Easy as π: The Importance Sampling Method

Oleg Mazonka, *2016*

Consider a program that simulates a particular model with randomness. Let the result of a simulation be an observable value A, where this value can be either continuous, discrete, or binary. For example, if the result ends with some event then we say that $A = 1$, otherwise $A = 0$. The probability of such an event is the expected value $\mathcal{A}$, which is found by taking the average of the results over N simulations. When N becomes large the result approaches the ”true” value:

$$\mathcal{A} = \langle A \rangle_p = \lim_{N\to\infty} \frac{1}{N} \sum A_p \quad (1)$$

The index p which appeared in this equation shows that the sampling process is performed over the ensemble p. What is ensemble p? In $\langle A \rangle_p$ it signifies the probability distribution $p(x)$, where x is a combined random variable which uniquely determines the outcome A. For $\langle A \rangle_p$ all possible combinations of x are selected with probability $p(x)$, and then the average is taken. Index p on the right hand side of eq. 1 specifies that the simulations are executed according to distribution $p(x)$, which is a particular algorithm of selecting each x.

To compute value $\mathcal{A}$ we introduce the bias of sampling. However one sampling process can be more efficient than another. Importance Sampling is a method to replace simulation with a different ensemble while expecting an improved efficiency, i.e. obtaining faster convergence of eq. 1. The first step is to note that sampling x with probability $p(x)$ while calculating $\langle A \rangle_p$ is the same as sampling x from the uniform distribution and multiplying value A by probability $p(x)$:

$$\langle A \rangle_p = \langle A\ p(x) \rangle_x$$

The whole idea of Importance Sampling is to introduce a different distribution $q(x)$, and to make the following transformations:

$$\langle A \rangle_p = \langle A\ p \rangle_x = \left\langle A \frac{p}{q} q \right\rangle_x = \langle A\, w\, q \rangle_x = \langle A\, w \rangle_q \quad (2)$$

The ratio p/q (a function depending on x) is defined as w. In order to have this function well defined, $q(x)$ must not be equal to zero wherever $p(x)$ is not zero. The last term in eq. 2 means that the average of $(A\,w)$ is now taken over the ensemble q. Eq. 2 demonstrates that the simulations with the new distribution $q(x)$ can be used to calculate the original value $\langle A \rangle_p$. Similarly to eq. 1 the average over results A converges to

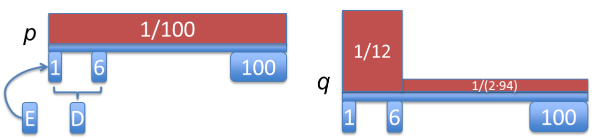
$$\mathcal{A} = \langle A\, w \rangle_q = \lim_{N\to\infty} \frac{1}{N} \sum A_q w \quad (3)$$

This time, however, the average of $(A\,w)$ is taken instead.

In the original model (eq. 1) we sample with distribution p, add values (A_p) obtained from different simulations, and divide the result by N. In the new model (eq. 3) we sample with distribution q, add values $(A_q w)$, and also divide by N.

This is the essence of the Importance Sampling method. It says that in both cases the result converges to the ”true” value $\mathcal{A}$, but it does not say how quickly it does. Depending on selection of new distribution q the new model may converge much faster. The improvement can be so great that it allows calculations which are impractical in the original model.

Example



Suppose we would like to calculate the probability of rolling a 1 on a die having only a random generator producing numbers from 1 to 100 (left figure). Let us call getting a 1 an event E and its probability P_E. Let us call getting any number from 1 to 6 an event D and its probability P_D. We would like to find the probability of E knowing that D occurred: $P(E|D)$. Bayes' theorem says

$$P(E|D) = P(D|E)P_E/P_D = P_E/P_D$$

where $P(D|E)$ is the probability of D knowing that E occurred that is obviously equal to 1. By simulating it 100 times we estimate:

$$P_E \to \frac{1}{100} \sum E_p \sim \frac{1}{100}; \quad P_D \to \frac{1}{100} \sum D_p \sim \frac{6}{100}$$

i.e. it is expected 1 outcome of a 1 out of 100 and about 6 of any number in the range $[1, 6]$. The estimate of the result is

$$P(E|D) = P_E/P_D \sim (1/100)/(6/100) = 1/6$$

Now let us change the random number generator so that there is a 50% chance of outputting numbers 1 to 6 (right figure). Hence, the probability of getting a number up to 6 is $(1/12)$ and a number greater than 6 is $1/(2\cdot(100-6))$. The function w inside the range $[1, 6]$ is $w = p/q = (1/100)/(1/12) = 12/100$. Now running only 12 simulations with this new random number generator we get

$$P_E \to \frac{1}{12} \sum E_q w = \frac{1}{12} \frac{12}{100} \sum E_q \sim \frac{1}{100}$$

$$P_D \to \frac{1}{12} \sum D_q w = \frac{1}{12} \frac{12}{100} \sum D_q \sim \frac{6}{100}$$

This time, however, E_q is expected about once out of 12, and D_q about 6 times out of 12. This means that changing the probability distribution from p to q we achieved a similar statistical estimate of the probability $P(E|D)$ by running only 12 simulations in the new model instead of 100 in the original model.