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 satificies 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 satifies 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 symetric 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 $D_{KL}(p_1\|p_2) = \dfrac{\xi^2}{\xi^1} - \ln\dfrac{\xi^2}{\xi^1} - 1$.

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

Exercise 3.15.8

Asume 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 $(w^\star, b^\star)$, which can be computed explicitly.

Exercise 3.15.9

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

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

Exercise 3.15.10

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

a. Show that $\phi_\sigma * \phi_\sigma = \phi_{\sigma\sqrt{2}}$;

b. 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

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

Show the following:

a. $D_{CS}(p,q) = 0$ if and only if $p = q$;

b. $D_{CS}(p,q) \geq 0$;

c. $D_{CS}(p,q) = D_{CS}(q,p)$;

d. $D_{CS}(p,q) = -\ln \displaystyle\int pq dx - \frac{1}{2}H_2(p) - \frac{1}{2}H_2(q)$, where $H_2(\cdot)$ denotes the quadratic Reyni entropy.

Exercise 3.15.12

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

b. 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

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

Consider the function $\phi : \mathcal{X} \to \mathbb{R}^2$ 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:

$$\begin{aligned} S(p,q_1q_2) &= -\int_{\mathbb{R}} p(x)\ln q_1(x)q_2(x)dx = -\int_{\mathbb{R}} p(x)\ln q_1(x)dx - \int_{\mathbb{R}} p(x)\ln q_2(x) \\ &= S(p,q_1)+S(p,q_2). \end{aligned}$$

□

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:

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

$\square$

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 arbitary 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:

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

$\square$

3.15.5 (a)

Given that $p(x)$ is a distribution, it follows that $p(X)$ as a r.v satisffies the inequality $0 \leq p(X) \leq 1$. This means $p(x) \leq 1, \forall x \in \sup(X)$. 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:

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

$\square$

3.15.6

This is an inmediate consecuence of exercise 3.15.5. Indeed, we have:

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

$\square$

3.15.7

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

3.15.8

In this case $f_{\omega,b}(\mathbf{x}) = \omega \cdot \mathbf{x} + b$, definied 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 $\mathbf{x}^i = (x_1^i, \ldots, x_n^i)$. Then, the normal equations for the ω_k (the componentes of the vector ω), $\forall k \in [n]$ and the bias parameter b are:

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

This system of equations has the following matricial expression:

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

Note that the entries of the matrix in equation 2, are exactly the partial derivatives $\partial^2_{\omega_k\omega_k} C(\omega, b), \partial^2_{bb} C(\omega, b)$, $\partial^2_{\omega_k\omega_j} C(\omega, b), \partial^2_{\omega_k b} C(\omega, b)$ 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: $\forall m, 0 \leq m \leq n, v_m := (x_m^1, \ldots, x_m^n)$. For $m = n+1$ we define: $v_{n+1} := \mathbf{1} = (1, \ldots, 1)$, and a vector $\varphi := (\phi(\mathbf{x}^1), \ldots, \phi(\mathbf{x}^n))$. The system of equations can be writen as:

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

This is $\mathcal{H}_{C(\omega,b)} = \mathrm{G}(v_1, \ldots v_{n+1})$ 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 $(\omega^\star, b^\star)$ to such system is unique and this solution gives indeed a minimum for $C(\omega, b)$. Furthermore, the values of the pair $(\omega^\star, b^\star)$ are explicitly computable because the system is linear. $\square$

3.15.9

Note that the Reyni entropy can be expressed as follows: $\mathrm{H}_\alpha(\mathrm{p}(X)) = \frac{\ln \mathbb{E}\left[(p(X))^{\alpha-1}\right]}{1-\alpha} = \frac{\ln \displaystyle\int_{\sup(X)} (p(x))^{\alpha-1} dP}{\alpha-1}$. In the last expression $\frac{dP}{dx} = p(x)$; This representation is a consecuence of the Radon-Nikodym's Theorem. Now, let's analyse the following function definied as a parametric integral. Let $I(\alpha)$ definied as: $I(\alpha) = \displaystyle\int_{\sup(X)} (p(x))^{\alpha-1} dP$

- CASE $0 < p(x) < 1$:

  Let $\{\alpha_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ a monotonicaly decreasing sequence of reals tending to 1. With this we can construct an increasing sequence of P-integrable functions of the form $\{(p(x))^{\alpha_n-1}\}_{n\in\mathbb{N}}$. Then, by construction the point limit will be the function $f(x) = 1$ which is P-integrable. Applying the Lebegue's monotone convergence theorem:
  $\lim_{k\to\infty} I(\alpha_k) = \displaystyle\lim_{k\to\infty} \int_{\sup(X)} (p(x))^{\alpha_k-1} dP = \int_{\sup(X)} \lim_{k\to\infty} (p(x))^{\alpha_k-1} dP = I(\lim_{k\to\infty} \alpha_k) = \lim_{\alpha\to1^+} I(\alpha) = 1$. This proves $I(\alpha)$ is right-continuous

  Taking $\{\alpha_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ a monotonicaly increasing sequence of reals tending to 1, let's construct the sequence of decreasing P-integrable functions of the form $\{(p(x))^{\alpha_n-1}\}_{n\in\mathbb{N}}$. Once again the point limit function is $f(x) = 1$. Applying the monotone convergence
  $\lim_{k\to\infty} I(\alpha_k) = \displaystyle\lim_{k\to\infty} \int_{\sup(X)} (p(x))^{\alpha_k-1} dP = \int_{\sup(X)} \lim_{k\to\infty} (p(x))^{\alpha_k-1} dP = I(\lim_{k\to\infty} \alpha_k) = \lim_{\alpha\to1^-} I(\alpha) = 1$. This proves $I(\alpha)$ is left-continuous.

- CASE $p(x) > 1$:

  The same is true in this case. It is worth to note that if the sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ is choosen to be decreasing then the sequence of functions is decreasing, the contrary situation also holds.

Because the function $f(\alpha, x) = (p(x))^{\alpha-1}$ is derivable, and its derivative is continuous (this can be proved by the monotone convergence theorem), the above can be then used to compute the limit $\lim_{\alpha\to1} \mathrm{H}_\alpha(\mathrm{p}(X))$. By a simple application $\alpha \to 1$ we find an indetermination of the type $0/0$; Using L'Hôpital's rule we get: $\displaystyle\lim_{\alpha\to1} \mathrm{H}_\alpha(\mathrm{p}(X)) = \lim_{\alpha\to1} \frac{1}{-1} \frac{1}{\displaystyle\int_{\sup(X)} (p(x))^{\alpha-1} dP} \int_{\sup(X)} (p(x))^{\alpha-1} \ln(p(x)) dP = \int_{\sup(X)} -\ln(p(x)) dP = \mathrm{H}(p)$.

□

3.15.10 (a)

This is a consecuence of exercise 3.15.9.*b*. By taking $\sigma_1 = \sigma_2 = \sigma$ we get that $\varphi_{\sigma_1}(x) \star \varphi_{\sigma_2}(x) = \varphi_{\sigma'}(x)$, with $\sigma' = \sqrt{2\sigma^2}$

3.15.10 (b)

We have $\phi_\sigma(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$; Computing $(\phi_{\sigma_1} \star \phi_{\sigma_2})(x) = \int_{\mathbb{R}} \phi_{\sigma_1}(t)\phi_{\sigma_2}(t-x)dt$ we get:

$$(\phi_{\sigma_1} \star \phi_{\sigma_2})(x) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma_1^2\sigma_2^2}\sqrt{2\pi}} e^{\dfrac{-t^2\sigma_2^2 - (t-x)\sigma_1^2}{2\sigma_1^2\sigma_2^2}} dt = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma_1^2\sigma_2^2}\sqrt{2\pi}} e^{-\dfrac{t^2(\sigma_1^2+\sigma_2^2) + 2tx\sigma_1^2 - x^2\sigma_1^2}{2\sigma_1^2\sigma_2^2}} dt$$

$$= \int_{\mathbb{R}} \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sqrt{2\pi\sigma_1^2\sigma_2^2}\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}} e^{-(\sigma_1^2+\sigma_2^2)\dfrac{t^2 - 2tx\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} + \frac{x^2(\sigma_1^2)^2}{(\sigma_1^2+\sigma_2^2)^2} + x^2\sigma_1^2 - \frac{x^2(\sigma_1^2)^2}{(\sigma_1^2+\sigma_2^2)^2}}{2\sigma_1^2\sigma_2^2}} dt$$

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-\dfrac{x^2}{2(\sigma_1^2+\sigma_2^2)}} \frac{1}{\sqrt{2\pi\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}} e^{-\dfrac{(t - x\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2})^2}{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}} dt$$

$$= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{-\dfrac{x^2}{2(\sigma_1^2+\sigma_2^2)}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}} e^{-\dfrac{(t - x\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2})^2}{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}} dt$$

$$= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} e^{\dfrac{-x^2}{2(\sigma_1^2+\sigma_2^2)}} \cdot 1 = \phi_{\sigma'}(x) \text{ , with } (\sigma')^2 = \sigma_1^2 + \sigma_2^2$$

□

3.15.11 (a)

- CASE $p = q$:

  Making the substitution in the formula for the Cauchy-Schwartz- divergence, we find:
  $D_{CS}(p,p) = -\ln\left(\frac{\int p(x)p(x)dx}{\sqrt{\int p(x)^2dx}\sqrt{\int p(x)^2dx}}\right) = -\ln(1) = 0.$

- CASE $D_{CS}(p,q) = 0$:

  If $D_{CS}(p,q) = 0$, then applying exponentials to both sides we obtain:
  $\dfrac{\displaystyle\int p(x)q(x)dx}{\sqrt{\displaystyle\int p(x)^2dx}\sqrt{\displaystyle\int q(x)^2dx}} = 1 \implies \|\displaystyle\int p(x)q(x)dx\| = \sqrt{\displaystyle\int p(x)^2dx}\sqrt{\displaystyle\int q(x)^2dx}$. This is the case of equality in the Cauchy–Bunyakovsky–Schwarz inequality. Then: $q = \lambda p, \lambda \neq 0$; Let's now prove $\lambda = 1$. Plugging this we get:
  $\dfrac{\lambda\displaystyle\int p(x)p(x)dx}{\lambda^2\sqrt{\displaystyle\int p(x)^2dx}\sqrt{\displaystyle\int p(x)^2dx}} = 1 = \dfrac{\lambda}{\lambda^2} \implies \lambda = 1 \implies p = q.$

□

3.15.11 (b)

By hypothesis both p, q are densities, then they are both nonnegative. This implies:

$\int p(x)q(x)dx = \int |\ p(x)q(x)\ | dx \leq \|p\|_2^2 \|q\|_2^2 \implies 0 < \frac{\int p(x)q(x)dx}{\|p\|_2^2 \|q\|_2^2} \leq 1$; Remember that $\ln(x) \leq 0$ if $0 < x < 1$ then, taking logarithms to both sides in the inequality and multiplying by -1 yields:

$$D_{CS}(p, q) = -\ln\left(\frac{\int p(x)q(x)dx}{\sqrt{\int p(x)^2 dx}\sqrt{\int q(x)^2 dx}}\right) \geq 0 \qquad \square$$

3.15.11 (c)

This property follows easily from the definition of the CS-divergence.

3.15.11 (d)

Applying the properties of the logarithms we obtain:

$$\begin{aligned} D_{CS}(p, q) &= -\ln\left(\frac{\int p(x)q(x)dx}{\sqrt{\int p(x)^2 dx}\sqrt{\int q(x)^2 dx}}\right) = -\ln \int p(x)q(x)dx - \frac{1}{2}\ln \int p(x)^2 dx - \frac{1}{2}\int q(x)^2 dx \\ &= -\ln \int p(x)q(x)dx - \frac{1}{2}\mathrm{H}_2(p) - \frac{1}{2}\mathrm{H}_2(q) \end{aligned}$$

$\square$

3.15.12 (a)

The claim follows from the inequality $|u| \geq |\tanh(u)|$, then making the substitution $u = f(x)$, with $f \in \mathcal{L}_1[0, 1]$ we get $|f(x)| \geq |\tanh(f(x))|$. On the other hand, the function tanh is derivable $\forall u \in \mathbb{R}$. Furthermore it's derivative is given by the identity: $\frac{d}{du}\tanh(u) = 1 - \tanh^2(u), \forall u \in \mathbb{R}$. Observe that $\frac{d}{du}\tanh(u) \leq 1, \forall u \in \mathbb{R}$. Applying the Intermediate Value Theorem yields:

- <u>CASE $u > 0$</u>

$$\begin{aligned} 0 < \frac{\tanh(u) - \tanh(0)}{u - 0} = \frac{\tanh(u)}{u} &= \frac{d}{du}\tanh(u)\Big|_{u=\theta}, \text{ where } \theta \in (0, u) \\ &= 1 - \tanh^2(\theta) \leq 1 \implies \tanh(u) \leq u = |u|\ , \forall u > 0 \end{aligned}$$

- <u>CASE $u < 0$</u>

$$\begin{aligned} 0 < \frac{\tanh(u) - \tanh(0)}{u - 0} = \frac{\tanh(u)}{u} &= \frac{d}{du}\tanh(u)\Big|_{u=\theta}, \text{ where } \theta \in (u, 0) \\ &= -1 + \tanh^2(\theta) \leq 1 \implies \tanh(u) \geq u = -|u|\ , \forall u < 0 \end{aligned}$$

Both cases can be summarized as $|u| \geq |\tanh(u)|$. If $u = f(x)$ this implies $|f(x)| \geq |\tanh(f(x))|$. Then, by the monotonicality of the integral operator we get the desired result. $\square$

3.15.12 (b)

In 3.15.12.a was proved the inequality $|u| \geq |\tanh(u)|$, $\forall u \in \mathbb{R}$ so if $f \in \mathcal{L}_2[0,1]$. Then, $|f(x)| \geq |\tanh(f(x))|$; Taking squares the latter inequality yieds: $|f(x)|^2 \geq |\tanh(f(x))|^2$ and once again, due to the monotonicality of the integral operator we get the desired inequality. $\square$

3.15.13

First let's find $\mathbb{E}_p\left[\phi(X)\right]$ and $\mathbb{E}_q\left[\phi(X)\right]$. Computing the expectancies we find:
$\mathbb{E}_p\left[\phi(X)\right] = (0,1)\cdot\frac{1}{2} + (1,0)\cdot\frac{1}{2} = (\frac{1}{2},\frac{1}{2})$ and $\mathbb{E}_q\left[\phi(X)\right] = (0,1)\cdot\frac{1}{3} + (1,0)\cdot\frac{2}{3} = (\frac{2}{3},\frac{1}{3})$. By direct application of the MMD formula we obtain:

$$\begin{aligned}
d^2_{MMD}(p,q) &= \mathbb{E}_p\left[\phi(X)\right]^T\mathbb{E}_p\left[\phi(X)\right] + \mathbb{E}_q\left[\phi(X)\right]^T\mathbb{E}_q\left[\phi(X)\right] - 2\mathbb{E}_p\left[\phi(X)\right]^T\mathbb{E}_q\left[\phi(X)\right] \\
&= (\frac{1}{2},\frac{1}{2})^T(\frac{1}{2},\frac{1}{2}) + (\frac{2}{3},\frac{1}{3})^T(\frac{2}{3},\frac{1}{3}) - 2(\frac{1}{2},\frac{1}{2})^T(\frac{2}{3},\frac{1}{3}) \\
&= \frac{1}{4} + \frac{1}{4} + \frac{4}{9} + \frac{1}{9} - 2\frac{1}{2}\frac{2}{3} - 2\frac{1}{2}\frac{1}{3} = \frac{1}{2} + \frac{5}{9} - 2\frac{3}{6} = \frac{19}{18} - \frac{18}{18} \\
&= \frac{1}{9} \implies d_{MMD}(p,q) = \frac{1}{3}.
\end{aligned}$$