Kullback-Leibler Divergence

- The Kullback-Leibler or KL Divergence is an approximate measure of the dissimilarity between two probability distributions R and Q, where R corresponds to an empirical distribution and Q is the distribution obtained from a model or theory

- It is denoted as $D_{KL}(R \| Q)$ and called the divergence from Q to R

- Definition (for discrete distbns): $D_{KL}(R \| Q) = \sum_i R(i) \log \frac{R(i)}{Q(i)}$ ... (8)

- It can be seen to be the Expectation of the logarithmic difference between the distributions R and Q, where the Expectation is based on the empirical distribution R.

* that which follows from observations and facts rather than from theory or logic.

Kullback-Leibler Divergence

- From the Gibb's inequality, it follows
- $$D_{KL}(R \| Q) = \sum_i R(i) \log R(i) - \sum_i R(i) \log Q(i) \geq 0 \qquad \ldots \qquad (11)$$
- Now, *R(i)* is fixed from the extracted sample values; we have to obtain *Q(i)* by choice of appropriate θ
- As discussed, that *Q(i)* will be the best which makes the inequality in eq. (11) closest to zero, and by corollary the associated θ will be the best choice of parameters
- Inequality in (11) will be closest to zero when the 2nd term, i.e.

$$\sum_i R(i) \log Q(i) \qquad \ldots \qquad (12)$$

maximizes. Note this term is exactly the same as the term we are seeking to *maximize* in eq. (7), i.e.

$$\theta_{ML} = \arg\max_{\theta} \mathrm{E}_{x \in \hat{p}_{data}} \log p_{model}(x;\theta)$$

- Thus from the KL Divergence relations one can also derive the Maximum Likelihood Estimation Principle for extraction of the best model parameters in any ML algorithm.