Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

Problem Set 6

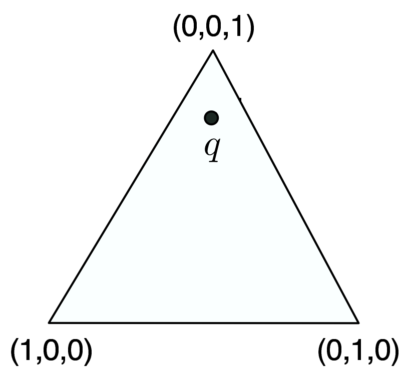
Issued: Tuesday, March 17, 2015 **Due:** Tuesday, April 7, 2015

Problem 6.1

Consider the set of distributions on $\Omega = \{0, 1, 2\}$ and note that they lie on the 2-simplex

$$\{p = (p_0, p_1, p_2) : p_0 + p_1 + p_2 = 1,\ p_0 \geq 0,\ p_1 \geq 0,\ p_2 \geq 0\}$$

represented by the triangular figure. Let y be a random variable such that $p_y(i) = p_i$, $i \in \{0, 1, 2\}$. Let $q = (1/6, 1/6, 2/3)$ be a particular probability mass function.



(a) Draw on the simplex the linear family corresponding to the expectation $\mathbb{E}[y] = 0$, i.e. draw $\mathcal{L}_0 = \{p : \mathbb{E}_p[y] = 0\}$.

(b) Draw $\mathcal{L}_{\frac{1}{2}} = \{p : \mathbb{E}_p[y] = 1/2\}$.

(c) Specify the exponential family $\mathcal{E}$ that passes through q and is orthogonal to $\mathcal{L}_{\frac{1}{2}}$, and draw the entire family on the 2-simplex.

(d) Calculate the I-projection p^* of q onto $\mathcal{L}_{\frac{1}{2}}$ and mark it on the simplex.

(e) Draw $\mathcal{P} = \{p : \mathbb{E}_p[y] \leq 1/2\}$.

(f) Calculate the I-projection p^* of q onto $\mathcal{P}$ and mark it. *Hint: $D(\cdot\|q)$ is convex in its first argument.*

Problem 6.2

Let $q(y) > 0$ $(y = 0, 1, \dots)$ be a probability mass function for a random variable y and let $\mathcal{P}$ be the set of all PMFs defined over $\{0, \dots, M-1\}$ for a known constant M:

$$\mathcal{P} \triangleq \{p(\cdot) | p(y) = 0 \text{ for all } y \geq M\}.$$

We can represent each element p of $\mathcal{P}$ as a M-dimensional vector $[p_0 \dots p_{M-1}]^{\mathrm{T}}$ that lies on a $(M-1)$-dimensional simplex, i.e., $\sum_{m=0}^{M-1} p_m = 1$.

(a) Show that, for all $p \in \mathcal{P}$, $D(q\|p) = \infty$.

(b) Show that, for all $p \in \mathcal{P}$, $D(p\|q) < \infty$.

(c) Find the I-projection of q onto $\mathcal{P}$, $p^* = \arg\min_p D(p\|q)$, and the corresponding divergence $D(p^*\|q)$ in terms of $Q(y) \triangleq \mathbb{P}(\mathsf{y} \leq y)$, the CDF of the random variable y.

Let $\mathcal{P}_\epsilon$ be the space of all PMFs with weight of ϵ on values M and above:

$$\mathcal{P}_\epsilon \triangleq \left\{ p(\cdot) \Big| \sum_{y=M}^{\infty} p(y) = \epsilon \right\}.$$

We can think of $\mathcal{P}_\epsilon$ as an extension of $\mathcal{P}$ to the distributions defined for all integers that only allows limited weight to be allocated to the values outside $\{0, \dots, M-1\}$.

(d) Find the I-projection of q onto $\mathcal{P}_\epsilon$, $p^*_\epsilon = \arg\min_p D(p\|q)$, and the corresponding divergence $D(p^*_\epsilon\|q)$ in terms of $Q(y)$. Show that $\lim_{\epsilon \to 0} D(p^*_\epsilon\|q) = D(p^*\|q)$.

(e) Show that $\mathcal{P}_\epsilon$ can be represented as a linear family of PMFs, i.e.,

$$\mathcal{P}_\epsilon = \{p(\cdot) | \mathbb{E}_p[t(\mathsf{y})] = c\},$$

and invent the appropriate statistic $t(\cdot)$ and constant c.

(f) Show that p^*_ϵ belongs to the exponential family $\mathcal{E} = \mathtt{E}(x; \lambda(x) = x, t(\cdot), \ln q(\cdot))$ and find the value of the parameter x that corresponds to p^*_ϵ.

Problem 6.3

Let $\mathcal{P}$ be the space of all distributions defined over alphabet $\mathcal{Y}$, $\{t_1(\cdot), \dots, t_{K+1}(\cdot)\}$ be $K+1$ functions defined over alphabet $\mathcal{Y}$, and $\{\bar{t}_1, \dots, \bar{t}_{K+1}\}$ be $K+1$ known constants.

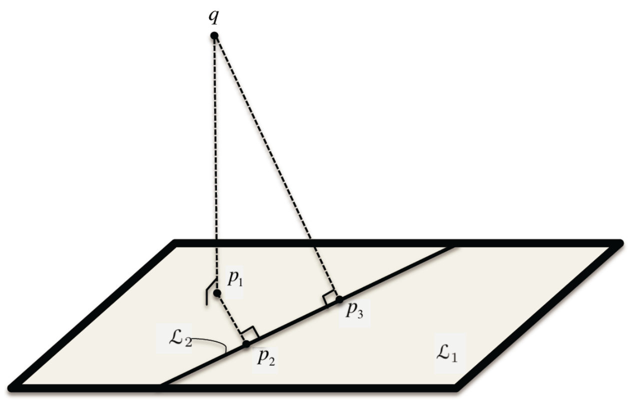
We define $\mathcal{L}_1$ to be a linear family of distributions characterized by $\{t_1(\cdot), \dots, t_K(\cdot)\}$ and $\{\bar{t}_1, \dots, \bar{t}_K\}$, i.e.,

$$\mathcal{L}_1 = \{p \in \mathcal{P} \,|\, \mathbb{E}_p[t_k(\mathsf{y})] = \bar{t}_k, \quad k = 1, \dots, K\},$$

and $\mathcal{L}_2 \subseteq \mathcal{L}_1$ to be a linear family of distributions characterized by $\{t_1(\cdot), \ldots, t_{K+1}(\cdot)\}$ and $\{\bar{t}_1, \ldots, \bar{t}_{K+1}\}$, i.e.,

$$\mathcal{L}_2 = \{p \in \mathcal{P} \,|\, \mathbb{E}_p [t_k(\mathsf{y})] = \bar{t}_k, \quad k = 1, \ldots, K+1\}.$$

Let $q \in \mathcal{P}$ be an arbitrary distribution defined over alphabet $\mathcal{Y}$, such that $q(y) > 0$ for all $y \in \mathcal{Y}$, p_1 be the I-projection of q on $\mathcal{L}_1$, p_2 be the I-projection of p_1 on $\mathcal{L}_2$, and p_3 be the I-projection of q on $\mathcal{L}_2$, as illustrated in the figure below:



Assume that $|\mathcal{Y}| >> K$ and that for all $y \in \mathcal{Y}$, there exists a distribution $p \in \mathcal{L}_2$, such that $p(y) > 0$. The latter guarantees that for all $y \in \mathcal{Y}$, $p_1(y) > 0$, $p_2(y) > 0$ and $p_3(y) > 0$, i.e., distributions p_1, p_2, and p_3 lie in the interior of the corresponding linear families.

The goal of this problem is to explore the relationship between p_2 and p_3. Two approaches are considered in part (a) and part (b), respectively. Please complete both parts in this problem.

(a) Approach I:

(i) Determine constants $\alpha, \beta, \gamma, \delta$ such that

$$D(p_2 \| p_3) = \alpha D(p_2 \| q) + \beta D(p_3 \| q),$$
$$D(p_3 \| p_2) = \gamma D(p_2 \| p_1) + \delta D(p_3 \| p_1).$$

(ii) Determine which of the two statements below is true and explain.

$$A : D(p_2 \| p_3) + D(p_3 \| p_2) > 0 \qquad B : D(p_2 \| p_3) + D(p_3 \| p_2) = 0$$

(iii) Determine which of the two statements below is true and explain.

A : p_2 is identical to p_3 $\qquad$ B : p_2 is not identical to p_3

(b) Approach II:

(i) Let

$$\mathcal{E}_1 = \mathtt{E}\left(\boldsymbol{\lambda}; \boldsymbol{\lambda}, \mathbf{u}(\cdot), \beta_1(\cdot)\right)$$

be the K-parameter exponential family that contains all distributions in $\mathcal{P}$ whose I-projection on $\mathcal{L}_1$ is identical to p_1. Determine the natural statistics $\{u_1(\cdot), \ldots, u_K(\cdot)\}$ and the log base distribution $\beta_1(\cdot)$ in terms of the distribution $p_1(\cdot)$ and functions $\{t_1(\cdot), \ldots, t_K(\cdot)\}$.

(ii) Similarly, let

$$\mathcal{E}_2 = \mathtt{E}\left(\boldsymbol{\eta}; \boldsymbol{\eta}, \mathbf{v}(\cdot), \beta_2(\cdot)\right)$$

be the $(K+1)$-parameter exponential family that contains all distributions in $\mathcal{P}$ whose I-projection on $\mathcal{L}_2$ is identical to the I-projection of p_1 on $\mathcal{L}_2$. Determine the natural statistics $\{v_1(\cdot), \ldots, v_{K+1}(\cdot)\}$ and the log base distribution $\beta_2(\cdot)$ in terms of the distribution $p_1(\cdot)$ and functions $\{t_1(\cdot), \ldots, t_{K+1}(\cdot)\}$.

(iii) Determine which of the two statements below is true and explain.

$$A : q \in \mathcal{E}_2 \qquad\qquad B : q \notin \mathcal{E}_2$$

(iv) Determine which of the two statements below is true and explain, without referring to your answers in part (a).

$$A : p_2 \text{ is identical to } p_3 \qquad\qquad B : p_2 \text{ is not identical to } p_3$$

Problem 6.4

A binary random variable x is observed through a symmetric error measurement mechanism, i.e., the measurement y is equal to x with probability $1-\epsilon$, and is equal to $1-\mathsf{x}$ with probability ϵ:

$$p_{\mathsf{y}|\mathsf{x}}(y \mid x) = \begin{cases} 1-\epsilon & y = x \\ \epsilon & y \neq x \end{cases}.$$

Let q be the unknown probability that $\mathsf{x} = 1$. Assume that $\epsilon < 1/2$.

(a) Show that the mutual information for the symmetric error measurement mechanism is

$$I_{\mathrm{S}} = I(\mathsf{x}; \mathsf{y}) = H_{\mathrm{B}}(q(1-\epsilon) + (1-q)\epsilon) - H_{\mathrm{B}}(\epsilon).$$

Recall that we defined $H_{\mathrm{B}}(p) = -p \log p - (1-p)\log(1-p)$.

(b) Determine the model capacity and the least informative prior for x.

(c) Suppose we have an alternative (erasure) measurement mechanism that produces an observation z. The observation z can take three possible values: 0, 1, and "NA"; "NA" means z is "erased" and no observation is available. Given $\mathsf{x} = x$, the observation z either is equal to x or gets erased and becomes "NA". The probabilities are $\mathbb{P}(\mathsf{z} = x) = (1 - 2\epsilon)$ and $\mathbb{P}(\mathsf{z} = \mathrm{NA}) = 2\epsilon$. It can be shown that the mutual information between x and z is

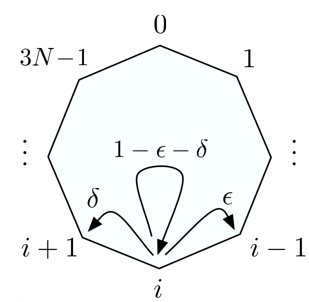
$$I_{\mathrm{E}} = I(\mathsf{x}; \mathsf{z}) = H_{\mathrm{B}}(q)(1 - 2\epsilon).$$

If your goal is to do the best inference possible (i.e., minimize the expected log-loss cost function), when would you prefer measurements from the original measurement mechanism over those from this new mechanism? How does your answer depend on the value of q? You can use the above result on $I(\mathsf{x}; \mathsf{z})$ directly without derivation.

Problem 6.5

Let x be a discrete random variable with prior $p_{\mathsf{x}}(\cdot)$ that takes values in the set $\{0, 1, \cdots, 3N - 1\}$. Let y be the noisy observation of x, defined as

$$p_{\mathsf{y}|\mathsf{x}}(y|x) = \begin{cases} \epsilon, & y = (x - 1) \bmod 3N \\ 1 - \epsilon - \delta, & y = x \\ \delta, & y = (x + 1) \bmod 3N, \end{cases}$$



where ϵ and δ are known positive constants such that $\epsilon + \delta \leq 1$ and the mod $3N$ operation wraps the values around the range:

$$z \bmod 3N = \begin{cases} 3N - 1, & z = -1 \\ z, & z = 0, \ldots, 3N - 1 \\ 0, & z = 3N. \end{cases}$$

(a) Determine $H(\mathsf{y}|\mathsf{x})$ as a function of prior $p_{\mathsf{x}}(\cdot)$, ϵ and δ.

(b) Determine the model capacity C and the least informative prior in terms of ϵ and δ.

(c) For this question, assume $0 < \epsilon = \delta < 1/2$. Our goal is to construct a prior that enables perfect estimation, i.e., it allows us to correctly identify x from the observed y with no chance of error. Find the maximal mutual information between variables x and y achievable under this constraint and the corresponding prior on x that achieves it.

Problem 6.6

Consider a discrete random value x taking on values in

$$\mathcal{X} = \left\{0, \frac{1}{M}, \frac{2}{M}, \ldots, \frac{M-1}{M}, 1\right\}$$

where M is a fixed integer. Further, let us assume that y is a Bernoulli random variable that takes on the values 0 and 1, such that

$$p_{\mathsf{y}|\mathsf{x}}(y|x) = \begin{cases} x & y = 1 \\ 1-x & y = 0 \end{cases}.$$

In this problem we will examine the model capacity C as a function of M. While hard to compute in closed form, both the model capacity and associated least informative prior can be calculated numerically using a algorithm known as the *Arimoto-Blahut* algorithm[1]:

1: Initialize $\hat{p}_{\mathsf{x}}^{(0)}(x)$ to a strictly positive probability distribution, and set $n = 1$.

2: Compute

$$c^{(n)}(x) \quad = \quad \exp\left(\sum_y p_{\mathsf{y}|\mathsf{x}}(y \mid x) \log \frac{p_{\mathsf{y}|\mathsf{x}}(y \mid x)}{\sum_{x'} \hat{p}_{\mathsf{x}}^{(n-1)}(x') p_{\mathsf{y}|\mathsf{x}}(y \mid x')}\right).$$

3: Update

$$\hat{p}_{\mathsf{x}}^{(n)}(x) = \hat{p}_{\mathsf{x}}^{(n-1)}(x) \frac{c^{(n)}(x)}{\sum_{x'} \hat{p}_{\mathsf{x}}^{(n-1)}(x') c^{(n)}(x')}.$$

4: Compare

$$\begin{aligned} I_L &= \log\left(\sum_x \hat{p}_{\mathsf{x}}^{(n)}(x) c^{(n)}(x)\right), \\ I_U &= \log\left(\max_x c^{(n)}(x)\right). \end{aligned}$$

If I_L and I_U are not close enough, increment n and go to step 2. Otherwise, set $\hat{C} = I_L$ and exit.

(a) Plot the model capacity C as a function of M, for $M = 1, 2, 3, \ldots$. Do you need to use Arimoto-Blahut?

[1] For complete details, see *Computation of Channel Capacity and Rate-Distortion Functions* on the course web site.

(b) Plot the entropy of the least informative prior as a function of M. This measures how close the prior is to a uniform distribution, in a relative entropy sense.

Now let us fix $M = 5$ in the alphabet for x, and consider N observations y_i, conditionally independent of x and each distributed according to

$$p_{\mathsf{y}_i|\mathsf{x}}(y_i|x) = \begin{cases} x & y_i = 1 \\ 1 - x & y_i = 0 \end{cases}.$$

Use the Arimoto-Blahut algorithm to compute the model capacity C_N and least informative prior as a function of N. Please notice that the observation in the second step of the algorithm should be the N-dimensional vector $\mathbf{y} = [y_1, \cdots, y_N]^{\mathrm{T}}$.

(c) Plot the normalized model capacity C_N/N as a function of N for $N = 1, 2, 3, \ldots$.

(d) Plot the entropy of the least informative prior as a function of N.

Problem 6.7 (practice)

Suppose the only way to get information about the true value of a particular binary random variable x is to ask an oracle. The oracle never lies, but with probability $1 - \delta$ gives no answer at all. Formally, the oracle's answer y is distributed according to,

$$p_{\mathsf{y}|\mathsf{x}}(y \mid x) = \begin{cases} \delta & y = x \\ 1 - \delta & y = \mathrm{NA} \end{cases}$$

where "NA" indicates the oracle did not answer, and $\delta \in [0, 1]$ is a known constant. Let q be the unknown probability that $\mathsf{x} = 1$. Assume all logs are base 2 in this problem.

(a) Show that the mutual information between the oracle's answer y and x is proportional to the entropy of the hidden random variable x, i.e.,

$$I(\mathsf{x}; \mathsf{y}) = \gamma(\delta) H_{\mathrm{B}}(q),$$

and find the proportionality coefficient $\gamma(\delta)$.

Recall that we defined $H_{\mathrm{B}}(p) = -p \log p - (1 - p) \log(1 - p)$.

(b) Determine the model capacity C and the least informative prior for x.

(c) To improve our chances of getting an answer, we ask the oracle the same questions n times. The oracle decides to answer or not on each trial independently of all other trials. Assume that the oracle's answers $\mathsf{y}_1, \ldots, \mathsf{y}_n$ are conditionally independent given x and are distributed according to the likelihood above.

(i) Find the mutual information for this observation model,

$$I(x; \mathbf{y}) = I(x; y_1, \ldots, y_n)$$

(ii) Find the model capacity C_n and the least informative prior.

(iii) Find $\lim_{n\to\infty}(C_n/n)$.

Problem 6.8 (practice)
A discrete random variable x takes values in the alphabet $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ where the subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ are *disjoint*, i.e., $\mathcal{X}_1 \cap \mathcal{X}_2 = \varnothing$. We observe x through a measurement process with observation y distributed according to

$$p_{y|x}(y|x) = \begin{cases} p^{(1)}_{y|x}(y|x) & x \in \mathcal{X}_1,\ y \in \mathcal{Y}_1 \\ p^{(2)}_{y|x}(y|x) & x \in \mathcal{X}_2,\ y \in \mathcal{Y}_2 \\ 0 & \text{otherwise,} \end{cases}$$

where the sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are also *disjoint*. In other words, the measurement process consists of two component measurement processes

$$p^{(1)}_{y|x}(\cdot|\cdot) \text{ for } x \in \mathcal{X}_1,\ y \in \mathcal{Y}_1 \qquad \text{and} \qquad p^{(2)}_{y|x}(\cdot|\cdot) \text{ for } x \in \mathcal{X}_2,\ y \in \mathcal{Y}_2.$$

The model capacities for the two component measurement processes are C_1 and C_2, respectively, and the least informative priors are $p_1^*(\cdot)$ and $p_2^*(\cdot)$, respectively. Our goal is to find the model capacity for the joint measurement process and the corresponding least informative prior.

(a) Show that any prior for x can be written in the form

$$p_x(x) = \begin{cases} \lambda p_1(x) & x \in \mathcal{X}_1 \\ (1-\lambda) p_2(x) & x \in \mathcal{X}_2 \end{cases} \tag{1}$$

for some $0 \le \lambda \le 1$, where $p_1(x)$ is some prior distribution on the elements of the set $\mathcal{X}_1$ and $p_2(x)$ is some prior on the set $\mathcal{X}_2$. Express each of λ, $p_1(\cdot)$, and $p_2(\cdot)$ in terms of the distribution $p_x(\cdot)$.

(b) By using the expression for a prior given in (1), the mutual information $I(x; y)$ for the joint measurement process can be written in the form

$$I(x; y) = f_1(\lambda) I_1 + f_2(\lambda) I_2 + g(\lambda),$$

where I_1 and I_2 are the corresponding mutual information for the two component measurement processes with priors $p_1(\cdot)$ and $p_2(\cdot)$, respectively. Determine $f_1(\cdot)$, $f_2(\cdot)$, and $g(\cdot)$.

(c) Determine the model capacity C and least informative prior $p^*(\cdot)$ for the joint measurement process. In your answer, describe $p^*(\cdot)$ by specifying the corresponding $p_1(\cdot)$, $p_2(\cdot)$ and λ when $p^*(\cdot)$ is expressed in the form (1). Express your answers in terms of the component capacities and least informative priors, i.e., in terms of C_1, C_2, $p_1^*(\cdot)$, and $p_2^*(\cdot)$.

Problem 6.9 (practice)
Let x be a ternary variable ($x \in \{0, 1, 2\}$) that parameterizes a likelihood family for a binary random variable y as follows:

$$\begin{aligned} p_{\mathsf{y}}(\cdot; 0) &= \epsilon^y (1-\epsilon)^{1-y}, \\ p_{\mathsf{y}}(\cdot; 1) &= (1-\epsilon)^y \epsilon^{1-y}, \\ p_{\mathsf{y}}(\cdot; 2) &= 1/2, \qquad \text{for } y = 0, 1, \end{aligned}$$

where ϵ is a known constant ($0 < \epsilon < 1/2$).

We use $I_p(\mathsf{x}; \mathsf{y})$ to denote mutual information between random variables x and y when the random variable x is distributed according to distribution $p(\cdot)$ and the likelihood of y given x is defined as $p_{\mathsf{y}}(\cdot; x)$. Using this notation, the model capacity can be expressed as $C = I_{p_{\mathsf{x}}^*}(\mathsf{x}; \mathsf{y})$, where p_{x}^* is the least informative prior.

(a) (i) Let $p_x(\cdot)$ be the weights associated with the variable x. Determine functions $f(\cdot)$ and $g(\cdot)$ such that

$$I_{p_x}(\mathsf{x}; \mathsf{y}) = H_{\mathrm{B}}\left(\frac{1}{2} + f(\epsilon)\left(p_x(1) - p_x(0)\right)\right) - g(\epsilon) p_x(2) - H_{\mathrm{B}}(\epsilon),$$

where $H_{\mathrm{B}}(\epsilon)$ is the entropy of a Bernoulli distribution with parameter ϵ.

(ii) Determine the least informative prior $p_{\mathsf{x}}^*(\cdot)$ for the likelihood model $p_{\mathsf{y}}(\cdot; x)$ and find constants α and β such that the model capacity can be expressed as

$$C = \alpha H_{\mathrm{B}}(1/2) + \beta H_{\mathrm{B}}(\epsilon).$$

(iii) Determine the mixture distribution $p_{\mathsf{y}}^*(\cdot)$ that uses the least informative prior $p_{\mathsf{x}}^*(\cdot)$ as weights.

(b) In this part, we use uniform weights (i.e., $p_{\mathsf{x}}(x) = 1/3$ for $x = 0, 1, 2$) to form the mixture distribution.

(i) Determine the resulting mixture distribution $p_{\mathsf{y}}(\cdot)$.

(ii) Determine constants γ and δ such that the reduction in mutual information due to this sub-optimal choice of prior can be expressed as

$$C - I_{p_{\mathsf{x}}}(\mathsf{x}; \mathsf{y}) = \gamma H_{\mathrm{B}}(1/2) + \delta H_{\mathrm{B}}(\epsilon).$$

In the remainder of the problem, we consider a general likelihood model $p_{\mathsf{y}}(\cdot; x)$ for random variable $\mathsf{y} \in \mathcal{Y}$ parameterized by parameter $x \in \mathcal{X}$. Let $p_{\mathsf{x}}^*(\cdot)$ be the least informative prior and $p_{\mathsf{y}}^*(\cdot)$ be the corresponding marginal distribution of the random variable y.

Let $q_{\mathsf{x}}(\cdot)$ be an arbitrary distribution defined over alphabet $\mathcal{X}$.

(c) Suppose $p_{\mathsf{x}}^*(x) > 0$ for all $x \in \mathcal{X}$. Determine distributions $q_1(\cdot)$ and $q_2(\cdot)$ defined over alphabet $\mathcal{Y}$ such that

$$C - I_{q_{\mathsf{x}}}(\mathsf{x}; \mathsf{y}) = D(q_1(\cdot) \| q_2(\cdot)).$$

Hint: Recall that the *Equidistance Property* implies that for a model $p_{\mathsf{y}|\mathsf{x}}(\cdot|\cdot)$, $D(p_{\mathsf{y}|\mathsf{x}}(\cdot|x) \| p_{\mathsf{y}}^*(\cdot)) = C$ for all $x \in \mathcal{X}$ such that $p_{\mathsf{x}}^*(x) > 0$.

(d) Let $q_1(\cdot)$ and $q_2(\cdot)$ be the distributions you determined in the previous part. Now suppose there exists $x \in \mathcal{X}$ such that $p_{\mathsf{x}}^*(x) = 0$. Determine which of the three statements below is true and explain.

$$\begin{aligned} A: &\quad C - I_{q_{\mathsf{x}}}(\mathsf{x}; \mathsf{y}) \le D(q_1(\cdot) \| q_2(\cdot)) \\ B: &\quad C - I_{q_{\mathsf{x}}}(\mathsf{x}; \mathsf{y}) \ge D(q_1(\cdot) \| q_2(\cdot)) \\ C: &\quad \text{None of the above is always true.} \end{aligned}$$