# Symmetrizing the Kullback-Leibler Distance

Don H. Johnson and Sinan Sinanović
Computer and Information Technology Institute
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77251–1892
dhj@rice.edu

March 18, 2001

**Abstract**

We define a new distance measure — the resistor-average distance — between two probability distributions that is closely related to the Kullback-Leibler distance. While the Kullback-Leibler distance is asymmetric in the two distributions, the resistor-average distance is not. It arises from geometric considerations similar to those used to derive the Chernoff distance. Determining its relation to well-known distance measures reveals a new way to depict how commonly used distance measures relate to each other.

## 1 Introduction

The Kullback-Leibler distance [15, 16] is perhaps the most frequently used information-theoretic "distance" measure from a viewpoint of theory. If $p_0$, $p_1$ are two probability densities, the Kullback-Leibler distance is defined to be

$$\mathcal{D}(p_1 \| p_0) = \int p_1(x) \log \frac{p_1(x)}{p_0(x)} \, dx \; . \tag{1}$$

In this paper, $\log(\cdot)$ has base two. The Kullback-Leibler distance is but one example of the Ali-Silvey class of information-theoretic distance measures [1], which are defined to be $d(p_0, p_1) = f(\mathcal{E}_0[c(\Lambda(X))])$, where $\Lambda(\cdot)$ represents the likelihood ratio $p_1(\cdot)/p_0(\cdot)$, $c(\cdot)$ is convex, $\mathcal{E}_0[\cdot]$ is expected value with respect to the distribution $p_0$ and $f(\cdot)$ is a non-decreasing function. To create the Kullback-Leibler distance within this framework, $c(x) = x \log x$ and $f(x) = x$. Another distance

measure of importance here is the Chernoff distance [5].

$$\mathcal{C}(p_0, p_1) = \max_{0 \leq t \leq 1} -\log \mu(t), \quad \mu(t) = \int [p_0(x)]^{1-t} [p_1(x)]^t \, dx \tag{2}$$

It too is in the Ali-Silvey class, with $c(x) = -x^t$ and $f(x) = -\log(-x)$. A special case of the Chernoff distance is the Bhattacharyya distance [3, 14] $\mathcal{B}(p_0, p_1) = -\log \mu(\frac{1}{2})$. These distances have the *additive* property: The distance between two joint distributions of statistically independent, identically distributed random variables equals the sum of the marginal distances. Note that because of the optimization step, the Chernoff distance is *not* additive when the random variables are not identically distributed; the Kullback-Leibler and Bhattacharyya distances are.

The term "distance" should not be taken rigorously; all of the distances defined here do not obey some of the fundamental axioms distances must satisfy. The Kullback-Leibler distance is not symmetric, and the Chernoff and Bhattacharyya distances do not satisfy the triangle inequality [14]. In fact, $\mathcal{D}(p_1 \| p_0)$ is taken to mean the distance from $p_0$ to $p_1$; because of the asymmetry, the distance from $p_1$ to $p_0$, $\mathcal{D}(p_0 \| p_1)$, is usually different. Despite these difficulties, recent work has shown that the Kullback-Leibler distance is geometrically important [4, 8]. If a manifold of probability distributions were created so that distribution pairs having equivalent optimal classifier defined manifold invariance, no distance metric can exist for the manifold because distance must be an asymmetric quantity. The Kullback-Leibler distance takes on that role for this manifold.

Delving further into this geometry yields a relationship between the Kullback-Leibler and Chernoff distance measures [7]. On the manifold, the geodesic curve $p_t$ linking two given probability distributions $p_0$ and $p_1$ is given by

$$p_t(x) = \frac{[p_0(x)]^{1-t} [p_1(x)]^t}{\mu(t)} \, , 0 \leq t \leq 1,$$

where $\mu(t)$ is defined in equation (2).[1] Consider the halfway point on the manifold defined according to the Kullback-Leibler distance as the distribution equidistant from the endpoints: $\mathcal{D}(p_{t^*} \| p_0) = \mathcal{D}(p_{t^*} \| p_1)$. Equating these distances yields $t^*$ as the parameter value that maximizes $-\log \mu(t)$ with the halfway-distance being the Chernoff distance: $\mathcal{C}(p_0, p_1) = \mathcal{D}(p_{t^*} \| p_0)$. The Bhattacharyya distance essentially chooses "halfway" to mean $t = \frac{1}{2}$.

These distance measures have three properties that make them important in information processing.

1. They (as do all others in the Ali-Silvey class by construction) satisfy a version of the Data Processing Inequality. If $\boldsymbol{\theta} \to X \to Y$ form a Markov chain, with $\boldsymbol{\theta}$ a nonrandom parameter

[1] This geometric theory, though written in terms of marginal distributions, applies to joint probability distributions as well.

vector and $X, Y$ random variables, then $d(X(\boldsymbol{\theta}_1), X(\boldsymbol{\theta}_0)) \geq d(Y(\boldsymbol{\theta}_1), Y(\boldsymbol{\theta}_0))$. This result says that no transformation can increase the distance between probability distributions. All distances that satisfy this inequality are said to be *information-theoretic*.

2. Through Stein's Lemma [6], the Kullback-Leibler and Chernoff distances are the exponential rates of optimal classifier performance probabilities. If $\mathbf{X}$ is a random vector having $N$ statistically independent and identically distributed components under both of the distributions $p_0$, $p_1$, the optimal (likelihood ratio) classifier results in error probabilities that obey the asymptotics

$$\begin{aligned}
\lim_{N\to\infty} \frac{\log P_F}{N} &= -\mathcal{D}(p_1 \| p_0) \ , \text{ fixed } P_M \\
\lim_{N\to\infty} \frac{\log P_e}{N} &= -\mathcal{C}(p_0, p_1) \\
\lim_{N\to\infty} \frac{\log P_e}{N} &\leq -\mathcal{B}(p_0, p_1)
\end{aligned}$$

   Here, $P_F$, $P_M$, and $P_e$ are the false-alarm, miss, and average-error probabilities, respectively. Loosely speaking, Stein's Lemma suggests that these error probabilities decay exponentially in the amount of data available to the likelihood-ratio classifier: for example, $P_F \sim 2^{-N\mathcal{D}(p_1\|p_0)}$ for a Neyman-Pearson classifier [13]. The relevant distance determines the rate of decay. Whether all Ali-Silvey distances satisfy a variant of Stein's Lemma is not known.

3. For perturbational changes in the parameter vector—$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \boldsymbol{\delta\theta}$, all Ali-Silvey distances having the property $f\,(c(1)) = 0$ yield

$$d(X(\boldsymbol{\theta}_0), X(\boldsymbol{\theta}_0 + \boldsymbol{\delta\theta})) = \frac{1}{2} f'\,(c(1))\, c''(1) \boldsymbol{\delta\theta}' \mathbf{F}_X(\boldsymbol{\theta}_0) \boldsymbol{\delta\theta} + O(\|\boldsymbol{\delta\theta}\|^4) \ .$$

   Thus, distance between perturbed stochastic models is proportional to a quadratic form consisting of the perturbation and the Fisher information matrix. The constant of proportionality equals $1/(2\ln 2)$ for the Kullback-Leibler distance, $\left(t^* - (t^*)^2\right)/(2\ln 2)$ for the Chernoff distance, and $1/(2\ln 2)$ for the Bhattacharyya distance.

These three properties directly relate information theoretic distances to the performances of optimal classifiers and optimal parameter estimators.

In addition to its geometric importance, the Kullback-Leibler distance is especially attractive because it can be evaluated. Johnson and Orsak [13] provide a table of Kullback-Leibler distances

between distributions differing in mean. Calculating the Chernoff distance requires solving a conceptually easy optimization problem: maximizing $-\log \mu(t)$ with $\mu(t)$ always being convex. However, analytic calculation of this maximum can be tedious. Recent work has focused on estimating information-theoretic distances from data [9, 11, 12]. Solving the optimization problem required to compute the Chernoff distance becomes a larger issue in this empirical case. For such empirical problems, one could eliminate the optimization by using the Bhattacharyya distance. However, neither of these can be easily computed in Markov situations, where we want to compute the distance between two sequences of random variables that have the same Markovian dependence order. The Kullback-Leibler distance is additive in such cases while the Chernoff and Bhattacharyya are not. Using first-order Markovian dependence among an ordered set of random variables as an example, wherein $p(\mathbf{x}) = p(x_1) \prod_{n=2}^{N} p(x_n \mid x_{n-1})$,

$$\mathcal{D}(p_1 \| p_0) = \mathcal{D}\left(p_1(x_1) \| p_0(x_1)\right) + \sum_{n=2}^{N} \mathcal{D}\left(p_1(x_n \mid x_{n-1}) \| p_0(x_n \mid x_{n-1})\right) \ ,$$

where

$$\mathcal{D}\left(p_1(x_n \mid x_{n-1}) \| p_0(x_n \mid x_{n-1})\right) = \int\int p_1(x_{n-1}, x_n) \log \frac{p_1(x_n \mid x_{n-1})}{p_0(x_n \mid x_{n-1})} \, dx_{n-1} \, dx_n \ .$$

This additivity property becomes important in applications because the distance between distributions having large $N$ can be computed from marginal, two-dimensional, and conditional distributions.

Despite the Kullback-Leibler distance's computational and theoretical advantages, what becomes a nuisance in applications is its lack of symmetry. Simple examples show that the ordering of the arguments in the Kullback-Leibler distance (1) can yield substantially different values. What is needed is symmetric distance measure that can be easily evaluated analytically and estimated, be information-theoretic, and be related to classifier and estimator performance. Ideally, we would like to use the Chernoff distance, but its computational difficulties and its inability to simplify under Markovian models make it unattractive. Consequently, we turn to symmetrizing the Kullback-Leibler distance.

## 2 Results

Although Jeffreys [10] did not develop it to symmetrize the Kullback-Leibler distance, the so-called $J$-divergence equals the average of the two possible Kullback-Leibler distances between two

probability distributions.[2] Assuming the component Kullback-Leibler distances exist,

$$\mathcal{J}(p_0, p_1) = \frac{\mathcal{D}(p_0\|p_1) + \mathcal{D}(p_1\|p_0)}{2} \ .$$

We now have a symmetric quantity that is easily calculated and estimated and is in the Ali-Silvey class ($c(x) = \frac{x-1}{2}\log x$). However, its relation to classifier performance is more tenuous than the other distances [2, 14].

$$\lim_{N\to\infty} \frac{\log P_e}{N} \geq -\mathcal{J}(p_0, p_1)$$

We have found this bound to be loose, with it not indicating well the exponential rate of the average-error probability $P_e$ (which is equal to the Chernoff distance).

To address the symmetry problem, we can consider alternate ways of "averaging" the two Kullback-Leibler distances. What immediately comes to mind are the geometric and harmonic means. The geometric mean $\mathcal{G}(p_0, p_1) = \sqrt{\mathcal{D}(p_0\|p_1)\mathcal{D}(p_1\|p_0)}$ does not seem have as interesting properties as the harmonic mean. We define a new symmetric distance, what we call the *resistor-average* distance, via the harmonic mean.

$$\frac{1}{\mathcal{R}(p_0, p_1)} \equiv \frac{1}{\mathcal{D}(p_1\|p_0)} + \frac{1}{\mathcal{D}(p_0\|p_1)} \tag{3}$$

This quantity gets its name from the formula for the equivalent resistance of a set of parallel resistors: $1/R_{\text{equiv}} = \sum_n 1/R_n$. It equals the harmonic sum (half the harmonic mean) of the component Kullback-Leibler distances. The relation among the various symmetric versions of the component Kullback-Leibler distances is

$$\max\{\mathcal{D}(p_0\|p_1), \mathcal{D}(p_1\|p_0)\} \geq \mathcal{J}(p_0, p_1) \geq \mathcal{G}(p_0, p_1) \geq \min\{\mathcal{D}(p_0\|p_1), \mathcal{D}(p_1\|p_0)\} \geq \mathcal{R}(p_0, p_1) \ .$$

The resistor-average is not an Ali-Silvey distance, but because of its direct relationship to the Kullback-Leibler distance, it does have properties 1 and 3 (as does the $J$-divergence and the geometric mean).[3] The resistor-average distance is not additive in either the Markov or the statistically independent cases; because it is directly computed from quantities that are (Kullback-Leibler distances), it shares the computational and interpretative attributes that additivity offers.

It is not as arbitrary as the $J$-divergence or the geometric mean as it arises by considering the halfway-point on the geodesic using the opposite sense of equidistant used in formulating the

[2] Many authors, including Jeffreys, define the $J$-divergence as the sum rather than the average. Using the average fits more neatly into the graphical relations developed subsequently.

[3] For distances that aren't in the Ali-Silvey class, "satisfying property 3" means $d(X(\boldsymbol{\theta}_0), X(\boldsymbol{\theta}_0 + \boldsymbol{\delta\theta})) = K\boldsymbol{\delta\theta}'\mathbf{F}_X(\boldsymbol{\theta}_0)\boldsymbol{\delta\theta} + O(\|\boldsymbol{\delta\theta}\|^4)$, where $K$ is a constant.

Chernoff distance: Rather than equating the distances from the endpoints to the halfway point, equate distances from the halfway-point to the ends. Denoting this point along the geodesic by $t^{**}$, we seek it by solving $\mathcal{D}(p_0\|p_{t^{**}}) = \mathcal{D}(p_1\|p_{t^{**}})$. In contrast to the notion of "halfway" that results in the Chernoff distance, this problem has a closed form solution.

$$t^{**} = \frac{\mathcal{D}(p_1\|p_0)}{\mathcal{D}(p_1\|p_0) + \mathcal{D}(p_0\|p_1)}$$

$$\mathcal{D}(p_0\|p_{t^{**}}) = \mathcal{R}(p_0, p_1) + \log \mu(t^{**})$$

Note that the term $\log \mu(t^{**})$ is negative. Because the Chernoff distance equals the maximum of $-\log \mu(t)$, we have the bound

$$\mathcal{D}(p_0\|p_{t^{**}}) + \mathcal{C}(p_0, p_1) \geq \mathcal{R}(p_0, p_1) \ .$$

The quantities $t^{**}$ and $t^*$ are equal when the probability distributions are symmetric and differ only in mean. In these special cases, the Kullback-Leibler distance is symmetric, making $\mathcal{R}(p_0, p_1) = \frac{1}{2}\mathcal{D}(p_1\|p_0)$. If in addition, $\mathcal{D}(p_0\|p_{t^{**}}) = \mathcal{C}(p_0, p_1)$, $\mathcal{R}(p_0, p_1) = 2\mathcal{C}(p_0, p_1)$.

The relation between the various distance measures can be visualized graphically (Figure 1). Because $-\log \mu(t)$ is concave, the resistor-average distance upper bounds the Chernoff distance: $\mathcal{R}(p_0, p_1) \geq \mathcal{C}(p_0, p_1)$. Consequently, $\lim_{N\to\infty} \frac{\log P_e}{N} \geq -\mathcal{R}(p_0, p_1)$. Computer experiments show this bound to usually be tight in terms of the exponential rate. In some cases, it can be loose: The curve $-\log \mu(t)$ can lie close to the $t$-axis, leaving the Chernoff and resistor-average distances far apart (Figure 2). Despite these extremes, in many realistic examples the Chernoff distance roughly equals half the resistor-average distance. Consequently, we have an easily computed quantity that can approximate the more difficult to compute Chernoff distance. The $J$-divergence can differ much more from the Chernoff distance while the more difficult-to-compute Bhattacharyya distance can be quite close. This graphical depiction of these distance measures suggests that as the two Kullback-Leibler distances differ more and more, the greater the discrepancy between the Chernoff distance from the $J$-divergence and the Bhattacharyya distance.

Table 1 shows some analytic examples. These examples illustrate some of the variations between the resistor-average and Chernoff distances. For the Laplacian example, the ratio of the resistor-average and Chernoff distances lies between 1 and 2, approaching 1 only when $d' \gg 1$. Despite appearances, the Chernoff and resistor-average distances are indeed symmetric functions of the parameters $\lambda_0$, $\lambda_1$ in the Poisson case. In this example, the resistor-average-Chernoff ratio lies between 1.65 and 2 over a hundred fold range of the ratio $\lambda_1/\lambda_0$.

Another approach to creating a symmetric distance measure was recently described by Topsøe [17]. What we call the Topsøe distance $\mathcal{T}(p_0, p_1)$ equals $\mathcal{D}(p_0\|q) + \mathcal{D}(p_1\|q)$, where

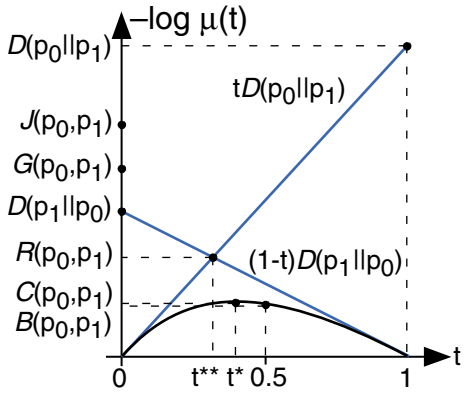


**Figure 1**: This figure portrays relations among many of the most frequently used information-theoretic distances. The focus is the function $-\log \mu(t)$, where $\mu(t) = \int [p_0(x)]^{1-t} [p_1(x)]^t \, dx$. The Chernoff distance is defined as the maximum of this function, with the maximum occurring at $t^*$. This curve's derivatives at $t = 0$ and $t = 1$ are $\mathcal{D}(p_0 \| p_1)$ and $-\mathcal{D}(p_1 \| p_0)$, respectively. The values of these tangent lines at their extremes thus correspond to the Kullback-Leibler distances. The tangent curves intersect at $t = t^{**}$, the value of which corresponds to the resistor-average distance defined in (3). The Bhattacharyya distance $\mathcal{B}(p_0, p_1)$ equals $-\log \mu\left(\frac{1}{2}\right)$. The $J$-divergence $\mathcal{J}(p_0, p_1)$ equals the average of the two Kullback-Leibler distances, with the geometric mean $\mathcal{G}(p_0, p_1)$ lying somewhere between the $J$-divergence and the smaller of the Kullback-Leibler distances.

| *Distribution* | *Kullback-Leibler* | *Chernoff* | *Resistor-Average* |
|---|---|---|---|
| Gaussian | $\frac{(m_1-m_0)^2}{2\sigma^2}$ | $\frac{(m_1-m_0)^2}{8\sigma^2}$ | $\frac{1}{2}\mathcal{D}(p_1 \| p_0)$ |
| Laplacian $d' = \frac{\lvert m_1-m_0 \rvert}{\sigma}$ | $\sqrt{2}d' + e^{-\sqrt{2}d'} - 1$ | $\frac{d'}{\sqrt{2}} - \ln(1 + \frac{d'}{\sqrt{2}})$ | $\frac{1}{2}\mathcal{D}(p_1 \| p_0)$ |
| Poisson $r = \frac{\lambda_1}{\lambda_0}$ | $\lambda_0(1 - r + r \ln r)$ | $\lambda_0 \frac{(r-1)\left\{\ln \frac{r-1}{\ln r} - 1\right\} + \ln r}{\ln r}$ | $\lambda_0 \frac{(r+1)\ln r + (1-r) + \frac{r}{1-r}(\ln r)^2}{\ln r}$ |

**Table 1**: Analytic examples of distance calculations for three common probability distributions. The Kullback-Leibler distance calculated in the first column is $\mathcal{D}(p_1 \| p_0)$. The Gaussian and Laplacian examples differed in mean only. All of these need to be divided by $\ln 2$ to correspond to base-2 logarithms.

$q = \frac{1}{2}(p_0 + p_1)$. Simple manipulations show that $\mathcal{T}(p_0, p_1) \leq \min\{\mathcal{D}(p_0 \| p_1), \mathcal{D}(p_1 \| p_0)\}$. Computer experiments indicated that the Topsøe distance was less than or equal to the resistor-average distance; however, we could not demonstrate this relationship mathematically. The simulations also indicated that the Topsøe distance could be above or below the Bhattacharyya and the Chernoff distances. Consequently, it probably cannot be directly related to the exponential rates of any error

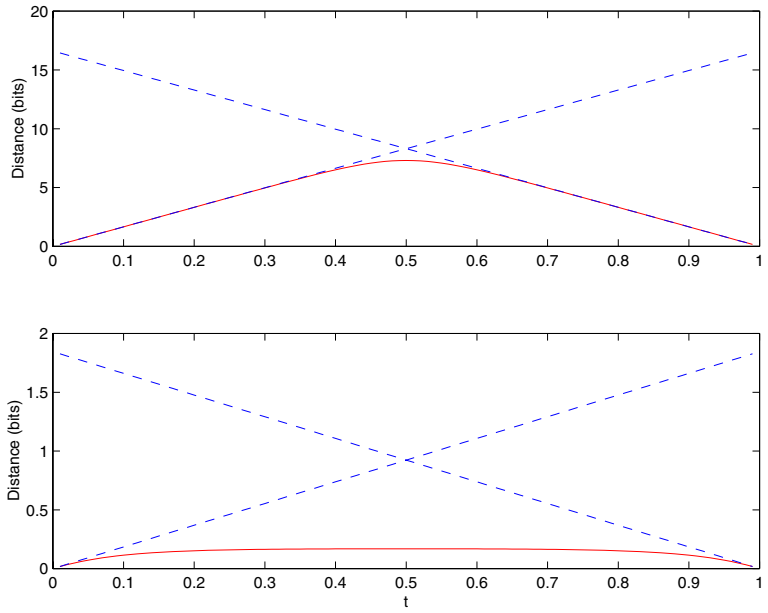


**Figure 2**: The plots show how different the resistor-average distance, located at the intersection of the lines, can be from the Chernoff distance. In the top panel, the two probability distributions were discrete: $p_0 = [.9999\ .0001], p_1 = [.0001\ .9999]$. In the bottom panel, the two discrete distributions were each uniformly distributed over 10 letters save for being very small at one letter, a different one in each case.

probability.

## 3 Conclusions

This work began as a search for an easily calculated symmetric distance that had many of the properties of the more fundamental Kullback-Leibler distance. We prefer to use the resistor-average distance rather than the other options described here because it more accurately reflects the average error probability of an optimal classifier and can be calculated easily from the two choices for Kullback-Leibler distance. Because of its kinship to the Kullback-Leibler distance, it satisfies the three properties described previously and bounds the Chernoff and Kullback-Leibler distances, the exponential rates of the average error and false-alarm probabilities respectively. Despite its close relation to the Kullback-Leibler distance, the resistor-average distance is not in the Ali-Silvey class. One side benefit of the geometric formulation is that the easily computed value $t^{**}$ can be used to initialize the required optimization for computing the Chernoff distance.

The graphical depiction of Figure 1 concisely and accurately summarizes relations among all

the distance measures of concern here. The two Kullback-Leibler distances control how skewed the curve $-\log \mu(t)$ might be: Nearly equal values suggest little skew while different values imply skew. The value of $t^{**}$ reveals where the maximizer $t^*$ might be relative to $t = \frac{1}{2}$. The closer $t^{**}$ is to $\frac{1}{2}$, the more closely the Bhattacharyya distance is to the Chernoff distance. When skew occurs, the $J$-divergence certainly departs markedly from a proportional relation to the Chernoff distance. The resistor-average distance, being closely related to the tangents of the curve $-\log \mu(t)$, more systematically tracks the Chernoff distance. However, as shown in Figure 2, distribution pairs do exist where the curve closely approximates its tangents, making $\mathcal{R}(p_0, p_1) \approx \mathcal{C}(p_0, p_1)$, and where the curve is nearly constant, making $\mathcal{R}(p_0, p_1) \gg \mathcal{C}(p_0, p_1)$. Given the computational difficulties with the Chernoff distance, we use the resistor-average distance because it generally approximates how the Chernoff distance changes within a distribution class. Furthermore, the resistor-average and Bhattacharyya distances always bracket the Chernoff distance; together they can be used to approximate accurately the Chernoff distance.

## References


[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Stat. Soc.*, 28:131–142, 1966.

[2] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18:349–369, 1989.

[3] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[4] N.N. Čencov. *Statistical Decision Rules and Optimal Inference*, volume 14 of *Translations in Mathematics*. American Mathematical Society, Providence, RI, 1982.

[5] H. Chernoff. Measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.*, 23:493–507, 1952.

[6] H. Chernoff. Large-sample theory: Parametric case. *Ann. Math. Stat.*, 27:1–22, 1956.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[8] A.G. Dabak. *A Geometry for Detection Theory*. PhD thesis, Dept. Electrical & Computer Engineering, Rice University, Houston, TX, 1992.

[9] C. M. Gruner. *Quantifying Information Coding Limits in Sensory Systems*. PhD thesis, Dept. Electrical & Computer Engineering, Rice University, Houston, Texas, 1998.

[10] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A*, 186:453–461, 1946.

[11] D. H. Johnson and C. M. Gruner. Information-theoretic analysis of neural coding. In *ICASSP '98*, Seattle, WA, 1998.

[12] D. H. Johnson and C. M. Gruner. Information-theoretic analysis of signal processing systems: Application to neural coding. In *Inter. Sympos. Information Theory*, MIT, Cambridge, MA, 1998.

[13] D. H. Johnson and G. C. Orsak. Relation of signal set choice to the performance of optimal non-Gaussian detectors. *IEEE Trans. Comm.*, 41:1319–1328, 1993.

[14] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. on Comm. Tech.*, COM-15(1):52–60, 1967.

[15] S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

[16] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

[17] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Info. Th.*, 46:1602–1609, 2000.