

## Exercises 3.15

### Exercise 3.15.1

Let  $p$ ,  $p_i$ ,  $q$ ,  $q_i$  be density functions on  $\mathbb{R}$  and  $\alpha \in \mathbb{R}$ . Show that the cross-entropy satisfies the following properties:

- a.  $S(p_1 + p_2, q) = S(p_1, q) + S(p_2, q)$ ;
- b.  $S(\alpha p, q) = \alpha S(p, q) = S(p, q^\alpha)$ ;
- c.  $S(p, q_1 q_2) = S(p, q_1) + S(p, q_2)$ .

### Exercise 3.15.2

Show that the cross entropy satisfies the following inequality

$$S(p, q) \geq 1 - \int p(x)q(x)dx$$

### Exercise 3.15.3

Let  $p$  a fixed density. Show that the symmetric relative entropy

$$D_{KL}(p||q) + D_{KL}(q||p)$$

reaches its minimum for  $p = q$ , and the minimum is equal to zero.

### Exercise 3.15.4

Consider two exponential densities,  $p_1 = \xi^1 e^{\xi^1 x}$  and  $p_2 = \xi^2 e^{\xi^2 x}$ ,  $x \geq 0$ .

- a. Show that  $D_{KL}(p_1||p_2) = \frac{\xi^2}{\xi^1} - \ln \frac{\xi^2}{\xi^1} - 1$ .
- b. Verify  $D_{KL}(p_1||p_2) \neq D_{KL}(p_2||p_1)$ .
- c. Show that the triangle inequality doesn't hold for three arbitrary densities.

### Exercise 3.15.5

Let  $X$  be a discrete random variable. Show the inequality

$$H(X) \geq 0.$$

### Exercise 3.15.6

Prove that if  $p$  and  $q$  are the densities of two discrete random variables, then  $D_{KL}(p||q) \leq S(p, q)$

### Exercise 3.15.7

We assume the target variable  $Z$  is  $\mathcal{E}$ -mesurable. What is mean squared error function in this case?

### Exercise 3.15.8

Assume that a neural network has an input-output function  $f_{w,b}$  linear in  $w$  and  $b$ . Show that the cost function (3.3.1) reaches its minimum for a unique pair  $(w^*, b^*)$ , which can be computed explicitly.

### Exercise 3.15.9

Show that the Shannon entropy can be retrieved from the Reyni entropy as

$$H(p) = \lim_{\alpha \rightarrow 1} H_\alpha(x).$$

### Exercise 3.15.10

Let  $\phi_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$ . Consider the convolution operation  $(f * g)(x) := \int f(t)g(x-t)dt$ .

- Show that  $\phi_\sigma * \phi_\sigma = \phi_{\sigma\sqrt{2}}$ ;
- Find  $\phi_\sigma * \phi_{\sigma'}$  in the case  $\sigma \neq \sigma'$ .

### Exercise 3.15.11

Consider two probability densities,  $p(x)$  and  $q(x)$ . The Cauchy-Schwartz divergence is defined by

$$D_{CS}(p, q) := -\ln\left(\frac{\int p(x)q(x)dx}{\sqrt{\int p(x)^2 dx} \sqrt{\int q(x)^2 dx}}\right).$$

Show the following:

- $D_{CS}(p, q) = 0$  if and only if  $p = q$ ;
- $D_{CS}(p, q) \geq 0$ ;
- $D_{CS}(p, q) = D_{CS}(q, p)$ ;
- $D_{CS}(p, q) = -\ln \int pq dx - \frac{1}{2}H_2(p) - \frac{1}{2}H_2(q)$ , where  $H_2(\cdot)$  denotes the quadratic Reyni entropy.

### Exercise 3.15.12

- Show that for any function  $f \in L^1[0, 1]$  we have the inequality  $\|\tanh(f)\|_1 \leq \|f\|_1$ .
- Show that for any function  $f \in L^2[0, 1]$  we have the inequality  $\|\tanh(f)\|_2 \leq \|f\|_2$ .

### Exercise 3.15.13

Consider two distributions on the sample space  $\mathcal{X} = \{x_1, x_2\}$  given by

$$p = \begin{pmatrix} x_1 & x_2 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad q = \begin{pmatrix} x_1 & x_2 \\ \frac{1}{2} & \frac{2}{3} \end{pmatrix}$$

Consider the function  $\phi : \mathcal{X} \rightarrow \mathbb{R}^2$  defined by  $\phi(x_1) = (0, 1)$   $\phi(x_2) = (1, 0)$ . Find the maximum mean discrepancy between  $p$  and  $q$ .

## SOLUTIONS

### 3.15.1 (a)

The claim follows from the linearity of the integral operator. In symbols we have:

$$\begin{aligned} S(p_1 + p_2, q) &= - \int_{\mathbb{R}} (p_1(x) + p_2(x)) \ln q(x) dx = - \int_{\mathbb{R}} p_1(x) \ln q(x) dx - \int_{\mathbb{R}} p_2(x) \ln q(x) dx \\ &= S(p_1, q) + S(p_2, q). \end{aligned}$$

□

### 3.15.1 (b)

From the linearity of the integral operator, and the property  $c \ln(x) = \ln(x^c)$  we have:

$$\begin{aligned} S(\alpha p, q) &= - \int_{\mathbb{R}} \alpha p(x) \ln q(x) dx = -\alpha \int_{\mathbb{R}} p(x) \ln q(x) dx = \alpha S(p, q) \\ &= - \int_{\mathbb{R}} \alpha p(x) \ln q(x) dx = - \int_{\mathbb{R}} p(x) \ln q(x)^\alpha dx = S(p, q^\alpha). \end{aligned}$$

□

### 3.15.1 (c)

Using the addition identity for the logarithm, we get:

$$\begin{aligned} S(p, q_1 q_2) &= - \int_{\mathbb{R}} p(x) \ln q_1(x) q_2(x) dx = - \int_{\mathbb{R}} p(x) \ln q_1(x) dx - \int_{\mathbb{R}} p(x) \ln q_2(x) dx \\ &= S(p, q_1) + S(p, q_2). \end{aligned}$$

□

### 3.15.2

By the inequality  $\ln(x) \leq x - 1$ ,  $\forall x \in \mathbb{R}^+$ , and the definition of cross-entropy follows:

$$\begin{aligned} S(p, q) &= - \int_{\mathbb{R}} p(x) \ln q(x) dx \geq - \int_{\mathbb{R}} p(x) (q(x) - 1) dx \\ &\geq - \int_{\mathbb{R}} -p(x) dx - \int_{\mathbb{R}} p(x) q(x) dx = 1 - \int_{\mathbb{R}} p(x) q(x) dx. \end{aligned}$$

□

### 3.15.3

From proposition 3.5.1 follows that  $D_{KL}(p||q) \geq 0$ ,  $D_{KL}(q||p) \geq 0$ , then  $D_{KL}(p||q) + D_{KL}(q||p) \geq 0$ . Clearly the value 0 is a minimum. Let's now prove that this minimum is attained when  $p = q$ . It is well known from the cross-entropy definition  $S(p, p) = H(p)$  and  $S(q, q) = H(q)$  then:

$D_{KL}(p||q) = D_{KL}(p||p) = S(p, p) - H(p) = 0$  and  $D_{KL}(q||p) = D_{KL}(q||q) = S(q, q) - H(q) = 0$ , which in turn imply  $D_{KL}(p||q) + D_{KL}(q||p) = 0$ .

□

### 3.15.4 (a)

By direct calculation we find:

$$\begin{aligned}
D_{KL}(p_1 \| p_2) &= S(p_1, p_2) - H(p_1) = - \int_{\mathbb{R}} \xi^1 e^{-\xi^1 x} \ln(\xi^2 e^{-\xi^2 x}) dx - \int_{\mathbb{R}} \xi^1 e^{-\xi^1 x} \ln(\xi^1 e^{-\xi^1 x}) \\
&= - \int_{\mathbb{R}} \xi^1 e^{-\xi^1 x} \ln(\xi^2) dx + \int_{\mathbb{R}} \xi^1 e^{-\xi^1 x} \xi^2 x dx + \int_{\mathbb{R}} \xi^1 e^{-\xi^1 x} \ln(\xi^1) dx - \int_{\mathbb{R}} \xi^1 e^{-\xi^1 x} \xi^1 x dx \\
&= -(\ln(\xi^2) - \ln(\xi^1)) \int_{\mathbb{R}} \xi^1 e^{-\xi^1 x} dx + (\xi^2 - \xi^1) \int_{\mathbb{R}} \xi^1 x e^{-\xi^1 x} dx \\
&= -(\ln(\xi^2) - \ln(\xi^1)) \mathbb{E}_{X \sim \exp(\xi^1)} [1] + (\xi^2 - \xi^1) \mathbb{E}_{X \sim \exp(\xi^1)} [X] = -\ln \frac{\xi^2}{\xi^1} + (\xi^2 - \xi^1) \frac{1}{\xi^1} \\
&= -\ln \frac{\xi^2}{\xi^1} + \frac{\xi^2}{\xi^1} - 1
\end{aligned}$$

□

### 3.15.4 (b)

Suppose the equality  $D_{KL}(p \| p) = D_{KL}(q \| p)$  holds and  $\xi^1 \neq \xi^2$ , then from exercise 3.14.4.a it follows:  
 $-\ln \frac{\xi^2}{\xi^1} + \frac{\xi^2}{\xi^1} - 1 = -\ln \frac{\xi^1}{\xi^2} + \frac{\xi^1}{\xi^2} - 1 \implies \frac{\xi^2}{\xi^1} = \frac{\xi^1}{\xi^2}$ . The later implies  $\frac{\xi^1}{\xi^2} = 1$  or equivalently  $\xi^1 = \xi^2$ , which is a contradiction.

### 3.15.4 (c)

Let  $p_1 = \exp(2)$ ,  $p_2 = \exp(3)$ ,  $p_3 = \exp(4)$ . Suppose the triangle inequality holds for these three arbitrary exponential distributions. This is:

$D_{KL}(p_1 \| p_3) \leq D_{KL}(p_1 \| p_2) + D_{KL}(p_2 \| p_3)$ . By exercise 3.15.4.b we would have:

$$\begin{aligned}
D_{KL}(p_1 \| p_3) &= \frac{4}{2} - \ln \frac{4}{2} - 1 \leq D_{KL}(p_1 \| p_2) + D_{KL}(p_2 \| p_3) = \frac{3}{2} - \ln \frac{3}{2} - 1 + \frac{4}{3} - \ln \frac{4}{3} - 1 \\
2 &\leq \frac{3}{2} + \frac{4}{3} - 1 = \frac{17}{6} - 1 = \frac{11}{6} = \frac{12}{6} - \frac{1}{6} = 2 - \frac{1}{6} \text{ (contradiction!)}
\end{aligned}$$

□

### 3.15.5 (a)

Given that  $p(x)$  is a distribution, it follows that  $p(X)$  as a r.v satisfies the inequality  $0 \leq p(X) \leq 1$ . This means  $p(x) \leq 1, \forall x \in \text{sup}(X)$ . Taking natural logs on both sides of the inequality  $p(x) \leq 1$  and multiplying by  $-1$ , we obtain:  $\ln p(x) \geq 0$ ; Multiplying by  $p(x)$  and summing over the support of  $X$ , we get:

$$\mathbb{E}[-\ln p(X)] = H(X) = \sum_{x \in \text{sup}(X)} -p(x) \ln(p(x)) \geq 0.$$

□

### 3.15.6

This is an immediate consequence of exercise 3.15.5. Indeed, we have:

$$D_{KL}(p \| q) = S(p, q) - H(p) \leq S(p, q) - 0 \leq S(p, q).$$

□

### 3.15.7

If the target variable  $Z$  happens to be  $\mathcal{E}$ -mesurable, then  $Y$  is independent of the sigma algebra  $\mathcal{E}$ . From this follows that  $C(\omega, b) = d(Z, Y)^2 = \mathbb{E}[(Z - \mathbb{E}[Z|\mathcal{E}])^2] = \mathbb{E}[(Z - Z)^2] = 0$ .

### 3.15.8

In this case  $f_{\omega, b}(\mathbf{x}) = \omega \cdot \mathbf{x} + b$ , defined on a compact subset of  $\mathbb{R}^n$ . Therefore, the cost function is given by:

$C(\omega, b) := \sum_{0 \leq i \leq n} (\omega \cdot \mathbf{x}^i + b - \phi(\mathbf{x}^i))^2$ . Obviously we have  $0 \leq C(\omega, b)$ . This means the function attains such

minimum inside the compact set; Let  $\mathbf{x}^i$  the  $n$ -dimensional observations, i.e  $\mathbf{x}^i = (x_1^i, \dots, x_n^i)$ . Then, the normal equations for the  $\omega_k$  (the componentes of the vector  $\omega$ ),  $\forall k \in [n]$  and the bias parameter  $b$  are:

$$\begin{cases} \sum_{0 \leq j \leq n} \omega_j \sum_{0 \leq i \leq n} x_j^i x_k^i + b \sum_{0 \leq i \leq n} x_k^i = \sum_{0 \leq i \leq n} \phi(\mathbf{x}^i) x_k^i, \forall k \in [n] \\ \sum_{0 \leq j \leq n} \omega_j \sum_{0 \leq i \leq n} x_j^i + nb = \sum_{0 \leq i \leq n} \phi(\mathbf{x}^i) \end{cases} \quad (1)$$

This system of equations has the following matricial expression:

$$\begin{bmatrix} \sum_{0 \leq i \leq n} x_1^i x_1^i & \sum_{0 \leq i \leq n} x_2^i x_1^i & \cdots & \sum_{0 \leq i \leq n} x_m^i x_1^i & \cdots & \sum_{0 \leq i \leq n} x_1^i \\ \sum_{0 \leq i \leq n} x_1^i x_2^i & \sum_{0 \leq i \leq n} x_2^i x_2^i & \cdots & \sum_{0 \leq i \leq n} x_m^i x_2^i & \cdots & \sum_{0 \leq i \leq n} x_2^i \\ \vdots & \vdots & & \vdots & & \vdots \\ \sum_{0 \leq i \leq n} x_1^i x_k^i & \sum_{0 \leq i \leq n} x_2^i x_k^i & \cdots & \sum_{0 \leq i \leq n} x_m^i x_k^i & \cdots & \sum_{0 \leq i \leq n} x_k^i \\ \vdots & \vdots & & \vdots & & \vdots \\ \sum_{0 \leq i \leq n} x_1^i & \sum_{0 \leq i \leq n} x_2^i & \cdots & \sum_{0 \leq i \leq n} x_m^i & \cdots & n \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \\ \vdots \\ b \end{bmatrix} = \begin{bmatrix} \sum_{0 \leq i \leq n} \phi(\mathbf{x}^i) x_1^i \\ \sum_{0 \leq i \leq n} \phi(\mathbf{x}^i) x_2^i \\ \vdots \\ \sum_{0 \leq i \leq n} \phi(\mathbf{x}^i) x_m^i \\ \vdots \\ \sum_{0 \leq i \leq n} \phi(\mathbf{x}^i) \end{bmatrix} \quad (2)$$

Note that the entries of the matrix in equation 2, are exactly the partial derivatives  $\partial_{\omega_k \omega_k}^2 C(\omega, b), \partial_{\omega_k b}^2 C(\omega, b), \partial_{\omega_k \omega_j}^2 C(\omega, b), \partial_{\omega_k b}^2 C(\omega, b)$  i.e such matrix is the hessian-matrix  $\mathcal{H}_{C(\omega, b)}$ ; Let  $\{v_m\}_{m \in [n+1]}$  be a collection of vectors in  $\mathbb{R}^n$  defined as follows:  $\forall m, 0 \leq m \leq n, v_m := (x_m^1, \dots, x_m^n)$ . For  $m = n+1$  we define:  $v_{n+1} := \mathbf{1} = (1, \dots, 1)$ , and a vector  $\varphi := (\phi(\mathbf{x}^1), \dots, \phi(\mathbf{x}^n))$ . The system of equations can be written as:

$$\begin{bmatrix} v_1 \cdot v_1 & v_1 \cdot v_2 & \cdots & v_1 \cdot v_m & \cdots & v_1 \cdot v_{n+1} \\ v_1 \cdot v_2 & v_2 \cdot v_2 & \cdots & v_2 \cdot v_m & \cdots & v_2 \cdot v_{n+1} \\ \vdots & \vdots & & \vdots & & \vdots \\ v_m \cdot v_1 & v_m \cdot v_2 & \cdots & v_m \cdot v_m & \cdots & v_m \cdot v_{n+1} \\ \vdots & \vdots & & \vdots & & \vdots \\ v_{n+1} \cdot v_1 & v_{n+1} \cdot v_2 & \cdots & v_{n+1} \cdot v_m & \cdots & v_{n+1} \cdot v_{n+1} \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \\ \vdots \\ b \end{bmatrix} = \begin{bmatrix} \varphi \cdot v_1 \\ \varphi \cdot v_2 \\ \vdots \\ \varphi \cdot v_m \\ \vdots \\ \varphi \cdot v_{n+1} \end{bmatrix} \quad (3)$$

This is  $\mathcal{H}_{C(\omega, b)} = G(v_1, \dots, v_{n+1})$  i.e the matrix in system 3 is a Gramm matrix, which by the postively semidefiniteness property of the Gramm matrices implies this matrix is postively defined. Then, the solution  $(\omega^*, b^*)$  to such system is unique and this solution gives indeed a minimum for  $C(\omega, b)$ . Furthermore, the values of the pair  $(\omega^*, b^*)$  are explicitly computable because the system is linear. □

### 3.15.9

First let's analyse the following function defined as a parametric integral. Let  $I(\alpha)$  defined as:

$I(\alpha) := \int_{\sup(X)} p(x)^\alpha dx = \int_{\sup(X)} p(x)^{\alpha-1} dP(x)$ . In last right expression  $\frac{dP}{dx} = p(x)$ . This representation is a consequence of the Radon-Nikodym Theorem. The function  $f(\alpha, x) := p(x)^\alpha$