1, x_2, x_3)$,
  - Sample x_1 from $p(x_1)$.
  - Using x_1, sample x_2 from $p(x_2|x_1)$.
  - Using x_1 and x_2, sample x_3 from $p(x_3|x_1, x_2)$.
- If children are conjuate to parents this is easy.
  - You might be able to build distribution out of conjugate parts.
- For non-conjugate models, hard to characterize all these conditionals.

Beyond Inverse Transform and Conjugacy

- We can't sample exactly from many distributions.
- But, we can use simple distributions to sample from complex distributions.

Beyond Inverse Transform and Conjugacy

- We can't sample exactly from many distributions.
- But, we can use simple distributions to sample from complex distributions.
- Method 1: Rejection sampling.
  - Example: sampling from a Gaussian subject to $x \in [-1, 1]$.

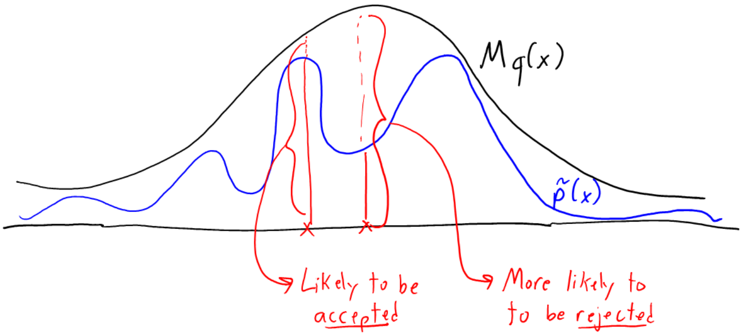
Rejection Sampling

- Ingredients of rejection sampling:
  1. Ability to evaluate unnormalized $\tilde{p}(x)$,
  $$p(x) = \frac{\tilde{p}(x)}{Z}.$$
  2. A distribution q that is easy to sample from.
  3. An upper bound M on $\tilde{p}(x)/q(x)$.

Rejection Sampling

- Ingredients of rejection sampling:
  1. Ability to evaluate unnormalized $\tilde{p}(x)$,
  $$p(x) = \frac{\tilde{p}(x)}{Z}.$$
  2. A distribution q that is easy to sample from.
  3. An upper bound M on $\tilde{p}(x)/q(x)$.
- Rejection sampling algorithm:
  1. Sample x from $q(x)$.
  2. Sample u from $\mathcal{U}(0, 1)$.
  3. Keep the sample if $u \leq \frac{\tilde{p}(x)}{Mq(x)}$.
- The accepted samples will be from $p(x)$.

Rejection Sampling



Rejection Sampling

- Examples
  - Sample from Gaussian q to sample from student t.
  - Sample from prior to sample from posterior ($M = 1$),

$$p(\theta|x) = \underbrace{p(x|\theta)}_{\leq 1} p(\theta).$$

Rejection Sampling

- Examples
  - Sample from Gaussian q to sample from student t.
  - Sample from prior to sample from posterior ($M = 1$),

$$p(\theta|x) = \underbrace{p(x|\theta)}_{\leq 1} p(\theta).$$

- Drawbacks:
  - You may reject a large number of samples.
    - Most samples are rejected for high-dimensional complex distributions.
  - You need to know M.

Rejection Sampling

- Examples
  - Sample from Gaussian q to sample from student t.
  - Sample from prior to sample from posterior ($M = 1$),

$$p(\theta|x) = \underbrace{p(x|\theta)}_{\leq 1} p(\theta).$$

- Drawbacks:
  - You may reject a large number of samples.
    - Most samples are rejected for high-dimensional complex distributions.
  - You need to know M.
- Extension in 1D for convex $-\log p(x)$:
  - Adaptive rejection sampling refines q after each rejection.

Importance Sampling

- Importance sampling is a variation that accepts all samples.

Importance Sampling

- Importance sampling is a variation that accepts all samples.
  - Key idea is similar to EM,

$$\begin{aligned}\mathbb{E}_p[f(x)] &= \sum_x p(x)f(x)\\ &= \sum_x q(x)\frac{p(x)f(x)}{q(x)}\\ &= \mathbb{E}_q\left[\frac{p(x)}{q(x)}f(x)\right],\end{aligned}$$

  and similarly for continuous distributions.

Importance Sampling

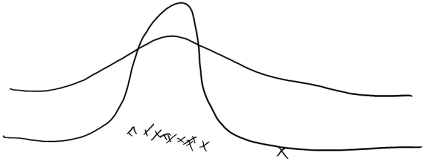
- Importance sampling is a variation that accepts all samples.
  - Key idea is similar to EM,

$$\begin{aligned}\mathbb{E}_p[f(x)] &= \sum_x p(x)f(x) \\ &= \sum_x q(x)\frac{p(x)f(x)}{q(x)} \\ &= \mathbb{E}_q\left[\frac{p(x)}{q(x)}f(x)\right],\end{aligned}$$

    and similarly for continuous distributions.
  - We can sample from q, and reweight by $p(x)/q(x)$ to sample from p.
  - Only assumption is that q is non-zero when p is non-zero.
  - If you only know unnormalized $\tilde{p}(x)$, variant gives approximation of Z.

Importance Sampling

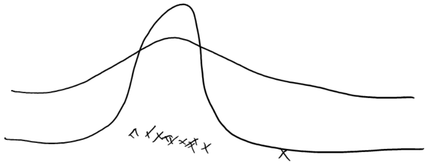
- As with rejection sampling, only efficient if q is close to p.
- Otherwise, weights will be huge for a small number of samples.
  - Even though unbiased, variance will be huge.



- In high-dimensions, these methods tend not to work well.

Importance Sampling

- As with rejection sampling, only efficient if q is close to p.
- Otherwise, weights will be huge for a small number of samples.
  - Even though unbiased, variance will be huge.



- In high-dimensions, these methods tend not to work well.
- For high dimensions, we often resort to methods based on dependent samples:
  1. Markov chain Monte Carlo.
     - Gibbs sampling, Metropolis-Hastings.
  2. Sequential Monte Carlo.
     - AKA sequential importance sampling or particle filtering.

Summary

- Hierarchical Bayesian models are useful in non-standard scenarios (non-IID).

Summary

- Hierarchical Bayesian models are useful in non-standard scenarios (non-IID).
- Conjugate priors are priors that lead to posteriors in the same family.
  - They make Bayesian inference much easier.

Summary

- Hierarchical Bayesian models are useful in non-standard scenarios (non-IID).
- Conjugate priors are priors that lead to posteriors in the same family.
    - They make Bayesian inference much easier.
- Exponential family distributions are the only distributions with conjugate priors.

Summary

- Hierarchical Bayesian models are useful in non-standard scenarios (non-IID).
- Conjugate priors are priors that lead to posteriors in the same family.
  - They make Bayesian inference much easier.
- Exponential family distributions are the only distributions with conjugate priors.
- Monte Carlo methods approximate distributions by samples.

Summary

- Hierarchical Bayesian models are useful in non-standard scenarios (non-IID).
- Conjugate priors are priors that lead to posteriors in the same family.
  - They make Bayesian inference much easier.
- Exponential family distributions are the only distributions with conjugate priors.
- Monte Carlo methods approximate distributions by samples.
- Inverse transform generates exact samples based on uniform samples.

Summary

- Hierarchical Bayesian models are useful in non-standard scenarios (non-IID).
- Conjugate priors are priors that lead to posteriors in the same family.
  - They make Bayesian inference much easier.
- Exponential family distributions are the only distributions with conjugate priors.
- Monte Carlo methods approximate distributions by samples.
- Inverse transform generates exact samples based on uniform samples.
- Rejection sampling and importance sampling use other distributions.

- Next time: MCMC, non-parametric Bayes, and the Automatic Statistician.