4

MONTE CARLO INTEGRATION

In this section, we introduce Monte Carlo integration and importance sampling in detail. We will show how these ideas can be applied to a variety of problems such as computing integrals, computing expectations, sampling from complex distributions, and computing marginal likelihoods.

4.1 INTRODUCTION

We have repeatedly highlighted that we are interested in sampling from a probability measure p. One reason we are interested in this is to *estimate expectations* of certain measures, i.e., we can estimate moments of distributions. Of course, so far, we have been considering drawing samples from known distributions (for which moments might be readily available). However, it is often the case in sampling applications that, in most cases, the primary goal *is* to compute expectations for distributions which are not available to us in closed form.

We will call this task as *Monte Carlo integration*. Briefly, given a probability distribution p, we are interested in computing expectations of the form

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

where $\varphi(x)$ is called a *test function*. For example, $\varphi(x) = x$ would give us the mean, $\varphi(x) = x^2$ the second moment, or $\varphi(x) = \log(x)$ would give us the entropy. For example, given $X^{(1)}, \ldots, X^{(N)} \sim p$ i.i.d, we know that (intuitively, at this point) the mean estimator is given by

$$\mathbb{E}_p[X] = \int xp(x)\mathrm{d}x \approx \frac{1}{N}\sum_{i=1}^{N} X^{(i)},$$

which is simply the empirical average of the samples. While this can be intuitive, it underlies a certain choice about the approximation of the probability distribution p using its samples.

In order to do this, we build an *empirical distribution* of the samples, using

$$p^N(x)\mathrm{d}x = \frac{1}{N}\sum_{i=1}^{N}\delta_{X^{(i)}}(x)\mathrm{d}x. \tag{4.1}$$

In order to understand how this works, we first need to understand the Dirac delta measure δ_x. The Dirac delta measure is defined as

$$f(y) = \int f(x)\delta_y(x)\mathrm{d}x. \tag{4.2}$$

Here, the Dirac can be thought as a *point mass* at y. In other words, the Dirac delta measure is a measure which is concentrated at a single point. To understand it intuitively, the object $\delta_y(x)$ can be informally thought as a function centered at y (and only takes value 1 at y) [1]

$$\delta_y(x) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise.} \end{cases}$$

One can see that then p^N is a *sample based approximation* of p, where the samples are equally weighted. While we never may use this particular approximation of a density, it is useful to build estimates of expectations. Generalising the above scenario, let us consider the estimation of the general expectation

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

Given samples $X^{(1)}, \ldots, X^{(N)}$, we can build p^N as in (4.1) and approximate this expectation as

$$\begin{aligned}
\bar{\varphi} &= \mathbb{E}_p[\varphi(x)] \\
&= \int \varphi(x)p(x)\mathrm{d}x \\
&\approx \int \varphi(x)p^N(x)\mathrm{d}x \\
&= \int \varphi(x)\frac{1}{N}\sum_{i=1}^{N}\delta_{X_i}(x)\mathrm{d}x \\
&= \frac{1}{N}\sum_{i=1}^{N}\int \varphi(x)\delta_{X_i}(x)\mathrm{d}x \\
&= \frac{1}{N}\sum_{i=1}^{N}\varphi(X_i) = \hat{\varphi}^N.
\end{aligned} \tag{4.3}$$

where we have used (4.2) in the approximate integral to arrive at the final expression. Note that this generalises the example above about the mean (which was $\varphi(x) = x$ case). In this course, we will also be interested in the properties of these estimators.

[1] This is not correct rigorously – just for intuition! Note that the Diracs always make sense with an integral attached to them.

Remark 4.1. As we can see that, the Monte Carlo estimator can be used to estimate expectations. We can also use this idea to estimate integrals. Consider a standard integration problem

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

where $f(x)$ is a function. We can use the Monte Carlo (MC) estimator to estimate this integral as

$$\begin{aligned} I &= \int \frac{f(x)}{p(x)} p(x)\mathrm{d}x \\ &\approx \int \frac{f(x)}{p(x)} p^N(x)\mathrm{d}x \qquad \text{where } p^N(x)\mathrm{d}x = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(x)\mathrm{d}x \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{f(X_i)}{p(X_i)}. \qquad \text{using (4.1)} \end{aligned}$$

In this case, we have $\varphi(x) = \frac{f(x)}{p(x)}$. This is particularly easy for the integrals of type

$$I = \int_0^1 f(x)\mathrm{d}x,$$

where $f(x)$ is a function. In this case, we can use the uniform distribution as the base distribution p and use the Monte Carlo estimator to estimate the integral without needing to compute any ratios.

In the following, we prove some results about the properties of the Monte Carlo estimator (4.3) when samples are i.i.d from p.

Proposition 4.1. *Let $X_1, \ldots, X_N$ be i.i.d samples from p. Then, the Monte Carlo estimator*

$$\hat{\varphi}^N = \frac{1}{N}\sum_{i=1}^{N} \varphi(X_i)$$

is unbiased, i.e.,

$$\mathbb{E}_p[\hat{\varphi}^N] = \bar{\varphi}.$$

Proof. We have

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

which proves the result. □

Next, we can also compute the variance of the Monte Carlo estimator.

Proposition 4.2. *Let $X_1, \ldots, X_N$ be i.i.d samples from p. Then, the Monte Carlo estimator*

$$\hat{\varphi}^N = \frac{1}{N}\sum_{i=1}^{N}\varphi(X_i)$$

has variance

$$\mathsf{var}_p[\hat{\varphi}^N] = \frac{1}{N}\mathsf{var}_p[\varphi(X)].$$

where

$$\mathsf{var}_p[\varphi(X)] = \int(\varphi(x) - \bar{\varphi})^2 p(x)\mathrm{d}x.$$

Proof. We have

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

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

Remark 4.2. The expression $\mathsf{var}_p[\hat{\varphi}^N]$ is the variance of the MC estimator but this expression requires the true mean $\bar{\varphi}$ to be known. In practice, we do not know the true mean but also have an MC estimator for it. We can plug this estimator into the variance in order to obtain an empirical variance estimator. Note that

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

This estimator then can be used to estimate the variance of the MC estimator.

We can therefore obtain a central limit theorem for our estimator, i.e.,

$$\frac{(\hat{\varphi}^N-\bar{\varphi})}{\sigma_{\varphi,N}} \to \mathcal{N}(0,1) \qquad \text{as} \qquad N\to\infty.$$

This can be used to build empirical confidence intervals for the estimators. However, this is not a principled estimate and may not be valid in many scenarios. We can also see that we have a standard deviation estimate (which follows from the variance estimate) given by

$$\mathsf{std}_p[\hat{\varphi}^N] = \sqrt{\mathsf{var}_p[\hat{\varphi}^N]} = \frac{\sigma_\varphi}{\sqrt{N}}.$$

This is a typical display of a convergence rate $\mathcal{O}(1/\sqrt{N})$.

Remark 4.3. One of the most common application of sampling is to estimate probabilities. We have seen that different choices of φ can lead to estimating different quantities such as the mean and nth moments. However, the MC estimators can also be used to estimate probabilities. In order to see this, assume that we would like to estimate $\mathbb{P}(X\in A)$ where $X\sim p$. We know that this is given as

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

see, e.g., Definition 3.2. For example, A can simply be an interval. Given the definition above, we can write

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

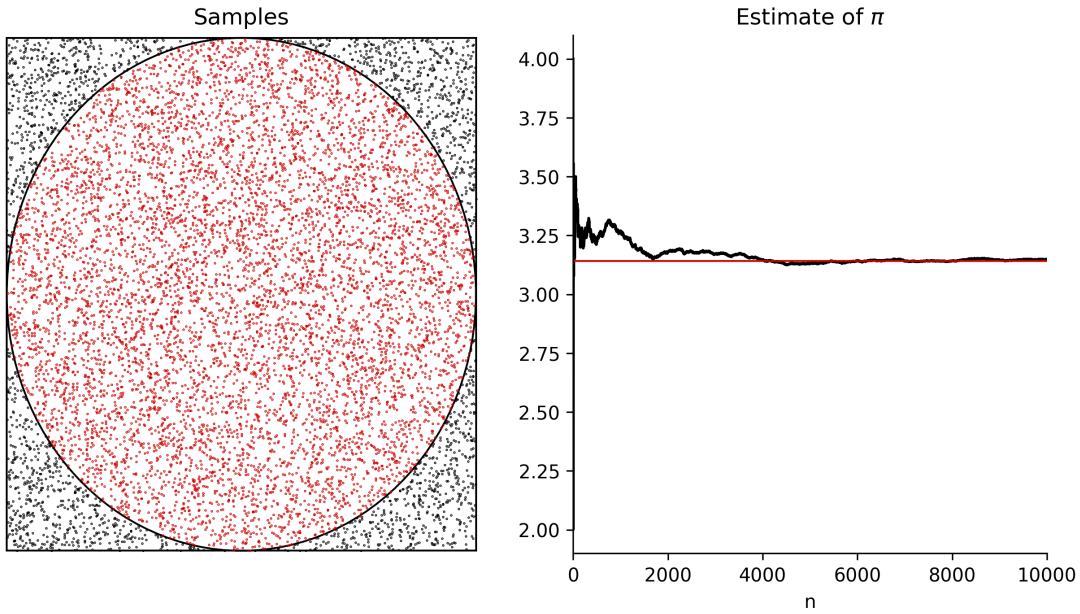

Figure 4.1: Estimating π using the Monte Carlo method.

where $\mathbf{1}_A(x)$ is the indicator function of A. We can therefore set $\varphi(x) = \mathbf{1}_A(x)$ and given the samples from p, we can build an estimator

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

This estimator also leads to an intuitive procedure: We sample $X_1, \ldots, X_N$ from p and we effectively just count the samples in A and divide it by N.

We can now return to the example of estimating π using the Monte Carlo method.

Example 4.1. We can recall the problem of estimating π using the Monte Carlo method. The logic that was used in this example was to estimate the area of a circle that lies within a square. To be precise, consider the square $[-1, 1] \times [-1, 1]$ and define the uniform distribution on this square as $p(x, y) = \mathrm{Unif}([-1, 1] \times [-1, 1])$. We can simply phrase the problem as estimating the area of the circle which we define as $\mathsf{A} \subset [-1, 1] \times [-1, 1]$. The

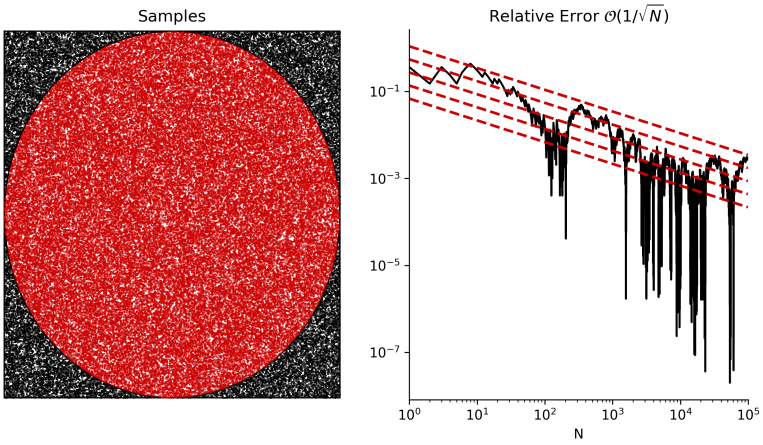

Figure 4.2: Relative error (see next section) of the Monte Carlo estimate provided by sampling within the circle.

set A is given as

$$\mathsf{A} = \{(x, y) \in [-1, 1] \times [-1, 1] \mid x^2 + y^2 \leq 1\}.$$

We can then formalise this problem as estimating the probability that a point lies within the circle. This is given as

$$\begin{aligned} \mathbb{P}(X \in \mathsf{A}) &= \int_{\mathsf{A}} p(x, y) \mathrm{d}x \mathrm{d}y, \\ &= \int \mathbf{1}_{\mathsf{A}}(x, y) p(x, y) \mathrm{d}x \mathrm{d}y. \end{aligned}$$

Sampling $(X_i, Y_i) \sim p(x, y)$ (a uniform sample within a square), we can estimate this integral using the standard MC method. More formally, we can write

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

A trajectory of the estimation procedure π can be seen from Fig. 4.1 w.r.t. varying sample size.

Nonasymptotic results showing the convergence rate of $\mathcal{O}(1/\sqrt{N})$ are also available (see, e.g., Akyildiz (2019, Corollary 2 .1)) – see Fig. 4.2 for a demonstration.

Example 4.2 (Example 3.4 from Robert and Casella (2004))**.** Let us consider an example of

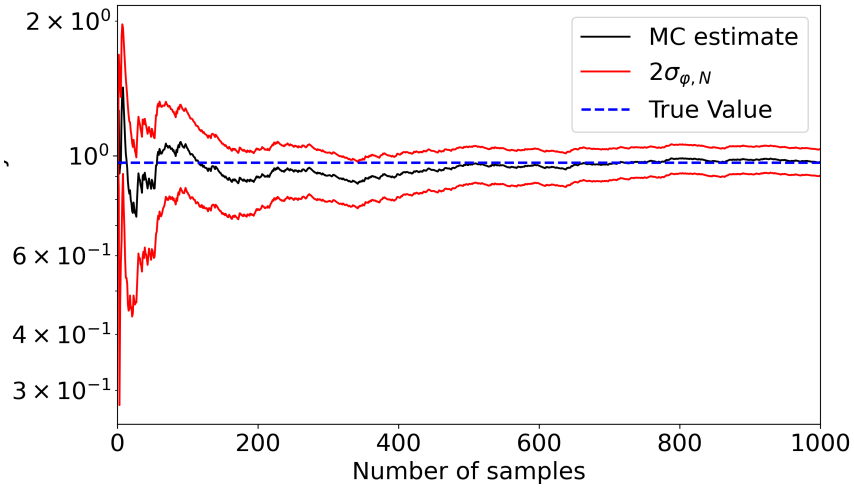

Figure 4.3: Monte Carlo integration of $h(x) = [\cos(50x) + \sin(20x)]^2$

estimating an integral:

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

The exact value of this integral is 0.965. We can use the MC method to estimate this integral. We can just choose $p(x) = \mathrm{Unif}(0, 1)$ and set $\varphi(x) = h(x)$. We can then write

$$\begin{aligned} I &= \int_0^1 h(x)\mathrm{d}x, \\ &= \int_0^1 \varphi(x)p(x)\mathrm{d}x, \end{aligned}$$

and apply the standard MC estimator. The results (together with the empirical variance estimate) can be seen from Fig. 4.3.

Finally, we provide an example of estimating the probability of a random variable.

Example 4.3. Consider $X \sim \mathcal{N}(0, 1)$ and we would like to estimate the probability that $X > 2$. The way to do this is to choose

$$p(x) = \mathcal{N}(0, 1), \quad \varphi(x) = \mathbf{1}_{\{x>2\}}(x).$$

We can then write that

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

where $X_1, \ldots, X_N \sim \mathcal{N}(0, 1)$. The results can be seen from Fig. 4.4.

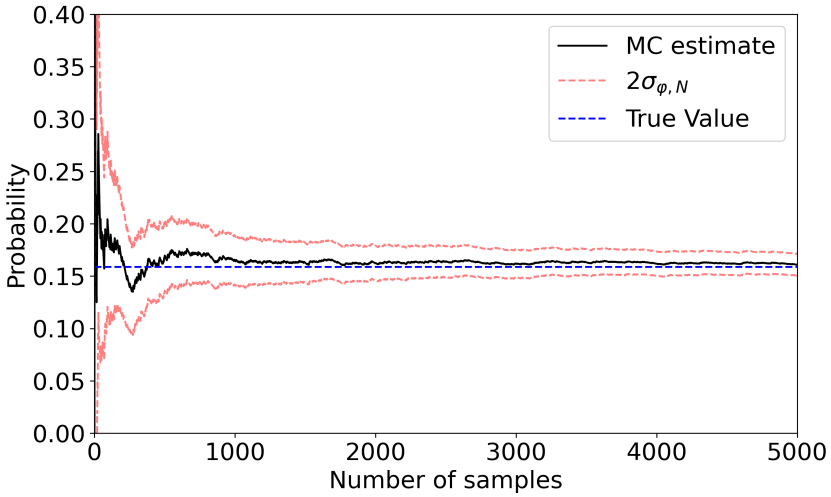

Figure 4.4: Monte Carlo estimation of the tail probability $X > 2$. The "true value" is computed via numerical integration.

4.2 ERROR METRICS

In order to quantify convergence, we will have a number of error metrics in the following section. We start with defining the bias as

$$\mathsf{bias}(\hat{\varphi}^N) = \mathbb{E}[\hat{\varphi}^N] - \bar{\varphi}, \tag{4.4}$$

where $\bar{\varphi}$ is the *true value*. We call an estimator *unbiased* if the bias is zero. In the case where we sample i.i.d from $p(x)$, we can build unbiased estimators of expectations and integrals. We recall the variance

$$\mathsf{var}(\hat{\varphi}^N) = \mathbb{E}[(\hat{\varphi}^N - \mathbb{E}[\hat{\varphi}^N])^2]. \tag{4.5}$$

If the estimator is unbiased, we can then replace $\mathbb{E}[\hat{\varphi}^N$ with $\bar{\varphi}$. Next, we define the mean squared error (MSE)

$$\mathsf{MSE}(\hat{\varphi}^N) = \mathbb{E}[(\hat{\varphi}^N - \bar{\varphi})^2]. \tag{4.6}$$

One can see that the MSE and the variance coincides if the estimator is unbiased. We have also the following decomposition of the MSE

$$\mathsf{MSE}(\hat{\varphi}^N) = \mathsf{bias}(\hat{\varphi}^N)^2 + \mathsf{var}(\hat{\varphi}^N). \tag{4.7}$$

We can define the root mean square error (RMSE) as

$$\mathsf{RMSE}(\hat{\varphi}^N) = \sqrt{\mathsf{MSE}(\hat{\varphi}^N)}. \tag{4.8}$$

Finally, we define the relative absolute error (RAE) as

$$\mathsf{RAE}(\hat{\varphi}^N) = \frac{|\hat{\varphi}^N - \bar{\varphi}|}{|\bar{\varphi}|}. \tag{4.9}$$

We usually plot the absolute error of the estimator, as we only run the experiment once in general[2]. We note that this absolute error $|\hat{\varphi}^N - \bar{\varphi}|$ is a random variable (since no

[2]However, if you were to do a proper experimentation, then you would have to run the same experiment M times (Monte Carlo runs) and average the error to estimate the RMSE.

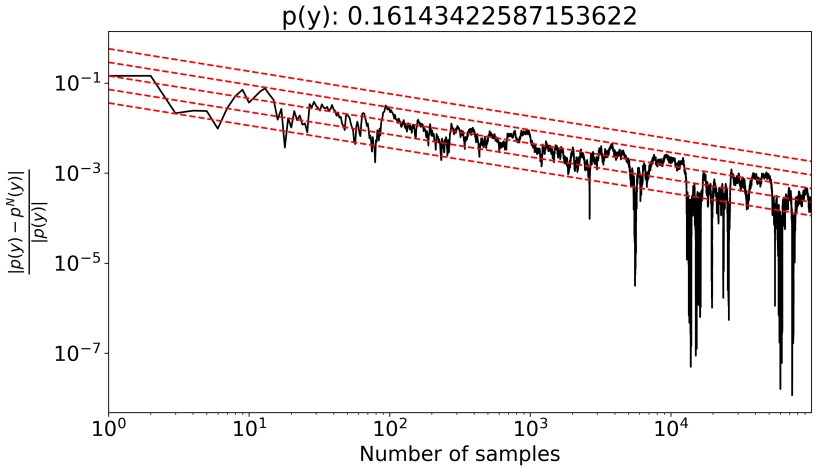

Figure 4.5: Estimating the marginal likelihood $p(y)$ for $y = 1$. One can clearly see the displayed error rate that is $\mathcal{O}(1/\sqrt{N})$.

expectations are taken). However, this quantity provably converges with a rate of $\mathcal{O}(1/\sqrt{N})$ (see, e.g., Akyildiz (2019, Corollary 2.1)). More precisely, we can write

$$|\hat{\varphi}^N - \bar{\varphi}| \leq \frac{V}{\sqrt{N}}, \tag{4.10}$$

where V is an almost surely finite random variable. This error rate will be displayed empirically in the following sections (see also Fig . 4.2).

Example 4.4 (Marginal Likelihood estimation)**.** Recall that, given a prior $p(x)$ and a likelihood $p(y|x)$, we can compute the marginal likelihood $p(y)$ as

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

This defines a nice integration problem that we can solve using MC. Assume that we are given the following model

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

Assume that $\mu_0 = 0$, $\sigma_0 = 1$, $\sigma = 2$, and $y = 1$. For fixed $p(y = 1)$, this integral becomes

$$p(y = 1) = \int p(y = 1|x)p(x)\mathrm{d}x,$$

where we can set $\varphi(x) = p(y = 1|x)$. We can then compute the integral using MC estimation procedure as

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

where $X_1, \ldots, X_N \sim p(x)$. The results can be seen from Fig. 4.5.

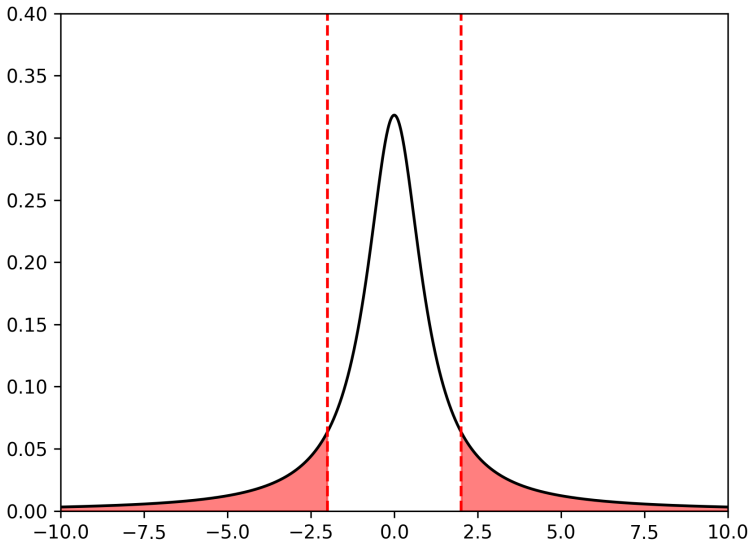

Figure 4.6: Cauchy density of Example 4.5.

We will next consider another example of estimating a probability where we show how to quantify the variance using the true value.

Example 4.5. Consider the following density

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

We would like to compute the probability of $X \sim p(x)$ being larger than 2, i.e., $\mathbb{P}(X > 2)$. We can compute this probability using MC estimation as

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

We can compute

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

We can also compute the real value of this integral as (see Example 2.3 for the CDF of this density)

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

Let us 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:

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

The variance of the estimator then

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

Could we do better? An idea is to use the fact that the density is symmetric around zero: This means $P(X > 2) = P(X < -2)$ (see Fig . 4.6). So we could compute:

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

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

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

As before, 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! This kind of variance improvements are crucial in safety critical applications.