CS 5135/6035 Learning Probabilistic Models

Lecture 20: Monte Carlo Integration

Gowtham Atluri

November 12, 2018

Reading Material

- Chapter 3. Monte Carlo Integration
  Christian Robert and George Casella. Introducing Monte Carlo Methods with R
- Chapter 5. Monte Carlo Integration
  http://www.math.chalmers.se/Stat/Grundutb/CTH/tms150/1516/MC_20151008.pdf
- Andrieu et al. An introduction to MCMC for machine learning, Machine learning, 2003.

Topics

- Monte Carlo Integration Methods
- Probability Interpretation
- Convergence
  - Estimate convergence
  - Error in the estimate
- Importance Sampling

Integrals in Bayesian approaches

Bayesian approaches require solving integrals in different scenarios:

1. Normalization (e.g., for determining the posterior distribution)
2. Marginalization (e.g., for averaging nuisance parameters)
3. Expectation (e.g., to obtain summary statistics of the posterior)

Challenges:

- Integrals in large dimensional spaces

$$p(\theta_1|y) = \int_{\theta_2 \ldots \theta_k} p([\theta_1, \theta_2, \ldots, \theta_k]|y) d\theta_2 \ldots d\theta_k$$

- Closed form solutions to integrals are not always possible

Solution: - Monte Carlo Methods

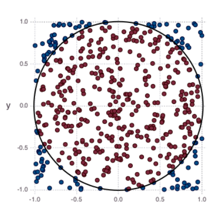
Monte Carlo Methods: a general introduction

- Monte Carlo methods are a broad class of computational algorithms
  - that rely on repeated random sampling to estimate a desired quantity

Example: Can we determine the value of π using MC method?

Approach:

1. Draw a square, and inscribe a circle in it
2. Uniformly scatter points over the square
3. Count the number of points inside the circle
4. Compute fraction of points inside the circle
   - Area of Circle/Square $= \pi r^2/(2r)^2 = \pi/4$
5. $\hat{\pi} = 4 \times$ fraction of points in circle

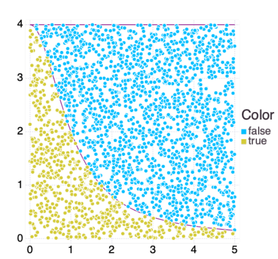


Monte Carlo Integration: Introduction

- Computing a definite integral $\int_a^b f(x)dx$ is equivalent to computing the area under the curve

$$\text{Example: compute} \int_0^5 \frac{4}{1+x^2} dx$$

- The same Monte Carlo approach for computing π applies here too!
  - We know value of integral $A_1 = \int_0^5 1dx = 5; A = 4A_1 = 20$
  - Scatter n points uniformly in the range $[0, 5]$
  - Compute proportion of points p in region of interest
  - Area under the curve is the area Ap



Monte Carlo Integration: Problem and Solution

Problem:

- We are interested in computing the value of the integral

$$I(f) = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} f(\mathbf{x}) d\mathbf{x}$$

  - $I(f)$ is a d-dimensional integral of a function f
  - $\mathbf{x}$ is a d-dimensional vector

$$I(f) = \int f(\mathbf{x}) d\mathbf{x} = \int_{x_1=x_1^{min}}^{x_1=x_1^{max}} \dots \int_{x_d=x_d^{min}}^{x_d=x_d^{max}} f(x_1, \dots, x_d) dx_1 \dots dx_d$$

Solution:

- Monte Carlo approximation of the integral $I(f)$ is given by

$$S_n = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{x}_i)$$

  - where $f(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$
  - n samples $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are drawn i.i.d. from $p(\mathbf{x})$

Monte Carlo Integration: Probability Interpretation

$$I(f) = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} f(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} g(\mathbf{x})p(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{p(x)}[g(\mathbf{x})] \approx \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{x}_i)$$

- Factorize $f(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$
- $p(\mathbf{x})$ can be interpreted as a probability density
  - $p(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x}) d\mathbf{x} = 1$
- Samples $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are drawn i.i.d. from density $p(\mathbf{x})$

Monte Carlo Integration: Probability Interpretation

$$I(f) = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} f(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} g(\mathbf{x})p(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{p(x)}[g(\mathbf{x})] \approx \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{x}_i)$$

- Factorize $f(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$
- $p(\mathbf{x})$ can be interpreted as a probability density
  - $p(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x}) d\mathbf{x} = 1$
- Samples $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are drawn i.i.d. from density $p(\mathbf{x})$
- This approach is similar to
  - *simulation approach* in nuisance parameter averaging
  - Inv-transform sampling from a mixture of distributions
  - Key difference is in factorization of $f(\mathbf{x})$
- Factorization of $f(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$ is *key* for MC to work
  - We need to find $g(\mathbf{x})$ and $p(\mathbf{x})$ such that $I(f) = \mathbb{E}_{p(x)}[g(\mathbf{x})]$

Monte Carlo Integration: Probability Interpretation

$$I(f) = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} f(\mathbf{x}) d\mathbf{x} \qquad \text{In MC integration } f(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$$

Often $p(x)$ is chosen to be Uniform

$$p(x) = \begin{cases} \frac{1}{\delta} & x^{min} \leq x \leq x^{max} \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } \delta = x^{max} - x^{min}$$

Then,

$$I(f) = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} f(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{x}^{min}}^{\mathbf{x}^{max}} g(\mathbf{x}) \frac{1}{\delta} d\mathbf{x} = \mathbb{E}_{p(x)}[g(\mathbf{x})] \approx \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{x}_i)$$

where $g(\mathbf{x}) = \delta f(\mathbf{x})$

This ($p(\mathbf{x})$ = *Uniform*) is called *ordinary* Monte Carlo Integration.

Monte Carlo Integration: Example

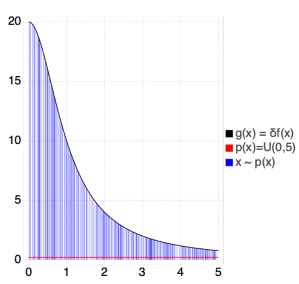
$$\text{Compute } I(f) = \int_0^5 \frac{4}{1+x^2} dx \qquad (\text{Here } d = 1)$$

Using *ordinary* MC method

- $p(x) = Uniform(0, 5) = \frac{1}{5} = \frac{1}{\delta}$ and $g(x) = \delta f(x)$
- $S_n = \frac{1}{n} \sum_{i=1}^{n} \delta f(x_i)$

Algorithm:

1. Initialize $x_1, \dots, x_n$ to 0s
2. **for** $i = 1, \dots, n$ times
3. Draw $x_i \sim U(0, 5)$
4. **end**
5. Compute $S_n = \frac{1}{n} \sum_{i=1}^{n} \delta f(x_i)$
6. Return S_n



Monte Carlo Integration: Example

$$\text{Compute } I(f) = \int_0^5 \frac{4}{1+x^2} dx \qquad (\text{Here } d = 1)$$

Using *ordinary* MC method

- $p(x) = Uniform(0, 5) = \frac{1}{5} = \frac{1}{\delta}$ and $g(x) = \delta f(x)$
- $S_n = \frac{1}{n} \sum_{i=1}^{n} \delta f(x_i)$

Algorithm:

1. Initialize $x_1, \dots, x_n$ to 0s
2. **for** $i = 1, \dots, n$ times
3. Draw $x_i \sim U(0, 5)$
4. **end**
5. Compute $S_n = \frac{1}{n} \sum_{i=1}^{n} \delta f(x_i)$
6. Return S_n

```
n=10000;
delta = 5;
f(x) = 4/(1+x^2);
x = rand(Uniform(0,5),n);
S = sum(delta.*f.(x))/n
```

```
## 5.436633068714979
```

Monte Carlo methods: Convergence

$$I(f) = \int f(\mathbf{x})d\mathbf{x} = \int g(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] \approx \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) = S_n$$

Questions:

1. Does the Monte Carlo integration method converge to the true value as larger and larger sets of samples are used?
   - We will *Law of Large Numbers* to answer this.
2. How to choose n in terms of desired accuracy and the confidence interval on the accuracy?
   - We will use *Central Limit Theorem* to answer this

Monte Carlo methods: Convergence (Q1)

$$I(f) = \int f(\mathbf{x})d\mathbf{x} = \int g(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] \approx \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) = S_n$$

- If the expectation $\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] = \mu$,

$$\mathbb{E}[S_n] = \mathbb{E}[\frac{1}{n}(g(\mathbf{x}_1) + \ldots + g(\mathbf{x}_n))] = \frac{1}{n}\mathbb{E}[g(\mathbf{x}_1) + \ldots + g(\mathbf{x}_n)] = \frac{n}{n}\mu = \mu$$

  - Expectation of S_n is the same as $\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})]$
- If the variance $Var[g(\mathbf{x})] = \sigma^2$,

$$Var[S_n] = Var[\frac{1}{n}(g(\mathbf{x}_1) + \ldots + g(\mathbf{x}_n))] = \frac{1}{n^2}Var[g(\mathbf{x}_1) + \ldots + g(\mathbf{x}_n)]$$
$$= \frac{1}{n^2}Var[g(\mathbf{x}_1)] + \ldots + Var[g(\mathbf{x}_n)] = \frac{\sigma^2 + \ldots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

  - Variance of the estimate S_n is $O(1/n)$

Monte Carlo methods: Convergence (Q1)

$$I(f) = \int f(\mathbf{x})d\mathbf{x} = \int g(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] \approx \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) = S_n$$

$$\mathbb{E}[S_n] = \mu = \mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] \qquad Var[S_n] = \frac{\sigma^2}{n} = \frac{Var[g(\mathbf{x})}{n}]$$

- Monte Carlo methods converge to the true value as $n \to \infty$.
- *Strong Law of Large Numbers:* Let $x_1, x_2, \ldots, x_n$ be i.i.d. with $\mathbb{E}[x_i] = \mu \in \mathbb{R}$, $Var(x_i) = \sigma^2 \in (0, \infty)$.

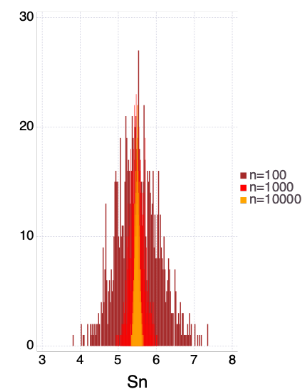
$$\text{If} \quad \bar{x}_i = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{then} \quad \bar{x}_i \to \mu$$

- LLN gives us the mean of the estimate S_n behavior when $n \to \infty$

Monte Carlo methods: Convergence (Q1)

Visualizing convergence for $I(f) = \int_0^5 \frac{4}{1+x^2}dx$, using $S_n = \frac{1}{n}\sum_{i=1}^{n} \delta f(x_i)$

```
n=[100, 1000, 10000];
delta = 5;
f(x) = 4/(1+x^2);
S1 = zeros(1000);
S2 = zeros(1000);
S3 = zeros(1000);
for i=1:1000
  x1 = rand(Uniform(0,5),n[1]);
  S1[i] = sum(delta.*f.(x1))/n[1];
  x2 = rand(Uniform(0,5),n[2]);
  S2[i] = sum(delta.*f.(x2))/n[2];
  x3 = rand(Uniform(0,5),n[3]);
  S3[i] = sum(delta.*f.(x3))/n[3];
end
plot(layer(x=S3, Geom.histogram),
     layer(x=S2, Geom.histogram),
     layer(x=S1, Geom.histogram));
```



Monte Carlo methods: Convergence (Q2)

- *Question:* How to choose n in terms of desired accuracy?
- *Approach:* We can estimate the error, for a chosen value of n, and work backwards

$$\epsilon_n = \mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)$$

- *Central Limit Theorem:*
  - Let $x_1, x_2, \ldots, x_n$ be i.i.d. with $\mathbb{E}[x_i^2] < +\infty$.
  - Let σ^2 denote the variance of x_i, i.e., $\sigma^2 = E((x_i - E(x_i))^2)$ and
  - $\epsilon_n = \mathbb{E}(x) - \frac{1}{n}\sum_{i=1}^{n} x_i$.

$$\text{then } (\frac{\sqrt{n}}{\sigma}\epsilon_n) \text{ converges in distribution to } \mathcal{N}(0,1)$$

Monte Carlo methods: Convergence (Q2)

- *Central Limit Theorem:*
  - Let $x_1, x_2, \ldots, x_n$ be i.i.d. with $\mathbb{E}[x_i^2] < +\infty$.
  - Let σ^2 denote the variance of x_i, i.e., $\sigma^2 = E((x_i - E(x_i))^2)$ and
  - $\epsilon_n = \mathbb{E}(x) - \frac{1}{n}\sum_{i=1}^{n} x_i$.

$$\text{then } (\frac{\sqrt{n}}{\sigma}\epsilon_n) \text{ converges in distribution to } \mathcal{N}(0,1)$$

- From this, it follows that for any a and b

$$\lim_{n \to +\infty} p(\frac{\sigma}{\sqrt{n}}a \le \epsilon_n \le \frac{\sigma}{\sqrt{n}}b) = \int_a^b \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx$$

- We observe that when $x \sim \mathcal{N}(0,1)$, $p(|x| \le 1.96) \approx 0.95$, using this we can say

$$|\epsilon_n| \le 1.96\frac{\sigma}{\sqrt{n}}, \quad \text{with a probability close to } 0.95$$

- Error ϵ_n is not dependent on the dimensionality of the integral d
  - It is of the order $O(1/\sqrt{n})$

Observations

- Error in the estimate of $I(f)$ for n samples is

$$|\epsilon_n| \leq 1.96 \frac{\sigma}{\sqrt{n}}, \quad \text{with a probability close to 0.95}$$

- If want to reduce the error in the estimate
  - Increase n significantly
    - when unlimited computing resources and time are available
  - (Somehow) reduce σ^2
    - useful when constraints are on computing resources and time
- Importance sampling
  - Reduces variance (σ^2)

Monte Carlo methods: Importance Sampling

- Importance Sampling is a **MC Integration** approach
  - not a *sampling approach*
- The *idea* is to sample random numbers from a density that is close to the shape of the integrand.
  - Shape of $f(\mathbf{x})$ and $q(\mathbf{x})$ should look similar, *support*$(f) \subset$ *support*(q)

$$I(f) = \int f(\mathbf{x})dx = \int \frac{f(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x})d\mathbf{x}$$

  - Choosing $q(\mathbf{x})$ requires some effort
    - $q(\mathbf{x})$ must be a probability density, i.e., $q(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x})d\mathbf{x} = 1$
- Using Monte Carlo integration on this 'factorization', we have Importance Sampling approach

Monte Carlo methods: Importance Sampling

$$I(f) = \int f(\mathbf{x})dx = \int \frac{f(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x})dx$$

Importance Sampling Approach:

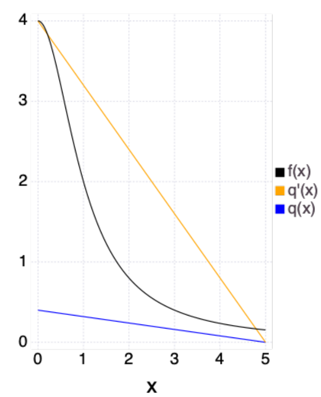
1. Initialize $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to 0s
2. **for** $i = 1, \ldots, n$ times
3. Draw $\mathbf{x}_i \sim q(\mathbf{x})$
4. **end**
5. Compute $S_n = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\mathbf{x}_i)}{q(\mathbf{x}_i)}$
6. Return S_n

Importance Sampling: Example

Compute $I(f) = \int_0^5 \frac{4}{1+x^2} dx$

$$I(f) = \int f(x)dx = \int \frac{f(x)}{q(x)} q(x)dx$$

- We need to select $q(x)$ such that
  - $q(x)$ and $f(x)$ are similar in shape
  - $q(x) \geq 0, \quad \text{for } x \in [0, 5]$
  - $\int_0^5 q(x)dx = 1$
- $q'(x) = \frac{100-20x}{25}$
- $\int q'(x)dx = 10$
- $q(x) = \frac{10-2x}{25}$



- We need to draw samples from $q(x)$ (accept-reject method?)

Importance Sampling

Accept-Reject Method

```
function accept_reject_method(n)
    x = 0:0.01:5;
    f(x) = (10-2x)/25;
    g(x) = pdf(Uniform(0,5),x);
    M = maximum(f.(x)./g.(x));
    count = 0;
    samples = [];
    while(count<n)
        y = rand(Uniform(0,5));
        u = rand(Uniform(0,1));
        if(u<f(y)/(M*g(y)))
            samples = [samples; y];
            count +=1;
        end
    end
    return samples;
end
```

Importance Sampling

1. Initialize $x_1, \ldots, x_n$ to 0s
2. **for** $i = 1, \ldots, n$ times
3. Draw $x_i \sim q(x)$
4. **end**
5. Compute $S_n = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{q(x_i)}$
6. Return S_n

```
f(x) = 4/(1+x^2);
q(x) = (10-2x)/25;

n = 10000;
x = accept_reject_method(n);
S = sum(f.(x)./(q.(x)))/length(x)
```

```
## 5.477582181147847
```


Importance Sampling: Variance reduction

- In *ordinary* MC: $I(f) = \int f(\mathbf{x})d\mathbf{x} = \int g(\mathbf{x})p(\mathbf{x})dx$
  - Variance of the estimate S_n

$$Var[S_n] = \frac{Var[g(\mathbf{x})]}{n}$$

  - In addition to n, variance depends on $Var[g(\mathbf{x})]$
- In Importance sampling:

$$I(f) = \int f(\mathbf{x})d\mathbf{x} = \int \frac{f(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x})d\mathbf{x}$$

  - Variance of the estimate is

$$Var[S_n] = \frac{Var[\frac{f(\mathbf{x})}{q(\mathbf{x})}]}{n}$$

- If the shape of q is similar to f, the ratio f/q will be (nearly) constant
  - This will keep the term $Var[\frac{f(\mathbf{x})}{q(\mathbf{x})}]$ small
  - Due to this estimate in Importance Sampling has low variance
    - when q is selected appropriately

Comparing variance *Ordinary* MC and IS

Ordinary MC Integration

```
n = 10000;
delta = 5;
f(x) = 4/(1+x^2);
S = zeros(100);
for i= 1:100
    x = rand(Uniform(0,5),n);
    S[i] = sum(delta.*f.(x))/n;
end

mean(S)
```

```
## 5.499100825974856
```

```
var(S)
```

```
## 0.0027350615959637506
```

Importance Sampling

```
n = 10000;
f(x) = 4/(1+x^2);
p(x) = (10-2x)/25;
S = zeros(100);
for i= 1:100
    x = accept_reject_method(n);
    S[i] = sum(f.(x)./(p.(x)))/n;
end

mean(S)
```

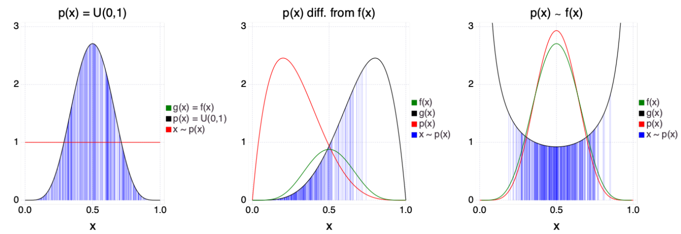
```
## 5.495252969036046
```

```
var(S)
```

```
## 0.0006740964975615764
```


Comparing variance *Ordinary* MC and IS

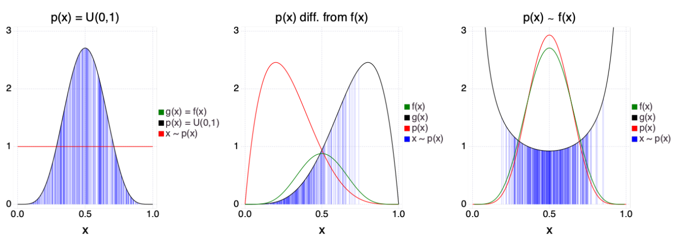
$$I(f) = \int f(x)dx = \int g(x)p(x)dx \qquad I(f) = \int f(x)dx = \int \frac{f(x)}{q(x)}q(x)dx$$



- Sampling well in places where $g(x)$ is high is critical to good approximation
- When $p(x) = Uniform$
  - Regions where $f(x)$ takes a high value are not given a priority
    - Takes more samples to get a good approximation in those regions

Comparing variance *Ordinary* MC and IS

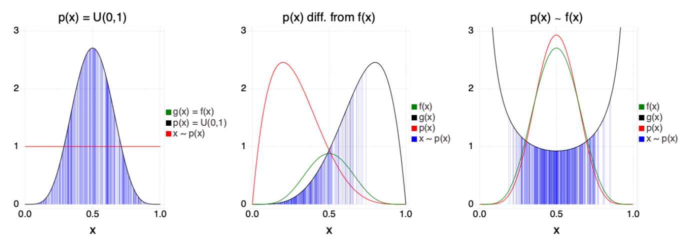
$$I(f) = \int f(x)dx = \int g(x)p(x)dx \qquad I(f) = \int f(x)dx = \int \frac{f(x)}{q(x)}q(x)dx$$



- When $p(x)$ has a shape different from $g(x)$
  - Regions where $g(x)$ is higher are poorly sampled
    - Takes a LOT of samples to get a good approximation in those regions

Comparing variance *Ordinary* MC and IS

$$I(f) = \int f(x)dx = \int g(x)p(x)dx \qquad I(f) = \int f(x)dx = \int \frac{f(x)}{q(x)}q(x)dx$$



- When $p(x)$ has a shape simular to $f(x)$
  - $g(x) = f(x)/p(x)$ is nearly a constant (when $f(x)$ takes high values)
  - Small number of samples can result in good approximation

Summary

- Monte Carlo Integration
  - Ordinary MC ($p(x) = U(a, b)$)
  - Importance Sampling ($q(x)$ has a similar shape as $f(x)$)
- Probability interpretation
- Convergence
  - Estimate converges
  - Variance of the estimate $Var(g(x))/n$
    - Depends on both $Var[g(x)]$ and n
- Importance Sampling
  - reduces variance of the estimate
  - by reducing the value of the term $Var[g(x)] = Var[\frac{f(x)}{q(x)}]$