Statistics for Data Science
Wintersemester 2024/25
Solutions to the 2$^{\text{nd}}$ exercise

1. The quantile function of f was computed in exercise 3 of last week:

$$F^{-1}(q) = \sqrt{\log \frac{q+1}{1-q}}, \quad q \in (0,1).$$

Please see the script `w2e1.py` on the course webpage.

2. (a) Please see the script `w2e2.py` on the course webpage.
(b) The PDF of $X \sim \mathcal{N}(0,1)$ is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}, \quad x \in \mathbb{R}.$$

and the task is to find the PDF f_Y for random variable $Y = g(X)$, $g(x) = \operatorname{arc}\tan x$, $x \in \mathbb{R}$. The inverse of g is given by

$$g^{-1}(y) = \tan y, \quad y \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2}),$$

and $(g^{-1})'(y) = \frac{1}{\cos^2 y}$ for $y \in (-\frac{\pi}{2}, \frac{\pi}{2})$. We can use the change of variables formula:

$$f_Y(y) = f_X(g^{-1}(y))|(g^{-1})'(y)| = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\tan^2 y} \frac{1}{\cos^2 y}, \quad y \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2}).$$

3. (a) Let $Z = \max(X, Y^2)$, where $X, Y \sim \mathcal{U}(0,1)$ are assumed to be independent. Then X and Y^2 are independent.[†] Thus

$$\begin{aligned} F_Z(t) = \mathbb{P}(Z \le t) &= \mathbb{P}(\max(X, Y^2) \le t) \\ &= \mathbb{P}(X \le t)\mathbb{P}(Y^2 \le t) \\ &= \mathbb{P}(X \le t)\mathbb{P}(-\sqrt{t} \le Y \le \sqrt{t}) \\ &= \left(\int_{-\infty}^{t} \mathbf{1}_{[0,1]}(x) \, \mathrm{d}x \right) \left(\int_{-\sqrt{t}}^{\sqrt{t}} \mathbf{1}_{[0,1]}(y) \, \mathrm{d}y \right), \end{aligned}$$

where $\int_{-\infty}^{t} \mathbf{1}_{[0,1]}(x) \, \mathrm{d}x = \begin{cases} 0 & \text{if } t < 0 \\ t & \text{if } 0 \le t \le 1 \\ 1 & \text{if } t > 1 \end{cases}$ and $\int_{-\sqrt{t}}^{\sqrt{t}} \mathbf{1}_{[0,1]}(x) \, \mathrm{d}x = \begin{cases} 0 & \text{if } t < 0 \\ \sqrt{t} & \text{if } 0 \le t \le 1 \\ 1 & \text{if } t > 1. \end{cases}$

Therefore

$$F_Z(t) = \begin{cases} 0 & \text{if } t < 0 \\ t^{3/2} & \text{if } 0 \le t \le 1 \\ 1 & \text{if } t > 1 \end{cases} \quad \Rightarrow \quad f_Z(t) = \frac{3}{2}\sqrt{t}\, \mathbf{1}_{[0,1]}(t).$$

We also infer from the above that $F_Z^{-1}(q) = q^{2/3}$, $q \in (0,1)$.
(b) Please see the script `w2e3.py` on the course webpage.

[†] *Claim:* Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be measurable (e.g., continuous) functions. If X and Y are independent real-valued RVs, then $f(X)$ and $g(Y)$ are independent real-valued RVs. *Proof:* $\mathbb{P}(f(X) \in A, g(Y) \in B) = \mathbb{P}(X \in f^{-1}(A), Y \in g^{-1}(B)) \overset{X \perp Y}{=} \mathbb{P}(X \in f^{-1}(A))\mathbb{P}(Y \in g^{-1}(B)) = \mathbb{P}(f(X) \in A)\mathbb{P}(g(Y) \in B)$ for all $A, B \subset \mathbb{R}$. Here, $f^{-1}(A) = \{x \in \mathbb{R} \mid f(x) \in A\}$ and $g^{-1}(B) = \{x \in B \mid g(x) \in B\}$ are preimages.

4. Let $(Y_1, Y_2) = g(X_1, X_2)$, where $g(x_1, x_2) = (\log x_1, \log x_2)$ for $x_1, x_2 > 0$. The mapping $g\colon \mathbb{R}_+^2 \to \mathbb{R}^2$ is a continuously differentiable bijection, and it has a continuously differentiable inverse $(x_1, x_2) \mapsto (\mathrm{e}^{x_1}, \mathrm{e}^{x_2})$ for $x_1, x_2 \in \mathbb{R}$. By the change of variables formula,

$$f_{X_1,X_2}(x_1, x_2) = f_{Y_1,Y_2}(g(x_1, x_2))|\det Dg(x_1, x_2)|,$$

where

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2\pi\sqrt{\det C}} \exp\left(-\frac{1}{2}[y_1 - 1, y_2 - 1]C^{-1}\begin{bmatrix} y_1 - 1 \\ y_2 - 1 \end{bmatrix}\right), \quad C = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

Note that the matrix C is positive definite since

$$[x_1, x_2]C\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 - 2x_1x_2 + 2x_2^2 = \underbrace{\frac{1}{2}(2x_1 - x_2)^2 + \frac{3}{2}x_2^2}_{\substack{\text{always nonnegative; and} \\ \text{zero iff } 2x_1 - x_2 = 0 \text{ and } x_2 = 0 \\ \text{iff } x_1 = x_2 = 0}} > 0$$

for all $(x_1, x_2) \in \mathbb{R}^2 \setminus \{(0, 0)\}$. (Alternatively, one can establish that C is positive definite by noticing that all of its eigenvalues, $\lambda_1 = 3$ and $\lambda_2 = 1$, are positive.) Moreover, $C^{-1} = \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix}$.

The Jacobian matrix of g is

$$Dg(x_1, x_2) = \begin{bmatrix} \frac{\partial}{\partial x_1}\log x_1 & \frac{\partial}{\partial x_2}\log x_1 \\ \frac{\partial}{\partial x_1}\log x_2 & \frac{\partial}{\partial x_2}\log x_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{x_1} & 0 \\ 0 & \frac{1}{x_2} \end{bmatrix} \quad \Rightarrow \quad \det Dg(x_1, x_2) = \frac{1}{x_1x_2}.$$

Therefore

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sqrt{\det C}}\frac{1}{x_1x_2} \exp\left(-\frac{1}{2}[\log x_1 - 1, \log x_2 - 1]C^{-1}\begin{bmatrix} \log x_1 - 1 \\ \log x_2 - 1 \end{bmatrix}\right) \tag{1}$$

for $x_1, x_2 > 0$. This is known as a multivariate *lognormal distribution*, defined by the fact that the (componentwise) logarithm of (X_1, X_2) is a multivariate Gaussian RV. In this case, we would denote $(X_1, X_2) \sim \mathrm{Lognormal}(\mu, C)$ with parameters $\mu = [1, 1]^{\mathrm{T}}$ and $C = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.

Remark. The PDF can be further simplified (by expanding the expression inside the exponent and evaluating the determinant of C) to

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\sqrt{3}\pi}\mathrm{e}^{-1-\frac{1}{3}\log^2 x_1 - \frac{1}{3}\log x_1 \log x_2 - \frac{1}{3}\log^2 x_2}.$$

However, the formula (1) is arguably more informative since it is easier to compare it to the PDF of a lognormal distribution by reading off the parameters $\mu = [1, 1]^{\mathrm{T}}$ and $C = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.