Introduction to Information Theory

Lecture 2

Lecturer: Haim Permuter Scribe: Morag Agmon, Tirza Routtenberg and Dor Tsur

I. Convexity

Definition 1 (Convex set) A set is *convex* if for every pair of points within the set, the whole straight line segment that joins them is also within the set.

In other words, let $\mathcal{A}$ be a set in a real or complex vector space. The set $\mathcal{A}$ is said to be convex if, for all $x_1 \in \mathcal{A}$ and $x_2 \in \mathcal{A}$ and for all λ in the interval $[0,1]$, the point $x_3 = \lambda x_1 + \bar{\lambda} x_2$ is in $\mathcal{A}$ (i.e., $x_3 \in \mathcal{A}$), where $\bar{\lambda} = 1 - \lambda$.

Example 1 (Convex sets) Examine the sets illustrated in Fig. 2. Part (a) illustrates a convex set while Part (b) illustrates a non-convex set.

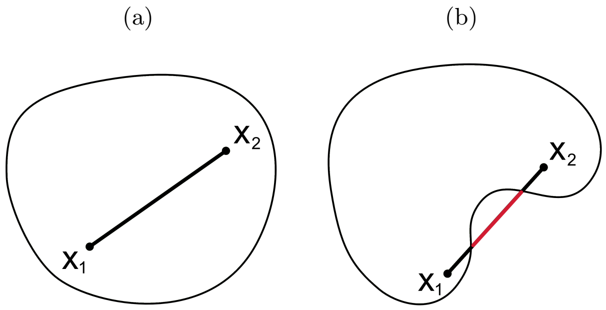

Fig. 1. (a) A convex set (b) a non-convex set

Example 2 (Convexity of a probability vector space) Show that the probability vector space is a convex set.
Answer: Consider a random variable X with alphabet $\mathcal{X} = 1, ..., k$. The probability vector space $P_X = [P_X(1), P_X(2), ..., P_X(k)] \in \mathbb{R}^k$, is the set of all vectors for which $P_X(i) \geq 0 \quad \forall i \in \mathcal{X}$, and $\sum_{i=1}^{K} P_X(i) = 1$. Now consider two probability vectors $P_X^{(1)}$ and $P_X^{(2)}$, and the vector

$$P_X^{(3)} = \lambda P_X^{(1)} + \bar{\lambda} P_X^{(2)}. \tag{1}$$

We need to show that $P_X^{(3)}$ is a probability vector. Since, $\sum_{i=1}^{K} P_X^{(3)}(i) = 1$ and $P_X^{(3)}(i) \geq 0 \quad \forall i \in \mathcal{X}$, indeed $P_X^{(3)}$ is a probability vector. Thus, the probability vector space is a convex set.

Definition 2 (Convex function.) Let $f(x)$ be a function of the form $f : \mathbb{R}^n \mapsto \mathbb{R}$, where $\mathbb{R}$ is the set of real numbers and $\mathbb{R}^n$ is an n dimensional real vector, hence $x \in \mathbb{R}^n$. A function $f(x)$ is a *convex function* if

$$f\left(\lambda x_1 + \bar{\lambda} x_2\right) \leq \lambda f(x_1) + \bar{\lambda} f(x_2) \tag{2}$$

for all x_1, x_2 in its domain, and for all $\lambda \in [0,1]$. A function is a *strictly convex function* if

$$f\left(\lambda x_1 + \bar{\lambda} x_2\right) < \lambda f(x_1) + \bar{\lambda} f(x_2) \tag{3}$$

for all x_1, x_2 in its domain, and for all $\lambda \in (0,1)$.

Definition 3 (Concave function) A function $f(x)$ is said to be *(strictly) concave function* if and only if $(-f(x))$ is (strictly) convex.

Example 3 (Convex/Concave functions) Examples of convex functions include x^2, $|x|$, e^x, and so on. Examples of concave functions include $\log x$ and $\sqrt{x}$ for $x > 0$. Note that linear functions $ax + b$ are both convex and concave. Figure 2 shows some examples of convex and concave functions.

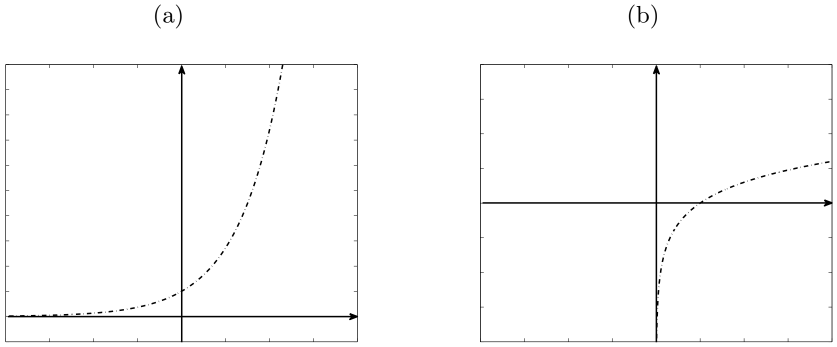

Fig. 2. (a) A convex function – e^x (b) a concave function – $\log x$

Lemma 1 (operations that preserve convexity) 1. *addition of functions* Let f_1 and f_2 be two convex functions, then $f_1 + f_2$ is also a convex function.
2. *matrix multiplication of the argument* Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $A \in \mathbb{R}^{n \times n}$ a matrix of dimension $n \times n$. Then $f(Ax)$ is also convex.

Exercise 1 Prove Lemma 1 using the definition of convex functions.

Lemma 2 (Second derivative test for scalar functions) Given a scalar function $f(x)$, i.e., $f : \mathbb{R} \mapsto \mathbb{R}$, where the second derivative exists we have

$$\forall x \in (a,b), \quad \frac{d^2 f(x)}{dx^2} \geq 0 \Longleftrightarrow f(x) \text{ is convex,} \tag{4}$$

and similarly

$$\forall x \in (a,b), \quad \frac{d^2 f(x)}{dx^2} \leq 0 \Longleftrightarrow f(x) \text{ is concave.} \tag{5}$$

In case we have strict inequality, then we have strict convexity, i.e.,

$$\forall x \in (a,b), \quad \frac{d^2 f(x)}{dx^2} > 0 \Longleftrightarrow f(x) \text{ is stricktly convex.} \tag{6}$$

and similarly for concave function.

Proof: First, we are proving one direction for the convexity (4), namely that if the derivative is non-negative it implies convexity. Then, proving the same direction for concavity in Eq. (5), namely, if the second derivative is non-positive it implies concavity, is similar. Finally, note that both direction complement each other to if and only if. In particular for (4) if the derivative is negative for some (arbitrary) small interval it implies that its not convex hence this complete the second direction.

We use the Taylor series expansion of the function around x_0:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x^*)\frac{(x^* - x_0)^2}{2}, \tag{7}$$

where x^* lies between x_0 and x. By hypothesis, $f''(x^*) \geq 0$, and thus the last term in (7) is nonnegative for all x, i.e.,

$$f(x) \geq f(x_0) + f'(x_0)(x - x_0). \tag{8}$$

We let $x_0 = \lambda x_1 + \bar{\lambda} x_2$, and take $x = x_1$, to obtain

$$\begin{aligned} f(x_1) &\geq f(x_0) + f'(x_0)\left(x_1 - \lambda x_1 - \bar{\lambda} x_2\right) \\ &= f(x_0) + \bar{\lambda} f'(x_0)(x_1 - x_2). \end{aligned} \tag{9}$$

Similarly, taking $x = x_2$, we obtain

$$f(x_2) \geq f(x_0) + \lambda f'(x_0)(x_2 - x_1). \tag{10}$$

Multiplying (9) by λ and (10) by $\bar{\lambda}$ and adding, we obtain (2). The proof for strict convexity proceeds along the same lines.

■

Example 4 (Showing convexity using second derivative test) Consider the function $f(x) = x \log x \quad x > 0$. Then,

$$\begin{aligned} f'(x) &= \log x + 1 \\ f''(x) &= \frac{1}{x} > 0. \end{aligned} \tag{11}$$

Thus, $f(x)$ is a convex function.

Lemma 2 on second derivative of scalar function can be extended to multivariate function (function of more than one variable, such as $f(x_1, x_2)$.

Lemma 3 (Condition for convexity of multivariate function) A function of several variables (multivariate function) is convex if and only if its Hessian matrix is positive semidefinite. For example, consider $\mathbb{R}^2$ space, then the Hessian matrix can be written as

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} \geq 0 \tag{12}$$

Recall, that a square matrix A of dimension $n \times n$ is *semidefinite* (i.e., $A \geq 0$) if for any vector $\mathbf{x}$ of length n, $\mathbf{x}^T A \mathbf{x} \geq 0$, where $\mathbf{x}^T$ is the transpose of $\mathbf{x}$.

There are several ways to check if a matrix is positive semidefinite such as
1. if and only if all the eigenvalues are non-negative
2. for a symmetric matrix (note that the Hessian matrix is symmetric), if and only if all the leading principal minors are non-negative.

Let's explain in more details the leading principal minors test. Let A be an $n \times n$ symmetric matrix, then it is positive-definite if and only if all the following matrices have a positive determinant:

- the upper left 1-by-1 corner of A,
- the upper left 2-by-2 corner of A,
- the upper left 3-by-3 corner of A,

$\vdots$

- A itself.

II. Jensen's inequality and its consequences

Theorem 1 (Jensen's inequality) Let X be a random variable and $f(x)$ a convex function. Then

$$\mathbb{E}\left[f(X)\right] \geq f\left(\mathbb{E}\left[X\right]\right), \tag{13}$$

and the inequality is reversed if $f(x)$ is concave. Moreover, if f is strictly convex, equality in (13) implies that X is deterministic, i.e., $X = E[X]$ with probability 1.

proof Assume for simplicity that the random variable gets two values in probabilities p_1 and p_2 (where $p_1 + p_2 = 1$). Then, using the convex function definition (2):

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) \tag{14}$$

■

Exercise 2 Prove (generalize) Jensen's inequality for arbitrary number n of probabilities p_n by induction. By convexity definition, the statement is true for $n = 2$. Suppose it is true also for some n, then prove it for $n + 1$.

Exercise 3 Prove that for a strictly convex function f, $\mathbb{E}\left[f(X)\right] = f\left(\mathbb{E}\left[X\right]\right)$ implies that X is deterministic. (Hint: assume that X is not deterministic - and show a contradiction).

The proof of Jensen's inequality presented above holds for any discrete random variable with finite alphabet. However, Jensen's inequality holds for any random variable, not necessarily with finite alphabet, such as continues random variable or a mixture of discrete and continuous random variable. An alternative proof is presented in the appendix (at the end of this lecture) and it does not assume that the random variable is discrete. However, it does use some convex analysis tools.

Theorem 2 (Non-negativity of $\mathbf{D}\left(\mathbf{P}||\mathbf{Q}\right)$) Let $P(x)$ and $Q(x)$ be two probability functions. Then

$$D\left(P||Q\right) \geq 0 \tag{15}$$

with equality if and only if $P\left(x\right) = Q\left(x\right)$ for all x.

Proof:

$$\begin{aligned} -D\left(P||Q\right) &= -\sum_x P\left(x\right)\log\frac{P\left(x\right)}{Q\left(x\right)} \\ &= \mathbb{E}_P\left[\log\frac{Q\left(x\right)}{P\left(x\right)}\right] \\ &\overset{(a)}{\leq} \log\left(\mathbb{E}_P\left[\frac{Q\left(x\right)}{P\left(x\right)}\right]\right) \\ &= \log\left(\sum_x P\left(x\right)\frac{Q\left(x\right)}{P\left(x\right)}\right) \\ &= 0 \end{aligned} \tag{16}$$

where step (a) follows from Jensen's inequality and the fact that log is a concave function. Thus,

$$D\left(P||Q\right) \geq 0 \tag{17}$$

■

Exercise 4 Prove that if $D\left(P||Q\right) = 0$, then $P = Q$. Hint: $\log\left(x\right)$ is a strictly concave function.

Corollary 1 (Non-negativity of mutual information.) For any two random variables, X and Y,

$$I\left(X;Y\right) \geq 0 \tag{18}$$

Proof: We saw that $I\left(X;Y\right) = D\left(P_{XY}||P_XP_Y\right)$, and since $D\left(P_{XY}||P_XP_Y\right) \geq 0$ with equality if and only if $X \perp\!\!\!\perp Y$ it follows that $I(X;Y) \geq 0$ and is equal to zero if and only if $P(x,y) = P(x)P(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

■

Exercise 5 Prove Corollary 1 using Theorem 2. (The proof is provided above, but as an exercise prove it without looking at the proof).

Corollary 2 (Upper bound on Entropy) Let X be a random variable with alphabet $\mathcal{X}$. Then,

$$H\left(X\right) \leq \log|\mathcal{X}| \tag{19}$$

with equality if and only if X has a uniform distribution.

proof Let $U(x) = \frac{1}{|\mathcal{X}|}$ be the uniform probability function over X, and let $P(x)$ be the probability function for X. Then

$$0 \overset{(a)}{\leq} D(P||U) = \sum P(x) \log \frac{P(x)}{U(x)} = \log |\mathcal{X}| - H(X), \tag{20}$$

where inequality (a) follows from Theorem 2. Thus,

$$H(X) \leq \log |\mathcal{X}| \tag{21}$$

■

Theorem 3 (Conditioning reduces entropy)

$$H(X|Y) \leq H(X) \tag{22}$$

with equality if and only if $X \perp\!\!\!\perp Y$.

Intuitively, the theorem says that knowing another random variable Y can only reduces the uncertainty in X. Note that this is true only on the average.

Exercise 6 Prove Theorem 3.

III. Convexity of the Divergence function

In this section we prove the convexity property of the divergence function, which will later help us prove the concavity of the Entropy function and some concave/convex property of the mutual information. We will introduce two different ways to prove the following theorem:

Theorem 4 (Convexity of divergence) The function $D(P||Q)$ is convex in the pair (P,Q); i.e. if (P_1,Q_1) and (P_2,Q_2) are two pairs of probability mass functions (PMF), then

$$D(\lambda P_1 + (1-\lambda)P_2 || \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D(P_1||Q_1) + (1-\lambda) D(P_2||Q_2) \tag{23}$$

for all $\lambda \in [0,1]$.

A. The Log-Sum Inequality

The first way to prove this theorem uses a simple consequence of the concavity of the logarithm function (the proof appears in [1, Ch 2.7])

Theorem 5 (Log sum inequality) For $a_i \geq 0, b_i \geq 0, \quad i = 1, \ldots, n$:

$$\sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{n} a_i\right) \log\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right) \tag{24}$$

with equality *iff* $\frac{a_i}{b_i} = const$.

proof
- For $\sum_i a_i = 0$ or $\sum_i b_i = 0$ the proof is trivial (recall that $0 \log 0 = 0, 0 \log \frac{0}{0} = 0$).
- Let assume that $\sum_i a_i > 0$ and $\sum_i b_i > 0$. The function $f(x) = x \log(x)$ is strictly convex for all $x > 0$ (see Example 4). Using Jensen's inequality:

$$\sum_{i=1}^{n} \alpha_i f(x_i) \geq f\left(\sum_{i=1}^{n} \alpha_i x_i\right) \tag{25}$$

or

$$\sum_{i=1}^{n} \alpha_i x_i \log(x_i) \geq \left(\sum_{i=1}^{n} \alpha_i x_i\right) \log\left(\sum_{i=1}^{n} \alpha_i x_i\right) \tag{26}$$

for all $\alpha_i \geq 0, \quad i = 1, \ldots, n$ and $\sum_{i=1}^{n} \alpha_i = 1$.
Setting

$$\alpha_i = \frac{bi}{\sum_{i=1}^{n} b_i}$$

and

$$x_i = \frac{a_i}{b_i},$$

one obtains the log sum inequality (note that under our assumption it can be seen that $\alpha_i, x_i \geq 0$ and that $\sum_{i=1}^{n} \alpha_i = 1$)).

■

Exercise 7 Using the log-sum inequality (Theorem 5) show that for any two vector-probabilities P and Q the divergence is non negative, i.e., $D(P||Q) \geq 0$ and is zero if and only if $P = Q$.

We now prove the convexity of divergence, i.e., Theorem 4, using the Log-Sum Inequality, i.e., Theorem 5.
proof **(Theorem 4)** Let $P_1 = [p_{1,1}, p_{1,2}, ..., p_{1,m}]$ and similarly $P_2 = [p_{2,1}, p_{2,2}, ..., p_{2,m}]$, $Q_1 = [q_{1,1}, q_{1,2}, ..., q_{1,m}]$, and $Q_2 = [q_{2,1}, q_{2,2}, ..., q_{2,m}]$. Let us consider $1 \leq i \leq m$. By substituting $a_1 = \lambda p_{1,i},\ a_2 = (1-\lambda)p_{2,i}, b_1 = \lambda q_{1,i}$ and $b_2 = (1-\lambda)q_{2,i}$ in the log sum inequality in (24), we obtain

$$\lambda p_{1,i} \log\left(\frac{p_{1,i}}{q_{1,i}}\right) + (1-\lambda)p_{2,i} \log\left(\frac{p_{2,i}}{q_{2,i}}\right) \geq (\lambda p_{1,i} + (1-\lambda)p_{2,i}) \log\left(\frac{\lambda p_{1,i} + (1-\lambda)p_{2,i}}{\lambda q_{1,i} + (1-\lambda)q_{2,i}}\right). \tag{27}$$

Since (27) holds for all $1 \leq i \leq m$ its also true if we sum the left hand side and the right hand side over $1 \leq i \leq m$ and the summation yields (23).

■

B. The Perspective Transform

For the alternative way to prove the convexity of the relative entropy we introduce an operation which preserves convexity:

Theorem 6 (Perspective transform preserve convexity) If $f(x)$ is convex in x, then $tf(\frac{x}{t})$ is convex in (x,t), for $t\geq 0$.

The proof of this theorem can be found in Appendix B of this lecture.

Alternative proof of **(Theorem 4)**: Divergence is defined as:

$$D\left(P||Q\right) = \sum_{x} P\left(x\right) \log \frac{P\left(x\right)}{Q\left(x\right)} \tag{28}$$

we know that $-\log Q$ is a convex function in Q. using the perspective transform with $t=P$ (probability vectors are non-negative) we get that $-P\log\frac{Q}{P}$ is convex is (Q,P). We can also observe that:

$$-P(x)\log\frac{Q(x)}{P(x)} = P(x)\log\frac{P(x)}{Q(x)} \tag{29}$$

Because $D\left(P||Q\right)$ is a non-negative sum of elements of this form, which we have shown that are convex, we conclude that it is convex in (P,Q)

■

Example:

- $D(P_X||U)$ is convex for any probability function P_X where U is the uniform distribution on $|X|$.

IV. Convexity and Concavity properties of Entropy and Mutual Information

Theorem 7 (concavity of entropy) $H(P_X)$ is a concave function of P_X.

proof We can write entropy in terms of divergence as

$$H(P_X) = \log|\mathcal{X}| - D(P_X||U),$$

where U is the uniform distribution on $|\mathcal{X}|$ outcomes. H is concave function because D is convex and $|\mathcal{X}|$ is a constant with respect to P_X.

■

Let us present an alternative proof for this theorem based on the fact that conditioning does not increase entropy.

alternative proof **(Theorem 7)** Let define the following random variable:

$$\theta = \begin{cases} 1 & \text{with probability } \lambda \\ 2 & \text{with probability } 1-\lambda \end{cases} \tag{30}$$

and $P(x|\theta=1) = P^1(x)$, $P(x|\theta=2) = P^2(x)$. Thus $P(x) = \lambda P^1(x) + (1-\lambda)P^2(x)$. It can be seen that for this problem

$$H(X) = H(P_X) = H(\lambda P^1(x) + (1-\lambda)P^2(x)) \tag{31}$$

$$H(X|\theta) = \lambda H(X|\theta=1) + (1-\lambda)H(X|\theta=2) = \lambda H(P_X^1) + (1-\lambda)H(P_X^2)\ . \tag{32}$$

By substituting (31)-(32) in the following inequality

$$H(X) \geq H(X|\theta), \tag{33}$$

we obtain that $H(P_X)$ is concave function as a function of the distribution, P_X.

■

Exercise 8 Prove that $D(p||q)$ is convex in p for a fixed q and similarly convex in q for a fixed p.

Theorem 8 (Convexity and concavity of the mutual information) The following holds:
1. The mutual information $I(X;Y)$ is a concave function of P_X for fixed $P_{Y|X}$
2. The mutual information $I(X;Y)$ is a convex function of $P_{Y|X}$ for fixed P_X.

Interpretation: For given (constant) system we can maximize the input such that the information will be maximized. For given input we can minimize the system (channel) such that the information will be maximized.
proof

Part 1: This part is proven by the concavity of entropy. By definition:

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x \in \mathcal{X}} P(x)H(Y|X = x). \tag{34}$$

According to Bayes rule:

$$P(y) = \sum_{x \in \mathcal{X}} P(y|x)P(x). \tag{35}$$

Thus, if $P_{Y|X}$ is fixed, P_Y is a linear function of P_X. Furthermore $H(Y)$ is a concave function of P_Y (Theorem 7). Because of the linear relation between P_Y and P_X it follows that $H(Y)$ is a concave function of P_X. In addition, $\sum_{x \in \mathcal{X}} P(x)H(Y|X = x)$ is also linear function of P_X. Hence, the difference is a concave function of P_X.

Part 2: This part is proven by the convexity of the divergence. For given P_X we consider two different joint distributions, $P^1_{X,Y}, P^2_{X,Y}$ with the corresponding conditional distributions, $P^1_{Y|X}$ and $P^2_{Y|X}$ and marginal distributions, P^1_Y, P^2_Y. We define[1]

$$P^\lambda_{Y|X} \triangleq \lambda P^1_{Y|X} + (1 - \lambda)P^2_{Y|X}, \quad 0 \leq \lambda \leq 1 \tag{36}$$

and

$$\begin{aligned} P^1_{X,Y} &\triangleq P_X P^1_{Y|X}, & (37)\\ P^2_{X,Y} &\triangleq P_X P^2_{Y|X}, & (38)\\ P^\lambda_{X,Y} &\triangleq P_X P^\lambda_{Y|X}, & (39) \end{aligned}$$

[1] Note that $P^\lambda_{Y|X}$ is also a conditional distribution measure.

$$(40)$$

Using this definition, it can be seen that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:

$$\begin{aligned} P^{\lambda}(x,y) &= \lambda P^{1}(x,y) + (1-\lambda)P^{2}(x,y), && (41) \\ P^{\lambda}(y) &= \lambda P^{1}(y) + (1-\lambda)P^{2}(y). && (42) \end{aligned}$$

Lets define

$$\begin{aligned} Q^{\lambda}(x,y) &\triangleq P(x)P^{\lambda}(y), \quad 0 \le \lambda \le 1 \\ Q^{1}(x,y) &\triangleq P(x)P^{1}(y), \\ Q^{2}(x,y) &\triangleq P(x)P^{2}(y), \end{aligned} \tag{43}$$

Then,

$$Q^{\lambda}(x,y) = \lambda Q^{1}(x,y) + (1-\lambda)Q^{2}(x,y), \quad 0 \le \lambda \le 1. \tag{44}$$

Let us define $I^{\lambda}(X;Y)$ to be the mutual information induced by $P^{\lambda}_{Y|X}P_X$ and $I^{1}(X;Y)$ and $I^{2}(X;Y)$, the mutual information induced by $P^{1}_{Y|X}P_X$ and $P^{2}_{Y|X}P_X$, respectively. We need to show that

$$I^{\lambda}(X;Y) \le \lambda I^{1}(X;Y) + (1-\lambda)I^{2}(X;Y). \tag{45}$$

To prove this consider the following:

$$\begin{aligned} I^{\lambda}(X;Y) &= D(P^{\lambda}_{X,Y}||Q^{\lambda}_{X,Y}) \\ &\overset{(a)}{=} D(\lambda P^{1}_{X,Y} + (1-\lambda)P^{2}_{X,Y}||\lambda Q^{1}_{X,Y} + (1-\lambda)Q^{2}_{X,Y}) \\ &\overset{(b)}{\le} \lambda D(P^{1}_{X,Y}||Q^{1}_{X,Y}) + (1-\lambda)D(P^{2}_{X,Y}||Q^{2}_{X,Y}) \\ &= \lambda I^{1}(X;Y) + (1-\lambda)I^{2}(X;Y). \end{aligned} \tag{46}$$

where (a) follows the definition of $P^{\lambda}_{X,Y}$ and $Q^{\lambda}_{X,Y}$ and (b) follows from the convexity of divergence (Theorem 4).

■

Appendix

I. Alternative proof of Theorem 1 (Jensen's inequality)

Before starting the proof let us state a lemma from convex analysis that we use. Let $\mathcal{L}$ denote a set of all linear functions that are below $\phi(x)$, i.e.,

$$\mathcal{L} = \{(a,b) : ax + b \le \phi(x), \forall x\}. \tag{47}$$

Lemma 4 (Alternative representation of a convex function) A convex function $\phi(x)$ equals the supremum over all linear function $l(x) = ax + b$ that satisfies $ax + b \leq \phi(x)$ for all x. In other words

$$\phi(x) = \sup_{(a,b)\in\mathcal{L}} \{ax + b\}. \tag{48}$$

proof **(Theorem 1)** (Jensen's inequality)
We need to prove that if $\phi(x)$ is convex then $\mathbb{E}\phi(x) \geq \phi(\mathbb{E}[x])$. Let $\mathcal{L}$ be defined as in (47). Now, choose a specific $(a, b) \in \mathcal{L}$. We have from the definition

$$\phi(x) \geq ax + b, \ \forall x. \tag{49}$$

Whenever we have two random variable that satisfy $U \geq V$, then $\mathbb{E}[U] \geq \mathbb{E}[V]$. Hence

$$\mathbb{E}\phi(x) \geq \mathbb{E}[ax + b], \tag{50}$$

and from the linearity of expectation we obtain that

$$\mathbb{E}\phi(x) \geq a\mathbb{E}[x] + b. \tag{51}$$

Equation (51) holds for all $(a, b) \in \mathcal{L}$, hence

$$\mathbb{E}\phi(x) \geq \sup_{(a,b)\in\mathcal{L}} a\mathbb{E}[x] + b. \tag{52}$$

Finally, follows from Lemma 4 that $\sup_{(a,b)\in\mathcal{L}} a\mathbb{E}[x] + b = \phi(\mathbb{E}[x])$, and therefore

$$\mathbb{E}\phi(x) \geq \phi(\mathbb{E}[x]). \tag{53}$$

■

II. proof of Theorem 6 (Perspective transform preserve convexity)

Let $f(x), f : \mathbb{R} \to \mathbb{R}$ be some convex of x. Let us define $g(x, t) \triangleq tf(\frac{x}{t})$. Now, we can check the condition for convexity of g in (x, t): Let $(x_1, t_1), (x_2, t_2) \in dom(f)$

$$\begin{aligned} g(\lambda(x_1, t_1) + \bar{\lambda}(x_2, t_2)) &= (\lambda t_1 + \bar{\lambda} t_2) f\left(\frac{\lambda t_1(\frac{x_1}{t_1}) + \bar{\lambda} t_2(\frac{x_2}{t_2})}{\lambda t_1 + \bar{\lambda} t_2}\right) \\ &\leq (\lambda t_1 + \bar{\lambda} t_2)\left(\frac{\lambda t_1}{\lambda t_1 + \bar{\lambda} t_2} f\left(\frac{x_1}{t_1}\right) + \frac{\bar{\lambda} t_2}{\lambda t_1 + \bar{\lambda} t_2} f\left(\frac{x_2}{t_2}\right)\right) \\ &= \lambda t_1 f\left(\frac{x_1}{t_1}\right) + \bar{\lambda} t_2 f\left(\frac{x_2}{t_2}\right) \\ &= \lambda g(x_1, t_1) + \bar{\lambda} g(x_2, t_2) \end{aligned} \tag{54}$$

So, $g(x, t)$ satisfies the convexity definition, and therefore a convex function in (x, t) ■

References

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New-York, 2nd edition, 2006.