Some inequalities for information divergence and related measures of discrimination

Flemming Topsøe
Department of Mathematics
University of Copenhagen

Abstract— **Inequalities which connect information divergence with other measures of discrimination or distance between probability distributions are used in information theory and its applications to mathematical statistics, ergodic theory and other scientific fields. We suggest new inequalities of this type, often based on underlying identities. As a consequence we obtain certain improvements of the well known Pinsker inequality. Our study depends on two measures of discrimination, called capacitory discrimination and triangular discrimination. The discussion contains references to related research and comparison with other measures of discrimination, e.g. Ali-Silvey-Csiszár divergences and, in particular, the Hellinger distance.**

Keywords— **Information divergence, Pinsker's inequality, capacitory discrimination, triangular discrimination, Hellinger distance.**

I. Introduction

We shall study two probability measures P and Q over a finite alphabet (for a more generel set–up, see the discussion, Section VI). The point probabilities corresponding to P and Q are denoted by (p_i), respectively (q_i). Apart from information divergence $D(P\|Q)$ and variational distance $V(P,Q)$, we consider two other measures of discrimination between P and Q, viz. *capacitory discrimination* given by $C(P,Q) = D(P\|M) + D(Q\|M)$ where $M = \frac{1}{2}(P+Q)$ and *triangular discrimination* given by

$$\Delta(P,Q) = \sum \frac{|p_i - q_i|^2}{p_i + q_i}.$$

We also consider a "directional" version of Δ, denoted Δ^*, and defined by

$$\Delta^*(P\|Q) = \sum_{k=0}^{\infty} 2^k \Delta(M_k, Q).$$

Here, $M_k = 2^{-k}P + (1-2^{-k})Q$.

As we shall see below, capacitory and triangular discrimination behave similarly. Indeed,

$$\tfrac{1}{2}\Delta(P,Q) \le C(P,Q) \le \log 2 \cdot \Delta(P,Q).$$

Via a general identity, this implies that

$$\tfrac{1}{2}\Delta^*(P\|Q) \le D(P\|Q) \le \log 2 \cdot \Delta^*(P\|Q).$$

Note that the inequality $D \ge \frac{1}{2}\Delta^*$ is a strengthening of Pinsker's inequality since $\Delta^* \ge V^2$.

The importance of inequalities for information divergence in mathematical statistics, information theory proper and other fields is well recognized. In particular, this is true for Pinsker's inequality.

II. Definitions and auxiliary results

By $A = \{a_i \mid 1 \le i \le n\}$ we denote an n-set, the *alphabet.* Distributions over A, always assumed to be probability distributions, are, typically, denoted by P, Q, Π, and the associated point probabilities are denoted by p_i, q_i and π_i, respectively. *Variational distance* (ℓ_1-distance) and *information divergence* (Kullback-Leibler divergence) are defined as usual:

$$V(P,Q) = \sum_{i=1}^{n} |p_i - q_i|,$$

$$D(P\|Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}.$$

Here, log denotes natural logarithm.

Information divergence is the basis for a measure of discrimination, called *capacitory discrimination*, which is defined by

$$C(P,Q) = D(P\|M) + D(Q\|M), \tag{1}$$

where $M = \frac{1}{2}(P+Q)$. Furthermore, we define *triangular discrimination* by

$$\Delta(P,Q) = \sum_{i=1}^{n} \frac{|p_i - q_i|^2}{p_i + q_i}. \tag{2}$$

A variant of this measure, depending on a natural number ν as parameter, and called *triangular discrimination of order* ν, will also be needed. It is defined by the equation

$$\Delta_\nu(P,Q) = \sum_{i=1}^{n} \frac{|p_i - q_i|^{2\nu}}{(p_i + q_i)^{2\nu - 1}}. \tag{3}$$

For $\nu = 1$ we are back to (2), i.e. $\Delta_1 = \Delta$. In the discussion we point out that Δ_ν is a power of a distance (a metric), but for the main results we do not need this fact.

Fix two probability distributions P and Q over A and consider the communication channel with two input letters, say 0 and 1, and with P and Q as the conditional distributions over A given 0 and 1, respectively. If (α, β) defines an input distribution ($\alpha, \beta \ge 0, \alpha + \beta = 1$), then the *information transmission rate*, here denoted $I(\alpha,\beta)$, is defined as usual, i.e.

$$I(\alpha,\beta) = \alpha D(P\|S) + \beta D(Q\|S) \tag{4}$$

where $S = \alpha P + \beta Q$ is the *output distribution induced by* (α, β). We see that

$$C(P,Q) = 2 \cdot I(\tfrac{1}{2}, \tfrac{1}{2}). \tag{5}$$

The *maximum transmission rate*, i.e. the capacity, is defined by

$$I_{\max} = \max_{(\alpha,\beta)} I(\alpha,\beta).$$

With any distribution Π over A, now conceived as a *predictor*, we associate a (*guaranteed*) *redundancy* $R(\Pi)$, defined by

$$R(\Pi) = \max(D(P\|\Pi), D(Q\|\Pi)). \tag{6}$$

The *minimum* (*guaranteed*) *redundancy* is the quantity

$$R_{\min} = \min_{\Pi} R(\Pi). \tag{7}$$

The above concepts are well known from the theory of universal coding and prediction. Here, we are dealing with a particularly simple instance of that theory. From the general theory we need a key fact, the *Gallager-Ryabko theorem*, cf. the discussion in Ryabko [34] for the history of this result, which tells us that

$$I_{\max} = R_{\min}. \tag{8}$$

We note that the "$\leq$-part" of (8), which is the part we need, follows immediately from a simple identity (called the *compensation identity* in [16]). In the present context this identity states that for any input distribution (α, β) and any predictor Π,

$$I(\alpha,\beta) + D(S\|\Pi) = \alpha D(P\|\Pi) + \beta D(Q\|\Pi) \tag{9}$$

where S is the output distribution induced by (α, β).

The elementary inequalities

$$\tfrac{1}{2}C \leq I_{\max} \leq R_{\min} \leq R(M) \leq C \tag{10}$$

serve as motivation for choosing the name "capacitory discrimination" for C. *Note:* Here, C stands for $C(P,Q)$; similar abbreviations are used in the sequal for $\Delta(P,Q), V(P,Q)$ and $D(P\|Q)$.

Regarding triangular discrimination [1] note that

$$\tfrac{1}{2}V^2 \leq \Delta \leq V. \tag{11}$$

The last inequality is trivial, and the first follows by writing Δ as the square of an ℓ_2-norm w.r.t. the probability measure $\frac{1}{2}(P+Q)$ and using the fact that this ℓ_2-norm dominates the corresponding ℓ_1-norm (for a generalization, see (36)).

III. A connection between capacitory and triangular discriminations

Our first observation displays a connection between capacitory and triangular discrimination (of varying orders).

Theorem 1: For any distributions P and Q over the alphabet A,

$$C(P,Q) = \sum_{\nu=1}^{\infty} \frac{1}{2\nu(2\nu-1)} \Delta_\nu(P,Q). \tag{12}$$

Proof: In addition to notation from Section II we introduce the following auxiliary quantities ($1 \leq i \leq n$):

$$m_i = \tfrac{1}{2}(p_i + q_i), \quad \varepsilon_i = |p_i - m_i|, \quad k_i = m_i/\varepsilon_i.$$

To simplify the exposition we assume that, for all i, $p_i \neq 0$, $q_i \neq 0$ and $p_i \neq q_i$ hold. Note that $1/k_i = |p_i - q_i|/(p_i + q_i)$ so that $0 < 1/k_i < 1$. We have:

$$\begin{aligned} C(P,Q) &= \sum_{i=1}^{n} p_i \log \frac{p_i}{m_i} + \sum_{i=1}^{n} q_i \log \frac{q_i}{m_i} \\ &= \sum_{i=1}^{n} \varepsilon_i \left((k_i+1)\log\frac{k_i+1}{k_i} + (k_i-1)\log\frac{k_i-1}{k_i} \right) \\ &= \sum_{i=1}^{n} \varepsilon_i \left(k_i \log\left(1 - \frac{1}{k_i^2}\right) \right. \\ &\quad \left. + \log\left(1+\frac{1}{k_i}\right) - \log\left(1 - \frac{1}{k_i}\right) \right) \\ &= \sum_{i=1}^{n} \varepsilon_i \sum_{\nu=1}^{\infty} \frac{1}{\nu(2\nu-1)} k_i^{-(2\nu-1)}. \end{aligned}$$

Reversing the order of summation gives (12). ∎

As a corollary we obtain:

Theorem 2: Capacitory and triangular discrimination behave similarly as they are connected by the inequalities

$$\tfrac{1}{2}\Delta(P,Q) \leq C(P,Q) \leq \log 2 \cdot \Delta(P,Q). \tag{13}$$

Proof: The first inequality follows by considering only the first term in (12). Noting that $\Delta_\nu(P,Q)$ decreases with increasing ν, we also deduce from (12) that

$$C(P,Q) \leq \sum_{\nu=1}^{\infty} \frac{1}{2\nu(2\nu-1)} \Delta_1(P,Q) = \log 2 \cdot \Delta(P,Q),$$

as claimed. We refer to the discussion for an alternative proof, not depending on Theorem 1. ∎

Thus, numerically, $0.50 \leq C/\Delta \leq 0.70$.

Combining with (10) and (11) we get:

$$\begin{aligned} \tfrac{1}{8}V^2 \leq \tfrac{1}{4}\Delta \leq \tfrac{1}{2}C \leq I_{\max} \leq R_{\min} \\ \leq C \leq \log 2 \cdot \Delta \leq \log 2 \cdot V \end{aligned}$$

IV. A connection between information divergence and capacitory discrimination

As in the previous sections we study two distributions P and Q over the n-letter alphabet A. With P and Q we associate the *successive midpoints* $(M_\nu)_{\nu\geq 0}$ *in the direction* Q which are given by $M_0 = P$ and $M_\nu = \frac{1}{2}(M_{\nu-1} + Q)$, i.e.

$$M_\nu = 2^{-\nu}P + (1 - 2^{-\nu})Q; \quad \nu \geq 0.$$

We intend to show that the following quantity:

$$\Delta^*(P\|Q) = \sum_{k=0}^{\infty} 2^k \Delta(M_k, Q).$$

behaves much like $D(P\|Q)$. Note that

$$\Delta^*(P\|Q) = \sum_{\nu=1}^{\infty}\sum_{i=1}^{n} \frac{|p_i - q_i|^2}{p_i + (2^\nu - 1)q_i}. \tag{14}$$

Theorem 3: Let P and Q be distributions over the alphabet A and consider the successive midpoints $(M_\nu)_{\nu\geq 0}$ in the direction Q. The following identity and inequalities hold:

$$D(P\|Q) = \sum_{\nu=0}^{\infty} 2^\nu C(M_\nu, Q), \tag{15}$$

$$\tfrac{1}{2}\Delta^*(P\|Q) \leq D(P\|Q) \leq \log 2 \cdot \Delta^*(P\|Q). \tag{16}$$

Proof: If $D(P\|Q) = \infty$, there exists i with $q_i = 0$ and $p_i > 0$. Then, by (13) and (14), $\sum_0^\infty 2^\nu C(M_\nu, Q) \geq \frac{1}{2}\sum_0^\infty 2^\nu \Delta(M_\nu, Q) = \infty$, hence (15) and (16) hold in this case.

Now assume that $D(P\|Q) < \infty$. From the compensation identity (9) with $\Pi = Q$ and $\alpha = \beta = \frac{1}{2}$ we get

$$D(P\|Q) = C(P,Q) + 2D(M\|Q). \tag{17}$$

Iterating this, we find, for every $k \geq 0$,

$$D(P\|Q) = \sum_{\nu=0}^{k} 2^\nu C(M_\nu, Q) + 2^{k+1} D(M_{k+1}\|Q). \tag{18}$$

A direct calculation shows that here, the last term tends to 0 as $k \to \infty$. Indeed,

$$2^k D(M_k\|Q) = \sum_{i=1}^{n} \left(\frac{(p_i - q_i)^2}{q_i} 2^{-k} + (p_i - q_i) \right)$$

[1] the motivation behind the chosen terminology is connected with the *triangular net* in $\mathbb{R}_+^n$ which consists of all points of the form $(t_{j_1}, \ldots, t_{j_n})$ where the j_i's are non-negative integers and $t_j = \frac{1}{2}j(j+1)$, the j'th *triangular number*. Then note that any set of neighbouring points in the triangular net are unit-distance apart measured by Δ, i.e. $\Delta(x,y) = 1$ for any such set of points (this requires an extension of the definition (2) to arbitrary points in $\mathbb{R}_+^n$).

$$\cdot \frac{\log(1+2^{-k\frac{p_i-q_i}{q_i}})}{2^{-k\frac{p_i-q_i}{q_i}}}$$

$$\to \sum_{i=1}^{n}(p_i-q_i)=0.$$

With this auxiliary result at hand, (15) follows readily from (18) and then, (16) follows from (13). ∎

Pinsker's inequality $D \geq \frac{1}{2}V^2$ is an immediate consequence of the left hand side inequalities of (11) and (16). Indeed, as $\Delta \geq \frac{1}{2}V^2$,

$$D(P\|Q) \geq \tfrac{1}{2}\sum_{\nu=0}^{\infty} 2^{\nu}\tfrac{1}{2}(V(M_\nu,Q))^2 = \tfrac{1}{2}V(P,Q)^2. \tag{19}$$

Due to the fact that D may be large, even infinite, while C, Δ and V are small, upper bounds for information divergence in terms of discrimination measures such as those discussed in this correspondance cannot hold without introducing extra terms.

Theorem 4: Let P and Q be distributions over A and put $c = \max_i(p_i/q_i)$. Then

$$D(P\|Q) \leq C(P,Q) + \log\left(\tfrac{1}{2}(1+c)\right) \tag{20}$$

Proof: Put $d = \frac{1}{2}(1+c)$. Then $p_i/q_i \leq dp_i/m_i$ for all i (again, $m_i = \frac{1}{2}(p_i+q_i)$). Therefore,

$$\begin{aligned} D(P\|Q) &\leq \sum_{i=1}^{n} p_i \log\frac{dp_i}{m_i} = D(P\|M) + \log d \\ &\leq C(P,Q) + \log d. \end{aligned}$$

∎

By previous results, similar bounds for triangular discrimination and for variational distance holds, e.g.

$$\begin{aligned} D(P\|Q) &\leq \log 2\cdot V(P,Q) + \log\left(\tfrac{1}{2}(1+c)\right) \\ &\leq \log 2\cdot V(P,Q) + \log c. \end{aligned} \tag{21}$$

Note that (33) of the discussion is a strengthening of both (20) and (21).

V. Some refinements

We shall focus on the lower bounds $C \geq \frac{1}{4}V^2$ and $D \geq \frac{1}{2}V^2$ (Pinsker's inequality). These inequalities can be conceived as giving only the first term in an infinite expansion. It appears to be more natural to use half the variational distance rather than variational distance itself as parameter. Choosing the parameter this way, it varies in the unit interval.

Theorem 5: For any distributions P and Q over the alphabet A,

$$C(P,Q) \geq \sum_{\nu=1}^{\infty} a_\nu (\tfrac{1}{2}V(P,Q))^{2\nu} \tag{22}$$

where the constants a_ν, all positive, are given by

$$a_\nu = \frac{2}{2\nu(2\nu-1)}; \quad \nu \geq 1. \tag{23}$$

Equality holds in (22) if and only if, either $P = Q$ or else, P and Q are supported by the same 2-set and are symmetrically placed (in the sense that $\frac{1}{2}(P+Q)$ is the uniform distribution over the 2-set in question).

The constants a_ν are best possible in the natural sense (cf. below).

Proof: We shall apply the datareduction inequality for information divergence (cf. Kullback and Leibler [24] or Csiszár [8] and Csiszár and Körner [10]). For another proof, see the discussion.

Let ∂ denote the reduction defined by decomposing A into $A^+ = \{i|p_i \geq q_i\}$, and $A^- = \{i|p_i < q_i\}$.

By this reduction, P and Q are replaced by ∂P and ∂Q, respectively, where these measures are concentrated on the 2-set $\partial A = \{A^+, A^-\}$ and have point masses $P(A^+), P(A^-)$ and $Q(A^+), Q(A^-)$, respectively. Note, firstly, that this reduction leaves the variational distance unchanged: $V(P,Q) = V(\partial P, \partial Q)$ and, secondly, that the two pairs (P,M) and (M,Q) (with $M = \frac{1}{2}(P+Q)$ as usual) lead to the same reduction as that defined by the pair (P,Q).

From Theorem 1 we can now conclude that

$$\begin{aligned} C(P,Q) \geq C(\partial P,\partial Q) &= \sum_{\nu=1}^{\infty}\frac{1}{2\nu(2\nu-1)}\Delta_\nu(\partial P,\partial Q) \\ &= \sum_{\nu=1}^{\infty}\frac{(\frac{1}{2}V(P,Q))^{2\nu}}{2\nu(2\nu-1)}\left(\frac{1}{(P(A^+)+Q(A^+))^{2\nu-1}}\right. \\ &\qquad \left. + \frac{1}{(P(A^-)+Q(A^-))^{2\nu-1}}\right) \\ &\geq 2\sum_{\nu=1}^{\infty}\frac{(\frac{1}{2}V(P,Q))^{2\nu}}{2\nu(2\nu-1)} \end{aligned}$$

which is the result stated. Note that the last inequality above follows from elementary considerations as it is easy to show that, for $n \in \mathbb{N}$, the minimal value of $x^{-n} + y^{-n}$, where x and y are non-negative numbers with sum 2, is 2, which is attained for $x = y = 1$ (and for no other values of x and y).

Really, the fact stated concerning equality in (22), follows by inspection of the above proof.

Concerning the last statement of the theorem, we first define the *best constants* $a_\nu^{\max}$ for an inequality like (22). This is done by recursion: For each $\nu_0 \geq 1$, $a_{\nu_0}^{\max}$ is defined as the maximum over all a_{ν_0} where a_{ν_0} fits into an inequality of the form (22) with $a_\nu = a_\nu^{\max}$ for all $\nu < \nu_0$.

The proof that $a_\nu = a_\nu^{\max}$ for all $\nu \geq 1$ with a_ν given by (23) is then carried out by induction: Let $\nu_0 \geq 1$ and assume that $a_\nu = a_\nu^{\max}$ for $\nu < \nu_0$. By (22), $a_{\nu_0}^{\max} \geq a_{\nu_0}$. To prove the reverse inequality, note that by the result proved regarding instances for which equality holds in (22) it follows that for all $0 \leq x \leq 1$,

$$\sum_{\nu=1}^{\infty} a_\nu x^{2\nu} \geq \sum_{\nu=1}^{\nu_0} a_\nu^{\max} x^{2\nu}.$$

This implies that for all $0 < x \leq 1$,

$$(a_{\nu_0} - a_{\nu_0}^{\max}) + \sum_{\nu>\nu_0} a_\nu x^{2(\nu-\nu_0)} \geq 0,$$

and $a_{\nu_0} \geq a_{\nu_0}^{\max}$ follows. Thus $a_{\nu_0} = a_{\nu_0}^{\max}$ must hold. ∎

Note that the discussion of Section VI contains a strengthening of the inequality (22).

By summing the infinite series in (22) we obtain a corollary which in essense is equivalent to (22). Indeed, putting $x = \frac{1}{2}V$, one finds that

$$\begin{aligned} C(P,Q) &\geq (1+x)\log(1+x) + (1-x)\log(1-x) \\ &= 2\cdot D((\alpha,\beta)\|(\tfrac{1}{2},\tfrac{1}{2})), \end{aligned} \tag{24}$$

with $\alpha = \frac{1}{2}(1+x) = \frac{1}{2} + \frac{1}{4}V$ and $\beta = \frac{1}{2}(1-x) = \frac{1}{2} - \frac{1}{4}V$. By combining with (15) one further finds that

$$D(P\|Q) \ge \sum_{\nu=1}^{\infty} b_\nu (\tfrac{1}{2}V(P,Q))^{2\nu} \tag{25}$$

with

$$b_\nu = \frac{2^{2\nu}}{2^{2\nu-1}-1} \cdot \frac{1}{2\nu(2\nu-1)}\,; \quad \nu \ge 1. \tag{26}$$

VI. Discussion

Capacitory discrimination can be viewed as a symmetrized and smoothed variant of information divergence. Furthermore, it has an interpretation as twice an information transmission rate, cf. (5). It is natural to compare capacitory discrimination with Jeffrey's measure of distance between two distributions given by $J(P,Q) = D(P\|Q) + D(Q\|P)$, cf. e.g. Csiszár and Körner [10] (Problem I.3.16). It is intuitively clear that $C \le J$. Even $C \le \frac{1}{2}J$ holds. In fact, by (17),

$$\frac{1}{2}J(P,Q) = C(P,Q) + \widetilde{C}(P,Q), \tag{27}$$

where $\widetilde{C}$ denotes the *dual capacitory distance* defined by

$$\widetilde{C}(P,Q) = D(M\|P) + D(M\|Q). \tag{28}$$

Though a little outside the scope of this correspondance, it is natural to note that capacitory discrimination can be expressed in terms of the entropy function H, even in two rather different ways. Indeed, from the following basic – but often overlooked – identity:

$$H(\sum \alpha_\nu P_\nu) = \sum \alpha_\nu H(P_\nu) + \sum \alpha_\nu D(P_\nu\|P^*),$$

where $P^* = \sum \alpha_\nu P_\nu$ is a convex combination, we find that

$$C(P,Q) = 2 \cdot H(M) - H(P) - H(Q).$$

But we may also rewrite C in the form

$$C(P,Q) = \log 4 - \sum (p_i + q_i) H\left(\frac{p_i}{p_i+q_i}, \frac{q_i}{p_i+q_i}\right)$$

and if we use this formula in conjunction with the following general inequalities

$$H(P) \ge \log 4 \cdot (1 - \sum p_i^2) \tag{29}$$

and

$$H(P) + \sum p_i^2 \le \log n + n^{-1}, \tag{30}$$

we are led to an alternative proof of Theorem 2. The inequality (29) has been announced on several occasions by the author, whereas (30) is new. Proofs of both inequalities, as well as of related ones, will be included in Harremoës and Topsøe [16], [17].

The introduction of capacitory discrimination $C(P,Q)$ is not new. It can even be said to be implicit in Shannon's original work due to the connection with concavity of the entropy function. In a somewhat sporadic form it occurs in You [41], Definition 13. More explicitly it is considered in Lin and Wong [27] where some simple properties are derived. This information is, essentially, repeted in Lin [26]. More detailed results were developed in Pecaric [32], in fact Pecaric developed the results of Theorems 1 and 2 independently of the author and of Boris Ryabko, cf. the acknowledgements. Pecaric used methods much resembling those presented here.

As demonstrated by Theorem 3, D is a weighted sum of C-type quantities, in particular, it follows that $C \le D$. This is also an immediate consequence of (17). However, the stronger inequality $C \le \log 2 \cdot D$ holds as will be shown in [17]. Here, $\log 2$ is the best, i.e. smallest, constant (consider $P = (1,0)$, $Q = (1-\varepsilon, \varepsilon)$). Inequalities of a similar type are the interrelated inequalities

$$2\log 2 \cdot D(P\|M) \le C(P,Q) \le \frac{2\log 2}{2\log 2 - 1} D(P\|M) \tag{31}$$

and

$$D(P\|M) \le \frac{1}{2\log 2 - 1} D(Q\|M). \tag{32}$$

They will also be proved in [17]. Again, it is easy to see that the constants in these inequalities are best possible.

Of course, (31) can be strengthened by replacing $D(P\|M)$ on the left hand side with the maximum of $D(P\|M)$ and $D(Q\|M)$ and by replacing $D(P\|M)$ on the right hand side with the minimum of $D(P\|M)$ and $D(Q\|M)$. While we are at it, we mention the more elementary inequality $D(P\|M) \le \frac{1}{2}D(P\|Q)$ (use convexity of $D(\cdot\|\cdot)$ in the second variable). The constant $\frac{1}{2}$ is best possible.

When we inspect the proof of Theorem 4, we see that (31) leads to the following inequalities (where D again stands for $D(P\|Q)$):

$$\begin{aligned} D &\le \frac{1}{2\log 2} C + \log\left(\frac{1}{2}(1 + \max_i \frac{p_i}{q_i})\right) \\ &\le \frac{1}{2} V + \log\left(\frac{1}{2}(1 + \max_i \frac{p_i}{q_i})\right) \end{aligned} \tag{33}$$

which give sharper bounds than those of (20) and (21).

Triangular discrimination arose in an attempt to sharpen a lower bound of ε-capacity given in Ryabko and Topsøe [35]. This theme may be pursued in a later publication. Here, the significance of the triangular discrimination measures lies in the strong connection with capacitory discrimination and information divergence (Theorems 1 and 3).

In fact, also triangular discrimination has occured before in the literature, perhaps for the first time in Vincze [40] where its statistical significance is briefly indicated. We also find triangular discrimination in LeCam [25] (cf. p.xvii and pp. 47–48) who refers to Δ as a "chi–square like distance", and LeCam indicates that this measure of discrimination was already used by Hellinger. However, the author has not been able to check up on this.

The importance for statistical research of discrimination measures like Δ_ν and, more generally, of arbitrary f-divergences (see below) is further discussed in Österreicher and Vajda [31]. For our purposes, the observation by Kafka, Österreicher and Vincze [18] (cf. also the announcement in Feldman and Österreicher [14]) that the triangular discrimination measures Δ_ν are all powers of metrics (indeed, $(\Delta_\nu)^{\frac{1}{2\nu}}$ is a metric, and the exponent here is the largest possible one) is most noteworthy. In particular, Δ is the square of a metric, a fact already noticed by LeCam [25]. Using identities and inequalities of Theorems 1–3, one can then relate information divergence to true metrics. This is of importance for (parts of) research as that of LeCam [25], Birgé [5], Birgé and Massart [6] and Yang and Barron [42], where the lack of metric properties for information divergence is compensated for by researching relations to true distances (typically then, the Hellinger distance is the preferred model metric).

Regarding further general properties of triangular discrimination, it was pointed out to the author in discussion with J.P.R. Christensen that the metric $\sqrt{\Delta}$ (even extended to cover other measures than probability measures) is isometric to a subset of Hilbert space. This property serves as yet another indication of the significance of triangular discrimination. Let us briefly indicate a proof of this fact. By Berg, Christensen and Ressel, [4], Chapter 3 (Proposition 3.2) it suffices to show that the kernel Δ is negative definite. So let there be given two finite sequences of real numbers, (c_i) and (p_i), and assume that $\sum c_i = 0$ and that all the p_i are positive. Then

$$\begin{aligned}\sum_{i,j} \frac{(p_i - p_j)^2}{p_i + p_j} c_i c_j &= -4 \sum_{i,j} \frac{p_i p_j}{p_i + p_j} c_i c_j \\ &= -4 \int_0^\infty \left(\sum_i c_i p_i e^{-tp_i}\right)^2 dt,\end{aligned}$$

a non-positive number. Thus Δ is negative definite as claimed.

Very recently, a family of discrimination measures given by the expression

$$\sum_i \frac{|p_i - q_i|^2}{bp_i + (1-b)q_i}$$

has been studied, cf. Györfi and Vajda [15]. These measures are of course closely related to triangular discrimination (Δ as well as Δ^*). A further recent and very detailed study which contains triangular discriminations of varying orders (Δ_ν) is given in Menéndez, Morales, Pardo and Vajda, [30].

Many of our results depend on the introduction of midpoints, in addition to the basic measures under study. Though related only weakly to results presented here, the reader may wish to consult Dembo [13], Proposition 1, where wider instances of qualitatively similar inequalities are proved in order to study the measure concentration problem.

The most important inequality discussed is without doubt Pinsker's inequality which has numerous applications. Some general references are Ahlswede and Wegener [2], Csiszár and Körner [10], Kullback [23] and Pinsker [33]. For more specific applications of Pinsker's inequality we can, for instance, point to Csiszár [11], Marton [28], [29], Ryabko and Topsøe [35] and Topsøe [36] for five different applications. The inequality originated with Pinsker [33] and was refined in Csiszár [8], cf. also Csiszár and Körner [10], Problem I.3.17. Note that our approach offers an alternative and rather elementary proof of Pinsker's inequality (for this remark note that the first inequality of (16) depends on (17) but does not need (15) for its proof).

The refined inequality for capacitory discrimination given in Theorem 5 is in a satisfactory form, but the corresponding result quoted for information divergence ((25) with constants as in (26)) is not, as the constants are not best possible. It is known that the best two-term inequality is $D \geq \frac{1}{2}V^2 + \frac{1}{36}V^4$ (Krafft [20]) and not $D \geq \frac{1}{2}V^2 + \frac{1}{84}V^4$ as one might think, considering (25). Let us briefly indicate a proof of this result (modifying the approach of Krafft [20]). We need only consider probability measures P and Q of the form $P = (\frac{1-\alpha}{2}, \frac{1+\alpha}{2})$, $Q = (\frac{1+\beta}{2}, \frac{1-\beta}{2})$. A straight forward expansion, cf. Kambo and Kotz [19], shows that

$$D(P\|Q) = \sum_1^\infty \frac{1}{2\nu(2\nu-1)} T_\nu(\alpha, \beta) \tag{34}$$

with the polynomials T_ν given by

$$T_\nu(\alpha, \beta) = \alpha^{2\nu} + 2\nu\alpha\beta^{2\nu-1} + (2\nu-1)\beta^{2\nu}. \tag{35}$$

Then $T_\nu \geq 0$ (evident if α and β are of the same sign, and, if not, factor out $(\alpha+\beta)^2$). Noting that $T_1 = (\alpha+\beta)^2 = V^2$ and that $T_2 = \frac{1}{3}V^2(V^2 + 2(\alpha - 2\beta)^2)$, it is easy to see that the two first best constants in an inequality like $D \geq \sum_1^\infty c_\nu V^{2\nu}$ are $c_1 = \frac{1}{2}$ and $c_2 = \frac{1}{36}$. For further details along these lines, we refer to Vajda [39], Krafft and Schmitz [21] and Toussaint [38]. The paper by Krafft and Schmitz essentially adds the term cV^6 with $c = \frac{1}{288}$, however, this is first stated explicitly in the paper by Toussaint. The author points out that the best term of the form cV^6 is obtained for $c = \frac{1}{270}$, a slight improvement over the result of Toussaint and predecessors. The author plans to return to the matter elsewhere.

The refined inequalities (22) and (25) can also be derived in a direct way from (12) and (15) by using the following inequalities which are of some independent interest:

$$\begin{aligned} 2^{-2\nu+1}V^{2\nu} &\leq 2^{-\nu+1}\Delta^\nu \\ &\leq \Delta_\nu \leq \Delta \leq V. \end{aligned} \tag{36}$$

The non-trivial parts follow as the numbers

$$\left(\tfrac{1}{2}\Delta_\nu(P,Q)\right)^{\frac{1}{2\nu}} = \left(\sum \left(\frac{|p_i - q_i|}{p_i + q_i}\right)^{2\nu} \left(\tfrac{1}{2}p_i + \tfrac{1}{2}q_i\right)\right)^{\frac{1}{2\nu}},$$

recognized as $\ell_{2\nu}$–norms w.r.t. a probability measure, are weakly increasing in ν (allowing also the value $\nu = \frac{1}{2}$ for which $\Delta_\nu = V$). Note that the constants in (36) are best possible as equality holds throughout for $P = (1,0), Q = (0,1)$.

The inequalities (36) can even be used to derive a strengthening of (22). Indeed, one finds that

$$C(P,Q) \geq \sum_{\nu=1}^\infty \frac{2^{-(\nu-1)}}{2\nu(2\nu-1)} \Delta(P,Q)^\nu, \tag{37}$$

again with best constans as coefficients (for this, observe that the discussion of equality in (37) is similar to the one in Theorem 5).

One may also derive a variant of (25) of the form $D \geq \sum c_\nu \cdot \Delta^\nu$, but a straight forward derivation gives constants far from the best possible ones. The most interesting inequality of this type is of the form $D \geq c \cdot \Delta$. Let c^{max} denote the largest such constant. From (16) and from (11) and Pinsker's inequality it follows that $\frac{1}{2} \leq c^{max} \leq 1$. But, in fact, $0.8437 \leq c^{\max} \leq 0.8961$ holds. The first inequality follows as one can show that, quite generally,

$$D \geq \frac{27}{32}\Delta$$

(indication of proof: Either argue directly via Lagrange multipliers or, more elegantly, reduce the problem to one involving a certain f-divergence). As to the second inequality, note that for each $0 < \rho < \infty$, $c^{\max}$ is bounded above by

$$\frac{\rho+1}{\rho-1}\left(1 - \frac{\log \rho}{\rho - 1}\right)$$

(consider $P = (1-\varepsilon, \varepsilon)$ and $Q = (1-\rho\varepsilon, \rho\varepsilon)$ for small ε's). The minimum of this quantity is attained for $\rho = \rho_0$ defined by $\rho_0 > 1$ and

$$\log \rho_0 = \frac{(\rho_0 - 1)(1 + 3\rho_0)}{\rho_0(3 + \rho_0)}$$

($\rho_0 \approx 4.447$). One finds that

$$c^{\max} \leq \frac{(\rho_0 + 1)^2}{\rho_0(3 + \rho_0)}$$

(≈ 0.8961). It is plausible that here, equality holds – numerical evidence points to this as a safe conjecture.

It may well be worth while to study the *directional modification* used for Δ in more general situations. Here we note only that the construction may also be introduced for C and Δ_ν:

$$C^*(P\|Q) = \sum_0^\infty 2^k C(M_k, Q), \Delta_\nu^*(P\|Q) = \sum_0^\infty 2^k \Delta_\nu(M_k, Q)$$

with $M_0, M_1, \cdots$ denoting the successive midpoints in direction Q, and then, from (12) and (15), we get:

$$D = C^* = \sum_1^\infty \frac{1}{2\nu(2\nu - 1)} \Delta_\nu^*. \tag{38}$$

The two inequalities of (13) are best possible in a natural sense: By considering $P = (1, 0)$ and $Q = (0, 1)$ it follows that $\log 2$ is the smallest possible constant in an inequality like $C \leq \alpha \cdot \Delta$. And, regarding the other inequality, we may either appeal to Pinsker's inequality or, more directly, we may consider the two functions $f(\varepsilon) = C(P, Q_\varepsilon)$ and $g(\varepsilon) = \Delta(P, Q_\varepsilon)$ with $P = (\frac{1}{2}, \frac{1}{2})$ and $Q_\varepsilon = (\frac{1}{2} + \varepsilon, \frac{1}{2} - \varepsilon)$ and note that $f(0) = f'(0) = g(0) = g'(0) = 0$ and that $f''(0) = 2$, $g''(0) = 4$. Hence $C(P, Q_\varepsilon)/\Delta(P, Q_\varepsilon) \to \frac{1}{2}$ as $\varepsilon \to 0$, and it follows that $\beta = \frac{1}{2}$ is the largest possible constant in an inequality of the form $\beta \cdot \Delta \leq C$.

The two inequalities of Theorem 3, i.e. $\frac{1}{2}\Delta^* \leq D \leq \log 2 \cdot \Delta^*$, are also best possible. That the constant $\frac{1}{2}$ is optimal follows since any larger constant would give rise to a strengthening of Pinsker's inequality beyond what is possible. It is more tricky to show that $c_{\min} = \log 2$ with $c_{\min}$ the smallest possible constant in an inequality of the form $D \leq c \cdot \Delta^*$ (note e.g., that it does not follow from the related inequality $C \leq \log 2 \cdot \Delta$ since $\Delta \leq D$ does not hold generally). Clearly, $c_{\min} \leq \log 2$. Furthermore, for each $0 < \rho < \infty$, we find that $c_{\min}$ is bounded below by

$$(1-\rho)^{-2} \cdot \left(-\log \rho - (1-\rho)) \cdot (\sum_1^\infty (1 - \rho + \rho \cdot 2^\nu)^{-1} \right)^{-1}$$

(consider $P = (1-\varepsilon, \varepsilon)$ and $Q = (1-\rho\varepsilon, \rho\varepsilon)$ for small ε's). Now, for $0 < \rho < 1$,

$$\sum_1^\infty (1 - \rho + \rho \cdot 2^\nu)^{-1} \leq \int_0^\infty (1 - \rho + \rho \cdot 2^t)^{-1} dt$$

$$= \left. \frac{1}{1-\rho} t - \frac{1}{(1-\rho)\log 2} \log(1 - \rho + \rho 2^t) \right]_0^\infty$$

$$\leq -\frac{\log \rho}{(1-\rho)\log 2},$$

hence

$$c_{\min} \geq \log 2 \cdot \left(\frac{1}{1-\rho} + \frac{1}{\log \rho} \right),$$

and by choosing ρ small, it follows that $c_{\min} \geq \log 2$, thus $c_{\min} = \log 2$ as claimed.

All measures of discrepancy considered in this correspondence are particular instances of Ali-Silvey-Csiszár divergences (below referred to simply as f-divergences), cf. Ali and Silvey [3], Csiszár [7], [8], [9]. Recall that for a convex function $f : [0, \infty[\to \mathbb{R}$, the *$f$-divergence* between P and Q is defined by

$$I_f(P, Q) = \sum q_i f(\frac{p_i}{q_i}) \tag{39}$$

(we need only instances with $f(0)$ finite; below we typically normalize so that $f(0) = 1$). The family of functions $(f_s)_{s \geq 1}$ with $f_s(u) = |u-1|^s \cdot (u+1)^{-(s-1)}$ gives rise to variational distance V ($s = 1$), triangular discrimination Δ ($s = 2$) and triangular discrimination of order ν, Δ_ν ($s = 2\nu$). And if we take $f(u) = (u-1)^2/(u + 2^{\nu+1} - 1)$, we are led to the map $(P, Q) \curvearrowright \Delta(M_\nu, Q)$ where, as usual, $M_\nu = 2^{-\nu}P + (1 - 2^{-\nu})Q$.

Among the most popular choices we mention $f(u) = \frac{1}{2}(\sqrt{u} - 1)^2$ which gives rise to the *Hellinger discrimination* h^2:

$$h^2(P, Q) = \tfrac{1}{2} \sum (\sqrt{p_i} - \sqrt{q_i})^2.$$

The square-root $h = \sqrt{h^2}$ is a true distance. Note that h^2, just as Δ, is negative definite (simple direct proof), thus h is a *hilbertian* metric, as is $\sqrt{\Delta}$.

The basic relations between V, Δ and h^2 are the following three: $\frac{1}{2}V^2 \leq \Delta \leq V$, cf. (11), then the relation $2h^2 \leq \Delta \leq 4h^2$ (derived e.g. by comparing the corresponding f-functions, cf. also LeCam [25] and Dacunha–Castelle [12]) and lastly, the relation $\frac{1}{8}V^2 \leq h^2 \leq \frac{1}{2}V$ (follows from the two first). The occuring coefficients are best possible as may be seen by considering the case $P = (1, 0), Q = (0, 1)$ or the case $P = (\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon), Q = (\frac{1}{2}, \frac{1}{2})$ for small values of ε.

Kraft [22] improved part of this by pointing out that $\frac{1}{8}V^2 \leq h^2(1 - \frac{1}{2}h^2)$ (follows by applying Cauchy–Schwartz' inequality). The inequalities between h^2 and Δ were also noted by LeCam [25].

In Dacunha–Castelle [12] we find the important inequality $D \geq -2\log(1 - h^2)$ (follows by Jensen's inequality), in particular, $D \geq 2h^2$. This is best possible as an inequality $D \geq ch^2$ cannot hold for any $c > 2$ (consider a large ρ and $P = (\varepsilon, 1-\varepsilon), Q = (\rho\varepsilon, 1 - \rho\varepsilon)$ for small $\varepsilon's$).

The above comments indicate that Δ and h^2 have similar properties. It may be true that Δ is closer to D in some sense (as our results have shown), but, on the other hand, h^2 has nice structural properties (discussed elsewhere) which are not shared by Δ. In conclusion then, triangular discrimination cannot offer a replacement of Hellinger distance, but does appear to provide a convenient supplement.

Those results of this correspondance which involve comparison of one measure of discrimination with another, especially those with best constants, go some way to establish a hierarchy among these measures. In this connection, the reader may want to consult Abrahams, [1].

Generalizations of identities and inequalities here presented (except those involving the entropy function) to a countably infinite alphabet or to distributions defined with reference to general measure spaces are, basically, straight forward and a matter of routine. This is important, e.g. if Δ will replace h^2 for certain investigations in statistics. However, one issue deserves special mention, viz. regarding the general validity of (15) under the condition $D(P\|Q) < \infty$. So assume that P and Q are probability measures on a general measurable space and that $D(P\|Q) < \infty$. In particular then, P is absolutely continuous w.r.t. Q. We put $M_\varepsilon = \varepsilon P + (1-\varepsilon)Q$; $0 < \varepsilon < 1$, and shall prove that

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} D(M_\varepsilon \| Q) = 0. \tag{40}$$

Put $\varphi = \frac{dP}{dQ}$. A simple computation shows that

$$\frac{1}{\varepsilon} D(M_\varepsilon\|Q) = \frac{1}{\varepsilon}\int \log \frac{dM_\varepsilon}{dQ} dM_\varepsilon = A + B + C$$

with

$$A = \frac{1}{\varepsilon}\log(1-\varepsilon), \quad B = \int \log(1 + \frac{\varepsilon}{1-\varepsilon}\varphi) dP,$$

$$C = \int \frac{1-\varepsilon}{\varepsilon} \log(1 + \frac{\varepsilon}{1-\varepsilon}\varphi) dQ.$$

Clearly, $A \to -1$ as $\varepsilon \to 0$. Concerning B, call the integrand f_ε, and note that f_ε decreases pointwise to 0 as ε decreases to 0 and that $f_{\frac{1}{2}} = \log(1+\varphi)$ is P-integrable (as $\log(1+x) \le 2\log x$ for $x \ge 2$ and as $D(P\|Q) < \infty$). It follows that $B \to 0$ as $\varepsilon \to 0$. Finally, concerning C, we observe that C is of the form $\int g_\varepsilon dQ$ with all $g_\varepsilon \le \varphi$ and that $g_\varepsilon \to \varphi$ hence, again by dominated convergence, $C \to 1$ as $\varepsilon \to 0$. All in all, $A + B + C \to 0$ as $\varepsilon \to 0$ and (40) follows.

Acknowledgements

The research presented arose in discussions with Boris Ryabko and I am much indepted to his suggestions. In particular, Theorem 1 is due to Ryabko. Just before finishing the first version of the manuscript, the author had the opportunity to discuss the findings at the workshop "Information theory, statistical learning and pattern recognition", CIRM, Marseille, december 1998. The exchange of ideas at the workshop with Andrew Barron, Lucien Birgé, Amir Dembo, Katalin Marton and Ofer Zeitouni resulted in improvements of the discussion. In particular, the proof of (40) was suggested by Ofer Zeitouni.

Some comments on the improvements of the manuscript which occured between the first and the final submission are in order. Apart from sending the first manuscript to colleagues and friends, it was posted on the RGMIA collection of research reports, cf.[37]. This resulted in detailed comments from Josip Pecaric, [32]. As mentioned already, main points of [32] are independent derivations of Theorems 1 and 2 but also relevant references ([26] and [27]) which had been overlooked by the author. Useful feedback was also received from Julia Abrahams, Jens Peter Reus Christensen and my co-worker Peter Harremoës. Finally, thanks goes to both referees for their encouragement and comments, e.g. regarding [30] and [15] which were not known to the author.

References

[1] J. Abrahams, *On the Selection of Measures of Distance between Probability Distributions*, Proceedings of the IEEE International Symposium on Information Theory, 1981, pp. 109–113,Elsevier, New York, 1982.

[2] R. Ahlswede, and I. Wegener, *Search problems*, Chichester: Wiley, 1987. (German original by Teubner, 1979).

[3] S.M. Ali, and S.D. Silvey,*A general class of coefficients of divergence of one distribution from another*, J. Roy. Stat. Soc. Ser. B, vol 28, pp. 131–142, 1996.

[4] C. Berg, J.P.R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups*,Springer, New York, 1984.

[5] L. Birgé, *Approximation dans les espaces métriques et theorie de l'estimation*, Z. Wahrscheinlichkeitstheorie verw.Geb., vol. 65, pp. 181–237, 1983.

[6] L. Birgé, and P. Massart, *Rates of convergence for minimum contrast estimator*, Probab. Theory Relat. Fields, vol. 97, pp. 113–150, 1993.

[7] I. Csiszár, *A note on Jensen's inequality*, Studia Sci. Math. Hungar., vol. 1, pp. 185–188, 1966.

[8] I. Csiszár, *Information-type measures of difference of probability distributions and indirect observation*, Studia Sci. Math. Hungar., vol. 2, pp. 299–318, 1967.

[9] I. Csiszár, *On topological properties of f-divergences*, Studia Sci. Math. Hungar., vol. 2, pp. 329–339, 1967.

[10] I. Csiszár, and J. Körner, *Information Theory: Coding Theorems for discrete memoryless systems*, New York: Academic, 1981.

[11] I. Csiszár, *Sanov property, generalized I-projection and a conditional limit theorem*, Ann. Prob. vol. 12, pp. 768–793, 1984.

[12] D. Dacunha-Castelle, *Ecole d'Ete de Probabilites de Saint-Flour VII–1977*, Berlin, Heidelberg, New York: Springer, 1978.

[13] A. Dembo, *Information inequalities and concentration of measure*, Ann. Prob., vol. 25, pp. 927–939, 1997.

[14] D. Feldman, and F. Österreicher, *A note on f-divergences*, Studia Sci. Math. Hungar., vol. 24, pp. 191–200, 1989.

[15] L. Györfi, and I. Vajda, *A class of modified Pearson and Neyman statistics*, Statistics and Decisions, to appear.

[16] P. Harremoës, and F. Topsøe, *Instances of exact prediction and universal coding*, in preparation

[17] P. Harremoës. and F. Topsøe, *Inequalities which connect entropy and index of coincidence*, in preparation

[18] P. Kafka, F. Österreicher, and I. Vincze, *On powers of f-divergences defining a distance*, Studia Sci. Math. Hungar., vol. 26, pp. 415–422, 1991.

[19] N. S. Kambo, and S. Kotz, *On exponential bounds for binomial probabilities*, Ann. Inst. Stat. Math., vol. 18, pp. 277–287, 1966.

[20] O. Krafft, *A note on exponential bounds for binomial probabilities*, Ann. Inst. Stat. Math., vol. 21, pp. 219–220, 1969.

[21] O. Krafft, and N. Schmitz, *A note on Hoefding's inequality*, J. Amer. Statist. Assoc., vol. 64 , pp. 907–912, 1969.

[22] C. Kraft, *Some conditions for consistency and uniform consistency of statistical procedures*, Univ. of California Publ. in Statistics, vol. 1 , pp. 125–142, 1955.

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

[24] S. Kullback, and R. Leibler, *On information and sufficiency*, Ann. Math. Statist., vol. 22, pp. 79–86, 1951

[25] L. LeCam, *Asymptotic Methods in Statistical Decision Theory*, New York: Springer, 1986.

[26] J. Lin, *Divergence Measures Based on the Shannon Entropy*, IEEE Trans. Infor. Theory, vol. 37, pp. 145–151, 1991.

[27] J. Lin and S.K.M. Wong, *A new directed divergence measure and its characterization*, Int. J. General Systems, vol. 17, pp. 73–81, 1990.

[28] K. Marton, *Bounding $\overline{d}$-distance by informational divergence: A method to prove measure concentration*, Ann. Prob., vol. 24, pp. 857–866, 1966.

[29] K. Marton, *A simple proof of the blowing-up lemma*, IEEE Trans. Inform. Theory, vol. 32 , pp. 445–446, 1986.

[30] M. Menéndez, D. Morales, L. Pardo and I. Vajda, *Two approaches to grouping of data and related disparaty statistics*, Communications in Statistics – Theory and Methods, vol. 27, pp. 609–633, 1998.

[31] F. Österreicher, and I. Vajda, *Statistical information and discrimination*, IEEE Trans. Inform. Theory, vol. 39, pp. 1036–1039, 1993.

[32] J. Pecaric, *On directed divergence measures*, manuscript, private communication, 1999.

[33] M.S. Pinsker, *Information and information stability of random variables and processes*, San Francisco: Holden-Day, 1964 (Russian original 1960).

[34] B.Ya. Ryabko, *Comments on "A source matching approach to finding minimax codes"*, IEEE Trans. Inform. Theory, vol. 27, pp. 780–781, 1981. (Including also the ensuing Editor's Note).

[35] B.Ya. Ryabko, and F. Topsøe, *On asymptotically optimal methods of prediction and adaptive coding for Markov sources*, manuscript submitted for publication.

[36] F. Topsøe, *Information theoretical optimization techniques*, Kybernetika, vol. 15, pp. 8–27, 1979.

[37] F. Topsøe, *Some inequalities for information divergence and related measures of discrimination*, RGMIA Research Report Collection, http://melba.vu.edu.au/ rgmia/v2n1.html, pp. 73–83, 1999.

[38] G.T. Toussaint, *Sharper lower bounds for discrimination information in terms of variation*, IEEE Trans. Inform. Theory, vol. 21, pp. 99–100, 1975

[39] I. Vajda, *Note on discrimination information and variation*, IEEE Trans. Inform. Theory, vol. 16, pp. 771–773, 1970.

[40] I. Vincze, *On the concept and measure of information contained in an observation*, Contributions to probability (ed. by J. Gani and V.K. Rohatgi), pp. 207–214, New York: Academic, 1981.

[41] A.K.C. Wong and M. You, *Entropy and distance of random graphs with application to structural pattern recognition*, IEEE Trans. Pattern Anal. Machine Intell., vol. 7, pp. 599–609, 1985.

[42] Y. Yang, and A. Barron, *Information-theoretic determination of minimax rates of convergence*, under publication.