Inequalities for the L_1 Deviation of the Empirical Distribution

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi,
Sergio Verdu[1], Marcelo J. Weinberger
Information Theory Research Group
HP Laboratories Palo Alto
HPL-2003-97 (R.1)
June 18th , 2003*

E-mail: tsachy@stanford.edu, eord@hpl.hp.com, seroussi@hpl.hp.com, verdu@princeton.edu, marcelo@hpl.hp.com

Sanov's theorem, Pinsker's inequality, large deviations, L_1 distance, divergence, variational distance, Chernoff bound

We derive bounds on the probability that the L_1 distance between the empirical distribution of a sequence of independent identically distributed random variables and the true distribution is more than a specified value. We also derive a generalization of Pinsker's inequality relating the L_1 distance to the divergence.

* Internal Accession Date Only
[1] Princeton University, Princeton, NJ 08544

Approved for External Publication

Inequalities for the L_1 Deviation of the Empirical Distribution

Tsachy Weissman[*], Erik Ordentlich[†], Gadiel Seroussi[†], Sergio Verdu[‡], Marcelo J. Weinberger[†]

June 13, 2003

Abstract

We derive bounds on the probability that the L_1 distance between the empirical distribution of a sequence of independent identically distributed random variables and the true distribution is more than a specified value. We also derive a generalization of Pinsker's inequality relating the L_1 distance to the divergence.

Key words: Sanov's theorem, Pinsker's inequality, large deviations, L_1 distance, divergence, variational distance, Chernoff bound.

1 Preliminaries

Let $\mathcal{A}$ denote the finite set $\{1, \ldots, a\}$. For two probability distributions P and Q on $\mathcal{A}$ let

$$\|P - Q\|_1 = \sum_{k=1}^{a} |P(k) - Q(k)|$$

denote the variational, or L_1, distance between P and Q. For a sequence of symbols $\mathbf{x^m} = x_1, \ldots, x_m \in \mathcal{A}^m$, let $\hat{P}_{\mathbf{x^m}}$ be the empirical probability distribution on $\mathcal{A}$ defined by

$$\hat{P}_{\mathbf{x^m}}(j) = \frac{1}{m} \sum_{i=1}^{m} 1(x_i = j), \tag{1}$$

[*]Hewlett-Packard Laboratories, 1501 Page Mill Rd., Palo Alto, CA 94304 and Stanford University, Stanford, CA 94305, tsachy@stanford.edu.

[†]Hewlett-Packard Laboratories, 1501 Page Mill Rd., Palo Alto, CA 94304, eord@hpl.hp.com, seroussi@hpl.hp.com, marcelo@hpl.hp.com.

[‡]Princeton University, Princeton, NJ 08544, verdu@princeton.edu.

where $1(\cdot)$ denotes the indicator function of the specified event.

For probability distributions P and Q on $\mathcal{A}$, let

$$D(P\|Q) = \sum_{k=1}^{a} P(k) \log \frac{P(k)}{Q(k)} \tag{2}$$

denote the divergence between P and Q, where throughout $\log(\cdot)$ denotes the natural logarithm.

For $0 \leq p_1, p_2 \leq 1$ let

$$D_B(p_1\|p_2) = p_1 \log \frac{p_1}{p_2} + (1-p_1) \log \frac{1-p_1}{1-p_2} \tag{3}$$

denote the binary divergence, whereas for $(p_1, p_2) \notin [0,1]^2$ we set $D_B(p_1\|p_2) = \infty$.

The following conventions implied by continuity are adopted: for $c > 0$, $c/0 = \infty$, $c/\infty = 0$, $c\infty = \infty$, $\log \infty = \infty$, $e^{-\infty} = 0$. Additionally, in (2) and (3) it is assumed that $0 \log(0/0) = 0$ and $0 \log 0 = 0$.

For $p \in [0, 1/2)$, we define

$$\varphi(p) = \frac{1}{1-2p} \log \frac{1-p}{p} \tag{4}$$

and, by continuity, set $\varphi(1/2) = 2$.

For a probability distribution P on $\mathcal{A}$, we define

$$\pi_P = \max_{A \subseteq \mathcal{A}} \min(P(A), 1 - P(A)). \tag{5}$$

Note that $\pi_P \leq 1/2$ for any P.

Finally, throughout we take the minimum of a function over an empty set to be ∞.

2 Results

In this work, we prove the following theorem on the probability that the L_1 distance between the empirical distribution of a sequence of independent identically distributed random variables and the true distribution is more than a specified value.

Theorem 2.1 *Let P be a probability distribution on the set $\mathcal{A} = \{1, \ldots, a\}$. Let $\mathbf{X^m} = X_1, X_2, ..., X_m$ be independent identically distributed random variables distributed according to*

P. Then, for all $\epsilon > 0$,

$$-\lim_{m\to\infty} \frac{1}{m} \log Pr(\|P - \hat{P}_{\mathbf{X^m}}\|_1 \geq \epsilon) = \min_{A\subseteq\mathcal{A}} D_B\left(P(A) + \frac{\epsilon}{2} \,\Big\|\, P(A)\right). \tag{6}$$

Additionally,

$$\begin{aligned} Pr(\|P - \hat{P}_{\mathbf{X^m}}\|_1 \geq \epsilon) &\leq (2^a - 2)e^{-m[\min_{A\subseteq\mathcal{A}} D_B(P(A)+\epsilon/2\|P(A))]} &(7)\\ &\leq (2^a - 2)e^{-m\varphi(\pi_P)\epsilon^2/4}. &(8)\end{aligned}$$

We also strengthen Pinsker's inequality ([3], Problem 3.17; [2], Lemma 12.6.1), relating the L_1 distance to the divergence, as follows.

Theorem 2.2 *Let P and Q be two probability distributions on the set $\mathcal{A} = \{1, \ldots, a\}$. Then*

$$\|P - Q\|_1 \leq 2\sqrt{\frac{D(P\|Q)}{\varphi(\pi_Q)}}. \tag{9}$$

Theorems 2.1 and 2.2 are discussed in Section 3 and proved in Section 4.

3 Discussion

Pinkser's inequality states that

$$\|P - Q\|_1 \leq \sqrt{2D(P\|Q)}. \tag{10}$$

Theorem 2.2 strengthens (10) since, by Proposition 4.3 below, $\varphi(\pi_Q) \geq 2$ with equality if and only if $\pi_Q = 1/2$.

The method-of-types argument underlying the proof of Sanov's Theorem ([2], Theorem 12.4.1), and Pinsker's inequality, can be used to directly derive the well-known bound for the L_1 norm

$$\begin{aligned} Pr(\|P - \hat{P}_{\mathbf{X^m}}\|_1 \geq \epsilon) &\leq (n+1)^{a-1}e^{-m[\min_{P':\|P'-P\|_1\geq\epsilon} D(P'\|P)]} &(11)\\ &\leq (n+1)^{a-1}e^{-m\epsilon^2/2}. &(12)\end{aligned}$$

It follows from Sanov's Theorem and (6) in Theorem 2.1 that the exponent in (11) is the same as in (7). A minor advantage of (7) is the improved factor multiplying the exponential decay, i.e., it avoids the polynomial factor of (11).[1] In comparing (8) and (12), we see that (8) not only retains the improved factor multiplying the exponential decay, but, more importantly, may also have a faster exponential decay. Again, this follows from the fact that $\varphi(\pi_P) \geq 2$ with equality if and only if $\pi_P = 1/2$ (Proposition 4.3).

One may wonder why weakening (7) to (8) is useful at all. In many situations, the complicated dependence on ϵ of the exponent in (7) leads to an intractable analysis. The usual simplifying step involves applying the Pinsker, Hoeffding, or Bernstein inequalities (see, e.g., [1]). The Pinsker and Hoeffding inequalities have the disadvantage of dropping the dependence on the underlying distribution by assuming a worst case behavior. Bernstein's inequality, on the other hand, is useful only for small ϵ. The bound (8) avoids these disadvantages by scaling the ϵ^2 form of the exponents of the Hoeffding and Pinsker approaches with the optimum distribution-dependent factor.

One concrete application of this tool is in the proof of Theorem 2 of [4], in which numerous bounds on $Pr(\|P - \hat{P}_{\mathbf{X^m}}\|_1 \geq \epsilon)$ for different P's and ϵ's must be aggregated through a series of union bounds and optimizations over the constituent ϵ's. The complicated dependence on ϵ of the exponent in (7) makes the aggregation of bounds of this form prohibitively complex. The simpler exponent of (8), on the other hand, leads to a tractable analysis. The presence of $\varphi(\pi_P)$ in the exponent of (8) allows the final aggregated exponent to retain a dependence on the constituent distributions P (unlike (11)), which in [4] correspond to the per input channel output probabilities of a discrete memoryless channel.

It should be noticed that other refinements of Pinsker's inequality considered in the literature (see [5], [6, Corollary 1.4], and references therein) are still independent of the underlying distribution. Rather, the idea is to add higher powers of $\|P - Q\|_1$ to the lower bound on $D(P\|Q)$ (e.g., one such result states that $D(P\|Q) \geq (1/2)\|P - Q\|_1^2 + (1/36)\|P - Q\|_1^4$).

[1] In the binary case, the Chernoff bounding technique, as opposed to the method-of-types, also yields the multiplicative factor of (7). In fact, the arguments behind our proof of (7) can be viewed as extending the Chernoff bounding technique to the non-binary case.

4 Proofs of main theorems

The proofs of Theorems 2.1 and 2.2 will make use of the following four propositions, which are proved in Appendix A.

Proposition 4.1 *For* $p \in [0, 1/2]$*,* $D_B(p + \epsilon \| p) \leq D_B(1 - p + \epsilon \| 1 - p)$*.*

Proposition 4.2 *For* $p \in [0, 1/2]$

$$\inf_{\epsilon \in (0, 1-p]} \frac{D_B(p + \epsilon \| p)}{\epsilon^2} = \varphi(p).$$

Proposition 4.3 *The function* $\varphi(p)$ *is strictly decreasing for* $p \in [0, 1/2]$*.*

Proposition 4.4 *For all distributions* Q *on* $\mathcal{A} = \{1, \ldots, a\}$*,*

$$\min_{P' : \|P' - Q\|_1 \geq \epsilon} D(P' \| Q) = \min_{A \subseteq \mathcal{A}} D_B\left(Q(A) + \frac{\epsilon}{2} \,\Big\|\, Q(A)\right).$$

Proof of Theorem 2.1: By Sanov's Theorem (cf., e.g., Equation (12.96) of Theorem 12.4.1 in [2]),

$$\begin{aligned} -\lim_{m \to \infty} \frac{1}{m} \log Pr(\|P - \hat{P}_{\mathbf{X^m}}\|_1 \geq \epsilon) &= \min_{P' : \|P' - P\|_1 \geq \epsilon} D(P' \| P) \\ &= \min_{A \subseteq \mathcal{A}} D_B\left(P(A) + \frac{\epsilon}{2} \,\Big\|\, P(A)\right), \qquad (13) \end{aligned}$$

where (13) follows from Proposition 4.4, proving (6).

To prove (7) and (8), we start with the well-known fact that for any distribution Q on $\mathcal{A}$

$$\|Q - P\|_1 = 2 \max_{A \subseteq \mathcal{A}} (Q(A) - P(A)), \qquad (14)$$

which, together with a union bound, implies that

$$Pr\left(\|\hat{P}_{\mathbf{X^m}} - P\|_1 \geq \epsilon\right) \leq \sum_{A \subseteq \mathcal{A}} Pr\left(\hat{P}_{\mathbf{X^m}}(A) - P(A) \geq \frac{\epsilon}{2}\right). \qquad (15)$$

For $\epsilon > 0$ and $A = \mathcal{A}, \emptyset$, clearly $Pr(\hat{P}_{\mathbf{X^m}}(A) - P(A) \geq \epsilon/2) = 0$. For the other subsets of $\mathcal{A}$, the standard Chernoff bounding technique applied to the binary random variable $1(X_i \in A)$ shows that

$$Pr\left(\hat{P}_{\mathbf{X^m}}(A) - P(A) \geq \frac{\epsilon}{2}\right) \leq e^{-mD_B(P(A)+\epsilon/2\|P(A))}. \tag{16}$$

Combining (16) with (15) results in

$$\begin{aligned} Pr\left(\|\hat{P}_{\mathbf{X^m}} - P\|_1 \geq \epsilon\right) &\leq \sum_{A \subset \mathcal{A}: A \neq \mathcal{A}, \emptyset} e^{-mD_B(P(A)+\epsilon/2\|P(A))} \\ &\leq (2^a - 2)e^{-m[\min_{A\subseteq\mathcal{A}} D_B(P(A)+\epsilon/2\|P(A))]}, \end{aligned}$$

proving (7).

Now,

$$\begin{aligned} &\min_{A\subseteq\mathcal{A}} D_B\left(P(A) + \frac{\epsilon}{2} \,\Big\|\, P(A)\right) \\ &= \min_{A\subseteq\mathcal{A}} D_B\left(\min(P(A), 1-P(A)) + \frac{\epsilon}{2} \,\Big\|\, \min(P(A), 1-P(A))\right) \quad (17) \\ &\geq \frac{\min_{A\subseteq\mathcal{A}} \varphi(\min(P(A), 1-P(A)))\epsilon^2}{4} \quad (18) \\ &= \frac{\varphi(\pi_P)\epsilon^2}{4}, \quad (19) \end{aligned}$$

where (17), (18), and (19) follow, respectively, from Proposition 4.1, Proposition 4.2 (the case $\epsilon/2 > 1 - \min(P(A), 1 - P(A))$ follows by our conventions), and Proposition 4.3 and (5), completing the proof of (8). □

Proof of Theorem 2.2: We have

$$\begin{aligned} D(P\|Q) &\geq \min_{P': \|P'-Q\|_1 \geq \|P-Q\|_1} D(P'\|Q) \\ &= \min_{A\subseteq\mathcal{A}} D_B\left(Q(A) + \frac{\|P-Q\|_1}{2} \,\Bigg\|\, Q(A)\right) \quad (20) \\ &\geq \frac{\varphi(\pi_Q)\|P-Q\|_1^2}{4}, \quad (21) \end{aligned}$$

where (20) follows from Proposition 4.4 and (21) follows from Propositions 4.1, 4.2, and 4.3, as in equations (17) through (19) in the proof of Theorem 2.1. The proof is completed by rearranging terms. □

Appendix

A Proofs of propositions

Proof of Proposition 4.1: If $p < \epsilon$, the result holds by the conventions supporting the definition (3). For $p \geq \epsilon$, let

$$f(p,\epsilon) = D_B(p+\epsilon \| p) - D_B(1-p+\epsilon \| 1-p).$$

We show that $f(p,\epsilon) \leq 0$ for $p \in [\epsilon, 1/2]$ by showing below that $\partial f/\partial p|_{p=1/2} \geq 0$ and $\partial^2 f/\partial p^2 \leq 0$ for $p \in (\epsilon, 1/2]$. The claim then follows for $p \in (\epsilon, 1/2]$ since $f(1/2, \epsilon) = 0$. The claim for $p = \epsilon$ follows by continuity.

By definition,

$$\begin{aligned} f(p,\epsilon) = (p+\epsilon)\log\frac{p+\epsilon}{p} + (1-p-\epsilon)\log\frac{1-p-\epsilon}{1-p} \\ -(1-p+\epsilon)\log\frac{1-p+\epsilon}{1-p} - (p-\epsilon)\log\frac{p-\epsilon}{p}. \end{aligned}$$

Differentiating we obtain

$$\frac{\partial f}{\partial p} = \log(p+\epsilon) - \log(1-p-\epsilon) + \log(1-p+\epsilon) - \log(p-\epsilon) - \frac{2\epsilon}{p} - \frac{2\epsilon}{1-p}, \tag{A.1}$$

and

$$\begin{aligned} \frac{\partial^2 f}{\partial p^2} &= \frac{1}{p+\epsilon} - \frac{1}{p-\epsilon} - \frac{1}{1-p+\epsilon} + \frac{1}{1-p-\epsilon} + \frac{2\epsilon}{p^2} - \frac{2\epsilon}{(1-p)^2} \\ &= 2\epsilon\left[-\frac{1}{p^2-\epsilon^2} + \frac{1}{(1-p)^2-\epsilon^2} + \frac{1}{p^2} - \frac{1}{(1-p)^2}\right] \\ &= 2\epsilon\left((1-p)^2 - p^2\right)\left[\frac{1}{p^2(1-p)^2} - \frac{1}{(p^2-\epsilon^2)((1-p)^2-\epsilon^2)}\right]. \end{aligned} \tag{A.2}$$

From (A.2) we see that $\partial^2 p/\partial f^2 \leq 0$ for $p \in (\epsilon, 1/2]$, as claimed above, since $x(1-x)$ is non-decreasing for $x \in [0, 1/2]$.

From (A.1) we have

$$\frac{\partial f}{\partial p}\Big|_{p=1/2} = 2\log(1+2\epsilon) - 2\log(1-2\epsilon) - 8\epsilon$$

so that $\partial f/\partial p|_{p=1/2,\epsilon=0} = 0$. Differentiating with respect to ϵ we have

$$\begin{aligned}\frac{\partial^2 f}{\partial\epsilon\partial p}\Big|_{p=1/2} &= \frac{4}{1+2\epsilon} + \frac{4}{1-2\epsilon} - 8 \\ &= \frac{8}{1-4\epsilon^2} - 8,\end{aligned}$$

which is non-negative for $\epsilon \in [0, 1/2)$. This implies that $\partial f/\partial p|_{p=1/2} \geq 0$ for all $\epsilon \in [0, 1/2)$, again, as claimed above. □

Proof of Proposition 4.2: For a given $p \in [0, 1/2]$, consider the function

$$\begin{aligned}f(\epsilon) &= D_B(p+\epsilon\|p) - \varphi(p)\epsilon^2 \\ &= (p+\epsilon)\log\frac{p+\epsilon}{p} + (1-p-\epsilon)\log\frac{1-p-\epsilon}{1-p} - \frac{\epsilon^2}{1-2p}\log\frac{1-p}{p}.\end{aligned}$$

Let $f'(\epsilon)$ be the derivative of $f(\epsilon)$. We will show below that:

1. $f'(\epsilon) \geq 0$ for $\epsilon \in [0, 1/2 - p]$.
2. $f'(\epsilon) \leq 0$ for $\epsilon \in [1/2 - p, 1 - 2p]$.
3. $f'(\epsilon) \geq 0$ for $\epsilon \in [1 - 2p, 1 - p]$.

The first property of $f'(\epsilon)$ and $f(0) = 0$, together, imply that $f(\epsilon) \geq 0$ for $\epsilon \in [0, 1/2 - p]$. Similarly, properties 2 and 3 and the readily verified $f(1 - 2p) = 0$ imply that $f(\epsilon) \geq 0$ for $[1/2 - p, 1 - p]$. Thus, $f(\epsilon) \geq 0$ with equality at $\epsilon = 1 - 2p$, which proves the proposition.

We now verify the above properties 1-3 of $f'(\epsilon)$. Let

$$g(\epsilon) \triangleq f'(\epsilon) = \log\frac{p+\epsilon}{p} - \log\frac{1-p-\epsilon}{1-p} - \frac{2\epsilon}{1-2p}\log\frac{1-p}{p}.$$

After differentiating we have

$$g'(\epsilon) = \frac{1}{p+\epsilon} + \frac{1}{1-p-\epsilon} - 2\varphi(p)$$

and

$$g''(\epsilon) = -\frac{1}{(p+\epsilon)^2} + \frac{1}{(1-p-\epsilon)^2}\ ,$$

from which we see that $g''(\epsilon) \leq 0$ for $\epsilon \in [0, 1/2 - p]$ and $g''(\epsilon) \geq 0$ for $\epsilon \in [1/2 - p, 1 - p]$. Thus, $g(\epsilon)$ is concave for $\epsilon \in [0, 1/2 - p]$, which, together with the fact that $g(0) = g(1/2 - p) = 0$, implies Property 1 above. Analogously, properties 2 and 3 follow from the convexity of $g(\epsilon)$ for $\epsilon \in [1/2-p, 1-p]$ and the fact that $g(1/2-p) = g(1-2p) = 0$. □

Proof of Proposition 4.3: Differentiating $\varphi(p)$ (see (4)) with respect to p yields

$$\varphi'(p) = \frac{1}{(1-2p)^2}\left[-\frac{1-2p}{1-p} - \frac{1-2p}{p} + 2\log\frac{1-p}{p}\right].$$

Thus, to show that $\varphi'(p) < 0$ for $p \in (0, 1/2)$, it suffices to show that

$$g(p) \triangleq -\frac{1-2p}{1-p} - \frac{1-2p}{p} + 2\log\frac{1-p}{p} < 0.$$

To this end, note that $g(1/2) = 0$ and that the derivative of $g(p)$ is

$$\begin{aligned} g'(p) &= -\frac{-2(1-p)+(1-2p)}{(1-p)^2} - \frac{-2p-(1-2p)}{p^2} - \frac{2}{1-p} - \frac{2}{p} \\ &= \frac{1}{(1-p)^2} + \frac{1}{p^2} - \frac{2}{(1-p)p} \\ &= \left[\frac{1}{1-p} - \frac{1}{p}\right]^2. \end{aligned}$$

In particular, $g'(p) > 0$ for $p \neq 1/2$. Continuity arguments complete the proof for $p = 1/2$ and $p = 0$. □

Proof of Proposition 4.4: The argument is similar to a step in the proof of Pinsker's inequality ([3], Problem 3.17; [2], Lemma 12.6.1). For distributions P and Q on $\mathcal{A}$ let $A(P,Q) = \{a \in \mathcal{A} : P(a) \geq Q(a)\}$. It is then not difficult to see that

$$\|P - Q\|_1 = 2(P(A(P,Q)) - Q(A(P,Q))). \tag{A.3}$$

Therefore

$$\begin{aligned} &\min_{P':\|P'-Q\|_1\geq\epsilon} D(P'\|Q) \\ &= \min_{A\subseteq\mathcal{A}}\left[\min_{P':\|P'-Q\|_1\geq\epsilon, A(P',Q)=A} D(P'\|Q)\right] \\ &\geq \min_{A\subseteq\mathcal{A}}\left[\min_{P':\|P'-Q\|_1\geq\epsilon, A(P',Q)=A} D_B(P'(A)\|Q(A))\right] && \text{(A.4)} \\ &= \min_{A\subseteq\mathcal{A}}\left[\min_{P':P'(A)-Q(A)\geq\epsilon/2, A(P',Q)=A} D_B(Q(A)+P'(A)-Q(A)\|Q(A))\right] && \text{(A.5)} \\ &= \min_{A\subseteq\mathcal{A}} D_B\left(Q(A)+\frac{\epsilon}{2}\;\Big\|\; Q(A)\right), && \text{(A.6)} \end{aligned}$$

where (A.4) follows from the data processing inequality ([3], Lemma 3.11), (A.5) follows from (A.3), and (A.6) follows from the fact that $D_B(p\|q)$ is non-decreasing in p for $p \geq q$, continuity, and, in the event that the set in (A.5) is empty, from our conventions.

The reverse inequality follows by letting

$$A^* = \arg\min_{A \subseteq \mathcal{A}} D_B\left(Q(A) + \frac{\epsilon}{2} \,\Big\|\, Q(A)\right).$$

If $Q(A^*) + \epsilon/2 > 1$ then $D_B(Q(A^*) + \epsilon/2 \| Q(A^*)) = \infty$, in which case the reverse inequality is trivial. If $Q(A^*) + \epsilon/2 \leq 1$, define the distribution P^* as

$$P^*(k) = \begin{cases} \frac{(Q(A^*)+\epsilon/2)Q(k)}{Q(A^*)} & \text{for } k \in A^* \\ \frac{(1-Q(A^*)-\epsilon/2)Q(k)}{1-Q(A^*)} & \text{for } k \in \mathcal{A} \setminus A^*, \end{cases}$$

and check that $D(P^*\|Q) = D_B(Q(A^*) + \epsilon/2 \| Q(A^*))$ and $\|P^* - Q\|_1 = \epsilon$. □

References

[1] L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition,* Springer-Verlag New York, Inc., 1996.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory,* Wiley, New York, 1991.

[3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems,* Akadémiai Kiadó, Budapest, 1981.

[4] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, M. J. Weinberger, "Universal Discrete Denoising: Known Channel", *HP Labs. Tech. Rep.*, HPL-2003-29, February 2003. Submitted to *IEEE Trans. Inform. Theory.*

[5] F. Topsøe, "Some Inequalities for Information Divergence and Related Measures of Discrimination", *IEEE Trans. Inform. Theory*, vol. IT-46, pp. 1602–1609, July 2000.

[6] F. Topsøe, "Bounds for Entropy and Divergence for Distributions over a Two-element Set", *J. of Inequalities in Pure and Applied Mathematics*, vol. 2, 2001.