ε.δε.μ^2

ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Pattern Recognition

1st Set of Analytical Problems

Nikos Stamatis
nikolaosstamatis@mail.ntua.gr
MSc Student
SN: 03400115

December 10, 2022

Exercise 1.1

The density of an Erlang distribution is

$$p(x\,|\,\vartheta) = \begin{cases} \vartheta^2 x e^{-\vartheta x}, & x > 0 \\ 0, & x \le 0, \end{cases}$$

where $\vartheta \in (0, +\infty)$ is a parameter. The support of an Erlang distribution is the set $I = [0, +\infty)$.[1]

Let $\tilde{x} = (x_1, \ldots, x_n)$ be a random sample from the Erlang distribution. Its likelihood is then

$$L(\tilde{x}\,|\,\vartheta) = \prod_{i=1}^{n} p(x_i\,|\,\vartheta) = \vartheta^{2n} e^{-\vartheta \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} x_i,$$

and if all x_i are strictly positive, its log-likelihood is

$$\ell(\vartheta) = \ln L(\tilde{x}\,|\,\vartheta) = 2n \ln \vartheta + \sum_{i=1}^{n} \ln x_i - \vartheta \sum_{i=1}^{n} x_i,$$

with

$$\ell'(\vartheta) = \frac{2n}{\vartheta} - \sum_{i=1}^{n} x_i \quad \text{and}$$
$$\ell''(\vartheta) = -\frac{2n}{\vartheta^2} < 0, \quad \text{for all } \vartheta > 0.$$

The derivative of the log-likelihood is zero for $\vartheta^* = \frac{2}{\bar{x}}$, and since its second derivative is negative, this ϑ^* corresponds to a local maximum. Since $\lim_{\vartheta \to \infty} \ell(\vartheta) = -\infty$, the maximum is also global.

If some sample point x_i is equal to zero, then $L(\tilde{x}|\vartheta) = 0$ for all ϑ, and the maximum is attained for any $\vartheta \in (0, +\infty)$. To sum up, the MLE of the Erlang is

$$\text{MLE} = \begin{cases} \frac{2}{\bar{x}}, & \text{if } x_i > 0 \text{ for all } i, \\ (0, +\infty), & \text{if } x_i = 0 \text{ for some } i. \end{cases} \tag{1}$$

Although the last case may seem redundant, it has to be included for the sake of completeness, since 0 belongs to the support of our distribution. The argument that an observation $x_i = 0$ has zero probability to occur is completely meaningless here; any particular point has zero probability to occur when drawn from a continuous distribution. Any point in the support must be considered as a possible sample.

The fact that in the last case the MLE is not unique is also not particularly strange. All three possible cases can occur for the MLE of a distribution: 1) It may exist and be unique, 2) there may exist many different MLE's, 3) an MLE may not exist at all.

[1] The support of a measure P is the largest closed set F with the property that $P(U_x) > 0$ for any open neighborhood U_x of any point $x \in F$ [AB06, p. 441]. Here, $p(x|\vartheta) > 0$ for every $x > 0$ and the closure of $(0, +\infty)$ is $[0, +\infty)$.

Exercise 1.2

The risk associated with a partition R_1, R_2 and prior probabilities $P(\omega_1), P(\omega_2)$ is

$$R(P(\omega_1), R_2) = P(\omega_1) \int_{R_2} p(x\,|\,\omega_1)\,dx + (1 - P(\omega_1)) \int_{R_2^c} p(x\,|\,\omega_2)\,dx \tag{2}$$

For the minimax criterion, we need to solve the following problem:

$$\min_{R_2} \max_{\lambda \in [0,1]} R(\lambda, R_2), \tag{3}$$

where the above minimum is taken over all the subsets R_2 of $\mathbb{R}$. Since $R(\lambda, R_2) = \lambda \int_{R_2} p(x\,|\,\omega_1)dx + (1 - \lambda) \int_{R_2^c} p(x\,|\,\omega_2)dx$ is just a convex combination of the quantities $\int_{R_2} p(x\,|\,\omega_1)dx$ and $\int_{R_2^c} p(x\,|\,\omega_2)dx$, its maximum will be equal to

$$\max_{\lambda \in [0,1]} R(\lambda, R_2) = \max\left\{ \int_{R_2} p(x\,|\,\omega_1), \int_{R_2^c} p(x\,|\,\omega_2) \right\} \tag{4}$$

and problem (3) can be restated as

$$\min_{R_2} \max\left\{ \int_{R_2} p(x\,|\,\omega_1), \int_{R_2^c} p(x\,|\,\omega_2) \right\}. \tag{5}$$

The solution $\tilde{R}_2$ of the previous minimization problem has to satisfy the following relation:

$$\int_{\tilde{R}_2} p(x\,|\,\omega_1)\,dx = \int_{\tilde{R}_2^c} p(x\,|\,\omega_2)\,dx. \tag{6}$$

If not, then we could decrease the highest value slightly, while increasing the lowest one (like the motion of a weighing scale) and obtain a better partition. Indeed, let R_2 be such that $\int_{R_2} p(x\,|\,\omega_1)dx > \int_{R_2^c} p(x\,|\,\omega_2)dx$ and let $\delta = \int_{R_2} p(x\,|\,\omega_1)dx - \int_{R_2^c} p(x\,|\,\omega_2)dx$. Pick some $A \subseteq R_2$ such that $0 < \int_A p(x\,|\,\omega_1) < \frac{\delta}{3}$ and $0 < \int_A p(x\,|\,\omega_2) < \frac{\delta}{3}$ both hold. Then

$$\begin{aligned} \int_{R_2} p(x\,|\,\omega_1)dx &> \int_{R_2 \setminus A} p(x\,|\,\omega_1)dx \\ &> \int_{(R_2 \setminus A)^c} p(x\,|\,\omega_2)dx \\ &> \int_{R_2^c} p(x\,|\,\omega_2)dx, \end{aligned}$$

which implies that the set $R_2 \setminus A$ achieves a smaller value for the minumum than R_2. Therefore, R_2 cannot be a minimizer for our problem.

We showed that every minimizer has to satisfy relation (6). We stress, however, that the opposite is not true. Namely, not any set R_2 which satisfies (6)

is necessarily a minimizer for our problem.[2] As an example, let A_1, A_2, A_3, A_4 be four disjoint intervals, each having length 1, and define $f(x) = \frac{1}{3} I_{A_1}(x) + \frac{2}{3} I_{A_2}(x)$ and $g(x) = \frac{1}{3} I_{A_3}(x) + \frac{2}{3} I_{A_4}(x)$. Now, let $A = I_1 \cup I_4$ and $B = I_2 \cup I_3$.

Then $\int_A f = \int_{A^c} g = \frac{1}{3}$, $\int_B f = \int_{B^c} g = \frac{2}{3}$, so although both A and B satisfy (6), they are not both minimizers, as $\frac{1}{3} \neq \frac{2}{3}$.

To return to our exercise, we need to find a counterexample where the solution is not unique. The A and B provided above do not constitute one, since they are not both minimizers. However, one can easily modify them.

A general rule to construct distinct minimizers is to consider distrbutions f, g with the property that $\operatorname{spt} f \subsetneq S_1$ and $\operatorname{spt} g \subsetneq S_2$, with S_1, S_2 disjoint. Here, a minimizer always exists but it is not unique: Any B such that $\operatorname{spt} f \subseteq B \subseteq S_1$ will minimize the risk.

The same can be achieved if $f \equiv g$ on $\operatorname{spt} f \cap \operatorname{spt} g$. In this case, partition $\operatorname{spt} f \cap \operatorname{spt} g = I \cup J$ into two sets such that $\int_I f = \int_J f$ and set $B = I \cup (\operatorname{spt} f \setminus \operatorname{spt} g)$ where B is the same as before. Such a partition for $\operatorname{spt} f \cap \operatorname{spt} g$ can also be achieved with several ways, always leading to distinct minimizers.

We conclude the exercise with an example where the minimizer R_2 is unique almost everywhere. Let $X|\omega_1$ be a half-normal distribution and $X|\omega_2$ be the symmetric of a half-normal distribution:

$$p(x|\omega_1) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi}} e^{-\frac{x^2}{2}}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

$$p(x|\omega_2) = \begin{cases} \frac{\sqrt{2}}{\sqrt{\pi}} e^{-\frac{x^2}{2}}, & x \leq 0, \\ 0, & x > 0. \end{cases}$$

Then the only minimizers are $R_2^* = (0, +\infty)$ and $R_2^{**} = [0, +\infty)$ which are equal a.s..

Exercise 1.3

We remind first that if $X : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$ is a random variable and $T : \mathbb{R}^n \to \mathbb{R}^m$ a measurable function, then we can define the random variable $Y = T \circ X$ from Ω to $\mathbb{R}^m$. If P_X is the distribution of X, namely the probability measure on $\mathbb{R}^n$ for which $P_X(A) = P(X^{-1}(A))$ for every $A \subseteq \mathbb{R}^n$ Borel, then the distribution P_Y of Y on $\mathbb{R}^m$ is just the measure P_Y for which

$$P_Y(B) = P(Y^{-1}(B)) = P(X^{-1}(T^{-1}(B))) = P_X(T^{-1}(B)).$$

[2] Interestingly, one can always find sets R_2 which satisfy (6): Fix some $x_0 \in \mathbb{R}$ and for every $\varepsilon > 0$ set $I_\varepsilon = [x_0 - \varepsilon, x_0 + \varepsilon]$. Define $F(\varepsilon) = \int_{I_\varepsilon} p(x|\omega_1)dx - \int_{I_\varepsilon^c} p(x|\omega_2)dx$. Then F is continuous with $F(0) = -1$, $\lim_{\varepsilon \to \infty} F(\varepsilon) = 1$, so there exists some ε^* with $F(\varepsilon^*) = 0$.

Consequently, the distributions of the two random variables X and Y satisfy the relation $P_Y(B) = P_X(T^{-1}(B))$ for every $B \subseteq \mathbb{R}^m$ Borel.

We return to the exercise. We will show something more general:

Proposition 1. *Let $X_1, X_2 : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$ be random variables with priors, p_1, p_2, let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a measurable function[3] and set $Y_1 = T \circ X_1, Y_2 = T \circ X_2$. Then the Bayes error of Y_1, Y_2 can not be smaller than the one of X_1, X_2.*

Proof. Suppose that $\tilde{C}$ is a classifier on $\mathbb{R}^m$ which picks ω_1 on $\tilde{R}_1 \subseteq \mathbb{R}^m$ and ω_2 on its complement $\tilde{R}_2$. Its Bayes error is

$$\begin{aligned} P_e(\tilde{C}) &= p_1 P(Y_1 \in \tilde{R}_2) + p_2 P(Y_2 \in \tilde{R}_1) \\ &= p_1 P_{Y_1}(\tilde{R}_2) + p_2 P_{Y_2}(\tilde{R}_1). \end{aligned} \tag{7}$$

The proof idea is very simple. With the help of the previous classifier, we will construct a new classifier C on the original space $\mathbb{R}^n$ with error equal to (7). The error of the new classifier can then be directly compared to the Bayes error on $\mathbb{R}^n$.

Let C be the classifier on $\mathbb{R}^n$ which classifies as follows: When a point $x \in \mathbb{R}^n$ arrives, it computes Tx. If $Tx \in \tilde{R}_1$, it classifies the point in ω_1, otherwise it classifies it in ω_2.

We will compute the Bayes error for C. Let $R_1 = T^{-1}(\tilde{R}_1)$[4] and $R_2 = T^{-1}(\tilde{R}_2)$. Then,

$$\begin{aligned} P_e(C) &= p_2 P(X_2 \in R_1) + p_1 P(X_1 \in R_2) \\ &= p_2 P(X_2 \in T^{-1}(\tilde{R}_1)) + p_1 P(X_1 \in T^{-1}(\tilde{R}_2)) \\ &= p_2 P_{X_2}(T^{-1}(\tilde{R}_1)) + p_1 P_{X_1}(T^{-1}(\tilde{R}_2)) \\ &= p_2 P_{Y_2}(\tilde{R}_1) + p_1 P_{Y_1}(\tilde{R}_2), \end{aligned}$$

so $P_e(C) = P_e(\tilde{C})$. However, since C is a classifier on $\mathbb{R}^n$, its error can not be smaller than the Bayes error $P_e(\text{Bayes})$. Therefore, $P_e(\text{Bayes}) \leq P_e(C) = P_e(\tilde{C})$ as we wanted. □

Both of the questions of the exercise are special cases of the previous proposition. The first one is stated for specific (normal) distributions, dimensions ($n = 2$) and also T is supposed to be a projection, while the second one relaxes only the distribution assumption.

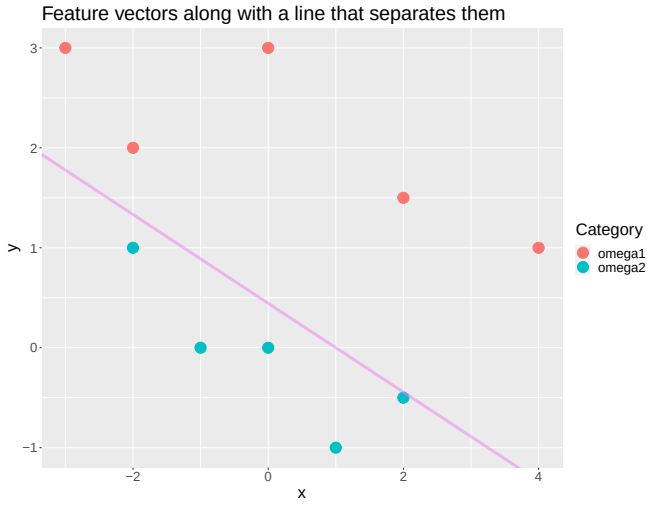

Figure 1. The 10 feature vectors along with the separating line, provided by the perceptron algorithm.

Exercise 1.4

The detailed computations are rather lengthy and are included in the end of our paper. Here, we explain the procedure we followed. Firstly, we drew the points and obtained Figure 1 (without the separating line obviously). We noticed that the points x_4, x_8, x_2 and x_{10} are the ones that play the most important role. This means that if we manage to find a separating line for the smaller dataset consisting only of these four points, then the same line will separate the original dataset.[5]

For this reason we made sure that these four points were fed first during each epoch. No matter the order in which the points are picked, the algorithm will always converge to a separating line, but by picking an appropriate order, one may hope to achieve the convergence much earlier, as was the case here.

The algorithm converged to the vector $w = (-2, 2, 4.5)$, which defines the line $y = \frac{2}{4.5} - \frac{2x}{4.5}$. The convergence was achieved very early in the second epoch. In total, it required to run the algorithm for two full epochs and 4 additional iterations in the third epoch to confirm that convergence was achieved.

The vectors were augmented by adding an extra coordinate in the beginning, which was set to 1 for all vectors. So x_1 became $x_1 = (1, 2, 1.5)$ etc. In general, the vector $w = (w_1, w_2, w_3)$ will define the line $w_1 + w_2 x + w_3 y = 0$. Without the augmentation, w_1 will be equal to zero, restricting the algorithm to only describe lines that pass through the origin. By inspecting Figure 1, it should be clear that no such line can separate our data points, so the augmentation was indeed necessary.

Exercise 1.5

As stated, the conclusion of the exercise does not hold, and we will provide a counterexample later. We will prove, however, that the conclusion is correct if one reverses the roles of p_1 and p_2.

We will prove the exercise without using Lagrange multipliers. Hopefully, it will be a pleasant change to see a different argument. We will rely on the following inequality:

Lemma 2. *Let Σ_1, Σ_2 be positive definite matrices. Then*

$$\ln \frac{|\Sigma_2|}{|\Sigma_1|} \geq \operatorname{tr}((\Sigma_1 - \Sigma_2)^{-1}\Sigma_1) \tag{8}$$

Proof. Let X, Y be multivariate normal variables with zero means and covariance matrices Σ_1, Σ_2 respectively. Let also p, q denote their respective densities. A direct consequence of the Jensen inequality is that the Kullback-Leibler divergence has the property that $d_{KL}(p, q) \geq 0$ for every distributions p, q. This gives us the inequality $\int p \ln p \geq \int p \ln q$. Lets compute these two expressions for the distributions of our Lemma:

$$\begin{aligned}\int p \ln p &= \int p(x)\left(-\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}x'\Sigma_1^{-1}x\right)dx \\ &= -\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}\mathbb{E}_{X\sim N(0,\Sigma_1)}[X'\Sigma_1^{-1}X],\end{aligned}$$

with

$$\begin{aligned}\mathbb{E}_{X\sim N(0,\Sigma_1)}[X'\Sigma_1^{-1}X] &= \mathbb{E}_{X\sim N(0,\Sigma_1)}[\operatorname{tr}(X'\Sigma_1^{-1}X)] \\ &= \mathbb{E}_{X\sim N(0,\Sigma_1)}[\operatorname{tr}(\Sigma_1^{-1}XX')] \\ &= \operatorname{tr}\left(\mathbb{E}_{X\sim N(0,\Sigma_1)}[\Sigma_1^{-1}XX']\right) \\ &= \operatorname{tr}\left(\Sigma_1^{-1}\mathbb{E}_{X\sim N(0,\Sigma_1)}[XX']\right) \\ &= \operatorname{tr}(\Sigma_1^{-1}\Sigma_1) \\ &= l.\end{aligned}$$

Therefore,

$$\int p \ln p = -\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_1| - \frac{l}{2}, \tag{9}$$

and similarly,

$$\int p \ln q = -\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_2| - \frac{1}{2}\operatorname{tr}(\Sigma_2^{-1}\Sigma_1). \tag{10}$$

Putting it all together,

$$\int p \ln p \geq \int p \ln q \qquad \Rightarrow$$

[3] Here T is not necessarily a projection, nor a linear map. We don't even assume that $m \leq n$, nor that X_1, X_2 should be absolutely continuous with respect to the Lebesgue measure.

[4] This is just the inverse image of the set $\tilde{R}_1$ under T. It is always well defined, regardless of whether T is invertible or not.

[5] One could even run the perceptron algorithm using only these four points and confirm our claim.

$$-\frac{1}{2}\ln|\Sigma_1| - \frac{1}{2}\operatorname{tr}(\Sigma_1^{-1}\Sigma_1) \geq -\frac{1}{2}\ln|\Sigma_2| - \frac{1}{2}\operatorname{tr}(\Sigma_2^{-1}\Sigma_1) \quad \Rightarrow$$
$$\ln\frac{|\Sigma_2|}{|\Sigma_1|} \geq \operatorname{tr}((\Sigma_1 - \Sigma_2)^{-1}\Sigma_1),$$

as we wanted. □

Using the previous Lemma, we can solve the exercise:

Proposition 3. *Let p_2 be the density of a random variable with mean μ and covariance matrix Σ. Then the problem*

$$\begin{aligned} \text{minimize} \quad & d_{KL}(p_2, p), \\ s.t. \quad & p \sim N(\tilde{\mu}, \tilde{\Sigma}), \end{aligned}$$

has $p^* \sim N(\mu, \Sigma)$ as its solution.

Proof. Let $p \sim N(\tilde{\mu}, \tilde{\Sigma})$ and $p^* \sim N(\mu, \Sigma)$. Then $d_{KL}(p_2, p^*) \leq d_{KL}(p_2, p)$, if and only if $\int p_2 \ln p^* \geq \int p_2 \ln p$. The second quantity is

$$\int p_2 \ln p = -\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\tilde{\Sigma}| - \frac{1}{2}\mathbb{E}_{X\sim p_2}[(X-\tilde{\mu})'\tilde{\Sigma}^{-1}(X-\tilde{\mu})],$$

with

$$\mathbb{E}_{X\sim p_2}\left[(X-\tilde{\mu})'\tilde{\Sigma}^{-1}(X-\tilde{\mu})\right] = \operatorname{tr}\left(\tilde{\Sigma}^{-1}\mathbb{E}_{X\sim p_2}\left[(X-\tilde{\mu})(X-\tilde{\mu})'\right]\right).$$

By adding and subtracting the term μ in the expectation, we obtain that

$$\begin{aligned} \mathbb{E}_{X\sim p_2}\left[(X-\tilde{\mu})(X-\tilde{\mu})'\right] &= \mathbb{E}_{X\sim p_2}\left[(X-\mu)(X-\mu)'\right] + \\ &+ 2\mathbb{E}_{X\sim p_2}\left[(X-\mu)(\mu-\tilde{\mu})'\right] + \\ &+ \mathbb{E}_{X\sim p_2}\left[(\mu-\tilde{\mu})(\mu-\tilde{\mu})'.\right] \\ &= \Sigma + (\mu-\tilde{\mu})(\mu-\tilde{\mu})'. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}_{X\sim p_2}\left[(X-\tilde{\mu})'\tilde{\Sigma}^{-1}(X-\tilde{\mu})\right] &= \operatorname{tr}\left(\tilde{\Sigma}^{-1}\Sigma + \tilde{\Sigma}^{-1}(\mu-\tilde{\mu})(\mu-\tilde{\mu})'\right) \\ &= \operatorname{tr}\left(\tilde{\Sigma}^{-1}\Sigma\right) + \|\mu-\tilde{\mu}\|_{\tilde{\Sigma}}, \end{aligned}$$

where $\|\mu-\tilde{\mu}\|_{\tilde{\Sigma}}$ is the Mahalanobis norm induced by $\tilde{\Sigma}$. To sum up,

$$\int p_2 \ln p = -\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\tilde{\Sigma}| - \frac{1}{2}\operatorname{tr}(\tilde{\Sigma}^{-1}\Sigma) - \frac{1}{2}\|\mu-\tilde{\mu}\|_{\tilde{\Sigma}}.$$

By working similarly, we can compute the first quantity as

$$\int p_2 \ln p^* = -\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\Sigma\right).$$

So,

$$\begin{aligned} & d_{KL}(p_2, p^*) \leq d_{KL}(p_2, p) \iff \\ & \int p_2 \ln p^* \geq \int p_2 \ln p \iff \\ & -\frac{1}{2}\ln|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\Sigma\right) \geq -\frac{1}{2}\ln|\tilde{\Sigma}| - \frac{1}{2}\operatorname{tr}(\tilde{\Sigma}^{-1}\Sigma) - \frac{1}{2}\|\mu-\tilde{\mu}\|_{\tilde{\Sigma}}. \end{aligned}$$

The last inequality is equivalent to

$$\operatorname{tr}\left((\Sigma-\tilde{\Sigma})^{-1}\Sigma\right) \leq \ln\frac{|\tilde{\Sigma}|}{|\Sigma|} + \|\mu-\tilde{\mu}\|_{\tilde{\Sigma}} \tag{11}$$

which clearly holds, since $\operatorname{tr}\left((\Sigma-\tilde{\Sigma})^{-1}\Sigma\right) \leq \ln\frac{|\tilde{\Sigma}|}{|\Sigma|}$ by the previous Lemma and the quantity $\|\mu-\tilde{\mu}\|_{\tilde{\Sigma}}$ is non-negative. □

We finish the exercise by providing a counterexample for the actual claim of the exercise, before our correction. Recall that the Kullback-Leibler divergence is not symmetric [Joy11], meaning that $d_{KL}(p, q)$ is not necessarily equal to $d_{KL}(q, p)$, so the order in which the two distributions are taken is important.

Suppose, for the sake of contradiction, that the conclusion of the exercise is correct. Let p_2 be a one dimensional distribution with mean $\mu = 0$ and variance σ^2, and let p^* be the corresponding Normal distribution. Then the inequality $d_{KL}(p^*, p_2) \leq d_{KL}(p, p_2)$, which must hold for every $p \sim N(0, \tilde{\sigma}^2)$, implies that

$$\ln\frac{\tilde{\sigma}}{\sigma} \leq \int p^*(x)\ln p_2(x)dx - \int p(x)\ln p_2(x)dx \tag{12}$$

We pick $p_2(x) = \frac{1}{2}e^{-|x|}$ to be the Laplace distribution with mean $\mu = 0$ and variance $\sigma^2 = 2$. If $q(x) \sim N(0, \tau^2)$, then

$$\begin{aligned} \int q(x)\ln p_2(x)dx &= -\frac{1}{2}\int_{-\infty}^{\infty}|x|q(x)dx \\ &= -\frac{1}{2}\int_0^{\infty}\frac{2x}{\sqrt{2\pi}\tau}e^{-\frac{x^2}{2\tau^2}}dx. \end{aligned}$$

The last integral is just the expectation of the Half-Normal distribution and is equal to $\frac{\tau\sqrt{2}}{\sqrt{\pi}}$. So, $\int p^*(x)\ln p_2(x)dx = -\frac{1}{2}\frac{\sqrt{2\tau}}{\sqrt{\pi}}$, therefore, relation (12) can be re-written as

$$\ln\sigma \leq \frac{\sqrt{2}}{2\sqrt{\pi}}\sigma + \ln\sqrt{2} - \frac{1}{\sqrt{\pi}}, \tag{13}$$

which must hold for any $\sigma > 0$. However, this is not the case, just substitute the value $\sigma = 2.5$ to obtain a contradiction.

Exercise 1.6

1) Given a point X_i, the observation Y_i can be written as $Y_i = b_0 + b_1 X_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$ independent. Therefore, every Y_i follows a normal distribution $N(b_0 + b_1 X_i, \sigma^2)$. The likelihood of a sample $\tilde{y} = (y_1, \ldots, y_n)$ is

$$L(\tilde{y}, b_0, b_1) = \prod_i^n \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i - b_0 - b_1 x_i)^2}{2\sigma^2}}$$

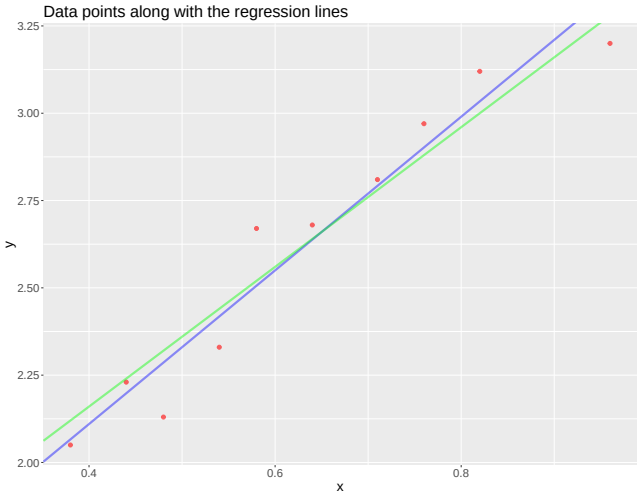

Figure 2. The data points along with the two regression lines. The blue one corresponds to the least squares estimate, whereas the green one to the one provided by the LMS algorithm.

$$= ce^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2}.$$

Additionally,

$$\ell(b_0,b_1) = \ln L(\tilde{y},b_0,b_1) = c' - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2$$

so

$$\begin{aligned}\arg\max_{b_0,b_1}\ell(b_0,b_1) &= \arg\max_{b_0,b_1}\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2\right\}\\ &= \arg\min_{b_0,b_1}\left\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2\right\}\\ &= \arg\min_{b_0,b_1}\left\{\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2\right\}. \qquad (14)\end{aligned}$$

Using the least square method, the goal is to minimize the quantity $\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2$, namely the computation of $\arg\min_{b_0,b_1}\{\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2\}$. This expression is identical to (14), so the two methods yield the same solution.

2) We will use the formulas $\widehat{b}_0 = \overline{y} - \widehat{b}_1\overline{x}$, and $\widehat{b}_1 = \frac{\sum_{i=1}^{n}x_iy_i - n\overline{xy}}{\sum_{i=1}^{n}x_i^2 - n\overline{x}^2}$. We have that

$$\begin{aligned}\sum_i x_iy_i &= 0.779+0.981+1.022+1.258+1.548+1.715+\\ &\quad +1.995+2.257+2.558+3.072\\ &= 17.187,\\ \sum_i x_i^2 &= 0.144+0.193+0.230+0.291+0.336+0.409+\\ &\quad +0.504+0.577+0.672+0.922\\ &= 4.282,\\ n\overline{x}\cdot\overline{y} &= 16.526, \quad n\overline{x}^2 = 3.982,\end{aligned}$$

and by substitution we find that $\widehat{b}_0 = 1.23$ and $\widehat{b}_1 = 2.2$. The resulting regression line is $Y = 1.23 + 2.2X$.

3) We executed the LMS algorithm, starting from $b_0 = (1,2)$ and renewing this value according to the formula $b \leftarrow b + \frac{1}{k}(y_i - b'\tilde{x}_i)\tilde{x}_i$, where $\tilde{x}_i = [1, x_i]$ is the augmented i-th observation and k the current epoch. After a single epoch we ended up with the regression line $Y = 1.12 + 1.94X$. Full details in the end of the paper.

We implemented the LMS algorithm in R:

```
x <- c(.38, .44, .48, .54, .58, .64, .71, .76, .82, .96)
y <- c(2.05, 2.23, 2.13, 2.33, 2.67, 2.68, 2.81, 2.97,
    3.12, 3.2)

ones = rep(1,10)
newx = t(rbind(ones, x))
newx = as.data.frame(newx)

hoff <- function(Y, X, b0 = c(1,1), epochs=1){
  b = b0
  N = length(X[,1])
  for (epoc in 1:epochs){
    for (i in 1:N){
      delta_b = (1/epoc) * (Y[i] - sum(X[i, ]*b)) * X[i, ]
      b <- b + delta_b
      b <- as.numeric(b)
      }
    }
  print(b) }

hoff(Y=y, X=newx, b0=c(1,2), epochs=10^4)

1.256728 2.159880
```

After 10^4 epochs, the resulting regression line was $Y = 1.26 + 2.16X$.

Exercise 1.7

1) The Bayes rule will pick the category which achieves the following maximum:

$$\begin{aligned}\arg\max_{i=1,2}p(\omega_i\,|\,x) &= \arg\max_{i=1,2}\frac{p(x\,|\,\omega_i)p(\omega_i)}{p(x)}\\ &= \frac{1}{2}\arg\max_{i=1,2}p(x\,|\,\omega_i)\\ &= \arg\max\{2x, 2-2x\}\\ &= \begin{cases}1, & x\in\left[\frac{1}{2},1\right],\\ 2, & x\in\left[0,\frac{1}{2}\right).\end{cases}\end{aligned}$$

The Bayes error is equal to

$$\begin{aligned}P_e &= P(\omega_1)\int_{R_2}p(x\,|\,\omega_1)dx + P(\omega_2)\int_{R_1}p(x\,|\,\omega_2)dx\\ &= \frac{1}{2}\int_0^{\frac{1}{2}}2xdx + \frac{1}{2}\int_{\frac{1}{2}}^{1}(2-2x)dx\end{aligned}$$

$$= \frac{1}{8} + \frac{1}{8}$$
$$= 0.25.$$

4) We will solve the general case $P_n(e)$ right away, so questions 2) and 3) can be omitted. The proof is far from straightforward and will be based on the following building blocks:

Lemma 4. *Let X, Y be independent random variables with cdfs F_X, F_Y and pdfs p_X, p_Y respectively. Then*

$$P[Y \le X] = \int F_Y(t) p_X(t) dt. \tag{15}$$

Proof. By conditioning on the value of X,

$$\begin{aligned} P[Y \le X] &= \int P[Y \le t \mid X = t] P[X = t] dt \\ &= \int F_Y(t) p_X(t) dt, \end{aligned}$$

as we wanted. □

Lemma 5. *Let $X_1, \dots, X_n$ be i.i.d. random variables with cdf F_X and pdf p_X. Set $Z = \min_{i=1,\dots,n} X_i$ to be their minimum. Then the cdf and pdf of Z are*

$$F_Z(t) = 1 - (1 - F_X(t))^n \quad \text{and} \tag{16}$$
$$p_Z(t) = n p_X(t) (1 - F_X(t))^{n-1} \tag{17}$$

respectively.

Proof. By the independence of the X_i's,

$$\begin{aligned} P[Z \le t] &= 1 - P[Z > t] \\ &= 1 - P[X_i > t, \quad \text{for all } i] \\ &= 1 - \prod_{i=1}^{n} P[X_i > t] \\ &= 1 - P[X > t]^n \\ &= 1 - (1 - F_X(t))^n. \end{aligned}$$

For the pdf, just compute the derivative of F_Z. □

Lemma 6. *Let X be a random variable with cdf F_X and pdf p_X, and fix some $a \in \mathbb{R}$. Then the random variable $Z = |X - a|$ has the following distribution:*

$$F_Z(t) = F_X(a + t) - F_X(a - t), \tag{18}$$
$$p_Z(t) = p_X(a + t) + p_X(a - t). \tag{19}$$

Proof. The proof is rather elementary, as

$$P[Z \le t] = P[a - t \le X \le a + t] = F_X(a + t) - F_X(a - t).$$

The formula for p_Z then follows. □

Putting all the previous Lemmas together, we get our main tool:

Lemma 7. *Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be i.i.d. with distributions F_X, p_X and F_Y, p_Y respectively, and fix some $a \in \mathbb{R}$. Set $p_n(a) = P[\min_i |Y_i - a| \le \min_i |X_i - a|]$. Then*

$$\begin{aligned} p_n(a) = n \int & \left(1 - (1 + F_Y(a - t) - F_Y(a + t))^n\right) \cdot \\ & \cdot (1 + F_X(a - t) - F_X(a + t))^{n-1} \cdot \\ & \cdot \left(p_X(a + t) + p_X(a - t)\right) dt. \end{aligned} \tag{20}$$

Proof. Combining all the previous lemmas,

$$\begin{aligned} p_n(a) &= \int F_{\min|Y_i - a|}(t) p_{\min|X_i - a|}(t) dt \\ &= n \int \left(1 - (1 - F_{|Y-a|}(t))^n\right) \left(1 - F_{|X-a|}(t)\right)^{n-1} p_{|X-a|}(t) dt \\ &= n \int \left(1 - (1 + F_Y(a - t) - F_Y(a + t))^n\right) \cdot \\ &\quad \cdot (1 + F_X(a - t) - F_X(a + t))^{n-1} \cdot \\ &\quad \cdot \left(p_X(a + t) + p_X(a - t)\right) dt, \end{aligned}$$

as we wanted. □

Lemma 7 provides us with a formula for the expected error given that a test point $x_{\text{test}} = a$ has arrived. To compute the expected error $P_n(e)$, we need to condition on the value of x_{test} and integrate:

Lemma 8. *With the same notation as in the exercise statement,*

$$P_n(e) = \int_0^1 p_n(a) p_X(a) da = \int_0^1 2a p_n(a) da, \tag{21}$$

where $p_n(a)$ is given by (20) for the general case and by Table I (line 2) for the distributions of the exercise.

We substituted the distributions given in the exercise, $p_X(x) = 2x$, $F_X(x) = x^2$, $p_Y(x) = 2 - 2x$ and $F_Y(x) = 2x - x^2$ into (20) and computed $p_n(a)$. Detailed computations can be found in the end of our paper. Here we only include the final formula for the general $p_n(a)$ as well as the special case for $n = 1$ (see Table I).

We did not manage to simplify these formulas any further, due to the nasty integrals that appear in it. However, for $n = 1$ these integrals simplified a lot. The error $P_1(e)$ was computed to be $P_1(e) = 0.35$, a value that was also confirmed through simulation (see Appendix A). Similarly, $P_2(e)$ was found to be 0.34.

Notice that these values are different from the ones obtained by Cover and Hart in their seminal paper [CH67, Paragraph VII], where they seemingly study the same problem. They manage to find a much more elegant

$$p_n(a) = \; n\int \left(1-(1+F_Y(a-t)-F_Y(a+t))^n\right)(1+F_X(a-t)-F_X(a+t))^{n-1}\left(p_X(a+t)+p_X(a-t)\right)dt.$$

$$p_n(a) = \begin{cases} 4an\int_0^a \left[1-(1-4t(1-a))^n\right](1-4at)^{n-1}dt + 2n\int_a^{1-a}\left[1-\left(1-2(a+t)+(a+t)^2\right)^n\right](a+t)\left(1-(a+t)^2\right)^{n-1}dt, & 0\le a<\frac{1}{2}, \\ 4an\int_0^{1-a} \left[1-(1-4t(1-a))^n\right](1-4at)^{n-1}dt + 2n\int_{1-a}^{a}\left[1-\left(2(a-t)-(a-t)^2\right)^n\right](a-t)\left(1+(a-t)^2\right)^{n-1}dt, & \frac{1}{2}\le a\le 1, \end{cases}$$

$$p_1(a) = \begin{cases} \frac{5}{6}-\frac{8}{3}a^3, & 0\le a<\frac{1}{2}, \\ 8a(1-a)^3+(2a-1)^2-\frac{4}{3}(2a-1)^3+\frac{(2a-1)^4}{2}, & \frac{1}{2}\le a\le 1. \end{cases}$$

Table I
The general formula for $p_n(a)$, the formula for $p_n(a)$ for the distributions given in Exercise 1.7, and the special case for $n=1$.

formula for the expected error, $P_n(e) = \frac{1}{3} + \frac{1}{(n+1)(n+2)}$, which gives $P_1(e) = 0.5$ and $P_2(e) = 0.42$.

Given their definition of $P_n(e)$, their $P_2(e) = 0.42$ should be comparable to our $P_1(e) = 0.35$, however they are not the same. The discrepancy is due to the fact that the actual problem they are solving is slightly different to ours. To compute $P_n(e)$ they use n total points which are drawn at random from $p(x)$ (and not equally divided among the two classes as we do).

5) The limit $\lim_{n\to\infty} P_n(e)$ is equal to

$$\begin{aligned}\lim_{n\to\infty} P_n(e) &= \int \left(1-p^2(\omega_1\,|\,x)-p^2(\omega_2\,|\,x)\right)p(x)dx \\ &= \int\left(1-\frac{1}{4}\frac{p^2(x\,|\,\omega_1)}{p^2(x)}-\frac{1}{4}\frac{p^2(x\,|\,\omega_2)}{p^2(x)}\right)dx \\ &= 1-\int_0^1 x^2dx-\int_0^1(1-x)^2dx \\ &= \frac{1}{3}.\end{aligned}$$

By the Cover-Hart inequality [CH67], for classification problems with two classes,

$$P^* \le \lim_{n\to\infty} P_n(e) \le 2P^*(1-P^*), \tag{22}$$

where P^* denotes the Bayes error. The quantities we found satisfy this inequality indeed:

$$P^* = 0.25 < \lim_{n\to\infty} P_n(e) = 0.33 < 2P^*(1-P^*) = 0.38.$$

Exercise 1.8

This exercise also has an obvious typo. The distribution $p(x_1)$ as given, integrates to 1 if and only if $\vartheta_1 = 1$, rendering any effort to estimate ϑ_1 from the sample, pointless. Not to mention that the initial choice for ϑ^0 picks a $\vartheta_1 = 2$ which does not belong to the parameter space $\Theta_1 = \{1\}$.

What the exercise meant to give, was an exponential distribution, $p(x) = \vartheta_1 e^{-\vartheta_1 x}$, for $x \ge 0$. We proceed with the computation of $Q(\vartheta;\vartheta^0)$:

$$\begin{aligned}Q(\vartheta;\vartheta^0) &= \mathbb{E}_{x_{32}}\left[\sum_{k=1}^3 \ln p(x_k;\vartheta)\,|\,\vartheta^0\right] \\ &= \int \sum_{k=1}^3 \ln p(x_k;\vartheta)\,p\left(x_{32}\,|\,\vartheta^0, x_{31}=1\right)dx_{32} \\ &= \sum_{k=1}^2 \ln p(x_k;\vartheta) + \int \ln p(x_3;\vartheta)\,p(x_{32}\,|\,\vartheta^0)\,dx_{32}.\end{aligned}$$

For the first two summands,

$$\begin{aligned}\ln p(x_k\,|\,\vartheta) &= \ln\left(\vartheta_1 e^{-\vartheta_1 x_{k1}}\frac{1}{\vartheta_2}I_{[0,\vartheta_2]}(x_{k2})\right) \\ &= \ln\vartheta_1 - \vartheta_1 x_{k1} - \ln\vartheta_2 + \ln I_{[0,\vartheta_2]}(x_{k2}),\end{aligned}$$

whereas for the integral, $I = \int \ln p(x_3;\vartheta)\,p(x_{32}\,|\,\vartheta^0)\,dx_{32}$,

$$\begin{aligned}I &= \int\left(\ln\vartheta_1 - \ln\vartheta_2 - \vartheta_1 + \ln I_{[0,\vartheta_2]}(x_{32})\right)\frac{1}{3}I_{[0,3]}(x_{32})\,dx_{32} \\ &= \ln\vartheta_1 - \ln\vartheta_2 - \vartheta_1,\end{aligned}$$

as clearly $x_{32} \in [0,\vartheta_2]$. Combining them together,

$$\begin{aligned}Q(\vartheta,\vartheta^0) &= 2\ln\frac{\vartheta_1}{\vartheta_2} - 6\vartheta_1 + \ln I_{[0,\vartheta_2]}(2) + \ln I_{[0,\vartheta_2]}(5) \\ &= \begin{cases} -\infty, & \vartheta_2 < 5, \\ 3\ln\vartheta_1 - 3\ln\vartheta_2 - 6\vartheta_1, & \vartheta_2 \ge 5. \end{cases}\end{aligned}$$

For the M-step, clearly Q can attain its possible maximum only when $\vartheta_2 \ge 5$, so we can confine our search into this half-line. Additionally, since Q is decreasing with respect to ϑ_2, the candidate maximizer ϑ^* of Q will have the form $\vartheta^* = (\vartheta_1^*, 5)$. Computing the derivative of $Q = 3\ln\vartheta_1 - 3\ln 5 - 6\vartheta_1$, with respect to ϑ_1, we obtain that $Q' = \frac{3}{\vartheta_1} - 6$ and $Q'' = -\frac{3}{\vartheta_1^2} < 0$, with a critical point $\vartheta_1^* = \frac{1}{2}$ which corresponds to a local maximum.

Lastly, the function $Q(\vartheta_1) = 3\ln\vartheta_1 - 3\ln 5 - 6\vartheta_1$ is coercive, meaning that $Q(\vartheta_1) \to -\infty$ as $\vartheta_1 \to +\infty$, and

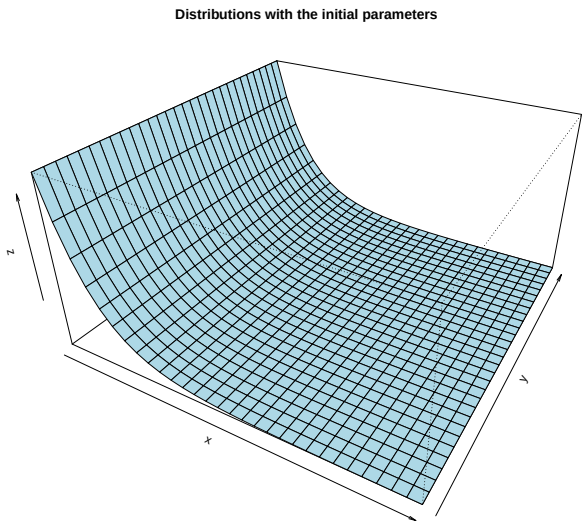

Figure 3. The distribution that corresponds to the initial parameter vector $\vartheta = (2,3)$.

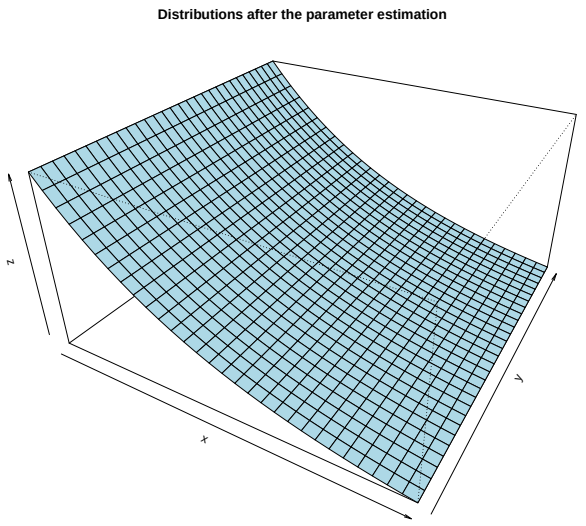

Figure 4. The distribution that corresponds to the parameter vector $\vartheta = (0.5,5)$, computed after one iteration of the EM-algorithm.

continuous. By a simple application of the Weierstrass theorem, it possesses a global maximum, and due to its differentiability, it must be a critical point of it. We conclude that $\vartheta^* = \left(\frac{1}{2},5\right)$ is the parameter vector that maximizes Q. In Figures 3 and 4 we can see the distributions before and after one iteration of the EM-algorithm.

Appendix

A) Estimating $P_n(e)$ in Exercise 1.7 using simulation, based on the inverse transform algorithm [Ros12], in R.

```
# Functions p1 and p2 generate one sample point
# from the distributions p(x|omega1) and
# p(x|omega2) respectively, using the inverse
# transform.

p1 <- function(x){return(sqrt(runif(x)))}
p2 <- function(x){return(1 - sqrt(runif(x)))}

my_sim <- function(n){
# Draws n-points from p_1, n_points from p_2
# and one test point from p_1.
# Classifies the test point according to the 1-NN
# rule. Returns 1 if the classification was correct,
# and 0 otherwise.

  pred = 0
  sam1 <- p1(n) # Draw n-points from omega_1
  sam2 <- p2(n) # Draw n-points from omega_2
  new_point = p1(1)  # Draw the test point
  full_sam = c(sam1, sam2)
  full_sam = abs(full_sam-new_point) # Distances
  min_ind=which.min(full_sam) # argmin for dist
  if (min_ind <= n){pred = 1} # label for argmin
  return(pred)
}

many_sim <- function(n, M){
# Executes M-repetitions of the my_sim simulation.
# Returns the average number of
# correct classifications.

  vec = rep(3, M)
  for (i in 1:M) {vec[i] = my_sim(n)}
  return(sum(vec) / M)
}


for (i in 1:3)
    {cat('P_',i,'=', 1 - many_sim(i, 10^8), '\n')}

P_1 = 0.3500071
P_2 = 0.3396602
P_3 = 0.3360257
```

References

[AB06] C. Aliprantis, K. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer, 3rd ed., 2006. doi: 10.1007/978-3-662-03004-2 Cited on p. 1

[CB01] G. Casella, R. Berger, *Statistical Inference*, Cengage Learning, 2001.

[CH67] T. Cover, P. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, 13, pp. 21-27, 1967. doi: 10.1109/TIT.1967.1053964 Cited on p. 6, 7

[Joy11] J. Joyce, Kullback-Leibler Divergence, In: Lovric M. (eds) International Encyclopedia of Statistical Science, Springer 2011. doi: 10.1007/978-3-642-04898-2_327 Cited on p. 4

[KT08] K. Koutroumbas, S. Theodoridis, *Pattern Recognition*, Academic Press, 2008. isbn: 978-1-59749-272-0

[KN05] Κουμουλλής, Γ., Νεγρεπόντης, Σ., *Θεωρία Μέτρου*, Εκδόσεις Συμμετρία, 2005. url: simmetria.gr

[Ros12] S. Ross, *Simulation*, Academic Press, 5th Edition, 2012. doi: 9780124158252 Cited on p. 8