# Generalised Pinsker Inequalities

**Mark D. Reid**
Australian National University
Canberra ACT 0200, Australia
Mark.Reid@anu.edu.au

**Robert C. Williamson**
Australian National University and NICTA
Canberra ACT 0200, Australia
Bob.Williamson@anu.edu.au

## Abstract

We generalise the classical Pinsker inequality which relates variational divergence to Kullback-Liebler divergence in two ways: we consider *arbitrary* $f$-divergences in place of KL divergence, and we assume knowledge of a *sequence* of values of generalised variational divergences. We then develop a best possible inequality for this doubly generalised situation. Specialising our result to the classical case provides a new and tight explicit bound relating KL to variational divergence (solving a problem posed by Vajda some 40 years ago). The solution relies on exploiting a connection between divergences and the Bayes risk of a learning problem via an integral representation.

## 1 Introduction

Divergences such as the Kullback-Liebler and variational divergence arise pervasively. They are a means of defining a notion of distance between two probability distributions. The question often arises: given knowledge of one, what can be said of the other? For all distributions $P$ and $Q$ on an arbitrary set, the classical Pinsker inequality relates the Kullback-Liebler divergence $KL(P,Q)$ and variational divergence $V(P,Q)$ by $\mathrm{KL}(\mathrm{P},\mathrm{Q}) \geq \frac{1}{2}[V(P,Q)]^2$. This simple classical bound is known not to be tight. Over the past several decades a number of refinements have been given (see Appendix A for a summary of past work).

Vajda [31] posed the question of determining a *tight lower bound* on KL-divergence in terms of variational divergence. This "best possible Pinsker inequality" takes the form

$$L(V) := \inf_{V(P,Q)=V} \mathrm{KL}(P,Q), \qquad V \in [0,2). \tag{1}$$

Recently Fedotov et al. [7] presented an *implicit* parametric solution of the form of the graph of the bound as $(V(t), L(t))_{t\in\mathbb{R}^+}$ where

$$\begin{aligned} V(t) &= t\left(1 - \left(\coth(t) - \tfrac{1}{t}\right)^2\right), \qquad (2)\\ L(t) &= \log\left(\frac{t}{\sinh(t)}\right) + t\coth(t) - \frac{t^2}{\sinh^2(t)}. \end{aligned}$$

One can generalise the notion of a Pinsker inequality in at least two ways: 1) replace KL divergence by a general $f$-divergence; and 2) bound the $f$-divergence in terms of the known values of a *sequence* of generalised variational divergences (defined later in this paper) $(V_{\pi_i})_{i=1}^n$, $\pi_i \in (0,1)$. In this paper we study this doubly generalised problem and provide a complete solution in terms of explicit, best possible bounds.

The main result is given below as Theorem 6. Applying it to specific $f$-divergences gives the following corollary[1].

**Corollary 1** *Let $V = V(P,Q)$ denote the variational divergence between the distributions $P$ and $Q$ and similarly for the other divergences in Table 1 below. Then the following bounds for the divergences hold and are tight:*

$$h^2 \geq 2 - \sqrt{4-V^2}; \quad \mathrm{J} \geq 2V\ln\left(\tfrac{2+V}{2-V}\right); \quad \Psi \geq \tfrac{8V^2}{4-V^2}$$

$$\begin{aligned} \mathrm{I} &\geq \left(\frac{1}{2}-\frac{V}{4}\right)\ln(2-V)+\left(\frac{1}{2}+\frac{V}{4}\right)\ln(2+V)-\ln(2)\\ \mathrm{T} &\geq \ln\left(\tfrac{4}{\sqrt{4-V^2}}\right) - \ln(2)\\ \chi^2 &\geq [\![V<1]\!]V^2 + [\![V\geq 1]\!]\tfrac{V}{(2-V)} \qquad (3)\\ \mathrm{KL} &\geq \min_{\beta\in[V-2,2-V]} \left(\tfrac{V+2-\beta}{4}\right)\ln\left(\tfrac{\beta-2-V}{\beta-2+V}\right) + \\ &\quad \left(\tfrac{\beta+2-V}{4}\right)\ln\left(\tfrac{\beta+2-V}{\beta+2+V}\right). \qquad (4) \end{aligned}$$

The proof of the main result depends in an essential way on a learning theory perspective. We make use of an integral representation of $f$-divergences in terms of DeGroot's statistical information—the difference between a prior and posterior Bayes risk[4]. By using the relationships between the generalised variational divergence and the 0-1 misclassification loss we are able to use an elementary but somewhat intricate geometrical argument to obtain the result.

The rest of the paper is organised as follows. Section 2 collects background results upon which we rely. The main result of the paper is stated in Section 3 and its proof presented in in Section 4. Appendix A summarises previous work.

[1] The terms $[\![V<1]\!]$ and $[\![V\geq 1]\!]$ are indicator functions and are defined below.

## 2 Background Results and Notation

In this section we collect notation and background concepts and results we need for the main result.

### 2.1 Notational Conventions

The substantive objects are defined within the body of the paper. Here we collect elementary notation and the conventions we adopt throughout. We write $x \wedge y := \min(x, y)$, $x \vee y := \max(x, y)$ and $[\![p]\!] = 1$ if $p$ is true and $[\![p]\!] = 0$ otherwise. The generalised function $\delta(\cdot)$ is defined by $\int_a^b \delta(x) f(x) dx = f(0)$ when $f$ is continuous at 0 and $a < 0 < b$. For convenience, we will define $\delta_c(x) := \delta(x - c)$. The real numbers are denoted $\mathbb{R}$, the non-negative reals $\mathbb{R}^+$; Sets are in calligraphic font: $\mathcal{X}$. Vectors are written in bold font: $\mathbf{a}, \boldsymbol{\alpha}, \boldsymbol{x} \in \mathbb{R}^m$. We will often have cause to take expectations ($\mathbb{E}$) over random variables. We write such quantities in blackboard bold: $\mathbb{I}$, $\mathbb{L}$, *etc.* The lower bound on quantities with an intrinsic lower bound (e.g. the Bayes optimal loss) are written with an underbar: $\underline{L}$, $\underline{\mathbb{L}}$. Quantities related by double integration recur in this paper and we notate the starting point in lower case, the first integral with upper case, and the second integral in upper case with an overbar: $\gamma$, $\Gamma$, $\bar{\Gamma}$.

### 2.2 Csiszár $f$-divergences

The class of *$f$-divergences* [1, 3] provide a rich set of relations that can be used to measure the separation of the distributions. An $f$-divergence is a function that measures the "distance" between a pair of distributions $P$ and $Q$ defined over a space $\mathcal{X}$ of observations. Traditionally, the $f$-divergence of $P$ from $Q$ is defined for any convex $f : (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$. In this case, the $f$-divergence is

$$\mathbb{I}_f(P, Q) = \mathbb{E}_Q \left[ f\left(\frac{dP}{dQ}\right)\right] = \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ \qquad (5)$$

when $P$ is absolutely continuous with respect to $Q$ and equal $\infty$ otherwise.[2]

All $f$-divergences are non-negative and zero when $P = Q$, that is, $\mathbb{I}_f(P, Q) \geq 0$ and $\mathbb{I}_f(P, P) = 0$ for all distributions $P, Q$. In general, however, they are not metrics, since they are not necessarily *symmetric* (*i.e.*, for all distributions $P$ and $Q$, $\mathbb{I}_f(P, Q) = \mathbb{I}_f(Q, P)$) and do not necessarily satisfy the triangle inequality.

Several well-known divergences correspond to specific choices of the function $f$ [1, §5]. One divergence central to this paper is the *variational divergence* $V(P, Q)$ which is obtained by setting $f(t) = |t-1|$ in Equation 5. It is the only $f$-divergence that is a true metric on the space of distributions over $\mathcal{X}$ [13] and gets its name from its equivalent definition in the variational form

$$V(P, Q) = 2\|P - Q\|_\infty := 2 \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)|. \quad (6)$$

(Some authors define $V$ without the 2 above.) Furthermore, the variational divergence is one of a family of "primitive" or "simple" $f$-divergences discussed in Section 2.3. These are primitive in the sense that all other $f$-divergences can be expressed as a weighted sum of members from this family.

Another well known $f$-divergence is the Kullback-Leibler (KL) divergence $\mathrm{KL}(P, Q)$, obtained by setting $f(t) = t \ln(t)$ in Equation 5. Others are given in Table 1.

As already mentioned in the introduction, the KL and variational divergences satisfy the classical Pinsker's inequality which states that for all distributions $P$ and $Q$ over some common space $\mathcal{X}$

$$\mathrm{KL}(P, Q) \geq \tfrac{1}{2}[V(P, Q)]^2. \qquad (7)$$

### 2.3 Integral Representations of $f$-divergences

The main tool in our proof of Theorem 6 is the integral representation of $f$-divergences, first articulated by Österreicher and Vajda [20] and Gutenbrunner [12]. They show that an $f$-divergence can be represented as a weighted integral of the "simple" divergence measures

$$V_\pi(P, Q) = \mathbb{I}_{f_\pi}(P, Q), \qquad (8)$$

where $f_\pi(t) := \min\{\pi, 1 - \pi\} - \min\{1 - \pi, \pi t\}$ for $\pi \in [0, 1]$.

**Theorem 2** *For any convex $f$ such that $f(1) = 0$, the $f$-divergence $\mathbb{I}_f$ can be expressed, for all distributions $P$ and $Q$, as*

$$\mathbb{I}_f(P, Q) = \int_0^1 V_\pi(P, Q)\, \gamma_f(\pi)\, d\pi \qquad (9)$$

*where the (generalised) function*

$$\gamma_f(\pi) := \frac{1}{\pi^3} f''\left(\frac{1 - \pi}{\pi}\right). \qquad (10)$$

Recently, this theorem has been shown to be a direct consequence of a generalised Taylor's expansion for convex functions [17, 22].

Even when $f$ is not twice differentiable, the convexity of $f$ implies its continuity and so its right-hand derivative $f'_+$ exists. In this case, $\gamma$ is interpreted distributionally in terms of $df'_+$. For example, when $f(t) = |t - 1|$ then $f''(t) = 2\delta(t - 1)$ and so $\gamma_f(\pi) = 2\frac{1}{\pi^3}\delta(1 - 2\pi) = 16\delta_{\frac{1}{2}}(\pi)$.

The divergences $V_\pi$ for $\pi \in [0, 1]$ can be seen as a family of generalised variational divergences since, $df'_+(t)$ for any member of this family is $\pi\delta(t - \frac{1-\pi}{\pi})$ and so $\gamma_{f_\pi} = \frac{1}{\pi^2}\delta_{\frac{1-\pi}{\pi}}$. Thus, for $\pi = \frac{1}{2}$ we have $\gamma_{f_{\frac{1}{2}}} = 4\delta_{\frac{1}{2}}$, that is, four times the $\gamma$ function for variational divergence and so by (9) we see that

$$V(P, Q) = 4V_{\frac{1}{2}}(P, Q). \qquad (11)$$

Theorem 2 shows that knowledge of the values of $V_\pi(P, Q)$ for all $\pi \in [0, 1]$ is sufficient to compute the value of $\mathbb{I}_f(P, Q)$ for any $f$-divergence, since the weight function $\gamma$ is dependent only on $f$, not $P$ and $Q$. All of the generalised Pinsker bounds we derive are found by asking how knowledge of a the value of a finite number of $V_\pi(P, Q)$ constrains the overall value of $\mathbb{I}_f(P, Q)$.

[2] Liese and Miescke [18, pg. 34] give a definition that does not require absolute continuity.

| Symbol | Divergence Name | $f(t)$ | $\gamma(\pi)$ |
|---|---|---|---|
| $V(P,Q)$ | Variational | $\lvert t-1\rvert$ | $16\delta\left(\pi-\frac{1}{2}\right)$ |
| $\mathrm{KL}(P,Q)$ | Kullback-Liebler | $t\ln t$ | $\frac{1}{\pi^2(1-\pi)}$ |
| $\Delta(P,Q)$ | Triangular Discrimination | $(t-1)^2/(t+1)$ | 8 |
| $\mathrm{I}(P,Q)$ | Jensen-Shannon | $\frac{t}{2}\ln t-\frac{(t+1)}{2}\ln(t+1)+\ln 2$ | $\frac{1}{2\pi(1-\pi)}$ |
| $\mathrm{T}(P,Q)$ | Arithmetic-Geometric Mean | $\left(\frac{t+1}{2}\right)\ln\left(\frac{t+1}{2\sqrt{t}}\right)$ | $\frac{(2\pi-\frac{1}{2})^2+\frac{1}{2}}{4\pi^2(\pi-1)^2}$ |
| $\mathrm{J}(P,Q)$ | Jeffreys | $(t-1)\ln(t)$ | $\frac{1}{\pi^2(1-\pi)^2}$ |
| $h^2(P,Q)$ | Hellinger | $(\sqrt{t}-1)^2$ | $\frac{1}{2[\pi(1-\pi)]^{3/2}}$ |
| $\chi^2(P,Q)$ | Pearson $\chi$-squared | $(t-1)^2$ | $\frac{2}{\pi^3}$ |
| $\Psi(P,Q)$ | Symmetric $\chi$-squared | $\frac{(t-1)^2(t+1)}{t}$ | $\frac{2}{\pi^3}+\frac{2}{(1-\pi)^3}$ |

Table 1: Divergences and their corresponding functions $f$ and weights $\gamma$; confer [25, 17]. Topsøe [27] calls $C(P,Q)=2\mathrm{I}(P,Q)$ and $\tilde{C}(P,Q)=2\mathrm{T}(P,Q)$ the Capacitory and Dual Capacitory discrimination respectively. Several of the above divergences are "symmetrised" versions of others. For example, $T(P,Q)=\frac{1}{2}[\mathrm{KL}(\frac{P+Q}{2},P)+\mathrm{KL}(\frac{P+Q}{2},Q)]$, $I(P,Q)=\frac{1}{2}[\mathrm{KL}(P,\frac{P+Q}{2})+\mathrm{KL}(Q,\frac{P+Q}{2})]$, $\mathrm{J}(P,Q)=\mathrm{KL}(P,Q)+\mathrm{KL}(Q,P)$, and $\Psi(P,Q)=\chi^2(P,Q)+\chi^2(Q,P)$.

Table 1 summarises the weight functions $\gamma$ for a number of $f$-divergences that appear in the literature. These are used in the proof of specific bounds in Corollary 1.

Before we can prove the main result, we need to establish some properties of the general variational divergences. In particular, we will make use of their relationship to Bayes risks for 0-1 loss.

### 2.4 Divergence and Risk

Let $\underline{\mathbb{L}}(\pi,P,Q)$ denote the 0-1 Bayes risk for a classification problem in which observations are drawn from $\mathcal{X}$ using the mixture distribution $M=\pi P+(1-\pi)Q$, and each observation $x\in\mathcal{X}$ is assigned a positive label with probability $\eta(x):=\pi\frac{dP}{dM}(x)$. If $r=r(x)\in\{0,1\}$ is a label prediction for a particular $x\in\mathcal{X}$, the 0-1 expected loss for that observation is

$$L(r,\pi,p,q)=(1-\pi)q[\![r=1]\!]+\pi p[\![r=0]\!].$$

where $q=\frac{dQ}{dM}(x)$ and $p=\frac{dP}{dM}(x)$ are densities. Thus, the full expected 0-1 loss of a predictor $r:\mathcal{X}\to\{0,1\}$ is given by $\mathbb{L}(r,\pi,P,Q):=\mathbb{E}_M[L(r(x),\pi,p(x),q(x))]$ and it is well known (*e.g.*, [5]) that its Bayes risk is obtained by the Bayes optimal predictor $r^*(x):=[\![\eta(x)\geq\frac{1}{2}]\!]$. That is,

$$\underline{\mathbb{L}}(\pi,P,Q):=\inf_r\mathbb{L}(r,\pi,P,Q)=\mathbb{L}(r^*,\pi,P,Q),\quad(12)$$

where the infimum is taken over all ($M$-measurable) predictors $r:\mathcal{X}\to\{0,1\}$. So, by the definition of $\eta(x)$ and noting that $\eta\geq\frac{1}{2}$ iff $\pi p\geq\frac{1}{2}(\pi p+(1-\pi)q)$ which holds iff $\pi p\geq(1-\pi)q$ we see that the 0-1 Bayes risk can be expressed as

$$\begin{aligned}&\underline{\mathbb{L}}(\pi,P,Q)\qquad(13)\\ &=\mathbb{E}_M[(1-\pi)q[\![\eta\geq\tfrac{1}{2}]\!]+\pi p[\![\eta<\tfrac{1}{2}]\!]]\\ &=(1-\pi)\mathbb{E}_Q[[\![\pi p\geq(1-\pi)q]\!]]+\pi\mathbb{E}_P[[\![\pi p<(1-\pi)q]\!]].\end{aligned}$$

We now observe that

$$qf_\pi\left(\frac{p}{q}\right)=((1-\pi)\wedge\pi)q-\begin{cases}(1-\pi)q, & q(1-\pi)\leq\pi p\\ \pi p, & q(1-\pi)>\pi p\end{cases}$$

and so by noting that $\mathbb{E}_Q\left[f_\pi\left(\frac{dP}{dQ}\right)\right]=\mathbb{E}_M\left[qf_\pi\left(\frac{p}{q}\right)\right]$ we have established the following lemma.

**Lemma 3** *For all $\pi\in[0,1]$ and all distributions $P$ and $Q$, the generalised variational divergence satisfies*

$$V_\pi(P,Q)\quad=\quad(1-\pi)\wedge\pi-\underline{\mathbb{L}}(\pi,P,Q).\qquad(14)$$

Thus, the value of $V_\pi(P,Q)$ can be understood via the 0-1 Bayes risk for a classification problem with label-conditional distributions $P$ and $Q$ and prior probability $\pi$ for the positive class. This relationship between $f$-divergence and Bayes risk is not new. It was established in a more general setting by Österreicher and Vajda [20] (who note that the term in (14) is the statistical information for 0-1 loss) and later by Nguyen *et al.* [19].

### 2.5 Concavity of 0-1 Bayes Risk Curves

For a given pair of distributions $P$ and $Q$ the set of values for $\underline{\mathbb{L}}(\pi,P,Q)$ as $\pi$ varies over $[0,1]$ can be visualised as a curve as in Figure 2.

**Lemma 4** *For all distributions $P$ and $Q$, the function $\pi\mapsto\underline{\mathbb{L}}(\pi,P,Q)$ is concave.*

**Proof:** By (12) we have that

$$\underline{\mathbb{L}}(\pi,P,Q)=\mathbb{E}_M[L(r^*,\pi,p,q)].$$

Observe that

$$\begin{aligned}L(r^*,\pi,p,q)&=(1-\pi)q[\![\eta\geq\tfrac{1}{2}]\!]+\pi p[\![\eta<\tfrac{1}{2}]\!]\\&=\begin{cases}(1-\pi)q, & q(1-\pi)\leq\pi p\\ \pi p, & q(1-\pi)>\pi p\end{cases}\\&=\min\{(1-\pi)q,\pi p\}\end{aligned}$$

and so for any $p, q$ is the minimum of two linear functions and thus concave in $\pi$. The full Bayes risk is the expectation of these functions and thus simply a linear combination of concave functions and thus concave. ■

The tightness of the bounds in the main result of the next section depend on the following corollary of a result due to Torgersen [28]. It asserts that any appropriate concave function can be viewed as the 0-1 risk curve for *some* pair of distributions $P$ and $Q$. A proof can be found in [22, §6.3].

**Corollary 5** *Suppose $\mathcal{X}$ has a connected component. Let $\psi\colon [0,1] \to [0,1]$ be an arbitrary concave function such that for all $\pi \in [0,1]$, $0 \le \psi(\pi) \le \pi \wedge (1-\pi)$. Then there exists $P$ and $Q$ such that $\underline{\mathbb{L}}(\pi, P, Q) = \psi(\pi)$ for all $\pi \in [0,1]$.*

## 3 Main Result

We will now show how viewing $f$-divergences in terms of their weighted integral representation simplifies the problem of understanding the relationship between different divergences and leads, amongst other things, to an explicit formula for (1).

Fix a positive integer $n$. Consider a sequence $0 < \pi_1 < \pi_2 < \cdots < \pi_n < 1$. Suppose we "sampled" the value of $V_\pi(P,Q)$ at these discrete values of $\pi$. Since $\pi \mapsto V_\pi(P,Q)$ is concave, the piece-wise linear concave function passing through points

$$\{(\pi_i, V_{\pi_i}(P,Q))\}_{i=1}^n$$

is guaranteed to be an upper bound on the variational curve $(\pi, V_\pi(P,Q))_{\pi\in(0,1)}$. This therefore gives a lower bound on the $f$-divergence given by a weight function $\gamma$. This observation forms the basis of the theorem stated below.

**Theorem 6** *For a positive integer $n$ consider a sequence $0 < \pi_1 < \pi_2 < \cdots < \pi_n < 1$. Let $\pi_0 := 0$ and $\pi_{n+1} := 1$ and for $i = 0, \ldots, n+1$ let*

$$\psi_i := (1-\pi_i) \wedge \pi_i - V_{\pi_i}(P,Q)$$

*(observe that consequently $\psi_0 = \psi_{n+1} = 0$). Let*

$$A_n := \left\{ \mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n : \frac{\psi_{i+1} - \psi_i}{\pi_{i+1} - \pi_i} \le a_i \le \frac{\psi_i - \psi_{i-1}}{\pi_i - \pi_{i-1}},\ i = 1, \ldots, n \right\}. \quad (15)$$

*The set $A_n$ defines the allowable slopes of a piecewise linear function majorizing $\pi \mapsto V_\pi(P,Q)$ at each of $\pi_1, \ldots, \pi_n$. For $\mathbf{a} = (a_1, \ldots, a_n) \in A_n$, let*

$$\tilde{\pi}_i := \frac{\psi_i - \psi_{i+1} + a_{i+1}\pi_{i+1} - a_i \pi_i}{a_{i+1} - a_i},\ i = 0, \ldots, n, \quad (16)$$

$$j := \{k \in \{1, \ldots, n\} : \tilde{\pi}_k < \tfrac{1}{2} \le \tilde{\pi}_{k+1}\}. \quad (17)$$

$$\bar{\pi}_i := [\![i < j]\!]\tilde{\pi}_i + [\![i = j]\!]\tfrac{1}{2} + [\![j < i]\!]\tilde{\pi}_{i-1}, \quad (18)$$

$$\alpha_{\mathbf{a},i} := [\![i \le j]\!](1 - a_i) + [\![i > j]\!](-1 - a_{i-1}), \quad (19)$$

$$\beta_{\mathbf{a},i} := [\![i \le j]\!](\psi_i - a_i\pi_i) + [\![i > j]\!](\psi_{i-1} - a_{i-1}\pi_{i-1}) \quad (20)$$

*for $i = 0, \ldots, n+1$ and let $\gamma_f$ be the weight corresponding to $f$ given by (10).*

*For arbitrary $\mathbb{I}_f$ and for all distributions $P$ and $Q$ the following bound holds. If in addition $\mathcal{X}$ contains a connected component, it is tight.*

$$\mathbb{I}_f(P,Q) \quad (21)$$

$$\ge \min_{\mathbf{a}\in A_n} \sum_{i=0}^{n} \int_{\bar{\pi}_i}^{\bar{\pi}_{i+1}} (\alpha_{\mathbf{a},i}\pi + \beta_{\mathbf{a},i})\gamma_f(\pi)d\pi \quad (22)$$

$$= \min_{\mathbf{a}\in A_n} \sum_{i=0}^{n} \left[ (\alpha_{\mathbf{a},i}\bar{\pi}_{i+1} + \beta_{\mathbf{a},i})\,\Gamma_f(\bar{\pi}_{i+1}) - \alpha_{\mathbf{a},i}\bar{\Gamma}_f(\bar{\pi}_{i+1}) - (\alpha_{\mathbf{a},i}\bar{\pi}_i + \beta_{\mathbf{a},i})\,\Gamma_f(\bar{\pi}_i) + \alpha_{\mathbf{a},i}\bar{\Gamma}_f(\bar{\pi}_i) \right], \quad (23)$$

*where $\Gamma_f(\pi) := \int^\pi \gamma_f(t)dt$ and $\bar{\Gamma}_f(\pi) := \int^\pi \Gamma_f(t)dt$.*

Equation 23 follows from (22) by integration by parts. The remainder of the proof is in Section 4. Although (23) looks daunting, we observe: (1) the constraints on $\mathbf{a}$ are convex (in fact they are a box constraint); and (2) the objective is a relatively benign function of $\mathbf{a}$.

When $n = 1$ the result simplifies considerably. If in addition $\pi_1 = \frac{1}{2}$ then by (11) we have $V_{\frac{1}{2}}(P,Q) = \frac{1}{4}V(P,Q)$. It is then a straightforward exercise to explicitly evaluate (22), especially when $\gamma_f$ is symmetric. The following theorem expresses the result in terms of $V(P,Q)$ for comparability with previous results. The result for $\mathrm{KL}(P,Q)$ is a (best-possible) improvement on the classical Pinsker inequality.

**Theorem 7** *For any distributions $P, Q$ on $\mathcal{X}$, let $V := V(P,Q)$. Then the following bounds hold and, if in addition $\mathcal{X}$ has a connected component, are tight.*

*When $\gamma$ is symmetric about $\frac{1}{2}$ and convex,*

$$\mathbb{I}_f(P,Q) \ge 2\left[\bar{\Gamma}_f\left(\tfrac{1}{2} - \tfrac{V}{4}\right) + \tfrac{V}{4}\Gamma_f\left(\tfrac{1}{2}\right) - \bar{\Gamma}_f\left(\tfrac{1}{2}\right)\right] \quad (24)$$

*and $\Gamma_f$ and $\bar{\Gamma}_f$ are as in Theorem 6.*

This theorem gives the first explicit representation of the optimal Pinsker bound.[3] By plotting both (2) and (4) one can confirm that the two bounds (implicit and explicit) coincide; see Figure 1.

## 4 Proof of Main Result

**Proof: (Theorem 6)** This proof is driven by the duality between the family of variational divergences $V_\pi(P,Q)$ and the 0-1 Bayes risk $\underline{\mathbb{L}}(\pi, P, Q)$ given in Lemma 3. Given distributions $P$ and $Q$ let

$$\phi(\pi) = V_\pi(P,Q) = \pi \wedge (1-\pi) - \psi(\pi),$$

where $\psi(\pi) = \underline{\mathbb{L}}(\pi, P, Q)$. We know that $\psi$ is non-negative and concave and satisfies $\psi(\pi) \le \pi \wedge (1-\pi)$ and thus $\psi(0) = \psi(1) = 0$.

Since

$$\mathbb{I}_f(P,Q) = \int_0^1 \phi(\pi)\gamma_f(\pi)d\pi, \quad (25)$$

[3] A summary of existing results and their relationship to those presented here is given in appendix A.

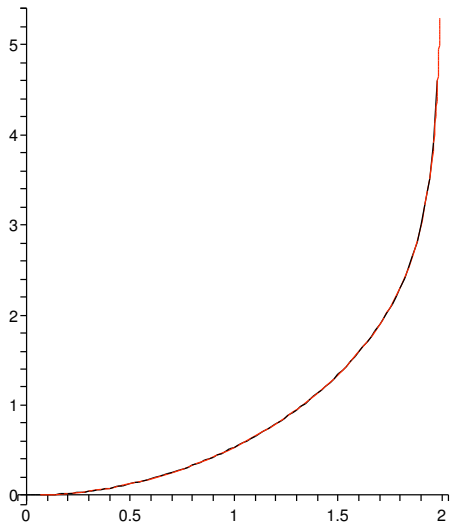


Figure 1: Lower bound on $\mathrm{KL}(P,Q)$ as a function of the variational divergence $V(P,Q)$. Both the explicit bound (4) and Fedotorev et al.'s implicit bound (2) are plotted.

$\mathbb{I}_f(P,Q)$ is minimised by minimising $\phi$ over all $(P,Q)$ such that

$$\phi(\pi_i) = \phi_i = \pi_i \wedge (1-\pi_i) - \psi(\pi_i).$$

Since $\psi_i := (1-\pi_i)\wedge\pi_i - V_{\pi_i}(P,Q) = \psi(\pi_i)$ the minimisation problem for $\phi$ can be expressed in terms of $\psi$ as:

$$\begin{aligned} &\text{Given } (\pi_i,\psi_i)_{i=1}^n \text{ find the maximal } \psi\colon [0,1]\to[0,\tfrac{1}{2}] && (26)\\ &\quad\text{such that}\quad \psi(\pi_i)=\psi_i,\quad i=0,\dots,n+1, && (27)\\ &\qquad\qquad \psi(\pi)\le \pi\wedge(1-\pi),\quad \pi\in[0,1], && (28)\\ &\qquad\qquad \psi \text{ is concave.} && (29)\end{aligned}$$

This will tell us the optimal $\phi$ to use since optimising over $\psi$ is equivalent to optimising over $\underline{\mathbb{L}}(\cdot,P,Q)$. Under the additional assumption on $\mathcal{X}$, Corollary 5 implies that for any $\psi$ satisfying (27), (28) and (29) there exists $P,Q$ such that $\underline{\mathbb{L}}(\cdot,P,Q)=\psi(\cdot)$. This establishes the tightness of our bounds.

Let $\Psi$ be the set of piece-wise linear concave functions on $[0,1]$ having $n+1$ segments such that $\psi\in\Psi \Rightarrow \psi$ satisfies (27) and (28). We now show that in order to solve (26) it suffices to consider $\psi\in\Psi$.

If $g$ is a concave function on $\mathbb{R}$, then let

$$\eth g(x) := \{s\in\mathbb{R}\colon g(y)\le g(x)+\langle s, y-x\rangle,\ y\in\mathbb{R}\}$$

denote the *sup-differential* of $g$ at $x$. (This is the obvious analogue of the *sub*-differential for convex functions [23].) Suppose $\tilde\psi$ is a general concave function satisfying (27) and (28). For $i=1,\dots,n$, let

$$G_i^{\tilde\psi} := \Big\{ [0,1]\ni g_i^{\tilde\psi} : \pi_i\mapsto\psi_i\in\mathbb{R} \text{ is linear and } \frac{\partial}{\partial\pi} g_i^{\tilde\psi}(\pi)\Big|_{\pi=\pi_i} \in \eth\tilde\psi(\pi_i)\Big\}.$$

Observe that by concavity, for all concave $\tilde\psi$ satisfying (27) and (28), for all $g\in\bigcup_{i=1}^n G_i^{\tilde\psi}$, $g(\pi)\ge\psi(\pi)$, for $\pi\in[0,1]$.

Thus given any such $\tilde\psi$, one can always construct

$$\psi^*(\pi) = \min(g_1^{\tilde\psi}(\pi),\dots,g_n^{\tilde\psi}(\pi)) \qquad (30)$$

such that $\psi^*$ is concave, satisfies (27) and $\psi^*(\pi)\ge\tilde\psi(\pi)$, for all $\pi\in[0,1]$. It remains to take account of (28). That is trivially done by setting

$$\psi(\pi) = \min(\psi^*(\pi), \pi\wedge(1-\pi)) \qquad (31)$$

which remains concave and piecewise linear (although with potentially one additional linear segment). Finally, the pointwise smallest concave $\psi$ satisfying (27) and (28) is the piecewise linear function connecting the points $(0,0),(\pi_1,\psi_1),(\pi_2,\psi_2),\dots,(\pi_m,\psi_m),(1,0)$.

Let $g\colon[0,1]\to[0,\frac{1}{2}]$ be this function which can be written explicitly as

$$g(\pi) = \left(\psi_i + \frac{(\psi_{i+1}-\psi)(\pi-\pi_i)}{\pi_{i+1}-\pi_i}\right)\cdot[\![\pi\in[\pi_i,\pi_{i+1}]]\!],$$

where we have defined $\pi_0:=0$, $\psi_0:=0$, $\pi_{n+1}:=1$ and $\psi_{n+1}:=0$.

We now explicitly parametrize this family of functions. Let $p_i\colon[0,1]\to\mathbb{R}$ denote the affine segment the graph of which passes through $(\pi_i,\psi_i)$, $i=0,\dots,n+1$. Write $p_i(\pi)=a_i\pi+b_i$. We know that $p_i(\pi_i)=\psi_i$ and thus

$$b_i = \psi_i - a_i\pi_i,\quad i=0,\dots,n+1. \qquad (32)$$

In order to determine the constraints on $a_i$, since $g$ is concave and minorizes $\psi$, it suffices to only consider $(\pi_{i-1},g(\pi_{i-1}))$ and $(\pi_{i+1},g(\pi_{i+1}))$ for $i=1,\dots,n$. We have (for $i=1,\dots,n$)

$$\begin{aligned} & p_i(\pi_{i-1}) && \ge g(\pi_{i-1})\\ \Rightarrow\quad & a_i\pi_{i-1}+b_i && \ge \psi_{i-1}\\ \Rightarrow\quad & a_i\pi_{i-1}+\psi_i-a_i\pi_i && \ge \psi_{i-1}\\ \Rightarrow\quad & a_i\underbrace{(\pi_{i-1}-\pi_i)}_{<0} && \ge \psi_{i-1}-\psi_i\\ \Rightarrow\quad & a_i && \le \frac{\psi_{i-1}-\psi_i}{\pi_{i-1}-\pi_i}. \end{aligned} \qquad (33)$$

Similarly we have (for $i=1,\dots,n$)

$$\begin{aligned} & p_i(\pi_{i+1}) && \ge g(\pi_{i+1})\\ \Rightarrow\quad & a_i\pi_{i+1}+b_i && \ge \psi_{i+1}\\ \Rightarrow\quad & a_i\pi_{i+1}+\psi_i-a_i\pi_i && \ge \psi_{i+1}\\ \Rightarrow\quad & a_i\underbrace{(\pi_{i+1}-\pi_i)}_{>0} && \ge \psi_{i+1}-\psi_i\\ \Rightarrow\quad & a_i && \ge \frac{\psi_{i+1}-\psi_i}{\pi_{i+1}-\pi_i}. \end{aligned} \qquad (34)$$

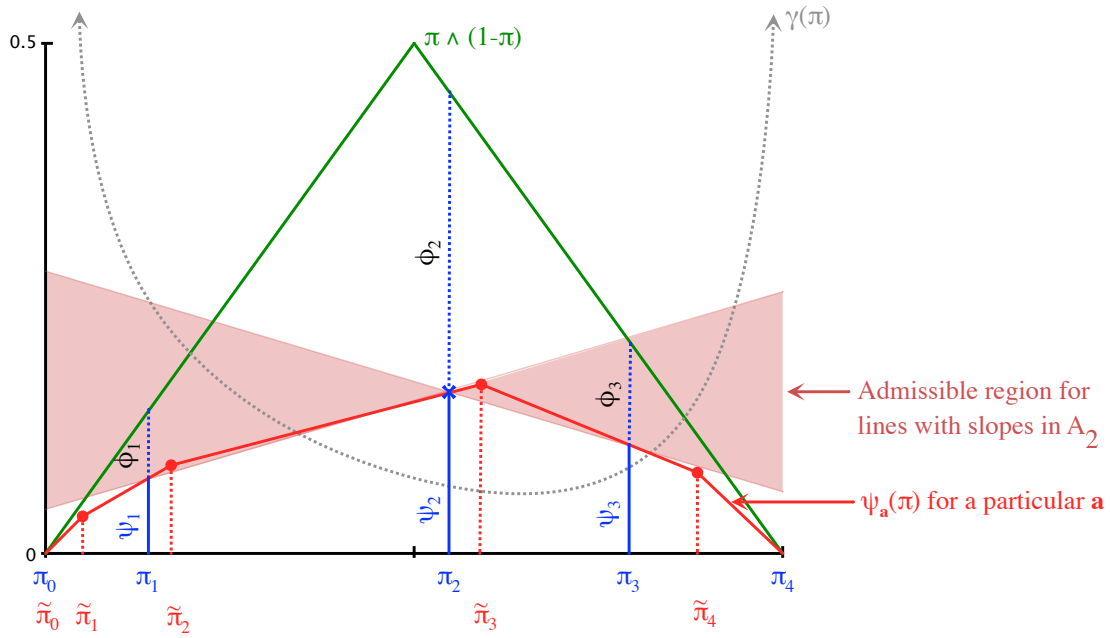


Figure 2: Illustration of construction of optimal $\psi(\pi) = \underline{\mathbb{L}}(\pi, P, Q)$ in the proof of Theorem 6. The optimal $\psi$ is piece-wise linear such that $\psi(\pi_i) = \psi_i$, $i = 0, \ldots, n+1$.

We now determine the points at which $\psi$ defined by (30) and (31) change slope. That occurs at the points $\pi$ when

$$\begin{aligned} & p_i(\pi) && = p_{i+1}(\pi) \\ \Rightarrow \quad & a_i\pi + \psi_i - a_i\pi_i && = a_{i+1}\pi + \psi_{i+1} - a_{i+1}\pi_{i+1} \\ \Rightarrow \quad & (a_{i+1} - a_i)\pi && = \psi_i - \psi_{i+1} + a_{i+1}\pi_{i+1} - a_i\pi_i \\ \Rightarrow \quad & \pi && = \frac{\psi_i - \psi_{i+1} + a_{i+1}\pi_{i+1}}{a_{i+1} - a_i} \\ & && =: \tilde{\pi}_i \end{aligned}$$

for $i = 0, \ldots, n$. Thus

$$\psi(\pi) = p_i(\pi), \quad \pi \in [\tilde{\pi}_{i-1}, \tilde{\pi}_i], \ i = 1, \ldots, n.$$

Let $\mathbf{a} = (a_1, \ldots, a_n)$. We explicitly denote the dependence of $\psi$ on $\mathbf{a}$ by writing $\psi_\mathbf{a}$. Let

$$\begin{aligned} \phi_\mathbf{a}(\pi) &:= \pi \wedge (1-\pi) - \psi_\mathbf{a}(\pi) \\ &= \alpha_{\mathbf{a},i}\pi + \beta_{\mathbf{a},i}, \ \pi \in [\bar{\pi}_{i-1}, \bar{\pi}_i], i = 1, \ldots, n+1, \end{aligned}$$

where $\mathbf{a} \in A_n$ (see (15)), $\bar{\pi}_i$, $\alpha_{\mathbf{a},i}$ and $\beta_{\mathbf{a},i}$ are defined by (18), (19) and (20) respectively. The extra segment induced at index $j$ (see (17)) is needed since $\pi \mapsto \pi \wedge (1-\pi)$ has a slope change at $\pi = \frac{1}{2}$. Thus in general, $\phi_\mathbf{a}$ is piece-wise linear with $n+2$ segments (recall $i$ ranges from 0 to $n+2$); if $\tilde{\pi}_{k+1} = \frac{1}{2}$ for some $k \in \{1, \ldots, n\}$, then there will be only $n+1$ non-trivial segments.

Thus

$$\left\{ \pi \mapsto \sum_{i=0}^{n} \phi_\mathbf{a}(\pi) \cdot [\![\pi \in [\bar{\pi}_i, \bar{\pi}_{i+1}]]\!] : \mathbf{a} \in A_n \right\}$$

is the set of $\phi$ consistent with the constraints and $A_n$ is defined in (15). Thus substituting into (25), interchanging the order of summation and integration and optimizing we have shown (22). The tightness has already been argued: under the additional assumption on $\mathcal{X}$, since there is no slop in the argument above since every $\psi$ satisfying the constraints in (26) is the Bayes risk function for some $(P, Q)$. ■

**Proof: (Theorem 7)** In this case $n = 1$ and the optimal $\psi$ function will be piecewise linear, concave, and its graph will pass through $(\pi_1, \psi_1)$. Thus the optimal $\phi$ will be of the form

$$\phi(\pi) = \begin{cases} 0, & \pi \in [0, L] \cup [U, 1] \\ \pi - (a\pi + b), & \pi \in [L, \frac{1}{2}] \\ (1-\pi) - (a\pi + b), & \pi \in [\frac{1}{2}, U]. \end{cases}$$

where $a\pi_1 + b = \psi_1 \Rightarrow b = \psi_1 - a\pi_1$ and $a \in [-2\psi_1, 2\psi_1]$ (see Figure 3). For variational divergence, $\pi_1 = \frac{1}{2}$ and thus by (11)

$$\psi_1 = \pi_1 \wedge (1 - \pi_1) - \frac{V}{4} = \frac{1}{2} - \frac{V}{4} \tag{35}$$

and so $\phi_1 = V/4$. We can thus determine $L$ and $U$:

$$\begin{aligned} & aL + b && = L \\ \Rightarrow \quad & aL + \psi_1 - a\pi_1 && = L \\ \Rightarrow \quad & L && = \frac{a\pi_1 - \psi_1}{a - 1}. \end{aligned}$$

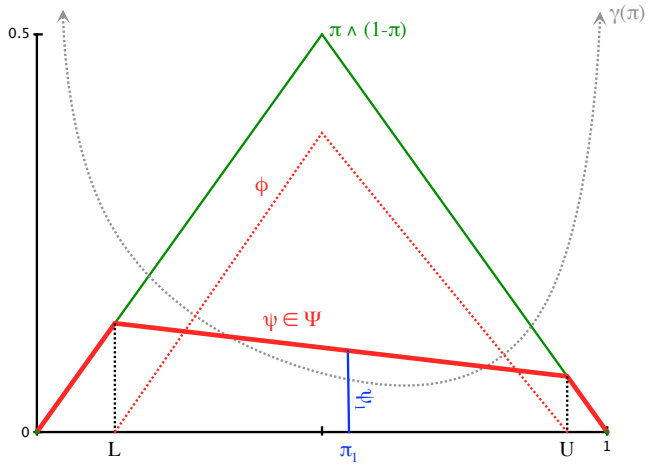


Figure 3: The optimisation problem when $n = 1$. Given $\psi_1$, there are many risk curves consistent with it. The optimisation problem involves finding the piece-wise linear concave risk curve $\psi \in \Psi$ and the corresponding $\phi = \pi \wedge (1-\pi) - \psi$ that maximises $\mathbb{I}_f$. $L$ and $U$ are defined in the text.

Similarly $aU + b = 1 - U \Rightarrow U = \frac{1-\psi_1+a\pi_1}{a+1}$ and thus

$$\mathbb{I}_f(P,Q) \geq \min_{a\in[-2\psi_1,2\psi_1]} \int_{\frac{a\pi_1-\psi_1}{a-1}}^{\frac{1}{2}} [(1-a)\pi - \psi_1 + a\pi_1]\gamma_f(\pi)d\pi + \int_{\frac{1}{2}}^{\frac{1-\psi_1+a\pi_1}{a+1}} [(-a-1)\pi - \psi_1 + a\pi_1 + 1]\gamma_f(\pi)d\pi. \quad (36)$$

If $\gamma_f$ is symmetric about $\pi = \frac{1}{2}$ and convex and $\pi_1 = \frac{1}{2}$, then the optimal $a = 0$. Thus in that case,

$$\begin{aligned}\mathbb{I}_f(P,Q) &\geq 2\int_{\psi_1}^{\frac{1}{2}} (\pi - \psi_1)\gamma_f(\pi)d\pi \quad (37)\\ &= 2\left[(\tfrac{1}{2} - \psi_1)\Gamma_f(\tfrac{1}{2}) + \bar{\Gamma}_f(\psi_1) - \bar{\Gamma}_f(\tfrac{1}{2})\right]\\ &= 2\left[\tfrac{V}{4}\Gamma_f(\tfrac{1}{2}) + \bar{\Gamma}_f\left(\tfrac{1}{2} - \tfrac{V}{4}\right) - \bar{\Gamma}_f(\tfrac{1}{2})\right] \quad (38)\end{aligned}$$

Combining the above with (35) leads to a range of Pinsker style bounds for symmetric $\mathbb{I}_f$:

**Jeffrey's Divergence** Since $J(P,Q) = \mathrm{KL}(P,Q) + \mathrm{KL}(Q,P)$ we have $\gamma(\pi) = \frac{1}{\pi^2(1-\pi)^2}$. (As a check, $f(t) = (t-1)\ln(t)$, $f''(t) = \frac{t+1}{t^2}$ and so $\gamma_f(\pi) = \frac{1}{\pi^3}f''\left(\frac{1-\pi}{\pi}\right) = \frac{1}{\pi^2(1-\pi)^2}$.) Thus

$$\begin{aligned}J(P,Q) &\geq 2\int_{\psi_1}^{1/2} \frac{(\pi-\psi_1)}{\pi^2(1-\pi)^2}d\pi\\ &= (4\psi_1 - 2)(\ln(\psi_1) - \ln(1-\psi_1)).\end{aligned}$$

Substituting $\psi_1 = \frac{1}{2} - \frac{V}{4}$ gives

$$J(P,Q) \geq V\ln\left(\frac{2+V}{2-V}\right).$$

Observe that the above bound behaves like $V^2$ for small $V$, and $V\ln\left(\frac{2+V}{2-V}\right) \geq V^2$ for $V \in [0,2]$. Using the traditional Pinkser inequality ($\mathrm{KL}(P,Q) \geq V^2/2$) we have

$$\begin{aligned}J(P,Q) &= \mathrm{KL}(P,Q) + \mathrm{KL}(Q,P)\\ &\geq \frac{V^2}{2} + \frac{V^2}{2} = V^2.\end{aligned}$$

**Jensen-Shannon Divergence** Here $f(t) = \frac{t}{2}\ln t - \frac{(t+1)}{2}\ln(t+1) + \ln 2$ and thus $\gamma_f(\pi) = \frac{1}{\pi^3}f''\left(\frac{1-\pi}{\pi}\right) = \frac{1}{2\pi(1-\pi)}$. Thus

$$\begin{aligned}I(P,Q) &= 2\int_{\psi_1}^{\frac{1}{2}} \frac{\pi-\psi_1}{2\pi(1-\pi)}d\pi\\ &= \ln(1-\psi_1) - \psi_1\ln(1-\psi_1) + \psi_1\ln\psi_1 + \ln(2).\end{aligned}$$

Substituting $\psi_1 = \frac{1}{2} - \frac{V}{4}$ leads to

$$I(P,Q) \geq \left(\tfrac{1}{2} - \tfrac{V}{4}\right)\ln(2-V) + \left(\tfrac{1}{2} + \tfrac{V}{4}\right)\ln(2+V) - \ln(2).$$

**Hellinger Divergence** Here $f(t) = (\sqrt{t}-1)^2$. Consequently $\gamma_f(\pi) = \frac{1}{\pi^3}f''\left(\frac{1-\pi}{\pi}\right) = \frac{1}{\pi^3}\frac{1}{2((1-\pi)/\pi)^{3/2}} = \frac{1}{2[\pi(1-\pi)]^{3/2}}$ and thus

$$\begin{aligned}h^2(P,Q) &\geq 2\int_{\psi_1}^{\frac{1}{2}} \frac{\pi-\psi_1}{2[\pi(1-\pi)]^{3/2}}d\pi\\ &= \frac{4\sqrt{\psi_1}(\psi_1-1) + 2\sqrt{1-\psi_1}}{\sqrt{1-\psi_1}}\\ &= \frac{4\sqrt{\frac{1}{2}-\frac{V}{4}}\left(\frac{1}{2}-\frac{V}{4}-1\right) + 2\sqrt{1-\frac{1}{2}+\frac{V}{4}}}{\sqrt{1-\frac{1}{2}+\frac{V}{4}}}\\ &= 2 - \frac{(2+V)\sqrt{2-V}}{\sqrt{2+V}}\\ &= 2 - \sqrt{4-V^2}.\end{aligned}$$

For small $V$, $2 - \sqrt{4-V^2} \approx V^2/4$.

**Arithmetic-Geometric Mean Divergence** In this case, $f(t) = \frac{t+1}{2}\ln\left(\frac{t+1}{2\sqrt{t}}\right)$. Thus $f''(t) = \frac{t^2+1}{4t^2(t+1)}$ and hence $\gamma_f(\pi) = \frac{1}{\pi^3}f''\left(\frac{1-\pi}{\pi}\right) = \gamma_f(\pi) = \frac{2\pi^2-2\pi+1}{\pi^2(\pi-1)^2}$ and thus

$$\begin{aligned}T(P,Q) &\geq 2\int_{\psi_1}^{\frac{1}{2}} (\pi-\psi_1)\frac{2\pi^2-2\pi+1}{\pi^2(\pi-1)^2}d\pi\\ &= -\frac{1}{2}\ln(1-\psi) - \frac{1}{2}\ln(\psi) - \ln(2).\end{aligned}$$

Substituting $\psi_1 = \frac{1}{2} - \frac{V}{4}$ gives

$$\begin{aligned}T(P,Q) &\geq -\frac{1}{2}\ln\left(\frac{1}{2}+\frac{V}{4}\right) - \frac{1}{2}\ln\left(\frac{1}{2}-\frac{V}{4}\right) - \ln(2)\\ &= \ln\left(\frac{4}{\sqrt{4-V^2}}\right) - \ln(2).\end{aligned}$$

**Symmetric $\chi^2$-Divergence** In this case $\Psi(P,Q) = \chi^2(P,Q)+\chi^2(Q,P)$ and thus (see below) $\gamma_f(\pi) = \frac{2}{\pi^3} + \frac{2}{(1-\pi)^3}$. (As a check, from $f(t) = \frac{(t-1)^2(t+1)}{t}$ we have $f''(t) = \frac{2(t^3+1)}{t^3}$ and thus $\gamma_f(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right)$ gives the same result.)

$$\begin{aligned}\Psi(P,Q) &\geq 2\int_{\psi_1}^{\frac{1}{2}} (\pi-\psi_1)\left(\frac{2}{\pi^3}+\frac{2}{(1-\pi)^3}\right)d\pi \\ &= \frac{2(1+4\psi_1^2-4\psi_1)}{\psi_1(\psi_1-1)}.\end{aligned}$$

Substituting $\psi_1 = \frac{1}{2} - \frac{V}{4}$ gives $\Psi(P,Q) \geq \frac{8V^2}{4-V^2}$.

When $\gamma_f$ is not symmetric, one needs to use (36) instead of the simpler (38). We consider two cases.

**$\chi^2$-Divergence** Here $f(t) = (t-1)^2$ and so $f''(t) = 2$ and hence $\gamma(\pi) = f''\left(\frac{1-\pi}{\pi}\right)/\pi^3 = \frac{2}{\pi^3}$ which is not symmetric. Upon substituting $2/\pi^3$ for $\gamma(\pi)$ in (36) and evaluating the integrals we obtain

$$\chi^2(P,Q) \geq 2 \min_{a\in[-2\psi_1,2\psi_1]} \underbrace{\frac{1+4\psi_1^2-4\psi_1}{2\psi_1-a} - \frac{1+4\psi_1^2-4\psi_1}{2\psi_1-a-2}}_{=:J(a,\psi_1)}.$$

One can then solve $\frac{\partial}{\partial a}J(a,\psi_1) = 0$ for $a$ and one obtains $a^* = 2\psi_1 - 1$. Now $a^* > -2\psi_1$ only if $\psi_1 > \frac{1}{4}$. One can check that when $\psi_1 \leq \frac{1}{4}$, then $a \mapsto J(a,\psi_1)$ is monotonically increasing for $a \in [-2\psi_1, 2\psi_1]$ and hence the minimum occurs at $a^* = -2\psi_1$. Thus the value of $a$ minimising $J(a,\psi_1)$ is

$$a^* = [\![\psi_1 > 1/4]\!](2\psi_1 - 1) + [\![\psi_1 \leq 1/4]\!](-2\psi_1).$$

Substituting the optimal value of $a^*$ into $J(a,\psi_1)$ we obtain

$$\begin{aligned}J(a^*,\psi_1) &= [\![\psi_1 > 1/4]\!](2+8\psi_1^2-8\psi_1) \\ &\quad + [\![\psi_1 \leq 1/4]\!]\left(\frac{1+4\psi_1^2-4\psi}{4\psi} - \frac{1+4\psi_1^2-4\psi}{4\psi_1-2}\right).\end{aligned}$$

Substituting $\psi_1 = \frac{1}{2} - \frac{V}{4}$ and observing that $V < 1 \Rightarrow \psi_1 > 1/4$ we obtain

$$\chi^2(P,Q) \geq [\![V<1]\!]V^2 + [\![V \geq 1]\!]\frac{V}{(2-V)}.$$

Observe that the bound diverges to $\infty$ as $V \to 2$.

**Kullback-Leibler Divergence** In this case we have $f(t) = t\ln t$ and thus $f''(t) = 1/t$ and consequently $\gamma_f(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right) = \frac{1}{\pi^2(1-\pi)}$ which is clearly not symmetric. From (36) we obtain

$$\begin{aligned}\mathrm{KL}(P,Q) \geq \min_{[-2\psi_1,2\psi_1]} &\left(1-\tfrac{a}{2}-\psi_1\right)\ln\left(\tfrac{a+2\psi_1-2}{a-2\psi_1}\right) \\ &+\left(\tfrac{a}{2}+\psi_1\right)\ln\left(\tfrac{a+2\psi_1}{a-2\psi_1+2}\right).\end{aligned}$$

Substituting $\psi_1 = \frac{1}{2} - \frac{V}{4}$ gives

$$\mathrm{KL}(P,Q) \geq \min_{a\in\left[\frac{V-2}{2},\frac{2-V}{2}\right]} \delta_a(V),$$

where

$$\delta_a(V) = \left(\tfrac{V+2-2a}{4}\right)\ln\left(\tfrac{2a-2-V}{2a-2+V}\right) + \left(\tfrac{2a+2-V}{4}\right)\ln\left(\tfrac{2a+2-V}{2a+2+V}\right).$$

Set $\beta := 2a$ and we have (4). ∎

# 5 Conclusion

We have generalised the classical Pinsker inequality and developed best possible bounds for the general situation. A special case of the result gives an explicit bound relating Kullback-Liebler divergence and variational divergence. The proof relied on an integral representation of $f$-divergences in terms of statistical information. Such representations are a powerful device as they identify the primitives underpinning general learning problems. These representations are further studied in [22].

# A History of Pinsker Inequalities

Pinsker [21] presented the first bound relating $\mathrm{KL}(P,Q)$ to $V(P,Q)$: $\mathrm{KL} \geq V^2/2$ and it is now known by his name or sometimes as the Pinsker-Csiszár-Kullback inequality since Csiszar [3] presented another version and Kullback [14] showed $\mathrm{KL} \geq V^2/2 + V^4/36$. Much later Topsøe [26] showed $\mathrm{KL} \geq V^2/2 + V^4/36 + V^6/270$. Non-polynomial bounds are due to Vajda [31]: $\mathrm{KL} \geq L_{\mathrm{Vajda}}(V) := \log\left(\frac{2+V}{2-V}\right) - \frac{2V}{2+V}$ and Toussaint [29] who showed $\mathrm{KL} \geq L_{\mathrm{Vajda}}(V) \vee (V^2/2 + V^4/36 + V^8/288)$.

Care needs to be taken when comparing results from the literature as different definitions for the divergences exist. For example Gibbs and Su [8] used a definition of $V$ that differs by a factor of 2 from ours. There are some isolated bounds relating $V$ to some other divergences, analogous to the classical Pinkser bound; Kumar [15] has presented a summary as well as new bounds for a wide range of *symmetric* $f$-divergences by making assumptions on the likelihood ratio: $r \leq p(x)/q(x) \leq R < \infty$ for all $x \in \mathcal{X}$. This line of reasoning has also been developed by Dragomir et al. [6] and Taneja [25, 24]. Topsøe [27] has presented some infinite series representations for capacitory discrimination in terms of triangular discrimination which lead to inequalities between those two divergences. Liese and Mieskce [18, p.48] give the inequality $V \leq h\sqrt{4-h^2}$ (which seems to be originally due to LeCam [16]) which when rearranged corresponds exactly to the bound for $h^2$ in theorem 7. Withers [32] has also presented some inequalities between other (particular) pairs of divergences; his reasoning is also in terms of infinite series expansions.

Arnold et al. [30] considered the case of $n = 1$ but arbitrary $\mathbb{I}_f$ (that is they bound an arbitrary $f$-divergence in terms of the variational divergence). Their argument is similar to the geometric proof of Theorem 6. They do not compute any of the explicit bounds in theorem 7 except they state (page 243) $\chi^2(P,Q) \geq V^2$ which is looser than (3).

Gilardoni [9] showed (via an intricate argument) that if $f'''(1)$ exists, then $\mathbb{I}_f \geq \frac{f''(1)V^2}{2}$. He also showed some fourth order inequalities of the form $\mathbb{I}_f \geq c_{2,f}V^2 + c_{4,f}V^4$ where the constants depend on the behaviour of $f$ at 1 in a complex way. Gilardoni [10, 11] presented a completely different approach which obtains many of the results of theorem 7.[4] Gilardoni [11] improved Va-

[4] We were unaware of these two papers until completing

jda's bound slightly to $\mathrm{KL}(P,Q) \geq \ln \frac{2}{2-V} - \frac{2-V}{2} \ln \frac{2+V}{2}$.

Gilardoni [10, 11] presented a general tight lower bound for $\mathbb{I}_f = \mathbb{I}_f(P,Q)$ in terms of $V = V(P,Q)$ which is difficult to evaluate explicitly in general:

$$\mathbb{I}_f \geq \frac{V}{2}\left(\frac{f[g_R^{-1}(k(1/V))]}{g_R^{-1}(k(1/V))-1} + \frac{f[g_L^{-1}(k(1/V))]}{1-g_L^{-1}(k(1/V))}\right),$$

where $k^{-1}(t) = \frac{1}{2}\left(\frac{1}{1-g_L^{-1}(t)} + \frac{1}{g_R^{-1}(t)-1}\right)$ and of course $k(u) = (k^{-1})^{-1}(u)$; and $g(u) = (u-1)f'(u) - f(u)$, $g_R^{-1}[g(u)] = u$ for $u \geq 1$ and $g_L^{-1}[g(u)] = u$ for $u \leq 1$. He presented a new parametric form for $\mathbb{I}_f = \mathrm{KL}$ in terms of Lambert's $W$ function. In general, the result is analogous to that of Fedotov et al. [7] in that it is in a parametric form which, if one wishes to evaluate for a particular $V$, one needs to do a one dimensional numerical search — as complex as (4). However, when $f$ is such that $\mathbb{I}_f$ is symmetric, this simplifies to the elegant form $\mathbb{I}_f \geq \frac{2-V}{2} f\left(\frac{2+V}{2-V}\right) - f'(1)V$. He presented explicit special cases for $h^2$, $J$,$\Delta$ and $I$ identical to the results in Theorem 7. It is not apparent how the approach of Gilardoni [10, 11] could be extended to more general situations such as that in Theorem 6 (i.e. $n > 1$).

Bolley and Villani [2] considered *weighted* versions of the Pinsker inequalities (for a weighted generalisation of Variational divergence) in terms of KL-divergence that are related to transportation inequalities.

**Acknowledgements**

This work was supported by the Australian Research Council and NICTA; an initiative of the Commonwealth Government under Backing Australia's Ability.

## References

[1] S.M. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.

[2] F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculte des Sciences de Toulouse*, 14(3):331–352, 2005.

[3] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[4] M.H. DeGroot. Uncertainty, Information, and Sequential Experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

[5] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Approach to Pattern Recognition*. Springer, New York, 1996.

[6] S.S. Dragomir, V. Gluščević, and C.E.M. Pearce. Csiszár $f$-divergence, Ostrowski's inequality and mutual information. *Nonlinear Analysis*, 47:2375–2386, 2001.

the results presented in the main paper.

[7] A.A. Fedotov, P. Harremoës, and F. Topsøe. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, June 2003.

[8] Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–435, 2002.

[9] G. L. Gilardoni. On Pinsker's Type Inequalities and Csiszár's $f$-divergences. Part I: Second and Fourth-Order Inequalities. arXiv:cs/0603097v2, April 2006.

[10] Gustavo L. Gilardoni. On the minimum $f$-divergence for a given total variation. *Comptes Rendus Académie des sciences, Paris, Series 1*, 343, 2006.

[11] Gustavo L. Gilardoni. On the relationship between symmetric $f$-divergence and total variation and an improved Vajda's inequality. Preprint, Departamento de Estatística, Universidade de Brasília, April 2006.

[12] Cornelius Gutenbrunner. On applications of the representation of $f$-divergences as averaged minimal Bayesian risk. In *Transactions of the 11th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 449–456, Dordrecht; Boston, 1990. Kluwer Academic Publishers.

[13] Mohammadali Khosravifard, Dariush Fooladivanda, and T. Aaron Gulliver. Confliction of the Convexity and Metric Properties in $f$-Divergences. *IEICE Transactions on Fundamentals of Electronics, Communication and Computer Sciences*, E90-A(9):1848–1853, 2007.

[14] S. Kullback. Lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, 13:126–127, 1967. Correction, volume 16, p. 652, September 1970.

[15] P. Kumar and S. Chhina. A symmetric information divergence measure of the Csiszár's $f$-divergence class and its bounds. *Computers and Mathematics with Applications*, 49:575–588, 2005.

[16] Lucien LeCam. *Asymptotic Methods in Statistical Decision Theory*. Springer, 1986.

[17] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[18] Friedrich Liese and Klaus-J. Miescke. *Statistical Decision Theory*. Springer, New York, 2008.

[19] X. Nguyen, M.J. Wainwright, and M.I. Jordan. On distance measures, surrogate loss functions, and distributed detection. Technical Report 695, Department of Statistics, University of California, Berkeley, October 2005.

[20] F. Österreicher and I. Vajda. Statistical information and discrimination. *IEEE Transactions on In-*

*formation Theory*, 39(3):1036–1039, 1993.

[21] M.S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, 1964.

[22] Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. arXiv preprint arXiv:0901.0356v1, 89 pages, January 2009.

[23] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970.

[24] I.J. Taneja. Bounds on non-symmetric divergence measures in terms of symmetric divergence measures. arXiv:math.PR/0506256v1, 2005.

[25] I.J. Taneja. Refinement inequalities among symmetric divergence measures. arXiv:math/0501303v2, April 2005.

[26] F. Topsøe. Bounds for entropy and divergence for distributions over a two-element set. *J. Ineq. Pure & Appl. Math*, 2(2), 2001.

[27] Flemming Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.

[28] E.N. Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991.

[29] G.T. Toussaint. Probability of error, expected divergence and the affinity of several distributions. *IEEE Transactions on Systems, Man and Cybernetics*, 8:482–485, 1978.

[30] Andreas Unterreiter, Anton Arnold, Peter Markowich, and Giuseppe Toscani. On generalized Csiszár-Kullback inequalities. *Monatshefte für Mathematik*, 131:235–253, 2000.

[31] I. Vajda. Note on discrimination and variation. *IEEE Transactions on Information Theory*, 16:771–773, 1970.

[32] Lang Withers. Some inequalities relating different measures of divergence between two probability distributions. *IEEE Transactions on Information Theory*, 45(5):1728–1735, 1999.