Solutions to Graded Homework 2

Exercise 1. a) use $B = A \cup (B \setminus A)$, where $A$ and $B \setminus A$ are disjoint, as well as $\Omega = A \cup A^c$ and $P(\Omega) = 1$.

b) use $A \cup B = A \cup (B \setminus (A \cap B))$ where $A$ and $B \setminus (A \cap B)$ are disjoint, as well as a).

c) use $\cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} B_n$, where $B_n = A_n \setminus (A_1 \cup \ldots \cup A_{n-1})$; the $B_n$ are disjoint, so by axiom (iii’)

$$P(\cup_{n=1}^{\infty} A_n) = P(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} P(B_n) \leq \sum_{n=1}^{\infty} P(A_n).$$

d) $P(\cup_{n \geq 1} A_n) = P(\cup_{n \geq 1} (A_n \cap (\cup_{i=1}^{n-1} A_i)^c)) = P(\cup_{n \geq 1} (A_n \cap A_i^c)) = \sum_{n=1}^{\infty} P(A_n \cap A_i^c)

$$lim_{n \to \infty} \sum_{i=1}^{n} P(A_i \cap A_i^c) = lim_{n \to \infty} P(\cup_{i=1}^{n} (A_i \cap A_i^c)) = lim_{n \to \infty} P(\cup_{i=1}^{n} A_i) = lim_{n \to \infty} P(A_n).$$
e) $P(\cap_{n \geq 1} A_n) = 1 - P((\cap_{n \geq 1} A_n)^c) = 1 - P(\cup_{n \geq 1} A_n^c) = 1 - lim_{n \to \infty} P(A_n^c) = lim_{n \to \infty} P(A_n).

Exercise 2. a) If $a > 0$, then $F_{Y_1}(t) = P\{aX \leq t\} = P\{X \leq t/a\} = F_X(t/a)$, which leads to

$$p_{Y_1}(t) = F'_{Y_1}(t) = \frac{1}{a} p_X(t/a)$$

when $X$ is a continuous random variable. If $a < 0$, then $F_{Y_1}(t) = P\{X \geq a\} = 1 - P\{X < a\}$. Notice that it is not necessarily the case that

$$P\{X < t/a\} = F_X(t/a), \quad \text{but it always holds that} \quad P\{X < t/a\} = \lim_{\varepsilon \downarrow 0} F_X((t - \varepsilon)/a)$$

When $X$ is a continuous random variable, then the left-hand side equality holds and consequently, $p_{Y_1}(t) = F'_{Y_1}(t) = \frac{1}{|a|} p_X(t/a)$. From this, we deduce that for all $a \neq 0$, $p_{Y_1}(t) = \frac{1}{|a|} p_X(t/a)$.

b) $F_{Y_2}(t) = P\{X + c \leq t\} = P\{X \leq t - c\} = F_X(t - c)$. When $X$ is a continuous random variable, we therefore obtain that $p_{Y_2}(t) = p_X(t - c)$.

c) $F_{Y_3}(t) = P\{X^2 \leq t\} = 0$ if $t < 0$ and $F_{Y_3}(t) = P\{|X| \leq \sqrt{t}\} = F_X(\sqrt{t}) - \lim_{\varepsilon \downarrow 0} F_X(-\sqrt{t} - \varepsilon)$ if $t \geq 0$. If $X$ is a continuous random variable, then

$$p_{Y_3}(t) = \begin{cases} 0 & \text{if } t < 0 \\
\frac{1}{2\sqrt{t}} (p_X(\sqrt{t}) + p_X(-\sqrt{t})) & \text{if } t > 0 
\end{cases}$$

d) $F_{Y_4}(t) = P\{e^X \leq t\} = 0$ if $t \leq 0$ and $F_{Y_4}(t) = P\{X \leq \log t\} = F_X(\log t)$ of $t > 0$. If $X$ is a continuous random variable, then

$$p_{Y_4}(t) = \begin{cases} 0 & \text{if } t \leq 0 \\
\frac{1}{2} p_X(\log t) & \text{if } t > 0 
\end{cases}$$

Exercise 3. The following are guaranteed to be cdfs: $F_1, F_3, F_5, F_6, F_7$ and $F_8$ ($F_2$ and $F_4$ are clearly not cdfs in general, $F_0$ is not right-continuous in $t = 0$ and $F_{10}(t) = \frac{1}{2}$ for all $t \in \mathbb{R}$).
Exercise 4. a) We have, by independence of the $X_j$'s:

$$F_Y(t) = \mathbb{P}(\{Y \leq t\}) = \mathbb{P}(\{\max(X_1, \ldots, X_n) \leq t\}) = \mathbb{P}(\{X_1 \leq t, \ldots, X_n \leq t\}) = \prod_{j=1}^{n} \mathbb{P}(\{X_j \leq t\}) = (F_X(t))^n$$

b) In a similar way, we obtain:

$$F_Z(t) = \mathbb{P}(\{Z \leq t\}) = 1 - \mathbb{P}(\{Z > t\}) = 1 - \mathbb{P}(\{\min(X_1, \ldots, X_n) > t\})$$

$$= 1 - \mathbb{P}(\{X_1 > t, \ldots, X_n > t\}) = 1 - \prod_{j=1}^{n} \mathbb{P}(\{X_j > t\}) = 1 - (1 - F_X(t))^n$$

c) $V = \max(U_1, U_2)$

d) $W = \begin{cases} U_1 & \text{with probability } \alpha \\ U_2 & \text{with probability } 1 - \alpha \end{cases}$

Coding Exercise 5. a) cf. code. In order to sample (approximately) from $F$, we pick uniformly at random $M$ numbers $a_1, \ldots, a_M \in \{0, 2\}$ and compute a sample $X$ as follows:

$$X = \sum_{k=1}^{M} \frac{a_k}{3^k}$$

For the experiments below, we have chosen $M = 20$. The code is written in Python 3, and uses the libraries numpy and matplotlib.

b-c) Here are the graphs we get with $n = 100$:

d) The test is as follows: let $F_n$ be the empirical cdf of the $n$ samples. The null hypothesis $H_0$ is that the samples are distributed according to $G$, the uniform distribution on $[0, 1]$. To test whether these are distributed according to $F$ (=alternative hypothesis $H_1$), the uniform distribution on the Cantor set $C$, compute the quantity

$$D_n = F_n(2/3) - F_n(1/3)$$
and observe that \( D_n = 0 \) if and only if no sample falls in the interval \( ]1/3, 2/3] \), which provides an indication, at least for large \( n \), that the samples are distributed according to \( F \).

Notice that the test proposed above is rudimentary. One could imagine more elaborate tests, checking more intervals (if not all? . . . up to the limit of the computer precision), that would allow to reduce the probability of a false positive below.

e) If it turns out that \( D_n > 0 \), then we can for sure reject the alternative hypothesis \( H_1 \), so the probability of a false negative is

\[
P(\text{the samples are distributed according to } F \mid D_n > 0) = 0
\]

But how confident should we be about the hypothesis \( H_1 \) if it turns out that \( D_n = 0 \)? Using Bayes’ formula and the prior assumption that the samples are distributed according to \( F \) with probability \( \alpha \), we obtain (abbreviating “the samples are distributed according to \( F \)” by “\( F \), simply):

\[
P(\text{the samples are distributed according to } F \mid D_n = 0) = \frac{P(F \cap \{D_n = 0\})}{P(\{D_n = 0\})} = \frac{1 \cdot \alpha}{1 \cdot \alpha + (2/3)^n \cdot (1 - \alpha)}
\]

since \( P(\{D_n = 0\} \mid G) = P(\{X_k \notin [1/3, 2/3], \forall 1 \leq k \leq n\}) = (2/3)^n \). So the probability of a false positive is given by:

\[
P(\text{the samples are distributed according to } G \mid D_n = 0) = 1 - \frac{\alpha}{\alpha + (2/3)^n \cdot (1 - \alpha)} = \frac{(2/3)^n}{\alpha + (2/3)^n}
\]

Notice that this probability can still be significant if \( \alpha \) is small (in particular, it can be close to 1/2 if \( \alpha \approx (2/3)^n \)).