probability redistribution models. The quantity $\alpha(x)$ is a normalization factor required to ensure that $\sum_y P_{BO}(y|x) = 1$:

$$\begin{aligned}\alpha(x) &= \frac{\tilde{\beta}(x)}{\sum_{y:C(x,y)=0} P_r(y|x)} \\ &= \frac{\tilde{\beta}(x)}{1 - \sum_{y:C(x,y)>0} P_r(y|x)}.\end{aligned}$$

The second formulation of the normalization is computationally preferable because it is generally the case that the total number of possible pairs far exceeds the number of observed pairs.

Should we use Jelinek-Mercer smoothing, Katz back-off smoothing, or perhaps some other technique? A thorough study by Chen and Goodman (1996) showed that back-off and Jelinek-Mercer smoothing perform consistently well, with back-off generally yielding better results for modeling pairs. Since the back-off formulation also contains a placeholder for us to apply similarity-based estimates, we will use Katz's estimation method whenever smoothed distributions are required.

2.3 Measures of Distributional Similarity

In this section, we consider theoretical and computational properties of several functions measuring the "similarity" between distributions. We refer to these functions as distance functions, rather than similarity functions, since most of them achieve their *minimum* when the two distributions being compared are *maximally* similar (i.e., identical). The work described in chapters 4 and 5 uses negative exponentials of distance functions when true similarity functions (that is, functions that increase as similarity increases) are required.

We certainly do not intend to give an exhaustive listing of all distance functions. (See Anderberg (1973) for an extensive survey.) Our purpose is simply to examine important properties of functions that we use or that are commonly employed by other researchers in natural language processing and machine learning.

We discuss the KL divergence in section 2.3.1 in detail, as it forms the basis for most of the work in this thesis. We also describe several other distance functions, including the total divergence to the mean (section 2.3.2), various geometric norms (section 2.3.3), and some similarity statistics (section 2.3.4). We will pay particular attention to the computational requirements of these functions. In view of the fact that we wish to use very large data sets, we will require that the time needed to calculate the distance between any two distributions be linear or near-linear in the number of attributes. This demand is not strictly necessary for the work described in this thesis – the clustering work of chapter 3 depends on the use of the KL divergence, and the similarity computations of chapters 4 and 5 are done in a preprocessing phrase. However, one of our future goals is to find adaptive versions of our algorithms, in which case we must use functions that can be computed efficiently.

We defer discussion of the *confusion probability*, defined by Essen and Steinbiss (1992), until chapter 4. This function is of great importance to us because Essen and Steinbiss's *co-occurrence smoothing* method is quite similar to our own work on language modeling. The reason we do not include the confusion probability in this chapter is that it is not a function of two distributions : each object x is described both by the conditional probability $P(y|x)$ and the marginal probability $P(x)$, so that comparing two objects involves four distributions.

For the remainder of this section, let x_1, x_2, and x_3 be three objects with associated distributions $P(\cdot|x_1)$, $P(\cdot|x_2)$, and $P(\cdot|x_3)$, respectively. It doesn't matter how these distributions were estimated. For notational convenience, we will call these distributions q, r, and s. We will occasionally refer to a distribution p by its corresponding attribute vector $(p(y_1), p(y_2), \ldots, p(y_N))$.

2.3.1 KL Divergence

We define the function $D(q||r)$ as

$$D(q||r) = \sum_{y \in \mathcal{Y}} q(y) \log \frac{q(y)}{r(y)} \tag{2.5}$$

(we will not specify the base of the logarithm). Limiting arguments lead us to set $0 \log \frac{0}{r} = 0$, even if $r = 0$, and $q \log \frac{q}{0} = \infty$ when q is not zero.

Function (2.5) goes by many names in the literature, including information gain (Rényi, 1970), error (Kerridge, 1961), relative entropy, cross entropy, and Kullback Leibler distance (Cover and Thomas, 1991). Kullback himself refers to the function as information for discrimination, reserving the term "divergence" for the symmetric function $D(q||r) + D(r||q)$ (Kullback, 1959). We will use the name *Kullback-Leibler (KL) divergence* throughout this thesis.

The KL divergence is a standard information-theoretic "measure" of the dissimilarity between two probability mass functions, and has been applied to natural language processing (as described in this thesis), machine learning, and statistical physics. It is not a metric in the technical sense, for it is not symmetric and does not obey the triangle inequality (see, e.g., theorem 12.6.1 of Cover and Thomas (1991)). However, it is non-negative, as shown in the following theorem.

Theorem 2.1 (Information inequality) *$D(q||r) \geq 0$, with equality holding if and only if $q(y) = r(y)$ for all $y \in \mathcal{Y}$.*

Proof. Most authors prove this theorem using *Jensen's inequality*, which deals with expectations of convex functions (notice that $D(q||r)$ is the expected value with respect to q of the quantity $\log(q/r)$). However, we present here a short proof attributed to Elizabeth Thompson (Green, 1996).

Let ln denote the natural logarithm, and let $b > 0$ be the base of the logarithm in (2.5). First observe that for any $z \geq 0$, $\ln(z) \leq z - 1$, with equality holding if and only if $z = 1$. Then, we can write

$$\begin{aligned}
-D(q||r) &= \sum_{y \in \mathcal{Y}} q(y) \log_b \frac{r(y)}{q(y)} \\
&= \frac{1}{\ln(b)} \sum_{y \in \mathcal{Y}} q(y) \ln \frac{r(y)}{q(y)} \\
&\leq \frac{1}{\ln(b)} \sum_{y \in \mathcal{Y}} q(y) \left(\frac{r(y)}{q(y)} - 1 \right) \\
&= \frac{1}{\ln(b)} \left(\sum_{y \in \mathcal{Y}} r(y) - \sum_{y \in \mathcal{Y}} q(y) \right) \\
&= \frac{1}{\ln(b)} (1 - 1) = 0,
\end{aligned}$$

with equality holding if and only if $\frac{r(y)}{q(y)} = 1$ for all $y \in \mathcal{Y}$. ∎

Since the KL divergence is 0 when the two distributions are exactly the same and greater than 0 otherwise, it is really a measure of dissimilarity, as mentioned above, rather than similarity. This yields an intuitive explanation of why we should not expect the KL divergence to obey the triangle inequality: as Hatzivassiloglou and McKeown (1993) observe, dissimilarity is not transitive.

What motivates the use of the KL divergence, if it is not a true distance metric? We appeal to statistics, information theory, and the maximum entropy principle.

The statistician Kullback (1959) derives the KL divergence from a Bayesian perspective. Let Y be a random variable taking values in $\mathcal{Y}$. Suppose we are considering exactly two hypotheses about Y: H_q is the hypothesis that Y is distributed according to q, and H_r is the hypothesis that Y is distributed according to r. Using Bayes' rule, we can write the posterior probabilities of the two hypotheses as

$$P(H_q|y) = \frac{P(H_q)q(y)}{P(H_q)q(y) + P(H_r)r(y)},$$

and

$$P(H_r|y) = \frac{P(H_r)r(y)}{P(H_q)q(y) + P(H_r)r(y)}.$$

Taking logs of both equations and subtracting, we obtain

$$\log \frac{q(y)}{r(y)} = \log \frac{P(H_q|y)}{P(H_r|y)} - \log \frac{P(H_q)}{P(H_r)}.$$

We can therefore consider $\log(q(y)/r(y))$ to be the information y supplies for choosing H_q over H_r: it is the difference between the logarithms of the posterior odds ratio and the prior odds ratio. $D(q||r)$ is then the average information for choosing H_q over H_r. Thus, the KL divergence does indeed measure the dissimilarity between two distributions, since the greater their divergence is, the easier it is, on average, to distinguish between them.

Another statistical rationale for using the KL divergence is given by Cover and Thomas (1991). Let the *empirical frequency distribution* of a sample $\mathbf{y}$ of length n be the probability mass function $p_{\mathbf{y}}$, where $p_{\mathbf{y}}(y)$ is simply the number of times y showed up in the sample divided by n.

Theorem 2.2 *Let r be a hypothesized source distribution. The probability according to r of observing a sample of length n with empirical frequency distribution q is approximately $b^{-nD(q||r)}$, where b is the base of the logarithm function.*

Therefore, we see that if we are trying to decide between hypotheses $r_1, r_2, \ldots, r_k$ when q is the empirical frequency distribution of the observed sample, then $D(q||r_i)$ gives the relative weight of evidence in favor of hypothesis r_i.

The KL divergence arises in information theory as a measure of coding inefficiency. If Y is distributed according to q, then the average codeword length of the best code for Y is the *entropy* $H(q)$ of q:

$$H(q) = -\sum_{y \in \mathcal{Y}} q(y) \log q(y).$$

However, if distribution r were (mistakenly) used to encode Y, then the average codeword length of the resulting code would increase by $D(q||r)$. Therefore, if the divergence between q and r is large, then q and r must be dissimilar, since it is inefficient (on average) to use r in place of q.

Finally, we look at the maximum entropy argument. The entropy of a distribution can be considered a measure of its uncertainty; distributions for which many outcomes are likely (so that one is "uncertain" which outcome will occur) can only be described by relatively complicated codes. The *maximum entropy principle*, first stated by Jaynes (1957), is to assume that the distribution underlying some observed data is the distribution with the highest entropy among all those consistent with the data – that is, one should pick the distribution that makes the fewest assumptions necessary. If one accepts the maximum entropy principle, then one can use it to motivate the use of the KL divergence in the following manner. The distribution $\tilde{r}(y) = 1/|\mathcal{Y}|$ is certainly the a priori maximum entropy distribution. We can write

$$\begin{aligned} D(q||\tilde{r}) &= \sum_{y \in \mathcal{Y}} q(y) \log q(y) - \sum_{y \in \mathcal{Y}} q(y) \log \tilde{r}(y) \\ &= -H(q) - \log \frac{1}{|\mathcal{Y}|} \\ &= \log |\mathcal{Y}| - H(q). \end{aligned}$$

Maximizing entropy is therefore equivalent to minimizing the KL divergence to the prior $\tilde{r}$ given above, subject to the constraint that one must choose a distribution that fits the data.

To summarize, we have described three motivations for using the KL divergence. For the sake of broad acceptability, we have given both Bayesian arguments (those that refer to priors) and

non-Bayesian ones.[2] These are by no means the only reasons. For further background, see Cover and Thomas (1991) and Kullback (1959) for general information, Aczél and Daróczy (1975) for an axiomatic development, and Rényi (1970) for a description of information theory that uses the KL divergence as a starting point.

Some authors (Brown et al., 1992; Church and Hanks, 1990; Dagan, Marcus, and Markovitch, 1995; Luk, 1995) use the *mutual information*, which is the KL divergence between the joint distribution of two random variables and their product distributions. Let A and B be two random variables with probability mass functions $f(A)$ and $g(B)$, respectively, and let $h(A, B)$ be their joint distribution function. Then

$$I(A, B) = D(h||f \cdot g) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} h(a, b) \log \frac{h(a, b)}{f(a) \cdot g(b)}, \tag{2.6}$$

where $\mathcal{A}$ and $\mathcal{B}$ denote the sets of possible values for A and B, respectively. The mutual information measures the dependence of A and B, for if A and B are independent, then $h = f \cdot g$, which implies that the KL divergence between h and $f \cdot g$ is zero by the information inequality (theorem 2.1). We will not give the mutual information further consideration because we do not wish to attempt to estimate joint distributions. Indeed, Church and Hanks (1990) consider two words to be associated if the words occur near each other in some sample of text; but Hatzivassiloglou and McKeown (1993) note that the occurrence of two adjectives in the same noun phrase means that the adjectives cannot be similar. Thus, the information that joint distributions carry about similarity varies too widely across different applications for it to be a generally useful notion for us.

While there are many theoretical reasons justifying the use of the KL divergence, there is a problem with employing it in practice. Recall that for distributions q and r, $D(q||r)$ is infinite if there is some $y' \in \mathcal{Y}$ such that $r(y') = 0$ but $q(y')$ is nonzero. If we know q and r exactly, then this is sensible, since the value y' allows us to distinguish between q and r with absolute confidence. However, often it is the case that we only have estimates $\hat{q}$ and $\hat{r}$ for q and r. If we are not careful with our estimates, then we may erroneously set $\hat{r}(y)$ to zero for some y for which $q(y) > 0$, with the effect that $D(\hat{q}||\hat{r})$ can be infinite when $D(q||r)$ is not.

There are several ways around this problem. One is to use smoothed estimates, as described above in section 2.2, for q and r; this is the approach taken in chapter 5. Another is to only calculate the KL divergence between distributions and average distributions. The work described in chapter 3 computes divergences to cluster centroids, which are created by averaging a whole class of objects. Chapter 4 describes experiments where we calculate the total divergence of q and r to their average; we examine some properties of the total divergence in the next subsection.

2.3.2 Total Divergence to the Mean

Equation (2.7) gives the definition of the *total (KL) divergence to the mean*, which appears in Dagan, Lee, and Pereira (1997) (A stands for "average"):

$$A(q, r) = D(q||\frac{q+r}{2}) + D(r||\frac{q+r}{2}), \tag{2.7}$$

where $((q+r)/2)(y) = (q(y)+r(y))/2$. If q and r are two empirical frequency distributions (defined just above theorem 2.2), then $A(q, r)$ can be used as a test statistic for the hypothesis that q and r are drawn from the same distribution.

Using theorem 2.1, we see that $A(q, r) \geq 0$, with equality if and only if $q = r$. $A(q, r)$ is clearly a symmetric function, but does not obey the triangle inequality, as will be shown below.

[2] The often heated debates between Bayesians and non-Bayesians are well known. For example, Skilling (1991, pg. 24) writes, "there is a valid defence [sic] of using non-Bayesian methods, namely incompetence."

We can write $A(q,r)$ in a more convenient form by observing that

$$\begin{aligned} D(q\|\frac{q+r}{2}) &= \sum_{y\in\mathcal{Y}} q(y)\log\frac{2q(y)}{q(y)+r(y)} \\ &= \log 2 + \sum_{y\in\mathcal{Y}} q(y)\log\frac{q(y)}{q(y)+r(y)}. \end{aligned}$$

The sum over $y \in \mathcal{Y}$ may be broken up into two parts, a sum over those y such that both $q(y)$ and $r(y)$ are greater than zero, and a sum over those y such that $q(y)$ is greater than zero but $r(y) = 0$. We call these sets *Both* and *Justq*, respectively: $Both = \{y : q(y) > 0, r(y) > 0\}$ and $Justq = \{y : q(y) > 0, r(y) = 0\}$. Then,

$$\begin{aligned} D(q\|\frac{q+r}{2}) &= \log 2 + \sum_{y\in Both} q(y)\log\frac{q(y)}{q(y)+r(y)} + \sum_{y\in Justq} q(y)\log\frac{q(y)}{q(y)+r(y)} \\ &= \log 2 + \sum_{y\in Both} q(y)\log\frac{q(y)}{q(y)+r(y)} + \sum_{y\in Justq} q(y)\log\frac{q(y)}{q(y)} \\ &= \log 2 + \sum_{y\in Both} q(y)\log\frac{q(y)}{q(y)+r(y)}. \end{aligned}$$

A similar decomposition of $D(r\|\frac{q+r}{2})$ into two sums over *Both* and $Justr = \{y : r(y) > 0, q(y) = 0\}$ holds. Therefore, we can write

$$A(q,r) = 2\log 2 + \sum_{y\in Both}\left\{q(y)\log\frac{q(y)}{q(y)+r(y)} + r(y)\log\frac{r(y)}{q(y)+r(y)}\right\}. \tag{2.8}$$

Equation (2.8) is computationally convenient, for it involves sums only over elements of *Both*, as opposed to over all the elements in $\mathcal{Y}$. We will typically consider situations in which *Both* is (estimated to be) much smaller than $\mathcal{Y}$.

Since the two ratios in (2.8) are both less than one, the sum over elements in *Both* is always negative. $A(q,r)$ therefore reaches its maximum when the set *Both* is empty, in which case $A(q,r) = 2\log 2$. This observation makes it easy to see that $A(q,r)$ does not obey the triangle inequality. Let $\mathcal{Y} = \{y_1, y_2\}$. Consider distributions $\tilde{q}$, $\tilde{r}$, and $\tilde{s}$, where

$$\tilde{q}(y_1) = 1,\ \tilde{q}(y_2) = 0; \qquad \tilde{r}(y_1) = \tilde{r}(y_2) = \frac{1}{2}; \qquad \tilde{s}(y_1) = 0,\ \tilde{s}(y_2) = 1.$$

Then $A(\tilde{q},\tilde{r}) + A(\tilde{r},\tilde{s}) = \log 2 + \log(2/3) + 2\log(4/3) = \log 2 + \log(32/27) < 2\log 2$, whereas $A(\tilde{q},\tilde{s}) = 2\log 2$, since the supports for $\tilde{q}$ and $\tilde{s}$ are disjoint. Therefore, $A(\tilde{q},\tilde{r}) + A(\tilde{r},\tilde{s}) \not\geq A(\tilde{q},\tilde{s})$, violating the triangle inequality.

2.3.3 Geometric Distances

If we think of probability mass functions as vectors, so that distribution p is associated with the vector $(p(y_1), p(y_2), \ldots, p(y_N))$ in $\Re^N$, then we can measure the distance between distributions by various geometrically-motivated functions, including the L_1 and L_2 norms and the cosine function. All three of these functions appear quite commonly in the clustering literature (Kaufman and Rousseeuw, 1990; Cutting et al., 1992; Schütze, 1993). The first two functions are true metrics, as the name "norm" suggests.

The L_1 *norm* (also called the "Manhattan" or "taxi-cab" distance) is defined as

$$L_1(q,r) = \sum_{y \in \mathcal{Y}} |q(y) - r(y)|. \tag{2.9}$$

Clearly, $L_1(q,r) = 0$ if and only if $q(y) = r(y)$ for all y. Interestingly, $L_1(q,r)$ bears the following relation, discovered independently by Csiszár and Kemperman, to $D(q||r)$:

$$L_1(q,r) \leq \sqrt{D(q||r) \cdot 2 \ln b}\,, \tag{2.10}$$

where b is the base of the logarithm function. Consequently, convergence in KL divergence implies convergence in the L_1 norm. However, we can find a much tighter bound, as follows. By dividing up the sum in equation (2.9) into sums over *Both*, *Justq*, and *Justr* as defined in section 2.3.2, we obtain

$$L_1(q,r) = \sum_{y \in Justq} q(y) + \sum_{y \in Justr} r(y) + \sum_{y \in Both} |q(y) - r(y)|.$$

Since

$$\sum_{y \in Justq} q(y) = 1 - \sum_{y \in Both} q(y) \qquad \text{and} \qquad \sum_{y \in Justr} r(y) = 1 - \sum_{y \in Both} r(y),$$

we can express $L_1(q,r)$ in a form depending only on the elements of *Both*:

$$L_1(q,r) = 2 + \sum_{y \in Both} \left(|q(y) - r(y)| - q(y) - r(y) \right). \tag{2.11}$$

Applying the triangle inequality to (2.11), we see that $L_1(q,r) \leq 2$, with equality if and only if the set *Both* is empty. Also, (2.11) is a convenient expression from a computational point of view, since we do not need to sum over all the elements of $\mathcal{Y}$. We describe experiments using L_1 as distance function in chapter 4.

The L_2 *norm* is the Euclidean distance between vectors. Let $||\cdot||$ denote the usual norm function, $||q(y)|| = \sqrt{\sum_y q(y)^2}$. Then,

$$L_2(q,r) = ||q(y) - r(y)|| = \left(\sum_{y \in \mathcal{Y}} (q(y) - r(y))^2 \right)^{\frac{1}{2}}.$$

Since the L_1 norm bounds the L_2 norm, the inequality of equation (2.10) also applies to the L_2 norm.

Although the L_2 norm appears quite often in the literature, Kaufman and Rousseeuw (1990) write that

> In many branches of univariate and multivariate statistics it has been known for a long time that methods based on the minimization of sums (or averages) of dissimilarities or absolute residuals (the so-called L_1 methods) are much more robust than methods based on sums of squares (which are called L_2 methods). The computational simplicity of many of the latter methods does not make up for the fact that they are extremely sensitive to the effect of one or more outliers. (pg. 117)

We therefore will not give further consideration to the L_2 norm in this thesis.

Finally, we turn to the *cosine function*. This symmetric function is related to the angle between two vectors; the "closer" two vectors are, the smaller the angle between them.

$$\cos(q,r) = \frac{\sum_{y \in \mathcal{Y}} q(y) r(y)}{||q|| ||r||} \tag{2.12}$$

Notice that the cosine is an inverse distance function, in that it achieves its maximum of 1 when $q(y) = r(y)$ for all y, and is zero when the supports of q and r are disjoint. For all the other functions described above, it is just the opposite: they are zero if and only if $q(y) = r(y)$ for all y, and are greater than zero otherwise. Further analysis of geometric properties of the cosine function and other geometric similarity functions used in information retrieval can be found in Jones and Furnas (1987).

The cosine function is not as efficient to compute as the other functions we have discussed. While the numerator in (2.12) requires only summing over elements of *Both*, the elements of *Justq* and *Justr* must be taken into account in calculating the denominator. It may be desirable to calculate the norms of all distributions as a preprocessing step (we cannot just normalize the vectors because we would violate the constraint that attribute vector components sum to one).

2.3.4 Similarity Statistics

There are many correlation statistics for measuring the association between random variables (Anderberg, 1973, Chapter 4.2). The most well-known of these is the Pearson correlation coefficient; some non-parametric measures are the gamma statistic, Spearman's correlation coefficient, and Kendall's τ coefficient (Gibbons, 1993). The Spearman statistic was used by Finch and Chater (1992) to find syntactic categories, and Kendall's statistic appears in work by Hatzivassiloglou and McKeown (1993) (henceforth H&M) on clustering adjectives. We concentrate on the latter statistic since we will discuss H&M's work in some detail in the next chapter.

Kendall's τ coefficient is based on pairwise comparisons. For every pair of contexts $(y_i,\ y_j)$, we consider the quantities $\alpha_q^{ij} = q(y_i) - q(y_j)$ and $\alpha_r^{ij} = r(y_i) - r(y_j)$. The pair is a *concordance* if both α_q^{ij} and α_r^{ij} have the same sign, and a *discordance* if their signs differ (if either of these quantities is zero, then the pair is a tie, which is neither a concordance nor a discordance). $\tau(q, r)$ is the difference between the probability of observing a concordance and the probability of observing a discordance, and so ranges between -1 and 1. A value of 1 corresponds to perfect concordance (but not necessarily equality) between q and r, -1 corresponds to perfect discordance, and 0 to no correlation. An unbiased estimator of $\tau(q, r)$ is

$$\hat{\tau}(q, r) = \frac{\text{number of observed concordances} - \text{number of observed discordances}}{\binom{|\mathcal{Y}|}{2}}.$$

In terms of computational efficiency, $\tau(q, r)$ is slightly more expensive then the total divergence to the mean or the L_1 norm. In order to calculate the number of discordances, H&M first order the y's in $\mathcal{Y}$ by their probabilities as assigned by q. Then, they rerank the y's according to the probabilities assigned by r. The number of discordances is then exactly the number of discrepancies between the two orderings. Since we need to sort the set $\mathcal{Y}$ and calculate the number of discrepancies between the two orderings, we spend $O(|\mathcal{Y}| \log_2 |\mathcal{Y}|)$ time to calculate the similarity between q and r. An optimization not noted by H&M is that for all $y' \in \mathit{Both} \cup \mathit{Justq} \cup \mathit{Justr}$ and $y'' \notin \mathit{Both} \cup \mathit{Justq} \cup \mathit{Justr}$ (that is, $q(y'') = r(y'') = 0$), the pair (y', y'') cannot be a discordance – it is a concordance if $y' \in \mathit{Both}$ and a tie otherwise. Therefore, we actually only need to sort $\mathcal{Y}' = \mathit{Both} \cup \mathit{Justq} \cup \mathit{Justr}$, a $O(|\mathcal{Y}'| \log_2 |\mathcal{Y}'|)$ operation. In the case of sparse data, this would be a significant time savings, although we would still be using more than linear time.

2.3.5 An Example

To aid in visualizing the behavior of the salient functions described above, we consider a two-dimensional example where $\mathcal{Y} = \{y_1, y_2\}$. In this situation, $q(y_2) = 1 - q(y_1)$ for any distribution q, so we only need to know the value of a distribution at y_1. In figure 2.1, we have plotted the values of various distance functions with respect to a fixed distribution $r = (.5, .5)$. The horizontal axis represents the probability of y_1, so that .75 on the horizontal axis means the distribution $q = (.75, .25)$. The fixed distribution r is at .5 on the horizontal axis.

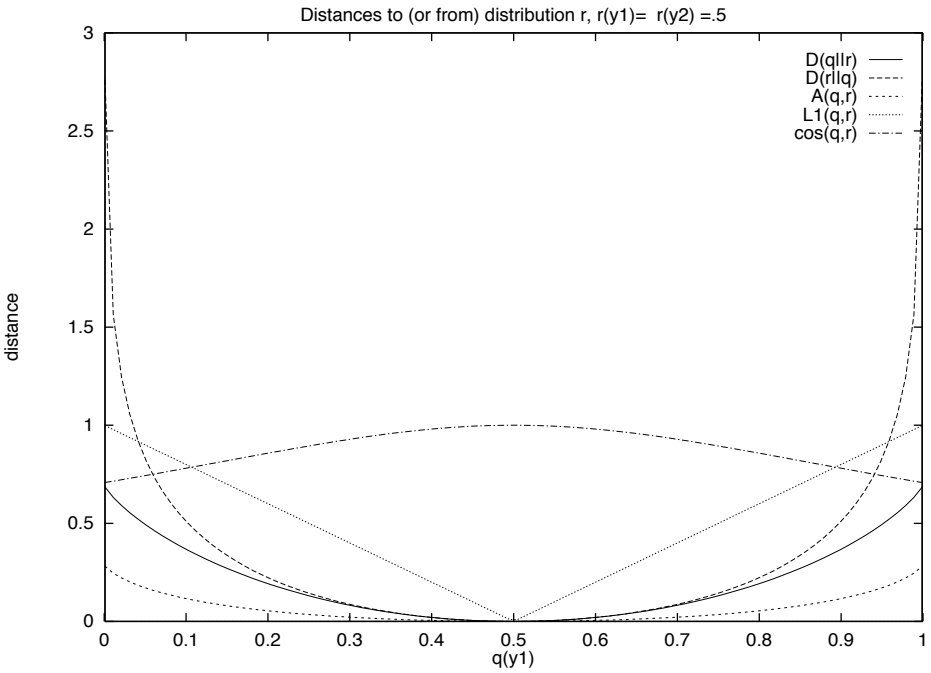

Figure 2.1: Comparison of distance functions

As observed above, the KL divergences, the total divergence to the mean, and the L_1 norm are all zero at r and increase as one travels away from r. The cosine function, on the other hand, is 1 at r and decreases as one travels away from r.

Figure 2.1 demonstrates that the KL divergence is not symmetric, for the curve $D(r||q)$ lies above the curve $D(q||r)$. In general, the KL divergence from a sharp to a flat distribution is less than the divergence from a flat to a sharp distribution – a sharp distribution (such as $(.9, .1)$) is one with relatively high values for some of the attributes, whereas a flat distribution resembles the uniform distribution. The intuition behind this behavior is as follows. If we assume that the source distribution (the second argument to $D(\cdot||\cdot)$) is flat, then it would be somewhat odd to observe a sharp sample distribution. However, it would be even more surprising to observe a flat sample if we believe that the source distribution is sharp. For instance, suppose the source distribution were (.5,.5). Then, the probability of observing 9 y_1's and 1 y_2 in a sample of length 10 (i.e., a sharp empirical distribution) would be

$$\binom{10}{9}(.5)^9(.5)^1 \approx .01.$$

However, if the source distribution were (.9,.1), then the probability of observing 5 y_1's and 5 y_2's (i.e., a flat empirical distribution) would be

$$\binom{10}{5}(.9)^5(.1)^5 \approx .001.$$

An interesting feature to note is that the curve for $A(q, r)$, the total divergence to the mean, is lower than the KL divergence curves, and that these, in turn, are for the most part lower than the L_1 curve. We speculate that the flatness of $D(q||r)$ and $A(q, r)$ relative to $L_1(q, r)$ around the point $q = r$ indicates that these two functions are somewhat more robust to sampling error, for using $q = r + \epsilon$ (for small ϵ) instead of $q = r$ results in a much greater change in the value of the L_1 norm than in the value of the KL divergence or the total divergence to the mean.

2.4 Summary and Preview

We have now established the groundwork for the results of this thesis. We have explained why we want to use distributions to represent objects, and have described ways to estimate these distributions and to measure the similarity between distributions.

We have been working with conditional probabilities induced by objects over contexts. As mentioned above, "objects" and "contexts" are fairly general notions; for instance, an object might be a document and the contexts might be the set of words that can occur in a document. We will confine our attention to modeling pairs of words, so that $\mathcal{X}$ and $\mathcal{Y}$ are sets of words. In chapters 3 and 4, $\mathcal{X}$ is a set of nouns and $\mathcal{Y}$ is a set of transitive verbs; $C(x, y)$ indicates the number of times x was the direct object of verb y. Chapter 5 considers the bigram case, where $\mathcal{X}$ is the set of all possible words, $\mathcal{Y} = \mathcal{X}$, and $C(x, y)$ denotes the number of times word x occurred immediately before the word y.