Statistics for Data Science
Wintersemester 2024/25
Solutions to the 3rd exercise

1. Let $X \sim \mathcal{N}(0,1)$ and $t \in \mathbb{R}$. Then

$$\mathbb{P}(X^2 \le t) = \mathbb{P}(-\sqrt{t} \le X \le \sqrt{t}) = \begin{cases} 0 & \text{if } t < 0, \\ \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{t}}^{\sqrt{t}} \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x & \text{if } t \ge 0. \end{cases}$$

If $t < 0$, then $f(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(X^2 \le t) = 0$. If $t \ge 0$, define $\Phi(t) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x$. Since $\Phi'(t) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}t^2}$, we obtain

$$f(t) = \frac{\mathrm{d}}{\mathrm{d}t}(\Phi(\sqrt{t}) - \Phi(-\sqrt{t})) = \Phi'(\sqrt{t})\frac{1}{2\sqrt{t}} + \Phi'(-\sqrt{t})\frac{1}{2\sqrt{t}} = \frac{1}{\sqrt{2\pi t}}\mathrm{e}^{-t/2},$$

as desired.

2. Let $X \sim \mathcal{N}(0,1)$. Then by the law of the unconscious statistician,

$$\mathbb{E}[|X|] = \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} x\mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x = -\sqrt{\frac{2}{\pi}}\left[\mathrm{e}^{-\frac{1}{2}x^2}\right]_{x=0}^{x=\infty} = \sqrt{\frac{2}{\pi}}.$$

Moreover,

$$\mathrm{Var}(|X|) = \mathbb{E}[|X|^2] - \mathbb{E}[|X|]^2 = \mathbb{E}[X^2] - \mathbb{E}[|X|]^2 = 1 - \frac{2}{\pi} = \frac{\pi - 2}{\pi}.$$

Discussion. There is an interesting phenomenon which happens for d-variate Gaussian distributions as the dimension d increases. See the Appendix of this document for further information.

3. Since $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^3] = 0$, we find that

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathbb{E}[(X - \mathbb{E}[X])(X^2 - \mathbb{E}[X^2])] \\ &= \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] - \mathbb{E}[X]\mathbb{E}[X^2] + \mathbb{E}[X]\mathbb{E}[X^2] = 0. \end{aligned}$$

The random variables X and X^2 are clearly dependent since the realization of X completely determines X^2. One can also easily check that

$$\mathbb{P}(X \le t, X^2 \le t) \ne \mathbb{P}(X \le t)\mathbb{P}(X^2 \le t) \quad \text{for any } t > 0.$$

4. The covariance matrix is defined by

$$C = \mathbb{E}[(X - \mu)(X - \mu)^{\mathrm{T}}],$$

which implies that $C_{i,j} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$, where we denote $X = [X_1, \ldots, X_n]^{\mathrm{T}}$ and $\mu = [\mu_1, \ldots, \mu_n]^{\mathrm{T}}$. Then

$$\begin{aligned} \mathbb{E}[\|X - \mu\|^2] &= \mathbb{E}[(X_1 - \mu_1)^2 + (X_2 - \mu_2)^2 + \ldots + (X_n - \mu_n)^2] \\ &= \mathbb{E}[(X_1 - \mu_1)^2] + \mathbb{E}[(X_2 - \mu_2)^2] + \ldots + \mathbb{E}[(X_n - \mu_n)^2] \\ &= \mathbb{E}[(X_1 - \mu_1)(X_1 - \mu_1)] + \mathbb{E}[(X_2 - \mu_2)(X_2 - \mu_2)] + \ldots \\ &\quad + \mathbb{E}[(X_n - \mu_n)(X_n - \mu_n)] \\ &= \mathrm{tr}(C). \end{aligned}$$

Appendix: Gaussian annulus theorem

We will consider the two-dimensional generalization of exercise 2, as well as the generalization to d dimensions. Before doing so, we note the following helpful integral identity:

$$\begin{aligned}
&\int_0^\infty r^k \mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}r \qquad\qquad \text{(change of variables: } t = \tfrac{1}{2}r^2\text{)}\\
&= 2^{k/2-1/2}\int_0^\infty t^{k/2-1/2}\mathrm{e}^{-t}\,\mathrm{d}t\\
&= 2^{k/2-1/2}\Gamma\big(\tfrac{k+1}{2}\big), \qquad\qquad (1)
\end{aligned}$$

where we used the *gamma function* $\Gamma(x) = \int_0^\infty t^{x-1}\mathrm{e}^{-t}\,\mathrm{d}t$, $x > 0$. The gamma function is a generalization of the factorial. In fact, it satisfies $\Gamma(x+1) = x\Gamma(x)$ for $x > 0$ and $\Gamma(k+1) = k!$ for $k \in \mathbb{N}_0$.

Bivariate case. Let $X \sim \mathcal{N}(0, I_2)$ and let $\|\cdot\|$ denote the Euclidean norm in $\mathbb{R}^2$. By using the change of variables to polar coordinates, i.e., $(x, y) = (r\cos\theta, r\sin\theta)$, with $\mathrm{d}x\,\mathrm{d}y = r\,\mathrm{d}\theta\,\mathrm{d}r$, we obtain

$$\begin{aligned}
\mathbb{E}[\|X\|] = \int_{\mathbb{R}^2}\|x\|\frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}\|x\|^2}\,\mathrm{d}x = \frac{1}{2\pi}\int_0^\infty\int_0^{2\pi} r^2\mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}\theta\,\mathrm{d}r = \int_0^\infty r^2\mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}\theta\,\mathrm{d}r\\
= \frac{\sqrt{2\pi}}{2} = \sqrt{\frac{\pi}{2}},
\end{aligned}$$

where the final integral follows from (1) or the computation carried out on pg. 114 of the lecture notes.

Similarly,

$$\mathbb{E}[\|X\|^2] = \int_{\mathbb{R}^2}\|x\|^2\frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}\|x\|^2}\,\mathrm{d}x = \frac{1}{2\pi}\int_0^\infty\int_0^{2\pi} r^3\mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}\theta\,\mathrm{d}r = \int_0^\infty r^3\mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}r \overset{(1)}{=} 2,$$

so

$$\mathrm{Var}(\|X\|) = \mathbb{E}[\|X\|^2] - \mathbb{E}[\|X\|]^2 = 2 - \frac{\pi}{2} = \frac{4-\pi}{2}.$$

***d*-variate case.** Let $X \sim \mathcal{N}(0, I_d)$ and let $\|\cdot\|$ denote the Euclidean norm in $\mathbb{R}^d$ for $d \in \mathbb{N}$. Then using the change of variables to d-dimensional spherical coordinates (see also: `https://en.wikipedia.org/wiki/Coarea_formula`), we obtain

$$\begin{aligned}
\mathbb{E}[\|X\|] &= \int_{\mathbb{R}^d} \|x\| \frac{1}{(2\pi)^{d/2}} \mathrm{e}^{-\frac{1}{2}\|x\|^2}\,\mathrm{d}x \\
&= \frac{1}{(2\pi)^{d/2}} \int_0^\infty \int_{\{\theta\in\mathbb{R}^d : \|\theta\|=1\}} r^d \mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}S(\theta)\,\mathrm{d}r \\
&\qquad\qquad \text{(surface area of the unit sphere in } \mathbb{R}^d\text{: } \tfrac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}) \\
&= \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})(2\pi)^{d/2}} \int_0^\infty r^d \mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}r \qquad \Big(\int_0^\infty r^d \mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}r \overset{(1)}{=} 2^{d/2-1/2}\Gamma(\tfrac{d+1}{2})\Big) \\
&= \frac{1}{\Gamma(\frac{d}{2})2^{d/2-1}} \cdot 2^{d/2-1/2}\Gamma\big(\tfrac{d+1}{2}\big) \\
&= \sqrt{2}\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} \qquad \Big(\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} = \frac{\sqrt{d}}{\sqrt{2}} - \frac{1}{4\sqrt{2d}} + \mathcal{O}(d^{-3/2}) \text{ as } d\to\infty\Big) \\
&= \sqrt{d} - \frac{1}{4}d^{-1/2} + \mathcal{O}(d^{-3/2}) \quad \text{as } d\to\infty,
\end{aligned}$$

where the last step follows from the asymptotic behavior of $\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})}$ as $d\to\infty$. Here, we denote $f(x) = \mathcal{O}(g(x))$ as $x\to\infty$ if there exists a constant $C>0$ such that $f(x)\le Cg(x)$ for all sufficiently large $x>0$.

Similarly,

$$\begin{aligned}
\mathbb{E}[\|X\|^2] &= \int_{\mathbb{R}^d} \|x\|^2 \frac{1}{(2\pi)^{d/2}} \mathrm{e}^{-\frac{1}{2}\|x\|^2}\,\mathrm{d}x \\
&= \frac{1}{(2\pi)^{d/2}} \int_0^\infty \int_{\{\theta\in\mathbb{R}^d : \|\theta\|=1\}} r^{d+1} \mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}S(\theta)\,\mathrm{d}r \\
&\qquad\qquad \text{(surface area of the unit sphere in } \mathbb{R}^d\text{: } \tfrac{2\pi^{d/2}}{\Gamma(\frac{d}{2})}) \\
&= \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})(2\pi)^{d/2}} \int_0^\infty r^{d+1} \mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}r \qquad \Big(\int_0^\infty r^{d+1} \mathrm{e}^{-\frac{1}{2}r^2}\,\mathrm{d}r \overset{(1)}{=} 2^{d/2}\Gamma(\tfrac{d}{2}+1)\Big) \\
&= 2\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2})} = d,
\end{aligned}$$

where we made use of the property $\Gamma(x+1) = x\Gamma(x)$ for $x>0$.

Therefore

$$\mathrm{Var}(\|X\|) = \mathbb{E}[\|X\|^2] - \mathbb{E}[\|X\|]^2 = \frac{1}{2} + \mathcal{O}(d^{-1/2}).$$

In conclusion:

$$\begin{aligned}
\mathbb{E}[\|X\|] &= \sqrt{d} + \mathcal{O}(d^{-1/2}) \quad \text{as } d\to\infty, \\
\mathrm{Var}(\|X\|) &= \tfrac{1}{2} + \mathcal{O}(d^{-1/2}) \quad \text{as } d\to\infty.
\end{aligned}$$

This is the so-called *Gaussian annulus theorem:* "Nearly all the mass of the PDF of a d-dimensional spherical Gaussian distribution with unit variance is concentrated in a thin annulus of constant width at radius $\sqrt{d}$ from the origin as the dimension d increases." That is, even though $X\sim\mathcal{N}(0,I_d)$ is centered at the origin, most of the realizations of this random variable will have norm close to $\sqrt{d}$ on average.