$$\int_{\mathscr{X}} p_0(x) \left(\log_2 \frac{p_0(x)}{p_1(x)} \right) dx \geq -\log_2(1) \tag{7.89}$$

The left hand side of Equation 7.89, based on Equation 7.84, is just $\mathscr{D}_{KL}(0 \to 1)$. Therefore, we can say,

$$\mathscr{D}_{KL}(0 \to 1) \geq 0 \tag{7.90}$$

which is a very important result, proving an important property of a *divergence*.

Note that Equation 7.90 may be written in terms of the *expected values* of the $f(q(x))$ and $f(p(x))$, where $f(x)$ is given by Equation 7.87,

$$-\int_{\mathscr{X}} p_0(x) \log_2 p_1(x) dx \geq -\int_{\mathscr{X}} p_0(x) \log_2 p_0(x) dx \tag{7.91}$$

where the left hand side of Equation 7.91 is known as the *cross entropy* of the true density of X with any other density, $p_1(x)$, and is denoted by $\hbar(p_0 \to p_1)$ for the continuous case and $\mathscr{H}(p_0 \to p_1)$ for the discrete case. Note the following formal definitions of *cross entropy*:

Definition 7.16 (Differential Cross Entropy). *The differential cross entropy, $\hbar(p_0 \to p_1)$, of two probability density functions, $p_0(x)$ and $p_1(x)$ is given by the following expression, when the Lebesgue measure is used,*

$$\hbar(p_0 \to p_1) \overset{\Delta}{=} -\int_{-\infty}^{\infty} p_0(x) \log_2 p_1(x) dx \tag{7.92}$$

Definition 7.17 (Cross Entropy). *Consider the discrete source of Section 7.3. The cross entropy, $\mathscr{H}(p_0 \to p_1)$, of two different probability mass functions, $p_0(X)$ and $p_1(X)$, for the discrete random variable X is given by,*

$$\mathscr{H}(p_0 \to p_1) \overset{\Delta}{=} -\sum_{i=1}^{n} p_0(X_i) \log_2 p_1(X_i) \tag{7.93}$$

Therefore,

$$\hbar(p_0) \leq \hbar(p_0 \to p_1) \tag{7.94}$$

for the continuous case and

$$\mathscr{H}(p_0) \leq \mathscr{H}(p_0 \to p_1) \tag{7.95}$$

for the discrete case.

Equation 7.94 is known as *Gibb's inequality* and it states that the Entropy is always less than or equal to the *cross entropy*, where $p_0(x)$ is the true probability density function of X and $p_1(x)$ is any other density function.

Before *Kullback and Leibler* [12], *Jeffreys* [10] defined a measure, now known as *Jeffreys' divergence*, which is related to the *Kullback-Leibler directed divergence* as follows,

$$\mathscr{D}_J\left(0 \leftrightarrow 1\right) = \int_{\mathscr{X}} \log_2 \frac{dP_0}{dP_1} d(P_0 - dP_1) \tag{7.96}$$

Jeffreys called it an invariant for expressing the difference between two distributions and denoted it as I_2. It is easy to see that this integral is really the sum of the two Kullback and Leibler directed divergences, one in favor of H_0 and the other in favor of H_1. Therefore,

$$\mathscr{D}_J\left(0 \leftrightarrow 1\right) = \mathscr{D}_{KL}\left(0 \rightarrow 1\right) + \mathscr{D}_{KL}\left(1 \rightarrow 0\right) \tag{7.97}$$

$$= \int_{\mathscr{X}} (p_0(x) - p_1(x)) \log_2 \frac{p_0(x)}{p_1(x)} dx \tag{7.98}$$

It is apparent that $\mathscr{D}_J\left(0 \leftrightarrow 1\right)$ is symmetric with respect to hypotheses H_0 and H_1, so it is a measure of the *divergence* between these hypotheses. Although $\mathscr{D}_J\left(0 \leftrightarrow 1\right)$ is *symmetric*, it still does not obey the *triangular inequality* property, so it cannot be considered to be a *metric*.

Throughout this book, we use $\mathscr{D}\left(0 \rightarrow 1\right)$ to denote a *directed divergence*, $\mathscr{D}\left(0 \leftrightarrow 0\right)$ to denote a (symmetric) *divergence* and $d\left(0, 1\right)$ for a distance. The subscripts, such as the KL in $\mathscr{D}_{KL}\left(0 \rightarrow 1\right)$, specify the type of *directed divergence*, *divergence* or *distance*.

It was mentioned that the nature of the measure is such that it may specify any type of random variable including a *discrete random variable*. In that case, the *KL-divergence* may be written as,

$$\mathscr{D}_{KL}\left(0 \rightarrow 1\right) = \sum_{x_i \in X} P_0(x_i) \log_2 \frac{P_0(x_i)}{P_1(x_i)} \tag{7.99}$$

See Section 8.2.1 for the expression for the *KL-divergence* between two normal density probability density functions.

7.6.1 Mutual Information

Consider a special case of *relative entropy* for a random variable defined in the *two-dimensional Cartesian product space* $(\mathscr{X}, \mathfrak{X})$, where $\{\mathscr{X} = \mathscr{R}^2\}$ – see Section 6.2.2. Then the *relative entropy* (*KL-divergence*) in favor of hypothesis H_0 ver-