Chapter 11: Sampling Methods

Lei Tang

Department of CSE
Arizona State University

Dec. 18th, 2007

Outline

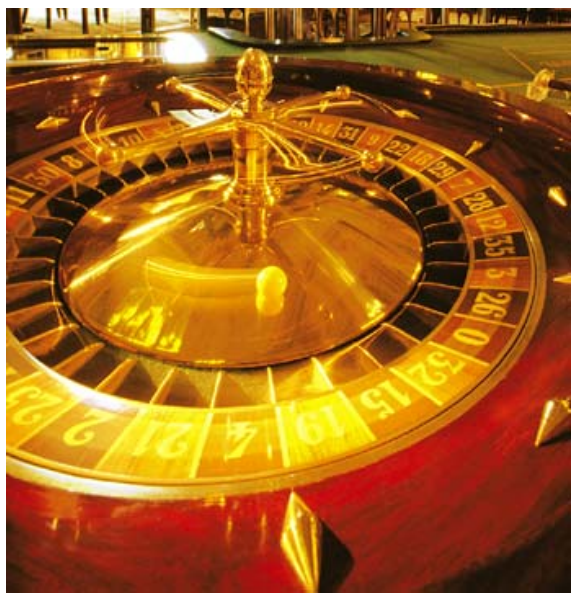
Introduction

- Exact inference is intractable for most probabilistic models of practical interest.
- We've already discussed deterministic approximations including *Variational Bayes* and *Expectation propagation*.
- Here we consider approximation based on *numerical sampling*, also known as Monte Carlo techniques.

What is Monte Carlo?

- *Monte Carlo* is a small hillside town in Monaco (near Italy) with casino since 1865 like Las Vegas.
- Stainslaw Marcin Ulam (Polish Mathematician) named the statistical sampling methods in honor of his uncle, who was a gambler and would borrow money from relatives because he "just had to go to Monte Carlo" (which is suggested by another mathematician Nicholas Metropolis).

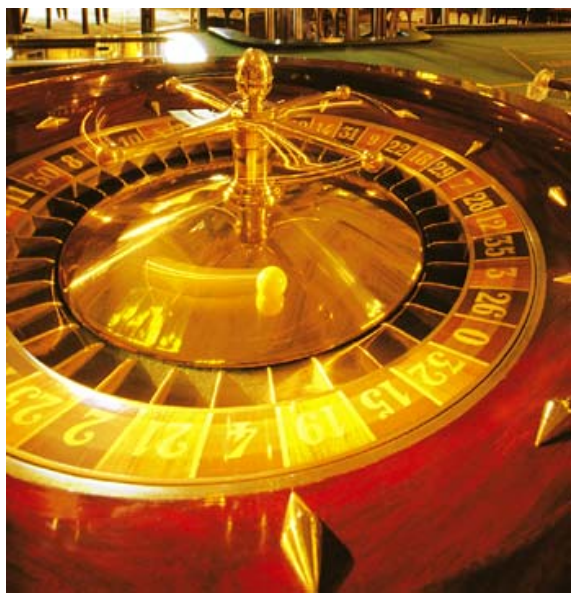
The magic is running dice.



What is Monte Carlo?

- *Monte Carlo* is a small hillside town in Monaco (near Italy) with casino since 1865 like Las Vegas.
- Stainslaw Marcin Ulam (Polish Mathematician) named the statistical sampling methods in honor of his uncle, who was a gambler and would borrow money from relatives because he "just had to go to Monte Carlo" (which is suggested by another mathematician Nicholas Metropolis).

The magic is running dice.



Common Questions

- Why do we need Monte Carlo techniques?
- Isn't it trivial to sample from a probability?
- Are Monte Carlo methods always slow?
- What can Monte Carlo methods do for me?

Common Questions

- Why do we need Monte Carlo techniques?
- Isn't it trivial to sample from a probability?
- Are Monte Carlo methods always slow?
- What can Monte Carlo methods do for me?

Common Questions

- Why do we need Monte Carlo techniques?
- Isn't it trivial to sample from a probability?
- Are Monte Carlo methods always slow?
- What can Monte Carlo methods do for me?

Common Questions

- Why do we need Monte Carlo techniques?
- Isn't it trivial to sample from a probability?
- Are Monte Carlo methods always slow?
- What can Monte Carlo methods do for me?

General Idea of Sampling

- Mostly, the posterior distribution is primarily required for prediction.
- Fundamental problem: find the *expectation* of some function $f(z)$ with respect to a probability $p(z)$.

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

- General idea: obtain a set of samples $z^{(l)}$ drawn independently from the distribution $p(z)$. So we can estimate the expectation:

$$\begin{aligned} \hat{f} &= \frac{1}{L}\sum_{l=1}^{L} f(z^{(l)}) \\ E[\hat{f}] &= E[f] \\ var[\hat{f}] &= \frac{1}{L}E\left[(f - E[f])^2\right] \end{aligned}$$

Note that the variance of estimate is independent of the sample dimensionality. Usually, 20+ independent samples may be sufficient.

General Idea of Sampling

- Mostly, the posterior distribution is primarily required for prediction.
- Fundamental problem: find the *expectation* of some function $f(z)$ with respect to a probability $p(z)$.

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

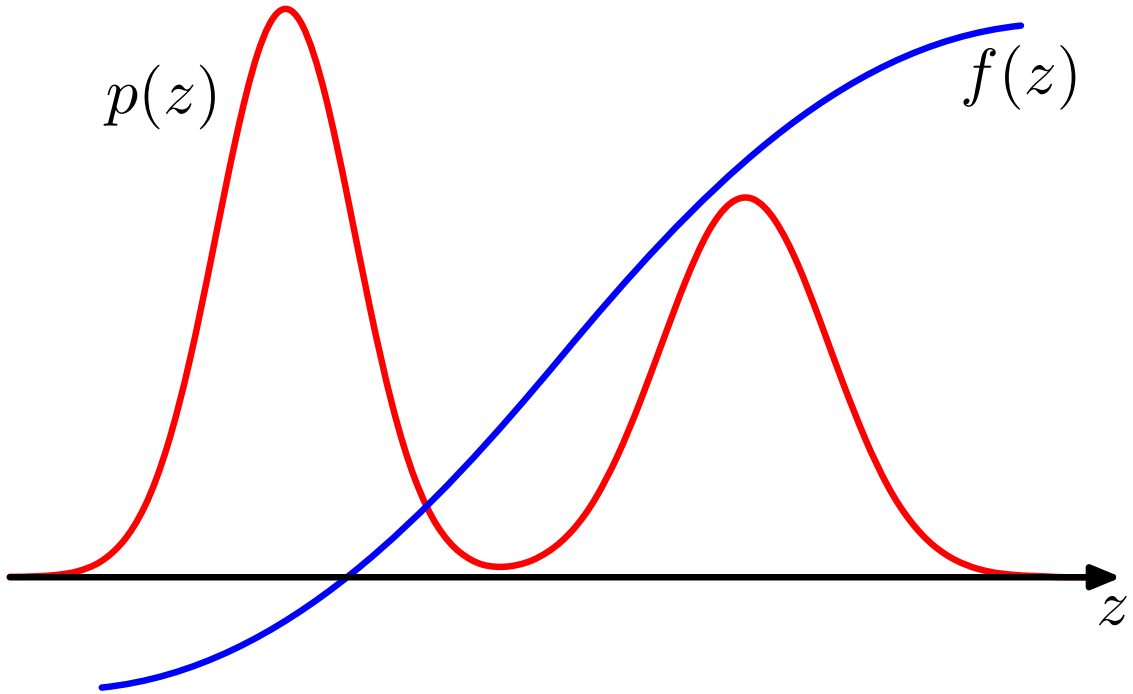
- General idea: obtain a set of samples $z^{(l)}$ drawn independently from the distribution $p(z)$. So we can estimate the expectation:

$$\begin{aligned} \hat{f} &= \frac{1}{L}\sum_{l=1}^{L} f(z^{(l)}) \\ E[\hat{f}] &= E[f] \\ var[\hat{f}] &= \frac{1}{L}E\left[(f - E[f])^2\right] \end{aligned}$$

Note that the variance of estimate is independent of the sample dimensionality. Usually, 20+ independent samples may be sufficient.

So sampling is trivial?

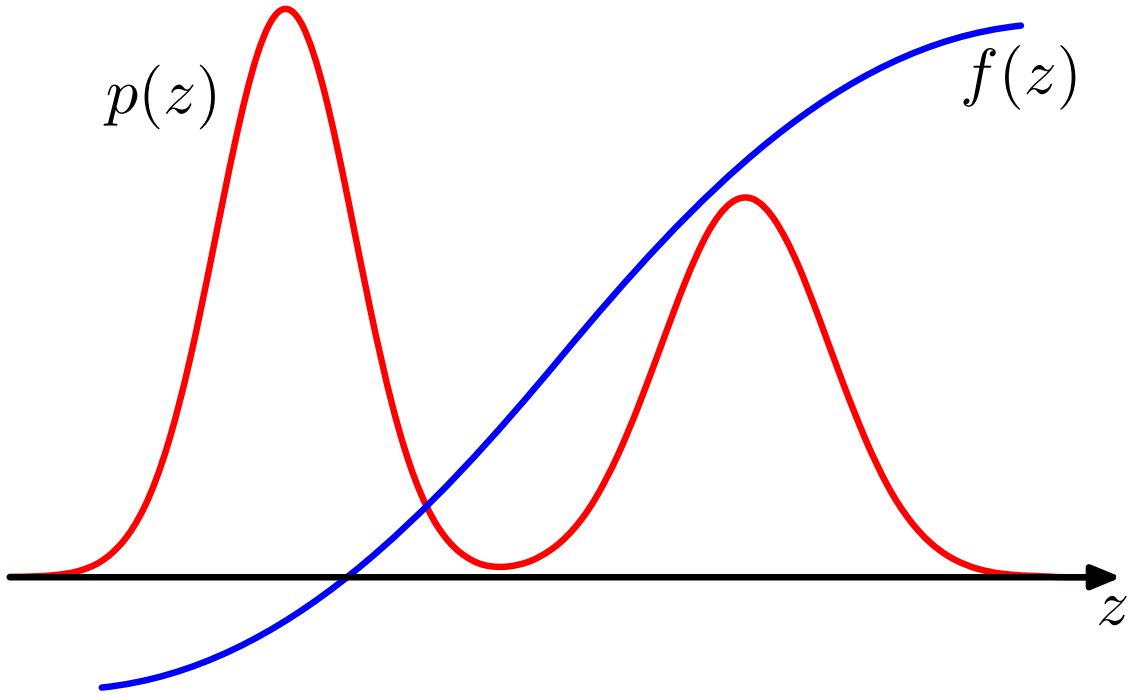
- Expectation might be dominated by regions of small probability.



- The samples might not be independent, so the effective sample size might be much smaller than the apparent sample size.
- In complicated distributions like $p(z) = \frac{1}{Z_p}\hat{p}(z)$, the normalization factor Z_p is hard to calculate directly.

So sampling is trivial?

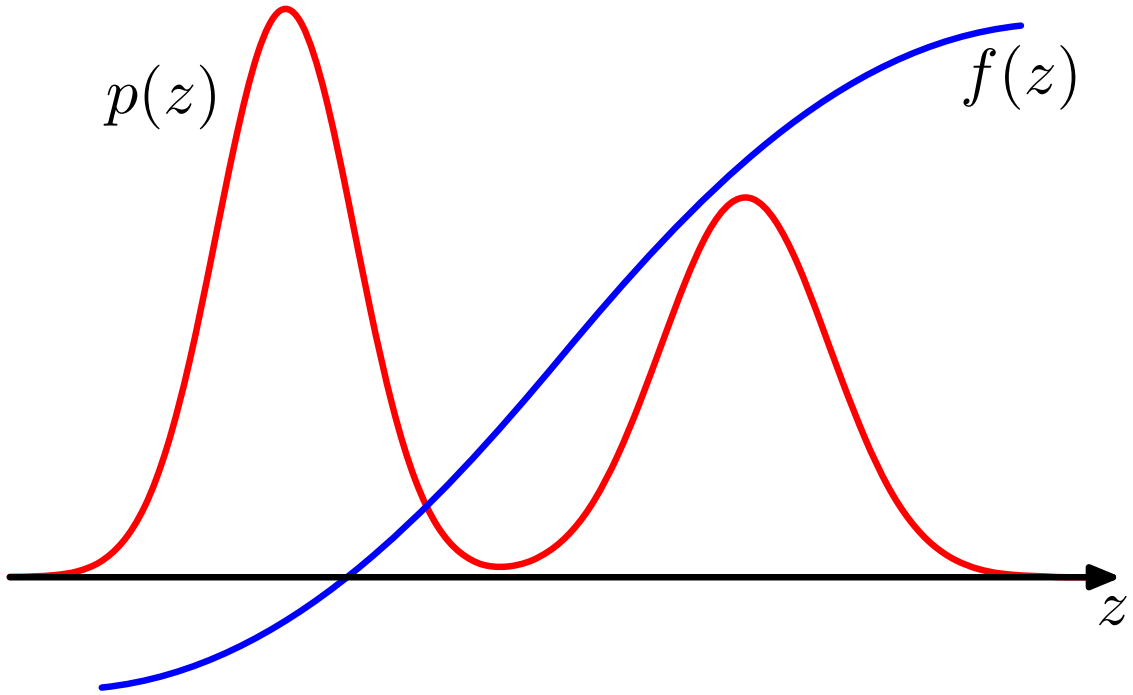
- Expectation might be dominated by regions of small probability.



- The samples might not be independent, so the effective sample size might be much smaller than the apparent sample size.
- In complicated distributions like $p(z) = \frac{1}{Z_p}\hat{p}(z)$, the normalization factor Z_p is hard to calculate directly.

So sampling is trivial?

- Expectation might be dominated by regions of small probability.



- The samples might not be independent, so the effective sample size might be much smaller than the apparent sample size.
- In complicated distributions like $p(z) = \frac{1}{Z_p}\hat{p}(z)$, the normalization factor Z_p is hard to calculate directly.

Sampling from Directed Graphical Models

- No variables are observed: Sample from the joint distribution using ancestral sampling.

$$p(z) = \prod p(z_i|pa_i)$$

  Make one pass through the set of variables in some order and sample from the conditional distribution $p(z_i|pa_i)$.

- Some nodes are observed: draw samples from the joint distribution and throw away samples which are not consistent with observations. Any serious problem?
- The overall probability of accepting a sample from the posterior decreases rapidly as the number of observed variables increases.

Sampling from Directed Graphical Models

- No variables are observed: Sample from the joint distribution using ancestral sampling.

$$p(z) = \prod p(z_i|pa_i)$$

  Make one pass through the set of variables in some order and sample from the conditional distribution $p(z_i|pa_i)$.
- Some nodes are observed: draw samples from the joint distribution and throw away samples which are not consistent with observations. Any serious problem?
- The overall probability of accepting a sample from the posterior decreases rapidly as the number of observed variables increases.

Sampling from Directed Graphical Models

- No variables are observed: Sample from the joint distribution using ancestral sampling.

$$p(z) = \prod p(z_i|pa_i)$$

  Make one pass through the set of variables in some order and sample from the conditional distribution $p(z_i|pa_i)$.
- Some nodes are observed: draw samples from the joint distribution and throw away samples which are not consistent with observations. Any serious problem?
- The overall probability of accepting a sample from the posterior decreases rapidly as the number of observed variables increases.

Sampling from Undirected Graphical Models

For undirected graph,

$$p(x) = \frac{1}{z} \prod_{C} \phi_C(x_C)$$

where C represents the maximal cliques.

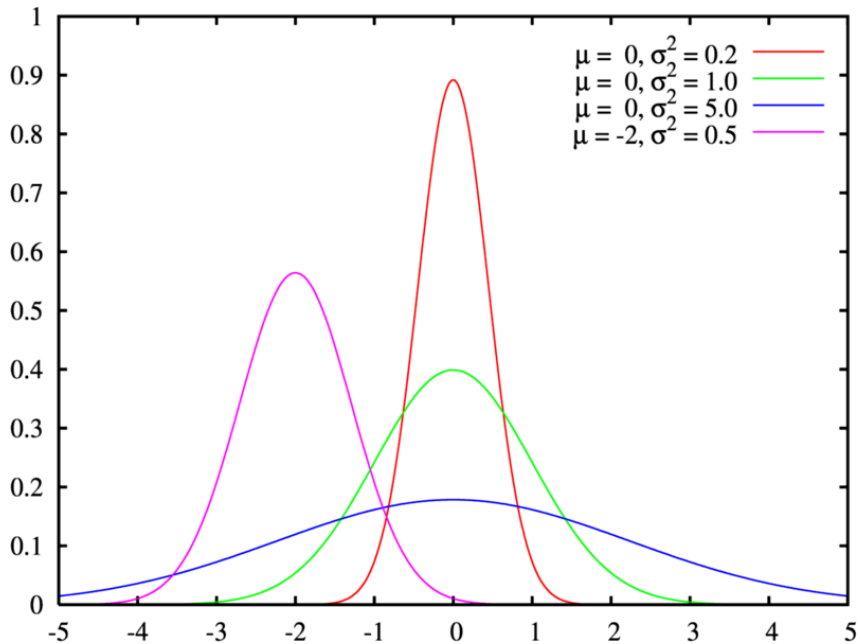
- No one-pass sampling strategy that will sample even from the prior distribution with no observed variables.
- More computational expensive techniques must be employed like Gibbs Sampling (covered later).

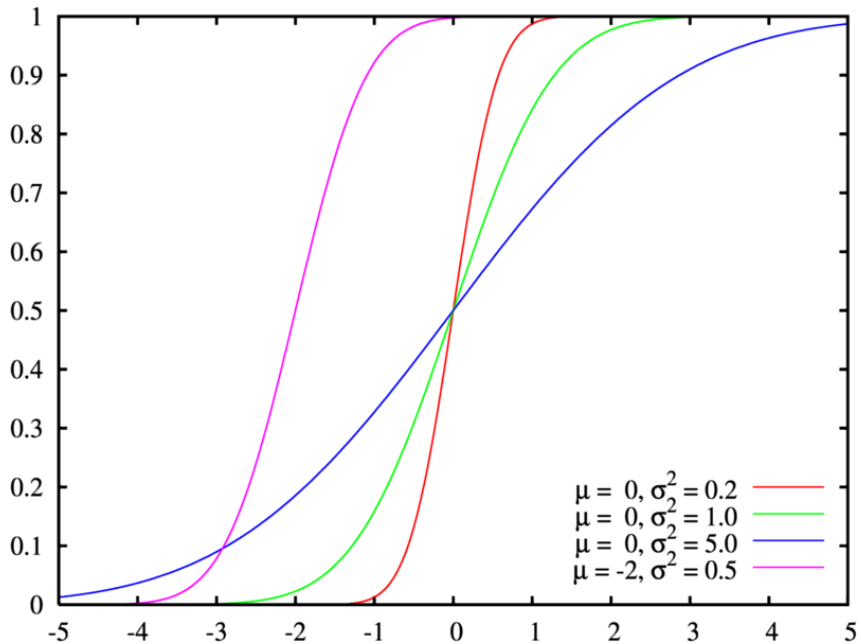
Sampling from marginal distribution

- Sample from joint distribution.
- Sample from conditional distribution (posterior).
- Sample from marginal distribution. If we already have a strategy to sample from a joint distribution $p(u, v)$, then we can obtain marginal distribution $p(u)$ simply by ignoring the values of v in each sample.
- This strategy is used in some sampling techniques.

Review of Basic Probability

- Probability distribution function (pdf)
- Cumulative distribution function (cdf)





Probability under Transformation

If we define a mapping $f(x)$ from the original sample space $\mathcal{X}$ to another sample space $\mathcal{Y}$:

$$f(x) : \mathcal{X} \to \mathcal{Y}$$
$$y = f(x)$$

What's $p(y)$ given $p(x)$?

$$\begin{aligned} F(y) &= P(Y \leq y) \\ &= P(f(X) \leq y) \\ &= \int_{\{x \in \mathcal{X} : f(x) \leq y\}} p(x) dx \end{aligned}$$

Probability under Transformation

If we define a mapping $f(x)$ from the original sample space $\mathcal{X}$ to another sample space $\mathcal{Y}$:

$$f(x) : \mathcal{X} \to \mathcal{Y}$$
$$y = f(x)$$

What's $p(y)$ given $p(x)$?

$$\begin{aligned} F(y) &= P(Y \leq y) \\ &= P(f(X) \leq y) \\ &= \int_{\{x \in \mathcal{X} : f(x) \leq y\}} p(x)dx \end{aligned}$$

For simplicity, we assume the function f is monotonic.

- Monotonic Increasing:

$$\begin{aligned} F_{\mathcal{Y}}(y) &= \int_{\{x \in \mathcal{X}: x \leq f^{-1}(y)\}} p(x)dx \\ &= \int_{-\infty}^{f^{-1}(y)} p(x)dx \\ &= F_{\mathcal{X}}(f^{-1}(y)) \end{aligned}$$

- Monotonic Decreasing:

$$\begin{aligned} F_{\mathcal{Y}}(y) &= \int_{\{x \in \mathcal{X}: x \geq f^{-1}(y)\}} p(x)dx \\ &= \int_{f^{-1}(y)}^{+\infty} p(x)dx \\ &= 1 - F_{\mathcal{X}}(f^{-1}(y)) \end{aligned}$$

For simplicity, we assume the function f is monotonic.

- Monotonic Increasing:

$$\begin{aligned} F_{\mathcal{Y}}(y) &= \int_{\{x \in \mathcal{X}: x \leq f^{-1}(y)\}} p(x)dx \\ &= \int_{-\infty}^{f^{-1}(y)} p(x)dx \\ &= F_{\mathcal{X}}(f^{-1}(y)) \end{aligned}$$

- Monotonic Decreasing:

$$\begin{aligned} F_{\mathcal{Y}}(y) &= \int_{\{x \in \mathcal{X}: x \geq f^{-1}(y)\}} p(x)dx \\ &= \int_{f^{-1}(y)}^{+\infty} p(x)dx \\ &= 1 - F_{\mathcal{X}}(f^{-1}(y)) \end{aligned}$$

$$
\begin{aligned}
p_{\mathcal{Y}}(y) &= \frac{d}{dy} F_Y(y) \\
&= \begin{cases} p_{\mathcal{X}}(f^{-1}(y)) \frac{d}{dy} f^{-1}(y) & \text{if f is increasing} \\ -p_{\mathcal{X}}(f^{-1}(y)) \frac{d}{dy} f^{-1}(y) & \text{if f is decreasing} \end{cases} \\
&= p_{\mathcal{X}}(f^{-1}(y)) \left| \frac{dx}{dy} \right|
\end{aligned}
$$

This can be generalized to multiple variables:

$$y_i = f_i(x_1, x_2, \cdots, x_M), i = 1, 2, \cdots, M.$$

Then $p(y_1, y_2, \cdots, y_M) = p(x_1, \cdots, x_M)|J|$ where J is the Jacobian matrix:

$$|J| = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_M}{\partial y_1} \\ \cdots & \cdots & \cdots \\ \frac{\partial x_1}{\partial y_M} & \cdots & \frac{\partial x_M}{\partial y_M} \end{vmatrix}$$

$$
\begin{aligned}
p_{\mathcal{Y}}(y) &= \frac{d}{dy} F_Y(y) \\
&= \begin{cases} p_{\mathcal{X}}(f^{-1}(y)) \frac{d}{dy} f^{-1}(y) & \text{if f is increasing} \\ -p_{\mathcal{X}}(f^{-1}(y)) \frac{d}{dy} f^{-1}(y) & \text{if f is decreasing} \end{cases} \\
&= p_{\mathcal{X}}(f^{-1}(y)) \left| \frac{dx}{dy} \right|
\end{aligned}
$$

This can be generalized to multiple variables:

$$y_i = f_i(x_1, x_2, \cdots, x_M), i = 1, 2, \cdots, M.$$

Then $p(y_1, y_2, \cdots, y_M) = p(x_1, \cdots, x_M)|J|$ where J is the Jacobian matrix:

$$|J| = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_M}{\partial y_1} \\ \cdots & \cdots & \cdots \\ \frac{\partial x_1}{\partial y_M} & \cdots & \frac{\partial x_M}{\partial y_M} \end{vmatrix}$$

Inversion Method

The Inversion Principle

Let F be a cdf on R with inverse F^{-1} defined by

$$F^{-1}(z) = inf\{x : F(x) = z, 0 \leq u \leq 1\}$$

If $Z \sim U(0,1)$, then $F^{-1}(Z)$ has cdf F; If X has cumulative distribution function F, then $F(X)$ is uniformly distributed on $[0,1]$.

$$P(F^{-1}(z) \leq x) = P(inf\{y : F(y) = z\} \leq x) = P(z \leq F(x)) = F(x)$$
$$P(F(x) \leq z) = P(x \leq F^{-1}(z)) = F(F^{-1}(z)) = z$$

Essentially, as long as we know the exact F^{-1}, we can generate samples for the desired distribution.

- Draw sample z uniformly from [0,1];
- return $F^{-1}(z)$

Inversion Method

The Inversion Principle

Let F be a cdf on R with inverse F^{-1} defined by

$$F^{-1}(z) = inf\{x : F(x) = z, 0 \leq u \leq 1\}$$

If $Z \sim U(0,1)$, then $F^{-1}(Z)$ has cdf F; If X has cumulative distribution function F, then $F(X)$ is uniformly distributed on $[0,1]$.

$$P(F^{-1}(z) \leq x) = P(inf\{y : F(y) = z\} \leq x) = P(z \leq F(x)) = F(x)$$
$$P(F(x) \leq z) = P(x \leq F^{-1}(z)) = F(F^{-1}(z)) = z$$

Essentially, as long as we know the exact F^{-1}, we can generate samples for the desired distribution.

- Draw sample z uniformly from [0,1];
- return $F^{-1}(z)$

Inversion Method

The Inversion Principle

Let F be a cdf on R with inverse F^{-1} defined by

$$F^{-1}(z) = inf\{x : F(x) = z, 0 \leq u \leq 1\}$$

If $Z \sim U(0,1)$, then $F^{-1}(Z)$ has cdf F; If X has cumulative distribution function F, then $F(X)$ is uniformly distributed on $[0,1]$.

$$P(F^{-1}(z) \leq x) = P(inf\{y : F(y) = z\} \leq x) = P(z \leq F(x)) = F(x)$$
$$P(F(x) \leq z) = P(x \leq F^{-1}(z)) = F(F^{-1}(z)) = z$$

Essentially, as long as we know the exact F^{-1}, we can generate samples for the desired distribution.

- Draw sample z uniformly from [0,1];
- return $F^{-1}(z)$

An Example

Suppose y follows an exponential distribution:

$$p(y) = \lambda exp(-\lambda), \quad y \geq 0$$

So

$$\begin{aligned} F(y) &= \int_0^y p(\hat{y}) d\hat{y} \\ &= \int_0^y \lambda exp(-\lambda \hat{y}) d\hat{y} \\ &= -exp(-\lambda \hat{y})|_0^y \\ &= 1 - exp(-\lambda y) \\ F^{-1}(z) &= -\lambda^{-1} ln(1-z) \end{aligned}$$

It follows that $y = -\lambda^{-1} ln(1-z)$.

1. Draw samples uniformly from $(0, 1)$.
2. Obtain the corresponding sample via the above equation.

An Example

Suppose y follows an exponential distribution:

$$p(y) = \lambda exp(-\lambda), \quad y \geq 0$$

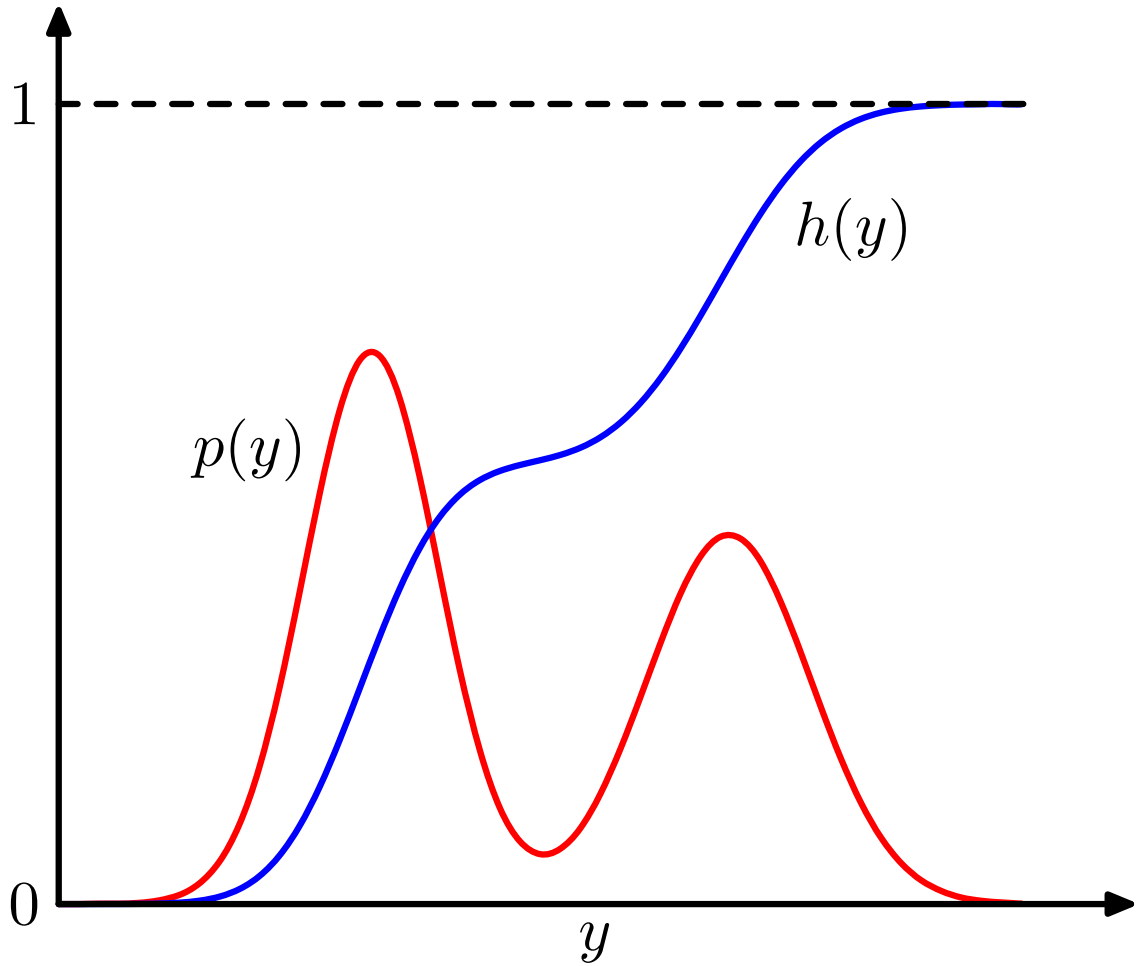
So

$$\begin{aligned} F(y) &= \int_0^y p(\hat{y}) d\hat{y} \\ &= \int_0^y \lambda exp(-\lambda \hat{y}) d\hat{y} \\ &= -exp(-\lambda \hat{y})|_0^y \\ &= 1 - exp(-\lambda y) \\ F^{-1}(z) &= -\lambda^{-1} ln(1-z) \end{aligned}$$

It follows that $y = -\lambda^{-1} ln(1-z)$.

1. Draw samples uniformly from $(0, 1)$.
2. Obtain the corresponding sample via the above equation.

Intuition



$h(y)$ is flat, then corresponding y should have low probability.

Sample from Gaussian Distribution

1. Use inversion method to draw samples. Unfortunatelly, the inverse function requires a lot of compuation and sometimes need approximation.
2. Use central-limit theorem. Draw n samples from $U(0, 1)$, calculate its average. Approximatelly, it follows a normal distribution.

Box-Muller method for generating Gaussian samples

Sample from Gaussian Distribution with zero mean and unit variance

- Generate pairs of uniformly distributed random numbers $z_1, z_2 \in (-1, 1)$.
- Discard each pair unless $z_1^2 + z_2^2 \leq 1$. Obtain a uniform distribution of points inside the unit circle with $p(z_1, z_2) = \frac{1}{\pi}$.
- $$y_1 = z_1 \left(\frac{-2 \ln r^2}{r^2} \right)^{\frac{1}{2}}$$
$$y_2 = z_2 \left(\frac{-2 \ln r^2}{r^2} \right)^{\frac{1}{2}}$$

  where $r^2 = z_1^2 + z_2^2$. Then, (y_1, y_2) follows a Gaussian distribution and unit variance.

Why it's Gaussian?

For multiple variables, we need the Jacobian of the change of variables:

$$p(y_1, y_2, \cdots, y_M) = p(z_1, \cdots, z_M) \left| \frac{\partial(z_1, \cdots, z_M)}{\partial(y_1, \cdots, y_M)} \right|$$

Thus, we only need to calculate the Jacobian matrix. As

$$\begin{aligned} y_1^2 + y_2^2 &= -2\ln(r^2) \Longrightarrow z_1^2 + z_2^2 = \exp(-\frac{y_1^2 + y_2^2}{2}) \\ \frac{y_1}{y_2} &= \frac{z_1}{z_2} \end{aligned}$$

Hence (tedious calculation skipped here, left as a homework)

$$\begin{aligned} p(y_1, y_2) &= p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right| \\ &= \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_1^2}{2} \right) \right] \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_2^2}{2} \right) \right] \end{aligned}$$

Other form of Gaussian Distribution

In previous example, it's a Gaussian Distribution with zero mean and unit variance. What if other mean and covariance matrix?

- If $y \sim N(0, 1)$, then $\sigma y + \mu \sim N(\mu, \sigma^2)$.
- To generate covariance matrix Σ, we can make use of *Cholesky decomposition* ($\Sigma = LL^T$). Then, if $\mu + Ly \sim N(\mu, \Sigma)$.

The previous examples show how to generate samples from standard distributions, but it's very limited. We encounter usually much more complicated distributions, especially in Bayesian inference. Need more elegant techniques.

Other form of Gaussian Distribution

In previous example, it's a Gaussian Distribution with zero mean and unit variance. What if other mean and covariance matrix?

- If $y \sim N(0, 1)$, then $\sigma y + \mu \sim N(\mu, \sigma^2)$.
- To generate covariance matrix Σ, we can make use of *Cholesky decomposition* ($\Sigma = LL^T$). Then, if $\mu + Ly \sim N(\mu, \Sigma)$.

The previous examples show how to generate samples from standard distributions, but it's very limited. We encounter usually much more complicated distributions, especially in Bayesian inference. Need more elegant techniques.

Rejection Sampling

Suppose we want to sample from distribution $p(z)$, and

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

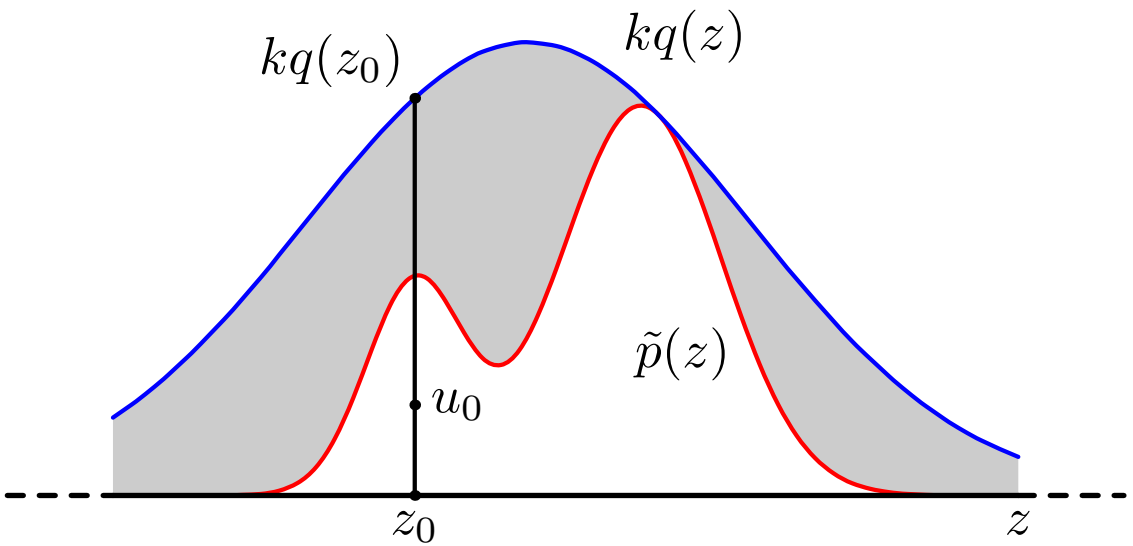
where $\hat{p}(z)$ can readily be evaluated, but Z_p is unknown.

Rejection Sampling

We need a simpler proposal distribution $q(z)$ such that there exists a constraint k such that $kq(z) \geq \hat{p}(z)$ for all z.

Algorithm

1. Draw a sample z_0 from $q(z)$.
2. Generate a number u_0 from uniform distribution over $[0, kq(z_0)]$;
3. If $u_0 \geq \hat{p}(z_0)$, the sample is rejected; Otherwise, z_0 is accepted.



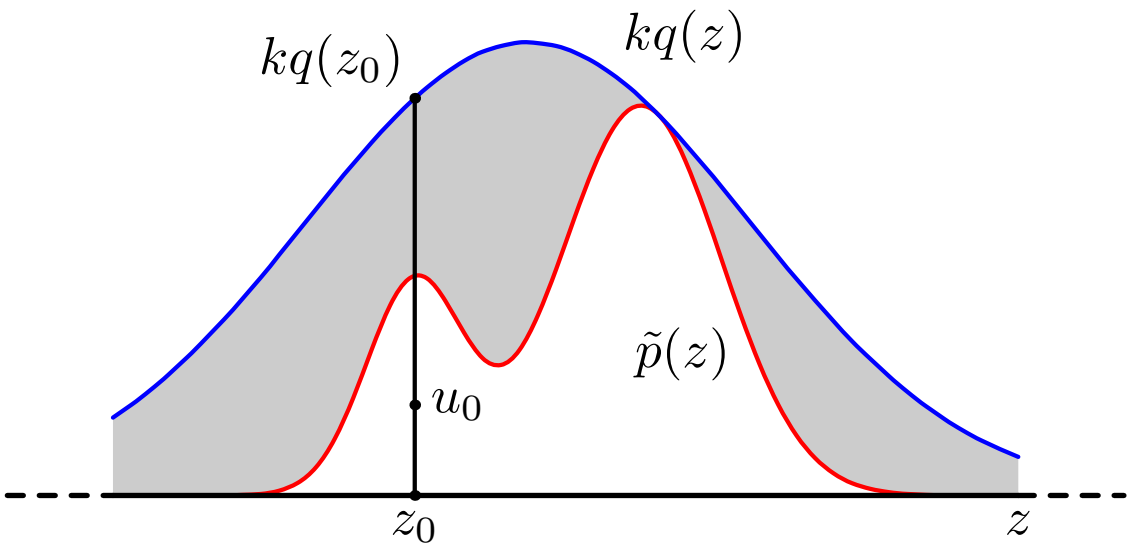
Note that the sample pair (z_0, u_0) has uniform distribution under the curve of $\hat{p}(z)$. Hence, the z values are distributed according to $p(z)$.

Rejection Sampling

We need a simpler proposal distribution $q(z)$ such that there exists a constraint k such that $kq(z) \geq \hat{p}(z)$ for all z.

Algorithm

1. Draw a sample z_0 from $q(z)$.
2. Generate a number u_0 from uniform distribution over $[0, kq(z_0)]$;
3. If $u_0 \geq \hat{p}(z_0)$, the sample is rejected; Otherwise, z_0 is accepted.



Note that the sample pair (z_0, u_0) has uniform distribution under the curve of $\hat{p}(z)$. Hence, the z values are distributed according to $p(z)$.

Disadvantages

- Sometimes, it's not so easy to find a k s.t. $kq(z) \geq \hat{p}(z), \forall z$.
- The ratio k must be as tight as possible.

$$p(\text{accept}) = \int \frac{\hat{p}(z)}{kq(z)} q(z) dz = \frac{1}{k} \int \hat{p}(z) dz$$

Larger k usually result in large portion of rejections :(

- As long as $\hat{p}(z)$ is under a envelope function $kq(z)$ for all z, this algorithm works. Is it possible to obtain relatively tight bound for different intervals of z?

Disadvantages

- Sometimes, it's not so easy to find a k s.t. $kq(z) \geq \hat{p}(z), \forall z$.
- The ratio k must be as tight as possible.

$$p(\text{accept}) = \int \frac{\hat{p}(z)}{kq(z)} q(z) dz = \frac{1}{k} \int \hat{p}(z) dz$$

  Larger k usually result in large portion of rejections :(
- As long as $\hat{p}(z)$ is under a envelope function $kq(z)$ for all z, this algorithm works. Is it possible to obtain relatively tight bound for different intervals of z?

Is a global k required?

- Essentially, we need to generate samples such that $p_{sampling}(z) \propto \hat{p}(z)$.
- So if a global k is used

$$p_{sampling}(z) \propto q(z) \frac{\hat{p}(z)}{k\ q(z)}$$

We get the required distribution.

- However, if we used different k in different intervals, this will result in some problem.
- Goal:sample from a Gaussian distribution p, we use $q = p$ as the proposal distribution
- Idealy, we should use a global $k = 1$. What if I set $k = 2$ for $z \leq 0$?
- All the positive samples will be accepted, but the negative samples will be accepted with only half chance. This is not our original Gaussian distribution!!

Is a global k required?

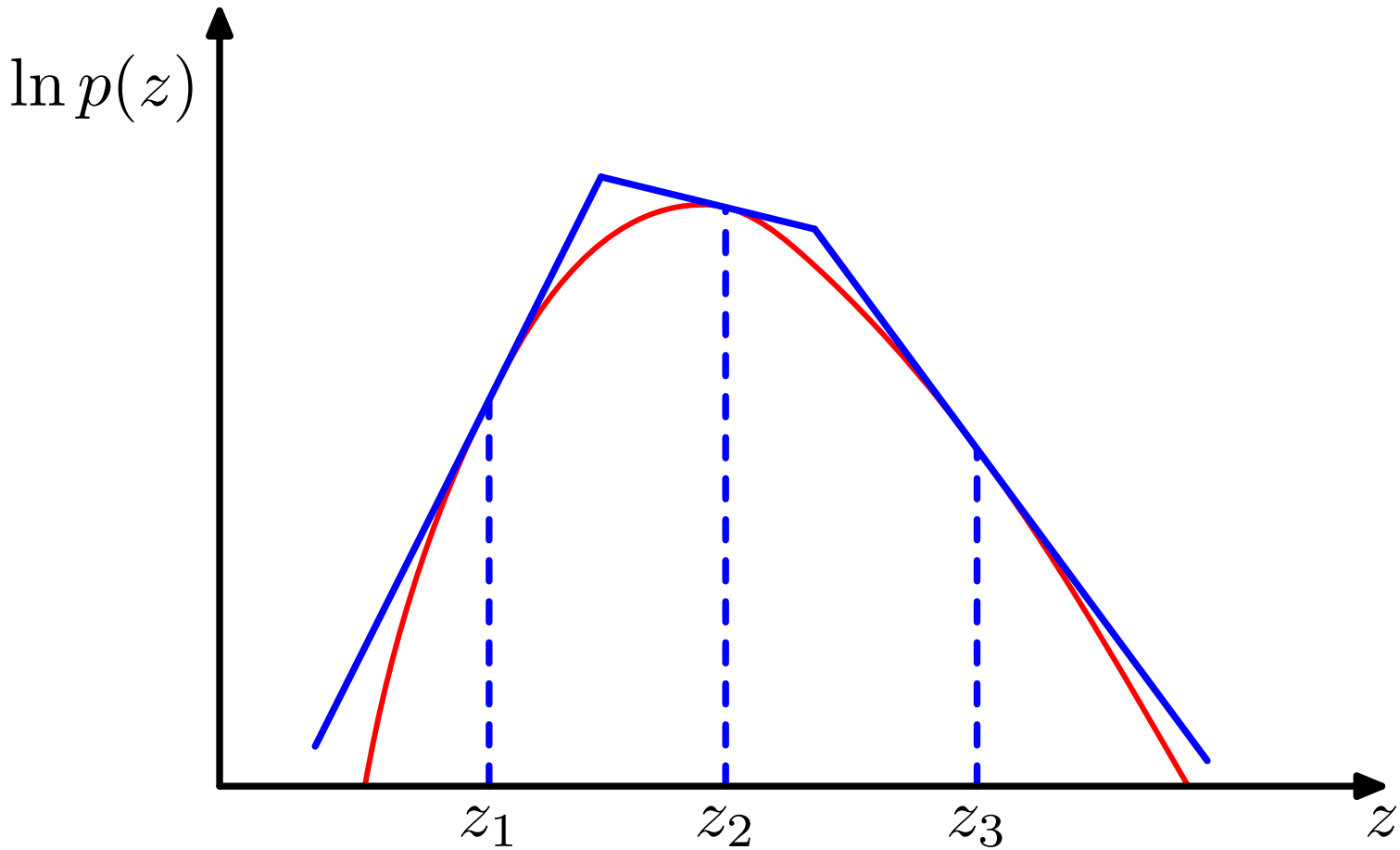
- Essentially, we need to generate samples such that $p_{sampling}(z) \propto \hat{p}(z)$.
- So if a global k is used

$$p_{sampling}(z) \propto q(z) \frac{\hat{p}(z)}{k\ q(z)}$$

  We get the required distribution.
- However, if we used different k in different intervals, this will result in some problem.
- Goal:sample from a Gaussian distribution p, we use $q = p$ as the proposal distribution
- Idealy, we should use a global $k = 1$. What if I set $k = 2$ for $z \leq 0$?
- All the positive samples will be accepted, but the negative samples will be accepted with only half chance. This is not our original Gaussian distribution!!

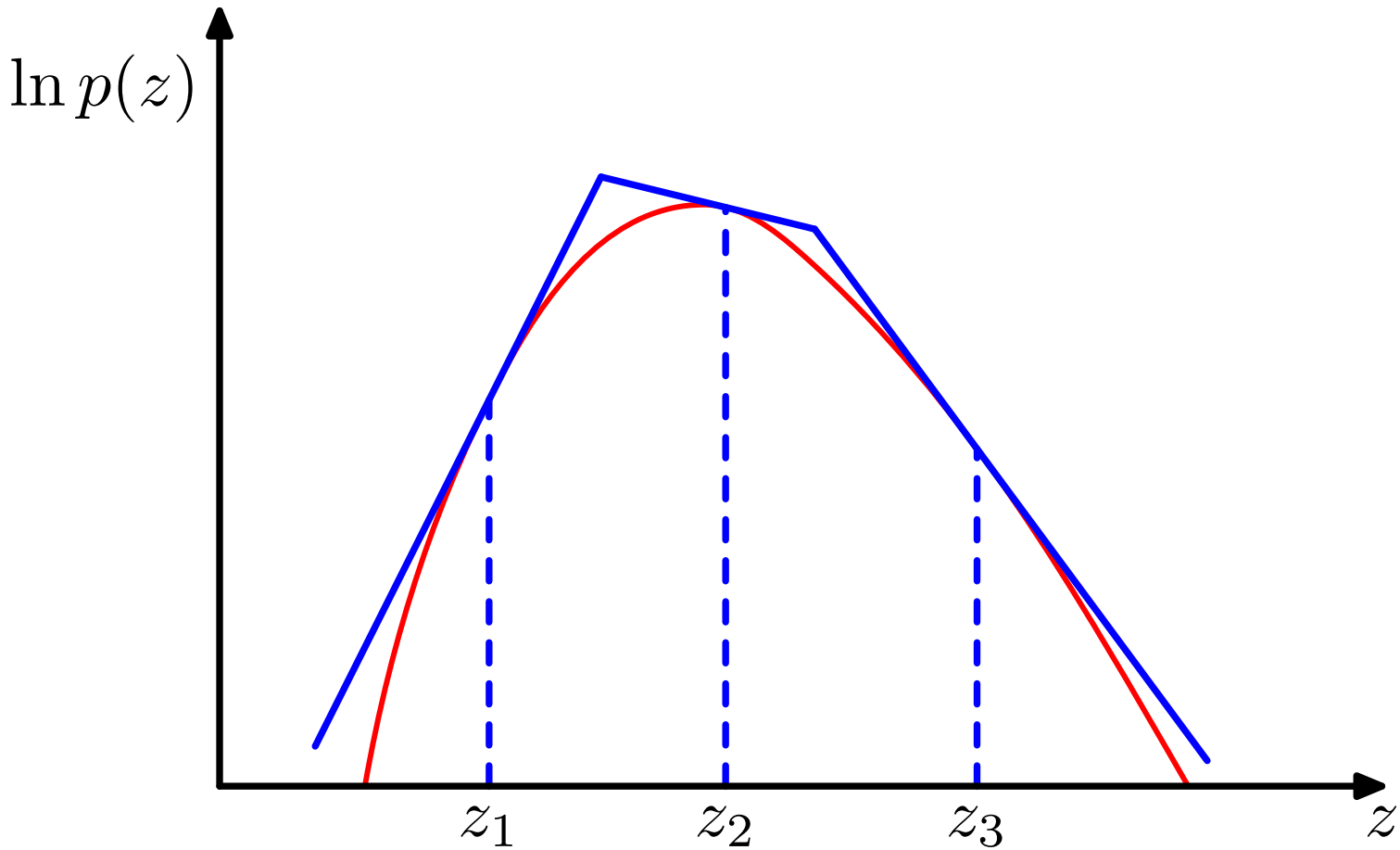
Adaptive Rejection Sampling

- Difficult to obtain suitable analytic form for the envelope distribution $q(z)$.
- Alternative Approach: Construct the envelope function on the fly.
- Particularly straightforward if $p(z)$ is log concave ($\log p(z)$ is concave).



Adaptive Rejection Sampling

- Difficult to obtain suitable analytic form for the envelope distribution $q(z)$.
- Alternative Approach: Construct the envelope function on the fly.
- Particularly straightforward if $p(z)$ is log concave ($\log p(z)$ is concave).



Construct Envelope On The Fly I

- The function $\ln p(z)$ and its gradient are evaluated at some initial set of grid points and the intersection of the resulting tangent lines are used to construct the envelope function.
- Suppose the tangent line between intersection z_{i-1} and z_i is

$$\begin{aligned} line(z) = ln\ E(z) &= -\lambda_i(z - z_{i-1}) + b_i \\ k\ q(z) = E(z) &= c_i exp\left\{-\lambda_i(z - z_{i-1})\right\} \\ q(z) &= \frac{E(z)}{\int_D E(z)dz} \quad \text{(Normalized envelope function)} \end{aligned}$$

- The envelope function comprises a piecewise exponential distribution of the form

$$q(z) = k_i \lambda_i exp\left\{-\lambda_i(z - z_{i-1})\right\} \qquad z_{i-1} \leq z \leq z_i$$

where $k_i = \frac{c_i}{\int_D E(z)dz}$.

Construct Envelope On The Fly I

- The function $\ln p(z)$ and its gradient are evaluated at some initial set of grid points and the intersection of the resulting tangent lines are used to construct the envelope function.
- Suppose the tangent line between intersection z_{i-1} and z_i is

$$\begin{aligned} line(z) = ln\ E(z) &= -\lambda_i(z - z_{i-1}) + b_i \\ k\ q(z) = E(z) &= c_i exp\left\{-\lambda_i(z - z_{i-1})\right\} \\ q(z) &= \frac{E(z)}{\int_D E(z)dz} \quad \text{(Normalized envelope function)} \end{aligned}$$

- The envelope function comprises a piecewise exponential distribution of the form

$$q(z) = k_i \lambda_i exp\left\{-\lambda_i(z - z_{i-1})\right\} \qquad z_{i-1} \leq z \leq z_i$$

where $k_i = \frac{c_i}{\int_D E(z)dz}$.

Construct Envelope On The Fly I

- The function $\ln p(z)$ and its gradient are evaluated at some initial set of grid points and the intersection of the resulting tangent lines are used to construct the envelope function.
- Suppose the tangent line between intersection z_{i-1} and z_i is

$$\begin{aligned} line(z) = ln\ E(z) &= -\lambda_i(z - z_{i-1}) + b_i \\ k\ q(z) = E(z) &= c_i exp\left\{-\lambda_i(z - z_{i-1})\right\} \\ q(z) &= \frac{E(z)}{\int_D E(z)dz} \quad \text{(Normalized envelope function)} \end{aligned}$$

- The envelope function comprises a piecewise exponential distribution of the form

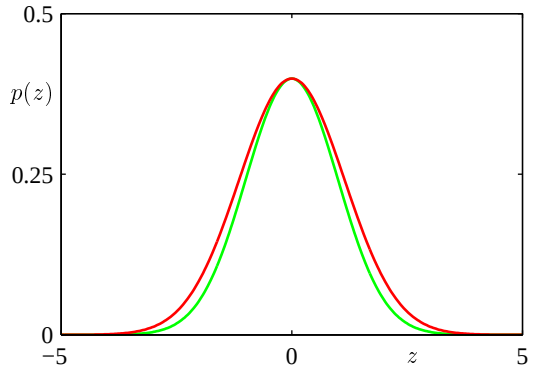
$$q(z) = k_i\lambda_i exp\left\{-\lambda_i(z - z_{i-1})\right\} \qquad z_{i-1} \leq z \leq z_i$$

where $k_i = \frac{c_i}{\int_D E(z)dz}$.

Construct Envelope on The Fly II

- A sample value z is drawn from the normalized envelope function $q(z)$. This could be achieved using inversion method.
- Draw a sample u from uniform distribution;
- If $u < exp(ln\hat{p}(z) - line(z))$, accept z;
- Otherwise, the tangent line of the new sample is computed to refine the envelope function.
- The envelope becomes tighter and tighter. Every rejected sample help refine the envelope function–It's adaptive!!

Curse of High Dimensionality for Rejection Sampling



Sample from a high-dimensional Gaussian distribution

- An artificial problem: wish to sample from $p(z) = N(0, \sigma_p^2 \mathbf{I})$.
- Suppose we have a proposal distribution $q(z) = N(0, \sigma_q^2 \mathbf{I})$ such that $\sigma_q^2 \geq \sigma_p^2$.
- The optimum bound k is obtained when $z = 0$.

$$k = \frac{p(z)}{q(z)} = \frac{|\sigma_p^2 \mathbf{I}|^{-1/2}}{|\sigma_q^2 \mathbf{I}|^{-1/2}} = \left(\frac{\sigma_q}{\sigma_p}\right)^D$$

Rejection is too much!

$$k = \left(\frac{\sigma_q}{\sigma_p}\right)^D$$

Remember that the acceptance rate is

$$p(accept) = \frac{1}{k}\int \hat{p}(z)dz = \frac{1}{k}$$

Here $\hat{p}(z) = p(z)$.

- The acceptance rate diminishes exponentially with dimensionality.
- If $D = 1000$, the acceptance ratio will be about $1/20,000$. Obtain 1 sample from 20,000 samples from $q(z)$.
- In practical examples, the desired distribution may be multi-modal or sharply peaked. It will be extremely difficult to find a good proposal distribution.
- Rejection sampling suffers from high-dimensionality. Usually act as a subroutine to sample from 1 or 2 dimensions in a more complicated algorithm.

Rejection is too much!

$$k = \left(\frac{\sigma_q}{\sigma_p}\right)^D$$

Remember that the acceptance rate is

$$p(accept) = \frac{1}{k}\int \hat{p}(z)dz = \frac{1}{k}$$

Here $\hat{p}(z) = p(z)$.

- The acceptance rate diminishes exponentially with dimensionality.
- If $D = 1000$, the acceptance ratio will be about $1/20,000$. Obtain 1 sample from 20,000 samples from $q(z)$.
- In practical examples, the desired distribution may be multi-modal or sharply peaked. It will be extremely difficult to find a good proposal distribution.
- Rejection sampling suffers from high-dimensionality. Usually act as a subroutine to sample from 1 or 2 dimensions in a more complicated algorithm.

Rejection is too much!

$$k = \left(\frac{\sigma_q}{\sigma_p}\right)^D$$

Remember that the acceptance rate is

$$p(accept) = \frac{1}{k}\int \hat{p}(z)dz = \frac{1}{k}$$

Here $\hat{p}(z) = p(z)$.

- The acceptance rate diminishes exponentially with dimensionality.
- If $D = 1000$, the acceptance ratio will be about $1/20,000$. Obtain 1 sample from 20,000 samples from $q(z)$.
- In practical examples, the desired distribution may be multi-modal or sharply peaked. It will be extremely difficult to find a good proposal distribution.
- Rejection sampling suffers from high-dimensionality. Usually act as a subroutine to sample from 1 or 2 dimensions in a more complicated algorithm.

- (Adaptive) Rejection Sampling might have to reject samples.
- A serious problem for high dimensionality.
- Is it possible to utilize all the samples?

- (Adaptive) Rejection Sampling might have to reject samples.
- A serious problem for high dimensionality.
- Is it possible to utilize all the samples?

- In practical cases, we usually only wish to calculate the expectation (e.g. Bayesian Prediction, E-step in EM algorithm).
- Consider the case where we know $p(z)$ but we can not draw samples from it directly.
- A simple strategy:

$$E[f] \approx \sum_{l=1}^{L} p(\mathbf{z}^{(l)}) f(\mathbf{z}^{(l)})$$

- The distribution of interest often have much of their mass confined to relatively small regions of $\mathbf{z}$. Uniform sampling would be very inefficient: only a very small proportion of the samples will make a significant contribution.
- We really like to choose the sample points to fall in regions where $p(\mathbf{z})$ is large, or ideally where the product $p(z)f(z)$ is large.

- In practical cases, we usually only wish to calculate the expectation (e.g. Bayesian Prediction, E-step in EM algorithm).
- Consider the case where we know $p(z)$ but we can not draw samples from it directly.
- A simple strategy:

$$E[f] \approx \sum_{l=1}^{L} p(\mathbf{z}^{(l)}) f(\mathbf{z}^{(l)})$$

- The distribution of interest often have much of their mass confined to relatively small regions of $\mathbf{z}$. Uniform sampling would be very inefficient: only a very small proportion of the samples will make a significant contribution.
- We really like to choose the sample points to fall in regions where $p(\mathbf{z})$ is large, or ideally where the product $p(z)f(z)$ is large.

Importance Sampling

Take a proposal distribution $q(z)$:

$$\begin{aligned} E[f] &= \int f(z)p(z)dz \\ &= \int f(z)\frac{p(z)}{q(z)}q(z)dz \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}f(z^{(l)}) \end{aligned}$$

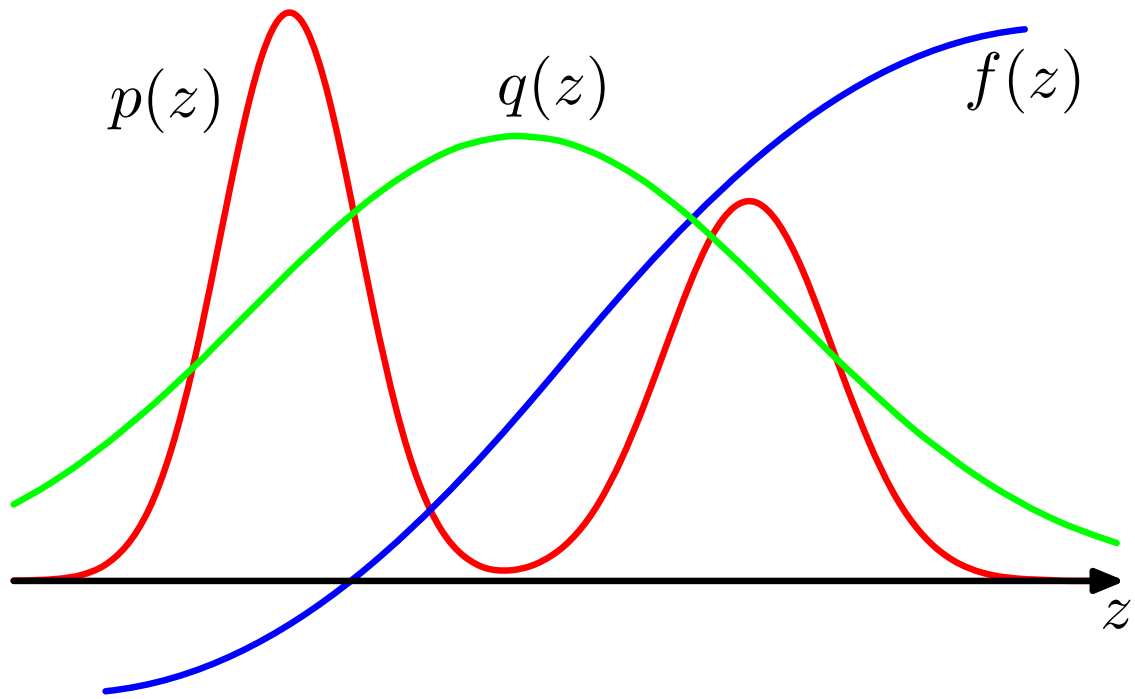
The quantities $r_l = \frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}$ are known as *importance weights*.

Importance Sampling

Take a proposal distribution $q(z)$:

$$\begin{aligned} E[f] &= \int f(z)p(z)dz \\ &= \int f(z)\frac{p(z)}{q(z)}q(z)dz \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}f(z^{(l)}) \end{aligned}$$

The quantities $r_l = \frac{p(\mathbf{z}^{(l)})}{q(\mathbf{z}^{(l)})}$ are known as *importance weights*.



- The importance weights correct the bias from a wrong distribution.
- There's no strict bound requirement as in rejection sampling.
- Unlike rejection sampling, all the samples are retained here.

Importance sampling without normalization factor

$p(z) = \hat{p}(z)/Z_p$ where $\hat{p}(z)$ can be evaluated easily but Z_p is unknown.
Suppose $q(z) = \hat{q}(z)/Z_q$:

$$\begin{aligned} E(f) &= \int f(z)p(z)dz \\ &= \frac{Z_q}{Z_p}\int f(z)\frac{\hat{p}(z)}{\hat{q}(z)}q(z)dz \\ &\approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\hat{r}_l f(z^{(l)}) \end{aligned}$$

where $\hat{r}_l = \hat{p}(z^{(l)})/\hat{q}(z^{(l)})$.

Quiz: But how to estimate $\frac{Z_q}{Z_p}$?

$$\begin{aligned}\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \hat{p}(z)dz = \int \frac{\hat{p}(z)}{\hat{q}(z)}q(z)dz \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\hat{r}_l\end{aligned}$$

So

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

where

$$w_l = \frac{\hat{r}_l}{\sum_m \hat{r}_m}$$

Here w_l can be considered as a *normalized* importance weight.
The core idea of using importance sampling is to transform a quantity to a expectation with respect to a distribution.

Basic Procedure

1. Use a proposal distribution $q(z)$ to generate samples;
2. Calculate the weights for each sample $\hat{r}_l = \hat{p}(z^{(l)})/\hat{q}(z^{(l)})$.
3. Calculate the normalized weight r_l.
4. Find out the expectation.

Importance sampling applied to Graphical Models (1)

How to calculate the expectation given some variables observed?

- Straightforward: ancestral sampling, throw away those inconsistent samples.
- *Uniform Sampling*: The joint distribution is obtained by first setting those variables z_i that are observed. Each remaining variables is then sampled independently from a uniform distribution over the probability space.
- Then the weight of each sample is proportional to $p(z)$. Essentially, use a uniform distribution as proposal distribution.
- Note that there's no ordering of variables for sampling.
- The posterior is far from uniform, so generally lead to poor result. For continuous values, the probabilitiy could be very low; For discrete values, the probability could be zero (as the sample might not be real).

Importance sampling applied to Graphical Models (1)

How to calculate the expectation given some variables observed?

- Straightforward: ancestral sampling, throw away those inconsistent samples.
- *Uniform Sampling*: The joint distribution is obtained by first setting those variables z_i that are observed. Each remaining variables is then sampled independently from a uniform distribution over the probability space.
- Then the weight of each sample is proportional to $p(z)$. Essentially, use a uniform distribution as proposal distribution.
- Note that there's no ordering of variables for sampling.
- The posterior is far from uniform, so generally lead to poor result. For continuous values, the probabilitiy could be very low; For discrete values, the probability could be zero (as the sample might not be real).

Importance sampling applied to Graphical Models (1)

How to calculate the expectation given some variables observed?

- Straightforward: ancestral sampling, throw away those inconsistent samples.
- *Uniform Sampling*: The joint distribution is obtained by first setting those variables z_i that are observed. Each remaining variables is then sampled independently from a uniform distribution over the probability space.
- Then the weight of each sample is proportional to $p(z)$. Essentially, use a uniform distribution as proposal distribution.
- Note that there's no ordering of variables for sampling.
- The posterior is far from uniform, so generally lead to poor result. For continuous values, the probabilitiy could be very low; For discrete values, the probability could be zero (as the sample might not be real).

Importance sampling applied to Graphical Models (1)

How to calculate the expectation given some variables observed?

- Straightforward: ancestral sampling, throw away those inconsistent samples.
- *Uniform Sampling*: The joint distribution is obtained by first setting those variables z_i that are observed. Each remaining variables is then sampled independently from a uniform distribution over the probability space.
- Then the weight of each sample is proportional to $p(z)$. Essentially, use a uniform distribution as proposal distribution.
- Note that there's no ordering of variables for sampling.
- The posterior is far from uniform, so generally lead to poor result. For continuous values, the probabilitiy could be very low; For discrete values, the probability could be zero (as the sample might not be real).

Importance sampling applied to Graphical Models (2)

- Likelihood Weighted Sampling: Based on ancestral sampling of variables.
- If the variable is observed, just set to its value for sampling; If not, sample from the conditional distribution.
- Essentially, a proposal distribution q such thtat

$$q(z_i) = \begin{cases} p(z_i|pa_i) & z_i \notin \mathbf{e} \\ 1 & z_i \in \mathbf{e} \end{cases}$$

- $$r(z) = \prod_{z_i \notin \mathbf{e}} \frac{p(z_i|pa_i)}{p(z_i|pa_i)} \prod_{z_i \in \mathbf{e}} \frac{p(z_i|pa_i)}{1} = \prod_{z_i \in \mathbf{e}} p(z_i|pa_i)$$

Limitations for Importance Sampling

- As with rejection sampling, the success of importance sampling depends crucially on how well the proposal distribution $q(z)$ matches the desired distribution $p(z)$.
- r_l is dominated by few if $p(z)f(z)$ is strongly varying, and has a significant proportion of its mass concentrated over relatively small region of z space. The effective sample size is actually much smaller than L.
- More severe if none of the sample falls into the regions where $p(z)f(z)$ is large. In this case, the variance of $r_l f(z^{(l)})$ could be small, but the expectation is totally wrong!!
- Key requirement for $q(z)$: Not be small or zero in regions where $p(z)$ may be significant. The shape of proposal distribution better be similar to the true distribution.

Limitations for Importance Sampling

- As with rejection sampling, the success of importance sampling depends crucially on how well the proposal distribution $q(z)$ matches the desired distribution $p(z)$.
- r_l is dominated by few if $p(z)f(z)$ is strongly varying, and has a significant proportion of its mass concentrated over relatively small region of z space. The effective sample size is actually much smaller than L.
- More severe if none of the sample falls into the regions where $p(z)f(z)$ is large. In this case, the variance of $r_l f(z^{(l)})$ could be small, but the expectation is totally wrong!!
- Key requirement for $q(z)$: Not be small or zero in regions where $p(z)$ may be significant. The shape of proposal distribution better be similar to the true distribution.

Limitations for Importance Sampling

- As with rejection sampling, the success of importance sampling depends crucially on how well the proposal distribution $q(z)$ matches the desired distribution $p(z)$.
- r_l is dominated by few if $p(z)f(z)$ is strongly varying, and has a significant proportion of its mass concentrated over relatively small region of z space. The effective sample size is actually much smaller than L.
- More severe if none of the sample falls into the regions where $p(z)f(z)$ is large. In this case, the variance of $r_l f(z^{(l)})$ could be small, but the expectation is totally wrong!!
- Key requirement for $q(z)$: Not be small or zero in regions where $p(z)$ may be significant. The shape of proposal distribution better be similar to the true distribution.

Rejection sampling

The determination of a suitable constant k might be impractical.

- Need to satisfy the bound requirement
- Large k leads to extremely low acceptance rate.

Is it possible to relax the "*tight bound*" requirement for sampling?

- Importance sampling does not require bound; and no rejection.
- But only for computing the expectation.
- Is it possible to combine importance weights with sampling?

Rejection sampling

The determination of a suitable constant k might be impractical.

- Need to satisfy the bound requirement
- Large k leads to extremely low acceptance rate.

Is it possible to relax the "*tight bound*" requirement for sampling?

- Importance sampling does not require bound; and no rejection.
- But only for computing the expectation.
- Is it possible to combine importance weights with sampling?

Sampling-importance-resampling(SIR)

SIR

- Recall the idea of Boosting algorithm: adjust the weight of each data point based on loss and then sample the data according to the weights.
- Similar idea for SIR:
  1. Draw L samples from $q(z)$: $(z^{(1)}, z^{(2)}, \cdots, z^{(L)})$.
  2. Weights are calculated the same as in importance sampling.
  3. A second set of L samples is drawn from the discrete distribution $(z^{(1)}, z^{(2)}, \cdots, z^{(L)})$.

Sampling-importance-resampling(SIR)

SIR

- Recall the idea of Boosting algorithm: adjust the weight of each data point based on loss and then sample the data according to the weights.
- Similar idea for SIR:
  1. Draw L samples from $q(z)$: $(z^{(1)}, z^{(2)}, \cdots, z^{(L)})$.
  2. Weights are calculated the same as in importance sampling.
  3. A second set of L samples is drawn from the discrete distribution $(z^{(1)}, z^{(2)}, \cdots, z^{(L)})$.

Why SIR works?

$$
\begin{aligned}
p(z \leq a) &= \sum_{l: z^{(l)} \leq a} w_l \\
&= \frac{\sum_l I(z^{(l)} \leq a) \hat{p}(z^{(l)}) / q(z^{(l)})}{\sum_l \hat{p}(z^{(l)}) / q(z^{(l)})}
\end{aligned}
$$

Take $L \longrightarrow \infty$, then

$$
\begin{aligned}
p(z \leq a) &= \frac{\int I(z \leq a) \{\hat{p}(z)/q(z)\} q(z) dz}{\int \{\hat{p}(z)/q(z)\} q(z) dz} \\
&= \frac{\int I(z \leq a) \hat{p}(z) dz}{\int \hat{p}(z) dz} \\
&= \int I(z \leq a) p(z) dz
\end{aligned}
$$

Here, the normalization factor of $p(z)$ is not required.

Comments

1. Sampling-Importance-Resampling is an approximation, but reject sampling is drawing samples from the true distribution.
2. Similar to rejection sampling, the approximation improves if the sampling distribution $q(z)$ get closer to the desired distribution.
3. When $q(z) = p(z)$, the initial samples $(z^{(1)}, z^{(2)}, \cdots, z^{(L)})$ have desired distribution and the weights $w_l = 1/L$.
4. If moments with respect to z is required, they can be evaluated similar to importance sampling.

Monte Carlo EM algorithm

- Sometimes, E-step in EM is intractable, especially proble Sampling methods can be used to approximate the E-step of the EM algorithm.
- Consider a model with hidden variables $\mathbf{Z}$, visible variables $\mathbf{X}$ and parameters θ. Then the expected complete-data log likelihood is

$$Q(\theta, \theta^{old}) = \int p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{Z}, \mathbf{X}|\theta) dz$$

We can approximate this integral by

$$Q(\theta, \theta^{old}) \simeq \frac{1}{L} \sum_{l=1}^{L} \ln p(\mathbf{Z}^{(l)}, \mathbf{X}|\theta)$$

- This procedure is called *Monte Carlo EM algorithm*.
- A typical side effect of this approach is lesser tendancy to get stuck into a local optima.

Stochastic EM

- A particular instance of Monte Carlo EM algorithm.
- Consider a finite mixture model, and draw just one sample at each E-step.
- The latent variable $\mathbf{Z}$ denotes the mixture membership for generating each data point.
- Essentially make a hard assignment of each data point to one of the components in the mixture.
- In the M-step, the sampled approximation to the posterior is used to update the model parameters in the usual way.
- Might take a long time to converge. But how to determine convergence?
- Sometimes, a smoothing scheme is employed.

$$Q(t) = \gamma Q(t-1) + (1-\gamma)\hat{Q}(t)$$

Stochastic EM

- A particular instance of Monte Carlo EM algorithm.
- Consider a finite mixture model, and draw just one sample at each E-step.
- The latent variable $\mathbf{Z}$ denotes the mixture membership for generating each data point.
- Essentially make a hard assignment of each data point to one of the components in the mixture.
- In the M-step, the sampled approximation to the posterior is used to update the model parameters in the usual way.
- Might take a long time to converge. But how to determine convergence?
- Sometimes, a smoothing scheme is employed.

$$Q(t) = \gamma Q(t-1) + (1-\gamma)\hat{Q}(t)$$

Stochastic EM

- A particular instance of Monte Carlo EM algorithm.
- Consider a finite mixture model, and draw just one sample at each E-step.
- The latent variable $\mathbf{Z}$ denotes the mixture membership for generating each data point.
- Essentially make a hard assignment of each data point to one of the components in the mixture.
- In the M-step, the sampled approximation to the posterior is used to update the model parameters in the usual way.
- Might take a long time to converge. But how to determine convergence?
- Sometimes, a smoothing scheme is employed.

$$Q(t) = \gamma Q(t-1) + (1-\gamma)\hat{Q}(t)$$

IP Algorithm

- Suppose we move from Maximum Likelihood approach to a full Bayesian treatment: sample form the posterior distribution $p(\theta, \mathbf{Z}|\mathbf{X})$.
- Suppose direct sample from the posterior is computationally difficult and it is relatively easy to sample from the complete-data parameter posterior $p(\theta|\mathbf{Z}, \mathbf{X})$.
- This inspires the *data augmentation algorithm* which alternates between imputation step and posterior step.

IP Algorithm

1. I-step:

$$p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z}|\theta, \mathbf{X}) p(\theta|\mathbf{X}) d\theta \tag{1}$$

   Draw $\theta^{(l)}$ from current estimate for $p(\theta|\mathbf{X})$, and then use this to draw a sample $\mathbf{Z}^{(l)}$ from $p(\mathbf{Z}|\theta^{(l)}, \mathbf{X})$.

2. P-step:

$$\begin{aligned} p(\theta|X) &= \int p(\theta|\mathbf{Z}, \mathbf{X}) p(\mathbf{Z}|\mathbf{X}) d\mathbf{Z} \\ &\simeq \frac{1}{L} \sum_{l=1}^{L} p(\theta|\mathbf{Z}^{(l)}, \mathbf{X}) \end{aligned}$$

   Use samples $\{Z^{(l)}\}$ obtained from the I-step to compute a revised estimate of the posterior distribution over θ.

Brief Summary

1. Why use sampling methods?
2. How to sample from distributions based on a uniform sample generator?
3. Rejection Sampling
4. Adaptive Rejection Sampling
5. Importance Sampling
6. Sampling-importance-resampling
7. Sampling and EM-algorithm