Journal of Machine Learning Research 12 (2011) 731-817 Submitted 8/09; Revised 11/10; Published 3/11

Information, Divergence and Risk for Binary Experiments

Mark D. Reid MARK.REID@ANU.EDU.AU
Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
Australian National University and NICTA
Canberra ACT 0200, Australia

Editor: Yoram Singer

Abstract

We unify f-divergences, Bregman divergences, surrogate regret bounds, proper scoring rules, cost curves, ROC-curves and statistical information. We do this by systematically studying integral and variational representations of these objects and in so doing identify their representation primitives which all are related to cost-sensitive binary classification. As well as developing relationships between generative and discriminative views of learning, the new machinery leads to tight and more general surrogate regret bounds and generalised Pinsker inequalities relating f-divergences to variational divergence. The new viewpoint also illuminates existing algorithms: it provides a new derivation of Support Vector Machines in terms of divergences and relates maximum mean discrepancy to Fisher linear discriminants.

Keywords: classification, loss functions, divergence, statistical information, regret bounds

1. Introduction

Some of the simplest machine learning problems concern binary experiments. There it is assumed that observations are drawn from a mixture of two distributions (one for each class). These distributions determine many important objects related to the learning problems they underpin such as risk, divergence and information. Our aim in this paper is to present all of these objects in a coherent framework explaining exactly how they relate to each other. Doing so brings conceptual clarity to the area as well as providing the means for a number of new technical results.

1.1 Motivation

There are many different notions that underpin the definition of machine learning problems. These include information, loss, risk, regret, ROC (Receiver Operating Characteristic) curves and the area under them, Bregman divergences and distance or divergence between probability distributions. On the surface, the problem of estimating whether two distributions are the same (as measured by, say, their Kullback-Leibler divergence) is different to the problem of minimisation of expected risk in a prediction problem. One goal of the present paper is to show how this superficial difference is indeed only superficial—deeper down they are the same problem and analytical and algorithmic insights for one can be transferred to the other.

Machine learning as an engineering discipline is still young.[1] There is no agreed language to describe machine learning problems (such is usually done with an informal mixture of English and

[1] Bousquet (2006) has articulated the need for an agreed vocabulary, a clear statement of the main problems, and to "revisit what has been done or discovered so far with a fresh look".

mathematics). There is very little in the way of composability of machine learning solutions. That is, given the solution to one problem, use it to solve another. Of course one would like to not merely be able to do this, but to be certain what one might lose in doing so. In order to do that, one needs to be able to provide theoretical guarantees on how well the original problem will be solved by solving the surrogate problem. Related to these issues is the fact that there are no well understood *primitives* for machine learning. Indeed, what does that even mean? All of these issues are the underlying motivation for this paper.

Our long term goal (towards which this paper is but the first step) is to turn the field of machine learning into a more well founded engineering discipline with an agreed language and well understood composition rules. Our motivation is that until one can start building systems modularly, one is largely restricted to starting from scratch for each new problem, rather than obtaining the efficiency benefits of re-use.[2]

We are comparing *problems*, not solutions or algorithms. Whilst there have been attempts to provide a degree of unification at the level of algorithms (Altun and Smola, 2006), there are intrinsic limits to such a research program. The most fundamental is that (surprisingly) there is no satisfactory formal definition of what an algorithm really is Blass and Gurevich (2003), nor how two algorithms can be compared with a view to determining if they are the same (Blass et al., 2009).

We have started with binary experiments because they are simple and widely used. As we will show, by pursuing the high level research agenda summarised above, we have managed to unify all of the disparate concepts mentioned and furthermore have simultaneously simplified and generalised two fundamental results: Pinsker inequalities between f-divergences and surrogate regret bounds. The proofs of these new results rely essentially on the decomposition into primitive problems.

1.2 Novelty and Significance

Our initial goal was to present existing material in a unified way. We have indeed done that. In doing so we have developed new (and simpler) proofs of existing results. Additionally we have developed some novel technical results. The key ones are:

1. A link between the weighted integral representations for proper scoring rules and those for f-divergences which allows the transformation from one to the other (Theorem 10);

2. A unified derivation of the integral representations in terms of Taylor series showing their equivalence (Theorem 18);

[2] Abelson et al. (1996) described the principles of constructing software with the aid of (Locke, 1690, Chapter 12, paragraph 1):

> The acts of the mind, wherein it exerts its power over simple ideas, are chiefly these three: (1) Combining several simple ideas into one compound one; and thus all *complex ideas* are made. (2) The second is bringing two ideas, whether simple or complex, together, and setting them by one another, so as to take a view of them at once, without uniting them into one; by which it gets all its *ideas of relations*. (3) The third is separating them from all other ideas that accompany them in their real existence; this is called abstraction: and thus all its *general ideas* are made

Modularity is central to computer hardware (Baldwin and Clark, 2006b,a) and other engineering disciplines (Gershenson et al., 2003) and plays a central role in some models of economic development (Varian, 2003; Weitzman, 1998; Mokyr, 1992). The reason modularity works is that components can be combined or *composed*.

3. Use of these representations to derive new bounds for divergences, Bayes risks and regrets: "surrogate regret bounds"(Theorem 25) and Pinsker inequalities (Theorem 30);
4. Showing that statistical information (and hence f-divergence) are both Bregman informations;
5. The derivation of SVMs from a variational perspective which provides a clearer explanation of the link between MMD (Maximum Mean Discrepancy) and SVMs (Support Vector Machines) §H;
6. Explicit formulae relating Bayes risk to the Neyman-Pearson function, which allows the transformation of risk curves to ROC curves and vice versa (Theorem 22).

The significance of these new connections is that they show that the choice of loss function (scoring rule), f-divergence and Bregman divergence (regret) are intimately related—choosing one implies choices for the others. Furthermore we show there are more intuitively usable parameterisations for f-divergences and scoring rules (their corresponding weight functions). The weight functions have the advantage that if two weight functions match, then the corresponding objects are identical. That is not the case for the f parameterising an f-divergence or the convex function parameterising a Bregman divergence. As well as the theoretical interest in such connections, these alternate representations suggest new algorithms for empirically estimating such quantities. We have represented all of the connections graphically in figure 1. The various symbols are defined below; the point of the picture here is to see the overall goal of the paper—the relating of a range of diverse concepts.

Given the broad scope of our work, there is of course much prior work, too much to summarise in this introduction. Appendix C summarises the main precursors and related work.

1.3 Paper Outline and Key Contributions

The following is an outline of the main structure of this paper section by section highlighting the contributions and novelty. A knowledgeable reader only interested in the core new results should be able to just read Sections 4–8 plus Appendix H with the aid of Table 1. More tedious and technical proofs and digressions are in the appendices.

§2 Many of the properties of the objects studied in this paper are directly derived from well-known properties of convex functions. In particular, a generalised form of Taylor's theorem and Jensen's inequality underpin many of the new results. Although elementary, we have started from this point because it shows how fundamental are the connections drawn later in the paper are. We rederive Savage's famous theorem (Theorem 7) from our perspective.

§3 One of the simplest type of statistical problems is that of distinguishing between two distributions. Such a problem is known as a *binary experiment*. Two classes of *measures of divergence* between the distributions are introduced: the class of Csiszár f-divergences and the class of Bregman divergences.

§4 When additional assumptions are made about a binary experiment—specifically, a prior probability for each of the two distributions—it becomes possible to talk about *risk and statistical information* of an experiment that is defined with respect to a loss function. A key result is Theorem 10 which shows that f-divergence, statistical information and Bregman divergence are all fundamentally equivalent.

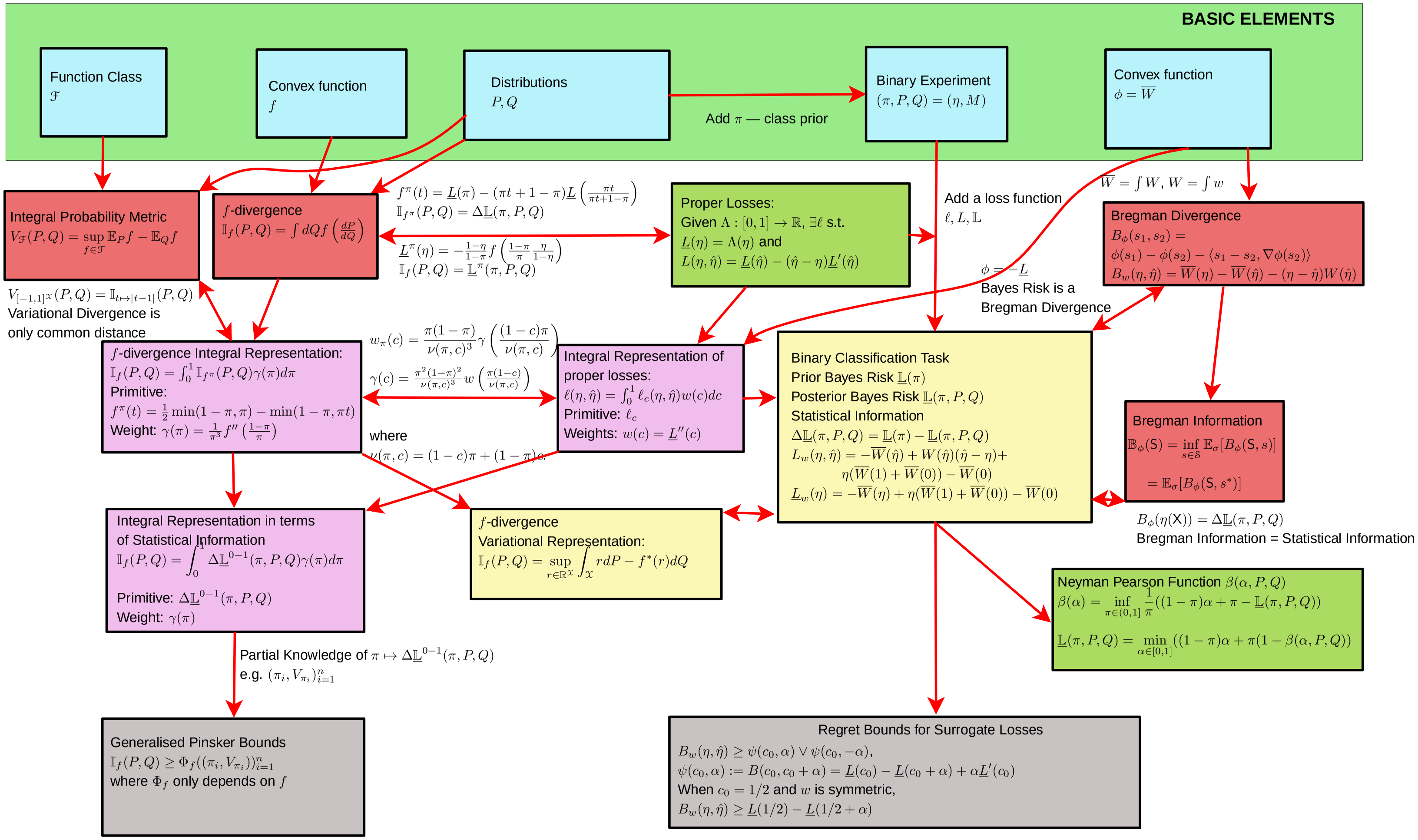

BASIC ELEMENTS
Function Class $\mathcal{F}$
Convex function f
Distributions P, Q
Add π — class prior
Binary Experiment $(\pi, P, Q) = (\eta, M)$
Convex function $\phi = \overline{W}$
Integral Probability Metric $V_{\mathcal{F}}(P,Q) = \sup_{f \in \mathcal{F}} \mathbb{E}_P f - \mathbb{E}_Q f$
$V_{[-1,1]^{\mathcal{X}}}(P,Q) = \mathbb{I}_{t \mapsto |t-1|}(P,Q)$
Variational Divergence is only common distance
f-divergence $\mathbb{I}_f(P,Q) = \int dQ f\left(\frac{dP}{dQ}\right)$
$f^{\pi}(t) = \underline{L}(\pi) - (\pi t + 1 - \pi)\underline{L}\left(\frac{\pi t}{\pi t + 1 - \pi}\right)$
$\mathbb{I}_{f^{\pi}}(P,Q) = \Delta\underline{\mathbb{L}}(\pi, P, Q)$
$\underline{L}^{\pi}(\eta) = -\frac{1-\eta}{1-\pi} f\left(\frac{1-\pi}{\pi}\frac{\eta}{1-\eta}\right)$
$\mathbb{I}_f(P,Q) = \underline{\mathbb{L}}^{\pi}(\pi, P, Q)$
Proper Losses:
Given $\Lambda : [0,1] \to \mathbb{R}$, $\exists \ell$ s.t.
$\underline{L}(\eta) = \Lambda(\eta)$ and
$L(\eta, \hat{\eta}) = \underline{L}(\hat{\eta}) - (\hat{\eta} - \eta)\underline{L}'(\hat{\eta})$
Add a loss function $\ell, L, \mathbb{L}$
$\overline{W} = \int W$, $W = \int w$
$\phi = -\underline{L}$
Bayes Risk is a Bregman Divergence
Bregman Divergence
$B_{\phi}(s_1, s_2) =$
$\phi(s_1) - \phi(s_2) - \langle s_1 - s_2, \nabla\phi(s_2)\rangle$
$B_w(\eta, \hat{\eta}) = \overline{W}(\eta) - \overline{W}(\hat{\eta}) - (\eta - \hat{\eta})W(\hat{\eta})$
f-divergence Integral Representation:
$\mathbb{I}_f(P,Q) = \int_0^1 \mathbb{I}_{f^{\pi}}(P,Q)\gamma(\pi)d\pi$
Primitive:
$f^{\pi}(t) = \frac{1}{2}\min(1-\pi, \pi) - \min(1-\pi, \pi t)$
Weight: $\gamma(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right)$
$w_{\pi}(c) = \frac{\pi(1-\pi)}{\nu(\pi,c)^3}\gamma\left(\frac{(1-c)\pi}{\nu(\pi,c)}\right)$
$\gamma(c) = \frac{\pi^2(1-\pi)^2}{\nu(\pi,c)^3} w\left(\frac{\pi(1-c)}{\nu(\pi,c)}\right)$
where
$\nu(\pi, c) = (1-c)\pi + (1-\pi)c$
Integral Representation of proper losses:
$\ell(\eta, \hat{\eta}) = \int_0^1 \ell_c(\eta, \hat{\eta})w(c)dc$
Primitive: ℓ_c
Weights: $w(c) = \underline{L}''(c)$
Binary Classification Task
Prior Bayes Risk $\underline{\mathbb{L}}(\pi)$
Posterior Bayes Risk $\underline{\mathbb{L}}(\pi, P, Q)$
Statistical Information
$\Delta\underline{\mathbb{L}}(\pi, P, Q) = \underline{\mathbb{L}}(\pi) - \underline{\mathbb{L}}(\pi, P, Q)$
$L_w(\eta, \hat{\eta}) = -\overline{W}(\hat{\eta}) + W(\hat{\eta})(\hat{\eta} - \eta) +$
$\eta(\overline{W}(1) + \overline{W}(0)) - \overline{W}(0)$
$\underline{L}_w(\eta) = -\overline{W}(\eta) + \eta(\overline{W}(1) + \overline{W}(0)) - \overline{W}(0)$
Bregman Information
$\mathbb{B}_{\phi}(\mathsf{S}) = \inf_{s \in \mathcal{S}} \mathbb{E}_{\sigma}[B_{\phi}(\mathsf{S}, s)]$
$= \mathbb{E}_{\sigma}[B_{\phi}(\mathsf{S}, s^*)]$
$B_{\phi}(\eta(\mathsf{X})) = \Delta\underline{\mathbb{L}}(\pi, P, Q)$
Bregman Information = Statistical Information
Integral Representation in terms of Statistical Information
$\mathbb{I}_f(P,Q) = \int_0^1 \Delta\underline{\mathbb{L}}^{0-1}(\pi, P, Q)\gamma(\pi)d\pi$
Primitive: $\Delta\underline{\mathbb{L}}^{0-1}(\pi, P, Q)$
Weight: $\gamma(\pi)$
f-divergence
Variational Representation:
$\mathbb{I}_f(P,Q) = \sup_{r \in \mathbb{R}^{\mathcal{X}}} \int_{\mathcal{X}} r dP - f^*(r)dQ$
Neyman Pearson Function $\beta(\alpha, P, Q)$
$\beta(\alpha) = \inf_{\pi \in (0,1]} \frac{1}{\pi}((1-\pi)\alpha + \pi - \underline{\mathbb{L}}(\pi, P, Q))$
$\underline{\mathbb{L}}(\pi, P, Q) = \min_{\alpha \in [0,1]} ((1-\pi)\alpha + \pi(1 - \beta(\alpha, P, Q))$
Partial Knowledge of $\pi \mapsto \Delta\underline{\mathbb{L}}^{0-1}(\pi, P, Q)$
e.g. $(\pi_i, V_{\pi_i})_{i=1}^n$
Generalised Pinsker Bounds
$\mathbb{I}_f(P,Q) \geq \Phi_f((\pi_i, V_{\pi_i}))_{i=1}^n$
where Φ_f only depends on f
Regret Bounds for Surrogate Losses
$B_w(\eta, \hat{\eta}) \geq \psi(c_0, \alpha) \vee \psi(c_0, -\alpha)$,
$\psi(c_0, \alpha) := B(c_0, c_0 + \alpha) = \underline{L}(c_0) - \underline{L}(c_0 + \alpha) + \alpha\underline{L}'(c_0)$
When $c_0 = 1/2$ and w is symmetric,
$B_w(\eta, \hat{\eta}) \geq \underline{L}(1/2) - \underline{L}(1/2 + \alpha)$

Figure 1: Diagrammatic summary of key relationships developed in the paper.

§5 A key technique we use is that of an integral representation. We show that integral representations of f-divergences and proper losses and statistical information are all essentially the same (Theorem 18). We explicitly compare the primitives for each of these representations and show their natural interpretation.

§6 The weight function view also illuminates various "graphical representations" of binary experiments, such as ROC curves. We unify several graphical representations for binary experiments and present new explicit formulae relating Bayes risk to the Neyman-Pearson function, which allows the transformation of risk curves to ROC curves and vice versa (Theorem 22).

§7 The various equivalences developed in the above sections are then used to derive new tight inequalities of interest in Machine Learning, The first is a We derive an explicit form for surrogate regret bounds for proper losses in terms of the weight function corresponding to the proper loss (Theorem 25). These are tight bounds on the conditional risk with respect to an arbitrary cost-sensitive misclassification loss when all is known is the value of the conditional risk with respect to an arbitrary proper loss. The result generalises existing results in two key ways. We also generalise the classical Pinsker inequality by deriving tight bounds on an arbitrary f-divergence when the value of several generalised variational divergences between the same distributions is known (Theorem 30). A side-effect is an explicit formula for the best possible bound on KL-divergence given knowledge of the classical variational divergence.

§8 Another representation of risks is a variational one. We systematically explore the relationship between Bayes risk and variational divergence, building upon classical results. An interesting consequence of our analysis is presented in Appendix H where we show that maximum mean discrepancy (MMD)—a kernel approach to hypothesis testing and divergence estimation—is essentially SVM learning in disguise. In doing so we present a novel, simple and interesting alternate derivation of the Support Vector Machine.

1.4 Notational Conventions

Here we record elementary notation and the conventions we adopt throughout the paper. Key notations are tabulated in table 1. We write $x \wedge y := \min(x,y)$, $x \vee y := \max(x,y)$, $(x)_+ := x \vee 0$, $(x)_- := x \wedge 0$ and the Iverson bracket $[\![p]\!] = 1$ if p is true and $[\![p]\!] = 0$ otherwise (Knuth, 1992). 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 $a < 0 < b$ (Antosik et al., 1973; Friedlander, 1982). The unit step $U(x) = \int_{-\infty}^{x} \delta(t) dt$. The real numbers are denoted $\mathbb{R}$, the non-negative reals $\mathbb{R}^+$ and the extended reals $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$; the rules of arithmetic with extended real numbers and the need for them in convex analysis are explained by Rockafellar (1970). Random variables are written in sans-serif font: S, X, Y. Sets are in calligraphic font: $\mathcal{X}$ (the "input" space), $\mathcal{Y}$ (the "label" space). Vectors are written in bold font: $\mathbf{a}, \boldsymbol{\alpha}, \boldsymbol{x} \in \mathbb{R}^m$. We will often have cause to take expectations ($\mathbb{E}$) of various functions over the random variable X. We write such quantities in blackboard bold: $\mathbb{I}$, $\mathbb{L}$, $\mathbb{B}$, $\mathbb{J}$. The elementary loss is ℓ, its conditional expectation w.r.t. Y is L and the full expectation (over the joint distribution $\mathbb{P}$ of (X,Y)) is $\mathbb{L}$. Lower bounds 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 appear 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: $w, W, \overline{W}$. Estimated quantities are hatted: $\hat{\eta}$. In several places we overload the notation. In all cases careful attention to the type of the arguments or subscripts reliably disambiguates.

Symbol	**Meaning**	**Defined**
I_ϕ	Perspective transform	(1)
(P,Q)	Binary experiment	§3
ℓ	Loss	§4.2
L	Conditional risk	§4.2
$\mathbb{L}$	Expected risk	§4.2
$\underline{L}$	Conditional Bayes risk	§4.2
$\underline{\mathbb{L}}$	Expected Bayes Risk	§4.2
$\mathbb{J}_\mu[\phi]$	Jensen gap	Th. 5
$\mathbb{I}_f(P,Q)$	f-divergence between P and Q	§3.2
$\phi^\diamond$	Csiszár dual of ϕ	(2)
$\phi^\star$	Legendre-Fenchel dual of ϕ	(3)
B_ϕ	Bregman divergence and regret	§4.4
TP_r, FN_r	True Positive / False Negative rate for test r	(10)
$\beta(\cdot,P,Q)$	Neyman-Pearson function for (P,Q)	(11)
r, τ	Test, Test statistic	§3.1
$\mathbb{B}_\phi(P,Q)$	Generative Bregman divergence	§3.3
$\mathbb{P}$	Joint distribution on $\mathcal{X}\times\mathcal{Y}$	§4.1
M	Reference measure for (P,Q) with prior π	§4.1
π	A priori probability of positive class	§4.1
η	Probability of positive class	§4.2
$\eta(\cdot)$	Conditional probability of positive class	§4.2
$T=(\eta,M;\ell)=(\pi,P,Q;\ell)$	Task	§4.2
$\hat{\eta}(\cdot)$	Estimator of $\eta(\cdot)$	§4.2
$\mathbb{B}_\phi(\mathsf{S})$	Bregman information of S	§4.5
$w(\cdot)$	Weight function for proper loss	§5.3
$\gamma(\cdot)$	Weight function for f-divergence	§5.1
$\Delta\underline{\mathbb{L}}(\eta,M)$	Statistical information	(20)
ℓ_c, L_c	Cost-sensitive mis-classification loss	(29),(30)
$\mathrm{ROC}(\tau)$	Receiver Operating Characteristic curve	(37)
$\mathrm{AUC}(\tau)$	Area Under the ROC Curve	(38)
$V_\pi(P,Q)$	Generalised Variational divergence	(49)

Table 1: Standard notation used throughout the paper.

2. Convex Functions and Their Representations

Many of the properties of divergences and losses are best understood through properties of the convex functions that define them. One aim of this paper is to explain and relate various divergences and losses by understanding the relationships between their primitive functions. The relevant definitions and theory of convex functions will be introduced as required. Any terms not explicitly defined can be found in books by Hiriart-Urruty and Lemaréchal (2001) or Rockafellar (1970).

A set $\mathcal{S} \subseteq \mathbb{R}^d$ is said to be *convex* if for all $\lambda \in [0,1]$ and for all points $s_1, s_2 \in \mathcal{S}$ the point $\lambda s_1 + (1-\lambda)s_2 \in \mathcal{S}$. A function $\phi : \mathcal{S} \to \mathbb{R}$ defined on a convex set $\mathcal{S}$ is said to be a (proper) *convex function* if[3] for all $\lambda \in [0,1]$ and points $s_1, s_2 \in \mathcal{S}$ the function ϕ satisfies

$$\phi(\lambda s_1 + (1-\lambda)s_2) \leq \lambda\phi(s_1) + (1-\lambda)\phi(s_2).$$

A function is said to be *concave* if $-\phi$ is convex.

The remainder of this section presents properties, representations and transformations of convex functions that will be used throughout this paper.

2.1 The Perspective Transform and the Csiszár Dual

When $\mathcal{S} = \mathbb{R}^+$ and $\phi : \mathbb{R}^+ \to \overline{\mathbb{R}}$ is convex, the *perspective transform* of ϕ is defined for $\tau \in \mathbb{R}^+$ via

$$I_\phi(s,\tau) := \begin{cases} \tau\phi(s/\tau), & \tau > 0, s > 0 \\ 0, & \tau = 0, s = 0 \\ \tau\phi(0), & \tau > 0, s = 0 \\ s\phi'_\infty, & \tau = 0, s > 0, \end{cases} \tag{1}$$

where $\phi(0) := \lim_{s\to 0}\phi(s) \in \overline{\mathbb{R}}$ and ϕ'_∞ is the *slope at infinity* defined as

$$\phi'_\infty := \lim_{s\to+\infty} \frac{\phi(s_0+s)-\phi(s_0)}{s} = \lim_{s\to+\infty}\frac{\phi(s)}{s}$$

for every $s_0 \in \mathcal{S}$ where $\phi(s_0)$ is finite. This slope at infinity is only finite when $\phi(s) = O(s)$, that is, when ϕ grows at most linearly as s increases. When ϕ'_∞ is finite it measures the slope of the linear asymptote. The function $I_\phi : [0,\infty)^2 \to \overline{\mathbb{R}}$ is convex in both arguments (Hiriart-Urruty and Lemaréchal, 1993b) and may take on the value $+\infty$ when s or τ is zero. It is introduced here because it will form the basis of the f-divergences described in the next section.[4]

The perspective transform can be used to define the *Csiszár dual* $\phi^\lozenge : [0,\infty) \to \overline{\mathbb{R}}$ of a convex function $\phi : \mathbb{R}^+ \to \mathbb{R}$ by letting

$$\phi^\lozenge(\tau) := I_\phi(1,\tau) = \tau\phi\left(\frac{1}{\tau}\right) \tag{2}$$

for all $\tau \in (0,\infty)$ and $\phi^\lozenge(0) := \phi'_\infty$. The original ϕ can be recovered from I_ϕ since $\phi(s) = I_f(s,1)$.

The convexity of the perspective transform I_ϕ in both its arguments guarantees the convexity of the dual $\phi^\lozenge$. Some simple algebraic manipulation shows that for all $s,\tau \in \mathbb{R}^+$

$$I_\phi(s,\tau) = I_{\phi^\lozenge}(\tau,s).$$

This observation leads to a natural definition of symmetry for convex functions. We will call a convex function $\lozenge$-*symmetric* (or simply *symmetric* when the context is clear) when its perspective transform is symmetric in its arguments. That is, ϕ is $\lozenge$-symmetric when $I_\phi(s,\tau) = I_\phi(\tau,s)$ for all $s,\tau \in [0,\infty)$. Equivalently, ϕ is $\lozenge$-symmetric if and only if $\phi^\lozenge = \phi$.

[3] The restriction of the values of ϕ to $\mathbb{R}$ will be assumed throughout unless explicitly stated otherwise. This implies the properness of ϕ since it cannot take on the values $-\infty$ or $+\infty$.

[4] The perspective transform is closely related to *epi-multiplication* which is defined for all $\tau \in [0,\infty)$ and (proper) convex functions ϕ to be $\tau \otimes \phi := s \mapsto \tau\phi(s/\tau)$ for $\tau > 0$ and is 0 when $\tau = s = 0$ and $+\infty$ otherwise. Bauschke et al. (2008) summarise the properties of this operation and its relationship to other operations on convex functions.

2.2 The Legendre-Fenchel Dual Representation

A second important dual operator for convex functions is the *Legendre-Fenchel (LF) dual*. The LF dual $\phi^\star$ of a function $\phi : \mathcal{S} \to \mathbb{R}$ is a function defined by

$$\phi^\star(s^\star) := \sup_{s\in\mathcal{S}}\{\langle s, s^\star\rangle - \phi(s)\}. \tag{3}$$

The LF dual of any function is convex and, if the function ϕ is convex and closed then the *LF bidual* is a faithful representation of the original function. That is,

$$\phi^{\star\star}(s) = \sup_{s^\star\in\mathcal{S}^\star}\{\langle s^\star, s\rangle - \phi^\star(s^\star)\} = \phi(s).$$

When $\phi\colon \mathcal{S} \to \mathbb{R}$, $\mathcal{S} \subseteq \mathbb{R}$, is a function of a real argument s and the derivative $\phi'(s)$ exists, the Legendre-Fenchel conjugate $\phi^\star$ is given by the *Legendre transform* (Hiriart-Urruty and Lemaréchal, 2001; Rockafellar, 1970)

$$\phi^\star(s) = s\cdot(\phi')^{-1}(s) - \phi\left((\phi')^{-1}(s)\right).$$

2.3 Integral Representations

In this paper we are primarily concerned with convex and concave functions defined on subsets of the real line. A central tool in their analysis is the integral form of their Taylor expansion. Here, ϕ' and ϕ'' denote the first and second derivatives of ϕ respectively.

Theorem 1 (Taylor's Theorem) *Let $\mathcal{S} = [s_0, s]$ be a closed interval of $\mathbb{R}$ and let $\phi : \mathcal{S} \to \mathbb{R}$ be differentiable on $[s_0, s]$ and twice differentiable on (s_0, s). Then*

$$\phi(s) = \phi(s_0) + \phi'(s_0)(s - s_0) + \int_{s_0}^{s}(s-t)\,\phi''(t)\,dt. \tag{4}$$

The argument s appears in the limits of integral in the above theorem and consequently can be awkward to work with. Also, it will be useful to expand ϕ about some point not at the end of the interval of integration. The following corollary of Taylor's theorem removes these problems by introducing piecewise linear terms of the form $(s-t)_+ = (s-t)\vee 0$.

Corollary 2 (Integral Representation I) *Suppose $-\infty < a < b < \infty$ and let $\phi : [a, b] \to \mathbb{R}$ be a twice differentiable function. Then, for all $s, s_0 \in [a, b]$ we have*

$$\phi(s) = \phi(s_0) + \phi'(s_0)(s - s_0) + \int_a^b \phi_{s_0}(s,t)\,\phi''(t)\,dt, \tag{5}$$

where

$$\phi_{s_0}(s,t) := \begin{cases} (s-t) & s_0 < t \le s \\ (t-s) & s < t \le s_0 \\ 0 & \textit{otherwise} \end{cases}$$

is piecewise linear and convex in s for each $s_0, t \in [a, b]$.

This result is a consequence of the way in which ϕ_t effectively restricts the limits of integration to the interval $(s_0,s) \subseteq [a,b]$ or $(s,s_0) \subseteq [a,b]$ depending on whether $s_0 < s$ or $s_0 \geq s$ with appropriate reversal of the sign of $(s-t)$.

When $a = 0$ and $b = 1$ a second integral representation for the unit interval can be derived from (5) that removes the term involving ϕ'.

Corollary 3 (Integral Representation II) *A twice differentiable function $\phi : [0,1] \to \mathbb{R}$ can be expressed as*

$$\phi(s) = \phi(0) + (\phi(1) - \phi(0))s - \int_0^1 \psi(s,t)\,\phi''(t)\,dt, \tag{6}$$

where $\psi(s,t) = (1-t)s \wedge (1-s)t$ is piecewise linear and concave in $s \in [0,1]$ for each $t \in [0,1]$.

The result follows by integration by parts of $t\phi''(t)$. The proof can be found in Appendix A.1. It is used in Section 5 below to obtain an integral representation of losses for binary class probability estimation. This representation can be traced back to Temple (1954) who notes that the kernel $\psi(s,t)$ is the Green's function for the differential equation $\psi'' = 0$ with boundary conditions $\psi(a) = \psi(b) = 0$.

Both these integral representations state that the non-linear part of ϕ can be expressed as a weighted integral of piecewise linear terms ϕ_{s_0} or ψ. When we restrict our attention to convex ϕ we are guaranteed the "weights" $\phi''(t)$ for each of these terms are non-negative. Since the measures of risk, information and divergence we examine below do not depend on the linear part of these expansions we are able to identify convex functions with the weights $w(t) = \phi''(t)$ that define their non-linear part. The sets of piecewise linear functions $\{\phi_{s_0}(s,t)\}_{t\in[a,b]}$ and $\{\psi(s,t)\}_{t\in[0,1]}$ can be thought of as families of "primitive" convex functions from which others can be built through their weighted combination. Representations like these are often called *Choquet representations* after work by Choquet (1953) on the representation of compact convex spaces (Phelps, 2001).

2.4 Representations for Non-Differentiable Convex Functions

It is possible to weaken the conditions on the representation results so they hold for continuous but not necessarily differentiable functions. As much of this paper deals with functions that fall into this category—namely general convex functions—being able to generalise these results is essential in order to understand the weight functions corresponding to the primitive f-divergences and loss functions. We will briefly discuss these generalisations and introduce some conventions for interpreting subsequent results in an effort to avoid too many distracting technicalities.

The convention for the remainder of this paper is that the *first derivative* of a convex function ϕ over $\mathbb{R}$ is to be interpreted as a right derivative. That is, we will take $\phi'(t)$ to be $\phi'_+(t) := \lim_{\varepsilon\downarrow 0} \frac{\phi(t)-\phi(t+\varepsilon)}{\varepsilon}$. Theorem 24.1 of Rockafellar (1970) guarantees that this derivative exists and is non-decreasing and right continuous on the domain of ϕ. It is therefore possible to define a Lebesgue-Stieltjes measure $\lambda_\phi((a,b]) := \phi'(b) - \phi'(a)$ for intervals in the domain of ϕ.

Second derivatives of convex ϕ are only ever used within integrals to "weight" the contribution of the non-negative, piecewise linear functions $\phi_{s_0}(\cdot,t)$ and $\psi(\cdot,t)$ discussed above. Thus, we write $\int_a^b f(t)\,\phi''(t)\,dt$ as a short-hand for the Lebesgue-Stieltjes integral $\int_a^b f(t)\,d\lambda_\phi(t)$. For simplicity, we will often speak of weight "functions" being equal to the second derivative of general convex functions. As we only ever consider linear operators on these weight functions, it is unproblematic to treat second derivatives as Schwartz distributions or "generalised functions" (Antosik et al., 1973;

Friedlander, 1982) and add, scale, and evaluate them like normal functions. The most exotic of these we will consider explicitly are the weight functions corresponding to the primitive ϕ_{s_0} and ψ functions. They correspond to Dirac delta distributions $\delta(\cdot)$ as defined in Section 1.4.

As Liese and Vajda (2006) carefully show, it is possible to derive generalised versions of the integral representations using the interpretations above. Of course, when the functions ϕ are twice differentiable these interpretations and generalised results coincide with those for the usual first and second derivatives.

2.5 Bregman Divergence

Bregman divergences are a generalisation of the notion of distances between points. Given a differentiable[5] convex function $\phi : \mathcal{S} \to \mathbb{R}$ and two points $s_0, s \in \mathcal{S}$ the *Bregman divergence*[6] *of s from s_0* is defined to be

$$B_\phi(s, s_0) := \phi(s) - \phi(s_0) - \langle s - s_0, \nabla\phi(s_0) \rangle \,, \tag{7}$$

where $\nabla\phi(s_0)$ is the gradient of ϕ at s_0. A concise summary of many of the properties of Bregman divergences is given by Banerjee et al. (2005b, Appendix A); see also Censor and Zenios (1997). In particular, Bregman divergences always satisfy $B_\phi(s, s_0) \geq 0$ and $B_\phi(s_0, s_0) = 0$ for all $s, s_0 \in \mathcal{S}$, regardless of the choice of ϕ. They are not always metrics, however, as they do not always satisfy the triangle inequality and their symmetry depends on the choice of ϕ.

When $\mathcal{S} = \mathbb{R}$ and ϕ is twice differentiable, comparing the definition of a Bregman divergence in (7) to the integral representation in (4) reveals that Bregman divergences between real numbers can be defined as the non-linear part of the Taylor expansion of ϕ. Rearranging (4) shows that for all $s, s_0 \in \mathbb{R}$

$$\int_{s_0}^{s} (s - t)\, \phi''(t) dt = \phi(s) - \phi(s_0) - (s - s_0)\phi'(s_0) = B_\phi(s, s_0) \tag{8}$$

since $\nabla\phi = \phi'$ and the inner product is simply multiplication over the reals. This result also holds for more general convex sets $\mathcal{S}$. Importantly, it intuitively shows why the following holds (because the Bregman divergence depends only on the *nonlinear* part of the Taylor expansion).

Theorem 4 *Let ϕ and ψ both be real-valued, differentiable convex functions over the convex set $\mathcal{S}$ such that $\phi(s) = \psi(s) + as + b$ for some $a, b \in \mathbb{R}$. Then, for all s and s_0, $B_\phi(s, s_0) = B_\psi(s, s_0)$.*

A proof can be obtained directly by substituting and expanding ψ in the definition of a Bregman divergence.

Equation 8 also shows why $B(s, s_0)$ is decreasing as $|s - s_0|$ decreases (a fact we will exploit later): since $\phi''(t) \geq 0$ for all t, if $s_0 < s$, then the integrand in (8) is always non-negative and the result is immediate by the nature of integration. If $s_0 > s$, a similar argument holds.

2.6 Jensen's Inequality and the Jensen Gap

A central inequality in the study of convex functions is Jensen's inequality. It relates the expectation of a convex function applied to a random variable to the convex function evaluated at its mean. We will denote by $\mathbb{E}_\mu[\cdot] := \int_{\mathcal{S}} \cdot \, d\mu$ expectation over $\mathcal{S}$ with respect to a probability measure μ over $\mathcal{S}$.

[5] Technically, ϕ need only be differentiable on the relative interior $\mathrm{ri}(\mathcal{S})$ of $\mathcal{S}$. We omit this requirement for simplicity and because it is not relevant to this discussion.

[6] Named in reference to Bregman (1967) although he was not the first to consider such an equation, at least in the one dimensional case; confer Brunk et al. (1957, p.838).

Theorem 5 (Jensen's Inequality) *Let $\phi : \mathcal{S} \to \mathbb{R}$ be a convex function, μ be a distribution and S be an $\mathcal{S}$-valued random variable (measurable w.r.t. μ) such that $\mathbb{E}_\mu[|\mathsf{S}|] < \infty$. Then*

$$\mathbb{J}_\mu[\phi] := \mathbb{E}_\mu[\phi(\mathsf{S})] - \phi(\mathbb{E}_\mu[\mathsf{S}]) \geq 0. \tag{9}$$

The proof is straight-forward and can be found in (Dudley, 2003, §10.2). Jensen's inequality can also be used to characterise the class of convex functions. If ϕ is a function such that (9) holds for all random variables and distributions then ϕ must be convex.[7] Intuitively, this connection between expectation and convexity is natural since expectation can be seen as an operator that takes convex combinations of random variables.

We will call the difference $\mathbb{J}_\mu[\phi]$ the *Jensen gap for ϕ when $\mathsf{S} \sim \mu$*. Many measures of divergence and information studied in the subsequent sections can be expressed as the Jensen gap of some convex function. Due to the linearity of expectation, the Jensen gap is insensitive to the addition of affine terms to the convex function that defines it:

Theorem 6 *Let $\phi : \mathcal{S} \to \mathbb{R}$ be convex function and S and μ be as in Theorem 5. Then for each $a, b \in \mathbb{R}$ the convex function $\psi(s) := \phi(s) + as + b$ satisfies $\mathbb{J}_\mu[\phi(\mathsf{S})] = \mathbb{J}_\mu[\psi(\mathsf{S})]$.*

The proof is a consequence of the definition of the Jensen gap and the linearity of expectations and can be found in Appendix A.2. An implication of this theorem is that when considering sets of convex functions as parameters to the Jensen gap operator they only need be identified by their non-linear part. Thus, the Jensen gap operator can be seen to impose an equivalence relation over convex functions where two convex functions are equivalent if they have the same Jensen gap, that is, if their difference is affine.

In light of the two integral representations in Section 2.3, this means the Jensen gap only depends on the integral terms in (5) and (6) and so is completely characterised by the weights provided by ϕ''. Specifically, for suitably differentiable $\phi : [a, b] \to \mathbb{R}$ we have

$$\mathbb{J}_\mu[\phi(\mathsf{S})] = \int_a^b \mathbb{J}_\mu[\phi_{s_0}(\mathsf{S}, t)]\, \phi''(t)\, dt.$$

Since several of the measures of divergence, information and risk we analyse can be expressed as a Jensen gap, this observation implies that these quantities can be identified with the weights provided by ϕ'' as it is these that completely determine the measure's behaviour.

3. Binary Experiments and Measures of Divergence

The various properties of convex functions developed in the previous section have many implications for the study of statistical inference. We begin by considering *binary experiments* (P, Q) where P and Q are probability measures[8] over a common space $\mathcal{X}$. We will consider P the distribution over *positive* instances and Q the distribution over *negative* instances. The densities of P and Q with respect to some third reference distribution M over $\mathcal{X}$ will be defined by $dP = p\, dM$ and $dQ = q\, dM$ respectively. Unless stated otherwise we will assume that P and Q are both absolutely continuous

[7] This can be seen by considering a distribution with a finite, discrete set of points as its support and applying Theorem 4.3 of Rockafellar (1970).

[8] We intentionally avoid too many measure theoretic details for the sake of clarity. Appropriate σ-algebras and continuity can be assumed where necessary.

with respect to M. (One can always choose M to ensure this by setting $M := (P+Q)/2$; but see the next section.)

There are several ways in which the "separation" of P and Q in a binary experiment can be quantified. Intuitively, these all measure the difficulty of distinguishing between the two distributions using instances drawn from their mixture. The further apart the distributions are the easier discrimination becomes. This intuition is made precise through the connections with risk and MMD later in Appendix H.

A central statistic in the study of binary experiments and statistical hypothesis testing is the likelihood ratio dP/dQ. As the following section outlines, the likelihood ratio is, in the sense of preserving the distinction between P and Q, the "best" mapping from an arbitrary space $\mathcal{X}$ to the real line.

3.1 Statistical Tests and the Neyman-Pearson Lemma

In the context of a binary experiment (P,Q), a *statistical test* is any function that assigns each instance $x \in \mathcal{X}$ to either P or Q. We will use the labels 1 and 0 for P and Q respectively and so a statistical test is any function $r : \mathcal{X} \to \{0,1\}$. In machine learning, a function of this type is usually referred to as a *classifier*. The link between tests and classifiers is explored further in Section 4.

Each test r partitions the instance space $\mathcal{X}$ into positive and negative *prediction sets*:

$$\begin{aligned} \mathcal{X}_r^+ &:= \{x \in \mathcal{X} : r(x) = 1\}, \\ \mathcal{X}_r^- &:= \{x \in \mathcal{X} : r(x) = 0\}. \end{aligned}$$

There are four *classification rates* associated with these predictions sets: the true positive rate (TP), true negative rate (TN), false positive rate (FP) and the false negative rate (FN). For a given test r they are defined as follows:

$$\begin{aligned} \mathrm{TP}_r &:= P(\mathcal{X}_r^+), \quad \mathrm{FP}_r := Q(\mathcal{X}_r^+), \\ \mathrm{FN}_r &:= P(\mathcal{X}_r^-), \quad \mathrm{TN}_r := Q(\mathcal{X}_r^-). \end{aligned} \tag{10}$$

The subscript r will be dropped when the test is clear by the context. Since P and Q are distributions over $\mathcal{X} = \mathcal{X}_r^+ \cup \mathcal{X}_r^-$ and the positive and negative sets are disjoint we have that $\mathrm{TP} + \mathrm{FN} = 1$ and $\mathrm{FP} + \mathrm{TN} = 1$. As a consequence, the four values in (10) can be summarised by choosing one from each column.

Often, statistical tests are obtained by applying a threshold τ_0 to a real-valued *test statistic* $\tau : \mathcal{X} \to \mathbb{R}$. In this case, the statistical test is $r(x) = [\![\tau(x) \geq \tau_0]\!]$. This leads to parameterised forms of prediction sets $\mathcal{X}_\tau^y(\tau_0) := \mathcal{X}^y_{[\![\tau \geq \tau_0]\!]}$ for $y \in \{+,-\}$, and the classification rates $\mathrm{TP}_\tau(\tau_0)$, $\mathrm{FP}_\tau(\tau_0)$, $\mathrm{TN}_\tau(\tau_0)$, and $\mathrm{TP}_\tau(\tau_0)$ which are defined analogously. By varying the threshold parameter a range of classification rates can be achieved. This observation leads to a well known graphical representation of test statistics known as the ROC curve, which is discussed further in Section 6.1.

A natural question is whether there is a "best" statistical test or test statistic to use for binary experiments. This is usually formulated in terms of a test's power and size. The *power* β_r of the test r for a particular binary experiment (P,Q) is a synonym for its true positive rate (that is, $\beta_r := \mathrm{TP}_r$ and so $1 - \beta_r := \mathrm{FN}_r$[9]) and the *size* α_r of same test is just its false positive rate $\alpha_r := \mathrm{FP}_r$. Here,

[9] This is opposite to the usual definition of β_r in the statistical literature. Usually, $1-\beta_r$ is used to denote the power of a test. We have chosen to use β_r for the power (true positive rate) as this makes it easier to compare with ROC curves and it is consistent with the usage of Torgersen (1991).

"best" is considered to be the *most powerful* (MP) test of a given size (Bickel and Doksum, 2001, §4.2). That is, a test r is considered MP of size $\alpha \in [0,1]$ if, $\alpha_r = \alpha$ and for all other tests r' such that $\alpha_{r'} \le \alpha$ we have $1-\beta_r \le 1-\beta_{r'}$. We will denote by $\beta(\alpha) := \beta(\alpha, P, Q)$ the true positive rate of an MP test between P (the alternative hypothesis) and Q (the null hypothesis) at Q with significance α. Torgersen (1991) calls $\beta(\cdot, P, Q)$ the *Neyman-Pearson function for the dichotomy* (P,Q). Formally, for each $\alpha \in [0,1]$, the Neyman-Pearson function β measures the largest true positive rate TP_r of any measurable classifier $r \colon \mathcal{X} \to \{-1,1\}$ that has false positive rate FP_r at most α. That is,

$$\beta(\alpha) = \beta(\alpha, P, Q) := \sup_{r \in \{-1,1\}^{\mathcal{X}}} \{\mathrm{TP}_r \ : \ \mathrm{FP}_r \le \alpha\}. \tag{11}$$

The Neyman-Pearson lemma (Neyman and Pearson, 1933) shows that the likelihood ratio $\tau^*(x) = dP/dQ(x)$ is the most powerful test for each choice of threshold τ_0. Since each choice of $\tau_0 \in \mathbb{R}$ results in a test $[\![dP/dQ \ge \tau_0]\!]$ of some size $\alpha \in [0,1]$ we have that[10]

$$\beta(\mathrm{FP}_{\tau^*}(\tau_0)) = \mathrm{TP}_{\tau^*}(\tau_0) \tag{12}$$

and so varying τ_0 over $\mathbb{R}$ results in a maximal ROC curve. This too is discussed further in Section 6.1.

The Neyman-Pearson lemma thus identifies the likelihood ratio dP/dQ as a particularly useful statistic. Given an experiment (P,Q) it is, in some sense, the best mapping from the space $\mathcal{X}$ to the reals. The next section shows how this statistic can be used as the basis for a variety of divergence measures between P and Q.

3.2 Csiszár f-divergences

The class of *f-divergences* (Ali and Silvey, 1966; Csiszár, 1967) provide a rich set of relations that can be used to measure the separation of the distributions in a binary experiment. 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

$$\mathbb{I}_f(P,Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] = \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ \tag{13}$$

when P is absolutely continuous with respect to Q and equals ∞ otherwise.[11]

The above definition is not completely well-defined as the behaviour of f is not specified at the endpoints of $(0,\infty)$. This is remedied via the perspective transform of f, introduced in Section 2.1 above which defines the limiting behaviour of f. Given convex $f : (0,\infty) \to \mathbb{R}$ such that $f(1) = 0$ the *f-divergence of P from Q* is

$$\mathbb{I}_f(P,Q) := \mathbb{E}_M\left[I_f(p,q)\right] = \mathbb{E}_{\mathsf{X}\sim M}\left[I_f(p(\mathsf{X}), q(\mathsf{X}))\right], \tag{14}$$

where I_f is the perspective transform of f (see (1)).

[10] Equation (43) in Section 6.3 below, shows that $\beta(\alpha)$ is the lower envelope of a family of linear functions of α and is thus concave and continuous. Hence, the equality in (12) holds.

[11] Liese and Miescke (2008, pg. 34) give a definition that does not require absolute continuity.

The restriction that $f(1)=0$ in the above definition is only present to normalise $\mathbb{I}_f$ so that $\mathbb{I}_f(Q,Q)=0$ for all distributions Q. We can extend the definition of f-divergences to all convex f by performing the normalisation explicitly. Since $f\left(\mathbb{E}_Q\left[dP/dQ\right]\right)=f(1)$ this is done most conveniently through the definition of the Jensen gap for the function f applied to the random variable dP/dQ with distribution Q. That is, for all convex $f:(0,\infty)\to\mathbb{R}$ and for all distributions P and Q

$$\mathbb{J}_Q\left[f\left(\frac{dP}{dQ}\right)\right]=\mathbb{I}_f(P,Q)-f(1). \tag{15}$$

Due to the issues surrounding the behaviour of f at 0 and ∞ the definitions in (13), (14) and (15) are not entirely equivalent. When it is necessary to deal with the limiting behaviour, the definition in (14) will be used. However, the version in (15) will be most useful when drawing connections between f-divergences and various definitions of information in Section 4 below.

Several properties of f-divergence can be immediately obtained from the above definitions. The symmetry of the perspective I_f in (2) means that

$$\mathbb{I}_f(P,Q)=\mathbb{I}_{f^\diamond}(Q,P) \tag{16}$$

for all distributions P and Q, where $f^\diamond$ is the Csiszár dual of f. The non-negativity of the Jensen gap ensures that $\mathbb{I}_f(P,Q)\geq 0$ for all P and Q. Furthermore, the affine invariance of the Jensen gap (Theorem 6) implies the same affine invariance for f-divergences.

Several well-known divergences correspond to specific choices of the function f (Ali and Silvey, 1966, §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 14. It is the only f-divergence that is a true metric on the space of distributions over $\mathcal{X}$ (Khosravifard et al., 2007) and gets its name from its equivalent definition in the variational form

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

(Some authors define V without the 2 above.) This form of the variational divergence is discussed further in Section 8. Furthermore, the variational divergence is one of a family of "primitive" f-divergences discussed in Section 5. 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 14. Others are given in Table 2 in Section 5.4.

3.3 Generative Bregman Divergences

Another measure of the separation of distributions can be defined as the expected Bregman divergence between the densities p and q with respect to the reference measure M. Given a convex function $\phi:\mathbb{R}^+\to\mathbb{R}$ the *generative Bregman divergence* between the distributions P and Q is (confer (14))

$$\mathbb{B}_\phi(P,Q):=\mathbb{E}_M\left[B_\phi(p,q)\right]=\mathbb{E}_{\mathsf{X}\sim M}\left[B_\phi(p(\mathsf{X}),q(\mathsf{X}))\right].$$

We call this Bregman divergence "generative" to distinguish it from the "discriminative" Bregman divergence introduced in Section 4 below, where the adjectives "generative" and "discriminative" are explained further.

Csiszár (1995) notes that there is only one divergence common to the class of f-divergences and the generative Bregman divergences. In this sense, these two classes of divergences are "orthogonal" to each other. Their only common point is when the respective convex functions satisfy $f(t) = \phi(t) = t\ln t - at + b$ (for $a, b \in \mathbb{R}$) in which case both $\mathbb{I}_f$ and $\mathbb{B}_\phi$ are the KL divergence.

4. Risk and Statistical Information

The above discussion of f-divergences assumes an arbitrary reference measure M over the space $\mathcal{X}$ to define the densities p and q. In the previous section, the choice of reference measure was irrelevant since f-divergences are invariant to this choice.

In this section an assumption is made that adds additional structure to the relationship between P and Q. Specifically, we assume that the reference measure M is a mixture of these two distributions. That is, $M = \pi P + (1-\pi)Q$ for some $\pi \in (0,1)$. In this case, by construction, P and Q are absolutely continuous with respect to M. Intuitively, this can be seen as defining a distribution over the observation space $\mathcal{X}$ by first tossing a coin with a bias π for heads and drawing observations from P on heads or Q on tails.

This extra assumption allows us to interpret a binary experiment (P,Q) as a generalised *supervised binary task* (π,P,Q) where the positive $(y=1)$ and negative $(y=-1)$ *labels* $y \in \mathcal{Y} := \{-1,1\}$ are paired with *observations* $x \in \mathcal{X}$ through a joint distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$. (We formally define a task later in terms of an experiment plus loss function.) Given an observation drawn from $\mathcal{X}$ according to M, it is natural to try to predict its corresponding label or estimate the probability it was drawn from P.

Below we will introduce risk, regret, and proper losses and show how these relate to discriminative Bregman divergence. We then show the connection between the generative view (f-divergence between the class conditional distributions) and Bregman divergence.

4.1 Generative and Discriminative Views

Traditionally, the joint distribution $\mathbb{P}$ of inputs $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$ is used as the starting point for analysing risk in statistical learning theory. In order to better link risks to divergences, in our analysis we will consider two related representations of $\mathbb{P}$.

The *generative* view decomposes the joint distribution $\mathbb{P}$ into two *class-conditional distributions* defined as $P(X) := \mathbb{P}(X|y=1)$, $Q(X) := \mathbb{P}(X|y=-1)$ for all $X \subseteq \mathcal{X}$ and a mixing probability or *prior* $\pi := \mathbb{P}(\mathcal{X}, y=1)$. The *discriminative* representation decomposes the joint distribution into an *observation distribution* $M(X) := \mathbb{P}(X,\mathcal{Y})$ for all $X \subseteq \mathcal{X}$ and an *observation-conditional density* or *posterior* $\eta(x) = \frac{dH}{dM}(x)$ where $H(X) := \mathbb{P}(X, y=1)$. The terms "generative" and "discriminative" are used here to suggest a distinction made by Ng and Jordan (2002): in the generative case, the aim is to model the class-conditional distributions P and Q and then use Bayes rule to compute the most likely class; in the discriminative case the focus is on estimating $\eta(x)$ directly. Although we are not directly interested in this paper in the problems of modelling or estimating we find the distinction a useful one.[12]

[12] The generative-discriminative distinction usually refers to whether one is modelling the process that generates each class-conditional distribution, or instead wishes solely to perform well on a discrimination task (Drummond, 2006; Lasserre et al., 2006; Minka, 2005; Rubinstein and Hastie, 1997). There has been some recent work relating the two in the sense that if the class conditional distributions are well estimated then will one perform well in discrimination (Long and Servedio, 2006; Long et al., 2006; Goldberg, 2001; Palmer and Goldberg, 2006).

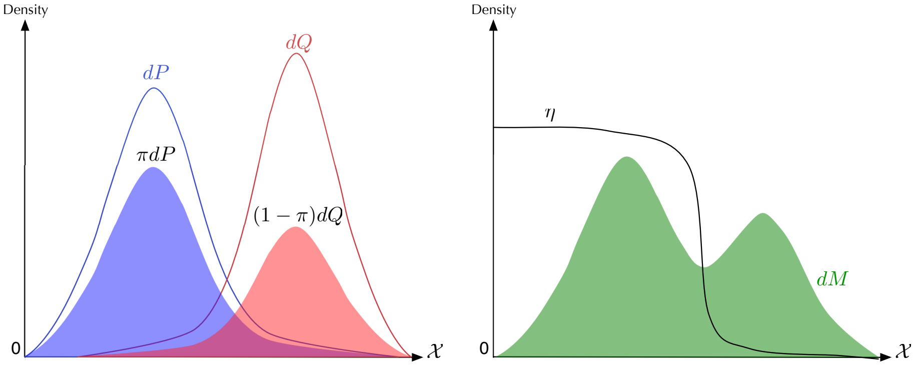


Figure 2: The generative and discriminative view of binary experiments.

Both these decompositions are exact since $\mathbb{P}$ can be reconstructed from either. Also, translating between them is straight-forward, since

$$M = \pi P + (1-\pi)Q \quad \text{and} \quad \eta = \pi \frac{dP}{dM},$$

so we will often swap between (η, M) and (π, P, Q) as arguments to functions for risk, divergence and information. A graphical representation of the generative and discriminative views of a binary experiment is shown in Figure 2.

The posterior η is closely related to the likelihood ratio dP/dQ in the supervised binary task setting. For each choice of $\pi \in (0,1)$ this relationship can be expressed by a mapping $\lambda_\pi : [0,1] \to [0,\infty]$ and its inverse λ_π^{-1} defined by

$$\begin{aligned} \lambda_\pi(c) &:= \frac{1-\pi}{\pi}\frac{c}{1-c}, \\ \lambda_\pi^{-1}(t) &= \frac{\pi t}{\pi t + 1 - \pi} \end{aligned} \tag{17}$$

for all $c \in [0,1)$ and $t \in [0,\infty)$, and $\lambda_\pi(1) := \infty$. Thus

$$\eta = \lambda_\pi^{-1}\left(\frac{dP}{dQ}\right) \quad \text{and, conversely,} \quad \frac{dP}{dQ} = \lambda_\pi(\eta).$$

These will be used later when relating f-divergences and risk.

4.2 Estimators and Risk

We will call a (M-measurable) function $\hat{\eta} : \mathcal{X} \to [0,1]$ a class probability *estimator*. Overloading the notation slightly, we will also use $\hat{\eta} = \hat{\eta}(x) \in [0,1]$ to denote an *estimate* for a specific observation $x \in \mathcal{X}$. Many of the subsequent arguments rely on this conditional perspective.

Estimate quality is assessed using a *loss function* $\ell : \mathcal{Y} \times [0,1] \to \mathbb{R}$ and the loss of the estimate $\hat{\eta}$ with respect to the label $y \in \mathcal{Y}$ is denoted $\ell(y, \hat{\eta})$. If $\eta \in [0,1]$ is the probability of observing the label $y = 1$ then the *point-wise risk* of the estimate $\hat{\eta} \in [0,1]$ is defined to be the η-average of the point-wise loss for $\hat{\eta}$:

$$L(\eta, \hat{\eta}) := \mathbb{E}_{\mathsf{Y}\sim\eta}[\ell(\mathsf{Y}, \hat{\eta})] = \ell(0, \hat{\eta})(1-\eta) + \ell(1, \hat{\eta})\eta. \tag{18}$$

(This is what Steinwart 2006 calls the *inner risk*.) When $\eta : \mathcal{X} \to [0,1]$ is an observation-conditional density, taking the M-average of the point-wise risk gives the *(full) risk* of the estimator $\hat{\eta}$:

$$\begin{aligned} \mathbb{L}(\eta, \hat{\eta}, M) &:= \mathbb{E}_M\left[L(\eta, \hat{\eta})\right] = \mathbb{E}_{\mathsf{X}\sim M}[L(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))] \\ &= \int_{\mathcal{X}} L(\eta(x), \hat{\eta}(x))\, dM(x) =: \mathbb{L}(\pi, \hat{\eta}, P, Q). \end{aligned}$$

The convention of using ℓ, L and $\mathbb{L}$ for the loss, point-wise and full risk is used throughout this paper. Any names or parameters associated to ℓ will be propagated to L and $\mathbb{L}$.

We call the combination of a loss ℓ and the distribution $\mathbb{P}$ a *task* and denote it discriminatively as $T = (\eta, M; \ell)$ or generatively as $T = (\pi, P, Q; \ell)$. A natural measure of the difficulty of a task is its minimal achievable risk, or *Bayes risk*:

$$\underline{\mathbb{L}}(\eta, M) = \underline{\mathbb{L}}(\pi, P, Q) := \inf_{\hat{\eta} \in [0,1]^{\mathcal{X}}} \mathbb{L}(\eta, \hat{\eta}, M) = \mathbb{E}_{\mathsf{X}\sim M}\left[\underline{L}(\eta(\mathsf{X}))\right],$$

where

$$[0,1] \ni \eta \mapsto \underline{L}(\eta) := \inf_{\hat{\eta} \in [0,1]} L(\eta, \hat{\eta})$$

is the *point-wise Bayes risk*. Note the use of the underline on $\underline{\mathbb{L}}$ and $\underline{L}$ to indicate that the corresponding functions $\mathbb{L}$ and L are minimised.

4.3 Proper Losses

If $\hat{\eta}$ is to be interpreted as an estimate of the true positive class probability η then it is desirable to require that $L(\eta, \hat{\eta})$ be minimised when $\hat{\eta} = \eta$ for all $\eta \in [0,1]$. Losses that satisfy this constraint are said to be *Fisher consistent* and are known as *proper scoring rules* (Buja et al., 2005; Gneiting and Raftery, 2007). To use common machine learning terminology we will refer to Fisher consistent losses as *proper losses*. This implies that a proper loss ℓ satisfies $\underline{L}(\eta) = L(\eta, \eta)$ for all $\eta \in [0,1]$.

There are a few properties of losses that we will require to establish certain key theorems below. The first of these is that we will say a loss is *fair* whenever $\eta \mapsto \ell(0, \eta)$ and $\eta \mapsto \ell(1, \eta)$ are, respectively, right continuous at 0 and left continuous at 1, and

$$\ell(0,0) = \ell(1,1) = 0.$$

That is, no loss incurred for perfect prediction and there are no sudden "jumps" in penalty for near-perfect prediction. The main place fairness is relied upon is in the integral representation of Theorem 16 where it is used to get rid of some constants of integration. In order to explicitly construct a proper loss from its associated "weight function" as shown in Theorem 17 we will require that the loss be *definite*, that is, its point-wise Bayes risk at 0 and 1 must be bounded from below:

$$\underline{L}(0) > -\infty\,,\ \underline{L}(1) > -\infty.$$

Since properness of a loss ensures $\underline{L}(\eta) = L(\eta, \eta)$ we see that a fair proper loss is necessarily definite since $L(0,0) = \ell(0,0) = 0 > -\infty$, and similarly for $L(1,1)$. Conversely, if a proper loss is definite then the finite values $\ell(0,0)$ and $\ell(1,1)$ can be subtracted from $\ell(0,\cdot)$ and $\ell(1,\cdot)$ to make it fair.

Finally, for Theorem 7 below to hold at the endpoints of the unit interval we require a loss to be *regular*, that is,

$$\lim_{\eta \searrow 0} \eta \ell(1, \eta) = \lim_{\eta \nearrow 1} (1-\eta) \ell(0, \eta) = 0. \tag{19}$$

Intuitively, this condition ensures that making mistakes on events that never happen should not incur a penalty. It is not difficult to show that any fair, definite loss is also regular (thus, a proper and fair loss is also regular) but the converse does not hold. Since properness and fairness imply definiteness and regularity, most of the situations we consider in the remainder of this paper will involve losses which are both proper and fair.

Proper losses for probability estimation and surrogate margin losses (confer Bartlett et al. 2006) for classification are closely related. (Surrogate margin losses are considered in more detail in Appendix D.) Buja et al. (2005) note that "the surrogate criteria of classification are exactly the primary criteria of class probability estimation" and that most commonly used surrogate margin losses are just proper losses mapped from $[0,1]$ to $\mathbb{R}$ via a link function. The main exceptions are hinge losses;[13] Buja et al. (2005, pg. 4) state that SVMs are "the only case that truly bypasses estimation of class probabilities and directly aims at classification." However, commonly used margin losses of the form $\phi(yF(x))$ are a more restrictive class than proper losses since, as Buja et al. (2005, §23) note, "[t]his dependence on the margin limits all theory and practice to a symmetric treatment of class 0 and class 1". The relation between link functions, proper losses and margin losses is considered in more detail by Reid and Williamson (2010).

The following important property of proper losses seems to be originally due to Savage (1971). It shows that a proper loss is completely characterised by a concave function defining its point-wise Bayes risk along with a simple structural relationship between its point-wise risk and Bayes risk.

Theorem 7 *A loss function ℓ is proper if and only if its point-wise Bayes risk $\underline{L}(\eta)$ is concave and for each $\eta, \hat{\eta} \in (0,1)$*

$$L(\eta,\hat{\eta}) = \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}).$$

Furthermore if ℓ is regular this characterisation also holds at the endpoints $\eta, \hat{\eta} \in \{0,1\}$.

For general concave functions $\underline{L}$ which may not be differentiable, $(-\underline{L})'$ is to be taken to be a right derivative as discussed in Section 2.4. The following proof uses an argument in Buja et al. (2005, §17) for the forward direction and the generalised Taylor's theorem due to Liese and Vajda (2006) for the converse.

Proof By definition, the point-wise Bayes risk $\underline{L}(\eta) = \inf_{\hat{\eta}} L(\eta,\hat{\eta})$ which, for each $\eta \in [0,1]$ is just the lower envelope of the lines $L(\eta,\hat{\eta}) = (1-\eta)\ell(0,\hat{\eta}) + \eta\ell(1,\hat{\eta})$ and thus $\underline{L}$ is concave.[14] The properness of ℓ means $\underline{L}(\eta) = L(\eta,\eta)$ and the $\hat{\eta}$-derivative of L is 0 when $\hat{\eta} = \eta$. Hence

$$\frac{\partial}{\partial\hat{\eta}} L(\eta,\hat{\eta})\bigg|_{\hat{\eta}=\eta} = (1-\eta)\ell'(0,\eta) + \eta\ell'(1,\eta) = 0$$

for all $\eta \in [0,1]$. Using this and expanding $\underline{L}'(\eta)$ via the product rule, a little algebra shows $\underline{L}'(\eta) = \ell(1,\eta) - \ell(0,\eta)$. Thus

$$\begin{aligned} \underline{L}(\hat{\eta}) + (\eta-\hat{\eta})\underline{L}'(\hat{\eta}) &= (1-\hat{\eta})\ell(0,\hat{\eta}) + \hat{\eta}\ell(1,\hat{\eta}) + (\eta-\hat{\eta})[\ell(1,\hat{\eta}) - \ell(0,\hat{\eta})] \\ &= (1-\eta)\ell(0,\hat{\eta}) + \eta\ell(1,\hat{\eta}), \end{aligned}$$

which is the definition of $L(\eta,\hat{\eta})$. The result holds at the endpoints if the loss is regular by applying the assumptions in (19).

[13] And powers of absolute divergence $|y-r|^\alpha$ for $\alpha \neq 2$.

[14] Since this argument made no use of the properness of ℓ we see the concavity of the Bayes risk holds for any loss.

Conversely, now suppose Λ is a concave function and let $\ell(y,\hat{\eta}) = \Lambda(\hat{\eta}) + (y-\hat{\eta})\Lambda'(\hat{\eta})$. The Taylor expansion of Λ is

$$\Lambda(\eta) \quad = \quad \Lambda(\hat{\eta}) + (\eta-\hat{\eta})\Lambda'(\hat{\eta}) + \int_{\hat{\eta}}^{\eta} (\eta - c)\,\Lambda''(c)\,dc$$

and so

$$L(\eta,\hat{\eta}) = \Lambda(\hat{\eta}) - \int_{\hat{\eta}}^{\eta} (\eta - c)\,\Lambda''(c)\,dc \geq \Lambda(\eta) = \underline{L}(\eta)$$

because the concavity of Λ means $\Lambda'' \leq 0$ and so the integral term is positive and is minimised to 0 when $\hat{\eta} = \eta$. This shows ℓ is proper, completing the proof. ■

This characterisation of the concavity of $\underline{L}$ means proper losses have a natural connection to Bregman divergences.

4.4 Discriminative Bregman Divergence

Recall from Section 2.5 that if $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set, then a convex function $\phi : \mathcal{S} \to \mathbb{R}$ defines a *Bregman divergence*

$$B_\phi(s,s_0) := \phi(s) - \phi(s_0) - \langle s - s_0, \nabla\phi(s_0)\rangle\,.$$

When $\mathcal{S} = [0,1]$, the concavity of $\underline{L}$ means $\phi(s) = -\underline{L}(s)$ is convex and so induces the Bregman divergence[15]

$$B_\phi(s,s_0) = -\underline{L}(s) + \underline{L}(s_0) - (s_0 - s)\underline{L}'(s_0) = L(s,s_0) - \underline{L}(s)$$

by Theorem 7. The converse also holds. Given a Bregman divergence B_ϕ over $\mathcal{S} = [0,1]$ the convexity of ϕ guarantees that $\underline{L} = -\phi$ is concave. Thus, we know that there is a proper loss ℓ with Bayes risk equal to $-\phi$. As noted by Buja et al. (2005, §19), the difference

$$B_\phi(\eta,\hat{\eta}) = L(\eta,\hat{\eta}) - \underline{L}(\eta)$$

is also known as the *point-wise regret* of the estimate $\hat{\eta}$ w.r.t. η. The corresponding *(full) regret* is the *M*-average point-wise regret

$$\mathbb{E}_{\mathsf{X}\sim M}[B_\phi(\eta(\mathsf{X}),\hat{\eta}(\mathsf{X}))] = \mathbb{L}(\eta,\hat{\eta},M) - \underline{\mathbb{L}}(\eta,M).$$

4.5 Bregman Information

Banerjee et al. (2005a) recently introduced the notion of the *Bregman information* $\mathbb{B}_\phi(\mathsf{S})$ of a random variable S drawn according to some distribution σ over $\mathcal{S}$. It is the minimal σ-average Bregman divergence that can be achieved by an element $s^* \in \mathcal{S}$ (the *Bregman representative*). In symbols,

$$\mathbb{B}_\phi(\mathsf{S}) := \inf_{s\in\mathcal{S}} \mathbb{E}_{\mathsf{S}\sim\sigma}\left[B_\phi(\mathsf{S},s)\right] = \mathbb{E}_{\mathsf{S}\sim\sigma}\left[B_\phi(\mathsf{S},s^*)\right].$$

The authors show that the mean $\overline{s} := \mathbb{E}_{\mathsf{S}\sim\sigma}[\mathsf{S}]$, is the unique Bregman representative. That is, $\mathbb{B}_\phi(\mathsf{S}) = \mathbb{E}_\sigma[B_\phi(\mathsf{S},\overline{s})]$. Surprisingly, this minimiser *only* depends on $\mathcal{S}$ and σ, not the choice of ϕ

[15] Technically, $\mathcal{S}$ is the 2-simplex $\{(s_1,s_2) \in [0,1]^2 : s_1 + s_2 = 1\}$ but we identify $s \in [0,1]$ with $(s, 1-s)$. Also, we once again interpret $(-\underline{L})'$ as a right derivative for general concave $\underline{L}$ as discussed in Section 2.4.

defining the divergence and is a consequence of Jensen's inequality and the form of the Bregman divergence.

Since regret is a Bregman divergence, it is natural to ask what is the corresponding Bregman information. In this case, $\phi = -\underline{L}$ and the random variable $\mathsf{S} = \eta(\mathsf{X}) \in [0,1]$, where $\mathsf{X} \in \mathcal{X}$ is distributed according to the observation distribution M. Noting that $\mathbb{E}_{\mathsf{X}\sim M}[\eta(\mathsf{X})] = \pi$, the proof of the following theorem stems from the definition of Bregman information and some simple algebra showing that $\inf_\eta \mathbb{L}(\eta,\pi,M) = \underline{\mathbb{L}}(\pi,M)$, since by assumption ℓ is a proper loss.

Theorem 8 *Suppose ℓ is a proper loss. Given a discriminative task (η, M) and letting $\phi = -\underline{L}$, the corresponding Bregman information of $\eta(\mathsf{X})$ satisfies*

$$\mathbb{B}_\phi(\eta(\mathsf{X})) = \mathbb{B}_\phi(\eta,M) := \underline{\mathbb{L}}(\pi,M) - \underline{\mathbb{L}}(\eta,M).$$

4.6 Statistical Information

The reduction in risk (from prior $\pi \in [0,1]$ to posterior $\eta \in [0,1]^{\mathcal{X}}$)

$$\Delta\underline{\mathbb{L}}(\eta,M) = \Delta\underline{\mathbb{L}}(\pi,P,Q) := \underline{\mathbb{L}}(\pi,M) - \underline{\mathbb{L}}(\eta,M) \tag{20}$$

is known as *statistical information* and was introduced by DeGroot (1962) motivated by Lindley (1956). This reduction can be interpreted as how much risk is removed by knowing observation-specific class probabilities η rather than just the prior π.

DeGroot originally introduced statistical information in terms of what he called an *uncertainty function* which, in the case of binary experiments, is any function $U : [0,1] \to [0,\infty)$. The statistical information is then the average reduction in uncertainty which can be expressed as a concave Jensen gap

$$-\mathbb{J}_M[U(\eta)] = \mathbb{J}_M[-U(\eta)] = U(\mathbb{E}_{\mathsf{X}\sim M}[\eta(\mathsf{X})]) - \mathbb{E}_{\mathsf{X}\sim M}[U(\eta(\mathsf{X}))].$$

DeGroot noted that Jensen's inequality implies that for this quantity to be non-negative the uncertainty function must be concave, that is, $-U$ must be convex.

Theorem 8 shows that statistical information is a Bregman information and corresponds to the Bregman divergence obtained by setting $\phi = -\underline{L}$. This connection readily shows that $\Delta\underline{\mathbb{L}}(\eta,M) \geq 0$ (DeGroot, 1962, Thm 2.1) since the minimiser of the Bregman information is $\pi = \mathbb{E}_{\mathsf{X}\sim M}[\eta(\mathsf{X})]$ regardless of loss and $B_\phi(\eta,\pi) \geq 0$ since it is a regret.

4.7 Unifying Information and Divergence

From a generative perspective, f-divergences can be used to assess the difficulty of a learning task by measuring the divergence between the class-conditional distributions P and Q. The more divergent the distributions for the two classes, the easier the classification task. Österreicher and Vajda (1993, Thm 2) made this relationship precise by showing that f-divergence and statistical information have a one-to-one correspondence:

Theorem 9 *If $(\pi,P,Q;\ell)$ is an arbitrary task and $\underline{L}$ is the associated conditional Bayes risk then defining*

$$f^\pi(t) := \underline{L}(\pi) - (\pi t + 1 - \pi)\underline{L}\left(\frac{\pi t}{\pi t + 1 - \pi}\right) \tag{21}$$

for $\pi \in [0,1]$ *implies* f^{π} *is convex,* $f^{\pi}(1) = 0$ *and*

$$I_{f^{\pi}}(P,Q) = \Delta\underline{\mathbb{L}}(\pi,P,Q)$$

for all distributions P and Q. Conversely, if f *is convex and* $f(1) = 0$ *then defining*

$$\underline{L}^{\pi}(\eta) := -\frac{1-\eta}{1-\pi} f\left(\frac{1-\pi}{\pi}\frac{\eta}{1-\eta}\right), \quad \pi \in [0,1]$$

implies

$$I_f(P,Q) = \Delta\underline{\mathbb{L}}^{\pi}(\pi,P,Q)$$

for all distributions P and Q, where $\Delta\underline{\mathbb{L}}^{\pi}$ *is the statistical information associated with* $\underline{L}^{\pi}$.

The proof, given in Appendix A.3, is a straight-forward calculation that exploits the relationships between the generative and discriminative views presented earlier. Combined with the link between Bregman and statistical information, this result means that they and f-divergences are *interchangeable* as measures of task difficulty. The theorem leads to some correspondences between well known losses and divergence: log-loss with $\mathrm{KL}(P,Q)$; square loss with triangular discrimination; and 0-1 loss with $V(P,Q)$. (See Section 5.5 for an explicitly worked out example.)

This connection generalises the link between f-divergences and F-errors (expectations of concave functions of η) in Devroye et al. (1996) and can be compared to the more recent work of Nguyen et al. (2005) who show that each f-divergence corresponds to the negative Bayes risk for a *family* of surrogate margin losses. The one-to-many nature of their result may seem at odds with the one-to-one relationship here. However, the family of margin losses given in their work can be recovered by combining the proper losses with link functions. Working with proper losses also addresses a limitation pointed out by Nguyen et al. (2005, pg. 14), namely that "asymmetric f-divergences cannot be generated by *any* (margin-based) surrogate loss function" and extends their analysis "to show that asymmetric f-divergences can be realized by general (asymmetric) loss functions".

4.8 Summary

The main results of this section can be summarised as follows.

Theorem 10 *Let* $f : \mathbb{R}^+ \to \mathbb{R}$ *be a convex function and for each* $\pi \in [0,1]$ *define for* $c \in [0,1)$*:*

$$\begin{aligned} \phi(c) &:= \frac{1-c}{1-\pi} f\left(\lambda_{\pi}(c)\right), \\ \underline{L}(c) &:= -\phi(c), \end{aligned}$$

where λ_{π} *is defined by (17). Then for every binary experiment* (P,Q) *we have*

$$\mathbb{I}_f(P,Q) = \Delta\underline{\mathbb{L}}(\eta,M) = \mathbb{B}_{\phi}(\eta,M),$$

where $M := \pi P + (1-\pi)Q$*,* $\eta := \pi dP/dM$ *and* $\underline{\mathbb{L}}$ *is the expectation (in* X*) of the conditional Bayes risk* $\underline{L}$*. Equivalently,*

$$\mathbb{J}_Q[f(dP/dQ)] = \mathbb{J}_M[-\underline{L}(\eta)] = \mathbb{J}_M[\phi(\eta)].$$

What this says is that for each choice of π the classes of f-divergences $\mathbb{I}_f$, statistical informations $\Delta\underline{\mathbb{L}}$ and (discriminative) Bregman informations $\mathbb{B}_\phi$ can all be defined in terms of the Jensen gap of some convex function. Additionally, there is a bijection between each of these classes due to the mapping λ_π that identifies likelihood ratios with posterior probabilities.

The class of f-divergences is "more primitive" than the other measures since its definition does not require the extra structure that is obtained by assuming that the reference measure M can be written as the convex combination of the distributions P and Q. Indeed, each $\mathbb{I}_f$ is invariant to the choice of reference measure and so is invariant to the choice of π. The results in the next section provide another way of looking at this invariance of $\mathbb{I}_f$. In particular, we see that every f-divergence is a weighted "average" of statistical informations or, equivalently, $\mathbb{I}_{f^\pi}$ divergences.

5. Primitives and Weighted Integral Representations

When given a class of functions like f-divergences, risks or measures of information it is natural to ask what the "simplest" elements of these classes are. We would like to know which functions are "primitive" in the sense that they can be used to express other measures but themselves cannot be so expressed.

The connections between risk, f-divergence, and statistical information discussed in Section 4 are all in terms of the convex functions that define each type of measurement. As discussed in Section 2.3, integral representations allow these convex functions to be expressed as weighted combinations of simple, convex, piecewise linear functions. By thinking of the set of these simple functions as a "basis" for convex functions, we are able to identify any convex function with its "coordinates"—that is, its weight function—relative to this basis.

The main result of this section essentially "lifts" this weight function representation of convex functions through the definitions of proper risks and f-divergence (and therefore also statistical and Bregman information) so they can be expressed as weighted integrals of primitive elements corresponding to the simple convex functions acting as the "basis". In the case of f-divergences and information the weight function in these integrals completely determines their behaviour. This means the weight functions can be used as a proxy for the analysis of these measures, or as a knob the user can adjust in choosing what to measure.

We also show that the close relationships between information and f-divergence in terms of their convex generators can be directly translated into a relationship between the respective weight functions associated with these measures. That is, given the weight function that determines an f-divergence there is, for each choice of the prior π, a simple transformation that yields the weight function for the corresponding statistical information, and *vice versa*.

This shift from "function as graph of evaluations" to "function as weighted combination of primitive functions" permeates the remainder of the paper and is (loosely!) analogous to the way the Fourier transform represents functions as sums of simple, periodic signals. In Section 6, risk curves are used to graphically summarise the values of all the primitive risks for a given binary experiment. In Section 7, surrogate regret bounds for proper losses and a tight generalisation of Pinsker's inequality are derived by considering the relationship between general regrets or divergences and the primitive ones comprising them. In both cases, the bounds are established by using weight functions to understand the relative contribution of each primitive to the weighted sum. In particular, the Pinkser-like inequalities in Appendix B for specific f-divergences are obtained via direct manipulation of their weight functions.

5.1 Integral Representations of f-divergences

The following result shows that the class of f-divergences (and, by the result of the previous section, statistical and Bregman information) is closed under conic combination.

Theorem 11 *For all convex functions $f_1, f_2\colon (0,\infty)\to\mathbb{R}$ and all $\alpha_1,\alpha_2\in[0,\infty)$, the function*

$$(0,\infty)\ni t\mapsto g(t) := \alpha_1 f_1(t)+\alpha_2 f_2(t) \tag{22}$$

is convex. Furthermore, for all distributions P and Q, we have

$$\mathbb{I}_g(P,Q)=\alpha_1\mathbb{I}_{f_1}(P,Q)+\alpha_2\mathbb{I}_{f_2}(P,Q). \tag{23}$$

Conversely, given f_1, f_2, α_1 and α_2, if (23) holds for all P and Q then g must be, up to affine additions, of the form (22).

The proof is a straight-forward application of the definition of convexity and of f-divergences.

One immediate consequence of this result is that the set of f-divergences is closed under conic combinations $\sum_i \alpha_i \mathbb{I}_{f_i}$. Furthermore, the arguments in Section 2.4 can be used to extend this observation beyond finite linear combination to generalised weight functions α. By Corollary 2, if f is a convex function then expanding it about 1 in (5) and setting $\alpha(s)=f''(s)$ means that

$$\mathbb{I}_f(P,Q)=\int_0^\infty \mathbb{I}_{F_s}(P,Q)\,\alpha(s)\,ds \tag{24}$$

where $F_s(t)=[\![s\le 1]\!](s-t)_+ + [\![s>1]\!](t-s)_+$.[16] The functions F_s, $s\in\mathbb{R}^+$ can therefore be seen as the generators of the class of primitive f-divergences. As a function of t, each F_s is piecewise linear, with a single “hinge” at s. Of course, any affine translation of any F_s is also a primitive. In fact, each F_s may undergo a different affine translation without changing the f-divergence $\mathbb{I}_f$. The weight function α is what completely characterises the behaviour of $\mathbb{I}_f$.

The integral in (24) need not always exist since the integrand may not be integrable. When the Cauchy Principal Value diverges we say the integral takes on the value ∞. We note that many (not all) f-divergences can sometimes take on infinite values.

The integral form in (24) can be readily transformed into an integral representation that does not involve an infinite integrand. This is achieved by mapping the interval $[0,\infty)$ onto $[0,1)$ via the change of variables $\pi=\frac{1}{1+s}\in[0,1]$. In this case, $s=\frac{1-\pi}{\pi}$ and so $ds=-\frac{d\pi}{\pi^2}$ and the integral of (24) becomes

$$\begin{aligned}\mathbb{I}_f(P,Q) &= -\int_1^0 \mathbb{I}_{F_{\frac{1-\pi}{\pi}}}(P,Q)\,\alpha(\tfrac{1-\pi}{\pi})\,\pi^{-2}\,d\pi \\ &= \int_0^1 \mathbb{I}_{\tilde{f}_\pi}(P,Q)\,\gamma(\pi)\,d\pi \end{aligned} \tag{25}$$

where

$$\tilde{f}_\pi(t) := \pi F_{\frac{1-\pi}{\pi}}(t) = \begin{cases}(1-\pi(1+t))_+\ , & \pi\ge\frac12\\ (\pi(1+t)-1)_+\ , & \pi<\frac12\end{cases} \tag{26}$$

[16] Technically, one must assume that f is twice differentiable for this result to hold. However, the convexity of f implies it has well-defined one-sided derivatives f'_+ and $\alpha(s)$ can be expressed as the measure corresponding to $d f'_+/d\lambda$ for the Lebesgue measure λ. Details can be found in Liese and Vajda (2006). The representation of a general f-divergence in terms of elementary ones is not new; see for example Österreicher and Feldman (1981) and Feldman and Österreicher (1989).

and

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

This observation forms the basis of the following restatement of a theorem by Liese and Vajda (2006). We include it here with a short proof to discuss the connection between f-divergences and statistical information.[17]

Theorem 12 *Let f be convex such that $f(1) = 0$. Then there exists a (generalised) function $\gamma : (0,1) \to \mathbb{R}$ such that, for all P and Q:*

$$\mathbb{I}_f(P,Q) = \int_0^1 \mathbb{I}_{f_\pi}(P,Q)\,\gamma(\pi)\,d\pi, \text{ where } f_\pi(t) = (1-\pi)\wedge\pi - (1-\pi)\wedge(\pi t).$$

Proof The earlier discussion giving the derivation of Equation (25) implies the result. The only discrepancy is over the form of f_π. We determine the precise form by noting that the family of $\tilde{f}_\pi$ given in (26) can be transformed by affine addition without affecting the representation of $\mathbb{I}_f$. Specifically,

$$\begin{aligned} f_\pi(t) &:= (1-\pi)\wedge\pi - (1-\pi)\wedge(\pi t) \\ &= \begin{cases} (1-\pi(1+t))_+\,, & \pi \geq \frac{1}{2} \\ (\pi(1+t)-1)_+ + \pi(1-t)\,, & \pi < \frac{1}{2} \end{cases} \\ &= \tilde{f}_\pi(t) + [\![\pi < \tfrac{1}{2}]\!]\pi(1-t), \end{aligned}$$

and so $\tilde{f}_\pi$ and f_π are in the same affine equivalence class for each $\pi \in [0,1]$. Thus, by Theorem 6 we have $\mathbb{I}_{f_\pi} = \mathbb{I}_{\tilde{f}_\pi}$ for each $\pi \in [0,1]$, proving the result. ■

The specific choice of f_π in the above theorem from all of the affine equivalents was made to make simpler the connection between integral representations for losses and f-divergences, discussed in Section 5.4.

One can easily verify that f_π are convex hinge functions of t with a hinge at $\frac{1-\pi}{\pi}$ and $f_\pi(1) = 0$. Thus $\{\mathbb{I}_{f_\pi}\}_{\pi\in(0,1)}$ is a family of primitive f-divergences; confer Österreicher and Feldman (1981) and Feldman and Österreicher (1989). This theorem implies an existing representation of f-divergences due to Österreicher and Vajda (1993, Theorem 1) and Gutenbrunner (1990). They show that an f-divergence can be represented as a weighted integral of statistical informations for 0-1 loss: for all P, Q

$$\mathbb{I}_f(P,Q) = \int_0^1 \Delta\underline{\mathbb{L}}^{0-1}(\pi,P,Q)\gamma(\pi)d\pi, \tag{27}$$

$$\gamma(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right). \tag{28}$$

An f divergence is *symmetric* if $\mathbb{I}_f(P,Q) = \mathbb{I}_f(Q,P)$ for all P, Q. The representation of $\mathbb{I}_f$ in terms of γ and Theorem 15 provides an easy test for symmetry:

[17] The $1/\pi^3$ term in the definition of γ seems a little unusual at first glance. However, it is easily understood as the product of two terms: $1/\pi^2$ from the second derivative of $(1-\pi)/\pi$, and $1/\pi$ from a transformation of variables within the integral to map the limits of integration from $(0,\infty)$ to $(0,1)$ via λ_π.

Corollary 13 *Suppose $\mathbb{I}_f$ is an f-divergence with corresponding weight function γ given by (28). Then $\mathbb{I}_f$ is symmetric iff $\gamma(\pi) = \gamma(1-\pi)$ for all $\pi \in [0,1]$.*

The proof is in Appendix A.4.

Corollary 13 provides a way of generating all convex f such that $\mathbb{I}_f$ is symmetric that is simpler than that proposed by Hiriart-Urruty and Martínez-Legaz (2007): let $\gamma(\pi) = \beta(\pi \wedge (1-\pi))$ where $\beta \in (\mathbb{R}^+)^{[0,\frac{1}{2}]}$ (i.e., all symmetric weight functions) and generate f from γ by inverting (28); explicitly,

$$f(s) = \int_0^s \left(\int_0^t \frac{1}{(\tau+1)^3} \gamma\left(\frac{1}{\tau+1}\right) d\tau \right) dt, \;\; s \in \mathbb{R}^+.$$

5.2 Proper Losses and Cost-Weighted Risk

We now consider a representation of proper losses in terms of primitive losses that originates with Shuford et al. (1966). Our discussion follows that of Buja et al. (2005) and then examines its implications in light of the connections between information and divergence just presented.

The *cost-weighted losses* are a family of losses parameterised by a false positive cost $c \in [0,1]$ that defines a loss for $y \in \{\pm 1\}$ and $\hat{\eta} \in [0,1]$ by

$$\ell_c(y, \hat{\eta}) = c [\![y = -1]\!] [\![\hat{\eta} \geq c]\!] + (1-c) [\![y = 1]\!] [\![\hat{\eta} < c]\!]. \tag{29}$$

Intuitively, a cost-weighted loss thresholds $\hat{\eta}$ at c and assigns a cost if the resulting classification disagrees with y. These correspond to the “signatures” for eliciting the probability η as described by Lambert et al. (2008). Substituting $c = \frac{1}{2}$ will verify that $2\ell_{\frac{1}{2}}$ is equivalent to 0-1 misclassification loss ℓ^{0-1}. Taking expectations with respect to Y we have

$$L_c(\eta, \hat{\eta}) = (1-\eta) c [\![\hat{\eta} \geq c]\!] + \eta (1-c) [\![\hat{\eta} < c]\!]. \tag{30}$$

We will use L_c, $\mathbb{L}_c$ and $\Delta \underline{\mathbb{L}}_c$ to denote the cost-weighted point-wise risk, full risk and statistical information associated with each cost-weighted loss. The following theorems collect some useful observations about these primitive quantities. The first shows that the point-wise Bayes risk is a simple, concave “tent” function. The second shows that cost-weighted statistical information is invariant under the switching of the classes provided the costs are also switched and that π and $1-c$ are interchangeable.

Theorem 14 *For all $\eta, c \in [0,1]$ the point-wise Bayes risk $\underline{L}_c(\eta) = (1-\eta)c \wedge (1-c)\eta$ and is therefore concave in both c and η.*

Proof From the definition of ℓ_c in Equation 29 and the definition of point-wise Bayes risk, we have, for $\eta \in [0,1]$,

$$\begin{aligned} \underline{L}_c(\eta) &= \inf_{\hat{\eta} \in [0,1]} L_c(\eta, \hat{\eta}) \\ &= \inf_{\hat{\eta} \in [0,1]} \{ (1-\eta) c [\![\hat{\eta} \geq c]\!] + \eta (1-c) [\![\hat{\eta} < c]\!] \} \\ &= \inf_{\hat{\eta} \in [0,1]} \{ \eta (1-c) + (c-\eta) [\![\hat{\eta} \geq c]\!] \}, \end{aligned}$$

where the last step makes use of the identity $[\![\hat{\eta} < c]\!] = 1 - [\![\hat{\eta} \geq c]\!]$. Since $(c-\eta)$ is negative if and only if $\eta > c$, the infimum is obtained by having $[\![\hat{\eta} \geq c]\!] = 1$ if and only if $\eta \geq c$, that is, by letting

$\hat{\eta} = \eta$. In this case, when $\hat{\eta} \geq c$ we have $\underline{L}_c(\eta) = c(1-\eta)$ and when $\hat{\eta} < c$ we have $\underline{L}_c(\eta) = (1-c)\eta$. The concavity of $\underline{L}_c$ is evident as this function is the minimum of two linear functions of c and η. ■

Theorem 15 *For all $c \in [0,1]$ and tasks $(\eta, M; \ell_c) = (\pi, P, Q; \ell_c)$ the statistical information satisfies 1)*

$$\Delta\underline{\mathbb{L}}_c(1-\eta, M) = \Delta\underline{\mathbb{L}}_{1-c}(\eta, M),$$

or equivalently,

$$\Delta\underline{\mathbb{L}}_c(1-\pi, Q, P) = \Delta\underline{\mathbb{L}}_{1-c}(\pi, P, Q);$$

and 2)

$$\Delta\underline{\mathbb{L}}_\pi(1-c, P, Q) = \Delta\underline{\mathbb{L}}_c(1-\pi, P, Q).$$

Proof By Theorem 14 we know $\underline{L}_c(\eta) = \min\{(1-\eta)c, (1-c)\eta\}$ and so $\underline{L}_c(1-\eta) = \underline{L}_{1-c}(\eta)$ for all $\eta, c \in [0,1]$. Therefore, $\underline{\mathbb{L}}_c(1-\eta, M) = \underline{\mathbb{L}}_{1-c}(\eta, M)$ for any $\eta : \mathcal{X} \to [0,1]$ including the constant function $\mathbb{E}_M[\eta]$. By definition, $\Delta\underline{\mathbb{L}}_c(\eta, M) = \underline{\mathbb{L}}(\mathbb{E}_M[\eta], M) - \underline{\mathbb{L}}(\eta, M)$ and so $\Delta\underline{\mathbb{L}}_{1-c}(\eta, M) = \Delta\underline{\mathbb{L}}_c(1-\eta, M)$ proving part 1.

Part 2 also follows from Theorem 14 by noting that $\underline{L}_c(1-\pi) = \underline{L}_\pi(1-c)$ and $\mathbb{E}_M[\underline{L}_c(\eta)] = \int_{\mathcal{X}} \min\{(1-c)\pi\, dP, (1-\pi)c\, dQ\}$. ■

5.3 Integral Representations of Proper Losses

The cost-weighted losses are primitive in the sense that they form the basis for a Choquet integral representation of proper losses. This representation is essentially a consequence of Taylor's theorem and was originally studied by Shuford et al. (1966) and later generalised by Schervish (1989). The recent presentation of this result by Lambert et al. (2008) gives yet a more general formulation in terms of the elicitability of properties of distributions, along with a geometric derivation. An historical summary of decompositions of scoring rules is given by Winkler et al. (1990, Section 4).

Theorem 16 *Let $\ell : \mathcal{Y} \times [0,1] \to \mathbb{R}$ be a fair, proper loss. Then for each $\hat{\eta} \in (0,1)$ and $y \in \mathcal{Y}$*

$$\ell(y, \hat{\eta}) = \int_0^1 \ell_c(y, \hat{\eta})\, w(c)\, dc \tag{31}$$

where the weight function[18] $w : (0,1) \to \mathbb{R}^+$ *satisfies*

$$w(c) = -\underline{L}''(c) \geq 0 \tag{32}$$

for all $c \in (0,1)$. Conversely, if ℓ is defined by (31) for some weight function $w : (0,1) \to \mathbb{R}^+$ then it is proper.

The proof is almost a direct consequence of Taylor's theorem.

[18] The weight function and second derivative of $-\underline{L}$ are to be interpreted distributionally as discussed in Section 2.4.

Proof We first assume ℓ is a proper loss so that $L(\eta,\hat{\eta}) = \mathbb{E}_{\mathsf{Y}\sim\eta}[\ell(\mathsf{Y},\hat{\eta})]$ and $\underline{L}(\eta) = L(\eta,\eta)$. Expanding $\underline{L}(\eta)$ about $\hat{\eta} \in (0,1)$ using Corollary 2 yields

$$\begin{aligned}\underline{L}(\eta) &= \underline{L}(\hat{\eta}) + (\eta-\hat{\eta})\underline{L}'(\hat{\eta}) + \int_0^1 \phi_c(\eta,\hat{\eta})\,\underline{L}''(c)\,dc \\ &= L(\eta,\hat{\eta}) + \int_0^1 \phi_c(\eta,\hat{\eta})\,\underline{L}''(c)\,dc \qquad (33)\end{aligned}$$

by Theorem 7. The generalised function $w(c) = -\underline{L}''(c) \geq 0$ by the concavity of $\underline{L}$. Rearranging (33) gives

$$L(\eta,\hat{\eta}) = \underline{L}(\eta) + \int_0^1 \phi_c(\eta,\hat{\eta})\,w(c)\,dc.$$

The definition of L in (18) implies $L(y,\hat{\eta}) = \ell(y,\hat{\eta})$ for $y \in \{0,1\}$ and so

$$\ell(y,\hat{\eta}) = \underline{L}(y) + \int_0^1 \phi_c(y,\hat{\eta})\,w(c)\,dc, \qquad (34)$$

where

$$\phi_c(y,\hat{\eta}) = [\![\hat{\eta} \leq c < y]\!](y-c) + [\![y \leq c < \hat{\eta}]\!](c-y),$$

which is equal to the definition of ℓ_c in (29) since the left (resp. right) term is only non-zero when $y = 1$ (resp. $y = 0$). Observe that $\underline{L}(0) = \underline{L}(1) = 0$ since $\underline{L}(0) = L(0,0) = \ell(0,0) = 0$ by the assumption that the loss is fair, and similarly for $\underline{L}(1)$.

This shows that (34) is equivalent to (31), completing the forward direction of the theorem.

If we now assume the function $w \geq 0$ is given and ℓ defined as in (31) then it suffices to show $\underline{L}(\eta) = L(\eta,\eta)$. First note that

$$\begin{aligned}L(\eta,\hat{\eta}) &= \mathbb{E}_{\mathsf{Y}\sim\eta}\left[\int_0^1 \ell_c(\mathsf{Y},\hat{\eta})\,w(c)\,dc\right] \\ &= \int_0^1 L_c(\eta,\hat{\eta})\,w(c)\,dc.\end{aligned}$$

Each of the L_c are proper and so are minimised when $\hat{\eta} = \eta$. Since $w(c) \geq 0$ this must also be sufficient to minimise L. ■

We will write ℓ_w, L_w and $\mathbb{L}_w$ to explicitly indicate the parameterisation of the loss, conditional loss and expected loss by the weight function w. A proper loss ℓ_w corresponding to a given weight function can be explicitly derived using the following theorem.

Theorem 17 *Given a weight function $w : [0,1] \to \mathbb{R}^+$, let $W(t) = \int^t w(c)\,dc$ and $\overline{W}(t) = \int^t W(c)\,dc$. Then the loss ℓ_w defined by*

$$\ell_w(y,\hat{\eta}) = -\overline{W}(\hat{\eta}) - (y-\hat{\eta})W(\hat{\eta})$$

is a proper loss. Additionally, if $\overline{W}(0)$ and $\overline{W}(1)$ are both finite then

$$(y,\hat{\eta}) \mapsto \ell_w(y,\hat{\eta}) + (\overline{W}(1) - \overline{W}(0))y + \overline{W}(0) \qquad (35)$$

is a fair, proper loss.

Proof First we define the loss

$$\ell(y,\hat{\eta}) := \int_0^1 \ell_c(y,\hat{\eta})\, w(c)\, dc$$

and proceed to show it is equal to the definition of ℓ_w. Theorem 16 guarantees that ℓ is proper and that $w = -\underline{L}''$. By definition of the improper integrals $\overline{W}$ and W and the fundamental theorem of calculus we know that $W' = w = -\underline{L}''$ and so $\overline{W}'(t) = W(t) = -\underline{L}'(t) + a$ and

$$\overline{W}(t) = -\underline{L}(t) + at + b, \tag{36}$$

where $a, b \in \mathbb{R}$ are constants of integration. Substituting these into the Savage representation of Theorem 7 for proper losses we see that

$$\begin{aligned} L(\eta,\hat{\eta}) &= \underline{L}(\hat{\eta}) + (\eta - \hat{\eta})\underline{L}'(\hat{\eta}) \\ &= -\overline{W}(\hat{\eta}) + a\hat{\eta} + b + (\eta - \hat{\eta})[-W(\hat{\eta}) + a] \\ &= -\overline{W}(\hat{\eta}) - (\eta - \hat{\eta})W(\hat{\eta}) + a\eta + b. \end{aligned}$$

Since $L(y,\hat{\eta}) = \ell(y,\hat{\eta})$ for $y \in \{0,1\}$ we have $\ell(0,\hat{\eta}) = \ell_w(0,\hat{\eta}) + b$ and $\ell(1,\hat{\eta}) = \ell_w(1,\hat{\eta}) + a + b$ for all $a, b \in \mathbb{R}$. Choosing $a = b = 0$ achieves the result.

If $\overline{W}(0)$ and $\overline{W}(1)$ are both finite then letting $a = \overline{W}(1) - \overline{W}(0)$ and $b = \overline{W}(0)$ means (36) implies $\overline{W}(0) = -\underline{L}(0) + \overline{W}(0)$ and so $\underline{L}(0) = 0$. Similarly, $\underline{L}(1) = 0$ showing that (35) is fair. ■

As an example of how this theorem lets us explicitly construct proper losses from weight functions, consider the weight function $w(c) = 1$. In this case, $W(t) = t$ and $\overline{W}(t) = \frac{t^2}{2}$. Thus, noting that $y^2 = y$ for $y \in \{0,1\}$ we have

$$\ell_w(y,\hat{\eta}) = -\tfrac{1}{2}\hat{\eta}^2 - (y - \hat{\eta})\hat{\eta} + \tfrac{1}{2}y = \tfrac{1}{2}(\hat{\eta} - y)^2$$

which is the square loss.

As a second example, consider $w(c) = \frac{1}{(1-c)c}$. In this case, $W(t) = \ln\left(\frac{t}{1-t}\right)$ and $\overline{W}(t) = (1-t)\ln(1-t) + t\ln(t)$. Since $\lim_{\varepsilon \to 0} \varepsilon \ln(\varepsilon) = 0$ we define $0\ln(0) := 0$ so that $b = \overline{W}(0) = 0$ and $a = \overline{W}(1) - \overline{W}(0) = 0$. This implies

$$\begin{aligned} \ell_w(y,\hat{\eta}) &= -(1-\hat{\eta})\ln(1-\hat{\eta}) - \hat{\eta}\ln(\hat{\eta}) - (y - \hat{\eta})\ln\left(\frac{\hat{\eta}}{1-\hat{\eta}}\right) \\ &= [-(1-\hat{\eta}) + (y - \hat{\eta})]\ln(1-\hat{\eta}) + [-\hat{\eta} - (y - \hat{\eta})]\ln(\hat{\eta}) \\ &= -(1-y)\ln(1-\hat{\eta}) - y\ln(\hat{\eta}) \end{aligned}$$

which is log loss.

5.4 Relating Integral Representations for $\mathbb{L}$ and $\mathbb{I}_f$

There is also the following direct relationship between the weight functions γ for an f-divergence and w for the corresponding statistical information. Since the weight functions are an attractive parameterization, it is convenient to be able to directly translate between the two respective weight functions. The proof is in Appendix A.5.

Theorem 18 *Let $f\colon \mathbb{R}^+ \to \mathbb{R}$ be convex (with $f(1)=0$) define $\mathbb{I}_f$ with corresponding weight function γ. Then for each $\pi \in (0,1)$ the weight function w^π in Theorem 16 for the loss ℓ^π given by Theorem 9 satisfies*

$$w^\pi(c) = \frac{\pi(1-\pi)}{\nu(\pi,c)^3}\gamma\left(\frac{(1-c)\pi}{\nu(\pi,c)}\right)$$

or, inversely,

$$\gamma(c) = \frac{\pi^2(1-\pi)^2}{\nu(\pi,c)^3} w\left(\frac{\pi(1-c)}{\nu(\pi,c)}\right),$$

where $\nu(\pi,c) = (1-c)\pi + (1-\pi)c$.

The representation (27,28) allows the determination of weights for standard f-divergences. Kullback-Liebler divergence $\mathrm{KL}(P,Q)$ corresponds to $\gamma(\pi) = \frac{1}{\pi^2(1-\pi)}$. Thus $J(P,Q) = \mathrm{KL}(P,Q) + \mathrm{KL}(Q,P)$ corresponds to $\gamma(\pi) = \frac{1}{\pi^2(1-\pi)^2}$. Several f-divergences are presented with their corresponding weight function in Table 2. The weight for $\mathrm{KL}(P,Q)$ has a double pole at $\pi = 0$ which is why KL-divergence is hard to estimate—it puts a lot of weight on $\Delta\underline{\mathbb{L}}^{0-1}(\pi,PQ)$ for $\pi \approx 0$ which by Theorem 15 means a lot of weight on $\Delta\underline{\mathbb{L}}_c(\frac{1}{2})$ for $c \approx 1$ which requires a good estimate of $\mathbb{L}_c(\eta,M)$ which is difficult with modest data sample sizes.[19]

A loss function corresponding to each f-divergence in Table 2 is also shown. The weight function $w(c)$ for the loss is for the case when $\pi = \frac{1}{2}$, that is, it is a loss for a binary classification problem with equal proportions of positive and negative examples. In this case, the relationship between w and γ simplifies to $w^{\frac{1}{2}}(c) = 2\gamma(1-c)$ since $\nu(\frac{1}{2},c) = \frac{1}{2}c + \frac{1}{2}(1-c) = \frac{1}{2}$.

The entries in Table 2 without a name for the loss correspond to losses that are not definite. It turns out that weight functions whose tail behaviour is not $o(c^{-2})$ or $o((1-c)^{-2})$ as c goes to 0 or 1, respectively (confer Buja et al., 2005, §6) imply non-definiteness of a proper loss.

5.5 Example—Squared Loss

We illustrate some of the above concepts with a simple example. Consider squared loss. We have

$$L(\eta,\hat{\eta}) = \hat{\eta}^2(1-\eta) + (\hat{\eta}-1)^2\eta$$

and thus $\underline{L}(\eta) = L(\eta,\eta) = \eta(1-\eta)$ and $\underline{L}''(\eta) = -2$ and thus by (32) $w(\eta) = 2$. From (21) we thus have

$$f^\pi(t) = \frac{\pi(1-\pi)(\pi t + 1 - \pi) - (1-\pi)\pi t}{\pi t + 1 - \pi}.$$

Choosing $\pi = \frac{1}{2}$ this becomes $f^{\frac{1}{2}}(t) = \frac{1-t}{4t+4}$. One can check that $8 \cdot f^{\frac{1}{2}}(t) + t - 1 = \frac{(t-1)^2}{t+1}$ which agrees with the f corresponding to Triangular Discrimination in Table 2. Scaling is just a question of normalisation and we have already seen that $\mathbb{I}_f$ is insensitive to affine offsets in f. This illustrates the awkwardness of parameterising $\mathbb{I}_f$ in terms of f: at first sight $\frac{1-t}{4t+4}$ and $\frac{(t-1)^2}{t+1}$ seem quite

[19] Considering KL-divergence from the weight function perspective suggests a scheme to estimate it: avoid attempting to estimate the regions near zero and one where the weight function diverges. A particular example of this is the divergence $\mathrm{KL}_\varepsilon(P,Q)$ which has weight function $\gamma(\pi) = \frac{1}{\pi^2(1-\pi)}[\![\pi \in [\varepsilon, 1-\varepsilon]]\!]$. The corresponding f can be worked out but has the rather less intuitively clear form $f(t) = [\![t < \frac{\varepsilon}{1-\varepsilon}]\!](t(\ln(\frac{\varepsilon}{1-\varepsilon})+1) - \frac{\varepsilon}{1-\varepsilon}) + [\![\frac{\varepsilon}{1-\varepsilon} \le t \le \frac{1-\varepsilon}{\varepsilon}]\!] t\ln t + [\![\frac{1-\varepsilon}{\varepsilon} < t]\!](t(\ln(\frac{1-\varepsilon}{\varepsilon})+1) - \frac{1-\varepsilon}{\varepsilon})$, $\varepsilon \in [0,1)$. This approach to regularizing the estimation of the KL-divergence was suggested by Gutenbrunner (1990, page 454).

Symbol	Divergence	$f(t)$	$\gamma(\pi)$	$w^{\frac{1}{2}}(c)$	$\ell(0,\hat{\eta})$	$\ell(1,\hat{\eta})$	Loss
$V(P,Q)$	Variational	$\lvert t-1\rvert$	$16\delta\left(\pi-\frac{1}{2}\right)$	$32\delta\left(\frac{1}{2}-c\right)$	$16[\![\hat{\eta}>\frac{1}{2}]\!]$	$16[\![\hat{\eta}\leq\frac{1}{2}]\!]$	0-1
$\mathrm{KL}(P,Q)$	Kullback-Leibler	$t\ln t$	$\frac{1}{\pi^2(1-\pi)}$	$\frac{2}{(1-c)^2c}$	$2\left[2\ln(1-\hat{\eta})+\frac{\hat{\eta}}{1-\hat{\eta}}\right]$	$2\left[\ln\frac{1-\hat{\eta}}{\hat{\eta}}-1\right]$	—
$\Delta(P,Q)$	Triangular Discrimination	$(t-1)^2/(t+1)$	8	16	$8\hat{\eta}^2$	$8(1-\hat{\eta})^2$	Square
$\mathrm{I}(P,Q)$	Jensen-Shannon	$\frac{t}{2}\ln\left(\frac{t}{t+1}\right)-\frac{1}{2}\ln\left(\frac{t+1}{4}\right)$	$\frac{1}{2\pi(1-\pi)}$	$\frac{1}{(1-c)c}$	$-\ln(1-\hat{\eta})$	$-\ln(\hat{\eta})$	Log
$\mathrm{T}(P,Q)$	Arith.-Geo. Mean	$\left(\frac{t+1}{2}\right)\ln\left(\frac{t+1}{2\sqrt{t}}\right)$	$\frac{2\pi^2-2\pi+1}{4\pi^2(1-\pi)^2}$	$\frac{2c^2-2c+1}{2c^2(1-c)^2}$	$\frac{1}{2}\left[\ln((1-\hat{\eta})\hat{\eta})-\frac{1-2\hat{\eta}}{1-\hat{\eta}}\right]$	$\frac{1}{2}\left[\ln((1-\hat{\eta})\hat{\eta})+\frac{1-2\hat{\eta}}{\hat{\eta}}\right]$	—
$\mathrm{J}(P,Q)$	Jeffreys	$(t-1)\ln(t)$	$\frac{1}{\pi^2(1-\pi)^2}$	$\frac{2}{(1-c)^2c^2}$	$2\left[\ln\left(\frac{\hat{\eta}}{1-\hat{\eta}}\right)+\frac{1}{1-\hat{\eta}}\right]$	$2\left[\ln\left(\frac{1-\hat{\eta}}{\hat{\eta}}\right)+\frac{1}{\hat{\eta}}\right]$	—
$h^2(P,Q)$	Hellinger	$(\sqrt{t}-1)^2$	$\frac{1}{2[\pi(1-\pi)]^{3/2}}$	$\frac{1}{[(1-c)c]^{3/2}}$	$2\sqrt{\frac{\hat{\eta}}{1-\hat{\eta}}}$	$2\sqrt{\frac{1-\hat{\eta}}{\hat{\eta}}}$	Boosting
$\chi^2(P,Q)$	Pearson χ^2	$(t-1)^2$	$\frac{2}{\pi^3}$	$\frac{4}{(1-c)^3}$	$2\frac{2\hat{\eta}-1}{(1-\hat{\eta})^2}$	$\frac{4}{1-\hat{\eta}}$	—
$\Psi(P,Q)$	Symmetric χ^2	$\frac{(t-1)^2(t+1)}{t}$	$\frac{2}{\pi^3}+\frac{2}{(1-\pi)^3}$	$\frac{4}{(1-c)^3}+\frac{4}{c^3}$	$2\frac{3\hat{\eta}-2}{(1-\hat{\eta})^2\hat{\eta}}$	$2\frac{1-3\hat{\eta}}{(1-\hat{\eta})\hat{\eta}^2}$	—

Table 2: Divergences and their corresponding functions f and weights γ along with the weights w and partial losses—see Section 5.4 ; confer Taneja (2005a); Liese and Vajda (2006). Topsøe (2000) 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 $\Psi(P,Q)=\chi^2(P,Q)+\chi^2(Q,P)$. The Boosting loss is also know as the "exponential" loss (Buja et al., 2005). Losses without a name are all indefinite losses and the forms given for $\ell(y,\cdot)$ in these cases are not normalised.

.

different. Using weight functions automatically filters out the effect of any affine offsets—if the weight functions corresponding to f_1 and f_2 match, then $\mathbb{I}_{f_1} = \mathbb{I}_{f_2}$. Finally observe that substituting $\gamma(\pi) = 8$ from the table into Theorem 18 we obtain $w^{\frac{1}{2}}(c) = \frac{1/4}{\nu(\pi,c)^3} \cdot 8 = 2$, consistent with the weight obtained above.

6. Graphical Representations

The last section described representations of risks and f-divergences in terms of weighted integrals of primitive functions. The values of the primitive functions lend themselves to a graphical interpretation that is explored in this section. In particular, a diagram called a *risk curve* is introduced. Risk curves are a useful aid to intuition when reasoning about risks, divergences and information and they are used in Section 7 to derive bounds between various divergences and risks.

Risk curves are closely related to the *cost curves* of Drummond and Holte (2006) as well as idealised *receiver operating characteristic, or ROC curves* (Fawcett, 2004). Proposition 20 makes this latter relationship explicit via a point-line duality between risk and ROC curves. Additionally, results about the Neyman-Pearson function by Torgersen (1981) allow us to establish a transformation between suitably smooth maximal ROC and minimal risk curves in Theorem 22. Despite the close ties between f-divergences and risks, and between risk curves and ROC curves, we show in Proposition 19 that the *area* under an ROC curve cannot be interpreted as an f-divergence.

6.1 ROC Curves

Plotting a *receiver operating characteristic curve* or *ROC curve* is a way of graphically summarising the performance of a test statistic. Recall from Section 3.1 that in the context of a binary experiment (P,Q) on a space $\mathcal{X}$, a test statistic τ is any function that maps points in $\mathcal{X}$ to the real line. Each choice of threshold $\tau_0 \in \mathbb{R}$ results in a classifier $r(x) = [\![\tau(x) \geq \tau_0]\!]$ and its corresponding classification rates. An ROC curve for the test statistic τ is simply a plot of the true positive rate of these classifiers as a function of their false positive rate as the threshold τ_0 varies over $\mathbb{R}$. Formally,

$$\mathrm{ROC}(\tau) := \{(\mathrm{FP}_\tau(\tau_0), \mathrm{TP}_\tau(\tau_0)) : \tau_0 \in \mathbb{R}\} \subset [0,1]^2. \tag{37}$$

A graphical example of an ROC curve is shown as the solid black line in Figure 3.

For a fixed experiment (P,Q), the Neyman-Pearson lemma provides an upper envelope for ROC curves. It guarantees that the ROC curve for the likelihood ratio $\tau^* = dP/dQ$ will lie above, or *dominate*, that of any other test statistic τ as shown in Figure 3. This is an immediate consequence of the likelihood ratio being the most powerful test since for each false positive rate (or size) α it will have the largest true positive rate (or power) β of all tests (Eguchi and Copas, 2001). Thus $\mathrm{ROC}(dP/dQ)$ is the *maximal* ROC curve.

The performance of a test statistic τ shown in an ROC curve is commonly summarised by the *Area Under the ROC Curve*, $\mathrm{AUC}(\tau)$, and is closely related to the Mann-Whitney-Wilcoxon statistic. Formally, if (P,Q) is a binary experiment and τ a test statistic the AUC is

$$\begin{aligned}\mathrm{AUC}(\tau) &:= \int_0^1 \beta_\tau(\alpha)\, d\alpha & (38)\\ &= \int_{-\infty}^{\infty} \mathrm{TP}_\tau(\tau_0)\, \mathrm{FP}'_\tau(\tau_0)\, d\tau_0, & (39)\end{aligned}$$

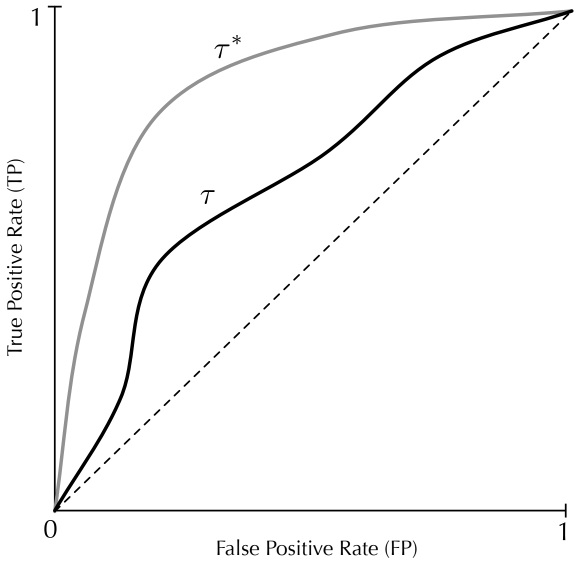


Figure 3: Example of an ROC diagram showing an ROC curve for an arbitrary statistical test τ (middle, bold curve) as well as an optimal statistical test τ^* (top, grey curve). The dashed line represents the ROC curve for a random, or uninformative statistical test.

where $\beta_\tau(\alpha) = \mathrm{TP}_\tau(\tau_0)$ for a $\tau_0 \in \mathbb{R}$ such that $\mathrm{FP}_\tau(\tau_0) = \alpha$.

In Section 3.1 the Neyman-Pearson lemma was used to argue that the curve $\beta(\alpha)$ for the likelihood ratio dominates all other curves. Since the likelihood ratio is used to define f-divergences, it is natural to ask whether the area under the maximal ROC curve is an f-divergence. Interestingly, the answer is "no".

Proposition 19 *There is no convex f such that $\mathbb{I}_f(P,Q) = \mathrm{AUC}(dP/dQ)$ for all distributions P and Q.*

Proof Note that an f-divergence's integral can be decomposed as follows

$$\mathbb{I}_f(P,Q) = \int_0^\infty f(t) \int_{\mathcal{X}_t} dQ\,dt, \tag{40}$$

where $\mathcal{X}_t := \{x \in \mathcal{X} : dP/dP(x) = t\} = (dP/dQ)^{-1}(t)$. Compare this to the definition of $\mathrm{AUC}(\tau)$ given in (39) when $\tau = dP/dQ$

$$\begin{aligned} \mathrm{AUC}(dP/dQ) &= \int_{-\infty}^\infty \mathrm{TP}_\tau(t)\,\mathrm{FP}'_\tau(t)\,dt \\ &= -\int_0^\infty (P \circ \tau^{-1})([t,\infty)) \int_{\mathcal{X}_t} dQ\,dt \end{aligned} \tag{41}$$

since $\mathrm{FP}'_\tau(t) = d/dt \int_t^\infty \int_{\mathcal{X}_t} dQ(x)\,dt = -\int_{\mathcal{X}_t} dQ$ and $dP/dQ \geq 0$. If we assume there exists an f such that for all binary experiments (P,Q), $\mathbb{I}_f(P,Q) = \mathrm{AUC}(dP/dQ)$ we would require the integrals in (40) and (41) to be equal for all (P,Q). This would require $f(t) = -(P \circ (dP/dQ)^{-1})([t,\infty))$ for all $t \in [0,\infty)$ which is not possible for all binary experiments (P,Q) simultaneously. ■

Although the maximal AUC for (P,Q) cannot be expressed as an f-divergence, Torgersen (1991) shows how it can be expressed as the variational divergence between the *product measures* $P \times Q$ and $Q \times P$. That is, $\mathrm{AUC}(dP/dQ) = V(P \times Q, Q \times P)$. Following up this connection and considering other f-divergences of product measures is left as future work.

It is important to realise that AUC is not a particularly intrinsic measure—just a common one. As the earlier discussion of integral representations has shown, there is value in considering weighted versions of integrals such as (38). As Hand (2008) notes in his commentary on a recent paper (outlining another type of performance curve): "To use all the values of the diagnostic instrument, when integrating to yield the overall AUC measure, it is necessary to decide what weight to give to each value in the integration. The AUC implicitly does this using a weighting derived empirically from the data." Along these lines, Xie and Priebe (2002) and Eguchi and Copas (2001) have suggested generalisations of the AUC that incorporates weights and show that certain choice of weight functions yield well-known losses.

A closer investigation of these generalisations of AUC and their connection to measures of divergence is also left as future work.

6.2 Risk Curves

Risk curves are a graphical representation closely related to ROC curves that take into account a prior π in addition to the binary experiment (P,Q). They provide a concise summary of the risk of an estimator $\hat{\eta}$ for the full range of costs $c \in [0,1]$ for a fixed prior $\pi \in [0,1]$, or, alternatively, for the full range of priors π given a fixed cost c.

A *risk curve for costs* for the estimator $\hat{\eta}$ is the set $\{(c, \mathbb{L}_c(\hat{\eta},\pi,P,Q)) : c \in [0,1]\}$ of points parameterised by cost.[20] A *risk curve for priors* for the estimator $\hat{\eta}$ is the set $\{(\pi, \mathbb{L}^{0\text{-}1}(\hat{\eta},\pi,P,Q)) : \pi \in [0,1]\}$.

Figure 4 shows an example of a *risk curve diagram*. On it is plotted the cost curves for an estimate $\hat{\eta}$ of a true posterior η on the same graph. The "tent" function also shown is the risk curve for the majority class predictor $\min\{(1-\pi)c, (1-c)\pi\}$. Here $\pi = \frac{1}{2}$. Other choices of $\pi \in (0,1)$ skew the tent and the curves under it towards 0 or 1.

In light of the weighted integral representations described in Theorem 16, several of the quantities can be associated with properties of a cost curve diagram. The weight function $w(c)$ associated with a loss ℓ can be interpreted as a weighting on the horizontal axis of a risk curve diagram. When the area under a risk curve is computed with respect to this weighting the result is the full risk $\mathbb{L}$ since $\mathbb{L}(\eta,\hat{\eta}) = \int_0^1 \mathbb{L}_c(\eta,\hat{\eta})\, w(c)\, dc$.

Furthermore, the weighted area between the risk curves for an estimate $\hat{\eta}$ and the true posterior η is the regret $\mathbb{L}(\eta,\hat{\eta}) - \underline{\mathbb{L}}(\eta)$ and the statistical information $\Delta\underline{\mathbb{L}}(\eta,M) = \underline{\mathbb{L}}(\pi,M) - \underline{\mathbb{L}}(\eta,M)$ is the weighted area between the "tent" risk curve for π and the risk curve for η.

The correspondence between ROC and risks curves is due to the relationship between the true class probability η and the likelihood ratio dP/dQ for a fixed π. As shown in Section 4.1, this

[20] Unlike the cost curves originally described by Drummond and Holte (2006), the version presented here does not normalise the risk, and plots the cost on the horizontal axis rather than the product of the prior probability and cost.

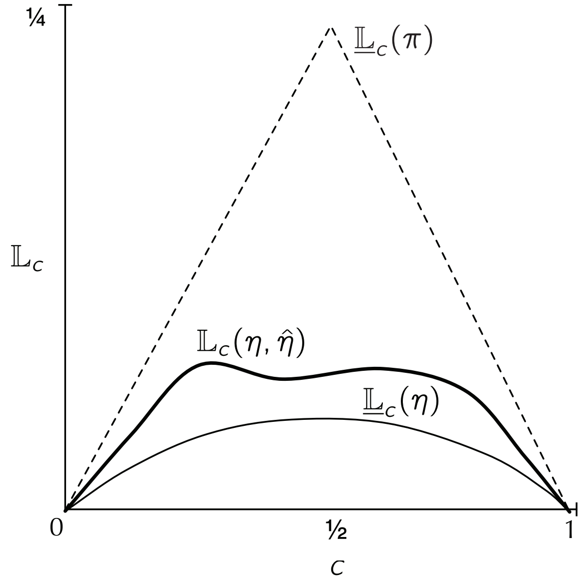


Figure 4: Example of a risk curve for costs diagram showing risk curves for costs for the true posterior probability η (bottom, solid curve), an estimate $\hat{\eta}$ (middle, bold curve) and the majority class or prior estimate (top, dashed curve).

relationship is

$$\frac{dP}{dQ} = \lambda_\pi(\eta) = \frac{1-\pi}{\pi}\frac{\eta}{1-\eta}.$$

Each cost $c \in [0,1]$ can be mapped to a corresponding test statistic threshold $\tau_0 = \lambda_\pi(c)$ and *vice versa*.

Drummond and Holte (2006) show that their cost curves have a point-line dual relationship with ROC curves. As can be established with some straight-forward algebra, the same result holds for our risk diagrams.

Proposition 20 *For a given point* $(\mathrm{FP},\mathrm{TP})$ *on an ROC diagram the corresponding line in a risk diagram is*

$$\mathbb{L}_c = (1-\pi)\,c\,\mathrm{FP} + \pi\,(1-c)\,(1-\mathrm{TP}), \quad c \in [0,1]$$

Conversely, the line in ROC space corresponding to a point $(c,\mathbb{L}_c)$ *in risk space is*

$$\mathrm{TP} = \frac{(1-\pi)c}{\pi(1-c)}\mathrm{FP} + \frac{(1-\pi)c - \mathbb{L}_c}{\pi(1-c)}, \quad \mathrm{FP} \in [0,1].$$

An example of this relationship is shown graphically[21] in Figure 5 between the point A and the line A*.

[21] An applet that demonstrates the relationship can be found at `http://mark.reid.name/iem/visualising-roc-and-cost-curve-duality.html`.

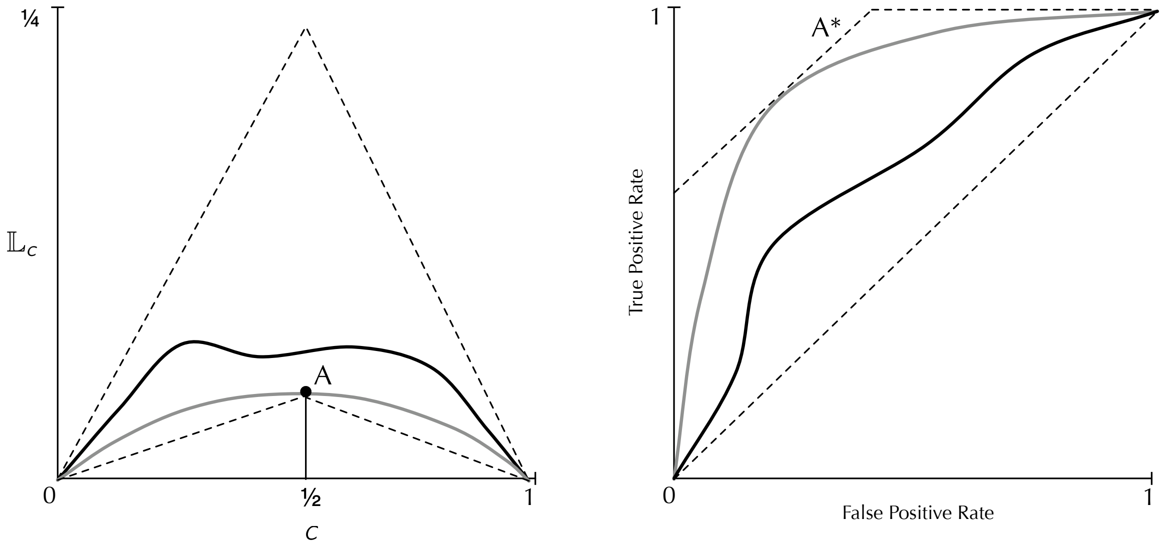


Figure 5: Cost curve diagram (left) and corresponding ROC diagram (right). The black curves on the left and right represent risk and classification rates of an example predictor. The grey Bayes risk curve on the left corresponds to the dominating grey ROC curve on the right for the likelihood statistic. Similarly, the dashed tent on the left corresponds to the dashed diagonal ROC line on the right. The point labelled A in the risk diagram corresponds to the line labelled A* in the ROC diagram.

6.3 Transforming from ROC to Risk Curves and Back

As mentioned earlier, the Neyman-Pearson lemma guarantees the ROC curve for η is maximal. This corresponds to the cost curve being minimal. In fact, these relationships are dual in the sense that there exists a transformation from one to the other as we shall now show. We make use of a connection between the Neyman-Pearson function in (11) and the maximal ROC curve due to Torgersen (1981). For completeness, a proof using our nomenclature can be found in Appendix A.7.

Theorem 21 *Let $\beta(\alpha,P,Q)$ be the Neyman-Pearson function for the binary experiment (P,Q) and let $\underline{\mathbb{L}}(\pi,P,Q)$ be the 0-1 Bayes risk on the same experiment for the prior π. Then, for any choice of $\pi\in[0,1]$ we have*

$$\underline{\mathbb{L}}(\pi,P,Q)=\underline{\mathbb{L}}=\min_{\alpha\in[0,1]}((1-\pi)\alpha+\pi(1-\beta(\alpha,P,Q)) \tag{42}$$

and conversely for any $\alpha\in[0,1]$,

$$\beta(\alpha,P,Q)=\inf_{\pi\in(0,1]}\frac{1}{\pi}((1-\pi)\alpha+\pi-\underline{\mathbb{L}}(\pi,P,Q)). \tag{43}$$

$\pi\mapsto\underline{L}(\pi,P,Q)$ is the lower envelope of a parameterized (by π) family of affine functions (in α) and is thus concave. When $\beta(\cdot)$ and $\underline{\mathbb{L}}(\cdot)$ are smooth, explicit closed form formulas can be found:

Theorem 22 *Suppose* β *and* $\underline{\mathbb{L}}$ *are differentiable on* $(0,1]$ *and* $[0,1]$ *respectively. Then*

$$\underline{\mathbb{L}}(\pi) = (1-\pi)\check{\beta}(\pi) + \pi(1-\beta(\check{\beta}(\pi))), \;\; \pi \in [0,1], \tag{44}$$

where

$$\check{\beta}(\pi) := \beta'^{-1}\left(\frac{1-\pi}{\pi}\right)$$

and

$$\beta(\alpha) = \frac{1}{\check{\underline{\mathbb{L}}}(\alpha)}\left[(1-\check{\underline{\mathbb{L}}}(\alpha))\alpha + \check{\underline{\mathbb{L}}}(\alpha) - \underline{\mathbb{L}}(\check{\underline{\mathbb{L}}}(\alpha))\right], \;\; \alpha \in (0,1], \tag{45}$$

where

$$\begin{aligned}\check{\underline{\mathbb{L}}}(\alpha) &:= \tilde{\underline{\mathbb{L}}}^{-1}(\alpha) \wedge 1, \\ \tilde{\underline{\mathbb{L}}}(\pi) &:= \underline{\mathbb{L}}(\pi) - \pi\underline{\mathbb{L}}'(\pi).\end{aligned}$$

The proof can be found in Appendix A.6.

Using (45) we present an example. Consider $\underline{\mathbb{L}}(\pi) = \gamma\pi(1-\pi)$ for $\gamma \in [0,1]$ One can readily check that $\tilde{\underline{\mathbb{L}}}_{(\gamma)}(\pi) = \gamma\pi^2$. Hence $\tilde{\underline{\mathbb{L}}}_{(\gamma)}^{-1}(\alpha) = \sqrt{\frac{\alpha}{\gamma}} \in \left[0, \frac{1}{\gamma}\right]$. Thus $\check{\underline{\mathbb{L}}}_{(\gamma)}(\alpha) = 0 \vee \tilde{\underline{\mathbb{L}}}_{(\gamma)}^{-1}(\alpha) \wedge 1 = \sqrt{\alpha/\gamma} \wedge 1$. Substituting and rearranging we find that the corresponding β is given by

$$\beta_\gamma(\alpha) = \frac{\alpha + \gamma + (\sqrt{\alpha/\gamma} \wedge 1)(1-\alpha-\gamma)}{\sqrt{\alpha/\gamma} \wedge 1}.$$

A graph of this $\beta(\cdot)$ is given in figure 6.

By construction $\beta(1) = 1$ and β is concave and continuous on $(0,1]$. The following lemma is due to Torgersen (1991). Given mild conditions on the space of instances, this gives a corollary which guarantees that all concave curves on a risk diagram can be realised by some pair of distributions. Their proofs can be found in Appendix A.8 and Appendix A.9, respectively.

Lemma 23 *Suppose* $\mathcal{X}$ *contains a connected component* $\mathcal{C}$. *Let* $\phi\colon [0,1] \to [0,1]$ *be an arbitrary function that is concave and continuous on* $(0,1]$ *such that* $\phi(1) = 1$. *Then there exists distributions* P *and* Q *on* $\mathcal{X}$ *such that* $\beta(\alpha, P, Q) = \phi(\alpha)$ *for all* $\alpha \in [0,1]$.

Corollary 24 *Suppose* $\mathcal{X}$ *contains a connected component. Let* $\psi\colon [0,1] \to [0,1]$ *be an arbitrary concave function such that for all* $\pi \in [0,1]$, $0 \leq \psi(\pi) \leq \pi \wedge (1-\pi)$. *Then there exists distributions* P *and* Q *on* $\mathcal{X}$ *such that* $\underline{\mathbb{L}}(\pi, P, Q) = \psi(\pi)$ *for all* $\pi \in [0,1]$.

The corollary shows that reasoning about cost-weighted risks for all possible binary experiments (P,Q) can be done purely geometrically. Each experiment can be associated with a concave curve and *vice versa* so that the existence of an experiment becomes equivalent to the existence of a concave curve with certain properties. This relationship is exploited in the next section to establish bounds for f-divergences in Theorem 30.

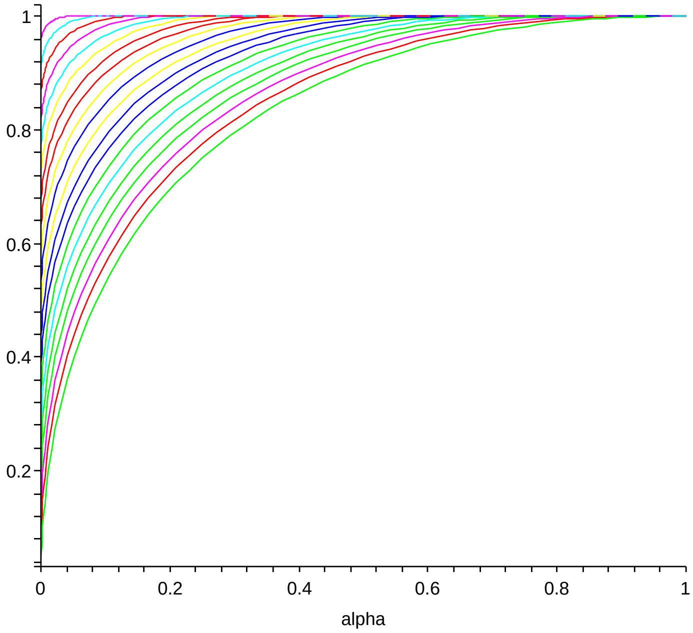


Figure 6: Graph of the parameterised Neyman-Pearson function $\alpha \mapsto \beta_\gamma(\alpha, P, Q)$ for $\gamma = i/20$, $i = 1, \ldots, 20$. (See text.)

7. Bounding General Objects in Terms of Primitives

All of the above results are exact—they are exact representations of particular primitives or general objects in terms of other primitives. Another type of relationship is an inequality. In this section we consider how we can (tightly) bound the value of a general object ($\mathbb{I}_f$ or B_w) in terms of primitive objects (V_π—the generalised variational divergence defined below—or B_c, the regret with respect to the cost weight loss (29)). Bounding $\mathbb{I}_f(P,Q)$ in terms of $V_\pi(P,Q)$ is a generalisation of the classical Pinsker inequality (Pinsker, 1964). Bounding $B_w(\eta,\hat{\eta})$ in terms of $B_c(\eta,\hat{\eta})$ is a generalisation of the so-called "surrogate regret bounds" (Zhang, 2004b; Bartlett et al., 2006).

As explained previously, we work with the *conditional* Bregman divergence $B_w(\eta,\hat{\eta})$. Results in terms of $B_w(\eta,\hat{\eta})$, $\eta,\hat{\eta} \in [0,1]$ immediately imply results for $\mathbb{B}_w(\eta,\hat{\eta})$, where $\eta,\hat{\eta} \in [0,1]^{\mathcal{X}}$ by taking expectations with respect to X.

7.1 Surrogate Regret Bounds

Suppose for some fixed $c_0 \in (0,1)$ that $B_{c_0}(\eta,\hat{\eta}) = \alpha$. What can be said concerning the value of $B_w(\eta,\hat{\eta})$ for an arbitrary weight function w? Surrogate regret bounds answer this question by showing how the value of B_{c_0} is controlled by a function of B_w. That is, $B_{c_0} \leq F(B_w)$ for some non-decreasing F. The main result of this subsection, Theorem 25, presents a general surrogate bound for proper losses implicitly as $B_w \geq F^{-1}(B_{c_0})$. However, as Corollary 28 shows, this implicit bound can always be inverted.

Previous work on this problem is summarised in Appendix D. Apart from their theoretical interest, these bounds have direct practical implications: it can often be much simpler to minimise $B_w(\eta, \hat{\eta})$ over $\hat{\eta}$ than to minimise $B_c(\eta, \hat{\eta})$. The bounds below will tell the user of such a scheme the maximum price they will have to pay, in terms of statistical performance, for using a particular surrogate.

Theorem 25 *Let $c_0 \in (0,1)$ and let $B_{c_0}(\eta, \hat{\eta})$ denote the point-wise regret for the cost-weighted loss ℓ_{c_0}. Suppose it is known that $B_{c_0}(\eta, \hat{\eta}) = \alpha$. Then the point-wise regret $B(\eta, \hat{\eta})$ for any proper surrogate loss ℓ with point-wise risk L and Bayes risk $\underline{L}$ satisfies*

$$B(\eta, \hat{\eta}) \geq \psi(c_0, \alpha) \vee \psi(c_0, -\alpha), \tag{46}$$

where

$$\psi(c_0, \alpha) := B(c_0, c_0 + \alpha) = \underline{L}(c_0) - \underline{L}(c_0 + \alpha) + \alpha \underline{L}'(c_0).$$

Furthermore (46) is tight.

The proof of this bound is almost a direct consequence of the fact that regrets for proper losses are Bregman divergences (see Section 4.4). This is a simplified version of an earlier proof by Reid and Williamson (2009). We will make use of the following expression for B_c derived by Buja et al. (2005). Its proof can be found in Appendix A.10.

Lemma 26 *Suppose L_c is the conditional risk for cost-sensitive misclassification loss (see 5.2). For any loss $c \in [0,1]$ the cost-weighted regret $B_c(\eta, \hat{\eta}) := L_c(\eta, \hat{\eta}) - \underline{L}_c(\eta)$ satisfies*

$$B_c(\eta, \hat{\eta}) = |\eta - c| [\![\eta \wedge \hat{\eta} < c \leq \eta \vee \hat{\eta}]\!].$$

Proof (Theorem 25) Let B be the conditional regret associated with some arbitrary proper loss ℓ and suppose that we know the cost-weighted regret $B_{c_0}(\eta, \hat{\eta}) = \alpha$. By Lemma 26, this implies that $\alpha = \eta - c_0$ when $\hat{\eta} \leq c_0 < \eta$ and $\alpha = c_0 + \eta$ when $\eta \leq c_0 < \hat{\eta}$. Since $B(\eta, \hat{\eta})$ is a Bregman divergence its value decreases as $|\eta - \hat{\eta}|$ decreases (see Section 2.5). Thus, in the first case we have $\hat{\eta} \leq c_0 < c_0 + \alpha = \eta$ and so $B(\eta, \hat{\eta}) = B(c_0 + \alpha, \hat{\eta}) \geq B(c_0 + \alpha, c_0)$ and is minimised when $\hat{\eta} = c_0$.

The proof of the second case, when $\eta = c_0 - \alpha \leq c_0 < \hat{\eta}$ proceeds identically. Thus, $B(\eta, \hat{\eta})$ is no smaller than each of $B(c_0 + \alpha, c_0)$ and $B(c_0 - \alpha, c_0)$, giving the required result. ■

By restricting attention to the case when $c_0 = \frac{1}{2}$ and symmetric losses we obtain, as a corollary, a result similar to that presented by Bartlett et al. (2006) for surrogate margin losses since $B_{\frac{1}{2}}$ is easily shown to be half the 0-1 regret. It is obtained by substituting $\alpha = \frac{1}{2}$ and noting the symmetry of L implies $\underline{L}'(\frac{1}{2}) = 0$; Appendix D contains some examples illustrating this special case.

Corollary 27 *If $\underline{L}$ is symmetric—that is, $\underline{L}(\frac{1}{2} - c) = \underline{L}(\frac{1}{2} + c)$ for $c \in [0, \frac{1}{2}]$—and $B_{\frac{1}{2}}(\eta, \hat{\eta}) = \alpha$, then*

$$B(\eta, \hat{\eta}) \geq \underline{L}(\tfrac{1}{2}) - \underline{L}(\tfrac{1}{2} + \alpha).$$

The bounds in Theorem 25 can be inverted to allow the approximate minimisation of a cost-weighted loss via the minimisation of a surrogate loss.

Corollary 28 *Minimising $B(\eta,\hat{\eta})$ w.r.t. $\hat{\eta}$ minimises the bound on $B_c(\eta,\hat{\eta})$ for each $c \in (0,1)$.*

Proof To see this, let $\psi'(c_0,\alpha) := \frac{\partial}{\partial\alpha}\psi(c_0,\alpha) = -\underline{L}'(c_0+\alpha)+\underline{L}'(c_0)$. Since $\underline{L}$ is concave, $\underline{L}'$ is non-increasing and hence $\underline{L}'(c_0+\alpha) \leq \underline{L}'(c_0)$ and so $\psi'(c_0,\alpha) \geq 0$ and therefore $\alpha \mapsto \psi(c_0,\alpha)$ is non-decreasing and thus invertible (although there may be non-uniqueness at points where $\psi(c_0,\alpha)$ is constant in α). This invertibility means minimising $B(\eta,\hat{\eta})$ w.r.t. $\hat{\eta}$, minimises the bound on $B_c(\eta,\hat{\eta})$. ■

Finally, Theorem 25 can be used to immediately establish a loose, second-order bound in α for symmetric losses in terms of their weight function, similar to a result due to Buja et al. (2005).

Corollary 29 *Suppose B_w is the regret for a symmetric proper loss ℓ with associated weight function w. Then*

$$B_w(\eta,\hat{\eta}) \geq \frac{w(\frac{1}{2})}{2}\left[B_{\frac{1}{2}}(\eta,\hat{\eta})\right]^2.$$

Proof A Taylor series expansion of the second term in the bound of Corollary 27 about $\alpha = \frac{1}{2}$ gives

$$B_w(\eta,\hat{\eta}) \geq \frac{w(\frac{1}{2})}{2}\alpha^2 + \frac{w''(\frac{1}{2})}{24}\alpha^4 + \cdots$$

since the linear term cancels and there is no third order term since w is symmetric and thus $w'(\frac{1}{2}) = 0$. Setting $\alpha = B_{\frac{1}{2}}(\eta,\hat{\eta})$ gives the result. ■

Some extensions to the above result have been recently presented by Scott (2010).

7.2 General Pinsker Inequalities for Divergences

The many different f divergences are single number summaries of the relationship between two distributions P and Q. Each f-divergence emphasises different aspects. Merely considering the functions f by which f-divergences are traditionally defined makes it hard to understand these different aspects, and harder still to understand how knowledge of $\mathbb{I}_{f_1}$ constrains the possible values of $\mathbb{I}_{f_2}$. When $\mathbb{I}_{f_1} = V$ (a special primitive for $\mathbb{I}_f$) and $\mathbb{I}_{f_2} = \mathrm{KL}$, this is a classical problem that has been studied for decades; Appendix E summarises the history.

Vajda (1970) posed the question of a *tight lower bound* on KL-divergence in terms of variational divergence. This "best possible Pinsker inequality" takes the form

$$L(V) := \inf_{V(P,Q)=V} \mathrm{KL}(P,Q), \quad V \in [0,2), \tag{47}$$

where the infimum is over all P and Q such that $V(P,Q) = V$. Recently Fedotov et al. (2003) presented an *implicit* (parametric) version of the form

$$(V(t),L(t))_{t\in\mathbb{R}^+}, \tag{48}$$
$$V(t) = t\left(1-\left(\coth(t)-\frac{1}{t}\right)^2\right), \quad L(t) = \ln\left(\frac{t}{\sinh(t)}\right)+t\coth(t)-\frac{t^2}{\sinh^2(t)}.$$

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 (47).

We make use of a generalised notion of variational divergence:

$$V_\pi(P,Q) := 2 \sup_{r\in[-1,1]^{\mathcal{X}}} |\pi\mathbb{E}_P r - (1-\pi)\mathbb{E}_Q r|, \tag{49}$$

where $\pi \in (0,1)$ and the supremum is over all measurable functions from $\mathcal{X}$ to $[-1,1]$.

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 π. Since $\pi \mapsto V_\pi(P,Q)$ is concave, the piecewise 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 γ. This observation forms the basis of the theorem stated below.

Theorem 30 *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

$$\begin{aligned} A_n \quad := \quad & \Bigg\{ \mathbf{a} = (a_1,\ldots,a_n) \in \mathbb{R}^n : \\ & \frac{\psi_{i+1}-\psi_i}{\pi_{i+1}-\pi_i} \le a_i \le \frac{\psi_i - \psi_{i-1}}{\pi_i - \pi_{i-1}},\ i = 1,\ldots,n \Bigg\}. \end{aligned} \tag{50}$$

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

$$\tilde{\pi}_i \quad := \quad \frac{\psi_i - \psi_{i+1} + a_{i+1}\pi_{i+1} - a_i\pi_i}{a_{i+1}-a_i},\ i = 0,\ldots,n, \tag{51}$$

$$j \quad := \quad \{k \in \{1,\ldots,n\} : \tilde{\pi}_k < \tfrac{1}{2} \le \tilde{\pi}_{k+1}\}, \tag{52}$$

$$\bar{\pi}_i \quad := \quad [\![i<j]\!]\tilde{\pi}_i + [\![i=j]\!]\tfrac{1}{2} + [\![j<i]\!]\tilde{\pi}_{i-1}, \tag{53}$$

$$\alpha_{\mathbf{a},i} \quad := \quad [\![i \le j]\!](1-a_i) + [\![i>j]\!](-1-a_{i-1}), \tag{54}$$

$$\beta_{\mathbf{a},i} \quad := \quad [\![i\le j]\!](\psi_i - a_i\pi_i) + [\![i>j]\!](\psi_{i-1} - a_{i-1}\pi_{i-1}) \tag{55}$$

for $i = 0,\ldots,n+1$ and let γ_f be the weight corresponding to f given by (28).

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

$$\mathbb{I}_f(P,Q) \quad \ge \quad \min_{\mathbf{a}\in A_n} \sum_{i=0}^n \int_{\bar{\pi}_i}^{\bar{\pi}_{i+1}} (\alpha_{\mathbf{a},i}\pi + \beta_{\mathbf{a},i})\gamma_f(\pi)d\pi \tag{56}$$

$$\begin{aligned} = \quad & \min_{\mathbf{a}\in A_n} \sum_{i=0}^n \big[(\alpha_{\mathbf{a},i}\bar{\pi}_{i+1} + \beta_{\mathbf{a},i})\,\Gamma_f(\bar{\pi}_{i+1}) - \alpha_{\mathbf{a},i}\bar{\Gamma}_f(\bar{\pi}_{i+1}) \\ & \quad - (\alpha_{\mathbf{a},i}\bar{\pi}_i + \beta_{\mathbf{a},i})\,\Gamma_f(\bar{\pi}_i) + \alpha_{\mathbf{a},i}\bar{\Gamma}_f(\bar{\pi}_i)\big], \end{aligned} \tag{57}$$

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

Equation 57 follows from (56) by integration by parts. The remainder of the proof is in Section A.12. Although (57) looks daunting, we observe: (1) the constraints on **a** are convex (in fact they are a box constraint); and (2) the objective is a relatively benign function of **a**.

When $n = 1$ the result simplifies considerably. If in addition $\pi_1 = \frac{1}{2}$ then $V_{\frac{1}{2}}(P,Q) = \frac{1}{4}V(P,Q)$. It is then a straightforward exercise to explicitly evaluate (56), especially when γ_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 31 *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 γ *is symmetric about* $\frac{1}{2}$ *and convex,*

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

and Γ_f *and* $\overline{\Gamma}_f$ *are as in Theorem 30.*

This theorem gives the first explicit representation of the optimal Pinsker bound.[22]

Corollary 32 *The following special cases hold (*γ *symmetric about* $1/2$*).*

$$\begin{aligned}
h^2(P,Q) &\geq 2 - \sqrt{4 - V^2}, \\
\mathrm{J}(P,Q) &\geq 2V \ln\left(\tfrac{2+V}{2-V}\right), \\
\Psi(P,Q) &\geq \frac{8V^2}{4 - V^2}, \\
\mathrm{I}(P,Q) &\geq \left(\tfrac{1}{2} - \tfrac{V}{4}\right)\ln(2 - V) + \left(\tfrac{1}{2} + \tfrac{V}{4}\right)\ln(2 + V) - \ln(2), \\
\mathrm{T}(P,Q) &\geq \ln\left(\tfrac{4}{\sqrt{4 - V^2}}\right) - \ln(2).
\end{aligned}$$

The following special cases hold (γ *is not symmetric)*

$$\chi^2(P,Q) \geq [\![V < 1]\!]V^2 + [\![V \geq 1]\!]\tfrac{V}{(2-V)}, \tag{58}$$

$$\mathrm{KL}(P,Q) \geq \min_{\beta \in [V-2,2-V]} \left(\tfrac{V+2-\beta}{4}\right)\ln\left(\tfrac{\beta-2-V}{\beta-2+V}\right) + \left(\tfrac{\beta+2-V}{4}\right)\ln\left(\tfrac{\beta+2-V}{\beta+2+V}\right). \tag{59}$$

By plotting both (48) and (59) one can confirm that the two bounds (implicit and explicit) coincide; see Figure 7.

The above theorem suggests a means by which one can *estimate* an f-divergence by estimating a sequence $(\mathbb{L}_{c_i}(\pi,P,Q))_{i=1}^{n}$. A simpler version of such an idea (more directly using the representation (27)) has been studied by Song et al. (2008).

[22] A summary of existing results and their relationship to those presented here is given in Appendix E.

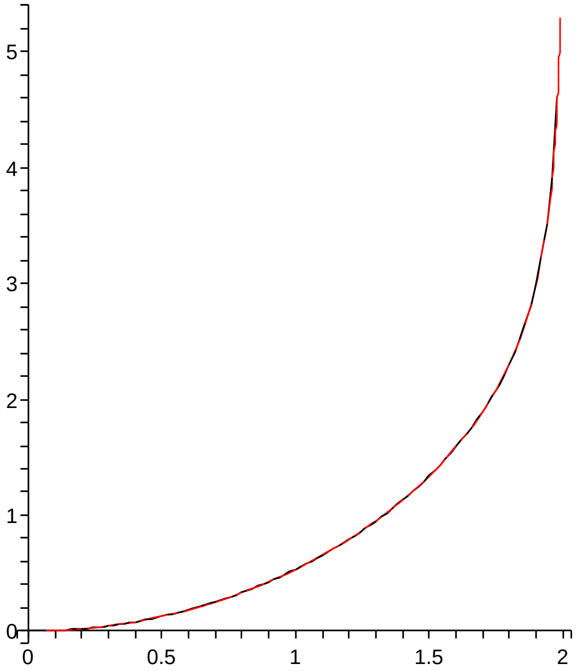


Figure 7: Lower bound on $\mathrm{KL}(P,Q)$ as a function of the variational divergence $V(P,Q)$. Both the explicit bound (59) and Fedotorev et al.'s implicit bound (48) are plotted.

8. Variational Representations

We have already seen a number of connections between the Bayes risk

$$\underline{\mathbb{L}}(\pi,P,Q)=\inf_{\hat{\eta}\in[0,1]^{\mathcal{X}}}\mathbb{E}_{\mathsf{X}\sim M}\left[\ell(\eta(\mathsf{X}),\hat{\eta}(\mathsf{X}))\right]$$

and the f-divergence

$$\mathbb{I}_f(P,Q)=\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]. \tag{60}$$

Comparing these definitions leads to an obvious and intriguing point: the definition of $\underline{\mathbb{L}}$ involves an optimisation, whereas that for $\mathbb{I}_f$ does not. Observe that the normal usage of these quantities is that one wishes to know not just the real number $\underline{\mathbb{L}}(\pi,P,Q)$, but also the estimate $\hat{\eta}\colon\mathcal{X}\to[0,1]$ that attains the minimal risk. In this section we will explore two views of $\mathbb{I}_f$—relating the standard definition to a *variational* one that explains where the optimisation is hidden in (60). We then explore some simpler relationships when using the linear "loss". In Appendix F we consider the variational representation of $\mathbb{I}_f$ obtained by representing f in terms of the LF dual $f^\star$. We also explore some generalisations that naturally arise from this representation and relate them to each other and to the standard f-divergence.

The easiest place to start, unsurprisingly, is with the variational divergence. Below we derive a straight-forward extension of the classical result relating $\underline{\mathbb{L}}^{0-1}(\frac{1}{2},P,Q)$ to $V(P,Q)$. We then explore variational representations for general f-divergences.

8.1 Generalised Variational Divergence

Let $\mathcal{C} \subseteq \{-1,1\}^{\mathcal{X}}$ denote a collection of measurable binary *classifiers* on $\mathcal{X}$. Consider the (constrained[23]) Bayes risk for 0-1 loss minimised over this set:

$$\underline{\mathbb{L}}_{\mathcal{C}}^{0-1}(\pi,P,Q) = \inf_{r\in\mathcal{C}} \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathbb{P}}[\ell^{0-1}(r(\mathsf{X}),\mathsf{Y})]. \tag{61}$$

The variational divergence is so called because it can be written

$$V(P,Q) = 2\sup_{A\subseteq\mathcal{X}} |P(A)-Q(A)|, \tag{62}$$

where the supremum is over all measurable subsets of $\mathcal{X}$. Since

$$V(P,Q) = \sup_{r\in[-1,1]^{\mathcal{X}}} |\mathbb{E}_P r - \mathbb{E}_Q r|,$$

consider the following generalisation of V:

$$V_{\mathcal{R},\pi}(P,Q) := 2\sup_{r\in\mathcal{R}\subseteq[-1,1]^{\mathcal{X}}} |\pi\mathbb{E}_P r - (1-\pi)\mathbb{E}_Q r|, \tag{63}$$

where $\pi \in (0,1)$ and the supremum is over all measurable functions from $\mathcal{X}$ to $[-1,1]$. (If $\mathcal{R} = [-1,1]^{\mathcal{X}}$ we just write $V_\pi(P,Q)$.) When $\pi = \frac{1}{2}$ this is a scaled version of what Müller (1997a,b) calls an *integral probability metric*.[24]

If $\mathcal{R}$ is *symmetric about zero* ($r \in \mathcal{R} \Rightarrow -r \in \mathcal{R}$), then the absolute value signs in (63) can be removed. To see this, suppose the supremum was attained at $\bar{r}$ and that $\alpha := \pi\mathbb{E}_P\bar{r} - (1-\pi)\mathbb{E}_Q\bar{r} < 0$. Choose $\bar{r}' := -\bar{r}$ and observe that $\pi\mathbb{E}_P\bar{r}' - (1-\pi)\mathbb{E}_Q\bar{r}' = -\alpha > 0$. Thus $V_{\mathcal{R},\pi}(P,Q) = 2\sup_{r\in\mathcal{R}\subseteq[-1,1]^{\mathcal{X}}}(\pi\mathbb{E}_P r - (1-\pi)\mathbb{E}_Q r)$.

Let $\operatorname{sgn}\mathcal{R} := \{\operatorname{sgn} r \colon r \in \mathcal{R}\}$ and for $a,b \in \mathbb{R}$, let $a\mathcal{R}+b := \{ar+b \colon r \in \mathcal{R}\}$.

Theorem 33 *Suppose $\mathcal{R} \subseteq [-1,1]^{\mathcal{X}}$ is symmetric about zero and $\operatorname{sgn}\mathcal{R} \subseteq \mathcal{R}$. For all $\pi \in (0,1)$ and all P and Q*

$$\underline{\mathbb{L}}_{(\operatorname{sgn}\mathcal{R}+1)/2}^{0-1}(\pi,P,Q) = \tfrac{1}{2} - \tfrac{1}{4}V_{\mathcal{R},\pi}(P,Q) \tag{64}$$

and the infimum in (61) corresponds to the supremum in (63).

The proof is in Appendix A.11.

[23] Tong and Koller (2000) call this the *restricted* Bayes risk.

[24] Zolotarev (1984) calls this a *probability metric with ζ-structure*. There are probability metrics that are neither f-divergences nor integral probability metrics. A large collection is due to Rachev (1991). A recent survey on relationships (inequalities and some representations) has been given by Gibbs and Su (2002). The idea of generalising variational divergence by restricting the set the supremum is taken over is also used by Ben-David et al. (2010).

8.2 The Linear "Loss" and the Generalised Variational Divergence

Theorem 33 shows that computing $V_{\mathcal{R},\pi}$ involves an optimisation problem equivalent to that arising in the determination of $\underline{\mathbb{L}}$. The arg min in the definition of $\underline{\mathbb{L}}$ is usually called the *hypothesis* (or *Bayes optimal hypothesis*). Following Borgwardt et al. (2006) we will call the arg max in (63) the *witness*.

When $\mathcal{R}=[-1,1]^{\mathcal{X}}$ and $\pi=\frac{1}{2}$, $\operatorname{sgn}\mathcal{R}\subseteq\mathcal{R}$ and furthermore $\mathcal{C}=(\operatorname{sgn}\mathcal{R}+1)/2=\{0,1\}^{\mathcal{X}}$ and so Theorem 33 reduces to the classical result that $\underline{\mathbb{L}}^{0-1}(\frac{1}{2},P,Q)=\frac{1}{2}-\frac{1}{4}V(P,Q)$ (Devroye et al., 1996).

The requirement that $\operatorname{sgn}\mathcal{R}\subseteq\mathcal{R}$ is unattractive. It is necessitated by the use of 0-1 loss. It can be removed by instead considering the *linear loss*

$$\ell^{\text{lin}}(r(x),y):=1-yr(x),\quad y\in\{-1,1\}.$$

If r is unrestricted, then there is no guarantee that $\ell^{\text{lin}}>-\infty$ and is thus a legitimate loss function. Below we will always consider $r\in\mathcal{R}$ such that the linear loss is bounded from below. Observe that the common hinge loss (Steinwart and Christmann, 2008) is simply $\ell^{\text{hinge}}(f(x),y)=0\vee\ell^{\text{lin}}(f(x),y)$.

Theorem 34 *Assume that $\mathcal{R}\subseteq[-a,a]^{\mathcal{X}}$ for some $a>0$ and is symmetric about zero. Then for all $\pi\in(0,1)$ and all distributions P and Q on $\mathcal{X}$*

$$\underline{\mathbb{L}}^{\text{lin}}_{\mathcal{R}}(\pi,P,Q)=1-\frac{1}{2}V_{\mathcal{R},\pi}(P,Q)$$

and the r that attains $\underline{\mathbb{L}}^{\text{lin}}_{\mathcal{R}}(\pi,P,Q)$ corresponds to the r that obtains the supremum in the definition of $V_{\mathcal{R},\pi}(P,Q)$.

Proof

$$\begin{aligned}
\underline{\mathbb{L}}^{\text{lin}}_{\mathcal{R}}(\pi,P,Q) &= \inf_{r\in\mathcal{R}}\left(\pi\mathbb{E}_{\mathsf{X}\sim P}\ell^{\text{lin}}(r(\mathsf{X}),-1)+(1-\pi)\mathbb{E}_{\mathsf{X}\sim Q}\ell^{\text{lin}}(r(\mathsf{X}),+1)\right)\\
&= \inf_{r\in\mathcal{R}}\left(\pi\mathbb{E}_{\mathsf{X}\sim P}(1+r(\mathsf{X}))+(1-\pi)\mathbb{E}_{\mathsf{X}\sim Q}(1-r(\mathsf{X}))\right)\\
&= \inf_{r\in\mathcal{R}}\left(\pi+\pi\mathbb{E}_P r+(1-\pi)-(1-\pi)\mathbb{E}_Q r\right)\\
&= 1+\inf_{r\in\mathcal{R}}\left(\pi\mathbb{E}_P r-(1-\pi)\mathbb{E}_Q r\right)\\
&= 1-\sup_{r\in\mathcal{R}}\left(\pi\mathbb{E}_P(-r)-(1-\pi)\mathbb{E}_Q(-r)\right)\\
&= 1-\sup_{r\in\mathcal{R}}\left(\pi\mathbb{E}_P r-(1-\pi)\mathbb{E}_Q r\right)\\
&= 1-\frac{1}{2}V_{\mathcal{R},\pi}(P,Q),
\end{aligned}$$

where the penultimate step exploits the symmetry of $\mathcal{R}$. ■

Now suppose that $\mathcal{R}=B_{\mathcal{H}}:=\{r\colon\|r\|_{\mathcal{H}}\leq 1\}$, the unit ball in $\mathcal{H}$, a Reproducing Kernel Hilbert Space (RKHS) (Schölkopf and Smola, 2002). Thus for all $r\in\mathcal{R}$ there exists a *feature map* $\phi\colon\mathcal{X}\to\mathcal{H}$ such that $r(x)=\langle r,\phi(x)\rangle_{\mathcal{H}}$ and $\langle\phi(x),\phi(y)\rangle_{\mathcal{H}}=k(x,y)$, where k is a positive definite *kernel* function. Borgwardt et al. (2006) show that

$$V^2_{B_{\mathcal{H}},\frac{1}{2}}(P,Q)=\frac{1}{4}\|\mathbb{E}_P\phi-\mathbb{E}_Q\phi\|^2_{\mathcal{H}}. \tag{65}$$

Given	Assumed	Derived
(P,Q)	$f \leftrightarrow \gamma$	$\mathbb{I}_f(P,Q)$
(π,P,Q)	$U \leftrightarrow w, W, \overline{W}$	$\mathbb{J}(U(\eta)) = \Delta\underline{\mathbb{L}}(\pi,P,Q)$ $\underline{\mathbb{L}}(\eta)$
	$\hat{\eta}$	$L_w(\eta,\hat{\eta})$, $B_w(\eta,\hat{\eta})$

Table 3: Summary relationships between key objects arising in Binary Experiments. “Given” indicates the object is given or provided by the world; “Assumed” is something the user of assumes or imposes in order to create a well defined problem; “Derived” indicates quantities that are derived from the primitives.

Thus

$$\underline{\mathbb{L}}^{\mathrm{lin}}_{\mathcal{R}}(\pi,P,Q) = 1 - \frac{1}{4}\|\mathbb{E}_P\phi - \mathbb{E}_Q\phi\|_{\mathcal{H}}. \tag{66}$$

Empirical estimators derived from the correspondence between (65) and (66) lead to the ν-Support Vector Machine and Maximum Mean Discrepancy; see Appendix H. Further generalizations of variational representations of $\mathbb{I}_f$ are explored in Appendix F.

9. Conclusions

There are several existing concepts that can be used to quantify the amount of information in a task and its difficulty: Uncertainty, Bregman information, statistical information, Bayes risk and regret, and f-divergences. Information is a difference in uncertainty; regret is a difference in risk. In the case of supervised binary class probability estimation, we have connected and extended several existing results in the literature to show how to translate between these perspectives. The representations allow a precise answer to the question of what are the primitives for binary experiments.

We have derived the integral representations in a simple and unified manner, and illustrated the value of the representations. Along the way we have drawn connections to a diverse set of concepts related to binary experiments: risk curves, cost curves, ROC curves and the area under them; variational representations of f-divergences, risks and regrets.

Two key consequences are surrogate regret bounds that are at once more general and simpler than those in the literature, and a generalisation of the classical Pinkser inequality providing, *inter alia*, an explicit form for the best possible Pinsker inequality relating Kullback-Leibler divergence and Variational divergence. We have also presented a new derivation of support vector machines and their relationship to Maximum Mean Discrepancy (integral probability metrics).

The key relationships between the basic objects of study are summarised in Table 3 and Figure 1 in §1.2.

All of the results we have presented demonstrate the fundamental and elementary nature of the cost-weighted misclassification loss, which is becoming increasingly appreciated in the Machine Learning literature (Bach et al., 2006; Beygelzimer et al., 2008). The viewpoint developed in this paper has also recently been used to better understand the structure of composite binary losses (losses involving a link function)—see Reid and Williamson (2010).

More generally, the present work is small part of a larger structural research agenda to understand the whole field of machine learning in terms of *relations* between problems. We envisage these relations being richer and more powerful than the already valuable *reductions* between learning problems. Much of the present literature on machine learning is highly solution focussed. Of course one does indeed like to *solve* problems, and we do not suggest otherwise. But it is hard to see structure in the panoply of solutions which continue to grow each year. The present paper is a first step to a pluralistic unification of a diverse set of machine learning problems. The goal we have in mind can be explained by analogy. There are several such analogies:

Computational Complexity Within the field of NP-completeness (Garey and Johnson, 1979; Johnson, 1982–1992; 2005–2007) lead to a detailed and structured understanding of the *relationships* between many fundamental problems and consequently guides the search for solutions for new problems.

Functional Analysis Compare Machine Learning problems with mathematical *functions*. In the 19th century, each function was considered separately. Functional Analysis (Lindström, 2008) *catalogued* them by considering *sets* of functions and *relations* (mappings) between them and subsequently developed many new and powerful tools. The increasing abstraction and focus on relations has remained a powerful force in mathematics (Wikipedia, 2007).

Biology A systematic *cataloging* (taxonomy) resonates with Biology's Linnean past—and taxonomies can indeed lead to standardisation and efficiency (Bowker and Star, 1999). But taxonomies alone are inadequate—it seems necessary to understand the relationships in a manner analogous to *Systems Biology* which "is about putting together rather than taking apart, integration rather than reduction.... Successful integration at the systems level must be built on successful reduction, but reduction alone is far from sufficient" (Noble, 2006).

Geology Finally, Lyell's *Principles of Geology* (Lyell, 1830) was a watershed in Geology's history (Bowker, 2005); prior work is *pre*-historical. Lyell's key insight was to explain the huge diversity of geological formations in terms of a relatively simple set of transformations applied repeatedly.

These analogies encourage our aspiration that by more systematically understanding the *relationships* between machine learning problems and how they can be *transformed* into each other, we will develop a better organised and more powerful toolkit for solving existing and future problems, and will make progress along the lines suggested by Hand (1994).

Acknowledgments

This work was motivated in part by discussions with Alekh Agarwal, Arthur Gretton, Ulrike von Luxburg and Bernhard Schölkopf in 2006. The authors were also aided by helpful conversations with Suvrit Sra and Bharath Sriperumbudur in Tübingen in 2008. Thanks also to Marcus Hutter for reading an early version of this paper and suggesting some useful changes and to Tim van Erven for finding several errors. Thanks also to the editor and reviewers for going along with our plan of a single (large!) integrated paper and their detailed and insightful reviews.

Parts of this work appeared earlier in the International Conference on Machine Learning 2009, and the Conference on Learning Theory, 2009.

This work was supported by the Australian Research Council and NICTA. NICTA is supported by the Australian Government through Backing Australia's Ability.

Appendix A. Proofs

This appendix presents the proofs that were omitted in the main body of the paper.

A.1 Proof of Corollary 3

Integration by parts of $t\phi''(t)$ gives $\int_0^1 t\,\phi''(t)\,dt = \phi'(1) - (\phi(1) - \phi(0))$ which can be rearranged to give

$$\phi'(1) = \int_0^1 t\,\phi''(t)\,dt + (\phi(1) - \phi(0)).$$

Substituting this into the Taylor expansion of $\phi(s)$ about 1 yields

$$\begin{aligned}
\phi(s) &= \phi(1) + \phi'(1)(s-1) + \int_s^1 (t-s)\,\phi''(t)\,dt \\
&= \phi(1) + \left[\int_0^1 t\,\phi''(t)\,dt + (\phi(1)-\phi(0))\right](s-1) + \int_0^1 (t-s)_+\,\phi''(t)\,dt \\
&= \phi(1) + (\phi(1)-\phi(0))(s-1) + \int_0^1 t(s-1)\,\phi''(t)\,dt + \int_0^1 (t-s)_+\,\phi''(t)\,dt \\
&= \phi(0) + (\phi(1)-\phi(0))s - \int_0^1 \psi(s,t)\,\phi''(t)\,dt,
\end{aligned}$$

where $\psi(s,t) := \min\{(1-t)s, (1-s)t\}$. This form of ψ is valid since

$$\begin{aligned}
-(t(s-1) + (t-s)_+) &= \begin{cases} -ts + t - t + s, & t \geq s \\ -ts + t, & t < s \end{cases} \\
&= \begin{cases} s - ts, & t \geq s \\ t - ts, & t < s \end{cases} \\
&= \min\{(1-t)s, (1-s)t\}
\end{aligned}$$

as required.

A.2 Proof of Theorem 6

Expanding the definition of the Jensen gap using the definition of ψ gives

$$\begin{aligned}
\mathbb{J}_\mu[\psi(\mathsf{S})] &= \mathbb{E}_\mu[\psi(\mathsf{S})] - \psi(\mathbb{E}_\mu[\mathsf{S}]) \\
&= \mathbb{E}_\mu[\phi(\mathsf{S}) + b\mathsf{S} + a] - (\phi(\mathbb{E}_\mu[\mathsf{S}]) + b\mathbb{E}_\mu[\mathsf{S}] + a) \\
&= \mathbb{E}_\mu[\phi(\mathsf{S})] + b\mathbb{E}_\mu[\mathsf{S}] + a - \phi(\mathbb{E}_\mu[\mathsf{S}]) - b\mathbb{E}_\mu[\mathsf{S}] - a \\
&= \mathbb{J}_\mu[\phi(\mathsf{S})]
\end{aligned}$$

as required.

A.3 Proof of Theorem 9

Proof Given a task $(\pi, P, Q; \ell)$ we need to first check that

$$f^{\pi}(t) := \underline{L}(\pi) - (\pi t + 1 - \pi)\underline{L}\left(\frac{\pi t}{\pi t + 1 - \pi}\right) \tag{67}$$

is convex and that $f^{\pi}(1) = 0$. This latter fact is obtained immediately by substituting $t = 1$ into $f^{\pi}(t)$ yielding $\underline{L}(\pi) - \underline{L}(\pi) = 0$. The convexity of f^{π} is guaranteed by Theorem 7, which shows that $\underline{L}$ is concave and the fact that the perspective transform of a convex function is always convex (see Section 2.1). Thus the function

$$t \mapsto I_{-\underline{L}}(\pi t, \pi t + 1 - \pi) = -(\pi t + 1 - \pi)\underline{L}\left(\frac{\pi t}{\pi t + 1 - \pi}\right)$$

is the composition of a convex function and an affine one and therefore convex.

Substituting (67) into the definition of f-divergence in (13) yields

$$\begin{aligned} \mathbb{E}_Q\left[f^{\pi}(dP/dQ)\right] &= \mathbb{E}_Q\left[\underline{L}(\pi) - \left(\pi\frac{dP}{dQ} + 1 - \pi\right)\underline{L}\left(\frac{\pi dP}{\pi dP + (1-\pi)dQ}\right)\right] \\ &= \underline{L}(\pi) - \int_{\mathcal{X}} \underline{L}\left(\pi\frac{dP}{dM}\right) dM \end{aligned}$$

since $dM = \pi dP + (1-\pi)dQ$. Recall that $\eta = \pi dP/dM$. Since $\underline{L}(\pi)$ is constant we note that $\underline{L}(\pi) = \mathbb{E}_M\left[\underline{L}(\pi)\right] = \underline{\mathbb{L}}(\pi, M)$ and so

$$\begin{aligned} \mathbb{E}_Q\left[f^{\pi}(dP/dQ)\right] &= \underline{L}(\pi) - \mathbb{E}_M\left[\underline{L}(\eta)\right] \\ &= \underline{\mathbb{L}}(\pi, M) - \underline{\mathbb{L}}(\eta, M) \\ &= \Delta\underline{\mathbb{L}}(\eta, M) \end{aligned}$$

as required for the forward direction.

Starting with

$$\underline{L}^{\pi}(\eta) := -\frac{1-\eta}{1-\pi} f\left(\frac{1-\pi}{\pi}\frac{\eta}{1-\eta}\right)$$

and substituting into the definition of statistical information in (20) gives us

$$\begin{aligned} \Delta\underline{\mathbb{L}}^{\pi}(\eta, M) &= \mathbb{E}_M\left[\underline{L}^{\pi}(\pi)\right] - \mathbb{E}_M\left[\underline{L}^{\pi}(\eta)\right] \\ &= \int_{\mathcal{X}} -\frac{1-\pi}{1-\pi} f(1)\, dM - \int_{\mathcal{X}} -\frac{1-\eta}{1-\pi} f\left(\frac{1-\pi}{\pi}\frac{\eta}{1-\eta}\right) dM \\ &= 0 + \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ \end{aligned}$$

since $f(1) = 0$, $dQ = (1-\eta)/(1-\pi)dM$ and

$$dP/dQ = \frac{1-\pi}{\pi}\frac{\eta}{1-\eta}$$

by the discussion in Section 4.1. This proves the converse statement of the theorem. ■

A.4 Proof of Corollary 13

Proof Let $f^{\diamond}(t) := t f(1/t)$ denote the Csiszár-dual of f as described in Section 2.1 above. It is known (see (16) and, for example, Liese and Vajda, 2006) that

$$\mathbb{I}_f(P,Q) = \mathbb{I}_{f^{\diamond}}(Q,P) \quad \text{if and only if} \quad f(t) = f^{\diamond}(t) + c_1 t + c_2$$

for some $c_1, c_2 \in \mathbb{R}$. Since f and γ are related by $f''\left(\frac{1-\pi}{\pi}\right) = \pi^3\gamma(\pi)$ we can argue as follows. Observe that $f^{\diamond\prime}(t) = f(1/t) - f'(1/t)/t$ and $f^{\diamond\prime\prime}(t) = f''(1/t)/t^3$. Hence $f^{\diamond\prime\prime}\left(\frac{1-\pi}{\pi}\right) = f''\left(\frac{\pi}{1-\pi}\right)\left(\frac{\pi}{1-\pi}\right)^3$. Let $\pi' = 1-\pi$. Thus $\frac{1-\pi}{\pi} = \frac{\pi'}{1-\pi'}$. Hence

$$\begin{aligned} f^{\diamond\prime\prime}\left(\frac{1-\pi}{\pi}\right) &= f''\left(\frac{1-\pi'}{\pi'}\right)\left(\frac{\pi}{1-\pi}\right)^3 \\ &= \pi'^3\gamma(\pi')\left(\frac{\pi}{1-\pi}\right)^3 \\ &= \pi^3\gamma(1-\pi). \end{aligned}$$

Thus if $\gamma(1-\pi) = \gamma(\pi)$, we have shown $\pi \mapsto \gamma(1-\pi)$ is the weight corresponding to $f^{\diamond}$. Observing that $\frac{\partial^2}{\partial t^2}(f^{\diamond}(t) + c_1 t + c_2) = f^{\diamond\prime\prime}$ concludes the proof. ■

A.5 Proof of Theorem 18

Proof Theorem 9 shows that

$$\underline{L}^{\pi}(\eta) = -\frac{1-\eta}{1-\pi} f\left(\frac{1-\pi}{\pi}\frac{\eta}{1-\eta}\right). \tag{68}$$

and we have seen from (32) that $w^{\pi}(c) = -(\underline{L}^{\pi})''(c)$. The remainder of this proof involves taking the second derivative of $\underline{L}$, doing some messy algebra and matching the result to the relationship between γ and f'' in (Equation 28).

Letting $r_\pi = r_\pi(\eta) = \frac{1-\pi}{\pi}\frac{\eta}{1-\eta}$ and taking derivatives of (68) yields

$$\begin{aligned} -(\underline{L}^{\pi})'(\eta) &= (1-\pi)^{-1}[-f(r_\pi) + (1-\eta)f'(r_\pi)r'_\pi] \\ -(\underline{L}^{\pi})''(\eta) &= (1-\pi)^{-1}[-f'(r_\pi)r'_\pi + (1-\eta)(f'(r_\pi)r''_\pi + f''(r_\pi)(r'_\pi)^2) - f'(r_\pi)r'_\pi] \\ &= (1-\pi)^{-1}[(-2r'_\pi + (1-\eta)r''_\pi)f'(r_\pi) + (1-\eta)(r''_\pi)^2 f''(r_\pi)]. \end{aligned}$$

However, the form of r_π means $r'_\pi = \frac{1-\pi}{\pi}\frac{1}{(1-\eta)^2}$ and so $r''_\pi = \frac{1-\pi}{\pi}\frac{2}{(1-\eta)^3}$. This means the coefficient of $f'(r_\pi)$ in the above expression vanishes

$$(-2r'_\pi + (1-\eta)r''_\pi) = \frac{1-\pi}{\pi}\left[\frac{-2}{(1-\eta)^2} + (1-\eta)\frac{2}{(1-\eta)^3}\right] = 0.$$

Substituting this back into $-(\underline{L})''$ gives us

$$\begin{aligned} -(\underline{L}^{\pi})''(\eta) &= \frac{1-\eta}{1-\pi} f''(r_\pi)(r'_\pi)^2 \\ &= \frac{1-\eta}{1-\pi} f''\left(\frac{1-\pi}{\pi}\frac{\eta}{1-\eta}\right)\frac{(1-\pi)^2}{\pi^2}\frac{1}{(1-\eta)^4} \\ w(\eta) &= \frac{1-\pi}{\pi^2(1-\eta)^3} f''\left(\frac{1-\pi}{\pi}\frac{\eta}{1-\eta}\right). \end{aligned}$$

By Equation 28 we have

$$\gamma(t) = \frac{1}{t^3} f''\left(\frac{1-t}{t}\right).$$

Letting $t = \frac{(1-c)\pi}{(1-c)\pi+(1-\pi)c}$ in that expression gives

$$\gamma\left(\frac{(1-c)\pi}{\nu(\pi,c)}\right) = \frac{\nu(\pi,c)^3}{(1-c)^3\pi^3} f''\left(\frac{1-\pi}{\pi}\frac{c}{1-c}\right).$$

Thus

$$\frac{\pi(1-\pi)}{\nu(\pi,c)^3}\gamma\left(\frac{(1-c)\pi}{\nu(\pi,c)}\right) = \frac{1-\pi}{\pi^2(1-c)^3} f''\left(\frac{1-\pi}{\pi}\frac{c}{1-c}\right) = w(c)$$

as required. The argument to show the inverse relationship is essentially the same. ■

A.6 Proof of Theorem 22

Proof Consider the right side of (42) and differentiate with respect to α:

$$\frac{\partial}{\partial\alpha}(1-\pi)\alpha+\pi(1-\beta(\alpha)) = (1-\pi)-\pi\beta'(\alpha).$$

Setting this to zero we have $(1-\pi) = \pi\beta'(\alpha)$ and thus $\beta'(\alpha) = \frac{1-\pi}{\pi}$. Since β is monotonically increasing and concave, β' is monotonically decreasing and non-negative. Thus we can set

$$\alpha = \beta'^{-1}\left(\frac{1-\pi}{\pi}\right) \in [0,1].$$

Substituting back into $(1-\pi)\alpha+\pi(1-\beta(\alpha))$ we obtain (44).

Now consider the right side of (43):

$$\frac{1}{\pi}((1-\pi)\alpha+\pi-\underline{\mathbb{L}}(\pi)). \tag{69}$$

Differentiating with respect to π we have $\frac{-\alpha}{\pi}-\frac{\underline{\mathbb{L}}'(\pi)}{\pi}+\frac{\underline{\mathbb{L}}(\pi)}{\pi^2}$. Setting this equal to zero we obtain

$$\begin{aligned} & \frac{-\alpha}{\pi}-\frac{\underline{\mathbb{L}}'(\pi)}{\pi}+\frac{\underline{\mathbb{L}}(\pi)}{\pi^2} = 0, \;\; \pi\in(0,1] \\ \Rightarrow \qquad & \alpha+\pi\underline{\mathbb{L}}'(\pi)-\underline{\mathbb{L}}(\pi) = 0. \end{aligned}$$

Observing the definition of $\tilde{\underline{\mathbb{L}}}$ we thus have that $\tilde{\underline{\mathbb{L}}}(\pi) = \alpha$. Now

$$\begin{aligned} \tilde{\underline{\mathbb{L}}}'(\pi) &= \frac{\partial}{\partial\pi}(-\pi\underline{\mathbb{L}}'(\pi)+\underline{\mathbb{L}}(\pi)) \\ &= -\pi\underline{\mathbb{L}}''(\pi)-\underline{\mathbb{L}}'(\pi)+\underline{\mathbb{L}}'(\pi) \\ &= -\pi\underline{\mathbb{L}}''(\pi) \\ &\geq 0 \end{aligned}$$

since $\underline{\mathbb{L}}$ is concave. Thus $\tilde{\underline{\mathbb{L}}}(\cdot)$ is monotonically non-decreasing and we can write $\pi = \tilde{\underline{\mathbb{L}}}^{-1}(\alpha)$. In order to ensure $\pi\in[0,1]$ we substitute $\pi = \check{\underline{\mathbb{L}}}(\alpha)$ into (69) to obtain (45). ■

A.7 Proof of Theorem 21

Proof Since the true positive rate for $r \in \{-1,1\}^{\mathcal{X}}$ is $\mathrm{TP}_r = P(r^{-1}(1))$ and the false positive rate for r is $\mathrm{FP}_r = Q(r^{-1}(1))$ we have

$$\beta(\alpha, P, Q) = \sup_{r\in\{-1,1\}^{\mathcal{X}}} \{P(\mathcal{X}_r^+) \ : \ Q(\mathcal{X}_r^+) \leq \alpha\},$$

where $\mathcal{X}_r^+ := r^{-1}(1)$.

Noting that the 0-1 loss of r is simply its probability of error—that is, the average of the false positive and false negative rates—we have for each $\pi \in [0,1]$ that the Bayes optimal 0-1 loss is

$$\underline{\mathbb{L}}(\pi, P, Q) = \inf_{r\in\{-1,1\}^{\mathcal{X}}} \{(1-\pi)Q(\mathcal{X}_r^+) + \pi(1-P(\mathcal{X}_r^+))\},$$

since the false negative rate $\mathrm{FN}_r = P(\mathcal{X} \setminus \mathcal{X}_r^+) = 1 - P(\mathcal{X}_r^+)$. Thus for all $\pi, \alpha \in [0,1]$, and all measurable functions $r\colon \mathcal{X} \to \{-1,1\}$,

$$\begin{aligned}
\underline{\mathbb{L}}(\pi, P, Q) &\leq (1-\pi)Q(\mathcal{X}_r^+) + \pi(1-P(\mathcal{X}_r^+)) \\
&\leq (1-\pi)\alpha + \pi(1-P(\mathcal{X}_r^+)) \\
&\leq (1-\pi)\alpha + \pi(1-\beta(\alpha, P, Q)).
\end{aligned}$$

Thus, we see that $\underline{\mathbb{L}}(\pi, P, Q)$ is the largest number $\underline{\mathbb{L}}$ such that $(1-\pi)\alpha + \pi(1-\beta(\alpha)) \geq \underline{\mathbb{L}}$ for all $\alpha \in [0,1]$ and hence one can set

$$\underline{\mathbb{L}}(\pi, P, Q) = \underline{\mathbb{L}} = \min_{\alpha\in[0,1]} ((1-\pi)\alpha + \pi(1-\beta(\alpha))$$

for each $\pi \in [0,1]$.

Conversely, we can express the Neyman-Pearson function β in terms of the Bayes risk. That is, for any $\alpha \in [0,1]$, $\beta(\alpha, P, Q)$ is the largest number β such that

$$\begin{aligned}
& \forall \pi \in [0,1] && (1-\pi)\alpha + \pi(1-\beta) \geq \underline{\mathbb{L}}(\pi) \\
\Leftrightarrow\ & \forall \pi \in [0,1] && (1-\pi)\alpha - \underline{\mathbb{L}}(\pi) \geq \pi(\beta - 1) \\
\Rightarrow\ & \forall \pi \in (0,1] && \frac{1}{\pi}((1-\pi)\alpha - \underline{\mathbb{L}}(\pi)) \geq \beta - 1 \\
\Leftrightarrow\ & \forall \pi \in (0,1] && \beta \leq \frac{1}{\pi}((1-\pi)\alpha + \pi - \underline{\mathbb{L}}(\pi)).
\end{aligned}$$

Thus we can set

$$\beta(\alpha) = \inf_{\pi\in(0,1]} \frac{1}{\pi}((1-\pi)\alpha + \pi - \underline{\mathbb{L}}(\pi)), \ \ \alpha \in [0,1].$$

■

A.8 Proof of Lemma 23

Proof Let $\mathcal{X}' = [0,1]$ and P be the uniform distribution on $\mathcal{X}'$. Overload P and Q to also denote the respective cumulative distribution functions (i.e., $P(x) = P([0,x])$). Thus $P(\pi) = \pi$). Set $Q(\pi) = \phi(\pi)$. Since $\phi(\cdot)$ is increasing it suffices to consider $r(\cdot)$ of the form $r_\pi(x) = [\![x < \pi]\!]$. Hence

$$\beta(\alpha) = \max\{\phi(\pi)\colon 0 \le \pi \le 1,\ \pi \le \alpha\},\ \alpha \in [0,1].$$

The maximum will always be obtained for $\pi = \alpha$ and thus $\beta(\alpha) = \phi(\alpha)$ for $\alpha \in [0,1]$. Finally, a pair of distributions on $\mathcal{X}$ can be constructed by embedding the connected component $\mathcal{C} \subset \mathcal{X}$ into $\mathcal{X}'$. Choose $g\colon \mathcal{C} \to \mathcal{X}'$ such that g is invertible. Such a g always exists since $\mathcal{C}$ is connected. Then g^{-1} induces distributions P' and Q' on $\mathcal{C}$ and thus on $\mathcal{X}$ by subsethood. ■

A.9 Proof of Corollary 24

Proof Choose a ψ satisfying the conditions and substitute into (43). This gives a corresponding $\phi(\cdot)$. We know from the preceding lemma that there exist P and Q such that $\beta(\cdot, P, Q) = \phi(\cdot)$ which corresponds to $\underline{\mathbb{L}}(\cdot, P, Q)$. Thus it remains to show that the function ϕ defined by

$$\phi(\alpha) = \inf_{\pi\in(0,1]} \frac{1}{\pi}((1-\pi)\alpha + \pi - \psi(\pi))$$

is concave and satisfies $\phi(1) = 1$. Observe that $\beta(1) = \inf_{\pi\in(0,1]} \frac{1-\psi(\pi)}{\pi}$. Now by the upper bound on ψ, we have $\frac{1-\psi(\pi)}{\pi} \ge \frac{1-1+\pi}{\pi} = \frac{1}{\pi} \ge 1$. But $\lim_{\pi\to 1} \frac{1-\psi(\pi)}{\pi} = 1$ and thus $\beta(1) = 1$. Finally note that

$$\beta(\alpha) = \inf_{\pi\in(0,1]} \left(\frac{1-\pi}{\pi}\right)\alpha + (1-\psi(\pi)).$$

This is the lower envelope of a parameterized (by π) family of affine functions (in α) and is thus concave. ■

A.10 Proof of Lemma 26

Proof From Theorem 14 we know that $\underline{L}_c(\eta) = \min\{(1-\eta)c, (1-c)\eta\}$ and note that $(1-\eta)c \le (1-c)\eta \iff c \le \eta$. Then, by the definition of L_c and the identity $1 - [\![p]\!] = [\![\neg p]\!]$ we have

$$\begin{aligned}
B_c(\eta,\hat\eta) &= (1-\eta)c[\![\hat\eta \ge c]\!] + (1-c)\eta[\![\hat\eta < c]\!] - \min\{(1-\eta)c, (1-c)\eta\}\\
&= (1-\eta)c[\![\hat\eta \ge c]\!] + (1-c)\eta[\![\hat\eta < c]\!] - (1-\eta)c[\![\eta \ge c]\!] - (1-c)\eta[\![\eta < c]\!]\\
&= (1-\eta)c([\![\hat\eta \ge c]\!] - [\![\eta \ge c]\!]) + (1-c)\eta([\![\hat\eta < c]\!] - [\![\eta < c]\!]).
\end{aligned}$$

Note that $[\![\hat\eta \ge c]\!] - [\![\eta \ge c]\!]$ is either 1 or -1 depending on whether $\hat\eta \ge c > \eta$ or $\hat\eta < c \le \eta$ and is zero otherwise. Similarly, $[\![\hat\eta < c]\!] - [\![\eta < c]\!]$ is 1 when $\hat\eta < c \le \eta$, is -1 when $\hat\eta \ge c > \eta$ and is zero otherwise. This means

$$\begin{aligned}
B_c(\eta,\hat{\eta}) &= \begin{cases}(1-\eta)c-(1-c)\eta, & \hat{\eta}\geq c>\eta\\ -(1-\eta)c+(1-c)\eta, & \eta\geq c>\hat{\eta}\end{cases}\\
&= \begin{cases}c-\eta, & \hat{\eta}\geq c>\eta\\ \eta-c, & \eta\geq c>\hat{\eta}\end{cases}\\
&= |\eta-c|[\![\min\{\eta,\hat{\eta}\}\leq c<\max\{\eta,\hat{\eta}\}]\!]
\end{aligned}$$

as required. ■

A.11 Proof of Theorem 33

Proof Let $\mathcal{C}:=(\operatorname{sgn}\mathcal{R}+1)/2\subseteq\{0,1\}^{\mathcal{X}}$ and so $\operatorname{sgn}\mathcal{R}=2\mathcal{C}-1$. Then

$$\begin{aligned}
\underline{\mathbb{L}}^{0-1}_{\mathcal{C}}(\pi,P,Q) &= \inf_{r\in\mathcal{C}}\mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathbb{P}}\ell^{0-1}(r(\mathsf{X}),\mathsf{Y})\\
&= \inf_{r\in\mathcal{C}}\left(\pi\mathbb{E}_{\mathsf{X}\sim P}\ell^{0-1}(r(\mathsf{X}),0)+(1-\pi)\mathbb{E}_{\mathsf{X}\sim Q}\ell^{0-1}(r(\mathsf{X}),1)\right)\\
&= \inf_{r\in\mathcal{C}}(\pi\mathbb{E}_{\mathsf{X}\sim P}[\![r(\mathsf{X})=1]\!]+(1-\pi)\mathbb{E}_{\mathsf{X}\sim Q}[\![r(\mathsf{X})=0]\!])\\
&= \inf_{r\in\mathcal{C}}(\pi\mathbb{E}_P r+(1-\pi)\mathbb{E}_Q(1-r))
\end{aligned}$$

since $\operatorname{Ran} r=\{0,1\}\Rightarrow\mathbb{E}_{\mathsf{X}\sim P}[\![r(\mathsf{X})=1]\!]=\mathbb{E}_{\mathsf{X}\sim P}r(\mathsf{X})$ and $\mathbb{E}_{\mathsf{X}\sim Q}[\![r(\mathsf{X})=0]\!]=\mathbb{E}_{\mathsf{X}\sim Q}(1-r(\mathsf{X}))$. Let $\rho=2r-1\in 2\mathcal{C}-1$. Thus $r=\frac{\rho+1}{2}$. Hence

$$\begin{aligned}
\underline{\mathbb{L}}^{0-1}_{\mathcal{C}}(\pi,P,Q) &= \inf_{\rho\in 2\mathcal{C}-1}\left(\pi\mathbb{E}_P\left(\frac{\rho+1}{2}\right)-(1-\pi)\mathbb{E}_Q\left(1-\frac{\rho+1}{2}\right)\right)\\
&= \frac{1}{2}\inf_{\rho\in 2\mathcal{C}-1}(\pi\mathbb{E}_P(\rho+1)+(1-\pi)\mathbb{E}_Q(1-\rho))\\
&= \frac{1}{2}\inf_{\rho\in 2\mathcal{C}-1}(\pi\mathbb{E}_P\rho+(1-\pi)\mathbb{E}_Q(-\rho)+\pi+(1-\pi))\\
&= \frac{1}{2}+\frac{1}{2}\inf_{\rho\in 2\mathcal{C}-1}(\pi\mathbb{E}_P\rho-(1-\pi)\mathbb{E}_Q\rho)\\
&= \frac{1}{2}-\frac{1}{2}\sup_{\rho\in 2\mathcal{C}-1}(\pi\mathbb{E}_P(-\rho)-(1-\pi)\mathbb{E}_Q(-\rho)).
\end{aligned}$$

Since $\mathcal{R}$ is symmetric about zero, $\operatorname{sgn}(\mathcal{R})=2\mathcal{C}-1$, $\mathcal{C}\subseteq\{0,1\}^{\mathcal{X}}$ is symmetric about $\frac{1}{2}$; that is, $\rho\in\mathcal{C}\Rightarrow(1-\rho)\in\mathcal{C}$. Thus

$$\begin{aligned}
\underline{\mathbb{L}}^{0-1}_{\mathcal{C}}(\pi,P,Q) &= \frac{1}{2}-\frac{1}{2}\sup_{\rho\in 2\mathcal{C}-1}(\pi\mathbb{E}_P\rho-(1-\pi)\mathbb{E}_Q\rho)\\
&= \frac{1}{2}-\frac{1}{4}V_{2\mathcal{C}-1,\pi}(P,Q)\\
&= \frac{1}{2}-\frac{1}{4}V_{\operatorname{sgn}\mathcal{R},\pi}(P,Q). \qquad (70)
\end{aligned}$$

Since by assumption $\operatorname{sgn}\mathcal{R} \subseteq \mathcal{R}$, the supremum in (63) will be ± 1-valued everywhere. Thus $V_{\operatorname{sgn}\mathcal{R},\pi}(P,Q) = V_{\mathcal{R},\pi}(P,Q)$. Combining this fact with (70) leads to (64).

Finally observe that by replacing inf and sup by arg min and arg max the final part of the theorem is apparent. ■

A.12 Pinsker Theorems

Proof (Theorem 30) Given a binary experiment (P,Q) denote the corresponding statistical information as

$$\phi(\pi) = \phi_{(P,Q)}(\pi) := \Delta\underline{\mathbb{L}}^{0-1}(\pi,P,Q) = \pi \wedge (1-\pi) - \psi_{(P,Q)}(\pi),$$

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

Since

$$\mathbb{I}_f(P,Q) = \int_0^1 \phi(\pi)\gamma_f(\pi)d\pi, \tag{71}$$

$\mathbb{I}_f(P,Q)$ is minimized by minimizing $\phi_{(P,Q)}$ over all (P,Q) such that

$$\phi(\pi_i) = \phi_i = \pi_i \wedge (1-\pi_i) - \psi_{(P,Q)}(\pi_i).$$

Let $\psi_i := \psi(\pi_i) = \frac{1}{2} - \frac{1}{4}V_{\pi_i}(P,Q)$. The problem becomes:

$$\begin{aligned} \text{Given } (\pi_i,\psi_i)_{i=1}^n \qquad & \text{find the maximal } \psi\colon [0,1] \to [0,\tfrac{1}{2}] \text{ such that} && (72)\\ & \psi(\pi_i) = \psi_i, \ \ i = 0,\ldots,n+1, && (73)\\ & \psi(\pi) \leq \pi \wedge (1-\pi), \ \ \pi \in [0,1], && (74)\\ & \psi \text{ is concave.} && (75) \end{aligned}$$

This will tell us the optimal ϕ to use since optimising over ψ is equivalent to optimizing over $\underline{\mathbb{L}}(\cdot,P,Q)$. Under the additional assumption on $\mathcal{X}$, Corollary 24 implies that for any ψ satisfying (73), (74) and (75) there exists P,Q such that $\underline{\mathbb{L}}(\cdot,P,Q) = \psi(\cdot)$.

Let Ψ be the set of piecewise linear concave functions on $[0,1]$ having $n+1$ segments such that $\psi \in \Psi \Rightarrow \psi$ satisfies (73) and (74). We now show that in order to solve (72) it suffices to consider $\psi \in \Psi$.

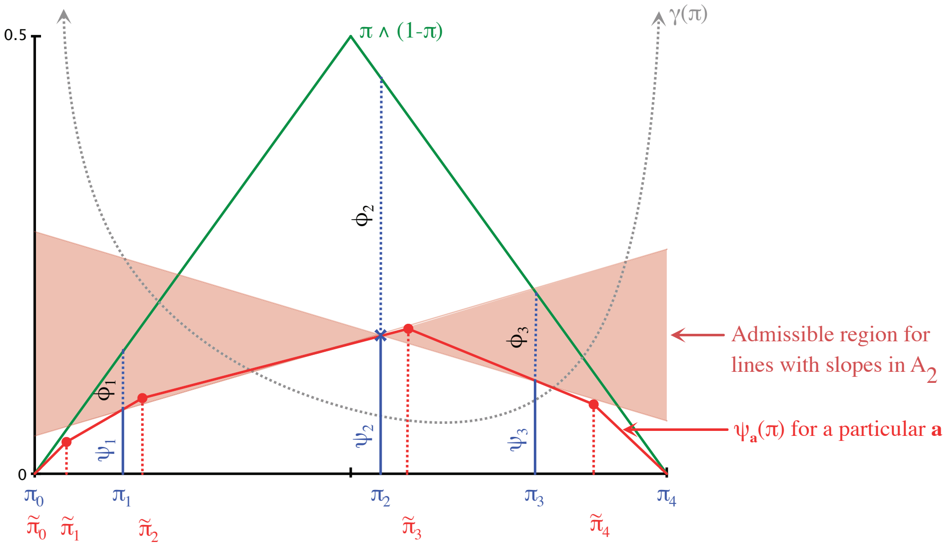
If g is a concave function on $\mathbb{R}$, then

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

denote the *sup-differential* of g at x. (This is the obvious analogue of the *sub*-differential for convex functions Rockafellar, 1970.) Suppose $\tilde{\psi}$ is a general concave function satisfying (73) and (74). For $i = 1,\ldots,n$, let

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

Observe that by concavity, for all concave $\tilde{\psi}$ satisfying (73) and (74), for all $g \in \bigcup_{i=1}^n G_i^{\tilde{\psi}}$, $g(\pi) \geq \psi(\pi)$, $\pi \in [0,1]$.



Admissible region for lines with slopes in A_2

$\psi_{\mathbf{a}}(\pi)$ for a particular **a**

Figure 8: Illustration of construction of optimal $\psi(\pi) = \underline{\mathbb{L}}(\pi, P, Q)$. The optimal ψ is piecewise linear such that $\psi(\pi_i) = \psi_i$, $i = 0, \ldots, n+1$.

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

$$\psi^*(\pi) = \min(g_1^{\tilde{\psi}}(\pi), \ldots, g_n^{\tilde{\psi}}(\pi)) \tag{76}$$

such that ψ^* is concave, satisfies (73) and $\psi^*(\pi) \geq \tilde{\psi}(\pi)$, for all $\pi \in [0,1]$. It remains to take account of (74). That is trivially done by setting

$$\psi(\pi) = \min(\psi^*(\pi), \pi \wedge (1-\pi)) \tag{77}$$

which remains concave and piecewise linear (although with potentially one additional linear segment). Finally, the pointwise smallest concave ψ satisfying (73) and (74) is the piecewise linear function connecting the points $(0,0), (\pi_1, \psi_1), (\pi_2, \psi_2), \ldots, (\pi_m, \psi_m), (1,0)$.

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

$$g(\pi) = \left(\psi_i + \frac{(\psi_{i+1} - \psi)(\pi - \pi_i)}{\pi_{i+1} - \pi_i}\right) \cdot [\![\pi \in [\pi_i, \pi_{i+1}]]\!], \quad i = 0, \ldots, n,$$

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

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

$$b_i = \psi_i - a_i\pi_i, \quad i = 0, \ldots, n+1.$$

In order to determine the constraints on a_i, since g is concave and minorizes ψ, 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 $i = 1, \dots, n$)

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

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

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

We now determine the points at which ψ defined by (76) and (77) change slope. That occurs at the points π 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 $i = 0, \dots, n$. Thus

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

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

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

where $\mathbf{a} \in A_n$ (see (50)), $\bar{\pi}_i$, $\alpha_{\mathbf{a},i}$ and $\beta_{\mathbf{a},i}$ are defined by (53), (54) and (55) respectively. The extra segment induced at index j (see (52)) 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 piecewise linear with $n+2$ segments (recall i ranges from 0 to $n+2$); if $\tilde{\pi}_{k+1} = \frac{1}{2}$ for some $k \in \{1, \dots, n\}$, then there will be only $n+1$ non-trivial segments.

Thus

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

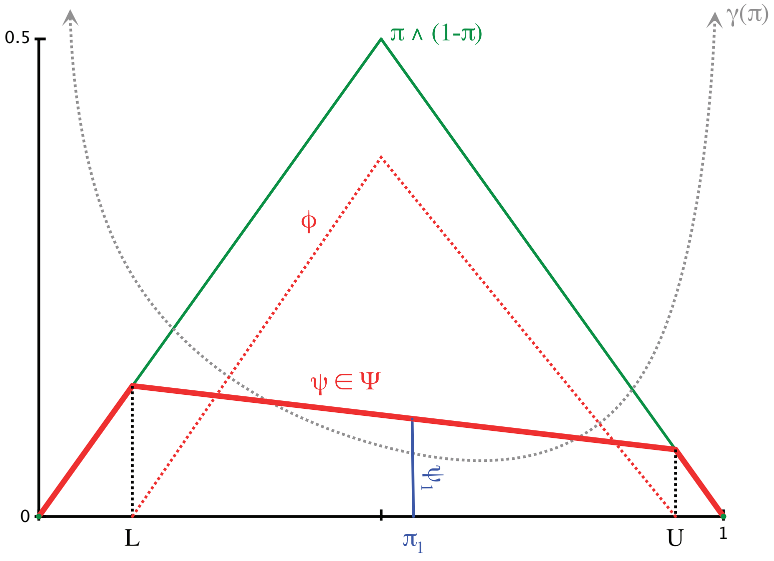


Figure 9: The optimisation problem when $n = 1$. Given ψ_1, there are many risk curves consistent with it. The optimisation problem involves finding the piecewise linear concave risk curve $\psi \in \Psi$ and the corresponding $\phi = \pi \wedge (1-\pi)$ that maximises $\mathbb{I}_f$. L and U are defined in the text.

is the set of ϕ consistent with the constraints and A_n is defined in (50). Thus substituting into (71), interchanging the order of summation and integration and optimizing we have shown (56). The tightness has already been argued: under the additional assumption on $\mathcal{X}$, since there is no slop in the argument above since every ϕ satisfying the constraints is the Bayes risk function for some (P,Q). ■

Proof (Theorem 31) In this case $n = 1$ and the optimal ψ function will be piecewise linear, concave, and its graph will pass through (π_1, ψ_1). Thus the optimal ϕ 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 9). For variational divergence, $\pi_1 = \frac{1}{2}$ and thus

$$\psi_1 = \pi_1 \wedge (1-\pi_1) - \frac{V}{4} = \frac{1}{2} - \frac{V}{4} \tag{78}$$

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

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

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

$$\mathbb{I}_f(P,Q) \geq \min_{a\in[-2\psi_1, 2\psi_1]} \int_{\frac{a\pi_1-\psi_1}{a-1}}^{\frac{1}{2}} [(1-a)\pi - \psi_1 + a\pi_1]\gamma_f(\pi)d\pi + \int_{\frac{1}{2}}^{\frac{1-\psi_1+a\pi_1}{a+1}} [(-a-1)\pi - \psi_1 + a\pi_1 + 1]\gamma_f(\pi)d\pi. \tag{79}$$

If γ_f is symmetric about $\pi = \frac{1}{2}$ (so by Corollary 13 $\mathbb{I}_f$ is symmetric) and convex and $\pi_1 = \frac{1}{2}$, then the optimal $a = 0$. Thus in that case,

$$\begin{aligned} \mathbb{I}_f(P,Q) & \geq 2\int_{\psi_1}^{\frac{1}{2}} (\pi - \psi_1)\gamma_f(\pi)d\pi \\ & = 2\left[(\tfrac{1}{2} - \psi_1)\Gamma_f(\tfrac{1}{2}) + \overline{\Gamma}_f(\psi_1) - \overline{\Gamma}_f(\tfrac{1}{2})\right] \\ & = 2\left[\tfrac{V}{4}\Gamma_f(\tfrac{1}{2}) + \overline{\Gamma}_f\left(\tfrac{1}{2} - \tfrac{V}{4}\right) - \overline{\Gamma}_f(\tfrac{1}{2})\right]. \end{aligned} \tag{80}$$

Appendix B. Examples of Generalised Pinsker Inequality

Combining the above with (78) leads to a range of Pinsker style bounds for symmetric $\mathbb{I}_f$:

Jeffrey's Divergence $J(P,Q) = \mathrm{KL}(P,Q) + \mathrm{KL}(Q,P)$. Thus $\gamma(\pi) = \frac{1}{\pi^2(1-\pi)} + \frac{1}{\pi(1-\pi)^2} = \frac{1}{\pi^2(1-\pi)^2}$. (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 the weight function $\gamma_f(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right) = \frac{1}{2\pi(1-\pi)}$. Thus

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

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

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

Hellinger Divergence Here $f(t) = (\sqrt{t}-1)^2$. Consequently the weight function

$$\gamma_f(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right) = \frac{1}{\pi^3}\frac{1}{2\left((1-\pi)/\pi\right)^{3/2}} = \frac{1}{2[\pi(1-\pi)]^{3/2}}$$

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 Here $f(t) = \frac{t+1}{2}\ln\left(\frac{t+1}{2\sqrt{t}}\right)$. 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 χ^2-Divergence Here $\Psi(P,Q) = \chi^2(P,Q) + \chi^2(Q,P)$ 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 γ_f is not symmetric, one needs to use (79) instead of the simpler (80). We consider two special cases.

χ^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 (79) 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

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

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 ∞ as $V \to 2$.

Kullback-Leibler Divergence In this case $f(t) = t\ln t$ and thus $f''(t) = 1/t$ and the weight function $\gamma_f(\pi) = \frac{1}{\pi^3} f''\left(\frac{1-\pi}{\pi}\right) = \frac{1}{\pi^2(1-\pi)}$ which is clearly not symmetric. From (79) we obtain

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

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

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

Set $\beta := 2a$ and we have (59).

■

Appendix C. Background and Prior Work

Specific prior results are referred to in the body of the paper. We now briefly indicate the broad sweep of prior work along the lines of the present paper.

The most important precursors and inspiration are the three nearly simultaneous[25] works by Buja et al. (2005), Liese and Vajda (2006) and Nguyen et al. (2005). The work by Dawid (2007) is also very similar in spirit to that presented here. A crucial difference is that he relies on a parametric viewpoint, and can use the machinery of Riemannian geometry. Zhang (2004a); Zhang and Matsuzoe (2009) have developed a number of connections between convex functions, the Bregman divergences they induce, and Riemannian geometry. All of the results in the present paper are, in contrast, "coordinate-free." The *motivation* of the present work is closely aligned with that of Hand (1994) whose avowed aim was to "stimulate debate about the need to formulate research questions sufficiently precisely that they may be unambiguously and correctly matched with statistical techniques." Hand and Vinciotti (2003) develop some refined machine learning tasks that can be viewed as weighted problems (in the sense of the weight functions we make extensive use of in this paper); confer Buja et al. (2005).

The paper presents a unification of sorts. This, in itself, is hardly new in machine learning. There are different approaches to unification. One distinction is between *Monistic* and *Pluralistic* approaches (James, 1909; Turkle and Papert, 1992); this corresponds to the hedgehog/fox distinction of Berlin (1953).

Monistic approaches aim for a single all encompassing theory.[26] A problem with most monistic approaches is that you have to accept it "all or nothing." There are many unifying approaches developed in Statistics and Machine learning that have left little trace; For example, Nelson's use of non-standard analysis (Nelson, 1987; Lutz and Musio, 2005) as the foundations for probability; Topsøe's (2006), Shafer and Vovk's (2001) game theory as a basis, and Le Cam's use of Riesz measures on a vector lattice to replace the traditional sample space (LeCam, 1964).

Pluralistic approaches are closer to what is proposed here (where, instead of searching for a single master representation, we study relationships and translations between a range of different representations). It resonates with Kiefer's assertion that "Statistics is too complex to be codified in terms of a simple prescription that is a panacea for all settings, and . . . one must look as carefully as possible at a variety of possible procedures. . . " (Kiefer, 1977). Examples of existing pluralistic attempts include limited problem catalogs such as for different notions of *cost* (Turney, 2000) or a restricted set of problems (Raudys, 2001).

The decision theoretic approach (DeGroot, 1970; Berger, 1985; Kiefer, 1987) due to Wald (1950, 1949) is central to the present paper. The idea of seeking *primitives* for statistics dates back at least to the elementary experiments of Birnbaum (1961). The relationship between risks and Bregman divergences is studied by Grünwald and Dawid (2004) and Buja et al. (2005).

[25] Nguyen et al. (2005) is dated 13 October, 2005, Liese and Vajda (2006) was received on 26 October 2005 and Buja et al. (2005) is dated 3 November 2005. Shen's PhD thesis (Shen, 2005), which contains most of the material in Buja et al. (2005), is dated 16 October 2005. The paper by Nguyen et al. (2005) has now appeared as Nguyen et al. (2009).

[26] Monistic approaches can be categorised into at least four distinct categories. They are briefly summarised in Appendix C.1.

There are numerous possible definitions of information. Many of them are sterile; Csiszár (1978) and Aczél (1984) provide a critical analysis. Floridi (2004) discusses pluralistic versus monistic approach: is there one single definition of information, or should there be many different definitions depending on the particular problem? Our view, like Shannon (1948), is that there are many types. Shannon information was developed with communications problems in mind—there is no reason why it is the only notion of information that makes sense for learning and inference.

There are many known relationships between risks and divergences between distributions many of which we explicitly discuss later in the paper. General results include those due to Österreicher (2003), Österreicher and Vajda (1993), Gutenbrunner (1990), Liese and Vajda (2006), Goel and De-Groot (1979) and Golic (1987). Particular relations between risk in binary classification problems and f-divergences are not new (Poor and Thomas, 1977; Kailath, 1967). Some more general results that relate the choice of loss function in a binary learning problem to particular f-divergences between the class-conditional distributions have been (re)-discovered (Eguchi and Copas, 2001; Nguyen et al., 2005; Österreicher and Vajda, 1993). Known results relating different distances between probability distributions are summarised by Gibbs and Su (2002).

The idea of solving a machine learning problem by using a solution to some other learning problem is now called a *machine learning reduction* (Beygelzimer et al., 2008, 2005) The idea is not new. Equivalences are a natural structuring device and were explicit in Ashby's foundational work on cybernetics (Ashby, 1956), a precursor to Machine Learning. Ben-Bassat (1978) studied the concept of ε-equivalence, Conover and Iman (1981) showed how rank tests can be derived by applying nonparametric tests to order statistics, and Goldman et al. (1989) and Bartlett et al. (1996) used reductions for theoretical purposes. However recently there has been a large number of explicit constructions of reductions (Zadrozny et al., 2003; Langford, 2006; Beygelzimer et al., 2005; Langford and Beygelzimer, 2005; Langford and Zadrozny, 2005; Langford et al., 2006; Li and Lin, 2007; Beygelzimer et al., 2007; Langford, 2007; Scott and Davenport, 2007),or development of results which although not explicitly called reductions are effectively so (Brown et al., 2002; Brown and Low, 1996; Brown and Zhao, 2003; Chaudhuri and Loh, 2002; Cossock and Zhang, 2006; Cuevas and Fraiman, 1997; Domingos, 1999; Steinwart et al., 2005; Tasche, 2001). Two key differences between the recent machine learning reductions literature and the present paper is that our relationships between problems are (usually) exact (instead of approximate) and we work with the true underlying distributions (rather than finite sample distributions).

The theory of *Comparison of Experiments*, developed by Blackwell (1951, 1953), and significantly extended by LeCam (1964, 1986) is also related to the overall goal set out here. It has been used to define notions of isomorphism for statistical problems (Morse and Sacksteder, 1966; Sacksteder, 1967) and is the subject of three books (Strasser, 1985; Torgersen, 1991; Heyer, 1982) and a recent review (Goel and Ginebra, 2003). The key difference with the present work is that the comparison of experiments theory seeks results that hold for *all* loss functions rather than for a particular one; with a few exceptions (Torgersen, 1991, Chapter 10). Blackwell related comparisons to sufficient statistics and characterised comparisons. LeCam (1964) quantified comparisons in terms of the degree to which one experiment is "better than" another (the deficiency distance). There are very few known examples of deficiency distance (Carter, 2002). Furthermore LeCam's theory is formulated in a particularly abstract way to make its theorems elegant (Yang and Le Cam, 1999). Renowned probabilists concur that its arcane formulation has made it inaccessible (van der Vaart, 2002; Pollard, 2000; Strasser, 2000). Consequently the subject has had relatively limited impact.

Graphical representations have been used for a long while to better understand binary experiments. In the main body of the paper we develop connections between Receiver Operating Characteristic (ROC) curves, (Fawcett, 2006, 2004; Flach, 2003; Flach and Wu, 2005; Maxion and Roberts, 2004) the Area Under ROC Curve (AUC), (Cortes and Mohri, 2004; Hand, 2008; Hand and Till, 2001; Hanley and McNeil, 1982) and Cost Curves (Drummond and Holte, 2006; Torgersen, 1991). These can be seen as *representations* of Binary Experiments.

C.1 Summary of Previous "Monistic" Approaches to Unification

There are are range of different approaches to unifying machine learning from a monistic perspective:

Low level data interchange: There is a small amount of work on developing standards for interchanging data sets (Grossman et al., 2002; Carey et al., 2007; Wettschereck and Muller, 2001)—this is analogous to PDDL (Ghallab et al., 1998). There are also some limited higher level attempts such as ontologies (Soldatova and King, 2006) and general frameworks (Fayyad et al., 1996).

Modelling frameworks: To *solve* a machine learning problem, one needs models. There is a rich literature on graphical models (Jordan, 1999), factor graphs (Kschischang et al., 2001) and Markov logic networks (Domingos and Richardson, 2004; Richardson and Domingos, 2006) which have allowed the unification of sets of problems (Worthen and Stark, 2001), with a focus on the modelling and computational techniques for particular problems.

Comparison of frameworks: There are several philosophical frameworks/approaches to designing inference and learning algorithms. Barnett (1999), Bayarri and Berger (2004) and Berger (2003) compare and contrast these. They are effectively comparing different monistic frameworks, not comparing problems.

Overarching frameworks: These include frameworks such as Bayesian (Robert, 1994), information theoretic (Jenssen, 2005b; Harremoës, 1993), game-theoretic (Vovk et al., 2005; Grünwald and Dawid, 2004), MDL (Grünwald, 2007; Rissanen, 2007), regularised distance minimisation (Borwein and Lewis, 1991; Altun and Smola, 2006; Broniatowski, 2004), and more narrowly focussed "unifying frameworks" such as information geometry (Dawid, 2007; Eguchi, 2005), exponential families (Canu and Smola, 2006) and the information bottleneck (Tishby et al., 2000).

Appendix D. Examples and Prior Work on Surrogate Regret Bounds

Surrogate regret bounds have garnered interest in the machine learning community (Zhang, 2004b; Bartlett et al., 2006; Steinwart, 2007; Steinwart and Christmann, 2008). Steinwart and Christmann (2008, Chapter 3) have presented a good summary of recent work.

All of the recent work has been in terms of *margin losses* of the form

$$L^{\phi}(\eta,\hat{h}) = \eta\phi(\hat{h}) + (1-\eta)\phi(-\hat{h}).$$

As Buja et al. (2005) discuss, such margin losses can not capture the richness of all possible proper losses. Bartlett et al. (2006) prove that for any $\hat{h}$

$$\psi\left(L^{0-1}(\eta,\hat{h}) - \underline{L}^{0-1}(\eta)\right) \leq L^{\phi}(\eta,\hat{h}) - \underline{L}^{\phi}(\eta),$$

where $\psi = \tilde{\psi}^{\star\star}$ is the LF biconjugate of $\tilde{\psi}$,

$$\tilde{\psi}(\theta) = H^{-}\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right),$$

$H(\eta) = \underline{L}^{\phi}(\eta)$ and

$$H^{-}(\eta) = \inf_{\alpha\colon\, \alpha(2\eta-1)\leq 0} (\eta\phi(\alpha) + (1-\eta)\phi(-\alpha))$$

is the optimal conditional risk under the constraint that the sign of the argument α disagrees with $2\eta - 1$.

We will consider two examples presented by Bartlett et al. (2006) and show that the bounds we obtain with the above theorem match the results we obtain with Theorem 25.

Exponential Loss Consider the link $\hat{h} = \psi(\hat{\eta}) = \frac{1}{2}\ln\frac{\hat{\eta}}{1-\hat{\eta}}$ with corresponding inverse link $\hat{\eta} = \frac{1}{1+e^{-2\hat{h}}}$. Buja et al. (2005) showed that this link function combined with exponential margin loss $\phi(\gamma) = e^{-\gamma}$ results in a proper scoring rule

$$L(\eta, \hat{\eta}) = \eta\left(\frac{1-\hat{\eta}}{\hat{\eta}}\right)^{\frac{1}{2}} + (1-\eta)\left(\frac{\hat{\eta}}{1-\hat{\eta}}\right)^{\frac{1}{2}}.$$

From (32) we obtain

$$w(\eta) = \frac{1}{2[\eta(1-\eta)]^{\frac{3}{2}}}.$$

(Note Buja et al., 2005 have missed the factor of $\frac{1}{2}$.) Thus $W(\eta) = \frac{2\eta-1}{\sqrt{\eta(1-\eta)}}$ and $\overline{W}(\eta) = -2\sqrt{\eta(1-\eta)}$. Hence we obtain

$$\underline{L}(\eta) = 2\sqrt{\eta(1-\eta)} \tag{81}$$

and from (46) we obtain that if $B_{\frac{1}{2}}(\eta, \hat{\eta}) = \alpha$ then

$$B(\eta, \hat{\eta}) \geq 1 - \sqrt{1-4\alpha^2}. \tag{82}$$

Equations 81 and 82 match the results presented by Bartlett et al. (2006) upon noting that $B_{\frac{1}{2}}(\eta, \hat{\eta})$ measures the loss in terms of $\ell_{\frac{1}{2}}$ and Bartlett et al. (2006) used $\ell^{0-1} = 2\ell_{\frac{1}{2}}$.

Truncated Quadratic Loss Consider the margin loss $\phi(\hat{h}) = (1 + \hat{h} \vee 0)^2 = (2\hat{\eta} \vee 0)^2$ with link function $\hat{h}(\hat{\eta}) = 2\hat{\eta} - 1$. From (32) we obtain $\underline{L}(\eta) = 4\eta(1-\eta)$ and from (46) the regret bound $B(\eta, \hat{\eta}) \geq 4\alpha^2$. These match the results presented by Bartlett et al. (2006) when again it is noted we used $\ell_{\frac{1}{2}}$ and they used ℓ^{0-1}.

The above results are for $c_0 = \frac{1}{2}$. Generalisations of margin losses to the case of uneven weights are presented by Steinwart and Christmann (2008, Section 3.5). Nevertheless, since the same ϕ function is still used for both components of the loss (albeit with unequal weights) such a scheme can still not capture the full generality of all proper scoring rules in the manner achieved by the results in Section 7.1.

Appendix E. History of Pinsker Inequalities

Pinsker (1964) 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 Csiszár (1967) presented another version and Kullback (1967) showed $\mathrm{KL} \geq V^2/2 + V^4/36$. Much later Topsøe (2001) showed $\mathrm{KL} \geq V^2/2 + V^4/36 + V^6/270$. Non-polynomial bounds are due to Vajda (1970): $\mathrm{KL} \geq L_{\mathrm{Vajda}}(V) := \ln\left(\frac{2+V}{2-V}\right) - \frac{2V}{2+V}$ and Toussaint (1978) 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 (2002) use 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 and Chhina (2005) have 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. (2001) and Taneja (2005a,b). Topsøe (2000) has presented some infinite series representations for capacitory discrimination in terms of triangular discrimination which lead to inequalities between those two divergences. Liese and Miescke (2008, p.48) give the inequality $V \leq h\sqrt{4-h^2}$ (which seems to be originally due to LeCam, 1986) which when rearranged corresponds exactly to the bound for h^2 in theorem 31. Withers (1999) has also presented some inequalities between other (particular) pairs of divergences; his reasoning is also in terms of infinite series expansions.

Unterreiter et al. (2000) 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 30. They do not compute any of the explicit bounds in theorem 31 except they state (page 243) $\chi^2(P,Q) \geq V^2$ which is looser than (58).

Gilardoni (2006a) 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 (2006b,c) presented a completely different approach which obtains many of the results of theorem 31.[27] Gilardoni (2006c) improved Vajda's bound slightly to $\mathrm{KL}(P,Q) \geq \ln\frac{2}{2-V} - \frac{2-V}{2}\ln\frac{2+V}{2}$.

Gilardoni (2006b,c) presented a general tight lower bound for $\mathbb{I}_f(P,Q)$ in terms of $V(P,Q)$ 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 $k^{-1}(t) = \frac{1}{2}\left(\frac{1}{1-g_L^{-1}(t)} + \frac{1}{g_R^{-1}(t)-1}\right)$, $g(u) = (u-1)f'(u) - f(u)$, $g_R^{-1}[g(u)] = u$ 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. (2003) 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 (59). 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, Δ and I identical to the results in Theorem 31. It is not apparent to us how the approach of Gilardoni (2006b,c) could be extended to more general situations such as that in Theorem 30 (i.e., $n > 1$).

Finally Bolley and Villani (2005) have considered *weighted* versions of the Pinsker inequalities (bounds for a weighted generalisation of Variational divergence) in terms of KL-divergence that are related to transportation inequalities.

[27] We were unaware of these two papers until completing the results presented in the main paper.

Appendix F. Variational Representation of $\mathbb{I}_f$ and its Generalizations

The variational representation of the Variational divergence (62) suggests the question of whether there is a variational representation for a general f-divergence. This has been considered previously. We briefly summarise the approach, and then explore some (new) implications of the representation.

One can obtain a variational representation for $\mathbb{I}_f$ by substituting a variational representation for f into the definition of $\mathbb{I}_f$ (Keziou, 2003a,b; Broniatowski, 2004; Broniatowski and Keziou, 2009). Let p and q denote the densities corresponding to P and Q and assume for now they exist. Recall from Section 2.2 above, that the Legendre-Fenchel conjugate of f is given by $f^\star(s) = \sup_{u\in\mathrm{Dom} f} us - f(u)$. In general $\mathrm{Ran}\, f^\star = \mathbb{R}^\star := \mathbb{R}\cup\{+\infty\}$. Since $f(u) = \sup_{\rho\in\mathbb{R}} u\rho - f^\star(\rho)$, we can write

$$\begin{aligned}
\mathbb{I}_f(P,Q) &= \int_{\mathcal{X}} q(x) \sup_{\rho\in\mathbb{R}}\left(\rho\frac{p(x)}{q(x)} - f^\star(\rho)\right)dx \\
&= \sup_{\rho\in\mathbb{R}^{\mathcal{X}}} \int_{\mathcal{X}} \rho(x)p(x) - f^\star(\rho(x))q(x)dx. \\
&= \sup_{\rho\in\mathbb{R}^{\mathcal{X}}} (\mathbb{E}_P\rho - \mathbb{E}_Q f^\star(\rho)). \qquad (83)
\end{aligned}$$

We make this concrete by considering the variational divergence. The corresponding f is given by $f(t) = |t-1|$ and (adopting the convention that $[\![\text{false}]\!]$ is a "very strong zero" so $[\![\text{false}]\!]\cdot\infty = 0$; confer Knuth, 1992)

$$f^\star(x) = [\![x\notin[-1,1]]\!]\infty + [\![x\in[-1,1]]\!]x.$$

Since the supremum in (83) will not be attained if the second term is infinite, one can restrict the supremum to be over $\mathcal{F} = \{\rho\in\mathbb{R}^{\mathcal{X}} \colon \|\rho\|_\infty \le 1\}$. Thus

$$\begin{aligned}
V(P,Q) &= \sup_{\rho\colon \|\rho\|_\infty\le 1} (\mathbb{E}_P\rho - \mathbb{E}_Q\rho) = \sup_{\rho\in\{-1,1\}^{\mathcal{X}}} (\mathbb{E}_P\rho - \mathbb{E}_Q\rho) \\
&= \sup_{\rho\in\{0,2\}^{\mathcal{X}}} (\mathbb{E}_P\rho - \mathbb{E}_Q\rho) = 2 \sup_{\rho\in\{0,1\}^{\mathcal{X}}} (\mathbb{E}_P\rho - \mathbb{E}_Q\rho) \\
&= 2\sup_A |P(A) - Q(A)|,
\end{aligned}$$

since the supremum will be attained for functions ρ taking on values only in $\{-1,1\}$ and the remaining steps are simply a shift and rescaling (to $\{0,2\}$ by adding 1, and then to $\{0,1\}$).

The representation (83) suggests the generalisation

$$\begin{aligned}
\mathbb{I}_{f,\mathcal{F}}(P,Q) &:= \sup_{\rho\in\mathcal{F}\subseteq\mathbb{R}^{\mathcal{X}}} \int_{\mathcal{X}} \rho(x)p(x) - f^\star(\rho(x))q(x)dx \\
&= \sup_{\rho\in\mathcal{F}} (\mathbb{E}_P\rho - \mathbb{E}_Q f^\star(\rho)).
\end{aligned}$$

Observing this is not symmetric in p and q suggests a further generalisation:

$$\begin{aligned}
\mathbb{I}_{f,g,\mathcal{F}}(P,Q) &:= \sup_{\rho\in\mathcal{F}\subseteq\mathbb{R}^{\mathcal{X}}} \int_{\mathcal{X}} -g^\star(\rho(x))p(x) - f^\star(\rho(x))q(x)dx \\
&= \sup_{\rho\in\mathcal{F}} (-\mathbb{E}_P g^\star(\rho) - \mathbb{E}_Q f^\star(\rho)).
\end{aligned}$$

Here $g^\star$ is the $\mathbb{R}^\star$-valued LF conjugate of a convex function g. Set $\mathbb{I}_{f,g} := \mathbb{I}_{f,g,\mathbb{R}^{\mathcal{X}}}$.

An alternative generalisation of $\mathbb{I}_f$ is

$$\tilde{\mathbb{I}}_{f,g,\mathcal{F}}(P,Q) := \sup_{\rho\in\mathcal{F}} \left(\mathbb{E}_P g^\star(\rho) - \mathbb{E}_Q f^\star(\rho)\right)$$

which is identical to (84) except for removal of the minus sign preceding $g^\star$. Set $\tilde{\mathbb{I}}_{f,g} := \tilde{\mathbb{I}}_{f,g,\mathbb{R}^{\mathcal{X}}}$. If $\rho \in \mathcal{F}$ are such that $\|\rho\|_\infty$ is unbounded, then in general $\tilde{I}_{f,g,\mathcal{F}}(P,Q)$ will be infinite. Properties of the alternative definition relate to the extended infimal convolution between two convex functions.

Definition 35 *Suppose $f,g\colon \mathbb{R}^+ \to \mathbb{R}^*$ are convex. The extended infimal convolution is*

$$(f\Box g)(\tau) := \inf_{x\in\mathbb{R}^+} f(x) + \tau g(x/\tau), \;\; \tau\in\mathbb{R}^+.$$

Note that the second term in this convolution is the perspective function (Section 2.1) applied to g, that is, $I_g(x,\tau)$.

Theorem 36 *Suppose $f,g\colon \mathbb{R}^+ \to \mathbb{R}^*$ are convex. Then*

1. $\mathbb{I}_f(P,Q) = \mathbb{I}_{f,\mathbb{R}^{\mathcal{X}}}(P,Q)$, $\tilde{\mathbb{I}}_{f,\mathrm{id},\mathcal{F}}(P,Q) = \mathbb{I}_{f,\mathcal{F}}(P,Q)$, *and*
$$\mathbb{I}_{t\mapsto|t-1|,\mathcal{F}}(P,Q) = 2V_{\mathcal{F},\frac{1}{2}}(P,Q).$$
2. $\tilde{\mathbb{I}}_{f_1,g_1,\mathcal{F}} = \mathbb{I}_{f_2,g_2,\mathcal{F}}$ *only if* $f_1 - f_2 = f_a$ *and* $g_1 - g_2 = g_a$ *and* $f_1, f_2, f_a, g_1, g_2, g_a$ *are affine.*
3. $\mathbb{I}_{f,f,\mathcal{F}} = \mathbb{I}_{\mathrm{id},\mathrm{id},f^\star(\mathcal{F})}(P,Q)$.
4. $\tilde{\mathbb{I}}_{f,f,\mathcal{F}} = \tilde{\mathbb{I}}_{\mathrm{id},\mathrm{id},f^\star(\mathcal{F})}(P,Q) = 2V_{f^\star(\mathcal{F})}(P,Q)$.
5. $\mathbb{I}_{f,g} = \mathbb{I}_{f\Box g}$.

Proof Part 1 follows immediately from the various definitions. Since affine functions are the only functions that are simultaneously convex and concave, $\tilde{\mathbb{I}}_{f_1,g_1,\mathcal{F}} = \mathbb{I}_{f_2,g_2,\mathcal{F}}$ only if f_1, f_2 (resp. g_1, g_2) are affine and their differences are affine (since an affine offset will not change $\tilde{\mathbb{I}}$). This proves part 2.

We have by change of variables

$$\tilde{\mathbb{I}}_{f,f,\mathcal{F}}(P,Q) = \sup_{\rho\in\mathcal{F}}(\mathbb{E}_P f^\star(\rho) - \mathbb{E}_Q f^\star(\rho)) = \sup_{\psi\in f^\star(\mathcal{F})} (\mathbb{E}_P\psi - \mathbb{E}_Q\psi) = \tilde{\mathbb{I}}_{\mathrm{id},\mathrm{id},f^\star(\mathcal{F})}(P,Q),$$

where $f^\star(\mathcal{F}) := \{f^\star\circ\rho\colon \rho\in\mathcal{F}\}$. (The same argument applies to $\mathbb{I}_{f,f,\mathcal{F}}$ although $\sup_{\psi\in g^\star(\mathcal{F})}(-\mathbb{E}_P\psi - \mathbb{E}_Q\psi)$ does not correspond to a generalised variational divergence.) This proves parts 3 and 4. ■

In order to prove 5 we need the following lemma.

Lemma 37 *Let $f\colon \mathbb{R}\to\mathbb{R}$ and $K\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be convex and bounded from below. Then the extended infimal convolution*

$$(f\Box K)(x) = \inf_{y\in\mathbb{R}} f(y) + K(x,y), \quad x\in\mathbb{R}$$

is convex in $x\in\mathbb{R}$.

Observe that if $K(x,y) = g(x-y)$ for convex g, then $f\Box K = f \oplus g$, the standard infimal convolution (Hiriart-Urruty and Lemaréchal, 1993b). This extended infimal convolution seems little studied with the exception owith the exception of Cepedello-Boiso (1998).

Proof Let $\tilde{f}(x,y) := f(y)$, $x \in \mathbb{R}$. Clearly $\tilde{f}$ is convex on $\mathbb{R}\times\mathbb{R}$. Let $\tilde{h}(x,y) = \tilde{f}(x,y) + K(x,y)$. Hiriart-Urruty and Lemaréchal (1993b, Proposition 2.1.1) show that $\tilde{h}$ is convex on $\mathbb{R}\times\mathbb{R}$. Observe that $(f\Box K)(x) = \inf\{\tilde{h}(x,y)\colon y \in \mathbb{R}\}$, that is, the *marginal* function of $\tilde{h}$. Since by construction $\tilde{h}$ is bounded from below, using the result of Hiriart-Urruty and Lemaréchal (1993b, p.169) proves the result. ■

Corollary 38 *For any convex f and g, $f\Box g$ is convex.*

Proof Observe that $(f\Box g)(x) = \inf_{y\in\mathbb{R}^+} f(y) + xg(y/x) = \inf_{y\in\mathbb{R}^+} f(y) + I_g(x,y)$, $x \in \mathbb{R}^+$, where I_g is the perspective function (1). Hiriart-Urruty and Lemaréchal (1993b, Proposition 2.2.1) show that if $g\colon \mathbb{R}^n \to \mathbb{R}$ is convex then the perspective I_g is convex on $\mathbb{R}^{n+1}$. The corollary then follows from the lemma. ■

Proof (part 5 of Theorem 36) Observe that if $h(x) = t\phi(x)$ then the LF conjugate $h^*(s) = t\phi(s/t)$. Thus using the Fenchel duality theorem (Rockafellar, 1970) we have, using (Rockafellar and Wets, 2004, Theorem 14.60) to justify the swapping the order of the supremum and integration,

$$
\begin{aligned}
\mathbb{I}_{f,g}(P,Q) &= \sup_{\rho\in\bar{\mathbb{R}}^{\mathcal{X}}} \int_{\mathcal{X}} -g^\star(\rho(x))p(x) - f^\star(\rho(x))q(x)dx \\
&= \int_{\mathcal{X}} \sup_{\rho\in\bar{\mathbb{R}}} -g^\star(\rho)p(x) - f^\star(\rho)q(x)dx \\
&= \int_{\mathcal{X}} \inf_{\rho\in\bar{\mathbb{R}}} f\left(\frac{\rho}{q(x)}\right) + g\left(\frac{\rho}{p(x)}\right)dx \\
&= \int_{\mathcal{X}} \inf_{\rho\in\bar{\mathbb{R}}} q(x)f\left(\frac{\rho}{q(x)}\right) + p(x)g\left(\frac{\rho}{p(x)}\right)dx \\
&= \int_{\mathcal{X}} i_{f,g}(p,q)(x)dx,
\end{aligned}
$$

where

$$
i_{f,g}(p,q)(\cdot) := \inf_{\rho\in\bar{\mathbb{R}}} q(\cdot)f\left(\frac{\rho}{q(\cdot)}\right) + p(\cdot)g\left(\frac{\rho}{p(\cdot)}\right).
$$

Let $x := \frac{\rho}{q} \in \bar{\mathbb{R}}^+$. Thus $\rho = xq$ and

$$
i_{f,g}(p,q) = \inf_{x\in\bar{\mathbb{R}}^+} qf(x) + pg(xq/p).
$$

Let $\tau = \frac{p}{q} \in \bar{\mathbb{R}}^+$. Thus

$$
\begin{aligned}
i_{f,g}(p,q)(\tau) &= \inf_{x\in\mathbb{R}^+} qf(x) + pg(x/\tau) \\
&= q\left[\inf_{x\in\bar{\mathbb{R}}^+} f(x) + \tau g(x/\tau)\right] \\
&= q\cdot(f\Box g)(\tau).
\end{aligned}
\tag{84}
$$

Let $h := f \Box g$. Observe from (84) that $i_{f,g}(p,q) = qh(p/q)$ and thus

$$\mathbb{I}_{f,g}(p,q) = \int_{\mathcal{X}} q(x) h\left(\frac{p(x)}{q(x)}\right) dx = \mathbb{I}_h(p,q)$$

if h is convex, which we know to be the case from Corollary 38. ■

It suggests the question: given a suitable convex f, does there always exist g such that $f = g \Box g$? This is analogous to the question of spectral factorisation (Sayed and Kailath, 2001) for ordinary linear convolution. We do not know the answer to this question, but have collected a few examples in Appendix G that demonstrates it is certainly true for *some* f. There does not appear to be a result analogous to part 5 of Theorem 36 for $\tilde{\mathbb{I}}_{f,g}$.

We have seen how f-divergences are related to integral probability metrics $V_{\mathcal{F}}$. It turns out that the variational divergence is special in being both. Many integral probability metrics are true metrics (Müller, 1997a,b). The only f-divergence that is a metric is the variational divergence. Whether there exist $\mathcal{F}$ such that $V_{\mathcal{F}}(\cdot,\cdot)$ is not a metric but equals $\mathbb{I}_f(\cdot,\cdot)$ for some $f \neq t \mapsto |t-1|$ (or affine transformation thereof) is left as an open problem.[28]

We end with another open problem. We have seen how $\underline{\mathbb{L}}_{\mathcal{F}}$ and $V_{\mathcal{F}}$ are related. This begs the question whether there is a representation of the form

$$\mathbb{I}_{f,\mathcal{F}}(P,Q) \overset{?}{=} \int_0^1 \Delta \underline{\mathbb{L}}_{\mathcal{F}}^{0-1}(\pi,P,Q)\gamma_f(\pi) d\pi.$$

Appendix G. Examples of Extended Convolution Factorisation

In this section we present three examples of f which can be written as $f = g \Box g$.

If $g(t) = (t-1)^2$ (corresponding to Pearson χ^2 divergence), $(g \Box g)(\tau) = \inf_{x \in \mathbb{R}^+}(x-1)^2 + \tau(x/\tau - 1)^2$. Differentiating the right-hand side with respect to x, setting to zero and solving for x gives $x = \frac{4}{2(1+1/\tau)}$. Substituting we obtain $(g \Box g)(\tau) = \frac{(\tau-1)^2}{\tau-1}$ which is the f for $\Delta(P,Q)$, the triangular discrimination.

If $g(t) = t\ln(t)$, a similar straightforward calculation yields $(g \Box g)(\tau) = \frac{-2\sqrt{\tau}}{e}$.

If $g(t) = (\sqrt{t} - 1)^2$ (corresponding to Hellinger divergence) then a similar calculation yields $(g \Box g)(\tau) = \frac{1}{2}(\sqrt{\tau} - 1)^2 = g(\tau)/2$. Thus this g plays a role analogous to a gaussian kernel in ordinary convolution. The significance of this is unclear.

We summarise the results (and the associated $g^\star$) in the following table.

$g(t)$	$(g \Box g)(\tau)$	$g^\star(s)$
$(t-1)^2$	$\frac{(\tau-1)^2}{\tau-1}$	$\frac{s^2}{4} + s$
$t\ln t$	$\frac{-2\sqrt{\tau}}{e}$	e^{s-1}
$(\sqrt{t}-1)^2$	$\frac{1}{2}(\sqrt{\tau}-1)^2$	$\frac{s}{1-s}[\![s<1]\!] + \infty[\![s \geq 1]\!]$

[28] This has in fact been solved by Sriperumbudur et al. (2009) since an earlier version of the present paper was published as an ArXiV preprint.

Whilst it is indeed straightforward to compute $(g \Box g)$ given g (although a simple closed form is not always possible), it is far from obvious how to go from a given f to a g such that $f = g \Box g$.

Hiriart-Urruty and Lemaréchal (1993a, page 69) show that for f convex on $\mathbb{R}^+$, g convex and increasing on $\mathbb{R}^+$,

$$(g \circ f)^\star(s) = \inf_{\alpha > 0} \alpha f^\star(\tfrac{s}{\alpha}) + g^\star(\alpha) = f^\star \Box g^\star.$$

This illuminates the difficulty of the above "factorisation problem". It is equivalent to: given a convex increasing $f^\star$, find a convex increasing $g^\star$ such that $f^\star = g^\star \circ g^\star$.

Appendix H. Empirical Estimators of $V_{B_{\mathcal{H}}, \frac{1}{2}}(P,Q)$ and SVMs

This appendix further develops the observations made in Section 8.2 regarding the relationship between divergence and risk when $\mathcal{R} = B_{\mathcal{H}}$, a unit ball in a reproducing kernel Hilbert space $\mathcal{H}$. In contrast to the rest of the paper (which focussed on relationships involving the underlying distributions), in this appendix we will consider the practical situation where there is only an empirical sample. We will see how the general results have interesting implications for sample based machine learning algorithms.

If we require an empirical estimate of $V_{\mathcal{R},\pi}(P,Q)$ we can replace P and Q by empirical distributions. We will use *weighted* empirical distributions. Given an independent identically distributed sample $\mathbf{w} = (w_1, \ldots, w_m) \in \mathcal{X}^m$ the $\boldsymbol{\alpha}$-weighted empirical distribution $\hat{P}_{\mathbf{w}}^{\boldsymbol{\alpha}}$ with respect to $\mathbf{w}$ is defined by

$$d\hat{P}_{\mathbf{w}}^{\boldsymbol{\alpha}} := \sum_{i=1}^m \alpha_i \delta(\cdot - w_i)$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, $\alpha_i \geq 0$, $i = 1, \ldots, m$ and $\sum_{i=1}^m \alpha_i = 1$. We will write $\hat{\mathbb{E}}_{\mathbf{w}}^{\boldsymbol{\alpha}} \phi := \mathbb{E}_{\hat{P}_{\mathbf{w}}^{\boldsymbol{\alpha}}} \phi = \sum_{i=1}^m \alpha_i \phi(w_i)$. Thus

$$V^2_{\mathcal{R},\frac{1}{2}}(\hat{P}_{\mathbf{w}}^{\boldsymbol{\alpha}}, \hat{P}_{\mathbf{z}}^{\boldsymbol{\beta}}) = \frac{1}{2} \| \hat{\mathbb{E}}_{\mathbf{w}}^{\boldsymbol{\alpha}} \phi - \hat{\mathbb{E}}_{\mathbf{z}}^{\boldsymbol{\beta}} \|^2_{\mathcal{H}}.$$

Suppose now that P and Q correspond to the positive and negative class conditional distributions. Let $\mathbf{x} := (x_1, \ldots, x_m)$ be a sample drawn from $M = \pi P + (1-\pi)Q$ with corresponding label vector $\mathbf{y} = (y_1, \ldots, y_m)$. Let $I := \{1, \ldots, m\}$, $I^+ := \{i \in I \colon y_i = 1\}$, $I^- := \{i \in I \colon y_i = -1\}$. Consider a weight vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ over the whole sample. Thus

$$\hat{\mathbb{E}}_P \phi = \sum_{i \in I^+} \alpha_i \phi(x_i) \quad \text{and} \quad \hat{\mathbb{E}}_Q \phi = \sum_{i \in I^-} \alpha_i \phi(x_i)$$

where we also require

$$\sum_{i \in I^+} \alpha_i = \frac{m^+}{m} \quad \text{and} \quad \sum_{i \in I^-} \alpha_i = \frac{m^-}{m}$$

and hence

$$\sum_{i \in I} \alpha_i y_i = \frac{m^+ - m^-}{m}.$$

Substituting into (65) we have

$$\begin{aligned}
2V_{B_{\mathcal{H}},\frac{1}{2}}(\hat{P},\hat{Q}) &= \left\langle \hat{\mathbb{E}}_P\phi - \hat{\mathbb{E}}_Q\phi, \hat{\mathbb{E}}_P\phi - \hat{\mathbb{E}}_Q\phi \right\rangle \\
&= \left\langle \sum_{i\in I^+}\alpha_i\phi(x_i) - \sum_{i\in I^-}\alpha_i\phi(x_i), \sum_{j\in I^+}\alpha_j\phi(x_j) - \sum_{j\in I^-}\alpha_j \right\rangle \\
&= \left\langle \sum_{i\in I}\alpha_i y_i\phi(x_i), \sum_{j\in I}\alpha_j y_j\phi(x_j) \right\rangle \\
&= \sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j y_i y_j \langle\phi(x_i),\phi(x_j)\rangle \\
&= \sum_{i\in I}\sum_{j\in I}\alpha_i\alpha_j y_i y_j k(x_i,x_j) =: J(\boldsymbol{\alpha},\mathbf{x}). \qquad (85)
\end{aligned}$$

We now consider three different choices of $\boldsymbol{\alpha}$.

Uniform weighting If we set $\alpha_i = \frac{1}{m}$, $i = 1,\dots,m$, then (85) becomes

$$\frac{1}{m^2}\sum_{i,j\in I} y_i y_j k(x_i,x_j) = \mathrm{MMD}_b^2[B_{\mathcal{H}}, \boldsymbol{x}^+, \boldsymbol{x}^-]$$

where $\boldsymbol{x}^+ := (x_i)_{i\in I^+}$, $\boldsymbol{x}^- := (x_i)_{i\in I^-}$ and MMD_b is the biased estimator of the *Maximum Mean Discrepancy* (Gretton et al., 2008), an alternate name for $V_{\mathcal{R}}$. Observe that from theorem 34, this case corresponds to using a Fisher linear discriminant in feature space (Devroye et al., 1996) when it is assumed that the within-class covariance matrices are both the identity matrix. This follows by observing that the constructed hypothesis is identical in both cases.

Pessimistic Weighting Instead of weighting each sample equally, one can optimise over $\boldsymbol{\alpha}$. By theorem 34, minimizing $J(\boldsymbol{\alpha},\mathbf{x})$ over $\boldsymbol{\alpha}$ will maximize $\underline{\mathbb{L}}^{\mathrm{lin}}$ and is thus the most pessimistic choice. Explicitly, we have

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m}\sum_{i=1}^{m}\alpha_i\alpha_j y_i y_j k(x_i,x_j) && (86)\\
\text{s.t.} \quad & \alpha_i \geq 0, \;\; i = 1,\dots,m && (87)\\
& \sum_{i=1}^{m}\alpha_i y_i = \frac{m^+ - m^-}{m} && (88)\\
& \sum_{i=1}^{m}\alpha_i = 1 && (89)
\end{aligned}$$

which can be recognized as the support vector machine (Cortes and Vapnik, 1995). The SVM uses the sign of the "witness" (Gretton et al., 2008), $x \mapsto \sum_{i=1}^{m}\alpha_i y_i k(x_i,x)$ as its predictor.

Interpolation between above two cases A parameterized interpolation between the above two cases can be constructed by the addition of the constraints

$$\alpha_i \leq \frac{1}{\nu m}, \;\; i = 1,\dots,m, \qquad (90)$$

where $\nu \in (0,1]$ is an adjustable parameter. Observe that ν controls the sparsity of $\boldsymbol{\alpha}$ since (90), (87) and (89) together imply that $|\{i \in I \colon \alpha_i \neq 0\}| \geq \nu m$. Crisp and Burges (2000) have shown that (86),...,(90) is equivalent to the ν-SVM algorithm (Schölkopf et al., 2000).

While “information-theoretic” approaches to the SVM and weighted kernel representations are hardly new,[29] the results presented here are novel and provide a simple and direct derivation of the SVM via the generalised variational divergence.

If $V_{B_{\mathcal{H}},\frac{1}{2}}(\hat{P}_{\boldsymbol{w}},\hat{Q}_{\boldsymbol{z}})$ is used as a test statistic to infer whether two samples $\boldsymbol{w}$ and $\boldsymbol{z}$ are drawn from the same distribution (as Gretton et al., 2008 do), then when the distributions from which $\boldsymbol{w}$ and $\boldsymbol{z}$ are drawn are close, the classification performance of the corresponding classifier (i.e., the classifier that uses the sign of the witness function) will be close to the worst possible. Thus one will be operating in a regime distinct from the normal situation, where the risk is typically small.

Finally observe that the derivation of the SVM presented here could be viewed as an application of an alternate “inductive principle”—a general recipe for constructing learning algorithms from learning task specification (Vapnik, 1989, 2006). The traditional Empirical Risk Minimization principle entails replacing (P,Q) with $(\hat{P}_{\boldsymbol{x}^+},\hat{Q}_{\boldsymbol{x}^-})$ in the definition of $\underline{\mathbb{L}}(\pi,P,Q)$. Then, in order to not overfit, one restricts the class of functions from which hypotheses are drawn. That is, there are two approximations:

$$\underline{\mathbb{L}}(\pi,P,Q) \xrightarrow{\text{Empirical Approximation (uniform)}} \underline{\mathbb{L}}(\pi,\hat{P}_{\boldsymbol{x}^+},\hat{Q}_{\boldsymbol{x}^-}) \xrightarrow{\text{Restrict Class}} \underline{\mathbb{L}}_{\mathcal{R}}(\pi,\hat{P}_{\boldsymbol{x}^+},\hat{Q}_{\boldsymbol{x}^-}).$$

Upon setting $\boldsymbol{\alpha}^+ = (\alpha_i)_{i\in I^+}$ and $\boldsymbol{\alpha}^- = (\alpha_i)_{i\in I^-}$, the derivation presented above, in contrast, can be summarised schematically by

$$\text{“}\underline{\mathbb{L}}(\pi,P,Q)\text{”} \xrightarrow{\text{Restrict Class}} \underline{\mathbb{L}}_{\mathcal{R}}(\pi,P,Q) \xrightarrow{\text{Empirical Approximation ($\boldsymbol{\alpha}$-weighted)}} \underline{\mathbb{L}}_{\mathcal{R}}(\pi,\hat{P}^{\boldsymbol{\alpha}^+}_{\boldsymbol{x}^+},\hat{Q}^{\boldsymbol{\alpha}^-}_{\boldsymbol{x}^-}),$$

where a different loss (the “linear” loss) was used at the start. With that loss function, reversing the order of the two approximations would not work, and is (thus) not equivalent to the ERM inductive principle. The first step makes $\underline{\mathbb{L}}$ well defined—with no restriction it is not, hence the quotes; and will avoid overfitting in any case. The second step is the more general ($\boldsymbol{\alpha}$-weighted) empirical approximation.

We believe that this alternate derivation of the SVM is of interest because it is simpler (avoids the need to introduces margins) and it elucidates the connection between the kernel methods for

[29] The use of kernel representations for classification is of course not new: from the classical kernel classifier (where $\alpha_i = 1/m$ for all $i \in I$) (Devroye et al., 1996, Chapter 10) to the Generalised Portrait (Aizerman et al., 1964), the Generalised Discriminant (Baudat and Anouar, 2000) and the panoply of techniques inspired by Support Vector Machines (Schölkopf and Smola, 2002; Herbrich, 2002). None of these techniques is designed from the perspective of minimising a f-divergence.

Principe et al. (2000a) have developed an approach to machine learning problems based on information theoretic criteria (Principe et al., 2000b; Jenssen et al., 2004; Xu et al., 2005; Jenssen, 2005a; Jenssen et al., 2006; Pavia et al., 2006). Jenssen et al. (2004, 2006) considered kernel methods from the perspective of Renyi’s quadratic entropy. They do not exploit the formal relationship between maximising divergence and minimising risk. They interpret the SVM as being constructed from weighted Parzen windows density estimates. Gretton et al. (2008) explained the relationship between their MMD estimators and those derived from (unweighted) Parzen windows estimates of the class-conditional distributions. Weighted Parzen windows estimates were used as a basis for building a classifier by Babich and Camps (1996). Weighted empirical distributions are widely used in particle filtering (Crisan and Doucet, 2002).

McDermott and Katagiri (2002) considered the direct optimisation of a classifier built on top of Parzen windows density estimates. They showed that the minimum classification error criterion is equivalent to a Parzen windows estimate of the theoretical Bayes risk. They re-derive the traditional approach of minimising an estimate of the expected loss. McDermott and Katagiri (2003) extended their approach to the multi-class setting in a way that takes account of all the “other” classes better in estimating the probability of error of a given class.

classification and MMD—indeed MMD is nothing but the Fisher linear discriminant applied to a binary problem induced by the given distributions P and Q.

References

H. Abelson, G.J. Sussman, and J. Sussman. *Structure and Interpretation of Computer Programs*. MIT Press, 1996.

J. Aczél. Measuring information beyond communication theory — why some generalized information measures maybe be useful, others not. *Aequationes Mathematicae*, 27:1–19, 1984.

M. A. Aizerman, É. M. Braverman, and L. I. Rozonoér. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

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.

Y. Altun and A. Smola. Unifying divergence minimization and statistical inference via convex duality. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pages 139–153, 2006.

P. Antosik, J. Mikusinski, and R. Sikorski. *Theory of Distributions: The Sequential Approach*. American Elsevier, 1973.

W. R. Ashby. *An Introduction to Cybernetics*. Chapman and Hall, 1956.

G.A. Babich and O.I. Camps. Weighted Parzen windows for pattern classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5):567–570, 1996.

F.R. Bach, D. Heckerman, and E. Horvitz. Considering cost asymmetry in learning classifiers. *Journal of Machine Learning Research*, 7:1713–1741, 2006.

C.Y. Baldwin and K.B. Clark. Modularity in the design of complex engineering systems. In Dan Braha Ali Minai and Yaneer Bar Yam, editors, *Complex Engineered Systems: Science Meets Technology*. Springer, 2006a.

C.Y. Baldwin and K.B. Clark. Between 'knowledge' and 'the economy': Notes on the scientific study of designs. In D. Forey and B. Kahin, editors, *Advancing Knowledge and the Knowledge Economy*, pages 299–328, Cambridge, Mass., 2006b. MIT Press.

A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005a.

A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005b.

V. Barnett. *Comparative Statistical Inference*. John Wiley and Sons, Chichester, 3rd edition, 1999.

P.L. Bartlett, P.M. Long, and R.C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, March 2006.

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

H. H. Bauschke, R. Goebel, Y. Lucet, and X. Wang. The proximal average: Basic theory. *SIAM Journal on Optimization*, 19:766–785, 2008.

M.J. Bayarri and J.O. Berger. The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

M. Ben-Bassat. epsilon-equivalence of feature selection rules. *IEEE Transactions on Information Theory*, 24(6):769–772, 1978.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman-Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985.

J.O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1): 1–32, 2003.

I. Berlin. *The Hedgehog and the Fox: An Essay on Tolstoy's view of History*. Weidenfeld and Nicolson, London, 1953.

A. Beygelzimer, V. Dani, T. Hayes, J. Langford, and B. Zadrozny. Error limiting reductions between classification tasks. In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, 2005.

A. Beygelzimer, J. Langford, and P. Ravikumar. Multiclass classification with filter trees. Preprint, June 2007. URL `http://hunch.net/~jl/projects/reductions/mc_to_b/invertedTree.pdf`.

A. Beygelzimer, J. Langford, and B. Zadrozny. Machine learning techniques — reductions between prediction quality metrics. In Zhen Liu and Cathy H. Xia, editors, *Performance Modeling and Engineering*, pages 3–28. Springer US, April 2008. URL `http://hunch.net/~jl/projects/reductions/tutorial/paper/chapter.pdf`.

P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*, volume 1. Prentice-Hall, 2nd edition, 2001.

A. Birnbaum. On the foundations of statistical inference: Binary experiments. *The Annals of Mathematical Statistics*, 32(2):414–435, June 1961.

D. Blackwell. Comparison of experiments. In J. Neyman, editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 93–102, Berkeley and Los Angeles, 31 July – 12 August 1951. University of California Press.

D. Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953.

A. Blass and Y. Gurevich. Algorithms: A quest for absolute definitions. *Bulletin of the European Association for Theoretical Computer Science*, 81:195–225, 2003.

A. Blass, N. Dershowitz, and Y. Gurevich. When are two algorithms the same? *Bull. Symbolic Logic*, 15(2):145–168, 2009.

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.

K.M. Borgwardt, A. Gretton, M.J. Rasch, H.P. Kriegel, B. Schölkopf, and A.J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14), 2006.

J.M. Borwein and A.S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal of Control and Optimization*, 29(2):325–338, 1991.

O. Bousquet. Making machine learning more scientific. Blog Post., 2006. URL `http://ml.typepad.com/machine_learning_thoughts/2006/06/making_machine_.html`.

G.C. Bowker. *Memory practices in the sciences*. MIT Press, 2005.

G.C. Bowker and S.L. Star. *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge, Mass., 1999.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

M. Broniatowski. Minimum divergence in inference and testing. In *Atti della XLII Riunione Scientifica*, Università di Bari, 2004. Società Italiana Statistica.

M. Broniatowski and A. Keziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16–36, 2009.

L.D. Brown and M.G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics*, 24(6):2384–2398, 1996.

L.D. Brown and L. Zhao. Direct asymptotic equivalence of nonparametric regression and the infinite dimensional location problem. Preprint, 2003. URL `http://www-stat.wharton.upenn.edu/~lzhao/papers/`.

L.D. Brown, T.T. Cai, M.G. Low, and C.H. Zhang. Asymptotic equivalence theory for nonparametric regression with random design. *Annals of Statistics*, 30(3):688–707, 2002.

H.D. Brunk, G.M. Ewing, and W.R. Utz. Minimizing integrals in certain classes of monotone functions. *Pacific Journal of Mathematics*, 7(1):833–847, 1957.

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, November 2005.

S. Canu and A. Smola. Kernel methods and the exponential family. *Neurocomputing*, 69(7-9): 714–720, 2006.

V.J. Carey, J. Mar, and R. Gentleman. MLInterfaces: Towards uniform behaviour of machine learning tools in R. Technical report, Harvard University, 2007.

A.V. Carter. Deficiency distance between multinomial and multivariate normal experiments. *Annals of Statistics*, 30(3):708–730, 2002.

Y. Censor and S.A. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, 1997.

M. Cepedello-Boiso. On regularization in superreflexive Banach spaces by infimal convolution formulas. *Studia Mathematica*, 129(3):265–284, 1998.

P. Chaudhuri and W.Y. Loh. Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 8(5):561–576, 2002.

G. Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5(54):131–295, 1953.

W.J. Conover and R.L. Iman. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124–129, 1981.

C. Cortes and M. Mohri. AUC optimization vs. error rate minimization. *Advances in Neural Information Processing Systems*, 16, 2004.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

D. Cossock and T. Zhang. Subset ranking using regression. In G. Lugosi and H.U. Simon, editors, *Proceedings of Conference on Learning Theory (COLT) 2006*, LNAI 4005, pages 605–619, 2006.

D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing,*, 50(3):736–746, 2002.

D. J. Crisp and C. J. C. Burges. A geometric interpretation of ν-SVM classifiers. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 244–250. MIT Press, 2000.

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

I. Csiszár. Information measures: A critical survey. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 73–86, Dordrecht, 1978. D. Riedel.

I. Csiszár. Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68 (1):161–186, March 1995.

A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25 (6):2300–2312, 1997.

A.P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, March 2007.

M.H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

M.H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Company, New York, 1970.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Approach to Pattern Recognition*. Springer, New York, 1996.

P. Domingos. MetaCost: a general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155–164, New York, NY, USA, 1999. ACM Press.

P. Domingos and M. Richardson. Markov logic: A unifying framework for statistical relational learning. In *Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields*, pages 49–54, 2004.

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.

C. Drummond. Discriminative vs. generative classifiers for cost sensitive learning. In *Proceedings of the Nineteenth Canadian Conference on Artificial Intelligence*, LNAI 4013, pages 479–490, Quebec, Canada, 2006.

C. Drummond and R.C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2nd edition, 2003.

S. Eguchi. Information geometry and statistical pattern recognition. To appear in *Sugaku Exposition*, American Mathematical Society, 2005.

S. Eguchi and J. Copas. Recent developments in discriminant analysis from an information geometric point of view. *Journal of the Korean Statistical Society*, 30:247–264, 2001.

T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31, 2004.

T. Fawcett. ROC graphs with instance-varying costs. *Pattern Recognition Letters*, 27(8):882–891, 2006.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 82–88. AAAI Press, 1996.

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

D. Feldman and F. Österreicher. A note on f-divergences. *Studia Scientiarum Mathematicarum Hungarica*, 24:191–200, 1989.

P.A. Flach. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*, pages 194–201, 2003.

P.A. Flach and S. Wu. Repairing concavities in ROC curves. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI-2005)*, page 702, 2005.

L. Floridi. Open problems in the philosophy of information. *Metaphilosophy*, 35(4):554–582, 2004.

F.G. Friedlander. *Introduction to the Theory of Distributions*. Cambridge University Press, 1982.

M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. WH Freeman & Co. New York, NY, USA, 1979.

J. K. Gershenson, G. J. Prasad, and Y. Zhang. Product modularity: definitions and benefits. *Journal of Engineering Design*, 14:295–313, 2003.

M. Ghallab, A. Howe, C. Knoblock, D. McDermott, A. Ram, M. Veloso, D. Weld, and D. Wilkins. PDDL specification, 1998. AIPS-98 Planning Committee.

A.L. Gibbs and F.E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–435, 2002.

G. L. Gilardoni. On Pinsker's type inequalities and Csiszár's f-divergences. part i: Second and fourth-order inequalities. arXiv:cs/0603097v2, April 2006a.

G.L. Gilardoni. On the minimum f-divergence for a given total variation. *Comptes Rendus Académie des sciences, Paris, Series 1*, 343, 2006b.

G.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 2006c. URL `http://www.unb.br/ie/est/docentes/curriculo/gustavo_homepage/artigos/improved.pdf`.

T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007.

P.K. Goel and M.H. DeGroot. Comparison of Experiments and Information Measures. *The Annals of Statistics*, 7(5):1066–1077, 1979.

P.K. Goel and J. Ginebra. When is one experiment 'always better than' another? *Journal of the Royal Statistical Society Series D (The Statistician)*, 52(4):515–537, 2003.

P.W. Goldberg. When can two unsupervised learners achieve PAC separation? *Proceedings of the 14th Annual Conference on Computational Learning Theory (COLT)*, pages 303–319, 2001.

S.A. Goldman, R.L. Rivest, and R.E. Schapire. Learning binary relations and total orders. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, pages 46–51, 1989.

J.D.J. Golic. On the relationship between the information measures and the Bayes probability of error. *IEEE Transactions on Information Theory*, 33(5):681–693, 1987.

A. Gretton, K.M. Borgwardt, M. Rasch, B. Schölkopf, and A.J. Smola. A kernel method for the two-sample-problem. *Journal of Machine Learning Research*, 2008. submitted.

R.L. Grossman, M.F. Hornick, and G. Meyer. Data mining standards initiatives. *Commun. ACM*, 45(8):59–61, 2002. ISSN 0001-0782.

P.D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

P.D. Grünwald and A.P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

C. 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.

D.J. Hand. Deconstructing statistical questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3):317–356, 1994.

D.J. Hand. Comment on "The skill-plot: A graphical technique for evaluating continuous diagnostic tests". *Biometrics*, 63:259, 2008.

D.J. Hand and R.J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

D.J. Hand and V. Vinciotti. Local versus global models for classification problems: Fitting models where it matters. *The American Statistician*, 57(2):124–131, 2003.

J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

P. Harremoës. *Time and Conditional Independence*. Ph.d. thesis, Roskilde University, 1993. Original in Danish entitled Tid og Betinget Uafhængighed. English translation partially available at `http://www.math.ku.dk/~moes/index.html`.

R. Herbrich. *Learning Kernel Classifiers*. MIT Press, Cambridge MA, 2002.

H. Heyer. *Theory of Statistical Experiments*. Springer, 1982.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms: Part II: Advanced Theory and Bundle Methods*. Springer, Berlin, 1993a.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms: Part I: Fundamentals*. Springer, Berlin, 1993b.

J-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001.

J.-B. Hiriart-Urruty and J.-E. Martínez-Legaz. Convex solutions of a functional equation arising in information theory. *Journal of Mathematical Analysis and Applications*, 328:1309–1320, 2007.

W. James. *A Pluralistic Universe*. Longmans, Green, and Co., London, 1909. Hibbert Lectures at Manchester College on the Present Situation in Philosophy.

R. Jenssen. An information theoretic approach to machine learning. Doctor Scientiarum Thesis, Department of Physics, Faculty of Science, University of Tromsø, 2005a.

R. Jenssen. *An Information Theoretic Approach to Machine Learning*. PhD thesis, Department of Physics, University of Tromsø, Norway, May 2005b.

R. Jenssen, D. Erdogmus, J.C. Principe, and T. Eltoft. Towards a unification of information theoretic learning and kernel methods. In *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP2004)*, pages 93–102, 2004.

R. Jenssen, D. Erdogmus, J.C. Príncipe, and T. Eltoft. Some equivalences between kernel methods and information theoretic methods. *Journal of VLSI Signal Processing*, 45(1):49–65, 2006. (Paper 3 of Jenssen (2005a)).

D.S. Johnson. NP-Completeness Columns. *Journal of Algorithms; ACM Transactions on Algorithms*, 2–13; 1–3, 1982–1992; 2005–2007. URL `http://www.research.att.com/~dsj/columns/`.

M.I. Jordan. *Learning in Graphical Models*. MIT Press, 1999.

T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communications*, 15(1):52–60, 1967.

A. Keziou. Dual representations of ϕ-divergences and applications. *Comptes Rendus Académie des sciences, Paris, Series 1*, 336:857–862, 2003a.

A. Keziou. *Utilisation des Divergences entre Mesures en Statistique Inférentielle*. PhD thesis, Université Paris 6, 2003b.

M. Khosravifard, D. Fooladivanda, and T.A. 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.

J. Kiefer. The foundations of statistics — are there any? *Synthese*, 36:161–176, 1977.

J.C. Kiefer. *Introduction to Statistical Inference*. Springer-Verlag, New York, 1987.

D.E. Knuth. Two notes on notation. *American Mathematical Monthly*, pages 403–422, 1992.

F.R. Kschischang, B.J. Frey, and H.A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

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.

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.

N. Lambert, J. Langford, J. Wortman, Y. Chen, D. Reeves, Y. Shoham, and D. Pennock. Self-financed wagering mechanisms for forecasting. In Lance Fortnow, John Riedl, and Tuomas Sandholm, editors, *Proceedings of the ACM Conference on Electronic Commerce*, pages 170–179, 2008.

J. Langford. Machine learning reductions tutorial. Slides presented at the *Machine Learning Summer School*, July 2006.

J. Langford. Alternative machine learning reductions definitions. Machine Learning (Theory), March 2007. URL `http://hunch.net/?p=255`.

J. Langford and A. Beygelzimer. Sensitive error correcting output codes. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT-2005)*, pages 158–172. Springer, 2005.

J. Langford and B. Zadrozny. Estimating class membership probabilities using classifier learners. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTAT'05)*, 2005.

J. Langford, R. Oliveira, and B. Zadrozny. Predicting conditional quantiles via reduction to classification. In *Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence*. AUAI Press, 2006.

J.A. Lasserre, C.M. Bishop, and T.P. Minka. Principled hybrids of generative and discriminative models. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Volume 1*, pages 87–94. IEEE Computer Society Washington, DC, USA, 2006.

L. LeCam. Sufficiency and Approximate Sufficiency. *The Annals of Mathematical Statistics*, 35(4): 1419–1455, 1964.

L. LeCam. *Asymptotic Methods in Statistical Decision Theory*. Springer, 1986.

L. Li and H.-T. Lin. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems 19*, pages 865–873, 2007.

F. Liese and K.-J. Miescke. *Statistical Decision Theory*. Springer, New York, 2008.

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

D.V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

J. Lindström. On the origin and early history of functional analysis. Technical Report U.U.D.M. Projrect Report 2008:1, Uppsala Universitet, January 2008.

J. Locke. *An Essay Concerning Human Understanding*. Thomas Basset, London, 1690.

P.M. Long and R.A. Servedio. Discriminative learning can succeed where generative learning fails. In G. Lugosi and H.U. Simon, editors, *Proceedings of the Conference on Learning Theory (COLT)*, LNAI 4005, pages 319–334, Berlin, 2006. Springer-Verlag.

P.M. Long, R.A. Servedio, and H.U. Simon. Discriminative learning can succeed where generative learning fails. Preprint, correction to Long and Servedio (2006), November 2006.

R. Lutz and M. Musio. An alternative mathematical foundation for statistics. *Acta Applicandae Mathematicae: An International Survey Journal on Applying Mathematics and Mathematical Applications*, 89(1):217–249, 2005.

C. Lyell. *Principles of Geology*. John Murray, London, 1830.

R. A. Maxion and R. R. Roberts. Proper use of ROC curves in intrusion/anomaly detection. Technical Report CS-TR-871, School of Computing Science, University of Newcastle upon Tyne, November 2004.

E. McDermott and S. Katagiri. A Parzen window based derivation of minimum classification error from the theoretical Bayes classification risk. In *Proceedings of International Conference on Spoken Language Processing*, pages 2465–2468, 2002.

E. McDermott and S. Katagiri. A new formalization of minimum classification error using a Parzen estimate of classification chance. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2003. Paper 4406.

T. Minka. Discriminative models, not discriminative training. Technical Report MSR-TR-2005-144, Microsoft Research, Cambridge, October 2005.

J. Mokyr. *The Lever of Riches: Technological Creativity and Economic Progress*. Oxford University Press, USA, 1992.

N. Morse and R. Sacksteder. Statistical isomorphism. *The Annals of Mathematical Statistics*, 37(1): 203–214, 1966.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997a.

A. Müller. Stochastic orders generated by integrals: A unified study. *Advances in Applied Probability*, 29:414–428, 1997b.

E. Nelson. *Radically Elementary Probability Theory*. Princeton University Press, 1987.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Math. or Phys. Character (1896-1934)*, 231:289–337, 1933. URL `http://dx.doi.org/10.1098/rsta.1933.0009`.

A. Y. Ng and M.I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes. In *Advances in Neural Information Processing Systems (NIPS)*, volume 15, 2002.

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.

X. Nguyen, M.J. Wainwright, and M.I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37:876–904, 2009.

D. Noble. *The Music of Life: Biology Beyond the Genome*. Oxford University Press, 2006.

F. Österreicher. f-divergences — representation theorem and metrizability. Technical Report, Institute of Mathematics, University of Salzburg, September 2003.

F. Österreicher and D. Feldman. Divergenzen von Wahrscheinlichkeitsverteilungen — Integralgeometrisch Betrachtet. *Acta Mathematica Academaiae Scientiarum Hungarica*, 37(4):329–337, 1981.

F. Österreicher and I. Vajda. Statistical information and discrimination. *IEEE Transactions on Information Theory*, 39(3):1036–1039, 1993.

N. Palmer and P.W. Goldberg. PAC classification based on PAC estimates of label class distributions. *Arxiv preprint cs.LG/0607047*, 2006.

A.R.C. Pavia, J.-.W Xu, and J.C. Príncipe. Kernel principal components are maximum entropy projections. In *ICA 2006. Springer Lecture Notes in Computer Science 3889*, pages 846–853, 2006.

R.R. Phelps. *Lectures on Choquet's Theorem*, volume 1757 of *Lecture Notes in Mathematics*. Springer, 2nd edition, 2001.

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

D. Pollard. Some thoughts on LeCam's statistical decision theory. Preprint, Statistics Department, Yale University, May 2000. URL `http://www.stat.yale.edu/~pollard/Papers/thoughts.pdf`.

H. Vincent Poor and John B. Thomas. Applications of Ali-Silvey distance measures in the design of generalized quantizers for binary decision systems. *IEEE Transactions on Communications*, 25(9):893–900, September 1977.

J.C. Principe, D. Xu, and J. Fisher. Information theoretic learning. In Simon Haykin, editor, *Unsupervised Adaptive Filtering: Volume 1, Blind Source Seperation*, pages 265–319, New York, 2000a. Wiley.

J.C. Principe, D. Xu, Q. Zhao, and J.W. Fisher III. Learning from examples with information theoretic criteria. *The Journal of VLSI Signal Processing*, 26(1–2):61–77, 2000b.

S.T. Rachev. *Probability metrics and the stability of stochastic models*. Wiley series in probability and mathematical statistics, Chichester, 1991.

S. Raudys. *Statistical and Neural Classifiers: An Integrated Approach to Design*, chapter Taxonomy of Pattern Classification Algorithms. Springer, London, 2001.

M.D. Reid and R.C. Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the International Conference on Machine Learning*, pages 897–904, 2009.

M.D. Reid and R.C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1):107–136, 2006.

J. Rissanen. *Information and Complexity in Statistical Modelling*. Springer, New York, 2007.

C.P. Robert. *The Bayesian Choice: A Decision-Theoretic Motivation*. Springer, New York, 1994.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Springer-Verlag, Berlin, 2004.

Y.D. Rubinstein and T. Hastie. Discriminative vs informative learning. In *Knowledge Discovery and Data Mining*, pages 49–53. AAAI Press, 1997.

R. Sacksteder. A note on statistical equivalence. *The Annals of Mathematical Statistics*, 38(3): 787–794, 1967.

L.J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

A.H. Sayed and T. Kailath. A survey of spectral factorization methods. *Numerical Linear Algebra with Applications*, 8(6-7):467–496, 2001.

M.J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17 (4):1856–1879, 1989.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge MA, 2002.

B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

C. Scott. Calibrated surrogate losses for classification with label-dependent costs. arXiv:1009:2718v1, September 2010.

C. Scott and M. Davenport. Regression level set estimation via cost-sensitive classification. *IEEE Transactions on Signal Processing*, 55(6 Part 1):2752–2757, 2007.

G. Shafer and V. Vovk. *Probability and Finance: It's Only a Game!* J. Wiley & Sons, 2001.

C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423; 623–656, 1948.

Y. Shen. *Loss Functions for Binary Classification and Class Probability Estimation*. PhD thesis, Department of Statistics, University of Pennsylvania, October 2005.

E. Shuford, A. Albert, and H.E. Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, June 1966.

L.N. Soldatova and R.D. King. An ontology of scientific experiments. *Interface: The Journal of The Royal Society*, 3(11):795–803, 2006.

L. Song, M.D. Reid, A.J. Smola, and R.C. Williamson. Discriminative estimation of f-divergence. Submitted to AISTATS09, October 2008.

B.K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. On integral probability metrics, ϕ-divergences and binary classification. Arxiv preprint arXiv:0901.2698v4, October 2009.

I. Steinwart. How to compare different loss functions and their risks. Preprint, Modeling, Algorithms and Informatics Group, CCS-3, Los Alamos National Laboratory, September 2006.

I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, August 2007.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

I. Steinwart, D. Hush, and C. Scovel. Density level detection is classification. *Advances in Neural Information Processing Systems*, 17, 2005.

H. Strasser. *Mathematical Theory of Statistics*. Walter de Gruyter, Berlin, 1985.

H. Strasser. Reduction of complexity. In Josef A. Mazanec and Helmut Strasser, editors, *A Nonparametric Approach to Perceptions-Based Market Segmentation: Foundations*, pages 99–140, Wien, 2000. Springer.

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

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

D. Tasche. Conditional expectation as quantile derivative. Arxiv preprint math.PR/0104190, 2001.

W. B. Temple. Stieltjes integral representation of convex functions. *Duke Mathematical Journal*, 21:527–531, 1954.

N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. Arxiv preprint physics/0004057, 2000.

S. Tong and D. Koller. Restricted Bayes optimal classifiers. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pages 658–664, 2000.

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

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

F. Topsøe. Between truth and description. Presented at Prague Stochastics 2006, June 2006.

E.N. Torgersen. Measures of information based on comparison with total information and with total ignorance. *The Annals of Statistics*, 9(3):638–657, 1981.

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

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.

S. Turkle and S. Papert. Epistemological pluralism and the revaluation of the concrete. *Journal of Mathematical Behavior*, 11(1):3–33, March 1992.

P. Turney. Types of cost in inductive concept learning. In *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning*, pages 15–21, 2000.

A. Unterreiter, A. Arnold, P. Markowich, and G. Toscani. On generalized Csiszár-Kullback inequalities. *Monatshefte für Mathematik*, 131:235–253, 2000.

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

A. van der Vaart. The statistical work of Lucien Le Cam. *Annals of Statistics*, 30(3):631–682, 2002.

V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 2006. 2nd edition.

V. N. Vapnik. Inductive principles of the search for empirical dependences (methods based on weak convergence of probability measures). In *COLT '89: Proceedings of the second annual workshop on computational learning theory*, pages 3–21, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

H.R. Varian. Innovation, components and complements. University of California, Berkeley, October 2003. URL `http://www.almaden.ibm.com/coevolution/pdf/varian_paper.pdf`.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, 2005.

A. Wald. Statistical decision functions. *The Annals of Mathematical Statistics*, 20(2):165–205, June 1949.

A. Wald. *Statistical Decision Functions*. John Wiley & Sons, New York, 1950.

M.L. Weitzman. Recombinant growth. *Quarterly Journal of Economics*, 113(2):331–360, 1998.

D. Wettschereck and S. Muller. Exchanging data mining models with the predictive modelling markup language. *International Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, 2001.

Wikipedia. Grothendieck's relative point of view. Wikipedia, September 2007. URL `http://en.wikipedia.org/wiki/Grothendieck's_relative_point_of_view`. Accessed 17/09/2007.

R. Winkler, J. Muñoz, J. Cervera, J. Bernardo, G. Blattenberger, J. Kadane, D. Lindley, A. Murphy, R. Oliver, and D. Ríos-Insua. Scoring rules and the evaluation of probabilities. *TEST*, 5(1):1–60, March 1990.

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

A.P. Worthen and W.E. Stark. Unified design of iterative receivers using factor graphs. *IEEE Transactions on Information Theory*, 47(2):843–849, 2001.

J. Xie and C.E. Priebe. A weighted generalization of the Mann-Whitney-Wilcoxon statistic. *Journal of Statistical Planning and Inference*, 102(2):441–466, 2002.

J.-W. Xu, D. Erdogmus, R. Jenssen, and J.C. Príncipe. An information-theoretic perspective to kernel independent component analysis. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, volume 5, pages 249–252, 2005.

G.L. Yang and L. Le Cam. A conversation with Lucien Le Cam. *Statistical Science*, 14(2):223–241, 1999.

B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, pages 435–442, 2003.

J. Zhang. Divergence function, duality, and convex analysis. *Neural Computation*, 16(1):159–195, 2004a.

J. Zhang and H. Matsuzoe. Dualistic riemannian manifold structure induced from convex functions. *Advances in Mechanics and Mathematics.*, 17:437–464, 2009.

T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Annals of Mathematical Statistics*, 32:56–134, 2004b.

V.M. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 38(2):278–302, 1984.