Thanks for THOMAS M.COVER et al. This proof is a part of their book 《Elements of Information Theory》.

Proof

$$
\begin{aligned}
&D(p(x, y)||q(x, y)) \\
&= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{q(x, y)} && (2.68) \\
&= \sum_x \sum_y p(x, y) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} && (2.69) \\
&= \sum_x \sum_y p(x, y) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x, y) \log \frac{p(y|x)}{q(y|x)} && (2.70) \\
&= D(p(x)||q(x)) + D(p(y|x)||q(y|x)). \quad \square && (2.71)
\end{aligned}
$$

Proof Begin

2.6 JENSEN'S INEQUALITY AND ITS CONSEQUENCES

In this section we prove some simple properties of the quantities defined earlier. We begin with the properties of convex functions.

Definition A function $f(x)$ is said to be *convex* over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \le \lambda \le 1$,

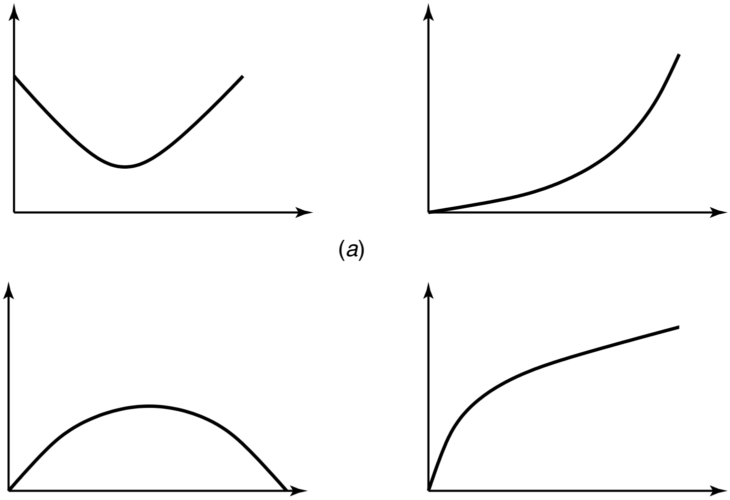
$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2). \tag{2.72}$$

A function f is said to be *strictly convex* if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Definition A function f is *concave* if $-f$ is convex. A function is convex if it always lies below any chord. A function is concave if it always lies above any chord.

Examples of convex functions include x^2, $|x|$, e^x, $x \log x$ (for $x \ge 0$), and so on. Examples of concave functions include $\log x$ and $\sqrt{x}$ for $x \ge 0$. Figure 2.3 shows some examples of convex and concave functions. Note that linear functions $ax + b$ are both convex and concave. Convexity underlies many of the basic properties of information-theoretic quantities such as entropy and mutual information. Before we prove some of these properties, we derive some simple results for convex functions.

Theorem 2.6.1 *If the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval.*



(*a*)

(*b*)

FIGURE 2.3. Examples of (a) convex and (b) concave functions.

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

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2, \tag{2.73}$$

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

We let $x_0 = \lambda x_1 + (1 - \lambda)x_2$ and take $x = x_1$, to obtain

$$f(x_1) \geq f(x_0) + f'(x_0)((1 - \lambda)(x_1 - x_2)). \tag{2.74}$$

Similarly, taking $x = x_2$, we obtain

$$f(x_2) \geq f(x_0) + f'(x_0)(\lambda(x_2 - x_1)). \tag{2.75}$$

Multiplying (2.74) by λ and (2.75) by $1 - \lambda$ and adding, we obtain (2.72). The proof for strict convexity proceeds along the same lines. □

Theorem 2.6.1 allows us immediately to verify the strict convexity of x^2, e^x, and $x \log x$ for $x \geq 0$, and the strict concavity of $\log x$ and $\sqrt{x}$ for $x \geq 0$.

Let E denote expectation. Thus, $EX = \sum_{x \in \mathcal{X}} p(x)x$ in the discrete case and $EX = \int x f(x)\, dx$ in the continuous case.

The next inequality is one of the most widely used in mathematics and one that underlies many of the basic results in information theory.

Theorem 2.6.2 (*Jensen's inequality*) *If f is a convex function and X is a random variable,*

$$Ef(X) \geq f(EX). \tag{2.76}$$

Moreover, if f is strictly convex, the equality in (2.76) implies that $X = EX$ with probability 1 (i.e., X is a constant).

Proof: We prove this for discrete distributions by induction on the number of mass points. The proof of conditions for equality when f is strictly convex is left to the reader.

For a two-mass-point distribution, the inequality becomes

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2), \tag{2.77}$$

which follows directly from the definition of convex functions. Suppose that the theorem is true for distributions with $k-1$ mass points. Then writing $p_i' = p_i/(1-p_k)$ for $i = 1, 2, \ldots, k-1$, we have

$$\sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i) \tag{2.78}$$

$$\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) \tag{2.79}$$

$$\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p_i' x_i\right) \tag{2.80}$$

$$= f\left(\sum_{i=1}^{k} p_i x_i\right), \tag{2.81}$$

where the first inequality follows from the induction hypothesis and the second follows from the definition of convexity.

The proof can be extended to continuous distributions by continuity arguments. □

We now use these results to prove some of the properties of entropy and relative entropy. The following theorem is of fundamental importance.

Theorem 2.6.3 (*Information inequality*) *Let $p(x), q(x), x \in \mathcal{X}$, be two probability mass functions. Then*

$$D(p||q) \geq 0 \tag{2.82}$$

with equality if and only if $p(x) = q(x)$ for all x.

Proof: Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$-D(p||q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \tag{2.83}$$

$$= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \tag{2.84}$$

$$\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \tag{2.85}$$

$$= \log \sum_{x \in A} q(x) \tag{2.86}$$

$$\leq \log \sum_{x \in \mathcal{X}} q(x) \tag{2.87}$$

$$= \log 1 \tag{2.88}$$

$$= 0, \tag{2.89}$$

where (2.85) follows from Jensen's inequality. Since $\log t$ is a strictly concave function of t, we have equality in (2.85) if and only if $q(x)/p(x)$ is constant everywhere [i.e., $q(x) = cp(x)$ for all x]. Thus, $\sum_{x \in A} q(x) = c \sum_{x \in A} p(x) = c$. We have equality in (2.87) only if $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, which implies that $c = 1$. Hence, we have $D(p||q) = 0$ if and only if $p(x) = q(x)$ for all x. Proof End

Corollary (*Nonnegativity of mutual information*) *For any two random variables, X, Y,*

$$I(X; Y) \geq 0, \tag{2.90}$$

with equality if and only if X and Y are independent.

Proof: $I(X; Y) = D(p(x, y)||p(x)p(y)) \geq 0$, with equality if and only if $p(x, y) = p(x)p(y)$ (i.e., X and Y are independent). □