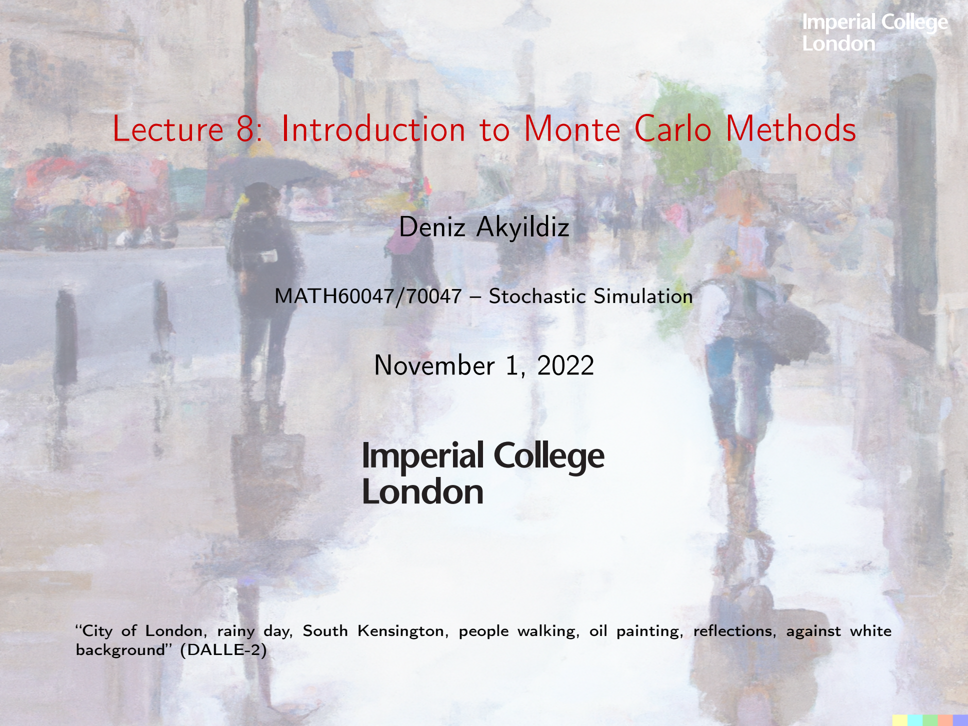


Imperial College London

Lecture 8: Introduction to Monte Carlo Methods

Deniz Akyildiz

MATH60047/70047 – Stochastic Simulation

November 1, 2022

Imperial College London

"City of London, rainy day, South Kensington, people walking, oil painting, reflections, against white background" (DALLE-2)

What we have done yesterday?

What we have done yesterday?

- Bayes rule for conditionally iid case (our *target* distributions could be posteriors)

What we have done yesterday?

- Bayes rule for conditionally iid case (our *target* distributions could be posteriors)
- Application of rejection sampling to posterior sampling and high dimensional targets (?)

What we have done yesterday?

- Bayes rule for conditionally iid case (our *target* distributions could be posteriors)
- Application of rejection sampling to posterior sampling and high dimensional targets (?)

Today, we will discuss

What we have done yesterday?

- Bayes rule for conditionally iid case (our *target* distributions could be posteriors)
- Application of rejection sampling to posterior sampling and high dimensional targets (?)

Today, we will discuss

- Marginal likelihoods

What we have done yesterday?

- Bayes rule for conditionally iid case (our *target* distributions could be posteriors)
- Application of rejection sampling to posterior sampling and high dimensional targets (?)

Today, we will discuss

- Marginal likelihoods
- An introduction to Monte Carlo methods

What we have done yesterday?

- Bayes rule for conditionally iid case (our *target* distributions could be posteriors)
- Application of rejection sampling to posterior sampling and high dimensional targets (?)

Today, we will discuss

- Marginal likelihoods
- An introduction to Monte Carlo methods
  - A set of general methods for sampling from distributions, estimating expectations, computing integrals

Recall the posterior computation:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}.$$

We did not discuss the denominator $p(y)$ in detail:

Recall the posterior computation:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}.$$

We did not discuss the denominator $p(y)$ in detail:

- It is called the *marginal likelihood*

Recall the posterior computation:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}.$$

We did not discuss the denominator $p(y)$ in detail:

- It is called the *marginal likelihood*
- It is the probability of the data y (now fixed) given the model

Recall the posterior computation:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}.$$

We did not discuss the denominator $p(y)$ in detail:

- It is called the *marginal likelihood*
- It is the probability of the data y (now fixed) given the model
- It is the normalization constant for the posterior

Recall the posterior computation:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}.$$

We did not discuss the denominator $p(y)$ in detail:

- It is called the *marginal likelihood*
- It is the probability of the data y (now fixed) given the model
- It is the normalization constant for the posterior

It is an extremely important quantity for choosing models.

We may not need to compute $p(y) = \int p(y|x)p(x)\mathrm{d}x$ for *posterior* inference as we discussed.

We may not need to compute $p(y) = \int p(y|x)p(x)\mathrm{d}x$ for *posterior* inference as we discussed.

But we will need it for *model* selection and *prediction*.

We may not need to compute $p(y) = \int p(y|x)p(x)\mathrm{d}x$ for *posterior* inference as we discussed.

But we will need it for *model* selection and *prediction*.

A simple example: Choosing the right prior.

Consider two different models:

$$\begin{aligned} x &\sim p_0(x) = \mathcal{N}(x; \mu_0, \sigma_0^2) \\ y|x &\sim \mathcal{N}(y; x, \sigma_y^2) \end{aligned}$$

and

$$\begin{aligned} x &\sim p_1(x) = \mathcal{N}(x; \mu_1, \sigma_1^2) \\ y|x &\sim \mathcal{N}(y; x, \sigma_y^2) \end{aligned}$$

Consider observing y (a single data point). Which model is more likely?

Recall that, for these models, we have computed $p(y)$ analytically before. We can compute for both models:

$$\begin{aligned}
p_0(y) &= \int p(y|x)p_0(x)\mathrm{d}x \\
&= \int \mathcal{N}(y; x, \sigma_y^2)\mathcal{N}(x; \mu_0, \sigma_0^2)\mathrm{d}x \\
&= \mathcal{N}(y; \mu_0, \sigma_0^2 + \sigma_y^2)
\end{aligned}$$

and

$$\begin{aligned}
p_1(y) &= \int p(y|x)p_1(x)\mathrm{d}x \\
&= \int \mathcal{N}(y; x, \sigma_y^2)\mathcal{N}(x; \mu_1, \sigma_1^2)\mathrm{d}x \\
&= \mathcal{N}(y; \mu_1, \sigma_1^2 + \sigma_y^2)
\end{aligned}$$

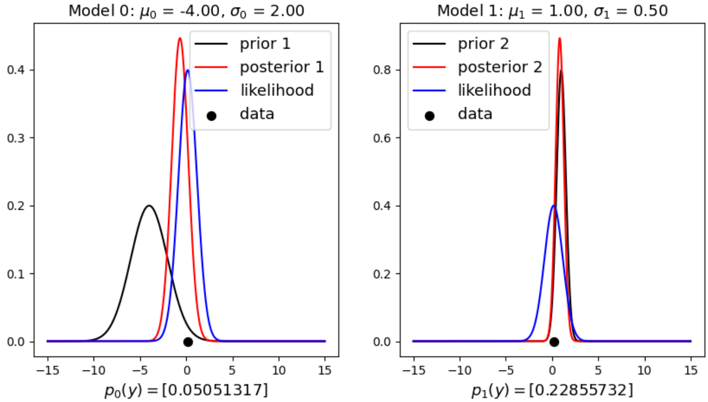
Recall that, for these models, we have computed $p(y)$ analytically before. We can compute for both models:

$$\begin{aligned} p_0(y) &= \int p(y|x)p_0(x)\mathrm{d}x \\ &= \int \mathcal{N}(y; x, \sigma_y^2)\mathcal{N}(x; \mu_0, \sigma_0^2)\mathrm{d}x \\ &= \mathcal{N}(y; \mu_0, \sigma_0^2 + \sigma_y^2) \end{aligned}$$

and

$$\begin{aligned} p_1(y) &= \int p(y|x)p_1(x)\mathrm{d}x \\ &= \int \mathcal{N}(y; x, \sigma_y^2)\mathcal{N}(x; \mu_1, \sigma_1^2)\mathrm{d}x \\ &= \mathcal{N}(y; \mu_1, \sigma_1^2 + \sigma_y^2) \end{aligned}$$

We will say Model 1 is better than Model 0 if $p_1(y) > p_0(y)$ for fixed y.


Model 0: μ_0 = -4.00, σ_0 = 2.00
prior 1
posterior 1
likelihood
data
0.0
0.1
0.2
0.3
0.4
−15
−10
−5
0
5
10
15
$p_0(y) = [0.05051317]$
Model 1: μ_1 = 1.00, σ_1 = 0.50
prior 2
posterior 2
likelihood
data
0.0
0.2
0.4
0.6
0.8
−15
−10
−5
0
5
10
15
$p_1(y) = [0.22855732]$

Bayesian no more...

- We will now look into the core concepts of the course

Bayesian no more...

- We will now look into the core concepts of the course

A large class of problems in statistics, machine learning, numerical analysis can be formulated as

Bayesian no more...

- We will now look into the core concepts of the course

A large class of problems in statistics, machine learning, numerical analysis can be formulated as

- Sampling from a distribution (e.g. a posterior, an unknown data distribution)

Bayesian no more...

- We will now look into the core concepts of the course

A large class of problems in statistics, machine learning, numerical analysis can be formulated as

- Sampling from a distribution (e.g. a posterior, an unknown data distribution)
- Estimating an expectation (e.g. the expectation of a function of a random variable)

Bayesian no more...

- We will now look into the core concepts of the course

A large class of problems in statistics, machine learning, numerical analysis can be formulated as

- Sampling from a distribution (e.g. a posterior, an unknown data distribution)
- Estimating an expectation (e.g. the expectation of a function of a random variable)
- Computing an integral

Bayesian no more...

- We will now look into the core concepts of the course

A large class of problems in statistics, machine learning, numerical analysis can be formulated as

- Sampling from a distribution (e.g. a posterior, an unknown data distribution)
- Estimating an expectation (e.g. the expectation of a function of a random variable)
- Computing an integral

Monte Carlo methods are a set of general methods for solving these problems by generating random samples (rejection sampling counts as a Monte Carlo method).

Consider a distribution $p(x)$ and estimating the expectation

Consider a distribution $p(x)$ and estimating the expectation

$$\mathbb{E}_p[X] = \int xp(x)dx \tag{1}$$

Given samples $X_1, \ldots, X_N \sim p(x)$, we can estimate the expectation as

Consider a distribution $p(x)$ and estimating the expectation

$$\mathbb{E}_p[X] = \int xp(x)dx \tag{1}$$

Given samples $X_1, \ldots, X_N \sim p(x)$, we can estimate the expectation as

$$\mathbb{E}_p[X] \approx \frac{1}{N}\sum_{i=1}^{N} X_i \tag{2}$$

The idea of estimating an expectation with random samples is more general. Indeed, for any function $\varphi(x)$, the expectation

The idea of estimating an expectation with random samples is more general. Indeed, for any function $\varphi(x)$, the expectation

$$\bar{\varphi} = \mathbb{E}_p[\varphi(x)] = \int \varphi(x) p(x) \mathrm{d}x \tag{3}$$

can be estimated using $X_1, \ldots, X_N \sim p(x)$ as

The idea of estimating an expectation with random samples is more general. Indeed, for any function $\varphi(x)$, the expectation

$$\bar{\varphi} = \mathbb{E}_p[\varphi(x)] = \int \varphi(x) p(x) \mathrm{d}x \tag{3}$$

can be estimated using $X_1, \ldots, X_N \sim p(x)$ as

$$\bar{\varphi} = \mathbb{E}_p[\varphi(x)] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi(X_i) = \hat{\varphi}^N \tag{4}$$

How to approximate probability distributions using samples? This is done via Dirac measure.

How to approximate probability distributions using samples? This is done via Dirac measure.

Define the Dirac measure:

$$\varphi(y) = \int \varphi(x)\delta_y(x)\mathrm{d}x \tag{5}$$

How to approximate probability distributions using samples? This is done via Dirac measure.

Define the Dirac measure:

$$\varphi(y) = \int \varphi(x)\delta_y(x)\mathrm{d}x \tag{5}$$

Intuition: δ_y is a measure concentrated at y: It evaluates the integral at y and is zero everywhere else. An incorrect way to write it:

$$\delta_y(x) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$$

How to approximate probability distributions using samples? This is done via Dirac measure.

Define the Dirac measure:

$$\varphi(y) = \int \varphi(x)\delta_y(x)\mathrm{d}x \tag{5}$$

Intuition: δ_y is a measure concentrated at y: It evaluates the integral at y and is zero everywhere else. An incorrect way to write it:

$$\delta_y(x) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$$

Note that, the use of Dirac only makes sense with an integral!

Given samples $X_1, \ldots, X_N \sim p(x)$, a probability distribution is approximated as

$$p(x)\mathrm{d}x \approx \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}(x)\mathrm{d}x = p^N(x)\mathrm{d}x$$

The estimator is called the *empirical* distribution.

Then, we can prove that

$$\bar{\varphi} = \mathbb{E}_p[\varphi(x)]$$

Then, we can prove that

$$\begin{aligned}\bar{\varphi} &= \mathbb{E}_p[\varphi(x)] \\ &= \int \varphi(x)p(x)\mathrm{d}x\end{aligned}$$

Then, we can prove that

$$\begin{aligned}\bar{\varphi} &= \mathbb{E}_p[\varphi(x)] \\ &= \int \varphi(x)p(x)\mathrm{d}x \\ &\approx \int \varphi(x)p^N(x)\mathrm{d}x\end{aligned}$$

Then, we can prove that

$$\begin{aligned}
\bar{\varphi} &= \mathbb{E}_p[\varphi(x)] \\
&= \int \varphi(x)p(x)\mathrm{d}x \\
&\approx \int \varphi(x)p^N(x)\mathrm{d}x \\
&= \int \varphi(x)\frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}(x)\mathrm{d}x
\end{aligned}$$

Then, we can prove that

$$
\begin{aligned}
\bar{\varphi} &= \mathbb{E}_p[\varphi(x)] \\
&= \int \varphi(x) p(x) \mathrm{d}x \\
&\approx \int \varphi(x) p^N(x) \mathrm{d}x \\
&= \int \varphi(x) \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}(x) \mathrm{d}x \\
&= \frac{1}{N} \sum_{i=1}^{N} \int \varphi(x) \delta_{X_i}(x) \mathrm{d}x
\end{aligned}
$$

Then, we can prove that

$$\begin{aligned}
\bar{\varphi} &= \mathbb{E}_p[\varphi(x)] \\
&= \int \varphi(x) p(x) \mathrm{d}x \\
&\approx \int \varphi(x) p^N(x) \mathrm{d}x \\
&= \int \varphi(x) \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}(x) \mathrm{d}x \\
&= \frac{1}{N} \sum_{i=1}^{N} \int \varphi(x) \delta_{X_i}(x) \mathrm{d}x \\
&= \frac{1}{N} \sum_{i=1}^{N} \varphi(X_i) = \hat{\varphi}^N
\end{aligned}$$

by the definition of the Dirac measure.

The same idea can be used to estimate an integral. Consider the problem:

$$I = \int f(x)\mathrm{d}x \tag{6}$$

The same idea can be used to estimate an integral. Consider the problem:

$$I = \int f(x)\mathrm{d}x \tag{6}$$

We can rephrase this problem as

$$I = \int \frac{f(x)}{p(x)} p(x)\mathrm{d}x \tag{7}$$

The same idea can be used to estimate an integral. Consider the problem:

$$I = \int f(x)\mathrm{d}x \tag{6}$$

We can rephrase this problem as

$$I = \int \frac{f(x)}{p(x)} p(x)\mathrm{d}x \tag{7}$$

Rewriting $\varphi(x) = f(x)/p(x)$ and using the estimator we derived

$$I \approx \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{p(X_i)}. \tag{8}$$

The same idea can be used to estimate an integral. Consider the problem:

$$I = \int f(x)\mathrm{d}x \tag{6}$$

We can rephrase this problem as

$$I = \int \frac{f(x)}{p(x)} p(x)\mathrm{d}x \tag{7}$$

Rewriting $\varphi(x) = f(x)/p(x)$ and using the estimator we derived

$$I \approx \frac{1}{N} \sum_{i=1}^{N} \frac{f(X_i)}{p(X_i)}. \tag{8}$$

This technique is sometimes called the Crude Monte Carlo estimator of the integral.

Claim: The estimator $\hat{\varphi}^N$ is unbiased.

Claim: The estimator $\hat{\varphi}^N$ is unbiased. *Proof:*

$$\mathbb{E}_p[\hat{\varphi}^N] \quad = \quad \mathbb{E}_p\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(X_i)\right]$$

Claim: The estimator $\hat{\varphi}^N$ is unbiased. *Proof:*

$$
\begin{aligned}
\mathbb{E}_p[\hat{\varphi}^N] &= \mathbb{E}_p\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(X_i)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_p[\varphi(X_i)]
\end{aligned}
$$

Claim: The estimator $\hat{\varphi}^N$ is unbiased. *Proof:*

$$
\begin{aligned}
\mathbb{E}_p[\hat{\varphi}^N] &= \mathbb{E}_p\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(X_i)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_p[\varphi(X_i)] \\
&= \frac{1}{N}\sum_{i=1}^{N}\int \varphi(x)p(x)\mathrm{d}x
\end{aligned}
$$

Claim: The estimator $\hat{\varphi}^N$ is unbiased. *Proof:*

$$\begin{aligned}
\mathbb{E}_p[\hat{\varphi}^N] &= \mathbb{E}_p\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(X_i)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_p[\varphi(X_i)] \\
&= \frac{1}{N}\sum_{i=1}^{N}\int \varphi(x)p(x)\mathrm{d}x \\
&= \int \varphi(x)p(x)\mathrm{d}x
\end{aligned}$$

Claim: The estimator $\hat{\varphi}^N$ is unbiased. *Proof:*

$$
\begin{aligned}
\mathbb{E}_p[\hat{\varphi}^N] &= \mathbb{E}_p\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(X_i)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_p[\varphi(X_i)] \\
&= \frac{1}{N}\sum_{i=1}^{N}\int \varphi(x)p(x)\mathrm{d}x \\
&= \int \varphi(x)p(x)\mathrm{d}x \\
&= \bar{\varphi}.
\end{aligned}
$$

Let us compute the variance of the estimator:

$$\mathsf{var}_p[\hat{\varphi}^N] \quad = \quad \mathsf{var}_p\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(X_i)\right]$$

Let us compute the variance of the estimator:

$$\begin{aligned}\mathsf{var}_p[\hat{\varphi}^N] &= \mathsf{var}_p\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(X_i)\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\mathsf{var}_p[\varphi(X_i)]\end{aligned}$$

Let us compute the variance of the estimator:

$$\begin{aligned}
\mathsf{var}_p[\hat{\varphi}^N] &= \mathsf{var}_p\left[\frac{1}{N}\sum_{i=1}^N \varphi(X_i)\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N \mathsf{var}_p[\varphi(X_i)] \\
&= \frac{1}{N^2}\sum_{i=1}^N \int (\varphi(x) - \bar{\varphi})^2 p(x)\mathrm{d}x
\end{aligned}$$

Let us compute the variance of the estimator:

$$
\begin{aligned}
\mathsf{var}_p[\hat{\varphi}^N] &= \mathsf{var}_p\left[\frac{1}{N}\sum_{i=1}^N \varphi(X_i)\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N \mathsf{var}_p[\varphi(X_i)] \\
&= \frac{1}{N^2}\sum_{i=1}^N \int (\varphi(x) - \bar{\varphi})^2 p(x)\mathrm{d}x \\
&= \frac{1}{N}\mathsf{var}_p[\varphi(X)] = \frac{\sigma_\varphi^2}{N}
\end{aligned}
$$

Provided that $\mathsf{var}_p[\varphi(X)] < \infty$ and the estimator is consistent as $N \to \infty$.

Given

$$\begin{aligned}\mathsf{var}_p[\hat{\varphi}^N] &= \frac{1}{N}\mathsf{var}_p[\varphi(X)] \\ &= \frac{1}{N}\int(\varphi(x)-\bar{\varphi})^2 p(x)\mathrm{d}x\end{aligned}$$

Given

$$
\begin{aligned}
\mathsf{var}_p[\hat{\varphi}^N] &= \frac{1}{N}\mathsf{var}_p[\varphi(X)] \\
&= \frac{1}{N}\int (\varphi(x) - \bar{\varphi})^2 p(x)\mathrm{d}x \\
&\approx \frac{1}{N^2}\sum_{i=1}^{N}(\varphi(X_i) - \hat{\varphi}^N)^2
\end{aligned}
$$

Given

$$
\begin{aligned}
\mathsf{var}_p[\hat{\varphi}^N] &= \frac{1}{N}\mathsf{var}_p[\varphi(X)] \\
&= \frac{1}{N}\int (\varphi(x) - \bar{\varphi})^2 p(x)\mathrm{d}x \\
&\approx \frac{1}{N^2}\sum_{i=1}^{N}(\varphi(X_i) - \hat{\varphi}^N)^2 \\
&= \sigma^2_{\varphi,N}.
\end{aligned}
$$

The variance can also be estimated using Monte Carlo method.

By the central limit theorem:

$$\frac{\hat{\varphi}^N - \bar{\varphi}}{\sigma_{\varphi,N}} \to \mathcal{N}(0,1). \tag{9}$$

By the central limit theorem:

$$\frac{\hat{\varphi}^N - \bar{\varphi}}{\sigma_{\varphi,N}} \to \mathcal{N}(0,1). \tag{9}$$

Therefore, this can be used to build confidence intervals (however, not very trustable).

We can see that the standard deviation of the estimator is given by

$$\mathsf{std}_p(\hat{\varphi}^N) = \frac{\sigma_\varphi}{\sqrt{N}}.$$

This is one reason you may hear much about $\mathcal{O}(1/\sqrt{N})$ convergence rate (this is optimal).

We can see that the standard deviation of the estimator is given by

$$\mathsf{std}_p(\hat{\varphi}^N) = \frac{\sigma_\varphi}{\sqrt{N}}.$$

This is one reason you may hear much about $\mathcal{O}(1/\sqrt{N})$ convergence rate (this is optimal).

More on convergence later.

Different choices of φ can be used to estimate different quantities.

- $\varphi(x) = x$ to estimate the mean

Different choices of φ can be used to estimate different quantities.

- $\varphi(x) = x$ to estimate the mean
- $\varphi(x) = x^2$ to estimate the second moment

Different choices of φ can be used to estimate different quantities.

- $\varphi(x) = x$ to estimate the mean
- $\varphi(x) = x^2$ to estimate the second moment
- $\varphi(x) = \log x$ to estimate entropy

Different choices of φ can be used to estimate different quantities.

- $\varphi(x) = x$ to estimate the mean
- $\varphi(x) = x^2$ to estimate the second moment
- $\varphi(x) = \log x$ to estimate entropy

Can we estimate probabilities?

Assume, we would like to estimate

$$\mathbb{P}(X \in A) = \int_A p(x) \mathrm{d}x \tag{10}$$

where A can be an interval, i.e., $A \subset \mathbb{R}$.

Assume, we would like to estimate

$$\mathbb{P}(X \in A) = \int_A p(x)\mathrm{d}x \tag{10}$$

where A can be an interval, i.e., $A \subset \mathbb{R}$. Then

$$\begin{aligned}\mathbb{P}(X \in A) &= \int_A p(x)\mathrm{d}x \\ &= \int \mathbf{1}_A(x)p(x)\mathrm{d}x,\end{aligned}$$

where $\mathbf{1}_A(x)$ is the indicator function of A.

$$
\begin{aligned}
\mathbb{P}(X \in A) &= \int \mathbf{1}_A(x) p(x) \mathrm{d}x, \\
&\approx \int \mathbf{1}_A(x) p^N(x) \mathrm{d}x, \\
&= \int \mathbf{1}_A(x) \frac{1}{N} \sum_{i=1}^N \delta_{X_i}(x) \mathrm{d}x, \\
&= \frac{1}{N} \sum_{i=1}^N \int \mathbf{1}_A(x) \delta_{X_i}(x) \mathrm{d}x, \\
&\approx \frac{1}{N} \sum_{i=1}^N \mathbf{1}_A(X_i).
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}(X \in A) &= \int \mathbf{1}_A(x) p(x) \mathrm{d}x, \\
&\approx \int \mathbf{1}_A(x) p^N(x) \mathrm{d}x, \\
&= \int \mathbf{1}_A(x) \frac{1}{N} \sum_{i=1}^N \delta_{X_i}(x) \mathrm{d}x, \\
&= \frac{1}{N} \sum_{i=1}^N \int \mathbf{1}_A(x) \delta_{X_i}(x) \mathrm{d}x, \\
&\approx \frac{1}{N} \sum_{i=1}^N \mathbf{1}_A(X_i).
\end{aligned}
$$

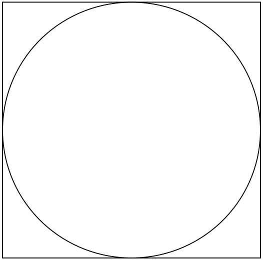
We effectively just count the samples in A and divide it by N.

$$
\begin{aligned}
\mathbb{P}(X \in A) &= \int \mathbf{1}_A(x) p(x) \mathrm{d}x, \\
&\approx \int \mathbf{1}_A(x) p^N(x) \mathrm{d}x, \\
&= \int \mathbf{1}_A(x) \frac{1}{N} \sum_{i=1}^N \delta_{X_i}(x) \mathrm{d}x, \\
&= \frac{1}{N} \sum_{i=1}^N \int \mathbf{1}_A(x) \delta_{X_i}(x) \mathrm{d}x, \\
&\approx \frac{1}{N} \sum_{i=1}^N \mathbf{1}_A(X_i).
\end{aligned}
$$

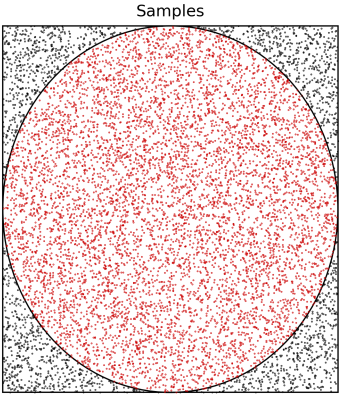
We effectively just count the samples in A and divide it by N.

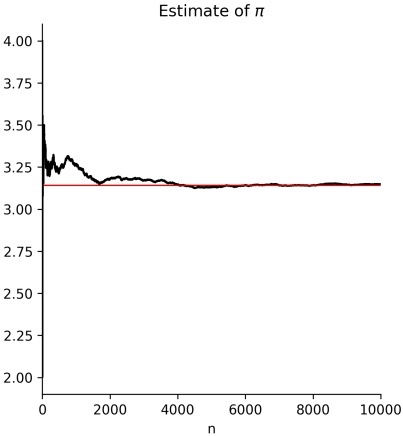
$\varphi(x) = \mathbf{1}_A(x)$ to estimate $\mathbb{P}(X \in A)$.

Our old problem:



Can we estimate the probability of this set, if we had access to samples from Unif$([-1,1] \times [-1,1])$?


Samples


Estimate of π
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2.00
0
2000
4000
6000
8000
10000
n

For probabilists, let $\mathsf{X} = [-1, 1] \times [-1, 1]$ and define the uniform measure such that $\mathbb{P}(\mathsf{X}) = 1$.

Let A be the "circle" s.t. $A \subset \mathsf{X}$. Now, the probability of A is given

$$
\begin{aligned}
\mathbb{P}(A) &= \int_A \mathbb{P}(\mathrm{d}x) \\
&= \int \mathbf{1}_A(x)\mathbb{P}(\mathrm{d}x), \\
&\approx \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_A(x_i) \to \frac{\pi}{4} \qquad \text{as } N \to \infty.
\end{aligned}
$$

where $x_i \sim \mathbb{P}$.

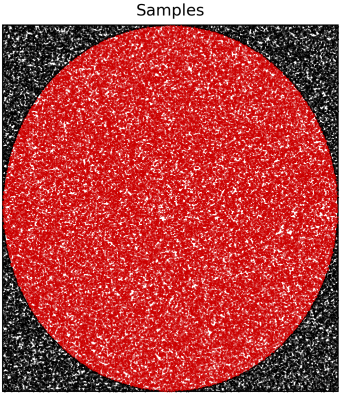
For probabilists, let $\mathsf{X} = [-1, 1] \times [-1, 1]$ and define the uniform measure such that $\mathbb{P}(\mathsf{X}) = 1$.

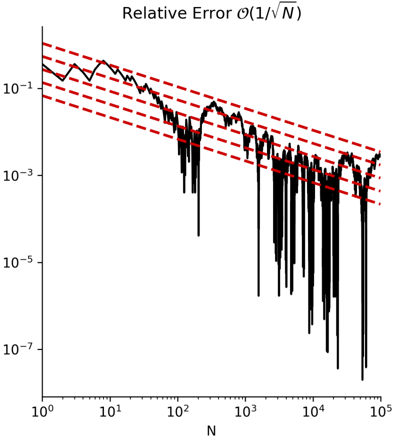
Let A be the "circle" s.t. $A \subset \mathsf{X}$. Now, the probability of A is given

$$\begin{aligned} \mathbb{P}(A) &= \int_A \mathbb{P}(\mathrm{d}x) \\ &= \int \mathbf{1}_A(x)\mathbb{P}(\mathrm{d}x), \\ &\approx \frac{1}{N}\sum_{i=1}^N \mathbf{1}_A(x_i) \to \frac{\pi}{4} \qquad \text{as } N \to \infty. \end{aligned}$$

where $x_i \sim \mathbb{P}$.

What about the convergence rate?


Samples


Relative Error $\mathcal{O}(1/\sqrt{N})$
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
N

Let us consider the integral (Example 3.4 from Robert and Casella, 2004)

$$\int_0^1 h(x)\mathrm{d}x = \int_0^1 [\cos(50x) + \sin(20x)]^2 \mathrm{d}x. \tag{11}$$

The exact value of this integral is 0.965 (yes, you can compute it).

Let us consider the integral (Example 3.4 from Robert and Casella, 2004)

$$\int_0^1 h(x)\mathrm{d}x = \int_0^1 [\cos(50x) + \sin(20x)]^2 \mathrm{d}x. \tag{11}$$

The exact value of this integral is 0.965 (yes, you can compute it).

We can use Monte Carlo integration to estimate this integral. Choose $p(x) = \mathsf{Unif}(0,1)$, then

$$\int_0^1 h(x)\mathrm{d}x = \int_0^1 h(x)p(x)\mathrm{d}x$$

so a uniform sample from $[0,1]$ is sufficient to estimate this integral.

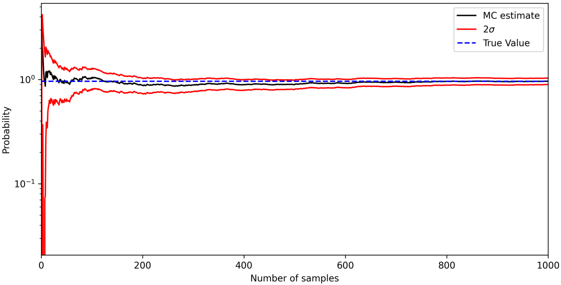


Figure: Monte Carlo integration of $h(x) = [\cos(50x) + \sin(20x)]^2$

Assume $X \sim \mathcal{N}(0, 1)$ and we are interested in

$$\mathbb{P}(X > 2) = \int_2^\infty \mathcal{N}(x; 0, 1)\mathrm{d}x. \tag{12}$$

Assume $X \sim \mathcal{N}(0, 1)$ and we are interested in

$$\mathbb{P}(X > 2) = \int_2^\infty \mathcal{N}(x; 0, 1)\mathrm{d}x. \tag{12}$$

We can formulate this as a Monte Carlo problem by choosing

$$\varphi(x) = \mathbf{1}_{\{x>2\}}. \tag{13}$$

Our estimation task is then sample $X_1, \ldots, X_n \sim \mathcal{N}(0,1)$ then

$$
\begin{aligned}
\mathbb{P}(X > 2) &= \int_2^\infty \mathcal{N}(x; 0, 1)\mathrm{d}x \\
&= \int_{-\infty}^\infty \mathbf{1}_{\{x>2\}} \mathcal{N}(x; 0, 1)\mathrm{d}x \\
&\approx \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i > 2\}}.
\end{aligned}
$$

Let us investigate the mean and the variance of this estimate against a numerical integrator (ground truth).

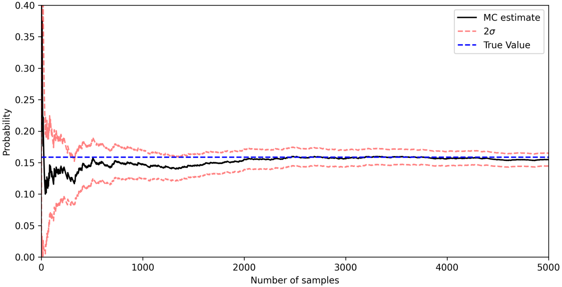


Figure: Monte Carlo estimation of the tail probability $X > 2$

In order to talk about estimators comfortably, we will now introduce various concepts of statistical estimators.

In order to talk about estimators comfortably, we will now introduce various concepts of statistical estimators.

- Error metrics (bias, variance, MSE, etc.)

In order to talk about estimators comfortably, we will now introduce various concepts of statistical estimators.

- Error metrics (bias, variance, MSE, etc.)
- Examples of estimating integrals, probabilities, and marginal likelihoods

In order to talk about estimators comfortably, we will now introduce various concepts of statistical estimators.

- Error metrics (bias, variance, MSE, etc.)
- Examples of estimating integrals, probabilities, and marginal likelihoods
- Explicit computations for variance

We will give a general review about how to compute the error of a generic estimator.

We will give a general review about how to compute the error of a generic estimator.

We will then apply these ideas to the Monte Carlo estimators.

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

We define the bias:

$$\text{bias}_p(\hat{\varphi}^N) = \mathbb{E}[\hat{\varphi}^N] - \bar{\varphi}. \tag{14}$$

In the perfect Monte Carlo case (samples exactly from p), the bias is zero.

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

We define the variance

$$\mathsf{var}_p(\hat{\varphi}^N) = \mathbb{E}_p\left[\left(\hat{\varphi}^N - \mathbb{E}_p[\hat{\varphi}^N]\right)^2\right],$$
$$= \mathbb{E}_p\left[\left(\hat{\varphi}^N\right)^2\right] - \mathbb{E}_p\left[\hat{\varphi}^N\right]^2.$$

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

We define the mean squared error:

$$\mathsf{MSE}(\hat{\varphi}^N) = \mathbb{E}[\big(\hat{\varphi}^N - \bar{\varphi}\big)^2].$$

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

We define the mean squared error:

$$\mathsf{MSE}(\hat{\varphi}^N) = \mathbb{E}[\big(\hat{\varphi}^N - \bar{\varphi}\big)^2].$$

Prove that (in-class exercise) this is also

$$\mathsf{MSE}(\hat{\varphi}^N) = \mathsf{var}_p(\hat{\varphi}^N) + \mathsf{bias}_p(\hat{\varphi}^N)^2.$$

Proof.

$$\mathsf{MSE}(\hat{\varphi}^N) \quad = \quad \mathbb{E}[\big(\hat{\varphi}^N - \bar{\varphi}\big)^2]$$

Proof.

$$\begin{aligned}\mathsf{MSE}(\hat{\varphi}^N) &= \mathbb{E}[\big(\hat{\varphi}^N - \bar{\varphi}\big)^2] \\ &= \mathbb{E}[\big(\hat{\varphi}^N\big)^2] - 2\bar{\varphi}\mathbb{E}[\hat{\varphi}^N] + \bar{\varphi}^2\end{aligned}$$

Proof.

$$
\begin{aligned}
\mathsf{MSE}(\hat{\varphi}^N) &= \mathbb{E}[\big(\hat{\varphi}^N - \bar{\varphi}\big)^2] \\
&= \mathbb{E}[\big(\hat{\varphi}^N\big)^2] - 2\bar{\varphi}\mathbb{E}[\hat{\varphi}^N] + \bar{\varphi}^2 \\
&= \mathsf{var}_p(\hat{\varphi}^N) + \mathbb{E}[\hat{\varphi}^N]^2 - 2\bar{\varphi}\mathbb{E}[\hat{\varphi}^N] + \bar{\varphi}^2
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\mathsf{MSE}(\hat{\varphi}^N) &= \mathbb{E}[\big(\hat{\varphi}^N - \bar{\varphi}\big)^2] \\
&= \mathbb{E}[\big(\hat{\varphi}^N\big)^2] - 2\bar{\varphi}\mathbb{E}[\hat{\varphi}^N] + \bar{\varphi}^2 \\
&= \mathsf{var}_p(\hat{\varphi}^N) + \mathbb{E}[\hat{\varphi}^N]^2 - 2\bar{\varphi}\mathbb{E}[\hat{\varphi}^N] + \bar{\varphi}^2 \\
&= \mathsf{var}_p(\hat{\varphi}^N) + \big(\mathbb{E}[\hat{\varphi}^N] - \bar{\varphi}\big)^2,
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\mathsf{MSE}(\hat{\varphi}^N) &= \mathbb{E}[\left(\hat{\varphi}^N - \bar{\varphi}\right)^2] \\
&= \mathbb{E}[\left(\hat{\varphi}^N\right)^2] - 2\bar{\varphi}\mathbb{E}[\hat{\varphi}^N] + \bar{\varphi}^2 \\
&= \mathsf{var}_p(\hat{\varphi}^N) + \mathbb{E}[\hat{\varphi}^N]^2 - 2\bar{\varphi}\mathbb{E}[\hat{\varphi}^N] + \bar{\varphi}^2 \\
&= \mathsf{var}_p(\hat{\varphi}^N) + \left(\mathbb{E}[\hat{\varphi}^N] - \bar{\varphi}\right)^2, \\
&= \mathsf{var}_p(\hat{\varphi}^N) + \mathsf{bias}_p(\hat{\varphi}^N)^2.
\end{aligned}
$$

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

We define the root mean squared error:

$$\mathsf{RMSE}(\hat{\varphi}^N) = \sqrt{\mathsf{MSE}(\hat{\varphi}^N)}.$$

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

We define the root mean squared error:

$$\mathsf{RMSE}(\hat{\varphi}^N) = \sqrt{\mathsf{MSE}(\hat{\varphi}^N)}.$$

For the **unbiased case**,

$$\mathsf{RMSE}(\hat{\varphi}^N) = \sqrt{\mathsf{var}_p(\hat{\varphi}^N)} = \mathsf{std}_p(\hat{\varphi}^N).$$

Note though we will have biased cases.

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

We define the root mean squared error:

$$\mathsf{RMSE}(\hat{\varphi}^N) = \sqrt{\mathsf{MSE}(\hat{\varphi}^N)}.$$

For the **unbiased case**,

$$\mathsf{RMSE}(\hat{\varphi}^N) = \sqrt{\mathsf{var}_p(\hat{\varphi}^N)} = \mathsf{std}_p(\hat{\varphi}^N).$$

Note though we will have biased cases. Do not use them interchangeably.

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

We generally plot

$$\mathsf{RAE}(\hat{\varphi}^N) = \frac{|\hat{\varphi}^N - \bar{\varphi}|}{|\bar{\varphi}|}.$$

Note that this is *random*, as no expectations are taken.

Consider

- the target quantity: $\bar{\varphi}$
- the estimator: $\hat{\varphi}^N$

We generally plot

$$\mathsf{RAE}(\hat{\varphi}^N) = \frac{|\hat{\varphi}^N - \bar{\varphi}|}{|\bar{\varphi}|}.$$

Note that this is *random*, as no expectations are taken.

It can be proven that the random error has

$$|\hat{\varphi}^N - \bar{\varphi}| \leq \mathcal{O}(U/\sqrt{N}),$$

see, e.g., Akyildiz, 2019, Corollary 2.1 for a proof.

Consider the following model:

$$
\begin{aligned}
p(x) &= \mathcal{N}(x; \mu_0, \sigma_0^2), \\
p(y|x) &= \mathcal{N}(y; x, \sigma^2).
\end{aligned}
$$

Assume that we have observed $y = 4$ and want to estimate the marginal likelihood $p(y)$.

Consider the following model:

$$p(x) = \mathcal{N}(x; \mu_0, \sigma_0^2),$$
$$p(y|x) = \mathcal{N}(y; x, \sigma^2).$$

Assume that we have observed $y = 4$ and want to estimate the marginal likelihood $p(y)$.

From our past experience, we have access to this density:

$$\begin{aligned} p(y) &= \int p(y|x)p(x)\mathrm{d}x, \\ &= \mathcal{N}(y; \mu_0, \sigma_0^2 + \sigma^2). \end{aligned}$$

Therefore, we know exactly the value of $p(4)$ (i.e. $p(y = 4)$).

Given that the problem is an integral w.r.t. $p(x)$,

$$p(y) = \int p(y|x)p(x)\mathrm{d}x,$$

we could use our Monte Carlo estimator to get unbiased estimates of $p(y)$.

Given that the problem is an integral w.r.t. $p(x)$,

$$p(y) = \int p(y|x)p(x)\mathrm{d}x,$$

we could use our Monte Carlo estimator to get unbiased estimates of $p(y)$.

Write down the estimator (in-class exercise).

Given that the problem is an integral w.r.t. $p(x)$,

$$p(y) = \int p(y|x)p(x)\mathrm{d}x,$$

we could use our Monte Carlo estimator to get unbiased estimates of $p(y)$.

Given that the problem is an integral w.r.t. $p(x)$,

$$p(y) = \int p(y|x)p(x)\mathrm{d}x,$$

we could use our Monte Carlo estimator to get unbiased estimates of $p(y)$.

Note here that $\varphi(x) = p(y|x)$ is a deterministic function of x for fixed y. Therefore, given $X_1, \ldots, X_N \sim p(x)$

$$\int \varphi(x)p(x)\mathrm{d}x \approx \frac{1}{N}\sum_{i=1}^{N}\varphi(X_i) = \frac{1}{N}\sum_{i=1}^{N}p(y|X_i).$$

For $y = 4$, we can compute this (given μ_0, σ_0, σ).

Set

$$\mu_0 = 0, \quad \sigma_0 = 1, \quad \sigma = 2, \quad y = 1$$

and generate $N = 100000$ samples from $p(x)$. We estimate

$$p^N(y) = \frac{1}{N} \sum_{i=1}^{N} p(y|X_i).$$

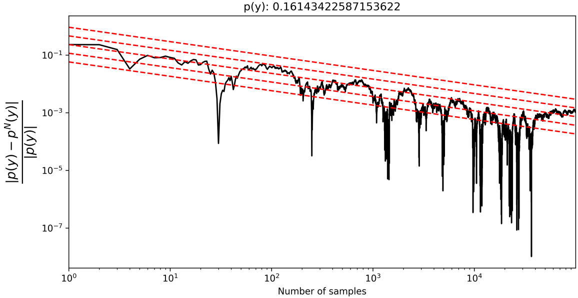


Figure: Estimating the marginal likelihood $p(y)$ for $y = 1$

Let

$$p(x) = \frac{1}{\pi(1+x^2)},$$

and let

$$\varphi(x) = \mathbf{1}_{\{x \geq 2\}}.$$

Let

$$p(x) = \frac{1}{\pi(1 + x^2)},$$

and let

$$\varphi(x) = \mathbf{1}_{\{x \geq 2\}}.$$

We would like to compute

$$\begin{aligned}
\mathbb{P}(X > 2) &= \int_2^\infty p(x) \mathrm{d}x \\
&= \int \mathbf{1}_{\{x > 2\}}(x) p(x) \mathrm{d}x \\
&= \int \varphi(x) p(x) \mathrm{d}x.
\end{aligned}$$

We can compute the real value of this integral:

$$I = \bar{\varphi} = \int_2^{\infty} p(x)\mathrm{d}x = F(\infty) - F(2) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}(2) = 0.1476$$

Compute the variance of the Monte Carlo estimator for $N = 10$ samples.

We can compute the real value of this integral:

$$I = \bar{\varphi} = \int_2^\infty p(x)\mathrm{d}x = F(\infty) - F(2) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}(2) = 0.1476$$

Compute the variance of the Monte Carlo estimator for $N = 10$ samples.

$$\mathsf{var}(\hat{\varphi}^N) = \frac{\mathsf{var}_p(\varphi)}{N}$$

So we need to compute:

$$\mathsf{var}_p(\varphi)$$

Now, we compute

$$\begin{aligned}
\mathsf{var}_p(\varphi) &= \int \varphi(x)^2 p(x)\mathrm{d}x - \left(\int \varphi(x)p(x)\mathrm{d}x\right)^2 \\
&= \int \mathbf{1}_{\{x>2\}}(x)^2 p(x)\mathrm{d}x - \left(\int \mathbf{1}_{\{x>2\}}(x)p(x)\mathrm{d}x\right)^2 \\
&= \int \mathbf{1}_{\{x>2\}}(x) p(x)\mathrm{d}x - \left(\int \mathbf{1}_{\{x>2\}}(x)p(x)\mathrm{d}x\right)^2 \\
&= 0.1476 - 0.1476^2 = 0.125.
\end{aligned}$$

Now, we compute

$$\begin{aligned}
\mathsf{var}_p(\varphi) &= \int \varphi(x)^2 p(x)\mathrm{d}x - \left(\int \varphi(x)p(x)\mathrm{d}x\right)^2 \\
&= \int \mathbf{1}_{\{x>2\}}(x)^2 p(x)\mathrm{d}x - \left(\int \mathbf{1}_{\{x>2\}}(x)p(x)\mathrm{d}x\right)^2 \\
&= \int \mathbf{1}_{\{x>2\}}(x) p(x)\mathrm{d}x - \left(\int \mathbf{1}_{\{x>2\}}(x)p(x)\mathrm{d}x\right)^2 \\
&= 0.1476 - 0.1476^2 = 0.125.
\end{aligned}$$

The variance of the estimator then

$$\mathsf{var}(\hat{\varphi}^N) = \frac{0.125}{10} = 0.0125.$$

Could we do better? An idea, the density is symmetric around zero

- This means $P(X > 2) = P(X < -2)$.

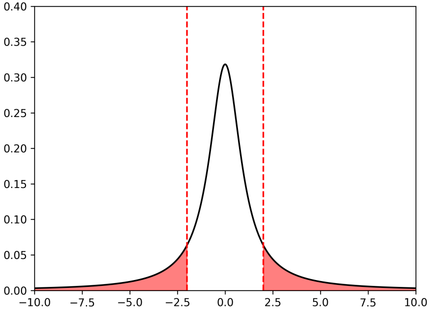


Figure: Cauchy density

We could compute:

$$\mathbb{P}(|X| > 2) = \mathbb{P}(X > 2) + \mathbb{P}(X < -2) = 2I.$$

Therefore, our new problem is $I = \frac{1}{2}\mathbb{P}(|X| > 2)$.

We could compute:

$$\mathbb{P}(|X| > 2) = \mathbb{P}(X > 2) + \mathbb{P}(X < -2) = 2I.$$

Therefore, our new problem is $I = \frac{1}{2}\mathbb{P}(|X| > 2)$.

Write it as

$$\begin{aligned} I &= \frac{1}{2}\int_{|x|>2} p(x)\mathrm{d}x, \\ &= \int \frac{1}{2}\mathbf{1}_{\{|x|>2\}}(x)p(x)\mathrm{d}x, \end{aligned}$$

Now define the test function

$$\varphi(x) = \frac{1}{2}\mathbf{1}_{\{|x|>2\}}(x).$$

We need to compute $\mathrm{var}_p(\varphi)$.

We need to compute $\mathsf{var}_p(\varphi)$.

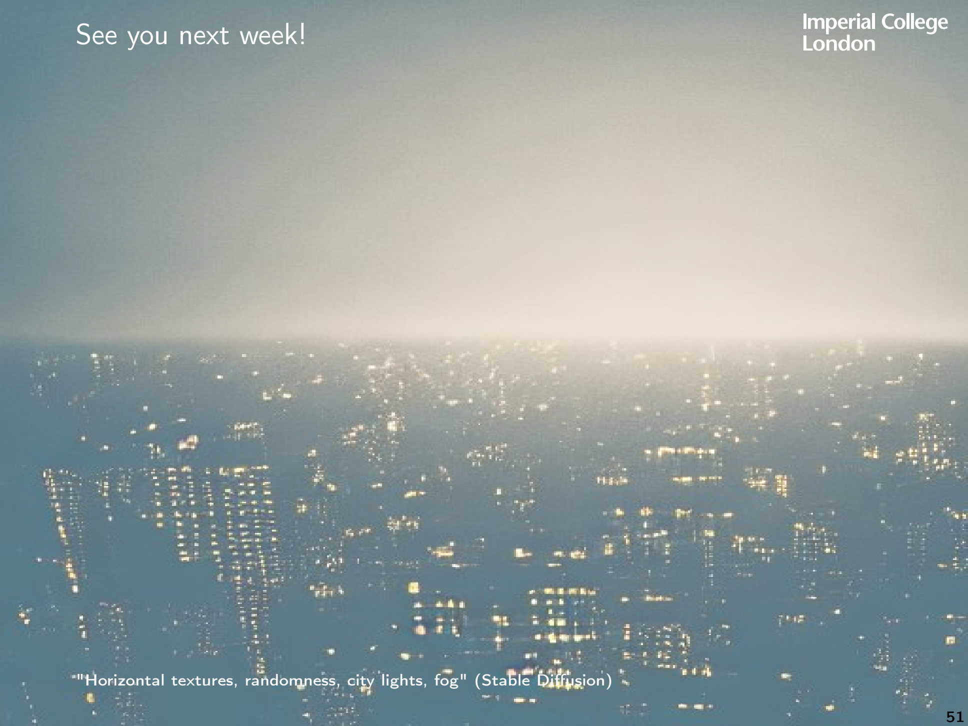
$$\begin{aligned}
\mathsf{var}_p(\varphi) &= \int \varphi(x)^2 p(x)\mathrm{d}x - \left(\int \varphi(x)p(x)\mathrm{d}x\right)^2 \\
&= \int \frac{1}{4}\mathbf{1}^2_{\{|x|>2\}}p(x)\mathrm{d}x - \left(\int \frac{1}{2}\mathbf{1}_{\{|x|>2\}}p(x)\mathrm{d}x\right)^2 \\
&= \int \frac{1}{4}\mathbf{1}_{\{|x|>2\}}p(x)\mathrm{d}x - \left(\int \frac{1}{2}\mathbf{1}_{\{|x|>2\}}p(x)\mathrm{d}x\right)^2 \\
&= \frac{1}{4} \times 2 \times 0.1476 - \frac{1}{4} \times (2 \times 0.1476)^2, \\
&= 0.052.
\end{aligned}$$

Therefore, the variance of the estimator for $N = 10$ samples is

$$\mathsf{var}(\hat{\varphi}^N) = \frac{0.052}{10} = 0.0052.$$

Improvement over the previous estimator! That was 0.0125.

See you next week!



"Horizontal textures, randomness, city lights, fog" (Stable Diffusion)

References I

- Robert, Christian P and George Casella (2004). *Monte Carlo statistical methods*. Springer.
- Akyildiz, Omer Deniz (2019). "Sequential and adaptive Bayesian computation for inference and optimization". Can be accessed from: http://akyildiz.me/works/thesis.pdf. PhD thesis. Universidad Carlos III de Madrid.