Statistics for Data Science Exercise 5
Wintersemester 2023/24
Please complete these problems before the exercise session on Tuesday 21 November, 2023, 8:30. Please be prepared to present your solutions to any problems that you completed successfully.

1. Let $X \sim \mathcal{N}(\mu, C)$, where $\mu \in \mathbb{R}^d$ and $C \in \mathbb{R}^{d\times d}$ is a symmetric, positive definite matrix for $d \in \mathbb{N}$. Show that

$$\mathbb{E}[\|X-\mu\|^2] = \operatorname{tr}(C),$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$ and $\operatorname{tr}(C)$ is the matrix trace, defined as the sum over the diagonal elements of matrix C.

2. Recall that the Bernoulli random variable $X \sim \operatorname{Ber}(p)$ with parameter $p \in (0,1)$ has only two possible outcomes: 1 ("success") with probability p and 0 ("failure") with probability $1-p$. The probability mass function of X is

$$p_X(x) = \begin{cases} 1-p & \text{if } x = 0, \\ p & \text{if } x = 1. \end{cases}$$

Let $X_1, \dots, X_n \sim \operatorname{Ber}(p)$ be i.i.d. random variables and let H be a random variable denoting the number of successful trials, i.e.,

$$H = \sum_{k=1}^{n} X_k.$$

Let $n = 10^3$ and $p = \frac{1}{3}$.

   (a) Visualize the distribution of H as a histogram. To achieve this, you can generate, e.g., 10^3 different realizations of H.[†]

   (b) Use the central limit theorem to approximate H using a Gaussian distribution. Plot this against the (appropriately normalized) histogram you derived in part (a) to validate your answer.

   (c) Use the Gaussian distribution you derived in part (b) to find an interval $\mathcal{I}$ such that

$$\mathbb{P}(H \in \mathcal{I}) \approx 0.95.$$

3. Let $X \sim \mathcal{U}([0,1]^d)$ be a uniform random variable in the hypercube $[0,1]^d$ with $d \geq 1$. Recall that this distribution has the probability density function

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0,1]^d, \\ 0 & \text{if } x \in \mathbb{R}^d \setminus [0,1]^d. \end{cases}$$

[†]To draw a sample from $\operatorname{Ber}(p)$, you may use, e.g., `numpy.random.binomial(1,p=p,size=n)` in Python.

Let $X_1, \dots, X_n \sim \mathcal{U}([0,1]^d)$ be i.i.d. random variables and let $g\colon [0,1]^d \to \mathbb{R}$ be a function. Then it is a consequence of the (strong) law of large numbers that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g(X_i) = \int_{[0,1]^d} g(x)\,\mathrm{d}x \quad \text{almost surely.}$$

Let $g(x) = \cos\left(\sum_{i=1}^{d} x_i\right)$ for $x = (x_1, \dots, x_d) \in [0,1]^d$. Estimate the value of the high-dimensional integral

$$I_d(g) = \int_{[0,1]^d} g(x)\,\mathrm{d}x$$

by using sample averages $Q_{d,n}(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$, where $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1]^d)$ is a random sample. This method of numerical integration is known as the *Monte Carlo method*.

In this case, the exact value of the integral is $I_d(g) = 2^d \cos(\frac{d}{2})(\sin\frac{1}{2})^d$ (you do not need to prove this). Compute the Monte Carlo numerical integration error $|I_d(g) - Q_{d,n}(g)|$ for sample sizes $n = 2^k$, $k = 0, 1, 2, \dots, 20$.† Try out several values for the dimension d, for example, $d \in \{10, 100, 1\,000\}$. What convergence rate do you observe for the numerical integration error as a function of n? Does increasing the dimension d affect the convergence rate?

4. Let $k \in \mathbb{N}$ and consider the function $f_k\colon \mathbb{R} \to (0,\infty)$,

$$f_k(x) = \begin{cases} 0 & \text{if } x < 1, \\ \frac{k}{x^{k+1}} & \text{if } x \geq 1. \end{cases} \tag{1}$$

   (a) Show that f_k is a probability density function for each $k \in \mathbb{N}$.

   (b) Let X be a random variable with probability density function (1) for $k \in \mathbb{N}$. Show that the moments $\mathbb{E}[X^\ell]$ exist for $0 \leq \ell \leq k-1$ and do not exist for $\ell \geq k$.

   (c) Let $X_1, \dots, X_n$ be i.i.d. random variables with the probability density function (1). What can you say about the convergence of the sample average $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ to $\mathbb{E}[X]$ for different values of k?

†If $n = 2^{20}$ takes too long, you can also compute the errors up to, say, $n = 2^{17}$.