# Basic Sampling Methods

Sargur Srihari srihari@cedar.buffalo.edu

### **Topics**

1. Motivation

Intractability in ML
How sampling can help

- Ancestral Sampling Using BNs
- 3. Transforming a Uniform Distribution
- 4. Rejection Sampling
- 5. Importance Sampling
- 6. Sampling-Importance-Resampling

#### 1. Motivation

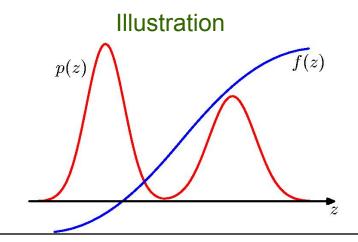
- When exact inference is intractable, we need some form of approximation
  - True of probabilistic models of practical significance
- Inference methods based on numerical sampling are known as Monte Carlo techniques
- Most situations will require evaluating expectations of unobserved variables, e.g., to make predictions
  - Rather than the posterior distribution

#### Common Task

- Find expectation of some function f(z)
  - with respect to a probability distribution p(z)
  - Components of z can be discrete, continuous or combination
- In case of continuous we wish to evaluate

$$E[f] = \int f(z)p(z)dz$$

 In discrete case, integral replaces by summation



In Bayesian regression, From sum and product rules  $p(t \mid x) = \int p(t \mid x, w) p(w) dw$ 

where  $p(t | x, w) \sim N(y(x, w), \beta^{-1})$ 

In general such expectations are too complex to be evaluated analytically

# Sampling: main idea

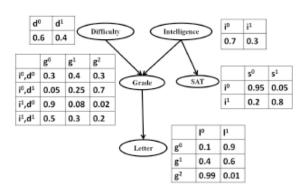
p(z) f(z)

- Obtain set of samples  $z^{(l)}$  where i = 1,..., L
- Drawn independently from distribution p(z)
- Allows expectation  $E[f] = \int f(z)p(z)dz$  to be approximated by  $\hat{f} = \frac{1}{L}\sum_{i=1}^{L} f(z^{(i)})$  Called an estimator
- Then  $E[\hat{f}] = E[f]$ , i.e., estimator has the correct mean
  - And  $\operatorname{var}[\hat{f}] = \frac{1}{L} E[(f E(f))^2]$  which is the variance of the estimator
- Accuracy independent of dimensionality of z
  - High accuracy can be achieved with few (10 or twenty samples)
- However samples may not be independent
  - Effective sample size may be smaller than apparent sample size
  - In example f(z) is small when p(z) is high and vice versa
    - Expectation may be dominated by regions of small probability thereby requiring large sample sizes

## 2. Ancestral Sampling

- If joint distribution is represented by a BN
  - no observed variables
  - a straightforward method exists
- Distribution is specified by  $p(z) = \prod_{i=1}^{M} p(z_i | pa_i)$ 
  - where  $z_i$  are set of variables associated with node i and
  - pa<sub>i</sub> are set of variables associated with node parents of node i
- To obtain samples from joint
  - we make one pass through set of variables in order  $z_l,...z_M$  sampling from conditional distribution  $p(z|pa_i)$
- After one pass through the graph we obtain one sample
- Frequency of different values defines the distribution
  - E.g., allowing us to determine marginals

$$P(L,S) = \sum_{D,I,G} P(D,I,G,L,S) = \sum_{D,I,G} P(D)P(I)P(G \mid D,I)P(L \mid G)P(S \mid I)$$



# Ancestral sampling with some nodes instantiated

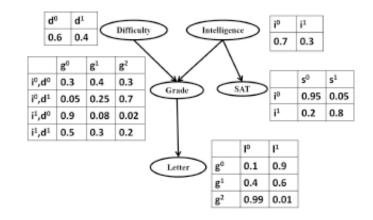
Directed graph where some nodes instantiated

with observed values

$$P(L = l^{0}, s = s^{1}) =$$

$$\sum_{D,I,G} P(D)P(I)P(G \mid D,I) \times$$

$$P(L = l^{0} \mid G)P(S = s^{1} \mid I)$$



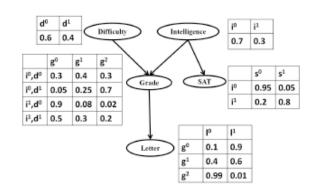
- Called Logic sampling
- Use ancestral sampling, except
  - When sample is obtained for an observed value:
    - if they agree then sample value is retained and proceed to next variable
    - If they don't agree, whole sample is discarded

## Properties of Logic Sampling

$$P(L = l^{0}, s = s^{1}) =$$

$$\sum_{D,I,G} P(D)P(I)P(G \mid D, I) \times$$

$$P(L = l^{0} \mid G)P(S = s^{1} \mid I)$$



- Samples correctly from posterior distribution
  - Corresponds to sampling from joint distribution of hidden and data variables
- But prob. of accepting sample decreases as
  - no of variables increase and
  - number of states that variables can take increases
- A special case of Importance Sampling
  - Rarely used in practice

#### **Undirected Graphs**

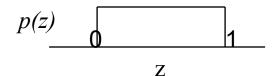
- No one-pass sampling strategy even from prior distribution with no observed variables
- Computationally expensive methods such as Gibbs sampling must be used
  - Start with a sample
  - Replace first variable conditioned on rest of values, then next variable, etc

### 3. Basic Sampling Algorithms

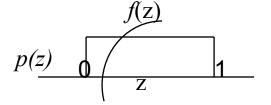
- Simple strategies for generating random samples from a given distribution
  - They will be pseudo random numbers that pass a test for randomness
- Assume that algorithm is provided with a pseudo-random generator for uniform distribution over (0,1)
- For standard distributions we can use transformation method of generating nonuniformly distributed random numbers

#### **Transformation Method**

- Goal: generate random numbers from a simple non-uniform distribution
- Assume we have a source of uniformly distributed random numbers
- z is uniformly distributed over (0,1), i.e., p(z) = 1 in the interval



#### Transformation for Standard Distributions



$$p(y) = f(z)$$

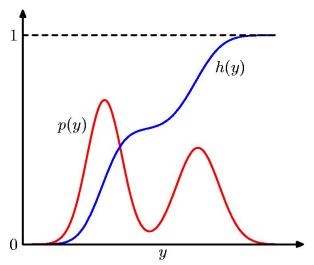
- If we transform values of z using f() such that y=f(z)
- Distribution of y is governed by  $p(y) = p(z) \left| \frac{dz}{dy} \right|$  (1)
- Goal is to choose f(z) such that values of y have distribution p(y)
- Integrating (1) above
  - Since p(z)=1 and integral of dz/dy wrt y is z

$$z = h(y) \equiv \int_{0}^{y} p(\hat{y}) d\hat{y}$$

- which is an indefinite integral of p(y)
- Thus  $y = h^{-1}(z)$
- So we have to transform uniformly distributed random numbers
  - Using function which is inverse of the indefinite integral of the distribution

### Geometry of Transformation

- We are interested in generating r.v. s from p(y)
  - non-uniform random variables



- h(y) is indefinite integral of desired p(y)
- $z \sim U(0,1)$  is transformed using  $y = h^{-1}(z)$
- Results in y being distributed as p(y)

#### Transformations for Exponential & Cauchy

We need samples from the Exponential Distribution

$$p(y) = \lambda \exp(-\lambda y)$$
where  $0 \le y < \infty$ 

In this case the indefinite integral is

$$z = h(y) = \int_{0}^{y} p(\hat{y})d\hat{y}$$
$$= 1 - \exp(-\lambda y)$$

- If we transform using  $y = -\lambda^{-1} \ln(1-z)$
- Then y will have an exponential distribution
- We need samples from the Cauchy Distribution
  - Cauchy Distribution  $p(y) = \frac{1}{\pi} \frac{1}{1+y^2}$
  - Inverse of the indefinite integral can be expressed as a "tan" function

#### Generalization of Transformation Method

Single variable transformation

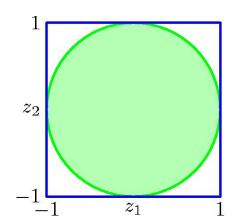
$$p(y) = p(z) \left| \frac{dz}{dy} \right|$$

- Where z is uniformly distributed over (0,1)
- Multiple variable transformation

$$p(y_1,...,y_M) = p(z_1,...,z_M) \left| \frac{d(z_1,...,z_M)}{d(y_1,...,y_M)} \right|$$

#### Transformation Method for Gaussian

- Box-Muller method for Gaussian
  - Example of a bivariate Gaussian
- First generate uniform distribution in unit circle



Generate pairs of uniformly distributed random numbers

$$z_1, z_2 \in (-1, 1)$$

- Can be done from U(0,1) using  $z \rightarrow 2z-1$
- Discard each pair unless  $z_1^2 + z_2^2 \le 1$
- Leads to uniform distribution of points inside unit circle with  $p(z_1, z_2) = \frac{1}{\pi}$

### Generating a Gaussian

• For each pair  $z_1, z_2$  evaluate the quantities

$$\begin{vmatrix} y_1 = z_1 \left( \frac{-2\ln z_1}{r^2} \right)^{1/2} & y_2 = z_2 \left( \frac{-2\ln z_2}{r^2} \right)^{1/2} \\ where & r^2 = z_1^2 + z_2^2 \end{vmatrix}$$

- Then  $y_1$  and  $y_2$  are independent Gaussians with zero mean and variance
- For arbitrary mean and covariance matrix

If 
$$y \sim N(0,1)$$
 then  $\sigma y + \mu$  has  $N(\mu, \sigma^2)$ 

In multivariate case

```
If components are independent and N(0,1)
then \mathbf{y} = \mathbf{\mu} + L\mathbf{z}
will have N(\mathbf{\mu}, \mathbf{\Sigma})
where \mathbf{\Sigma} = LL^T, is called the Cholesky decomposition
```

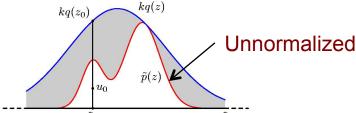
#### Limitation of Transformation Method

- Need to first calculate and then invert indefinite integral of the desired distribution
- Feasible only for a small number of distributions
- Need alternate approaches
- Rejection sampling and Importance sampling applicable to univariate distributions only
  - But useful as components in more general strategies

# 4. Rejection Sampling

- Allows sampling from a relatively complex distribution
- Consider univariate, then extend to several variables
- Wish to sample from distribution p(z)
  - Not a simple standard distribution
  - Sampling from p(z) is difficult
- Suppose we are able to easily evaluate p(z) for any given value of z, upto a normalizing constant Z

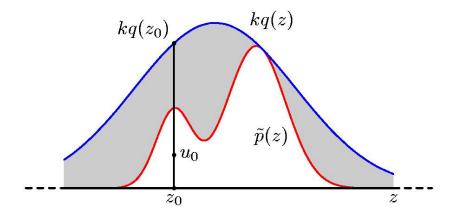
$$p(z) = \frac{1}{Z_p} \tilde{p}(z)$$



- where  $\tilde{p}(z)$  can readily be evaluated but  $Z_p$  is unknown
- e.g.,  $\tilde{p}(z)$  is a mixture of Gaussians
  - Note that we may know the mixture distribution but we need samples to generate expectations

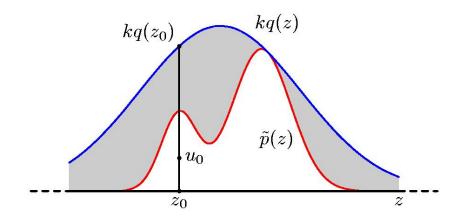
#### Rejection sampling: Proposal distribution

- Samples are drawn from simple distribution, called proposal distribution q(z)
- Introduce constant k whose value is such that  $kq(z) \ge \widetilde{p}(z)$  for all z
  - Called comparison function



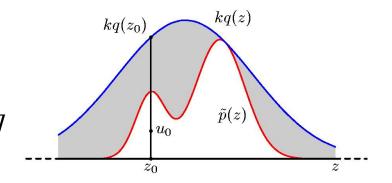
# Rejection Sampling Intuition

- Samples are drawn from simple distribution q(z)
- Rejected if they fall in grey area
  - Between un-normalized distribution  $p^{\sim}(z)$  and scaled distribution kq(z)
- Resulting samples are distributed according to p(z) which is the normalized version of p~(z)



#### Determining if sample is shaded area

- Generate two random numbers
  - $-z_0$  from q(z)
  - $-u_0$  from uniform distribution  $[0,kq(z_0)]$



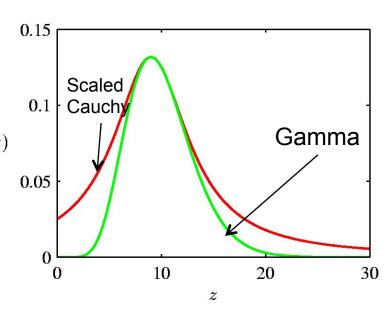
- This pair has uniform distribution under the curve of function kq(z)
- If  $u_0 > p(z_0)$  the pair is rejected otherwise it is retained
- Remaining pairs have a uniform distribution under the curve of p(z) and hence the corresponding z values are distributed according to p(z) as desired

#### Example of Rejection Sampling

Task of sampling from Gamma

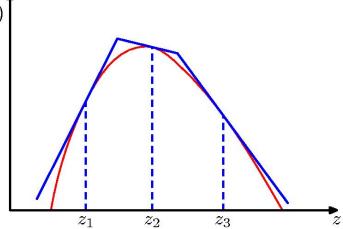
$$Gam(z \mid a, b) = \frac{b^a z^{a-1} \exp(-bz)}{\Gamma(a)}$$

- Since Gamma is roughly bell-shaped, proposal distribution is Cauchy
- Cauchy has to be slightly generalized
  - To ensure it is nowhere smaller than Gamma



## Adaptive Rejection Sampling

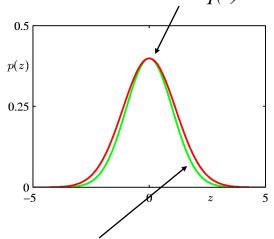
- When difficult to find suitable  $\ln p(z)$  analytic distribution p(z)
- Straight-forward when p(z) is log concave
  - When  $\ln p(z)$  has derivatives that are non-increasing functions of z
  - Function  $\ln p(z)$  and its gradient are evaluated at initial set of grid points
  - Intersections are used to construct envelope
    - A sequence of linear functions



#### Dimensionality and Rejection Sampling

- Gaussian example
- Acceptance rate is ratio of volumes under p(z) and kq(z)
  - diminishesexponentially withdimensionality

Proposal distribution q(z) is Gaussian Scaled version is kq(z)



True distribution p(z)

# 5. Importance Sampling

• Principal reason for sampling p(z) is evaluating

expectation of some f(z)

$$E[f] = \int f(z)p(z)dz$$

In Bayesian regression,  $p(t \mid x) = \int p(t \mid x, w) p(w) dw$  where  $p(t \mid x, w) \sim N(y(x, w), \beta^{-1})$ 

• Given samples  $z^{(l)}$ , l=1,...,L, from p(z), the finite sum approximation is  $\begin{bmatrix} 1 & L \\ c & (l) \end{bmatrix}$ 

$$\hat{f} = \frac{1}{L} \sum_{i=1}^{L} f(\mathbf{z}^{(l)})$$

- But drawing samples p(z) may be impractical
- Importance sampling uses:
  - a proposal distribution—like rejection sampling
    - But all samples are retained
  - Assumes that for any z, p(z) can be evaluated

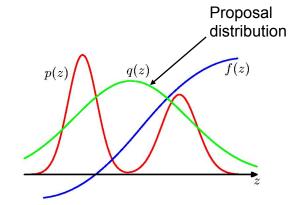
## Determining Importance weights

• Samples  $\{z^{(l)}\}$  are drawn from simpler dist. q(z)

$$E[f] = \int f(z)p(z)dz$$

$$= \int f(z)\frac{p(z)}{q(z)}q(z)dz$$

$$= \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^{(l)})}{q(z^{(l)})}f(z^{(l)})$$



Unlike rejection sampling All of the samples are retained

- Samples are weighted by ratios  $r_l = p(\mathbf{z}^{(l)}) / q(\mathbf{z}^{(l)})$ 
  - Known as importance weights
    - Which corrects the bias introduced by wrong distribution

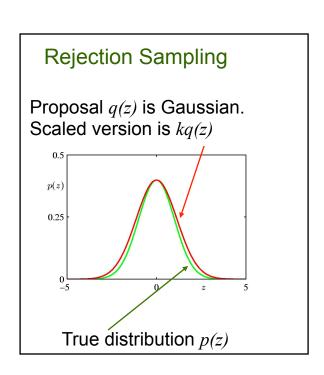
## Likelihood-weighted Sampling

- Importance sampling of graphical model using ancestral sampling
- Weight of sample z given evidence variables E

$$r(\mathbf{z}) = \prod_{\mathbf{z}_i \in E} p(\mathbf{z}_i \mid pa_i)$$

#### 6. Sampling-Importance Re-sampling (SIR)

- Rejection sampling depends on suitable value of k where  $kq(z) \ge p(k)$ 
  - For many pairs of distributions p(z) and q(z) it is impractical to determine value of k
  - If it is sufficiently large to guarantee a bound then impractically small acceptance rates
- Method makes use of sampling distribution q(z) but avoids having to determine k
- Name of method: Sampling proposal distribution followed by determining importance weights and then resampling



#### SIR Method

- Two stages
- Stage 1: L samples z<sup>(1)</sup>,...,z<sup>(L)</sup> are drawn from q(z)
- Stage 2: Weights w<sub>1</sub>,..,w<sub>L</sub> are constructed
  - As in importance sampling  $w_l = p(\mathbf{z}^{(l)}) / q(\mathbf{z}^{(l)})$
- Finally a second set of L samples are drawn from the discrete distribution  $\{\mathbf{z}^{(1)},...,\mathbf{z}^{(L)}\}$  with probabilities given by  $\{w_I,...,w_L\}$
- If moments wrt distribution p(z) are needed, use:

$$E[f] = \int f(z)p(z)dz$$
$$= \sum_{l=1}^{L} w_l f(z^{(l)})$$