MA3227 Numerical Analysis II

Tutorial 9: Sampling theorems

1 Sampling theorems

Assume we want to sample the two random variables $X \sim \mathrm{PDF}(p)$ and $Y \sim \mathrm{PDF}(q)$ where

$$p(x) = \begin{cases} 6\,x\,(1-x) & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases} \qquad q(x) = \begin{cases} \frac{\pi}{2}\,\sin(\pi x) & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let us begin by checking whether we can do so using inverse transform sampling.

1. Verify that the cumulative distribution functions associated with $p(x)$ and $q(x)$ are given by, respectively,

$$C(x) = \begin{cases} 0 & \text{if } x < 0, \\ 3x^2 - 2x^3 & \text{if } 0 \le x \le 1, \\ 1 & \text{if } 1 < x, \end{cases} \qquad D(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{2}\big(1 - \cos(\pi x)\big) & \text{if } 0 \le x \le 1, \\ 1 & \text{if } 1 < x. \end{cases}$$

We observe that there is a simple formula for $D^{-1}(u)$ but not for $C^{-1}(u)$. We will therefore use inverse transform sampling for simulating $Y \sim \mathrm{PDF}(q)$, but we will need another technique for simulating $X \sim \mathrm{PDF}(p)$.

2. Complete the function `rand_sin()` such that it simulates $Y \sim \mathrm{PDF}(q)$ using inverse transform sampling. You can test your answer using the provided function `histogram()`.

Since inverse transform sampling cannot be used to simulate $X \sim \mathrm{PDF}(p)$, we next try to do so using rejection sampling. This technique requires a proposal distribution which is close to $\mathrm{PDF}(p)$ and easy to sample, and `plot_pdfs()` combined with Task 2 shows that $\mathrm{PDF}(q)A$ is such a distribution. Therefore:

3. Complete the function `rand_quad()` such that it simulates $X \sim \mathrm{PDF}(p)$ using rejection sampling with $\mathrm{PDF}(q)$ as proposal distribution. You can test your answer using the provided function `histogram()`.

   Hint. You will need to compute

$$M = \sup_{x \in [0,1]} \frac{p(x)}{q(x)}.$$

   `plot_pdf_ratio()` reveals that

$$M = \lim_{x \to 0} \frac{p(x)}{q(x)},$$

   and this limit can be computed using L'Hôpital's rule.

Given `rand_sin()` and `rand_quad()`, we can estimate $\mathbb{E}[X]$ using either direct sampling or importance sampling with $\mathrm{PDF}(q)$ as sampling distribution. Let us first convince ourselves once more that the two approaches indeed yield the same result.

4. Complete the function `monte_carlo()` such that it prints the Monte Carlo estimates for $\mathbb{E}[X]$ using both direct and importance sampling. Verify that both approaches yield the same result up to sampling error.

Let us next try to assess which of the two methods is more efficient. Since Monte Carlo methods depend on an effort parameter (the number of samples), the correct way to address this question is to ask which method delivers better accuracy in less time, and we know that the accuracy of the Monte Carlo estimate $\tilde{\mathbb{E}}_N[X]$ depends on $\mathrm{Var}[X]$. The provided function `comparison()` therefore estimates the variances for both direct and importance sampling, and it also measures the runtimes of obtaining a single sample according to both approaches.

5. Determine which of the two sampling techniques is more efficient based on the output of `comparison()`. You may want to add code at the end of this function to facilitate your analysis.

On my machine, importance sampling turns out to be more efficient than direct sampling, and the reason for this is that the runtime per sample is about 1.5 times larger for direct sampling compared to importance sampling:

```
Runtime per sample:
        Direct sampling: 61 nanoseconds
    Importance sampling: 40 nanoseconds
```

Closer inspection of the two sampling algorithms reveals that direct sampling probably spends most of its time evaluating

$$\texttt{Y = rand\_sin()} \quad \text{and} \quad \texttt{rand() <= quad\_pdf(Y)/(M*sin\_pdf(Y))}, \tag{1}$$

while importance sampling spends most of its time evaluating

$$\texttt{Y = rand\_sin()} \quad \text{and} \quad \texttt{Y * quad\_pdf(Y) / sin\_pdf(Y)}. \tag{2}$$

6. Estimate the runtime-per-sample ratio between direct and importance sampling, assuming both sets of commands (1) and (2) take the same amount of time and hence the difference in runtime is only due to the fact that direct + rejection sampling requires several executions of (1) while importance sampling requires only a single execution of (2).