Journal of Machine Learning Research 24 (2023) 1-72 Submitted 9/22; Revised 8/23; Published 11/23

The Geometry and Calculus of Losses

Robert C. Williamson BOB.WILLIAMSON@UNI-TUEBINGEN.DE
University of Tübingen and Tübingen AI Center, Germany

Zac Cranko ZAC.CRANKO@GMAIL.COM
Sydney, Australia

Editor: Aryeh Kontorovich

Abstract

Statistical decision problems lie at the heart of statistical machine learning. The simplest problems are multiclass classification and class probability estimation. Central to their definition is the choice of loss function, which is the means by which the quality of a solution is evaluated. In this paper we systematically develop the theory of loss functions for such problems from a novel perspective whose basic ingredients are convex sets with a particular structure. The loss function is defined as the subgradient of the support function of the convex set. It is consequently automatically proper (calibrated for probability estimation). This perspective provides three novel opportunities. It enables the development of a fundamental relationship between losses and (anti)-norms that appears to have not been noticed before. Second, it enables the development of a calculus of losses induced by the calculus of convex sets which allows the interpolation between different losses, and thus is a potential useful design tool for tailoring losses to particular problems. In doing this we build upon, and considerably extend, existing results on M-sums of convex sets. Third, the perspective leads to a natural theory of "polar" loss functions, which are derived from the polar dual of the convex set defining the loss, and which form a natural universal substitution function for Vovk's aggregating algorithm.

Keywords: convex sets, support functions, gauges, polars, concave duality, proper loss functions, M-sums, distorted probabilities, polar losses, Shephard duality, anti-norms, Bregman divergences, semi inner products, Finsler geometry, aggregating algorithm, substitution functions, direct and inverse addition.

1. Introduction

Most machine learning research focusses on *methods* (algorithms). But these methods are designed to solve particular *problems*. Platt (1962) argued for the greater importance of problem-oriented research. Our premise is that we need to better understand the elements of machine learning problems, and their permissible transformations. We focus on some of the simplest possible machine learning problems, namely multiclass classification and probability estimation.

Stateless machine learning problems have three key ingredients:

1. the *loss function* ℓ: specifies how predictive performance is evaluated;
2. the *data generating process*: in the statistical setting this corresponds to an underlying probability distribution P from which samples are drawn;

3. the *model class* $\mathscr{F}$: the analyst's choice, informed by their prior knowledge[1].

Implicit in this is the protocol by which the learner or analyst interacts with the data; we presume the usual statistical batch setting for now, but most of the technical results of the paper are not so restricted. Thus an (idealised) machine learning problem can be parametrised by the triple $(\ell, \mathscr{F}, P)$.

Much research in machine learning focusses upon the classes of models $\mathscr{F}$ and methods for searching for the best element within $\mathscr{F}$ for data generated by P, on theoretical results concerning the complexity of $\mathscr{F}$ and its effect on convergence of empirical estimates (Vapnik, 1998), or the intrinsic geometry induced by $\mathscr{F}$ (Amari, 2016). Little attention has been paid to research on the loss function ℓ, and its interaction with the other ingredients $\mathscr{F}$ and P. A recent exception is (van Erven et al., 2015) which showed how the *joint* interaction of ℓ, $\mathscr{F}$ and P control the speed of convergence of learning algorithms. The lack of attention is surprising because the choice of loss function matters, especially when (as is typical) one has limited data, and so one cares about the speed of convergence of empirical estimates, and the best model in the class $\mathscr{F}$ has non-zero expected loss (again typically the case). Understanding the implications and options for the choice of loss function also matters when one considers the integration of machine learning technologies into larger socio-technical systems, since the loss function serves as an abstraction of what matters at the larger system level, and can be used, for example, to abstract a range of notions of fairness in ML problems (Menon and Williamson, 2018a).

Loss functions are central to statistical decision problems, and have a long history (Wald, 1950); see (Vernet et al., 2016; Williamson, 2013) for some pointers to the literature. The present paper focusses upon understanding at a deeper level the loss functions for multiclass probability estimation, and their transformations. Our results are, to use the apposite term of Rota (1997), "cryptomorphic" — an isomorphism that was previously hidden from view, which once decoded is illuminating. Our approach is parametrisation independent in the sense of the distinction made in (van Erven et al., 2015; Vernet et al., 2016) (in essence, what matters is the geometry of the set induced by the loss function which does not change under reparametrization).

Losses in machine learning play a role analogous to metrics in other applied problems. Menger (1928) introduced distance geometry (in order to view the world in terms of distances) and there is an incredible variety of distances to choose from (Deza and Deza, 2009). But as we shall see, it is the simpler notion of norms, and normed spaces, that are the most relevant in the study of losses. The development of functional analysis critically depended upon the development of finite dimensional normed spaces by Minkowski (1896). In his *Geometrie der Zahlen*, he developed the notion of a symmetric convex body and its equivalence to a norm ball $\{x \mid p(x) \leq 1\}$, as well as introducing the notion of a supporting hyperplane and the corresponding support function. He showed[2] it was the dual to the norm $p(x)$. We shall see that these concepts that were central to the development of

1. The claim that all the analyst brings is the model class $\mathscr{F}$ is a simplification that captures much of ML; in general the analyst provides a learning "algorithm" (a function) $A\colon S \mapsto f \in \mathscr{F}$ which given a sample S produces an f (Herbrich and Williamson, 2002). For the purposes of the present paper the simplification stated in the main text suffices.
2. See (Martini et al., 2001, Section 2) and (Thompson, 1996, Section 1.5) for a more detailed history. The extension of these concepts to infinite dimensional spaces underpinned the development of functional analysis; Dieudonné (1981, page 130) credits Helly (1921) with the idea of abstracting away from *particular* spaces such as ℓ^p, $\mathscr{L}^p$ or $\mathrm{C}([a,b])$ to the notion of *general* normed sequence spaces by methods which do not depend upon special features of the space. While apparently rather elementary, these finite dimensional normed spaces ("Minkowski Spaces" (Thompson, 1996)) underpin the general theory of Banach spaces. Pietsch (2007, Page xxii) quotes Dvoretzky (1960) inspired by Grothendieck: "many problems in the theory of Banach spaces may be reduced to the finite-dimensional case, i.e. to problems concerning Minkowski spaces."

normed spaces are, with minor modification, the foundation for an understanding of loss functions. Recapitulating history, we concentrate in this paper on the finite dimensional case, corresponding to multiclass classification and conditional probability estimation problems.

The rest of the paper is organised as follows. Section 2 introduces the mathematical machinery we utilise. Section 3 introduces (proper) loss functions, including the antipolar loss which is a natural "inverse" of a loss function. Section 4 presents examples, illustrating the new perspective, and the antipolar loss in particular. Section 5 presents some design strategies for loss functions in terms of their superprediction sets. Section 6 shows how to construct new proper losses from old ones by suitable combination of their superprediction sets. These results are based on the new results in Section 6.8 which substantially extend the theory of M-sums of convex sets, including a general duality result for M-sums of norms and anti-norms. Section 7 concludes.

1.1 Motivation, Expectations, Context and Significance

The goals and results of this paper are different in nature to those of the majority of papers in machine learning[3]. To that end, we give some context and set expectations. The paper contains no new algorithms and no experimental results. What it does contain is a new way of looking at loss functions which 1) illustrates the close connection between losses and norms and anti-norms; 2) presents the new idea of an antipolar loss; 3) develops a calculus for loss functions that allows multiple proper loss functions to be combined in a manner that the resulting loss is guaranteed proper; and 4) shows how the geometrical perspective can be used to design loss functions.

Why embark on this complex endeavour? Currently loss functions are widely used, but there is little insight to be had regarding the consequences of particular choices. This is especially true when these functions are identified with their algebraic formulas. There were insights derived by Reid and Williamson (2011) for the design of loss functions (following Hand and Vinciotti (2003)), and in (Menon and Williamson, 2018b, Appendix B), but these approaches, whilst tractable enough in the binary case, become intractable for the multiclass situation. As we will show, there is an intrinsic geometry to loss functions which controls the nature of the learning problem at hand. Evidence for this was already given by van Erven et al. (2012); Pacheco and Williamson (2023) who showed how the mixability constant of a loss (which appears in bounds for the regret in online learning) is directly controlled by the intrinsic geometry of the loss function.

Some of the value of the viewpoint developed in the paper is only realised in the companion paper (Williamson and Cranko, 2022) which uses the geometric approach developed here to derive, in a much simpler manner, the bridge between loss functions and measures of information that was previously presented by Reid and Williamson (2011) (binary case) and Garcia-Garcia and Williamson (2012) (multiclass case). In (Williamson and Cranko, 2022) we show that the geometric way of viewing information measures allows one to derive results seemingly unobtainable by others means. In particular, we derive a general data processing *equality* from which one can derive the classical strong data processing inequality. It turns out that the geometric viewpoint is central to these novel results.

The new perspective has been used by Kamalaruban et al. (2015) to show the connection between exp-concavity and mixability, which is relevant to online learning algorithms, as well as to the

3. But not different in nature to many papers in economics. Indeed, economists have, over a long period, conducted investigations on the foundations of their discipline (utilities). As we shall see below in footnote 12, the similarity turns out not to be just in style, but there is a remarkable parallel in content as well.

understanding of fast rates in statistical learning (van Erven et al., 2015). It was used by Mhammedi and Williamson (2018) to solve an open question regarding generalised mixability as well as to draw a connection to mirror descent. It also underpins the results of Cranko (2021), which develops the theory of proper composite losses in an infinite dimensional setting. The present paper is a substantially extended and improved version of (Williamson, 2014). Beyond the correction of some errors, the present paper fully develops the theory of M-sums of superprediction sets in a general and rigorous manner.

2. Preliminaries

We introduce some standard machinery from the theory of convex sets and functions (see Hiriart-Urruty and Lemaréchal, 2001; Rockafellar, 1970; Rockafellar and Wets, 2004; Schneider, 2014; Penot, 1997)[4]. The concave cases of some of these results can be found in the works of Pukelsheim (1983) and Barbara and Crouzeix (1994). In choosing our notational conventions, we have adopted notation more common in the mathematical literature, even though some of this may be unfamiliar to a machine learning audience (since we refer to the mathematical literature for a number of the results upon which we build).

2.1 Basic Notation

Let $\mathscr{X}$ be a finite dimensional Euclidean space over the reals. The space of linear functionals on $\mathscr{X}$ is $\mathscr{X}^*$, and the natural coupling is $\langle\cdot,\cdot\rangle\colon \mathscr{X}^*\times\mathscr{X}\to\mathbb{R}$; the usual inner product. Define the special sets $\mathbb{R}_{>0}\coloneqq(0,\infty)$; $\mathbb{R}_{\geq 0}\coloneqq[0,\infty)$; $\mathbb{R}_{<0}\coloneqq(-\infty,0)$; $\mathbb{R}_{\leq 0}\coloneqq(-\infty,0]$; $\overline{\mathbb{R}}\coloneqq[-\infty,+\infty]$; $\overline{\mathbb{R}}_{\geq 0}\coloneqq[0,+\infty]$. Denote the cardinality of a set $S\subset\mathscr{X}$ by $|S|$. If $S=\{T_\alpha\,|\,\alpha\in A\}$ is a set of sets, then $\bigcap S\coloneqq\bigcap_{\alpha\in A}T_\alpha$ and $\bigcup S\coloneqq\bigcup_{\alpha\in A}T_\alpha$. We refer to the components of $x\in\mathscr{X}$ by x_i and $x=(x_1,\ldots,x_n)$. Let $\Delta(S)$ denote the set of probability measures on a set S, and $[n]\coloneqq\{1,2,\ldots,n\}$; then $\Delta([n])\simeq\left\{x\in\mathbb{R}^n_{\geq 0}\;\middle|\;\sum_{i=1}^n x_i=1\right\}$. Let $(\mathrm{e}^i)_{i\in[n]}$ be the canonical basis vectors in $\mathbb{R}^n$. The family of p-norms (with $p\in[1,\infty]$) on the space $\mathscr{X}$ are defined by $\|x\|_p\coloneqq\left(\sum_{i\in[n]}|x_i|^p\right)^{1/p}$ for finite p, and $\|x\|_\infty\coloneqq\max_{i\in[n]}|x_i| \eqqcolon \bigvee_{i\in[n]}x_i$. The *p-unit ball* is $\mathrm{B}_p\coloneqq\{x\in\mathscr{X}\;|\;\|x\|_p\leq 1\}$, and if there is no subscript we take $\mathrm{B}\coloneqq\mathrm{B}_2$. The Iverson bracket $[\![\cdot]\!]$ takes a proposition and returns 1 if it is true, and 0 otherwise. We use the common conventions $\inf(\varnothing)\coloneqq+\infty$, $\sup(\varnothing)\coloneqq-\infty$, $\infty\cdot 0\coloneqq 0$ and $1/0\coloneqq\infty$. If $v\in\mathbb{R}^n$, then v' denotes its transpose. The all ones vector is defined as $1_n\coloneqq(1,\ldots,1)'\in\mathbb{R}^n$.

2.2 Convex Sets

Let $S,T\subseteq\mathscr{X}$, $x\in\mathscr{X}$, $\alpha\in\mathbb{R}$ and $U\subset\mathbb{R}$. Let $\alpha S\coloneqq\{\alpha s\,|\,s\in S\}$, $U\cdot S\coloneqq\{\alpha s\,|\,\alpha\in U,\ s\in S\}$, $S+x\coloneqq\{s+x\,|\,s\in S\}$. The *Minkowski sum* is $S+T\coloneqq\{s+t\,|\,s\in S \text{ and } t\in T\}$. For $\varnothing\subset S\subseteq\mathscr{X}$, $\mathrm{cl}(S)$ and $\overline{S}$ both denote its *closure*. The collection of *closed, nonempty, convex subsets of S* is $\mathcal{K}(S)$. The *interior* and *boundary* of S are

$$\mathrm{int}(S)\coloneqq\{x\in S\,|\,\exists\varepsilon>0,\ (\varepsilon\mathrm{B}+x)\subseteq S\}\quad\text{and}\quad\mathrm{bd}(S)\coloneqq S\setminus\mathrm{int}(S).$$

4. We recognise that there is a significant quantity of background material needed before we get to the machine learning problem and the results about loss functions. But this really illustrates the point of the paper: the deeper structure of loss functions arises from more fundamental geometrical concepts. And while some of the material in this section is widely known, the results for concave gauges and their polar duals, which are central to the analysis of loss functions, are both less well known and *not* a trivial variation of the convex case.

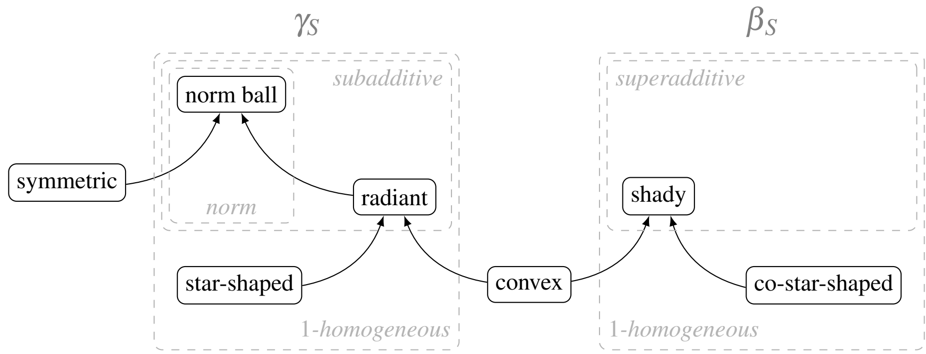


FIGURE 1: The relationship between various classes of sets $S \subseteq \mathcal{X}$ defined in §2.3, and the corresponding properties of the gauge γ_S and antigauge β_S defined in §2.7.

If S is convex its *relative interior* and *relative boundary* are

$$\operatorname{ri}(S) := \{x \in S \,|\, \forall y \in S,\ \exists \lambda > 1,\ \lambda x + (1-\lambda)y \in S\} \ \text{ and } \ \operatorname{rbd}(S) := S \setminus \operatorname{ri}(S). \tag{1}$$

Its *convex hull* is $\operatorname{co}(S) := \bigcap\{T \subseteq \mathcal{X} \,|\, S \subseteq T \text{ and } T \text{ convex}\}$; its *conic hull* is $\operatorname{cone}(S) := \bigcup_{t>0} tS$. For the closure of these operations we sometimes write $\operatorname{cl}\operatorname{co}(S) := \operatorname{cl}(\operatorname{co} S)$, and $\operatorname{cl}\operatorname{cone}(S) := \operatorname{cl}(\operatorname{cone} S)$.

2.3 Starry, Radiant and Shady Sets

A nonempty, proper subset $S \subset \mathcal{X}$ is:

- *star-shaped* if $(0,1] \cdot S \subseteq S$ and $0 \in \operatorname{int} S$;
- *co-star-shaped* if $[1,\infty) \cdot S \subseteq S$ and $0 \notin S$;
- *radiant* if it is star-shaped and convex;
- *shady* if it is co-star-shaped and convex.

By convention the empty set is star-shaped (and radiant), and the entire space is co-star-shaped. Thus the star-shaped sets are the complements of the co-star-shaped sets and vice versa. If $-S = S$ we say S is *symmetric*; if S is symmetric and radiant we say it is a *norm ball*. Let $\mathbb{R}(X)$ and $\mathbb{S}(X)$ denote, respectively, the collections of *closed radiant* and *closed shady subsets* of $X \subseteq \mathcal{X}$. These definitions are illustrated in Figure 1.

2.4 Cones and Recession Cones

A set $C \subseteq \mathcal{X}$ is said to be a *cone* if $\mathbb{R}_{>0} \cdot C \subseteq C$. A cone C is *pointed* if $0 \in C$; *salient* if $x, -x \in C$ implies $x = 0$; and *blunt* if $0 \notin C$. Every closed cone is pointed. Every blunt, convex cone is salient, but this is not the case for pointed convex cones. If a convex cone C is salient, then $C \setminus \{0\}$ is also a convex cone. For a cone $C \subseteq \mathcal{X}$ there is a natural counterpart $C^* \subseteq \mathcal{X}^*$ called the *dual cone*, where

$$C^* := \{x^* \in \mathcal{X}^* \mid \forall x \in C,\ \langle x^*, x\rangle \geq 0\}. \tag{2}$$

A pointed convex cone $C \subseteq \mathcal{X}$ induces a partial ordering on $\mathcal{X}$ which we denote $\succeq_C$; for $x, y \in \mathcal{X}$ we say $x \succeq_C y$ if $x - y \in C$. For a set $S \subseteq \mathcal{X}$ we say $d \in \mathcal{X}$ is a *recession direction of S* if $S + d \subseteq S$. The

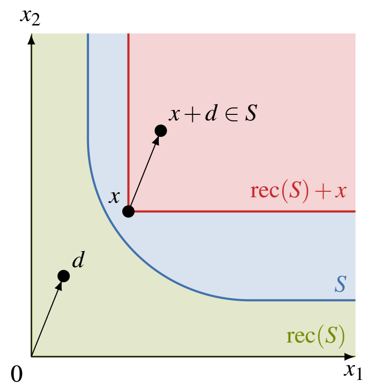


FIGURE 2: An illustration of the set $S \subseteq \mathcal{X}$ (which extends infinitely far to the "north-east", and thus only a finite portion is illustrated, a convention which we adopt throughout the paper) and its recession cone $\mathrm{rec}(S) = \mathbb{R}^2_{\geq 0}$. For all $x \in S$ and all $d \in \mathrm{rec}(S)$ the point $x+d$ is contained within S.

collection of recession directions of S is called its *recession cone* which we denote by

$$\mathrm{rec}(S) \coloneqq \{d \in \mathcal{X} \mid S + d \subseteq S\}. \tag{3}$$

The recession cone is illustrated in Figure 2. If $S \in \mathcal{K}(\mathcal{X})$ then it is easily verified that $\mathrm{rec}(S)$ is indeed a closed convex cone. We say a set $S \subseteq \mathcal{X}$ is *C-oriented* if $\mathrm{rec}(S) = C$. Unless otherwise stated we use X_+ to denote a salient, closed, convex cone that is a proper subset of $\mathcal{X}$: $X_+ \subset \mathcal{X}$.

Proposition 1 (Calculus of Recession Cones) *Let $S \in \mathcal{K}(\mathcal{X})$. The following hold:*

1. $\mathrm{rec}(S) = \{0\}$ *if and only if S is bounded;*
2. $\mathrm{rec}(S)$ *is a closed convex cone, thus* $S + \mathrm{rec}(S) = S$*;*
3. $d \in \mathrm{rec}(S)$ *if and only if there exist sequences* $(x_n)_{n\in\mathbb{N}}$, $x_n \in S$, *and* $(t_n)_{n\in\mathbb{N}} \searrow 0$, $t_n \in \mathbb{R}_{\geq 0}$, *such that* $x_n t_n \to d$*;*
4. *if* $C \subseteq \mathcal{X}$ *is a convex cone then* $\mathrm{rec}(C) = \mathrm{cl}(C)$*;*
5. *if* $A \subseteq B \subseteq \mathcal{X}$ *then* $\mathrm{rec}(A) \subseteq \mathrm{rec}(B)$.

If $(S_i)_{i\in I}$ *with* $S_i \in \mathcal{K}(\mathcal{X})$ *is a family with an arbitrary index set I with* $\bigcap_{i\in I} S_i \neq \varnothing$, *then*

6. $\mathrm{rec}(\bigcap_{i\in I} S_i) = \bigcap_{i\in I} \mathrm{rec}\, S_i$*;*
7. $\mathrm{rec}(\bigcup_{i\in I} S_i) \supseteq \bigcup_{i\in I} \mathrm{rec}\, S_i$.

If $[m] \subseteq I$ *is a finite subcollection, then*

8. $\mathrm{rec}\left(\sum_{i\in[m]} S_i\right) \supseteq \sum_{i\in[m]} \mathrm{rec}(S_i)$.

Proof These are all well-known and can be found in a variety of common references (Auslender and Teboulle, 2003; Rockafellar, 1970; Rockafellar and Wets, 2004). ■

2.5 Convex, Concave and Homogeneous Functions

For the remainder of this section let $f\colon \mathscr{X} \to \overline{\mathbb{R}}$; we define its *domain*, *epigraph* and *hypograph* respectively as

$$\begin{aligned}\operatorname{dom}(f) &:= \{x \in \mathscr{X} \mid f(x) \in \mathbb{R}\},\\ \operatorname{epi}(f) &:= \{(x,t) \in \operatorname{dom} f \times \mathbb{R} \mid f(x) \le t\},\\ \operatorname{hyp}(f) &:= \{(x,t) \in \operatorname{dom} f \times \mathbb{R} \mid f(x) \ge t\}.\end{aligned}$$

Let $\alpha \in \mathbb{R}$. The *below*, *level* and *above sets* are, respectively,

$$\begin{aligned}\operatorname{lev}_{\le\alpha}(f) &:= \{x \in \operatorname{dom}(f) \mid f(x) \le \alpha\},\\ \operatorname{lev}_{=\alpha}(f) &:= \{x \in \operatorname{dom}(f) \mid f(x) = \alpha\},\\ \operatorname{lev}_{\ge\alpha}(f) &:= \{x \in \operatorname{dom}(f) \mid f(x) \ge \alpha\}.\end{aligned}$$

We say f is *convex* if it satisfies

$$\forall x,y \in \operatorname{dom} f\ \forall t \in (0,1),\ f(tx+(1-t)y) \le tf(x)+(1-t)f(y).$$

If $-f$ is convex, then f is *concave*. Equivalently f is convex (resp. concave) if $\operatorname{epi}(f)$ is convex (resp. $\operatorname{hyp}(f)$ is convex). We say f is *quasi-convex* (resp. *quasi-concave*) if $\operatorname{lev}_{\le\alpha} f$ (resp. $\operatorname{lev}_{\ge\alpha} f$) is convex for all $\alpha \in \mathbb{R}$. We say a convex (resp. concave) function f is *closed* if $\operatorname{epi}(f)$ (resp. $\operatorname{hyp}(f)$) is closed. Thus if f is closed and convex, $-f$ is closed and concave. The *closure* of a convex (resp. concave) function f is the convex (resp. concave) function g such that $\operatorname{epi}(g) = \operatorname{cl}(\operatorname{epi} f)$ (resp. $\operatorname{hyp}(g) = \operatorname{cl}(\operatorname{hyp} f)$). The function g is denoted by $\operatorname{cl}(f)$. Finally for a set $S \subseteq \mathscr{X}$, define $\arg\sup_{x\in S} f(x) := \{x \in \mathscr{X} \mid f(x) = \sup_{x\in S} f(x)\}$ and $\arg\inf_{x\in S} f(x) := \{x \in \mathscr{X} \mid f(x) = \inf_{x\in S} f(x)\}$; either of these sets can be empty.

If for some $k \in \mathbb{R}$, f satisfies $f(\alpha x) = \alpha^k f(x)$ for $\alpha > 0$ and for all x, we say f is *homogeneous of degree k* or *k-homogeneous* (there is an obvious extension to set-valued functions). If f is 1-homogeneous, we also say f is *positively homogeneous*. If for all $x,y \in \operatorname{dom}(f)$ we have $f(x+y) \le f(x)+f(y)$ then f is called *subadditive*. If $-f$ is subadditive then f is called *superadditive*. If f is positively homogeneous and subadditive (resp. superadditive) then we say f is *sublinear* (resp. *superlinear*). All sublinear functions are convex and all superlinear functions are concave. Suppose $f_1,\dots,f_m\colon \mathscr{X} \to \overline{\mathbb{R}}$. Their *infimal convolution* is the function $\mathscr{X} \to \overline{\mathbb{R}}$ defined by

$$(f_1 \,\square \cdots \square\, f_m)(x) := \inf\{f_1(x_1)+\cdots+f_m(x_m) \,|\, x_1+\cdots+x_m = x\}. \tag{4}$$

2.6 Support Functions, Subdifferentials and Superdifferentials

For a set $S \subseteq \mathscr{X}$ we define its *convex support function*

$$\mathscr{X}^* \ni x^* \mapsto \sigma_S(x^*) := \sup\{\langle x^*, x\rangle \mid x \in S\} \in \overline{\mathbb{R}}. \tag{5}$$

However, in our setting, it will often be more natural to consider the *concave support function*

$$\mathscr{X}^* \ni x^* \mapsto \rho_S(x^*) := \inf\{\langle x^*, x\rangle \mid x \in S\} \in \overline{\mathbb{R}}. \tag{6}$$

The convex and concave support functions are related as follows:

$$\forall x^* \in \mathcal{X}^*,\ \sigma_S(x^*) = \sup_{x\in S}\langle x^*, x\rangle = \sup_{x\in -S}\langle x^*, -x\rangle = -\inf_{x\in -S}\langle x^*, x\rangle = -\rho_{-S}(x^*). \tag{7}$$

It is easy to see that σ and ρ are both 1-homogeneous, σ is subadditive and thus sublinear; and ρ is superadditive and thus superlinear.

We introduce the mappings $\partial_{\prime} f, \partial' f \colon \mathcal{X} \to 2^{\mathcal{X}^*}$ with

$$\begin{aligned}
\mathcal{X} \ni x &\mapsto \partial_{\prime} f(x) := \{x^* \in \mathcal{X}^* \mid \forall y \in \mathcal{X},\ f(y) - f(x) \geq \langle x^*, y - x\rangle\}, \\
\mathcal{X} \ni x &\mapsto \partial' f(x) := \{x^* \in \mathcal{X}^* \mid \forall y \in \mathcal{X},\ f(y) - f(x) \leq \langle x^*, y - x\rangle\}.
\end{aligned} \tag{8}$$

The first mapping is the classical *subdifferential*, and the second is the less-common *concave subdifferential*, or *superdifferential*. These sets are related by $\partial_{\prime}(f) = -\partial'(-f)$. The mapping

$$\mathcal{X} \ni x \mapsto \partial f(x) := \partial_{\prime} f(x) \cup \partial' f(x)$$

is known as the *symmetric subdifferential* (Mordukhovich and Shao, 1995). When f is convex $\partial f = \partial_{\prime} f$, and when f is concave $\partial f = \partial'(f)$. In the event that f is both convex and concave $\partial f = \partial_{\prime} f = \partial' f$. More importantly, the symmetric subdifferential satisfies $\partial(-f) = -\partial f$, which makes it a convenient choice for us since we deal with (sub/super)-differentials of both convex and concave functions. We refer to elements of ∂f as *subgradients* (recognising the slight terminological abuse in the choice of name) and write $\mathrm{dom}(\partial f) := \{x \in \mathcal{X} \mid \partial f \neq \varnothing\}$. If there is a function $g \colon \mathcal{X} \to \mathcal{X}^*$ that is always a subgradient of f in the sense that for all $x \in \mathrm{dom}(\partial f)$ we have $g(x) \in \partial f(x)$, then g is called a *selection of ∂f* and we write $g \in \partial f$. The following proposition is a standard result (see Penot, 2012; Bauschke and Combettes, 2011):

Lemma 2 *Let $f \colon \mathcal{X} \to \overline{\mathbb{R}}$ be convex with nonempty domain. Then* $\mathrm{ri}(\mathrm{dom}\, f) \subseteq \mathrm{dom}(\partial f)$.

It is easy to show that for some convex functions the inclusion in Lemma 2 is not strict; For example take $\partial\langle\cdot, s\rangle$, then $\mathrm{dom}(\partial\langle\cdot, s\rangle) = \mathrm{dom}(\langle\cdot, s\rangle)$.

Proposition 3 *Suppose $f \colon \mathcal{X} \to \overline{\mathbb{R}}$ is 1-homogeneous. Then ∂f is 0-homogeneous.*

Proof From the definition of the subdifferential, for all $x \in \mathcal{X}$ and all $\alpha > 0$,

$$\begin{aligned}
\partial_{\prime} f(\alpha x) &= \{x^* \in \mathcal{X}^* \mid \forall y \in \mathcal{X},\ f(y) - f(\alpha x) \geq \langle x^*, y - \alpha x\rangle\} \\
&= \{x^* \in \mathcal{X}^* \mid \forall \alpha y \in \mathcal{X},\ f(\alpha y) - f(\alpha x) \geq \langle x^*, \alpha y - \alpha x\rangle\} \\
&= \{x^* \in \mathcal{X}^* \mid \forall y \in \mathcal{X},\ \alpha f(y) - \alpha f(x) \geq \alpha\langle x^*, y - x\rangle\} \\
&= \partial_{\prime} f(x),
\end{aligned}$$

and $\partial_{\prime} f$ is thus 0-homogenous. The proof for the superdifferential is similar. ■

Lemma 4 (Zălinescu, 2013, Corollary 3) *Assume $C \in \mathcal{K}(\mathcal{X})$. Then σ_C is differentiable on* $\mathrm{dom}(\partial\sigma_C) \setminus \{0\}$ *if and only if either C is a singleton or* $\mathrm{int}(C) \neq \varnothing$ *and C is strictly convex.*

Lemma 4 together with (7) gives us the corollary for the concave case:

Corollary 5 *Assume $C \in \mathcal{K}(\mathcal{X})$. Then ρ_C is differentiable on $\operatorname{dom}(\partial \rho_C) \setminus \{0\}$ if and only if either C is a singleton or $\operatorname{int}(C) \neq \varnothing$ and C is strictly convex.*

The *gradient operator* on $\mathcal{X}$ is $\nabla := (\partial x_1, \ldots, \partial x_n)$. The following lemma has an obvious extension for the concave case:

Lemma 6 (Rockafellar, 1970, Theorem 25.1) *Let $f \colon \mathcal{X} \to \overline{\mathbb{R}}$ be convex. If f is differentiable at $x \in \operatorname{dom}(f)$, then $\partial f(x) = \{\nabla f(x)\}$. Conversely, if $\partial f(x)$ is a singleton at $x \in \operatorname{dom}(f)$ then f is differentiable at x.*

Support functions are oblivious to closures and convex hulls:

Lemma 7 (Hiriart-Urruty and Lemaréchal, 2001, Proposition C.2.2.1) *Suppose $S \subseteq \mathcal{X}$ is nonempty. Then $\sigma_S = \sigma_{\operatorname{cl}(S)} = \sigma_{\operatorname{co}(S)}$; whence $\sigma_S = \sigma_{\operatorname{cl}\operatorname{co}(S)}$.*

Using (7) we find:

Corollary 8 *Suppose $S \subseteq \mathcal{X}$ is nonempty. Then $\rho_S = \rho_{\operatorname{cl}(S)} = \rho_{\operatorname{co}(S)}$; whence $\rho_S = \rho_{\operatorname{cl}\operatorname{co}(S)}$.*

Lemma 9 *Let $S \subseteq \mathcal{X}$, $S \neq \varnothing$. Then $\overline{\operatorname{dom}(\sigma_S)} = -(\operatorname{rec}(\operatorname{cl}\operatorname{co} S))^*$ and $\overline{\operatorname{dom}(\rho_S)} = \operatorname{rec}(\operatorname{cl}\operatorname{co} S)^*$.*

Proof Firstly $\sigma_S = \sigma_{\operatorname{cl}\operatorname{co} S}$ (Lemma 7). Then from (Auslender and Teboulle, 2003, Theorem 2.2.1, p. 32) $-(\operatorname{dom}\sigma_S)^* = \operatorname{rec}(\operatorname{cl}\operatorname{co} S)$. Thus since $\operatorname{dom}(\sigma_S)$ is convex, $\overline{\operatorname{dom}(\sigma_S)} = (-\operatorname{rec}(\operatorname{cl}\operatorname{co} S))^* = -(\operatorname{rec}(\operatorname{cl}\operatorname{co} S))^*$. For the concave case using (7) we have $\operatorname{dom}(-\rho_{\operatorname{cl}\operatorname{co} S}) = \operatorname{dom}(\sigma_{-\operatorname{cl}\operatorname{co} S}) = (-\operatorname{rec}(-\operatorname{cl}\operatorname{co} S))^* = \operatorname{rec}(\operatorname{cl}\operatorname{co} S)^*$. ■

2.7 Gauge Functions

The theory of gauges (Minkowski functionals) and polars has been traditionally developed for radiant sets (see Hiriart-Urruty and Lemaréchal, 2001; Rockafellar, 1970; Schneider, 2014; Thompson, 1996), whereas the theory of gauges for shady sets is less well known (Rockafellar, 1967; Pukelsheim, 1983; Barbara and Crouzeix, 1994; Penot and Zălinescu, 2000). However gauge functions for shady sets have been used in statistics in a manner similar to that which we will use them (Pukelsheim, 1983) and in economics; see (e.g. Hasenkamp and Schrader, 1978) and footnote 12 below.

For convex sets the support function is a natural object to consider. Likewise, when working with star-shaped and co-star-shaped sets the *gauge* and *anti-gauge* are a natural parallel. For a set $S \subseteq \mathcal{X}$ we define its gauge and anti-gauge:

$$\mathcal{X} \ni x \mapsto \gamma_S(x) := \inf\{\lambda > 0 \mid x \in \lambda S\} \in \overline{\mathbb{R}}, \tag{9}$$

$$\mathcal{X} \ni x \mapsto \beta_S(x) := \sup\{\lambda > 0 \mid x \in \lambda S\} \in \overline{\mathbb{R}}. \tag{10}$$

If S is closed and radiant, then γ_S is closed and sublinear (Penot and Zălinescu, 2000, Proposition 2.3). Alternately, if S is closed and shady, then β_S is closed and superlinear (Penot and Zălinescu, 2000, Proposition 2.4). Thus γ_S is a *convex gauge* and β_S a *concave gauge*, as they are sometimes described in the literature. We list some properties of gauge functions and their associated sets in Table 1 and graphically in Figure 1. For closed S, the base star-shaped and co-star-shaped sets can be recovered

	Gauge function γ_S			
S	Non-negative	1-Homogeneous	Subadditive	Norm
star-shaped	•	•		
radiant	•	•	•	
norm ball	•	•	•	•

	Anti-gauge function β_S		
S	Non-negative	1-Homogeneous	Superadditive
co-star-shaped	•	•	
shady	•	•	•

TABLE 1: Properties of gauge and anti-gauge functions when restricted to their domains, as determined by their defining sets. See also Figure 1.

with the inverse mappings $\gamma_S \mapsto \mathrm{lev}_{\leq 1}(\gamma_S) = S$ and $\beta_S \mapsto \mathrm{lev}_{\geq 1}(\beta_S) = S$ respectively. Mirroring Lemma 9, Penot and Zălinescu (2000) observed that when S is closed

$$\mathrm{dom}(\gamma_S) = \mathrm{cone}(S) \cup \{0\} \subseteq \mathrm{rec}(S) \quad \text{and} \quad \mathrm{dom}(\beta_S) = \mathrm{cone}(S) \cup \{0\} \subseteq \mathrm{rec}(S).$$

As one might expect

$$\mathrm{dom}(\gamma_S) = \mathrm{cl}\,\mathrm{cone}(S) \quad \text{and} \quad \mathrm{dom}(\beta_S) = \mathrm{cl}\,\mathrm{cone}(S).$$

If $0 \in \mathrm{int}\, S$, then $\mathrm{dom}(\gamma_S) = \mathscr{X}$ for all star-shaped sets S. Conversely, for any co-star-shaped set $S \subset \mathscr{X}$ we have $\mathrm{cone}(S) \subset \mathscr{X}$. In general this is the key difference between gauge and anti-gauge functions. That is, while gauge functions can be finite on the whole space, anti-gauges are usually finite only on a conic subset of $\mathscr{X}$. This is significant with regard to the equivalence of gauges or norms. Recall two gauge functions $\gamma_{T_1}, \gamma_{T_2} \colon \mathscr{X} \to \mathbb{R}$ are equivalent if there exists constants $c, C \in \mathbb{R}_{>0}$ such that for all $x \in \mathscr{X}$, $c\gamma_{T_2}(x) \leq \gamma_{T_1}(x) \leq C\gamma_{T_2}(x)$. Whilst all convex gauges and norms in finite dimensional space are equivalent, that is *not* true in general for concave gauges or anti-norms $\beta_{S_1}, \beta_{S_2} \colon \mathscr{X} \to \overline{\mathbb{R}}$ if at least one of the concave gauges can take on the value $+\infty$ for some $x \in \mathscr{X}$. Unbounded concave gauges correspond to unbounded loss functions; a point which will be elaborated below.

The attentive reader will notice the similarity between the properties of the gauge of a norm ball and the properties of a norm on $\mathscr{X}$. Indeed every norm on $\mathscr{X}$, $\|\cdot\|$, can be written as a gauge of the norm ball $\mathrm{lev}_{\leq 1}\|\cdot\|$, and conversely the gauge of every norm ball, as defined in §2, is a norm. If one restricts analysis to the set $\mathrm{cone}(S)$ for a shady set $S \subset \mathscr{X}$—in effect ruling out multiplication by negative scalars—the function $\beta_S \colon \mathrm{cone}(S) \to \mathbb{R}$ is a natural counterpart to a norm on this space, which we call an *anti-norm*. As we shall see in §3, the conditional Bayes' risks associated with proper losses are in fact anti-norms[5].

5. Anti-gauges are sometimes called "anti-norms" (Berestovskiĭ and Gichev, 2004; Moszyńska and Richter, 2012; Merikoski, 1991), although confusingly this name is sometimes used to refer to the dual (polar) of a traditional norm (Horváth et al., 2017; Martini and Swanepoel, 2006). We will thus stick with the terminology "anti-gauge".

2.8 Legendre-Fenchel Conjugates

The *Legendre–Fenchel conjugate* or *convex conjugate* of f is the function

$$\mathscr{X}^* \ni x^* \mapsto f^*(x^*) := \sup_{x \in \mathscr{X}} (\langle x^*, x \rangle - f(x)) \in \overline{\mathbb{R}}. \tag{11}$$

In addition to the more common convex Legendre–Fenchel conjugate, we will make use of the concave conjugate, which, like the case of the concave support function, will be more appropriate for our purposes. The *concave conjugate* of f is

$$\mathscr{X}^* \ni x^* \mapsto f_*(x^*) := \inf_{x \in \mathscr{X}} (\langle x^*, x \rangle - f(x)) \in \overline{\mathbb{R}}, \tag{12}$$

and is related to the convex conjugate as follows:

$$f^*(x^*) = \sup_{x \in \mathscr{X}} (\langle x^*, x \rangle - f(x)) = - \inf_{x \in \mathscr{X}} (\langle -x^*, x \rangle - (-f)(x)) = -(-f)_*(-x^*).$$

The concave conjugate therefore satisfies a *reverse Fenchel–Young inequality*:

$$\forall x \in \mathscr{X}, \; \forall x^* \in \mathscr{X}^*, \; f(x) + f_*(x^*) \leq \langle x^*, x \rangle.$$

A function f is *lower semi-continuous* if for all $x \in \mathscr{X}$ and for all sequences $(x_n)_{n \in \mathbb{N}}$ with $x_n \in \mathscr{X}$ and $x_n \to x$ we have $f(x) \leq \liminf_{n \to \infty} f(x_n)$. If $-f$ is lower semi-continuous then f is said to be *upper semi-continuous*.

2.9 Polar Duality

While we do make some use of the above Legendre-Fenchel duality, more important for our purposes is the polar duality of convex sets (Moszyńska, 2006, Chapter 13). It arises from the classical polarity between points and lines relative to the unit circle in inversive geometry (Askwith, 1917). Following Berger (2010, §VII.5.C), define a bijection between the set of all points other than the origin onto the set of hyperplanes not containing the origin via $x \mapsto h_x := \{y \in \mathbb{R}^n \,|\, \langle x, y \rangle = 1\}$, which is known as the polar hyperplane of the point x. Associated with h_x is the halfspace $H_x := \{y \in \mathbb{R}^n \,|\, \langle x, y \rangle \leq 1\}$. Given a set $C \subset \mathbb{R}^n$ containing the origin, the *polar* of C is simply $\bigcap_{x \in C} H_x$.

More generally, there is a bijection between the sublinear functions σ_A and closed convex sets A; and another bijection between the superlinear functions ρ_B and closed convex sets B Hiriart-Urruty and Lemaréchal (2001). Noting Table 1 we see that the gauge and anti-gauge functions also satisfy the same criteria as the support functions when restricted to (respectively) the radiant and shady subsets. A natural question to ask then is: can we find sets A' and B' such that $\sigma_{A'} = \gamma_A$ and $\rho_{B'} = \beta_B$? The answer in the convex case with respect to the radiant sets is well known, but there is an equally rich, parallel structure in the concave case with respect to the shady sets. Penot (2012) and Zălinescu (2002) show the following results for the convex case, and Penot and Zălinescu (2000) prove the concave case.

For a set $S \subseteq \mathscr{X}$ its *polar* and *antipolar* are the sets

$$S^\circ := \{x^* \in \mathscr{X}^* \,|\, \forall x \in S, \; \langle x^*, x \rangle \leq 1\} \;\text{ and }\; S^\diamond := \{x^* \in \mathscr{X}^* \,|\, \forall x \in S, \; \langle x^*, x \rangle \geq 1\}. \tag{13}$$

Equivalently $S^\circ = \operatorname{lev}_{\leq 1} \rho_S$ and $S^\diamond = \operatorname{lev}_{\geq 1} \rho_S$. It is easy to show that the polarity operation $S \mapsto S^\circ$ takes closed radiant sets to closed radiant sets, and the antipolarity operation $S \mapsto S^\diamond$ takes closed

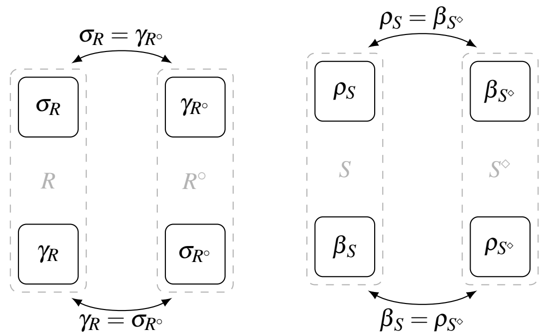


FIGURE 3: The polar duality of support functions and gauge functions for radiant (R) and shady (S) sets.

shady sets to closed shady sets such that if $S \in \mathcal{S}(X_+)$, then $S^\diamond \in \mathcal{S}(X_+^*)$. The polars and bipolars satisfy (Penot and Zălinescu, 2000, Lemma 4.2):

$$\begin{aligned} S^\circ &= (\operatorname{cl}\operatorname{co}((0,1]\cdot S))^\circ \\ S^{\circ\circ} &= \operatorname{cl}\operatorname{co}((0,1]\cdot S) \end{aligned} \quad \text{and} \quad \begin{aligned} S^\diamond &= (\operatorname{cl}\operatorname{co}([1,\infty)\cdot S))^\diamond \\ S^{\diamond\diamond} &= \operatorname{cl}\operatorname{co}([1,\infty)\cdot S), \end{aligned} \tag{14}$$

where the operations $S \mapsto \operatorname{cl}\operatorname{co}((0,1]\cdot S)$ and $S \mapsto \operatorname{cl}\operatorname{co}([1,\infty)\cdot S)$ are known as the *radiant hull* and *shady hull* respectively. The polar and antipolar operations also induce a natural duality relationship between the gauge and support functions:

$$\forall R \in \mathcal{R}(X), \quad \sigma_R = \gamma_{R^\circ}, \quad \sigma_{R^\circ} = \gamma_R; \quad \text{and} \quad \forall S \in \mathcal{S}(X), \quad \rho_S = \beta_{S^\diamond}, \quad \rho_{S^\diamond} = \beta_S, \tag{15}$$

such that we may define the *function polar and antipolar*:

$$\sigma_R^\circ := \sigma_{R^\circ}, \quad \gamma_R^\circ := \gamma_{R^\circ}; \quad \text{and} \quad \rho_S^\diamond := \rho_{S^\diamond}, \quad \beta_S^\diamond := \beta_{S^\diamond}. \tag{16}$$

The above relationships are presented diagrammatically in Figure 3.

The polar (antipolar) relationship between the convex (concave) support functions and gauge functions motivates a definition of the polar (antipolar) for sub/super-linear non-negative functions that is independent of its definition as a support function or gauge of a set. Let $f, g \colon \mathcal{X} \to \mathbb{R}$ with f sublinear, and g superlinear. Convex and concave polar duality correspondences for f and g are given by

$$\mathcal{X}^* \ni x^* \mapsto f^\circ(x^*) := \sup_{x \neq 0} \frac{\langle x^*, x\rangle}{f(x)} \in \overline{\mathbb{R}} \quad \text{and} \quad \mathcal{X}^* \ni x^* \mapsto g^\diamond(x^*) := \inf_{x \neq 0} \frac{\langle x^*, x\rangle}{g(x)} \in \overline{\mathbb{R}}.$$

Thus f° and $g^\diamond$ satisfy a *generalised Hölder* and *reverse Hölder* inequality respectively:

$$\forall x \in \mathcal{X}, x^* \in \mathcal{X}^*, \ \langle x^*, x\rangle \le f^\circ(x^*) f(x) \quad \text{and} \quad \langle x^*, x\rangle \ge g^\diamond(x^*) g(x). \tag{17}$$

The case of Hölder conjugate norms — $\|\cdot\|_p$ and $\|\cdot\|_q$ with $1/p + 1/q = 1$ — can easily be derived as a special case of the polar duality relationships above, with $\mathrm{B}_q = \mathrm{B}_p^\circ$ and $\mathrm{B}_p = \mathrm{B}_q^\circ$.

We will make use of the following result of Barbara and Crouzeix (1994) which can be seen to be analogous to the classical result (Hiriart-Urruty and Lemaréchal, 2001, Proposition E.1.4.3) regarding subdifferentials of Legendre–Fenchel conjugates. We express the result for the concave case because that is what we need for losses. An analogous result holds for convex gauges.

Lemma 10 (Barbara and Crouzeix 1994, Theorem 3.1) *Let $S \in \mathcal{S}(\mathscr{X})$. Then for $x \in \operatorname{dom} \beta_S$ and $x^* \in \operatorname{dom} \beta_{S^\diamond}$*

$$\frac{x}{\beta_S(x)} \in \partial \beta_{S^\diamond}(x^*) \iff \frac{x^*}{\beta_{S^\diamond}(x^*)} \in \partial \beta_S(x) \iff \beta_{S^\diamond}(x^*)\beta_S(x) = \langle x^*, x \rangle.$$

Barbara and Crouzeix (1994) provided a sketch of a proof. However as it is central to what follows we present a complete proof below.

Proof We first prove

$$\forall s \in S,\ \underbrace{\frac{x^*}{\beta_{S^\diamond}(x^*)} \in \partial \beta_S(s)}_{\text{(A)}} \iff \underbrace{\left(\frac{x^*}{\beta_{S^\diamond}(x^*)} \in S^\diamond \text{ and } \beta_S(s) = \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s \right\rangle \right)}_{\text{(B)}}. \tag{18}$$

For the sufficient condition suppose $s \in S$ and ${}^{x^*}\!/_{\beta_{S^\diamond}(x^*)} \in \partial \beta_S(s)$. Then from (8) we have

$$\forall y \in \mathscr{X},\ \beta_S(y) \le \beta_S(s) + \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, y - s \right\rangle. \tag{19}$$

Since by assumption S is shady, $s \neq 0$. Setting $y = 0$ and then $y = 2s$, and exploiting the 1-homogeneity of β_S and (19)

$$\beta_S(0) = 0 \le \beta_S(s) - \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s \right\rangle \implies \beta_S(s) \ge \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s \right\rangle, \tag{20}$$

$$\beta_S(2s) = 2\beta_S(s) \le \beta_S(s) + \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s \right\rangle \implies \beta_S(s) \le \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s \right\rangle. \tag{21}$$

Together, (20) and (21) give

$$\beta_S(s) = \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s \right\rangle. \tag{22}$$

Combining (19) with (22) gives

$$\begin{aligned} & \text{(A)} \implies \forall y \in \mathscr{X},\ \beta_S(y) \le \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s \right\rangle + \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, y - s \right\rangle = \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, y \right\rangle, \\ \iff\ & \text{(A)} \implies \forall y \in \mathscr{X},\ \beta_S(y) \le \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, y \right\rangle. \end{aligned} \tag{23}$$

Since S is closed and shady, $\operatorname{lev}_{\ge 1} \beta_S = S$ and

$$\forall y \in \mathscr{X},\ 1 \le \beta_S(s) \overset{(22)}{=} \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s \right\rangle \overset{(23)}{\le} \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, y \right\rangle \implies \forall s' \in S,\ 1 \le \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s' \right\rangle.$$

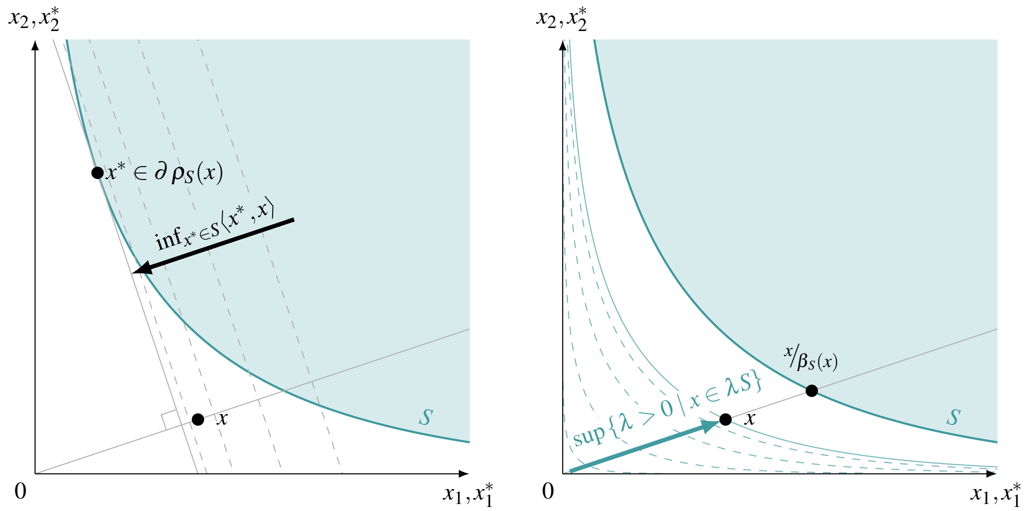
Thus ${}^{x^*}\!/_{\beta_{S^\diamond}(x^*)} \in S^\diamond$ by (13).

For the necessary condition suppose now that ${}^{x^*}\!/_{\beta_{S^\diamond}(x^*)} \in S^\diamond$ and let $s \in S$ be such that $\beta_S(s) = \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, s \right\rangle$. Then

$$0 = \beta_S(s) + \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, -s \right\rangle \iff \forall y \in \mathscr{X},\ \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, y \right\rangle = \beta_S(s) + \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, y - s \right\rangle. \tag{24}$$

The reverse Hölder inequality (17) gives

$$\forall y \in \mathscr{X},\ \left[\beta_{S^\diamond}(x^*)\beta_S(y) \le \langle x^*, y \rangle \iff \beta_S(y) \le \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, y \right\rangle \right]. \tag{25}$$



(a) (b)

FIGURE 4: Illustration of concave support function ρ_S and concave gauge function β_S. In (a) the set S is supported by the hyperplane normal to x at the point x^* with offset equal to $\rho_S(x)$. In (b), picking a point x, we consider scaled versions λS of S such that x ends up on the boundary of λS, thus giving the value of $\beta_S(x)$. The ray passing through x intersects the boundary of S at $x/\beta_S(x)$.

Combining (24) with (25) we have

$$(\mathrm{B}) \implies \forall y \in \mathcal{X},\ \beta_S(y) \le \beta_S(s) + \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, y - s \right\rangle.$$

Thus ${}^{x^*}\!/_{\beta_{S^\diamond}(x^*)} \in \partial \beta_S(s)$, and (18) is proved.

Let $x \in \operatorname{dom}(\beta_S)$. Then $\beta_S\left({}^{x}\!/_{\beta_S(x)}\right) = 1$. This follows since β_S is 1-homogenous. Since $S = \operatorname{lev}_{\ge 1} \beta_S$, we have ${}^{x}\!/_{\beta_S(x)} \in S$. Substituting $s = {}^{x}\!/_{\beta_S(x)}$ in (18) implies that for all $x^* \in \operatorname{dom}(\beta_{S^\diamond})$

$$\frac{x^*}{\beta_{S^\diamond}(x^*)} \in \partial \beta_S(x) \iff \left(\frac{x^*}{\beta_{S^\diamond}(x^*)} \in S^\diamond \text{ and } \frac{1}{\beta_S(x)} \cdot \beta_S(x) = \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, {}^{x}\!/_{\beta_S(x)} \right\rangle \right),$$

where we used the 0-homogeneity of $\partial \beta_S$ (Proposition 3) to obtain $\partial \beta_S({}^{x}\!/_{\beta_S(x)}) = \partial \beta_S(x)$. Since S is shady we can apply the bipolar theorem (14) to obtain the equivalent condition for all $x^* \in \operatorname{dom}(\beta_{S^\diamond})$:

$$\frac{x}{\beta_S(x)} \in \partial \beta_S^\diamond(x^*) \iff \left(\frac{x}{\beta_S(x)} \in S \text{ and } \frac{1}{\beta_{S^\diamond}(x^*)} \cdot \beta_{S^\diamond}(x^*) = \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, {}^{x}\!/_{\beta_S(x)} \right\rangle \right).$$

Finally observe

$$1 = \left\langle {}^{x^*}\!/_{\beta_{S^\diamond}(x^*)}, {}^{x}\!/_{\beta_S(x)} \right\rangle \iff \beta_{S^\diamond}(x^*)\beta_S(x) = \langle x^*, x \rangle,$$

which concludes the proof. ■

Lemma 11 *Suppose $S \in \mathcal{S}(\mathcal{X})$. Then β_S is strictly concave if and only if S is strictly convex.*

Proof We show the sufficient condition using a proof by contradiction. Assume (for the contradiction) β_S is strictly convex. If S is not strictly convex there exists $x, y \in S$ and $t \in (0,1)$ such that $tx + (1-t)y \in \mathrm{bd}(S)$. Since S is convex it follows that $\mathrm{ri}(S)$ is convex. And so it must be the case that $x, y \in \mathrm{bd}(S)$. Since S is shady, and by hypothesis β_S is strictly convex, it follows that

$$\underbrace{\beta_S(tx+(1-t)y)}_{1} < t\underbrace{\beta_S(x)}_{1} + (1-t)\underbrace{\beta_S(y)}_{1} = 1,$$

which is absurd. Thus S is strictly convex.

For necessity, choose arbitrary $x, y \in S$ with $x \neq y$. Then ${}^{x}\!/\!_{\beta_S(x)}, {}^{y}\!/\!_{\beta_S(y)} \in \mathrm{bd}(S)$ and $\beta_S({}^{x}\!/\!_{\beta_S(x)}) = \beta_S({}^{y}\!/\!_{\beta_S(y)}) = 1$. Since S is strictly convex $t\,{}^{x}\!/\!_{\beta_S(x)} + (1-t)\,{}^{y}\!/\!_{\beta_S(y)} \in \mathrm{ri}(S)$ for all $t \in (0,1)$. Thus

$$\begin{aligned}
& \beta_S\left(\frac{t}{\beta_S(x)}x + \frac{(1-t)}{\beta_S(y)}y\right) > 1 = t\beta_S\left(\frac{x}{\beta_S(x)}\right) + (1-t)\beta_S\left(\frac{y}{\beta_S(y)}\right) \\
\iff \quad & \beta_S\left(\frac{t}{\beta_S(x)}x + \frac{(1-t)}{\beta_S(y)}y\right) > \frac{t}{\beta_S(x)}\beta_S(x) + \frac{(1-t)}{\beta_S(y)}\beta_S(y),
\end{aligned}$$

where in the second line we exploited the 1-homogeneity of β_S. Multiplying both sides by $\frac{1}{{}^{t}\!/\!_{\beta_S(x)} + {}^{(1-t)}\!/\!_{\beta_S(y)}}$ and exploiting 1-homogeneity again gives $\forall t \in (0,1)$

$$\begin{aligned}
& \frac{1}{{}^{t}\!/\!_{\beta_S(x)} + {}^{(1-t)}\!/\!_{\beta_S(y)}}\beta_S\left(\frac{t}{\beta_S(x)}x + \frac{(1-t)}{\beta_S(y)}y\right) > \\
& \qquad \frac{{}^{t}\!/\!_{\beta_S(x)}}{{}^{t}\!/\!_{\beta_S(x)} + {}^{(1-t)}\!/\!_{\beta_S(y)}}\beta_S(x) + \frac{{}^{(1-t)}\!/\!_{\beta_S(y)}}{{}^{t}\!/\!_{\beta_S(x)} + {}^{(1-t)}\!/\!_{\beta_S(y)}}\beta_S(y) \\
\iff & \beta_S\left(\frac{{}^{t}\!/\!_{\beta_S(x)}}{{}^{t}\!/\!_{\beta_S(x)} + {}^{(1-t)}\!/\!_{\beta_S(y)}}x + \frac{{}^{(1-t)}\!/\!_{\beta_S(y)}}{{}^{t}\!/\!_{\beta_S(x)} + {}^{(1-t)}\!/\!_{\beta_S(y)}}y\right) > \\
& \qquad \frac{{}^{t}\!/\!_{\beta_S(x)}}{{}^{t}\!/\!_{\beta_S(x)} + {}^{(1-t)}\!/\!_{\beta_S(y)}}\beta_S(x) + \frac{{}^{(1-t)}\!/\!_{\beta_S(y)}}{{}^{t}\!/\!_{\beta_S(x)} + {}^{(1-t)}\!/\!_{\beta_S(y)}}\beta_S(y),
\end{aligned}$$

and thus β_S is strictly concave. ■

Corollary 12 *Let $S \in \mathcal{S}(\mathcal{X})$. Then ρ_S strictly concave if and only if $S^\diamond$ is strictly convex.*

Proof Apply Lemma 11 to $\rho_S = \beta_{S^\diamond}$. ■

Lemma 13 *Let $(s_i^n)_{n\in\mathbb{N}}$ with $s_i^n \in X_+ \setminus \{0\}$ for $i \in [m]$. Assume $\left(\sum_{i\in[m]} s_i^n\right)_{n\in\mathbb{N}}$ is a convergent sequence. Then each sequence $(s_i^n)_{n\in\mathbb{N}}$ for $i \in [m]$ has a convergent subsequence.*

Proof If each of the m sequences $(s_i^n)_{n\in\mathbb{N}}$ is bounded the proof is trivial. Assume that for $1 \leq i \leq l$ the sequences $(s_i^n)_{n\in\mathbb{N}}$ are unbounded. We now show there exists a linear functional z^* such that with $x_n := \sum_{i=1}^m s_i^n$,

$$\underbrace{\langle z^*, x_n\rangle}_{\text{bounded}} = \underbrace{\langle z^*, s_1^n\rangle + \cdots + \langle z^*, s_l^n\rangle}_{\text{unbounded}} + \underbrace{\langle z^*, s_{l+1}^n\rangle + \cdots + \langle z^*, s_m^n\rangle}_{\text{bounded}}, \tag{26}$$

	Convex case $R \in \mathcal{R}(\mathscr{X})$ f convex	Concave case $S \in \mathcal{S}(\mathscr{X})$ f concave	Definition
Support function	σ_R	ρ_R	(5), (6)
Gauge function	γ_R	β_S	(9), (10)
Conjugate function	f^*	f_*	(11), (12)
Polar set	R°	$S^\diamond$	(13)
Polar function	f°	$f^\diamond$	(16)

TABLE 2: Summary of the convex case and the (less known) concave case of support functions, gauge functions, polar sets, polar functions and Legendre-Fenchel conjugates.

producing a contradiction.

To show the existence of z^* define the set

$$U := \operatorname{cl}\operatorname{co}\left(\bigcup_{i\in[m]}\left\{\frac{s_i^n}{\|s_i^n\|}\,\middle|\, n\in\mathbb{N}\right\}\right) \subset X_+.$$

Observe that since X_+ is salient, closed and convex, we have $U \cap \{0\} = \varnothing$. Furthermore, we trivially have that $\{0\}$ is compact and U is closed. Then by the Hahn–Banach separation theorem (Penot, 2012, Theorem 1.79, p. 55) there exists $z^* \in \mathscr{X}^*$ such that $\langle z^*, u\rangle \geq \delta > 0$ for all $u \in U$. To see why the first l terms must be unbounded with this choice of z^* note that we can write

$$\forall i \in [l],\ \langle z^*, s_i^n\rangle = \left\langle z^*, \|s_i^n\|\, {}^{s_i^n}\!/\!_{\|s_i^n\|}\right\rangle = \|s_i^n\|\left\langle z^*, {}^{s_i^n}\!/\!_{\|s_i^n\|}\right\rangle,$$

where $\|s_i^n\| \to \infty$ and $\left\langle z^*, {}^{s_i^n}\!/\!_{\|s_i^n\|}\right\rangle \geq \delta$ for every $n \in \mathbb{N}$ since ${}^{s_i^n}\!/\!_{\|s_i^n\|} \in U$. This shows (26), which is absurd. ∎

2.10 Comparing the Convex and Concave Versions

Table 2 tabulates the convex and concave versions of the key mathematical objects we make use of in this paper. As can be seen, for every standard convex version, there is a corresponding concave version.

3. Loss Functions

In this section we will introduce proper losses; first in the traditional way, and then in terms of the superprediction set. We will then show some of the implications of the latter approach. A loss function is an "outcome contingent disutility": that is, for a given outcome y, it provides a measure of (dis)utility of a prediction as a function $-u(\cdot, y)$ (Berger, 1985). We introduce loss functions more formally by first introducing some concepts from statistical decision theory, to which we apply some of the geometric concepts introduced in §2. Let Z, Y be random variables taking values in the spaces $\mathscr{X}$ and $\mathscr{Y}$. We assume $\mathscr{Y}$ is finite with $n := |\mathscr{Y}|$, and therefore distributions that assign probability to every state of $\mathscr{Y}$ are isomorphic to probability vectors from $\Delta := \operatorname{ri}\left(\left\{p \in \mathbb{R}^n_{\geq 0} \,\middle|\, \sum_{i=1}^n p_i = 1\right\}\right)$, the relative interior of the *n-simplex*, that is $\operatorname{cl}(\Delta) \simeq \Delta(\mathscr{Y})$. Its dual cone, $\mathbb{R}^n_{\geq 0}$, is the associated collection of loss vectors. We use $X_+ \subseteq \mathbb{R}^n$ and $X_+^* \subseteq \mathbb{R}^n$ to denote an arbitrary pair of salient, closed, convex cones, dual to one another. The reader might find it helpful to identify X_+ with $\operatorname{cl}\operatorname{cone}(\Delta)$, and X_+^* with its dual, however most of our theorems merely depend on a dual pair of closed convex cones.

One can understand the effect of choice of loss in terms of the "conditional perspective" which allows one to ignore the distribution of Z, which is typically unknown.[6] We call mappings $\ell \colon \Delta \times \mathscr{Y} \to \mathbb{R}_{\geq 0}$ *loss functions*, and $\ell(p;y)$ is the penalty from predicting $p \in \Delta$ upon observing $y \in \mathscr{Y}$. (It will sometimes be convenient to consider the extension of ℓ defined as $\ell\colon \operatorname{cl}(\Delta) \times \mathscr{Y} \to \overline{\mathbb{R}}_{\geq 0}$, which now needs to map to $\overline{\mathbb{R}}_{\geq 0}$ to allow for infinite values; see Remark 22). It will be convenient to stack loss functions into a function over the second argument:

$$\forall p \in \Delta,\ \ell(p) \coloneqq y \mapsto \ell(p;y) \in \mathbb{R}^{\mathscr{Y}}_{\geq 0},$$

or equivalently we have a vector

$$\forall p \in \Delta,\ \ell(p) \simeq (\ell(p;y_1), \ldots, \ell(p;y_n))' \in \mathbb{R}^n_{\geq 0}.$$

For each fixed $y \in \mathscr{Y}$, the functions $\ell(\cdot\,;y)$ are called *partial losses*. If for some norm (the choice does not matter) $\|\ell(p)\| < \infty$ for all $p \in \Delta$ we say the loss is *bounded*.

The *conditional risk associated with ℓ* is defined via

$$L\colon \Delta \times \Delta \ni (q,p) \mapsto L(q,p) \coloneqq \mathbb{E}_q[\ell(p)] = \langle \ell(p), q \rangle \in \mathbb{R}_{\geq 0}.$$

The *conditional Bayes risk* $\underline{L} \coloneqq \Delta \ni q \mapsto \inf_{p\in\Delta} L(q,p)$ is always concave. For the next two definitions let $f\colon \mathscr{X} \to \Delta$, and $g(z) \coloneqq \Pr(Y|Z=z)$. That is, for $z \in \mathscr{X}$, $f(z)$ is a distribution over Y conditioned on $Z = z$ and $g(Z)$ is the true conditional distribution. The *full risk* is

$$\mathbb{E}_Z \mathbb{E}_{Y|Z}[\ell \circ f(Z)] = \mathbb{E}_Z \langle \ell \circ f(Z), g(Z) \rangle. \tag{27}$$

The most general framing of a supervised machine learning problem is to minimise (27) by choosing an appropriate function f. If we fix ℓ and g, the minimal value of the full risk (27) is bounded below by the *Bayes risk*[7]:

$$\inf_{f\colon \mathscr{X} \to \Delta} \mathbb{E}_Z \langle \ell \circ f(Z), g(Z) \rangle.$$

The *superprediction set* (Kalnishkan et al., 2004; Kalnishkan and Vyugin, 2002; Dawid, 2007) of a loss function $\ell\colon \Delta \to \mathbb{R}^n_{\geq 0}$ is

$$\operatorname{spr}(\ell) \coloneqq \bigcup_{l \in \ell(\Delta)} \left\{ x \in \mathbb{R}^n \,\middle|\, x \succeq_{\mathbb{R}^n_{\geq 0}} l \right\} \subseteq \mathbb{R}^n_{\geq 0}. \tag{28}$$

The set $\operatorname{spr}(\ell)$ (we write $\operatorname{spr}\ell$ when there is no ambiguity) consists of all the points x that incur no less loss than some point $l \in \ell(\Delta)$. In the parlance of game theory, this is the union of the points

6. See (Steinwart and Christmann, 2008; Reid and Williamson, 2011) for a discussion of this conditional perspective.
7. See (Williamson and Cranko, 2022) for a further discussion of the Bayes risk, its generalisation to restricted model classes $\mathscr{F}$, and the relationship to measures of information such as f-divergences and integral probability metrics.

p that are weakly dominated by some other point, together with the dominating points. Equivalently (28) can be written

$$\operatorname{spr}\ell = \bigcup_{l\in\ell(\Delta)} (l+\mathbb{R}^n_{\geq 0}) = \ell(\Delta)+\mathbb{R}^n_{\geq 0}. \tag{29}$$

In the next section we will be interested in the closed convex hull of the superprediction set mapping, for which it is useful to note

$$\operatorname{cl}\operatorname{co}(\operatorname{spr}\ell) \overset{(29)}{=} \operatorname{cl}\operatorname{co}(\ell(\Delta)+\mathbb{R}^n_{\geq 0}) = \operatorname{cl}\operatorname{co}(\ell(\Delta))+\mathbb{R}^n_{\geq 0} = \bigcup_{l\in\operatorname{cl}\operatorname{co}(\ell(\Delta))} (l+\mathbb{R}^n_{\geq 0}). \tag{30}$$

Lemma 14 is a special case of a result due to Choquet (1962), but as the original is in French we include a proof for our setting below.

Lemma 14 *Let $S_1,\ldots,S_m \subseteq X_+\setminus\{0\}$ each be closed. Then the set $S_1+\cdots+S_m$ is closed.*

Proof Let $S := S_1+\cdots+S_m$, and take a convergent sequence $(x_n)_{n\in\mathbb{N}} \to x$ with $x_n \in S$. Then for each x_n there exists $(s_i^n)_{i\in[m]}$ with $s_i^n \in S_i$ for $i\in[m]$ and $n\in\mathbb{N}$. By Lemma 13 each of the sequences $(s_i^n)_{n\in\mathbb{N}}$ has a convergent subsequence. And as S_i is closed $\lim_{n\to\infty} s_i^n \in S$ for each $i\in[m]$ and $x = \sum_{i\in[m]} \lim_{n\to\infty} s_i^n \in S$ (taking subsequences if need be). ∎

3.1 Proper Losses

A natural requirement to impose upon ℓ is that it is *proper*[8] (Hendrickson and Buehler, 1971), which means that

$$[\forall p,q\in\Delta,\ \langle \ell(q),q\rangle \leq \langle \ell(p),q\rangle] \iff \left[\forall q\in\Delta,\ q\in \operatorname*{arg\,inf}_{p\in\Delta}\langle \ell(p),q\rangle\right]. \tag{31}$$

That is, predicting the true probability minimises the expected loss. We say ℓ is *strictly proper* when the above inequality is strict for $p\neq q$, that is, $\{q\} = \operatorname{arg\,inf}_{p\in\Delta}\langle \ell(p),q\rangle$ for all $q\in\Delta$. If ℓ is proper, $\underline{L}(p) = L(p,p) = \langle \ell(p),p\rangle$. The superprediction set $\operatorname{spr}(\ell)$ of a proper loss ℓ has some useful properties: Theorem 15 makes explicit the link between the superprediction set and the convex geometry in §2. Proposition 21 below justifies that from a superprediction set we can construct a loss function. This motivates a shift in focus of analysis from loss functions ℓ to families of convex superprediction sets of proper losses.

Theorem 15 (Representation) *Let $\ell\colon \Delta\to\mathbb{R}^n$ be a loss function with the associated conditional Bayes risk $\underline{L}$. Then*

1. *$\underline{L} = \rho_{\operatorname{cl}\operatorname{co}(\operatorname{spr}\ell)}{}_{|\Delta}$ (the restriction of $\rho_{\operatorname{cl}\operatorname{co}(\operatorname{spr}\ell)}$ to Δ);*

8. See (Gneiting and Raftery, 2007; Reid and Williamson, 2011) for an elaboration of the notion of properness, which dates back at least to work of Shuford Jr. et al. (1966) and von Holstein (1970); early particular examples are due to Brier (1950) and Good (1952). Note that what we call a proper loss is often called a proper scoring rule; the case we consider corresponds to having an action space being a set of finite dimensional distributions; see for example (Grünwald and Dawid, 2004).

2. $[\forall p \in \Delta,\ \ell(p) \in \partial \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p)] \iff \ell$ *is proper*.

Proof The claim that $\underline{L} = \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}$ on Δ for all loss functions ℓ is straightforward:

$$\forall p \in \Delta,\ \underline{L}(p) = \inf_{q\in\Delta}\langle \ell(q), p\rangle = \inf_{l\in\overline{\ell(\Delta)}}\langle l, p\rangle = \inf_{l\in\overline{\ell(\Delta)+\mathbb{R}^n_{\geq 0}}}\langle l, p\rangle \overset{(30)}{=} \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p). \tag{32}$$

We now prove the second claim. Assume ℓ is proper. Then for $p \in \Delta$ we have $\rho_{\mathrm{spr}(\ell)}(p) = L(p,p) = \langle \ell(p), p\rangle$ and $\forall p \in \Delta,\ \forall q \in \mathbb{R}^n$,

$$\begin{aligned}
\langle \ell(p), q\rangle \geq \langle \ell(q), q\rangle &\iff \langle \ell(p), q\rangle - \langle \ell(p), p\rangle \geq \langle \ell(q), q\rangle - \langle \ell(p), p\rangle \\
&\overset{(32)}{\iff} \langle \ell(p), q-p\rangle \geq \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(q) - \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p).
\end{aligned}$$

Thus $\ell(p) \in \partial \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p)$ for $p \in \Delta$.

We use a proof by contradiction to show the reverse implication. Assume ℓ enjoys the subgradient representation $\ell(p) \in \partial \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p)$ for $p \in \Delta$, but is not proper. Since ℓ is not proper, by (31) there exists $p, q \in \Delta$ with

$$\begin{aligned}
\langle \ell(p), q\rangle < \langle \ell(q), q\rangle &\iff \langle \ell(p), q\rangle - \langle \ell(p), p\rangle < \langle \ell(q), q\rangle - \langle \ell(p), p\rangle \\
&\iff \langle \ell(p), q-p\rangle < \langle \ell(q), q\rangle - \langle \ell(p), p\rangle.
\end{aligned} \tag{33}$$

Since $\ell(p') \in \partial \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p')$ for $p' \in \Delta$, from (8) we have $\forall p', q' \in \mathbb{R}^n_{\geq 0}$,

$$\begin{aligned}
&\langle \ell(p'), q'-p'\rangle \geq \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(q') - \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p') \\
\implies\ & -\langle \ell(p'), p'\rangle \geq -\rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p') \\
\iff\ & \langle \ell(p'), p'\rangle \leq \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p'),
\end{aligned} \tag{34}$$

where in the implication we take $q' = 0$. This gives us, for our choice of p, q,

$$\begin{aligned}
\langle \ell(p), q-p\rangle &\overset{(33)}{<} \langle \ell(q), q\rangle - \langle \ell(p), p\rangle \\
&\overset{(34)}{\leq} \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(q) - \langle \ell(p), p\rangle \\
&\leq \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(q) - \inf_{r\in\Delta}\langle \ell(r), p\rangle \\
&\overset{(32)}{=} \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(q) - \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p),
\end{aligned}$$

which contradicts our assumption that $\ell(p) \in \partial \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p)$ for all $p \in \Delta$. ■

Thus in order to build a geometry of loss functions in terms of convex sets, with Theorem 15 we see that the propriety condition of the losses cannot be discarded; see also the discussion in section 3.6 below.

The definition of Δ as the relative interior of the probability simplex guarantees (via Corollary 17(2) below) that the subdifferential $\partial \rho_{\mathrm{cl\,co}(\mathrm{spr}\,\ell)}(p)$ is nonempty for all $p \in \Delta$. This is analogous to the differentiable case, where if one wishes to compute gradients of a differentiable function, the natural area of analysis is the interior of its domain of definition.

Proposition 16 *Suppose $\ell\colon \Delta \to \mathbb{R}^n_{\geq 0}$ is a loss function. Then* $\operatorname{cl}\operatorname{co}(\operatorname{spr}\ell)$ *is $\mathbb{R}^n_{\geq 0}$-oriented.*

Proof By hypothesis $\ell\colon \Delta \to \mathbb{R}^n$ and so from the definition of the superprediction set $\operatorname{spr}(\ell) \subseteq \mathbb{R}^n_{\geq 0}$ we have

$$
\begin{aligned}
\operatorname{rec}(\mathbb{R}^n_{\geq 0}) &\overset{\text{P1(5)}}{\supseteq} \operatorname{rec}(\operatorname{cl}\operatorname{co}(\operatorname{spr}\ell)) \\
&\overset{(30)}{=} \operatorname{rec}\left(\bigcup_{l \in \operatorname{cl}\operatorname{co}(\ell(\Delta))} (l + \mathbb{R}^n_{\geq 0})\right) \\
&\overset{\text{P1(7)}}{\supseteq} \bigcup_{l \in \operatorname{cl}\operatorname{co}(\ell(\Delta))} \operatorname{rec}(l + \mathbb{R}^n_{\geq 0}) \\
&= \operatorname{rec}(\mathbb{R}^n_{\geq 0}) \\
&\overset{\text{P1(4)}}{=} \mathbb{R}^n_{\geq 0}
\end{aligned}
$$

as desired. ■

Corollary 17 *Suppose $\ell\colon \Delta \to \mathbb{R}^n_{\geq 0}$ is a loss function. Then*

1. $\operatorname{dom}(\rho_{\operatorname{cl}\operatorname{co}(\operatorname{spr}\ell)}) \supseteq \mathbb{R}^n_{\geq 0}$, *and*
2. $\operatorname{dom}(\partial \rho_{\operatorname{cl}\operatorname{co}(\operatorname{spr}\ell)}) \supseteq \mathbb{R}^n_{>0}$.

Proof Take Proposition 16 and apply Lemma 9, this shows claim 1. Apply Lemma 2 to (1) to show claim 2. ■

Theorem 15 and Proposition 16 motivate the introduction of the following family of convex sets. (Recall that $\mathcal{X} = \mathbb{R}^n$ for some n and $X_+ \subset \mathcal{X}$ denotes a salient closed convex cone.) Take $C \subseteq \mathcal{X}$, let $\mathcal{P}(C)$ be the collection of C-oriented, closed, nonempty convex subsets of $C \setminus \{0\}$:

$$
\mathcal{P}(C) := \{S \in \mathcal{K}(C \setminus \{0\}) \mid \operatorname{rec}(S) = C\}. \tag{35}
$$

The construction (35) admits a lot of structure. In particular, Lemma 18 justifies our interest in the shady sets introduced in §2.7. Recall that $\mathcal{S}(\mathcal{X})$ denotes the collection of closed shady subsets of $\mathcal{X}$.

Lemma 18 $\mathcal{P}(X_+) \subseteq \mathcal{S}(\mathcal{X})$.

Proof Take $S \in \mathcal{P}(X_+)$. Since $S \subseteq \operatorname{rec}(S) = X_+$, we have

$$
\begin{aligned}
(\forall d \in X_+,\ \forall \alpha > 0,\ S + \alpha d \subseteq S) &\implies (\forall d \in S,\ \forall \alpha > 0,\ S + \alpha d \subseteq S) \\
&\implies (\forall \alpha \geq 1,\ \alpha S \subseteq S),
\end{aligned}
$$

and S is co-star-shaped. From (35), S is also convex, thus S is shady. ■

Lemma 19 *Let $S \in \mathcal{P}(X_+)$. Then* $\operatorname{ri}(S) \neq \varnothing$.

Proof The result follows from a proof by contradiction. Suppose $\mathrm{ri}(S) = \varnothing$. Then from (1) this means

$$\forall x \in S,\ \forall \lambda > 1,\ \exists y \in S,\ [\lambda x + (1-\lambda) y \notin S \iff \lambda x \notin S + (\lambda - 1) y]. \tag{36}$$

From the conditions, $\lambda - 1 > 0$ and $y \in S \subseteq \mathrm{rec}(S)$. Thus $(\lambda - 1)y \in \mathrm{rec}(S)$. It follows that $S + (\lambda - 1)y = S$ because for any $x \in \mathrm{rec}\, S$, $S + x = S$ (Proposition 1 (2)). Hence (36) reduces to

$$\forall x \in S,\ \forall \lambda > 1,\ \lambda x \notin S,$$

which contradicts the fact that S is co-star-shaped (Lemma 18). ■

Lemma 20 *Let $A_i \in \mathcal{P}(X_+)$ for $i \in [m]$. Then $\bigcap_{i \in [m]} A_i \neq \varnothing$.*

Proof Take an arbitrary sequence $(a_i)_{i \in [m]}$ with $a_i \in A_i$. Let $J_i := [m] \setminus \{i\}$, for $i \in [m]$. Then $\sum_{j \in J_i} a_j + a_i = \sum_{k \in [m]} a_k$ for each $i \in [m]$. Since X_+ is a cone $\sum_{j \in J_i} a_j \in X_+$ for each $i \in [m]$. By assumption $A_i \subseteq \mathrm{rec}(A_i) = X_+$, and from the definition of the recession cone (3) we always have $\sum_{k \in [m]} a_k = a_i + \sum_{j \in J_i} a_j \in a_i + X_+ \subseteq A_i$ for all $i \in [m]$, and thus $\bigcap_{i \in [m]} A_i \neq \varnothing$. ■

Proposition 21 completes the connection between superprediction sets and loss functions.

Proposition 21 *Take $S \in \mathcal{P}(\mathbb{R}^n_{\geq 0})$. There exists a 0-homogeneous selection $\ell : \mathbb{R}^n_{>0} \to \mathbb{R}^n_{\geq 0}$ of $\partial \rho_S$ in the sense that*

$$\forall p \in \mathbb{R}^n_{>0},\ \ell(p) \in \partial \rho_S(p), \tag{37}$$

and ℓ restricted to Δ is a proper loss.

Proof From Lemma 2, $\mathrm{dom}(\partial \rho_S) \supseteq \mathrm{ri}(\mathbb{R}^n_{\geq 0}) = \mathbb{R}^n_{>0}$, and so there exists $\ell(p) \in \underset{x \in S}{\arg\inf} \langle x, p \rangle$ for $p \in \mathbb{R}^n_{>0}$. Thus

$$\forall p, q \in \Delta \subseteq \mathbb{R}^n_{>0},\ \langle \ell(q), q \rangle \leq \langle \ell(p), q \rangle,$$

and ℓ is proper. Since ρ_S is 1-homogeneous, ℓ is 0-homogeneous (Proposition 3). ■

Remark 22 (From bounded to unbounded) In Theorem 15 and Proposition 21 we needed to be careful when talking about the domain of definition of a loss function ℓ. This was to ensure that $\partial \rho_{\mathrm{spr}(\ell)}$ is nonempty in order to have the inclusion $\ell(p) \in \partial \rho_{\mathrm{spr}(\ell)}(p)$. Recall our definition of Δ as the relative interior of the standard simplex. In practice since if there exists $q \in \mathrm{cl}(\Delta)$ with $\partial \rho_{\mathrm{spr}(\ell)}(q) = \varnothing$, we can define $\ell(q; y) := \lim_{(p_n) \to q} \ell(p_n; y)$, where the sequence (p_n) is chosen with $p_n \in \Delta$ and by Lemma 2 we know we have $\partial \rho_{\mathrm{spr}(\ell)}(p_n) \neq \varnothing$ (allowing us to take $\ell(p_n) \in \partial \rho_{\mathrm{spr}(\ell)}(p_n)$). This extends our loss function to a mapping $\ell\colon \mathrm{cl}(\Delta) \to \overline{\mathbb{R}}^n_{\geq 0}$. Many of the results on bounded loss functions can be similarly extended to unbounded loss functions, by restricting analysis to Δ; for example the range of log loss is $\overline{\mathbb{R}}^n_{\geq 0}$ when defined on $\mathrm{cl}(\Delta)$ but $\mathbb{R}^n_{\geq 0}$ when restricted to Δ.

Name		$\ell(p;y)$	$\underline{L}(p)$	Propriety	Bounded	Ref.
0/1	$\ell_{0/1}$	$\operatorname{co}\{e^j : j \in \arg\max_{y\in\mathscr{Y}} p_y\}$	$\min_{y\in\mathscr{Y}} p_y$	proper	•	§3.1
Logarithmic	$\ell_{\log}$	$-\log(p_y)$	$-\sum_{y\in\mathscr{Y}} p_y \log p_y$	strict		§3.4
Concave Norm $a \in [-\infty, 1] \setminus \{0\}$	ℓ_a	$\left(\frac{p_y}{\beta_{\frac{a}{a-1}}(p)}\right)^{\frac{1}{a-1}}$	$\beta_{\frac{a}{a-1}}(p)$	strict		§4.1
Brier	ℓ_{Br}	$(1+\\|p\\|_2^2)\cdot 1_n - 2p_y$	$1-\\|p\\|_2^2$	strict	•	§4.2
Cobb–Douglas $a \in \mathbb{R}^n_{\geq 0}$	ℓ_{CD_a}	$\psi_a(p/a)\cdot a_y/p_y$	$\psi_a(p)$	strict		§4.3

TABLE 3: Some common loss functions, their conditional Bayes risks, and their propriety and boundedness.

In light of Remark 22, the distinction between bounded and unbounded losses becomes less important. Some commonly used loss functions are listed in Table 3 along with their boundedness.

Corollary 23 below follows from the proof of Proposition 21 together with Corollary 5.

Corollary 23 *A loss function ℓ is strictly proper if and only if* $\operatorname{cl}\operatorname{co}(\operatorname{spr}\ell)$ *is strictly convex.*

Corollary 24 *There is a bijection between the equivalence class of loss functions $\ell : \Delta \to \mathbb{R}^n_{\geq 0}$ which agree almost everywhere and the family of convex sets $\mathcal{P}(\mathbb{R}^n_{\geq 0})$.*

Proof There is a bijection between superlinear functions ρ_S and closed convex sets S (Hiriart-Urruty and Lemaréchal, 2001, Theorem C.2.2.2), and the mapping $\partial\rho_S$ is a singleton almost everywhere (Hiriart-Urruty and Lemaréchal, 2001, Theorem B.4.2.3). The connection to loss functions (Theorem 15 and Proposition 21) completes the proof. ■

3.2 Starting with Sets

The above development motivates the key viewpoint of the present paper: *start with a set $S \in \mathcal{P}(\mathbb{R}^n_{\geq 0})$ and derive the loss (and other quantities) from it.* We will thus sometimes explicitly parametrise the loss theoretic functions as ℓ_S, L_S and $\underline{L}_S$. One immediate consequence of using a set $\operatorname{spr}(\ell) \in \mathcal{P}(\mathbb{R}^n_{\geq 0})$ to define a proper loss ℓ is that it may be the case that for two different loss functions $\ell \neq m$ we have $\operatorname{cl}\operatorname{co}(\operatorname{spr}\ell) = \operatorname{cl}\operatorname{co}(\operatorname{spr} m)$. This is the case whenever the conditional Bayes risk functions for ℓ and m coincide. However in such cases ℓ and m differ only on a set of measure zero (cf. Vernet et al., 2016, Proposition 8). That is, for some $S := \operatorname{spr}(\ell) = \operatorname{spr}(m)$, both ℓ and m satisfy (37). However when ℓ is strictly proper, $\operatorname{spr}(\ell)$ is strictly convex and so for strictly convex $\operatorname{spr}(\ell) = \operatorname{spr}(m)$ we always have $\ell = m$.

Remark 25 (Misclassification loss) Misclassification loss $\ell_{0/1}$ (also called $0/1$ loss) (Buja et al., 2005; Gneiting and Raftery, 2007) assigns zero loss when predicting correctly and a loss of 1 when predicting incorrectly. This can be extended to when one predicts with a distribution $p \in \Delta$, with $\ell_{0/1}(p) = e^j$ where $j = \arg\max_{i\in[n]} p_i$ and e^j is the jth canonical basis vector in $\mathbb{R}^n$, under the assumption that $\{p_i : i \in [n]\}$ has a unique maximum. We can extend $\ell_{0/1}$ to all of Δ in a manner that is consistent with Theorem 15 as follows: Define

$$\underline{L}_{0/1} : \Delta \ni p \mapsto \min_{i \in [n]} p_i,$$

and let $\ell_{0/1}(p) := \partial' \underline{L}_{0/1}(p) = -\partial_{\prime}(-\underline{L}_{0/1}(p)) = -\partial_{\prime}(-\min_i p_i) = -\partial_{\prime}(\max_i(-p_i))$. Using (Hiriart-Urruty and Lemaréchal, 2001, Example 3.4, page 182) we have

$$\ell_{0/1}(p) = -\partial_{\prime}(-\underline{L}(p)) = -\operatorname{co}\{-\mathrm{e}^j : j \in \arg\max_{i \in [n]} p_i\} = \operatorname{co}\{\mathrm{e}^j : j \in \arg\max_{i \in [n]} p_i\}. \tag{38}$$

When the argmax is unique, this reduces to just e^j as per the usual definition.

Remark 26 (Naturally 0-Homogeneous) In Theorem 15 we needed to make restrictions on the domain of the concave support function and subdifferential in order for the geometric functions from convex analysis to agree with the classical concept of a loss function. However when loss functions are viewed instead as the subgradient of a concave support function these restrictions are indeed arbitrary. The subgradient representation of a proper loss function (37) suggests via the 0-homogeneity result of Proposition 21 that it is more natural to take a proper loss function ℓ defined in the conventional way on Δ and instead consider its 0-homogeneous extension: $\mathbb{R}^n_{>0} \ni p \mapsto \ell'(p) := \ell\left(p \cdot {}^{1}\!/_{\|p\|_1}\right) \in \mathbb{R}^n_{\geq 0}$. Defined like this, ℓ' satisfies (37) and agrees with ℓ on Δ. Without loss of generality, for the remainder of our analysis we will refer to a *loss function* as such a 0-homogeneous mapping $\mathbb{R}^n_{>0} \to \mathbb{R}^n_{\geq 0}$.[9]

3.3 Bregman Divergences, Semi Inner Products and Finslerian Geometry

The relationship between ℓ and the conditional Bayes risk $\underline{L}$ (which we know is equal to $\rho_{\mathrm{spr}(\ell)}$) is usually credited to Savage (1971) and is intimately related to Bregman divergences[10]. Given a convex function ϕ, the *Bregman divergence* between $x, y \in \operatorname{dom}(\phi)$ is defined to be

$$\mathrm{B}_\phi(x,y) := \phi(x) - \phi(y) - \langle g(y), x - y \rangle,$$

where $g \in \partial\phi$ is a selection of the subgradient of ϕ. It is known that the *regret* $L(p,q) - \underline{L}(p)$ is a Bregman divergence with $\phi = -\underline{L}$, which is not only convex but is also 1-homogeneous. The additional structure of 1-homogeneity offers a nice simplification.

Proposition 27 *Give a Bayes risk $\underline{L}$, $\forall p, q \in \mathbb{R}^n_{>0}$,*

$$\mathrm{B}_{-\underline{L}}(p,q) = L(p,q) - \underline{L}(p) = \langle \ell(q) - \ell(p), p \rangle. \tag{39}$$

Proof We have $\forall p, q \in \mathbb{R}^n_{>0}$, $L(p,q) = \langle \ell(q), p \rangle$, and thus if ℓ is proper by Theorem 15, the general form of the Bregman divergence simplifies:

$$\mathrm{B}_{-\underline{L}}(p,q) = -\underline{L}(p) + \underline{L}(q) + \langle \ell(q), p - q \rangle$$

9. A formal alternative would be to work with the horizon $\operatorname{hzn}\mathbb{R}^n$ of directions $\operatorname{dir}\colon \mathbb{R}^n \to \operatorname{hzn}\mathbb{R}^n$ which can be considered pure (magnitudeless) direction vectors, so that for $\alpha > 0$, $\operatorname{dir}(\alpha x) = \operatorname{dir}(x)$ (Rockafellar and Wets, 2004, Chapter 3). It amounts to the same thing since 0-homogeneity of ℓ means that the magnitude of any vector $x \in \mathbb{R}^n$ does not affect the value of $\ell(x)$; all that matters is the direction of x. Thus one could in fact *define* $\tilde{\ell} : \operatorname{hzn}\mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ via $\tilde{\ell}(\tilde{x}) = \ell(x)$, where x is the unique point of intersection of the "infinite magnitude" direction vector $\tilde{x}$ with the unit sphere.

10. See (Reid and Williamson, 2011) for more context and background on Bregman divergences.

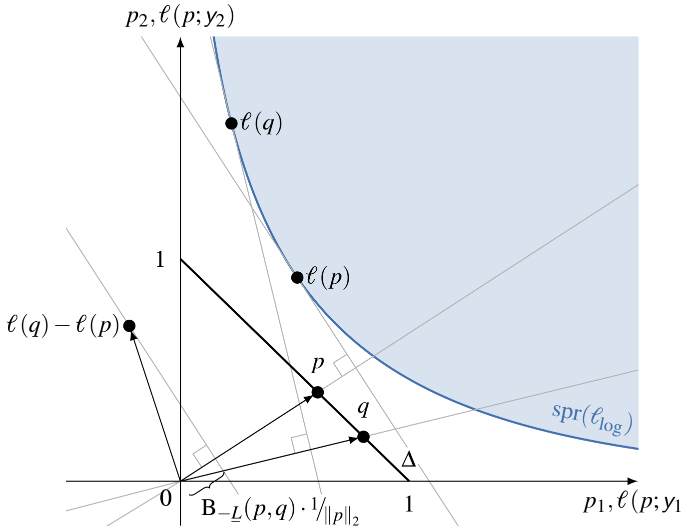


FIGURE 5: Geometrical interpretation of regret as the Bregman divergence $\mathrm{B}_{-\underline{L}}(p,q) = \langle \ell(q) - \ell(p), p\rangle$. As $q \to p$, so $\ell(q) \to \ell(p)$ and the vectors $\ell(q) - \ell(p)$ and p become orthogonal and $\mathrm{B}_{-\underline{L}}(p,q) \to 0$.

$$
\begin{aligned}
&= -\langle \ell(p), p\rangle + \langle \ell(q), q\rangle + \langle \ell(q), p\rangle - \langle \ell(q), q\rangle \\
&= \langle \ell(q) - \ell(p), p\rangle \\
&= L(p,q) - \underline{L}(p),
\end{aligned}
$$

where we have used the fact that since ℓ is proper $\underline{L}(p) = \langle \ell(p), p\rangle$. ■

The simpler form (39) provides for a geometrical interpretation of the Bregman divergence as the inner product of the vectors $(\ell(q) - \ell(p))$ and p, which we illustrate in Figure 5.

Considering loss functions as subgradients of concave support functions provides an intriguing geometrical perspective which we now sketch. It is based upon existing work on norm derivatives (Alsina et al., 2010) which we first summarise. Given a norm $\|\cdot\|$ on some vector space V, define the *normalised norm derivative* for $x, y \in V$ via

$$\tau'(x,y) := \lim_{\lambda\to 0} \frac{\|x+\lambda y\|^2 - \|x\|^2}{2\lambda} = \|x\| \lim_{\lambda\to 0} \frac{\|x+\lambda y\| - \|x\|}{\lambda}.$$

The function τ' is of interest because if the norm is derived from an inner product via $\|\cdot\| = \langle\cdot,\cdot\rangle^{\frac{1}{2}}$, then $\tau'(x,y) = \langle x, y\rangle$. For norms that are *not* derived from an inner product, the normalised norm derivative can be claimed to be "like" an inner product. This claim can be formalised as follows. First we slightly generalise τ' by writing it in terms of a gauge function γ with associated unit ball S° in V which we henceforth take to be $\mathbb{R}^n$ (recall every norm is a gauge function, but a gauge function is only a norm if its unit ball is centrally symmetric with respect to the origin). Suppose S° is smooth and strictly convex and thus S is too, and hence σ_S (and σ_{S°) is differentible everywhere since there is only one support point for a given hyperplane. We can generalise the definition of τ' as

$$
\begin{aligned}
\tau'_S(x,y) &:= \gamma_{S^\circ}(x) \lim_{\lambda\to 0} \frac{\gamma_{S^\circ}(x+\lambda y) - \gamma_{S^\circ}(x)}{\lambda} \qquad (40)\\
&= \sigma_S(x) \lim_{\lambda\to 0} \frac{\sigma_S(x+\lambda y) - \sigma_S(x)}{\lambda}
\end{aligned}
$$

$$= \sigma_S(x) D\sigma_S(x) \cdot y,$$

where $Df(x) = (\partial f(x)/\partial x_1, \ldots, \partial f(x)/\partial x_n)$, the Jacobian of f, is a row vector (Magnus and Neudecker, 1999, page 99).

Lumer (1961) introduced a *semi inner product* on a real vector space V as a real-valued two-place function $[\cdot,\cdot]$ satisfying the following axioms[11].

SIP1 $[x+y,z] = [x,z]+[y,z] \;\; \forall x,y,z \in V.$

SIP2 $[\lambda x,y] = \lambda[x,y] \;\; \forall \lambda \in \mathbb{R}, \; \forall x,y \in V.$

SIP3 $[x,x] \geq 0 \;\; \forall x \in V, \; x \neq 0.$

SIP4 $|[x,y]|^2 \leq [x,x][y,y] \;\; \forall x,y \in V.$

SIP5 $[x,\lambda y] = \lambda[x,y] \;\; \forall \lambda \in \mathbb{R}, \; \forall x,y \in V.$

SIP6 $[y,x+\varepsilon y] \to [y,x]$ for all real $\varepsilon \to 0$, $\forall x,y \in V$.

Unlike the standard inner product, $[\cdot,\cdot]$ is *not* symmetric: in general $[x,y] \neq [y,x]$.

Suppose $0 \in S$ then $0 \in S^\circ$. Suppose further that the principal curvatures of S and S° are all non-zero (ℓ_S and ℓ_{S° being strictly proper guarantee that). Then σ_S and σ_{S° are in C^2 (Schneider, 2014, page 115). Let $G_S := \frac{1}{2} H\gamma_{S^\circ}^2$, where H denotes the hessian: for $f\colon \mathbb{R}^n \to \mathbb{R}$, $Hf(x) := D((D(x))')$ (Magnus and Neudecker, 1999). Following Giles (1967) (who assumed γ was additionally a norm, and thus symmetric, an assumption which we drop) for $x,y \in \mathbb{R}^n$ we define

$$[y,x]_S := y' \cdot G_S(x) \cdot x.$$

Proposition 28 *Suppose S satisfies the assumptions above. Then $[\cdot,\cdot]_S$ is a semi inner product.*

Proof Since σ_S is convex, we have $H\sigma_S$ is positive semidefinite (in fact it has only one zero eigenvalue under the assumptions above (Schneider, 2014, page 118)). By properties of support functions we have that for all x

$$D\gamma_{S^\circ}(x) \cdot x = D\sigma_S(x) \cdot x = \sigma_S(x) = \gamma_{S^\circ}(x).$$

Furthermore, for all x

$$H\gamma_{S^\circ}(x) \cdot x = H\sigma_S(x) \cdot x = 0_n.$$

By the product rule, we thus have for all x

$$\tfrac{1}{2} H\gamma_{S^\circ}^2(x) = \tfrac{1}{2} D((D\gamma_{S^\circ}^2(x))') = D(D\gamma_{S^\circ}(x)' \cdot \gamma_{S^\circ}(x)) = D\gamma_{S^\circ}(x)' \cdot D\gamma_{S^\circ}(x) + \gamma_{S^\circ}(x) H\gamma_{S^\circ}(x). \tag{41}$$

Hence for all x

$$\begin{aligned} \tfrac{1}{2} H\gamma_{S^\circ}^2(x) \cdot x &= (D\gamma_{S^\circ}(x)' \cdot D\gamma_{S^\circ}(x) + \gamma_{S^\circ}(x) H\gamma_{S^\circ}(x)) \cdot x \\ &= D\gamma_{S^\circ}(x)' \cdot D\gamma_{S^\circ}(x) \cdot x \\ &= D\gamma_{S^\circ}(x)' \cdot \gamma_{S^\circ}(x), \end{aligned} \tag{42}$$

11. Semi-inner-products have been used previously in machine learning; see e.g. (Zhang et al., 2009; Der and Lee, 2007).

and consequently for all x

$$x' \cdot \frac{1}{2} H\gamma^2_{S^\circ}(x) \cdot x = x' \cdot D\gamma_{S^\circ}(x)' \cdot \gamma_{S^\circ}(x) = \gamma_{S^\circ}(x)' \cdot \gamma_{S^\circ}(x) = \gamma^2_{S^\circ}(x).$$

Since $0 \in S^\circ$, $\gamma_{S^\circ}(x) \geq 0$ for all x. Thus from (41) we see that for all x, $G_S(x)$ is the sum of two positive semidefinite matrices $\gamma_{S^\circ}(x)H\gamma_{S^\circ}(x)$ and a $D\gamma_{S^\circ}(x)' \cdot D\gamma_{S^\circ}(x)$. The positive definiteness of the rank one second term follows from the fact its only non-zero eigenvalue is $D\gamma_{S^\circ}(x) \cdot D\gamma_{S^\circ}(x)'$ and since $\gamma_{S^\circ}(0) = \sigma_S(0) = 0$, by positive homogeneity, and by convexity, for all x, $D\gamma_{S^\circ}(x) \geq 0$ (elementwise); consequently $D\gamma_{S^\circ}(x) \cdot D\gamma_{S^\circ}(x)' \geq 0$. Since $\lambda_{\min}(A+B) \geq \lambda_{\min}(A)$ for positive semidefinite A and B (Lütkepohl, 1996, 9.12.2(7)), we conclude that $G_S(x) = \frac{1}{2}H\gamma^2_{S^\circ}(x)$ is positive semidefinite for all x.

For arbitrary $x, y, z \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$ and $\varepsilon \to 0$, we have

$$\begin{aligned}
[x+y, z]_S &= (x+y)' \cdot G_S(z) \cdot z = x' \cdot G_S(z) \cdot z + y' \cdot G_S(z) \cdot z = [x,z]_S + [y,z]_S \\
[\lambda x, y]_S &= (\lambda x)' \cdot G_S(y) \cdot y = \lambda (x' \cdot G_S(y) \cdot y) = \lambda [x,y]_S \\
[x,x]_S &= x' \cdot G_S(x) \cdot x \geq 0 \text{ since } G_S(x) \text{ is positive semidefinite for all } x \\
[y,x]_S &= y' \cdot G_S(x) \cdot x = y' \cdot D\gamma_{S^\circ}(x)'\gamma_{S^\circ}(x) \leq \gamma_{S^\circ}(y)\gamma_{S^\circ}(x) = [x,x]_S^{\frac{1}{2}} [y,y]_S^{\frac{1}{2}} \\
[x, \lambda y]_S &= x' \cdot G_S(\lambda y) \cdot \lambda y = x' \cdot G_S(y) \cdot \lambda y = \lambda [x,y]_S \\
[y, x + \varepsilon y]_S &= x' \cdot G_S(x + \varepsilon y) \cdot (x + \varepsilon y) \to y' \cdot G_S(x) \cdot x \ \text{ since } \gamma_{S^\circ} \in C^2,
\end{aligned}$$

demonstrating that $[\cdot,\cdot]_S$ satisfies axioms SIP1–SIP6, where the antepenultimate line used the fact that $y' \cdot D\gamma_{S^\circ}(x)' \leq \gamma(y)$, and the penultimate line follows from γ_{S° being 1-homogeneous, implying $\gamma^2_{S^\circ}$ is 2-homogeneous and so by Euler's theorem, $H\gamma^2_{S^\circ}$ is 0-homogeneous. ■

Observe that by (42), we have

$$[y,x]_S = y' \cdot D\gamma_{S^\circ}(x)'\gamma_{S^\circ}(x) = \gamma_{S^\circ}(x)D\gamma_{S^\circ}(x) \cdot y = \sigma_S(x)D\sigma_S(x) \cdot y = \tau'_S(x,y), \tag{43}$$

by (40). One can reparametrise $[\cdot,\cdot]_S$ as follows. By (Schneider, 2014, Page 55), we have

$$\left(\frac{1}{2}\gamma^2_S\right)^* = \frac{1}{2}\gamma^2_{S^\circ},$$

where $(\cdot)^*$ is the Legendre-Fenchel conjugate (11). Combining this with the result from (Seeger, 1992) that $(Hf(x))^{-1} = Hf^*(y)$, where $y = (Df)(x)$ and $x = (Df)^{-1}(y)$, we can write

$$\frac{1}{2}H\gamma^2_{S^\circ}(u) = \frac{1}{2}(H\gamma^2_S(x))^{-1},$$

where $u = (\frac{1}{2}D\gamma^2_S(x))$ and $x = (\frac{1}{2}D\gamma^2_S)^{-1}(u)$. Observe that if $A = H\gamma^2_S$ is 0-homogeneous, then it is easy to see that A^{-1} is too. Combining (43) and (42) we have

$$[y,x]_S = y' \cdot \frac{1}{2}H\gamma^2_S(x)^{-1} \cdot x.$$

The semi-inner product $[x,y]_S$ induces a Finslerian geometry, a generalisation of Riemannanian geometry, where the norm varies throughout the space (as in the Riemannian case) but the unit balls are not necessarily ellipsoids (as in the Riemannian case). Rund (1959, page 16) notes that the metric of a Finsler space may be regarded as being locally Minkowskian — just like Riemannian geometry without the quadratic restriction (Chern, 1996).

For the situation of interest in the present paper (where we work with concave, rather than convex, gauge and support functions), we can mimic the above development to consider *anti semi inner products* $\langle\!\!\langle \cdot,\cdot \rangle\!\!\rangle$. We merely need replace the convex gauge γ by the concave gauge β, and the convex support function σ by the concave support function ρ and to reverse the inequality in the fourth axiom to read

SIP$\widehat{\mathbf{4}}$ $|\langle\!\!\langle x,y \rangle\!\!\rangle|^2 \geq \langle\!\!\langle x,x \rangle\!\!\rangle \langle\!\!\langle y,y \rangle\!\!\rangle, \quad \forall x,y \in V$.

We keep the other axioms the same. Define

$$\langle\!\!\langle y,x \rangle\!\!\rangle_S := y' \cdot \widehat{G}_S(x) \cdot x,$$

where $\widehat{G}_S := \frac{1}{2} H \beta^2_{S^\diamond}$. Then by a similar argument to the above, we have that $\langle\!\!\langle y,x \rangle\!\!\rangle_S$ is indeed an anti semi inner product. The only change is working with negative semidefiniteness instead of positive, and the use of the "reverse" inequality $D\rho_{S^\diamond} \cdot y \geq \rho_{S^\diamond}(x)$, which is a rephrasing of the result with loss functions that $L(x,y) \geq \underline{L}(x)$.

Following the argument in (Lumer, 1961), but reversing the inequalities, we have that $\langle\!\!\langle x,x \rangle\!\!\rangle^{\frac{1}{2}}$ is a concave gauge function. Translating to our loss notation we can write

$$\langle\!\!\langle y,x \rangle\!\!\rangle_S = \underline{L}_S(x) L_S(y,x), \tag{44}$$

which can be seen to be a weighted conditional risk. Utilising the formula for Bregman divergences (39) and by analogy with (43), we have (writing $\mathrm{B}_S := \mathrm{B}_{\underline{L}_S}$) that $\langle\!\!\langle y,x \rangle\!\!\rangle_S = \rho_S(x) D\rho(x) \cdot y = \underline{L}_S(x) \langle \ell_S(x), y \rangle = \underline{L}_S(x) L_S(y,x)$, and thus

$$\begin{aligned} \underline{L}(x)\, \mathrm{B}_S(y,x) &= \underline{L}_S(x)[L_S(y,x) - \underline{L}_S(y)] \\ &= \underline{L}_S(x) L_S(y,x) - \underline{L}_S(x) \underline{L}_S(y) \\ &= \langle\!\!\langle y,x \rangle\!\!\rangle_S - \sqrt{\langle\!\!\langle x,x \rangle\!\!\rangle_S \langle\!\!\langle y,y \rangle\!\!\rangle_S}. \end{aligned}$$

Expressing loss-theoretic quantities in terms of $\langle\!\!\langle \cdot,\cdot \rangle\!\!\rangle_S$ and hence $\widehat{G}_S$ may also be conceptually valuable, but we defer further investigation of this. The normalisation that naturally arises in (44) has not, to our knowledge, arisen previously. This does suggest that Riemannian geometry, the traditional foundation of "information geometry" (Amari, 2016), is not quite the right fit for the geometry of losses, and we instead need the richer, locally Minkowskian (Chern, 1996), notion of Finslerian geometry (Rund, 1959).

3.4 The Antipolar Loss

A loss function ℓ maps a distribution $p \in \mathbb{R}^n_{>0}$ to a loss vector $\ell(p) \in \mathbb{R}^n_{\geq 0}$. Given $\ell(p)$, one might ask if we can recover p? This problem arises naturally in a variety of settings (Vovk, 2001; Gneiting and Katzfuss, 2014). In practice it might be difficult to find or even show the existence of an inverse loss, but in light of Remark 26 we can show the existence of, and suggest several ways to calculate, a pseudoinverse, $\ell^\diamond$. For reasons that will become clear, we call the function $\ell^\diamond$ the *antipolar loss*.

The antipolar loss provides a universal substitution function (Kamalaruban et al., 2015) for the Aggregating Algorithm (Vovk, 2001, 1995, 1990). The substitution function needs to map an arbitrary superprediction $x \in \operatorname{spr}(\ell)$ to a prediction $p \in \Delta$ such that $l := \ell(p)$ dominates x in the sense that $l \preceq_{\mathbb{R}^n_{\geq 0}} x$ (that is, pointwise inequality not incurring more loss under each $y \in \mathscr{Y}$). Explicitly stating a substitution function is the primary difference between the (unrealisable) aggregating "pseudo-algorithm" and the aggregating "algorithm" (Vovk, 2001). Determining the substitution function even for simple cases can be difficult (Zhdanov, 2011). We show below that by making use of antipolars, we can determine such substitution functions; see (Friedlander et al., 2014; Aravkin et al., 2018) for further uses of antipolarity.

Proposition 29 *Let $\ell\colon \mathbb{R}^n_{>0} \to \mathbb{R}^n_{\geq 0}$ be a proper loss. There exists $\ell^\diamond\colon \mathbb{R}^n_{>0} \to \mathbb{R}^n_{\geq 0}$ with $\ell^\diamond \in \partial \rho_{(\operatorname{spr}\ell)^\diamond}$ that satisfies*

$$\forall p \in \mathbb{R}^n_{>0},\ \ell(p) = (\ell \circ \ell^\diamond \circ \ell)(p) \quad \textit{and} \quad \forall x \in \mathbb{R}^n_{>0},\ \ell^\diamond(x) = (\ell^\diamond \circ \ell \circ \ell^\diamond)(x). \tag{45}$$

Furthermore $\ell^\diamond$ is a proper loss.

Proof Let $S := \operatorname{spr}(\ell)$. Since the subdifferential of a support function is 0-homogenous (Proposition 3) applying Lemma 10 we have $\forall p \in \mathbb{R}^n_{>0}$,

$$\begin{aligned}\ell(p) \in \partial \rho_S(p) \iff & \frac{\ell(p) \cdot \rho_{S^\diamond}(\ell(p))}{\rho_{S^\diamond}(\ell(p))} \in \partial \rho_S(p) \\ \overset{\text{L10}}{\iff} & \frac{p}{\rho_S(p)} \in \partial \rho_{S^\diamond}(\ell(p) \cdot \rho_{S^\diamond}(\ell(p))) = \partial \rho_{S^\diamond}(\ell(p)).\end{aligned}$$

Let us choose $\ell^\diamond$ to satisfy $\ell^\diamond(\ell(p)) = p \cdot {}^1\!/_{\rho_S(p)}$. This defines $\ell^\diamond\colon \mathbb{R}^n_{>0} \to \mathbb{R}^n_{\geq 0}$. We can then extend $\ell^\diamond$ to $\mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ using the argument of Remark 22. Note that $\ell(\ell^\diamond(\ell(p))) = \ell\big(p \cdot {}^1\!/_{\rho_S(p)}\big) = \ell(p)$. For the second claim regarding $\ell^\diamond$, exchange the roles of ℓ and $\ell^\diamond$ in the above argument and apply Proposition 3.

We now argue that $\ell^\diamond$ is proper. By construction $\ell^\diamond(\ell(p)) = p/\rho_S(p) \in \partial \rho_{S^\diamond}(\ell(p))$. Let $q = \ell(p)$. Then $\ell^\diamond(q) \in \partial \rho_{S^\diamond}(q)$. Proposition 21 then implies that $\ell^\diamond|_\Delta$ is proper as long as $S^\diamond \in \mathcal{P}(\mathbb{R}^n_{\geq 0})$ which we will now show. We have

$$\mathcal{P}(\mathbb{R}^n_{\geq 0}) = \{S \in \mathcal{K}(\mathbb{R}^n_{\geq 0} \setminus \{0\}) \,|\, \operatorname{rec}(S) = \mathbb{R}^n_{\geq 0}\}.$$

By the observation following (13), the map $S \mapsto S^\diamond$ takes closed shady sets to closed shady sets and $S \in \mathcal{S}(X_+) \Rightarrow S^\diamond \in \mathcal{S}(X_+^*)$. Furthermore $S^\diamond$ is convex. (Suppose $x_0^*, x_1^* \in S^\diamond$ and for some $\lambda \in (0,1)$, let $x_\lambda^* := \lambda x_1^* + (1-\lambda)x_0^*$, then it is straightforward to check that $x_\lambda^* \in S^\diamond$ from the definition of $S^\diamond$.) Finally we have that

$$\begin{aligned}\operatorname{rec}(S^\diamond) &= \{d \in \mathbb{R}^n \,|\, \forall x \in S,\ \langle x^*, x\rangle \geq 1 \Rightarrow \langle x^* + d, x\rangle \geq 1\} \\ &= \{d \in \mathbb{R}^n \,|\, \forall x \in S,\ \langle x^*, x\rangle \geq 1 \Rightarrow \langle x^*, x\rangle + \langle d, x\rangle \geq 1\} \\ &= \{d \in \mathbb{R}^n \,|\, \forall x \in S,\ \langle d, x\rangle \geq 1\} \\ &= \{d \in \mathbb{R}^n \,|\, \forall x \in S,\ \langle d, x\rangle \geq 0\} \\ &= \bigcap_{x \in S} \{d \in \mathbb{R}^n \,|\, \langle d, x\rangle \geq 0\} \\ &= \mathbb{R}^n_{\geq 0},\end{aligned}$$

where the last line follows from the fact that $\operatorname{rec}(S) = \mathbb{R}^n_{\geq 0}$. ■

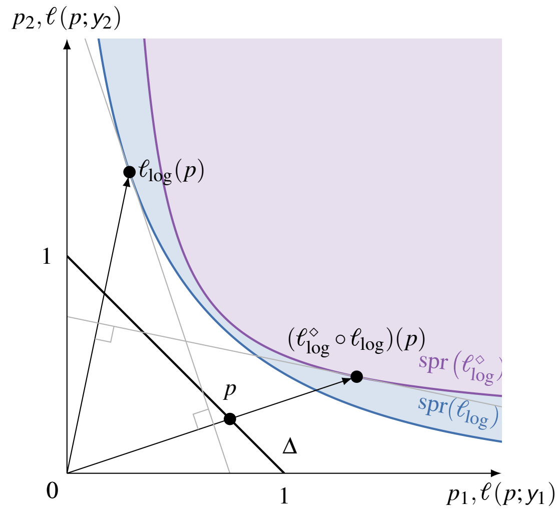


FIGURE 6: Illustration of Proposition 29 using $\ell_{\log}$ (Table 3) with $\mathscr{Y} = [2]$. Upon predicting $p \in \Delta$, we receive the loss vector $\ell_{\log}(p)$. Evaluating the antipolar loss $\ell_{\log}^{\diamond}$ at $\ell_{\log}(p)$ we have the point $(\ell_{\log}^{\diamond} \circ \ell_{\log})(p) = \alpha p$ for some $\alpha > 0$. However, $\ell_{\log}$ is 0-homogeneous (Proposition 21) and so $(\ell_{\log} \circ \ell_{\log}^{\diamond} \circ \ell_{\log})(p) = \ell_{\log}(p)$ for all $p \in \mathbb{R}^2_{>0}$.

Proposition 29 is illustrated with $\ell_{\log}$ in Figure 6, which should be compared with the work of Shephard (1953, pg. 23) which was the inspiration for the argument regarding antipolar losses in the present paper[12].

More generally, it follows from Lemma 10 that if ℓ is a mapping $\mathrm{ri}(X_+) \to X_+^*$ with $\mathrm{spr}(\ell) \in \mathcal{P}(X_+)$ then $\ell^{\diamond}$ is a mapping $\mathrm{ri}(X_+^*) \to X_+$ with $\mathrm{spr}(\ell^{\diamond}) \in \mathcal{P}(X_+^*)$. The pseudoinverse property of (45) can be expressed using the notion of the direction of a vector as in footnote 9, allowing us to write for all $p \in \mathbb{R}^n_{>0}$, $\mathrm{dir}(\ell^{\diamond} \circ \ell)(p) = \mathrm{dir}\, p$.

12. This has become known as “Shephard’s duality theorem” in the economics literature (Shephard, 1970; Jacobsen, 1972; McFadden, 1978; Hanoch, 1978; Cornes, 1992; Färe and Primont, 1994, 1995; Penot, 2005; Zălinescu, 2016) and appears in standard microeconomics texts (Varian, 1978). Shephard’s development of dual theory in economics in his 1953 book (Shephard, 1953) was described as “one of the most original contributions to economic theory of all time” (Jorgenson, 1981) due to three key ideas:

1. The duality between “cost” and “production” functions (essentially polar duality of concave gauge functions);
2. Shephard’s lemma (Varian, 1978, page 74), (Mas-Collel et al., 1995, Page 141): essentially the result (Schneider, 2014) that the subgradient of a support function (in economics terminology, “cost function” evaluated at some fixed price of input vectors) $\partial\sigma(x)$ is the support set (“conditional factor demand correspondence” evaluated at the same price vector), and furthermore if σ is differentiable at x, the support set is a singleton;
3. Homotheticity: essentially that the key functions of the theory are a composition of a positive monotone increasing scalar function and a positively homogeneous function of several variables.

The economic theory tends to obscure the simplicity of concave gauge duality because of the need to parametrise families of sets (either by the vector of inputs available to a firm or the vector of outputs) and the adoption of convoluted terminology (“conditional factor demand correspondence” instead of “support set”). The geometry in all cases is simply that of concave gauge duality, a point explicitly recognised by Hasenkamp and Schrader (1978) in the context of aggregation problems arising in production economics.

Proposition 30 *Let $\ell\colon \mathbb{R}^n_{>0} \to \mathbb{R}^n_{\geq 0}$ be a proper loss. Then $\rho_{\mathrm{spr}(\ell)}$ is strictly concave if and only if $\ell^\diamond$ is strictly proper.*

Proof This follows immediately from Corollaries 12 and 23. ■

Corollary 31 *If $\rho_{\mathrm{spr}(\ell)}$ is strictly concave then for any function $\mathcal{m}$ that satisfies $\mathcal{m} \in \partial\rho_{(\mathrm{spr}\,\ell)^\diamond}$ we have $\mathcal{m} = \ell^\diamond$.*

Proof It follows that $\mathrm{spr}(\ell)^\diamond$ is strictly convex (Corollary 12), thus $\partial\rho_{\mathrm{spr}(\ell)^\diamond}$ is always a singleton (Corollary 5). Thus there is only one possible selection $\mathcal{m}$ with $\mathcal{m} \in \partial\rho_{(\mathrm{spr}\,\ell)^\diamond}$. ■

Remark 32 If $\ell^\diamond$ is the antipolar of the loss ℓ then clearly so is the family $(\alpha\ell^\diamond)_{\alpha>0}$ since as ℓ is 0-homogeneous $\ell\circ(\alpha\ell^\diamond) = \ell\circ\ell^\diamond$.

There are three ways the antipolar loss $\ell^\diamond$ can be computed given a proper loss ℓ (cf. §2.7): we can

1. take the superdifferential of the concave support function of the antipolar superprediction set $\ell \mapsto \partial\rho_{(\mathrm{spr}\,\ell)^\diamond} \ni \ell^\diamond$;
2. compute the antipolar of the associated conditional Bayes risk function and superdifferentiate $\underline{L} \mapsto \partial\underline{L}^\diamond \ni \ell^\diamond$; or
3. solve the optimisation problem in

$$\ell^\diamond(p) \in \arg\inf\left\{x \in \mathbb{R}^n_{\geq 0} \,\middle|\, \langle \ell(p), x\rangle \geq 1\right\}.$$

The complete set of antipolar loss function relationships is presented in Figure 7. The notion of the antipolar loss and its relationship to the concave polar of the superprediction set provides conceptual insight. Furthermore, at least in some cases one can determine the inverse in closed form (see equation 46 in §4.1, as well as the other examples in §4.2 and §4.3).

3.5 Convexifying Proper Losses: The Canonical Link

All proper losses ℓ have convex superprediction sets, but that does not imply that the partial functions $\ell_i = \ell(\cdot;i)$ are convex for all $i \in [m]$ (Reid and Williamson, 2010; Vernet et al., 2016). However, such proper losses with non-convex partial losses can be made convex by reparametrisation.

A *composite proper loss* $\ell\circ\psi^{-1}$ is the composition of a proper loss ℓ and an (inverse) "link function" ψ^{-1} that reparametrizes the loss (Reid and Williamson, 2010; Vernet et al., 2016). The aforementioned papers studied such links using the tools of differential calculus. We will now show that the geometric perspective of the present paper, along with the properties of antipolar losses, allows a simpler proof of the fact that there is always a special link function which ensures the resulting composite loss is a convex function. This link function is called the "canonical link" in (Reid and Williamson, 2010) (binary case) and (Vernet et al., 2016) (general multiclass case); as we

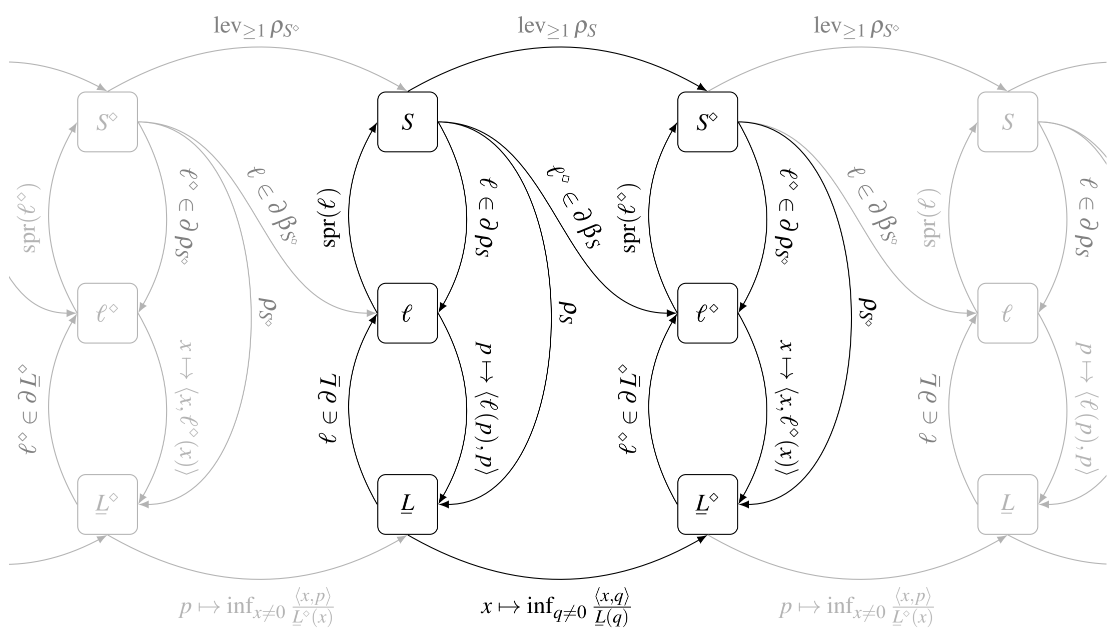


FIGURE 7: Illustration of the relationships superprediction sets S, associated loss functions ℓ, and conditional Bayes' risks $\underline{L}$, along with their antipolar counterparts. (The diagram, which is a finite graph, has been "unrolled" to make reading easier.)

shall see below, the canonical loss is indeed the composition of the loss with its associated antipolar loss.

We first need some additional notions (Pennanen, 1999; Gissler and Hoheisel, 2022). Suppose X and Y are sets, and $K \subset Y$ is a convex cone, and $f\colon X \to Y$ with $\operatorname{dom} f$ convex. Recalling from §2.4 the ordering $\succeq_K$, we say f is *K-convex* if for all $x_0, x_1 \in X$, and all $\alpha \in (0,1)$,

$$f(\alpha x_1 + (1-\alpha)x_0) \preceq_K \alpha f(x_1) + (1-\alpha) f(x_0).$$

That is, $f(\alpha x_1 + (1-\alpha)x_0) - \alpha f(x_1) + (1-\alpha) f(x_0) \in -K$. The *K-epigraph* of a function $f\colon X \to Y$ is

$$\operatorname{epi}_K f := \{(x,y) \in X \times Y \,|\, f(x) \preceq_K y\}.$$

For functions $f\colon X \to \mathbb{R}$, the traditional epigraph $\operatorname{epi} f$ corresponds to $\operatorname{epi}_K f$ with $K = \mathbb{R}_{\leq 0}$. Analogous to the result for the traditional epigraph that f is convex iff $\operatorname{epi} f$ is, we have (Jahn, 2011, Lemma 14.8):

Lemma 33 *Suppose $f\colon X \to Y$ and $\operatorname{dom} f$ is convex, and $K \subset Y$ is a cone. Then f is K-convex if and only if $\operatorname{epi}_K f$ is a convex subset of $X \times Y$.*

Suppose $f\colon \mathbb{R}^n \to \mathbb{R}^n$, and write $f(x) = (f_1(x), \ldots, f_n(x))$. Let $K = \mathbb{R}^n_{\geq 0}$. Then f is K-convex if

$$\begin{aligned} & f(\alpha x_1 + (1-\alpha)x_0) - \alpha f(x_1) - (1-\alpha) f(x_0) \in \mathbb{R}^n_{\leq 0} \\ \Rightarrow\ & \forall i \in [n],\ \ f_i(\alpha x_1 + (1-\alpha)x_0) - \alpha f_i(x_1) - (1-\alpha) f_i(x_0) \in (\infty, 0] \end{aligned}$$

$$
\begin{aligned}
&\Rightarrow \ \forall i \in [n], \quad f_i(\alpha x_1 + (1-\alpha)x_0) \le \alpha f_i(x_1) + (1-\alpha) f(x_0) \\
&\Rightarrow \ \forall i \in [n], \quad f_i \text{ is convex.}
\end{aligned}
$$

Hence (confer (Gissler and Hoheisel, 2022, Section 4.4.4)) f is *component-wise convex*:

Lemma 34 *$f\colon \mathbb{R}^n \to \mathbb{R}^n$ is $\mathbb{R}^n_{\ge 0}$-convex iff $f_i\colon \mathbb{R}^n \to \mathbb{R}$ is convex for all $i \in [n]$.*

Let $\tilde{\ell} := \ell \circ \ell^{\diamond}$ denote the *canonical loss* induced by composing an arbitrary proper loss ℓ with its associated antipolar loss $\ell^{\diamond}$. Proposition 29 implies there exists a function $\gamma_{\tilde{\ell}}\colon \mathbb{R}^n_{\ge 0} \to \mathbb{R}_{>0}$ such that for all $x \in \mathbb{R}^n_{\ge 0}$, $\tilde{\ell}(x) = \gamma_{\tilde{\ell}}(x)x$.

We can now prove the following result directly, without the need for differential calculus as used in (Vernet et al., 2016, Corollary 32 *et seq.*).

Theorem 35 *Suppose ℓ is a proper loss. Then the canonical loss $\tilde{\ell}$ is component-wise convex.*

Proof By Lemmas 33 and 34, $\tilde{\ell}$ is component-wise convex if $\operatorname{epi}_K \tilde{\ell}$ is convex, with $K = \mathbb{R}^n_{\ge 0}$, which we now proceed to show. Write $\gamma = \gamma_{\tilde{\ell}}$. Then

$$
\begin{aligned}
\operatorname{epi}_K \tilde{\ell} &= \{(x,y) \in \mathbb{R}^n_{\ge 0} \times \mathbb{R}^n_{\ge 0} \,|\, \tilde{\ell}(x) \preceq_K y\} \\
&= \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \{(x,y) \,|\, y \in \mathbb{R}^n_{\ge 0},\ \tilde{\ell}(x) \preceq_K y\} \\
&= \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \{(x,y) \,|\, y \in \mathbb{R}^n_{\ge 0},\ \gamma(x)x \preceq_K y\} \\
&= \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \{(x,y) \,|\, y \in \mathbb{R}^n_{\ge 0},\ x \preceq_K y/\gamma(x)\} \\
&= \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \{(x,\gamma(x)y') \,|\, y' \in \mathbb{R}^n_{\ge 0},\ x \preceq_K y'\} \\
&= \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \{x\} \times \gamma(x)(\{x\} + \mathbb{R}^n_{\ge 0}) \\
&= \mathbb{R}^n_{\ge 0} \times \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \gamma(x)(\{x\} + \mathbb{R}^n_{\ge 0}) \\
&= \mathbb{R}^n_{\ge 0} \times \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \gamma(x)\{y \in \mathbb{R}^n_{\ge 0} \,|\, x \preceq_K y\} \\
&= \mathbb{R}^n_{\ge 0} \times \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \{y \in \mathbb{R}^n_{\ge 0} \,|\, x \preceq_K y/\gamma(x)\} \\
&= \mathbb{R}^n_{\ge 0} \times \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \{y \in \mathbb{R}^n_{\ge 0} \,|\, \gamma(x)x \preceq_K y\} \\
&= \mathbb{R}^n_{\ge 0} \times \bigcup_{x \in \mathbb{R}^n_{\ge 0}} \{y \in \mathbb{R}^n_{\ge 0} \,|\, \tilde{\ell}(x) \preceq_K y\} \\
&= \mathbb{R}^n_{\ge 0} \times \bigcup_{l \in \ell(\Delta)} \{y \in \mathbb{R}^n_{\ge 0} \,|\, l \preceq_K y\} \\
&= \mathbb{R}^n_{\ge 0} \times \operatorname{spr} \tilde{\ell},
\end{aligned}
$$

by (29). By (Hiriart-Urruty and Lemaréchal, 2001, Proposition A.1.2.3), the Cartesian product of convex sets is convex. Since $\mathbb{R}^n_{\geq 0}$ and $\operatorname{spr}\tilde{\ell}$ are both convex, we have thus shown that $\operatorname{epi}_K \tilde{\ell}$ is also convex, concluding the proof. ■

3.6 The Naturalness of our Setup, and its Advantages

The above development presumes a particular form for S, namely that it is in the class $\mathcal{P}(\mathbb{R}^n_{\geq 0})$. This assumption is necessary for our proofs, but is it reasonable? Furthermore, the definition of a loss function as the subgradient of ρ_S is rather unusual. What advantage does it have? Finally, the introduction of the link function seems to just complicate matters even further. What additionality does it bring? In this brief subsection we provide succinct answers to these natural questions.

The choice of $\mathcal{P}(\mathbb{R}^n_{\geq 0})$ is indeed simply justified by the results obtainable: if S is not convex, then $\rho_S = \rho_{\operatorname{co} S}$ so there is nothing lost in assuming convexity. Since we only use the loss via its average (the risk) we are thus only interested in supporting hyperplanes with normals in the positive orthant; thus the assumption on the recession cone simply allows unbounded losses to exist, and ensures they are bounded on the relative interior of the simplex.[13] Some advantages of defining the loss as $\partial \rho_S$ will be elaborated in the remainder of the paper (in terms of designing losses), but it also allows a direct connection to the question of mixability of a loss, which is defined in terms of the geometry of S (van Erven et al., 2012), as well as to the dual theory of production economics (as elaborated in footnote 12). Without the inherent geometrical structure, it seems unlikely one would have stumbled across the notion of an antipolar loss. Finally, the representation of a general (not necessarily proper) loss ℓ, such that $\operatorname{spr}\ell \in \mathcal{P}(\mathbb{R}^n_{\geq 0})$, as $\ell = \lambda \circ \psi^{-1}$ (where λ is proper, and ψ^{-1} is an inverse link function) allows a very clean separation of concerns: the statistical properties of the loss are controlled by λ, and the convexity of the partial losses ℓ_i, $i \in [n]$ is controlled by ψ; confer (Vernet et al., 2016).

4. Examples of the Antipolar Loss

We now present some examples of the antipolar loss for some well-known standard loss functions.

4.1 Concave Norm Losses

In general the calculation of antipolars of superprediction sets, or equivalently the antipolar loss may be difficult to achieve in closed form. However, analogous to the case of classical l_p gauges (norms) with the duality property that $\mathrm{B}_p^\circ = \mathrm{B}_q$ with $\frac{1}{p} + \frac{1}{q} = 1$, there is a parametric family of antigauges which has an attractive self-closure property with respect to taking antipolars. Following Barbara and Crouzeix (1994), we define $\beta_a \colon \mathbb{R}^n_{\geq 0} \to \overline{\mathbb{R}}$ for $a \in [-\infty, 1] \setminus \{0\}$ as follows

$$\beta_{-\infty}(p) := \min_{y \in \mathscr{Y}} p_y$$

$$\beta_a(p) := \begin{cases} \left(\sum_{y \in \mathscr{Y}} p_y^a\right)^{1/a} & p \in \mathbb{R}^n_{>0} \\ 0 & p \in \operatorname{bd}\mathbb{R}^n_{\geq 0}, \end{cases} \qquad \forall a \in (-\infty, 0)$$

13. For a much more careful treatment of unbounded losses, see (Waggoner, 2021).

$$\beta_a(p) := \left(\sum_{y\in\mathscr{Y}} p_y^a\right)^{1/a} \qquad\qquad \forall a \in (0,1].$$

Barbara and Crouzeix (1994) showed for all $a \in [-\infty,1] \setminus \{0\}$ that β_a is indeed an antigauge and furthermore

$$\frac{1}{a}+\frac{1}{b}=1 \iff \beta_a^\diamond = \beta_b. \tag{46}$$

Note that if $a \in (0,1]$ then $b \in [-\infty,0)$, and if $a \in [-\infty,0)$ then $b \in (0,1]$. Thus no antigauge in the family $\{\beta_a \,|\, a \in [-\infty,1]\setminus\{0\}\}$ can be its own antipolar (unlike the classical result that $\mathrm{B}_2^\circ = \mathrm{B}_2$). The family of antigauges $\{\beta_a \,|\, a \in [-\infty,1]\setminus\{0\}\}$ can be used to define a family of proper losses on n outcomes. Since β_a is an antigauge, it can be written either as the antigauge of the set $\mathrm{lev}_{\geq 1}(\beta_a)$, or using polar duality as the concave support function of the set $\mathrm{lev}_{\geq 1}(\beta_a^\diamond)$, which is the convention we follow below.

In order to find the associated loss function we need ℓ_a such that $\mathrm{spr}(\ell_a) = \mathrm{lev}_{\geq 1}(\beta_a^\diamond)$. The self antipolar property (46) makes this easy. Since $\beta_{\frac{a}{a-1}}$ is differentiable on the interior of its domain we have

$$\forall a \in (-\infty,1)\setminus\{0\}, \forall p \in \mathbb{R}^n_{>0},\ \ell_a(p) = \nabla\beta_{\frac{a}{a-1}}(p) = \exp\left(\tfrac{1}{\beta_{\frac{a}{a-1}}(p)}\cdot p;\, \tfrac{1}{a-1}\right),$$

where the exponentiation of vectors is defined component wise, $\exp(p;a) := (p_1^a,\ldots,p_n^a)$. Applying the same procedure to the antipolar loss we have

$$\forall a \in (-\infty,1)\setminus\{0\},\ \forall p \in \mathbb{R}^n_{>0},\ \ell_a^\diamond(p) = \nabla\beta_a(p) = \exp\left(\tfrac{1}{\beta_a(p)}\cdot p;\, a-1\right).$$

There are two special values of $a \in [-\infty,1]\setminus\{0\}$ worth mentioning: When $a=1$, $b=-\infty$ and β_b is no longer differentiable, but it is superdifferentiable, with

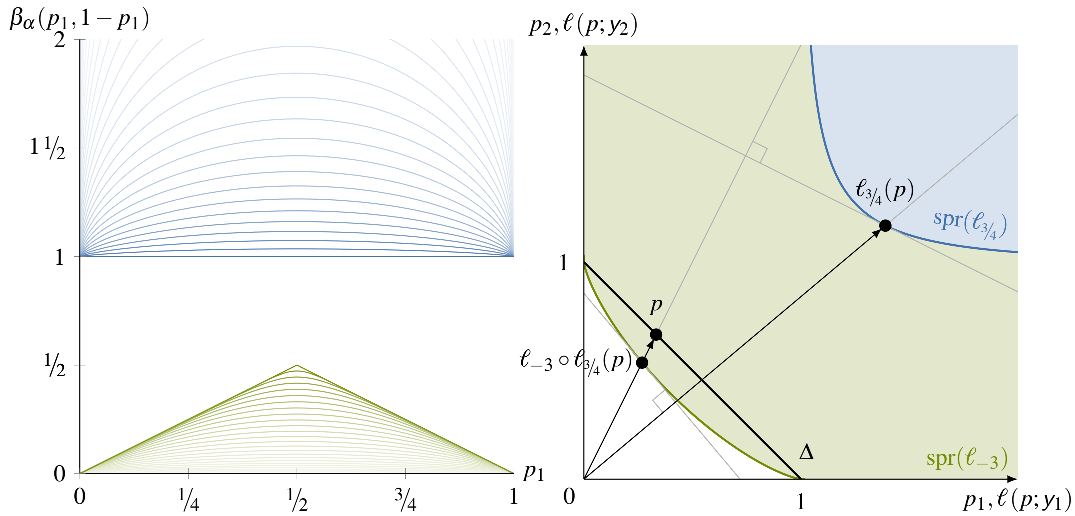
$$\forall p \in \mathbb{R}^n_{>0},\ \ell_{0/1}(p) \in \partial\beta_{-\infty}(p).$$

When $a=-\infty$, $b=1$ and $\nabla\beta_1(p) = 1_n$. Thus

$$\forall p \in \mathbb{R}^n_{>0},\ \ell_{-\infty}(p) := 1_n \in \partial\beta_1(p),$$

the *constant loss*. Note that $\ell_{-\infty} = \ell_{0/1}^\diamond$. The closure of the family $(\ell_a)_{a\in[-\infty,1]\setminus\{0\}}$ under the antipolar operation is illustrated in Figure 8.

Example 1 Misclassification loss is not *strictly* proper and so $\partial\rho_{\mathrm{spr}(\ell_{0/1})}(p)$ will not be a singleton for all $p \in \Delta$. This poses a problem for calculating $\ell_{0/1}^\diamond$, since the antipolar superprediction set subdifferential $\partial\rho_{\mathrm{spr}(\ell_{0/1}^\diamond)}(p)$ will not be a singleton. However, we can use the family $(\ell_a)_{a\in[-\infty,1]\setminus\{0\}}$ to approximate $\ell_{0/1}$, and therefore approximate the antipolar. That is we can come arbitrarily close to obtaining the pair $(\ell_{0/1}(p), \ell_{0/1}^\diamond(p))$ with the sequence $(\ell_a(p), \ell_{\frac{a}{a-1}}(p))_{a<0}$ for $p \in \mathbb{R}^n_{>0}$. The pointwise limit is illustrated explicitly in Figure 9.



(a) The functions $(\beta_a)_{a\in[-\infty,0)}$ are coloured from blue to white, and their antipolar counterparts, $(\beta_a)_{a\in(0,1]} \simeq (\beta_a^\diamond)_{a\in[-\infty,0)}$, are coloured from white to green. As $a \nearrow 1$, $\ell_a \to \ell_{0/1}$, and as $a \to -\infty$, $\ell_\alpha \to 1_2$, the constant loss.

(b) The loss $\ell_{3/4}$ and its antipolar ℓ_{-3} acting on the vector $p := ({}^1\!/_3, {}^2\!/_3)$.

FIGURE 8: The concave norm conditional Bayes risk functions β_a and losses ℓ_a, and an illustration of self-polarity of the family $\{\ell_a\}_{a\in[-\infty,1]\setminus\{0\}}$.

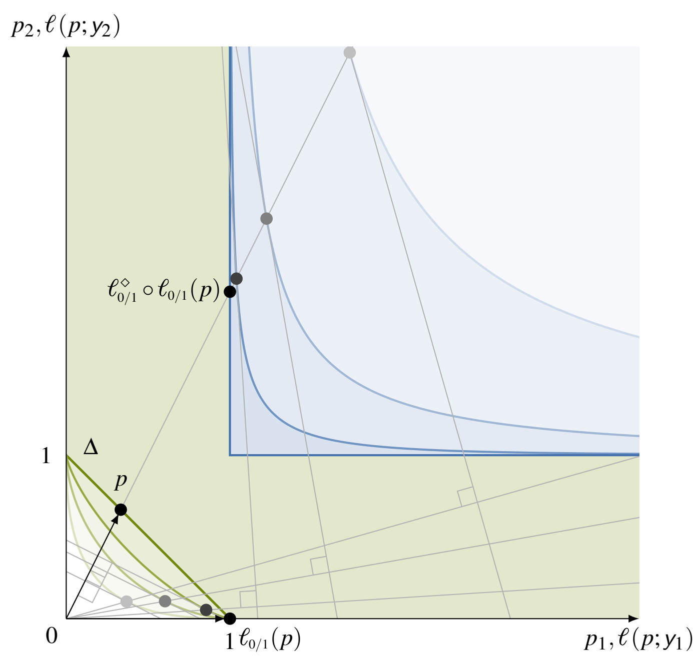


FIGURE 9: Illustration of Example 1 with $p := ({}^1\!/_3, {}^2\!/_3)$. We can simultaneously approximate misclassification loss, $\ell_{0/1}$, and its antipolar, $\ell_{0/1}^\diamond$, over Δ using the concave norm losses.

4.2 Brier Loss

The Brier score (Brier, 1950) is usually defined for $p \in \Delta$ in terms of its conditional Bayes risk (van Erven et al., 2012, Section 5), but for our purposes we need to work with the 1-homogeneous extension to $\mathbb{R}^n_{\geq 0}$:

$$\forall p \in \mathbb{R}^n_{\geq 0},\ \rho_{\mathrm{spr}(\ell_{\mathrm{Br}})}(p) = \begin{cases} \|p\|_1 - \frac{\|p\|_2^2}{\|p\|_1} & p \in \mathbb{R}^n_{\geq 0} \setminus \{0\} \\ 0 & p = 0. \end{cases} \tag{47}$$

Indeed we have $\ell_{\mathrm{Br}} \in \partial \rho_{\mathrm{spr}(\ell_{\mathrm{Br}})}$:

$$\begin{aligned} \forall p \in \mathbb{R}^n_{>0},\ \partial \rho_{\mathrm{spr}(\ell_{\mathrm{Br}})}(p) &= \partial(\|\cdot\|_1)(p) - \frac{1}{\|p\|_1^2}\Big(\|p\|_1 \cdot (\nabla \|\cdot\|_2^2)(p) + \partial(-\|p\|_2^2 \cdot \|\cdot\|_1)(p)\Big) \\ &= \partial(\|\cdot\|_1)(p) - \frac{1}{\|p\|_1^2}\Big(\|p\|_1 \cdot 2p - \|p\|_2^2 \cdot \partial(\|\cdot\|_1)(p)\Big) \\ &= \frac{1}{\|p\|_1}\left(\|p\|_1 + \frac{\|p\|_2^2}{\|p\|_1}\right) \cdot \partial(\|\cdot\|_1)(p) - \frac{2p}{\|p\|_1} \\ &\ni \left(1 + \frac{\|p\|_2^2}{\|p\|_1^2}\right) 1_n - 2\frac{p}{\|p\|_1}, \end{aligned}$$

where to compute the subdifferential of (47) we used the concave subdifferential quotient rule (Mordukhovich and Shao, 1995, Theorem 5.2). Thus

$$\forall p \in \mathbb{R}^n_{>0},\ \ell_{\mathrm{Br}}(p) = \left(1 + \frac{\|p\|_2^2}{\|p\|_1^2}\right) 1_n - 2\frac{p}{\|p\|_1}.$$

Example 2 When $n = 2$ we can determine the Brier loss antipolar explicitly. Since we know $\rho^{\diamond}_{\mathrm{spr}(\ell_{\mathrm{Br}})}$ is 1-homogeneous, it suffices to evaluate it on the 2-simplex Δ and then 1-homogeneously extend it. Parametrising an element $p \in \Delta$ as $p = (p_1, 1 - p_1)$ it follows that

$$\begin{aligned} \rho_{\mathrm{spr}(\ell^{\diamond}_{\mathrm{Br}})}(p) &= \inf_{q \neq 0} \frac{\langle p, q\rangle}{\rho_{\mathrm{spr}(\ell_{\mathrm{Br}})}(q)} \\ &= \inf_{q \in \mathrm{cl}\,\Delta} \frac{q_1 p_1 + q_2(1-p_1)}{1 - (q_1^2 + q_2^2)} \\ &= \inf_{0 \leq q_1 \leq 1} \frac{q_1 p_1 + (1-q_1)(1-p_1)}{1 - q_1^2 - (1-q_1)^2}. \end{aligned} \tag{48}$$

This can be computed directly, resulting in

$$\forall p \in \Delta,\ \rho_{\mathrm{spr}(\ell^{\diamond}_{\mathrm{Br}})}(p) = f(p_1) := \frac{(2p_1 - 1)^2 \sqrt{p_1(1-p_1)}}{4p_1^2 + 2\sqrt{p_1(1-p_1)} - 4p_1},$$

and thus for $p = \alpha(p_1, 1 - p_1) \in \mathbb{R}^2_{\geq 0}$, $\rho_{\mathrm{spr}(\ell^{\diamond}_{\mathrm{Br}})}(p) = \alpha f(p_1)$[14]. The Brier loss and its polar are illustrated in Figure 10.

14. The explicit form of the loss itself can be obtained by differentiation of $\rho_{\mathrm{spr}(\ell^{\diamond}_{\mathrm{Br}})}(p)$ and restriction to the simplex. Although little insight seems gleanable from the formula, we present it for completeness. We have for all $q \in [0,1]$,

$$\ell^{\diamond}_{\mathrm{Br},1}(q) = \frac{(2q-1)\Big(2q^2\sqrt{q(1-q)} - 2q^2(1-q) + 3q(1-q) - (q+1)\sqrt{q(1-q)}\Big)}{2\Big(2q^2 + \sqrt{q(1-q)} - 2q\Big)^2}$$

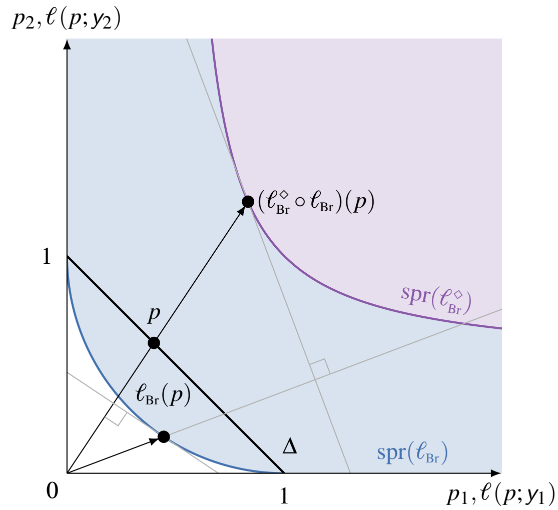


FIGURE 10: Brier loss and its antipolar acting on the vector $p \coloneqq (2/5, 3/5)$.

It does not seem possible to find a closed form for $\ell^{\diamond}_{\mathrm{Br}}$ when $n > 2$. However the objective function in (48) can be seen to be quasi-convex in p (since $\rho_{\mathrm{spr}(\ell_{\mathrm{Br}})}$ is concave and positive and thus $1/\rho_{\mathrm{spr}(\ell_{\mathrm{Br}})}$ is quasi-convex) and therefore is amenable to numerical solution.

4.3 Cobb-Douglas Loss

As a final example, let $a \in \mathbb{R}^n_{\geq 0}$ and consider the parametrised superlinear function

$$\mathbb{R}^n \ni p \mapsto \psi_a(p) \coloneqq \begin{cases} \prod_{i \in [n]} p_i^{a_i/\|a\|_1} & p \in \mathbb{R}^n_{\geq 0} \\ -\infty & \text{otherwise.} \end{cases}$$

Barbara and Crouzeix (1994) show that ψ_a is "self-polar" (cf. Remark 32) in the sense that

$$\forall a \in \mathbb{R}^n_{\geq 0},\ \forall p \in \mathbb{R}^n_{>0},\ \psi^{\diamond}_a(p) = \frac{\|a\|_1}{\psi_a(a)} \psi_a(p). \tag{49}$$

The function ψ_a can be seen to be the form of the Cobb-Douglas production function (Cobb and Douglas, 1928), the self-duality of which has been an object of considerable interest in microeconomics (Houthhakker, 1965; Samuelson, 1965; Sato, 1976)[15]. In order to find the associated loss function we need ℓ_{CD_a} such that $\mathrm{spr}(\ell_{\mathrm{CD}_a}) = \mathrm{lev}_{\geq 1}(\psi^{\diamond}_a)$. The self polar property (49) makes this easy since

$$\ell_{\mathrm{CD}_a} \in \partial \psi^{\diamond}_a \iff \ell_{\mathrm{CD}_a} \in \frac{\|a\|_1}{\psi_a(a)} \partial \psi_a.$$

$$\ell^{\diamond}_{\mathrm{Br},2}(q) = \frac{(2q-1)\left(2q^2\sqrt{q(1-q)} + 2q^2(1-q) - 3q\sqrt{q(1-q)} + q(1-q)\right)}{2\left(2q^2 + \sqrt{q(1-q)} - 2q\right)^2}.$$

15. It would be of interest to determine other self-dual losses using the results of (Houthhakker, 1965; Samuelson, 1965; Sato, 1976) and to ascertain the significance (if any) of the self-dual nature of the "boosting loss" (example 3). Observe that for all $a \in \mathbb{R}^n_{\geq 0}$ and all $p \in \mathbb{R}^n_{>0}$, $(\ell_{\mathrm{CD}_a} \circ \ell_{\mathrm{CD}_a} \circ \ell_{\mathrm{CD}_a})(p) = \ell_{\mathrm{CD}_a}(p)$, a fact one can verify directly by using (49).

Writing the quotient of vectors componentwise, $\frac{a}{p} := (a_1/p_1, \ldots, a_n/p_n)$, since ψ_a is differentiable on its domain we have

$$\forall p \in \mathbb{R}^n_{\geq 0},\ \ell_{\mathrm{CD}_a}(p) = \frac{\|a\|_1}{\psi_a(a)}(\nabla\psi_a)(p) = \frac{\|a\|_1}{\psi_a(a)}\frac{a}{p}\frac{\psi_a(p)}{\|a\|_1} = \frac{\psi_a(p)}{\psi_a(a)}\frac{a}{p}. \tag{50}$$

Applying the same procedure to the antipolar loss we have

$$\forall p \in \mathbb{R}^n_{\geq 0},\ \ell^{\diamond}_{\mathrm{CD}_a}(p) = (\nabla\psi_a)(p) = \frac{\psi_a(p)}{\|a\|_1}\frac{a}{p}. \tag{51}$$

Example 3 We illustrate the self-duality of ℓ_{CD_a} with a simple example. Set $n = 2$ and $a_1 = a_2 = 1$ and thus $\psi_a(x) = \sqrt{x_1 x_2}$, $\psi_a(a) = 1$, and $\|a\|_1 = 2$. Restricting to $\Delta([2])$, and writing $p = (p_1, p_2) \in \Delta([2])$, from (50) we have

$$\forall p \in \Delta([2]),\ \ell_{\mathrm{CD}_a}(p) = \frac{\sqrt{p_1 p_2}}{p} = \left(\sqrt{\frac{p_2}{p_1}}, \sqrt{\frac{p_1}{p_2}}\right),$$

which can be recognised as the "boosting loss" (Buja et al., 2005). This loss has as its *weight function* (Reid and Williamson, 2011) $w(p_1) := -\frac{\partial^2}{\partial p_1^2}\rho(p_1, 1-p_1)$, where $\rho = \psi_a$ is the concave support function of $\operatorname{spr}\ell_{\mathrm{CD}_a}$. We have

$$w(p_1) = \frac{1}{(p_1(1-p_1))^{3/2}}.$$

Using (51) to calculate the antipolar loss $\ell^{\diamond}_{\mathrm{CD}_a}$ we have

$$\ell^{\diamond}_{\mathrm{CD}_a}(p) = \frac{\sqrt{p_1 p_2}}{2}\frac{1}{p} = \frac{1}{2}\ell_{\mathrm{CD}_a}(p).$$

The superprediction sets associated with the loss ℓ_{CD_a} and its antipolar are illustrated in Figure 11.

5. Designing Losses via their Superprediction Sets

Loss functions are clearly essential for machine learning, but they are often taken for granted, their choice being primarily a consequence of convenience or familiarity. We posit that this is due in part to a lack of a tools for designing and tuning them. In this section we offer some starting points for such a tuning exercise.

The conventional approach to working with loss functions is to focus on the analytic form of the mappings $(p, y) \mapsto \ell(p, y)$. In this section we show some examples of the power of instead working with the family $\mathcal{P}(\mathbb{R}^n_{\geq 0})$ and deriving the associated loss functions via the subdifferential of the concave support functions.

5.1 Canonical Normalisation

One problem that presents itself when working with the family $\mathcal{P}(\mathbb{R}^n_{\geq 0})$ is that of normalisation. In §4 the lack of consistency of normalisation between (ℓ_a), ℓ_{Br} and ℓ_{CD_a} made it difficult to compare these losses side by side. Our proposed normalisation for a loss ℓ is to pick $p_\ell \in \mathbb{R}^n_{>0}$ and $\alpha > 0$

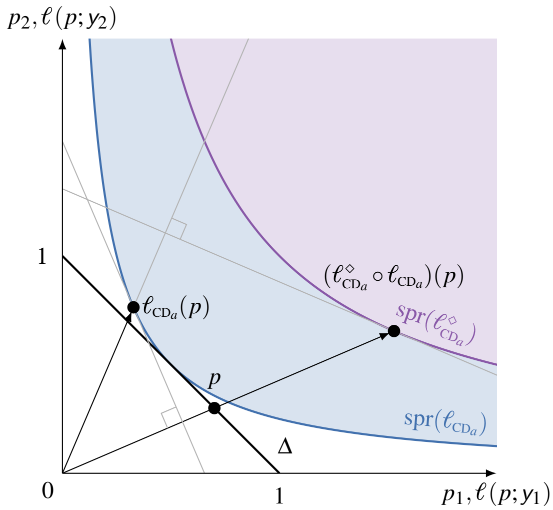


FIGURE 11: Illustration of the self-dual nature of the Cobb-Douglas loss ℓ_{CD_a} with $a := (1/2, 1/2)$ and $p := (7/10, 3/10)$.

such that $\rho_{\mathrm{spr}(\alpha\ell)}(p_\ell) = 1$, a task that the superprediction set machinery makes very simple. There are a couple of ways one might choose p_ℓ, the simplest is $p_\ell := 1_n$ for all loss functions ℓ. However, a more robust choice is $p_\ell \in \arg\max_{p\in\Delta} \rho_{\mathrm{spr}(\ell)}(p)$. We say a loss function ℓ is *normalised* if its associated conditional Bayes risk function attains a maximum value of 1. For several of our results we need a non-compact version of the Sion (1958) minimax theorem. The following result attributed to Ha (1981) is originally shown in a much more general setting and so we state it for the space $\mathscr{X}$ below.

Lemma 36 (Ha 1981, Theorem 2) *Let X and Y each be nonempty convex subsets of $\mathscr{X}$. Let $f\colon X \times Y \to \mathbb{R}$ be such that*

1. *For each $x \in X$, $y \mapsto f(x,y)$ is lower semi-continuous and quasi-convex;*
2. *For each $y \in Y$, $x \mapsto f(x,y)$ is upper semi-continuous and quasi-concave.*

If there exists a nonempty convex set $X' \subseteq X$ and a compact set $Y' \subseteq Y$ such that

$$\inf_{y\in Y} \sup_{x\in X} f(x,y) \leq \inf_{y\in Y'} \max_{x\in X'} f(x,y), \tag{52}$$

then

$$\inf_{y\in Y} \sup_{x\in X} f(x,y) = \sup_{x\in X} \inf_{y\in Y} f(x,y).$$

Theorem 37 *Let ℓ be a proper loss and let $p^\star \in \arg\max_{p\in\Delta} \rho_{\mathrm{spr}(\ell)}(p)$. Then*

$$\partial \rho_{\mathrm{spr}(\ell)}(p^\star) \ni \frac{1_n}{\beta_{\mathrm{spr}(\ell)}(1_n)} = \frac{1_n}{\rho_{\mathrm{spr}(\ell^\diamond)}(1_n)}.$$

Proof First we apply Lemma 36 to establish

$$\max_{p\in\Delta} \inf_{z\in\mathrm{spr}(\ell)} \langle z, p\rangle = \inf_{z\in\mathrm{spr}(\ell)} \max_{p\in\Delta} \langle z, p\rangle. \tag{53}$$

For $\alpha \geq {}^{1}\!/_{\beta_{\mathrm{spr}(\ell)}(1_n)}$ we have $\alpha 1_n \in \mathrm{spr}(\ell)$, and

$$\exists \beta \geq 1, \inf_{z \in \mathrm{spr}(\ell)} \max_{p \in \Delta} \langle z, p\rangle \leq \inf_{z \in \{\beta\alpha 1_n\}} \max_{p \in \Delta} \langle z, p\rangle,$$

since the left hand side is finite. Thus we have demonstrated the sufficient condition for (52). Since $\mathrm{spr}(\ell)$ and Δ are convex, we have satisfied the conditions for Lemma 36 and shown (53).

Therefore

$$\max_{p\in\Delta} \rho_{\mathrm{spr}(\ell)}(p) = \max_{p\in\Delta} \inf_{z\in \mathrm{spr}(\ell)} \langle z, p\rangle \overset{(53)}{=} \inf_{z\in\mathrm{spr}(\ell)} \max_{p\in\Delta} \langle z, p\rangle.$$

Observe $\max_{p\in\Delta}\langle z, p\rangle = \max_{y\in\mathscr{Y}} z_y$. A proof by contradiction easily confirms that for $z \in \partial \rho_{\mathrm{spr}(\ell)}(p^\star)$ we have $z_y = z_z$ for all $y, z \in \mathscr{Y}$. Thus the minimising $z \in \partial\rho_{\mathrm{spr}(\ell)}(p^\star)$ is a multiple of 1_n, more precisely

$$z = \inf\{\lambda > 0 \mid \lambda 1_n \in \mathrm{spr}(\ell)\} \cdot 1_n = 1_n \cdot \frac{1}{\beta_{\mathrm{spr}(\ell)}(1_n)},$$

which completes the proof. ■

By Theorem 37 we see that the maximum of $\rho_{\mathrm{spr}(\ell)}$ occurs for $p^\star \in \Delta$ such that $\ell(p^\star) = \alpha 1_n$ where $\alpha := {}^{1}\!/_{\beta_{\mathrm{spr}(\ell)}(1_n)}$. If we normalise a proper loss ℓ with the coefficient $c := \beta_{\mathrm{spr}(\ell)}(1_n)$, then evaluating the conditional Bayes risk at $p^\star \in \Delta$ we have

$$\rho_{\mathrm{spr}(c\ell)}(p^\star) = c\rho_{\mathrm{spr}(\ell)}(p^\star) \overset{\mathrm{T37}}{=} c\langle {}^{1_n}\!/_{\beta_{\mathrm{spr}(\ell)}(1_n)}, p^\star\rangle = \frac{c}{\beta_{\mathrm{spr}(\ell)}(1_n)} \langle 1_n, p^\star\rangle = 1.$$

The following corollary demonstrates another application of the polar loss, that is the uniform loss vector 1_n is achieved by the prediction that maximises the conditional Bayes risk.

Corollary 38 *Let ℓ be a proper loss and $p^\star := {}^{\ell^\diamond(1_n)}\!/_{\|\ell^\diamond(1_n)\|_1}$. Then $p^\star \in \arg\max_{p\in\Delta} \rho_{\mathrm{spr}(\ell)}(p)$.*

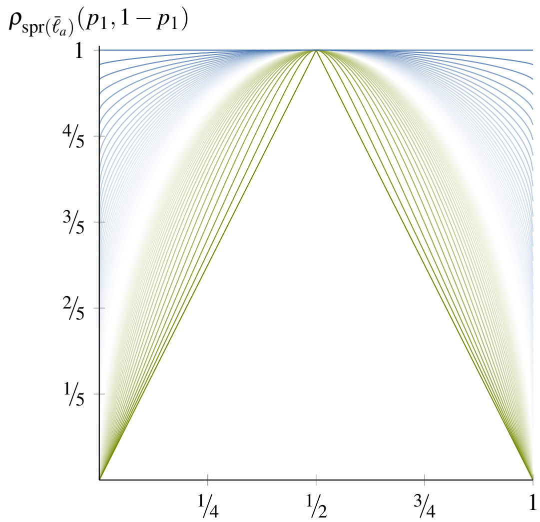
Corollary 39 *A proper loss function ℓ is normalised if and only if $1_n \in \mathrm{bd}(\mathrm{spr}\,\ell)$.*

We give the normalisation coefficients for the common losses from Table 3 in Table 4, and plot their conditional Bayes risk functions and superprediction sets in Figure 12 (the overbar denotes this normalisation). With the normalised versions of these loss functions we can now see that $\bar{\ell}_{\mathrm{CD}_{1_n}}$ is attained as the limit $\lim_{a\to 0} \bar{\ell}_a$.

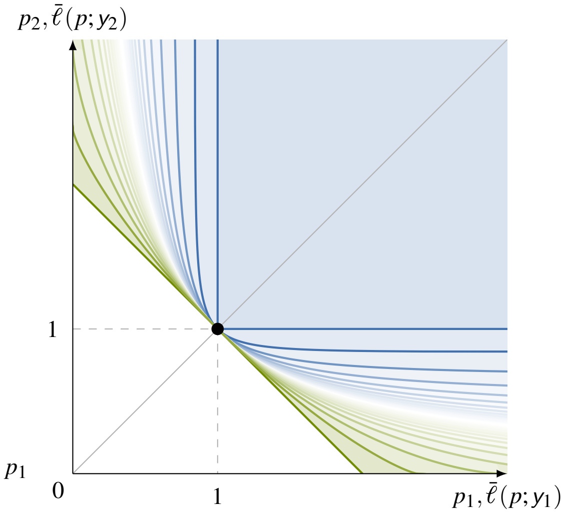
5.2 Shifting the Maximum

In §5.1 we saw that the conditional Bayes risk of a proper loss ℓ is maximised over the probability simplex at $p^\star := {}^{\ell^\diamond(1_n)}\!/_{\|\ell^\diamond(1_n)\|_1}$ (Corollary 38). The question naturally arises then of how one might one modify $\ell \mapsto \check{\ell}$ to reposition the maximum to an arbitrary $p^0 \in \Delta$.[16] That is, $p^\star \mapsto \check{\ell}^\diamond(1_n) = p^0$.

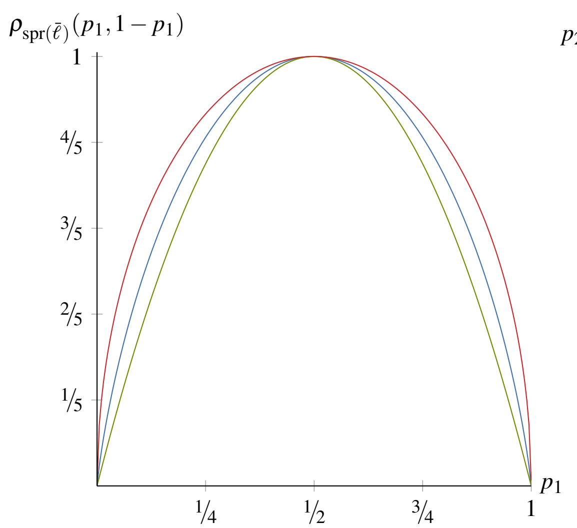
16. The motivation for doing so arises from considering the cost-sensitive missclassification losses ℓ_c, $c \in (0,1)$ (Reid and Williamson, 2011, Section 5.2), whose conditional Bayes risks are $\underline{L}_c(p) = (1-p)c \wedge (1-c)p$. The maximum of $\underline{L}_c(p)$ over p occurs at c (although the maximum value does not remain constant as c varies). The corresponding losses ℓ_c allow one to impose a different cost for false positives and false negatives. Thus shifting the maximum of a general loss allows one to reweight the costs for the different types of prediction error one might make.



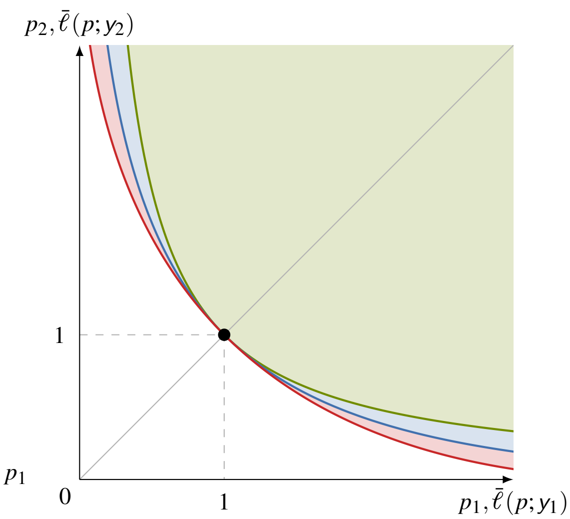
(a) Conditional Bayes risk for ℓ_{CD_a}: graph of $\rho_{\mathrm{spr}(\bar{\ell}_a)}$ for $a = (\alpha, \alpha)$, with $\alpha > 0$ and $\alpha \in I := \{0.9525^i \mid i \in [40]\}$ (blue); and $\alpha < 0$ with $\alpha = \beta/(\beta - 1)$, $\beta \in I$ (green).



(b) Superprediction sets for ℓ_{CD_a} for same parameter range as in (a).



(c) Comparison of conditional Bayes risks for $\bar{\ell}_{\mathrm{CD}_a}$ with $a = (1/2, 1/2)$ (green), $\bar{\ell}_{\mathrm{Br}}$ (red), and $\bar{\ell}_{\log}$ (blue).



(d) Comparison of superprediction sets for $\bar{\ell}_{\mathrm{CD}_a}$ with $a = (1/2, 1/2)$ (green), $\bar{\ell}_{\mathrm{Br}}$ (red), and $\bar{\ell}_{\log}$ (blue).

FIGURE 12: Normalised loss functions.

Name	Symbol	Maximiser	Coefficient	Normalised Loss
Misclassification Error (0/1)	$\bar{\ell}_{0/1}$	$(1/n,\ldots,1/n)$	n	$n\ell_{0/1}(p,y)$
Logarithmic	$\bar{\ell}_{\log}$	$(1/n,\ldots,1/n)$	$1/\log(n)$	$-\frac{1}{\log(n)}\log(p_y)$
Concave Norm $a\in[-\infty,1]\setminus\{0\}$	$\bar{\ell}_a$	$(1/n,\ldots,1/n)$	$\sqrt[a]{n}$	$\sqrt[a]{n}\cdot\left(\frac{p_y}{\beta_{\frac{a}{a-1}}(p)}\right)^{\frac{1}{a-1}}$
Brier	$\bar{\ell}_{\mathrm{Br}}$	$(1/n,\ldots,1/n)$	$\frac{n}{n-1}$	$\frac{n}{n-1}\cdot\left((1+\|p\|_2^2)\cdot 1_n-2p_y\right)$
Cobb–Douglas $a\in\mathbb{R}^n_{\geq 0}$	$\bar{\ell}_{\mathrm{CD}_a}$	$(a_1,\ldots,a_n)$	$\frac{\|a\|_1}{\psi_a(a)}$	$\|a\|_1\psi_a(p/a^2)\cdot\frac{1}{p_y}$

TABLE 4: Canonical normalisations of the loss functions from Table 3.

If one is given only ℓ or $\underline{L}$ this is not obvious. However, the answer is simple in terms of $S := \operatorname{spr}(\ell)$. Before proceeding we note that this problem is not well posed since we have not defined what exactly we are hoping to retain of the original function ℓ. That said, our construction entails—more or less—the minimum required perturbation of S in order to endow $\check{\ell}$ with the necessary desiderata.

To solve this question we construct a new super prediction set $\check{S}$ from S and define $\check{\ell}\in\partial\rho_{\check{S}}$. The family $\mathcal{P}(\mathbb{R}^n_{\geq 0})$ is a cone since it is closed under positive scalar multiplication and addition of sets from the family. Thus if we construct a mapping $S\mapsto\check{S} := \alpha S + x^*$, where $\alpha>0$ and $x^*\in\mathbb{R}^n_{\geq 0}$ we can easily ensure $\check{S}\in\mathcal{P}(\mathbb{R}^n_{\geq 0})$.

In order to move the maximiser from $p^\star$ to p^0 it suffices to translate the set S by $\ell(p^\star)-\ell(p^0)$. However $-\ell(p^0)\notin\mathbb{R}^n_{\geq 0}$, and so we "push" the vector $-\ell(p^0)$ into $\mathbb{R}^n_{\geq 0}$ by adding just enough of the constant loss: $\alpha\cdot 1_n$, where $\alpha := \max_{y\in\mathscr{Y}}\ell(p^0,y)$. This has a neutral effect on the arg max. We now have

$$\check{S} := \frac{1}{\beta(\ell(p^0))}\left(S+\ell(p^\star)-\ell(p^0)+\alpha\cdot 1_n\right), \tag{54}$$

where the term $1/\beta_S(\ell(p^0))$ normalises $\check{S}$ so that $\max_{p\in\Delta}\beta_S(p)=\max_{q\in\Delta}\beta_{\check{S}}(q)$. The normalisation can easily be calculated using the identity $\ell(p^0)=\beta_S(\ell(p^0))\cdot\ell(p^\star)$.

Using the calculus of support functions,

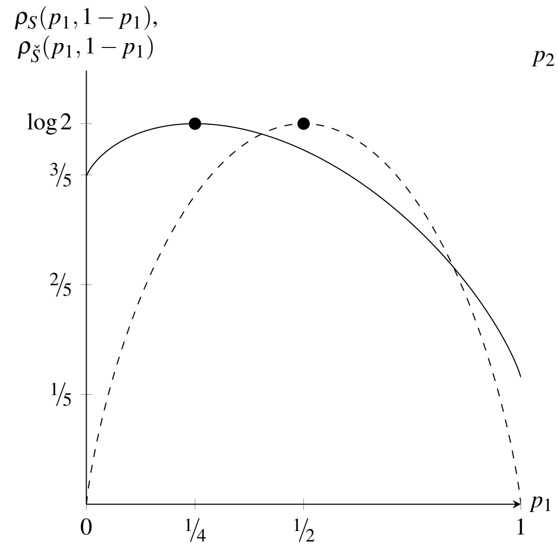
$$\rho_{\check{S}} = \frac{1}{\beta_S(\ell(p^0))}\left(\rho_S+\left\langle\ell(p^\star)-\ell(p^0),\cdot\right\rangle+\alpha\|\cdot\|_1\right),$$

and

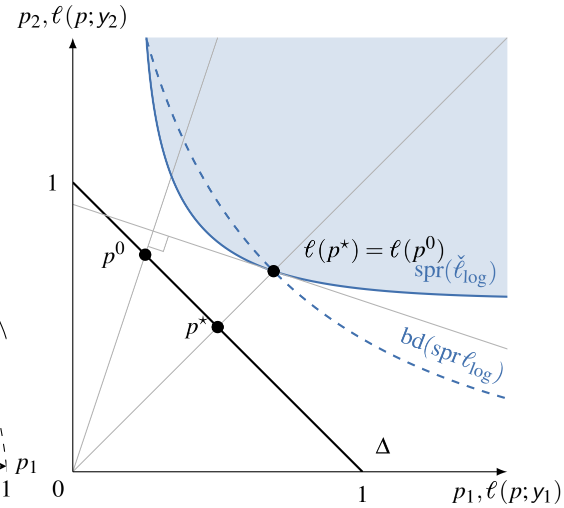
$$\check{\ell}(q) = \frac{1}{\beta_S(\ell(p^0))}\left(\ell(q)+\ell(p^\star)-\ell(p^0)+\alpha\cdot 1_n\right). \tag{55}$$

Example 4 We will now apply the operations (54) and (55) to $\ell_{\log}$ with $\mathscr{Y} := [2]$ and $p^0 := (1/4, 3/4)$. We know $\rho_{\operatorname{spr}(\ell_{\log})}$ achieves its maximum at the uniform prediction: $\ell^\diamond(1_n)=(1/2,1/2)$. To shift the maximum to p^0 we define the new loss $\forall p\in\Delta$,

$$\check{\ell}_{\log}(p) = \frac{1}{\beta_{\operatorname{spr}(\ell_{\log})}(p^0)}\left(\ell_{\log}(p)+\ell_{\log}(p^\star)-\ell_{\log}(p^0)-(\max_{y'\in\mathscr{Y}}\ell_{\log}(p^0;y'))\cdot 1_n\right)$$



(a) The effect of $\ell_{\log} \mapsto \check{\ell}_{\log}$ on the conditional Bayes risk function. The original conditional Bayes risk is dashed.



(b) The corresponding superprediction set operation.

FIGURE 13: Illustration of Example 4, shifting the maximum of the conditional Bayes risk function associated with $\ell_{\log}$. The modified conditional Bayes risk $\rho_{\mathrm{spr}(\check{\ell}_{\log})}$ attains its maximum at $(1/4, 3/4)$.

$$= 1.7783\left(\ell_{\log}(p) + \ell_{\log}(1_n/2) - \ell_{\log}(p^0) + \log(4) \cdot 1_n\right).$$

The effect of $\rho_{\mathrm{spr}(\ell_{\log})} \mapsto \rho_{\mathrm{spr}(\check{\ell}_{\log})}$ is illustrated in Figure 13a. The corresponding superprediction set operation is illustrated in Figure 13b.

5.3 Building Losses From Norms

In §2.7 we looked at norms as gauge functions, and saw that the antigauge functions naturally give rise to the notion of an antinorm. In §3 we saw that these antinorms are precisely the conditional Bayes risk functions. In this section we will see that there is a natural injection—or family thereof—between the symmetric radiant sets and the shady sets. In doing so we define a new family of bounded, proper loss functions: the *norm losses*.

Recall (§2.4) $X_+ \subset \mathcal{X}$ denotes a salient, closed, convex cone, and X_+^* denotes its dual cone. The following result provides a means to take a symmetric radiant set to generate a superprediction set for a proper loss.

Theorem 40 *Let $R \in \mathcal{R}(\mathcal{X})$ be symmetric. Choose $x \in X_+$ with $x \in \bigcap_{r \in R \cap X_+}(X_+ + r)$ and $x \notin R$. Then*

1. *$R + x + X_+ \subseteq X_+$, and*
2. *$R + x + X_+ \in \mathcal{P}(X_+)$.*

Corollary 41 *Let $x \in \bigcap_{r \in R}(X_+ + r)$ then $R + \alpha x + X_+ \in \mathcal{P}(X_+)$, where $\alpha > 1$.*

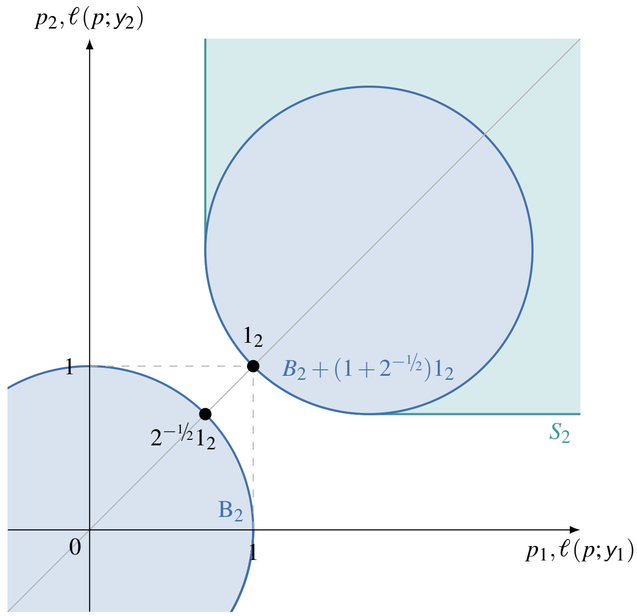


FIGURE 14: Illustration of Theorem 40 using B_2. We translate B_2 "northeast" using the vector $x^* := (1+2^{-1/2}) \cdot 1_2$. This ensures $S_2 := B_2 + (1+2^{-1/2}) \cdot 1_2 + \mathbb{R}^2_{\geq 0} \subseteq \mathbb{R}^2_{\geq 0}$.

Proof From the definition of the dual cone, this means

$$R + x \subseteq X_+ \iff \forall x^* \in X_+^*, \, \forall r \in R, \, \langle r+x, x^* \rangle \geq 0. \tag{56}$$

Minimising the inner product in (56) we have

$$\begin{aligned} \forall x \in X_+, \, \forall r \in R, \, \langle r+x, x^* \rangle &\geq \inf_{z \in X_+^* \setminus \{0\}} \min_{r \in R} \langle r+x, z \rangle \\ &= \inf_{z \in X_+^* \setminus \{0\}} \left(-\langle z/\gamma_R(z), z \rangle + \langle x, z \rangle \right) \\ &= \inf_{z \in X_+^* \setminus \{0\}} \left\langle x - z/\gamma_R(z), z \right\rangle. \end{aligned} \tag{57}$$

We exclude 0 from X_+^* since $z = 0$ satisfies (56) trivially. And $\min_{r \in R} \langle r, z \rangle = -\left\langle z/\gamma_R(z), z \right\rangle$ follows because since R is symmetric and convex, the maximum occurs at $\left\langle z/\gamma_R(z), z \right\rangle$.

Let us now choose $x \in X_+$ as described in the theorem. Then

$$\begin{aligned} x \in \bigcap_{r \in R \cap X_+^*} (X_+ + r) &\iff \forall r \in R \cap X_+^*, \quad x - r \in X_+ \\ &\implies \forall r \in \mathrm{bd}(R) \cap X_+^*, \quad x - r \in X_+ \\ &\iff \forall z \in X_+^* \setminus \{0\}, \quad x - \frac{z}{\gamma_R(z)} \in X_+ \end{aligned}$$

which gives

$$\inf_{z \in X_+^* \setminus \{0\}} \left\langle x - z/\gamma_R(z), z \right\rangle \geq 0. \tag{58}$$

Therefore

$$\forall x \in X_+^*,\ \forall r \in R,\ \langle r+x, x^* \rangle \overset{(57)}{\geq} \inf_{z \in X_+^* \setminus \{0\}} \left\langle x - {}^{z}\!/_{\gamma_R(z)}, z \right\rangle \overset{(58)}{\geq} 0.$$

Thus by (56), $R + x \subseteq X_+$. A closed convex cone is its own recession cone (Proposition 1(4)), which proves claim 1.

The only condition for $R + x + X_+ \in \mathcal{P}(X_+)$ that is non-trivial to show is $R + x + X_+ \ni 0$. As before assume $x \in X_+$ is chosen according to the conditions of the theorem. Then $R \ni x$. Since $x \in X_+$, let $x_0 := x \cdot {}^{1}\!/_{\gamma_R(x)}$ and $x_1 := x \cdot (1 - {}^{1}\!/_{\gamma_R(x)})$ and $x_0, x_1 \in X_+$. Then $x = x_0 + x_1$ and $x_0 \in \mathrm{bd}(R)$. Since R is radiant $x_0 \neq 0$.

$$x_0 \in \mathrm{bd}(R) \iff 0 \in \mathrm{bd}(R) - x_0.$$

By the symmetry of R, this is eqivalent to

$$\begin{aligned} 0 \in \mathrm{bd}(R) + x_0 &\iff x_1 \in \mathrm{bd}(R) + x_0 + x_1 \\ &\iff\ 0 \notin \mathrm{bd}(R) + x, \end{aligned}$$

which implies $R + x + X_+ \ni 0$, and claim 2 is proved. ■

In light of Theorem 40 it might be surprising to note that there is no obvious operation $\mathcal{S}(\mathbb{R}^n) \to \mathcal{R}(\mathbb{R}^n)$. The long flat portions of the set $S_2 \in \mathcal{S}(\mathbb{R}^2)$ in Figure 14 make it easy to see how we might reconstruct B_2 by translating S_2 and forcing the resulting set to be symmetric. But there is no such simple answer for the superprediction sets of unbounded losses. See, for example, $\mathrm{spr}(\ell_{\log})$ (Figure 7) and $\mathrm{spr}(\ell_{\mathrm{CD}_\alpha})$ (Figure 11).

5.4 The Norm Losses

Let $(\mathrm{B}_\alpha)_{\alpha \in [1,\infty]}$ be the family of closed unit α-norm balls in $\mathbb{R}^n$ with $\mathrm{B}_\alpha := \{x \in \mathbb{R}^n \mid \|x\|_\alpha \leq 1\}$. The family (B_α) is increasing in the sense that

$$\alpha \leq \gamma \iff \mathrm{B}_\alpha \subseteq \mathrm{B}_\gamma.$$

The point 1_n satisfies $1_n \in \bigcap_{r \in \mathrm{B}_\alpha} (X_+^* + r)$ for all $\alpha \in [1,\infty]$. For each B_α take the point $\big(1 + {}^{1}\!/_{\gamma_{\mathrm{B}_\alpha}(1_n)}\big) 1_n = (1 + n^{-1/\alpha}) 1_n$ and build the set $S_\alpha := \mathrm{B}_\alpha + (1 + n^{-1/\alpha}) 1_n + \mathbb{R}^n_{\geq 0}$. By Corollary 41 we have $S_\alpha \in \mathcal{P}(\mathbb{R}^n_{\geq 0})$, guaranteeing properness of the associated loss functions: $\ell_{\|\cdot\|_\alpha} \in \partial \rho_{S_\alpha}$. The set B_2 along with S_2 is shown in Figure 14.

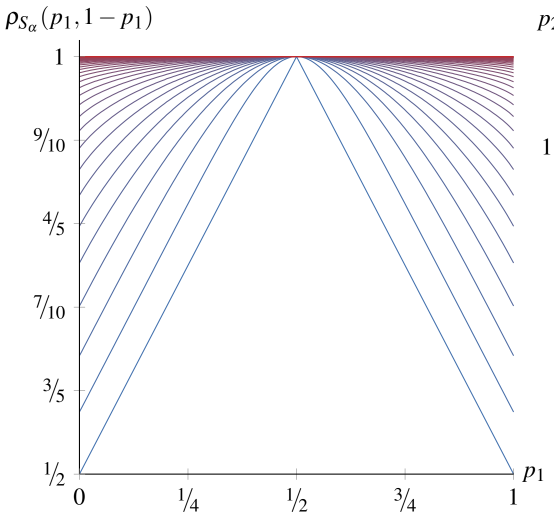
Our choice of construction of S_α has another convenient property:

$$\forall \alpha \in [1,\infty],\ \rho_{S_\alpha}(\ell^\diamond_{\|\cdot\|_\alpha}(1_n)) = 1. \tag{59}$$

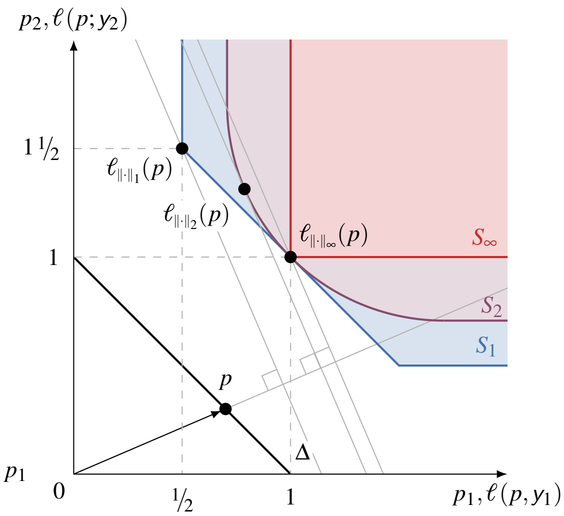
That is, the family $(\ell_{\|\cdot\|_\alpha})_{\alpha \in [1,\infty]}$ has the normalisation about the conditional Bayes risk from §5.1.

We can derive the closed form expression for the whole family on $\mathbb{R}^n_{\geq 0}$ as follows:

$$\rho_{S_\alpha} = \rho_{\mathrm{B}_\alpha} + (1 + n^{-1/\alpha}) \langle 1_n, \cdot \rangle + \rho_{\mathbb{R}^n_{\geq 0}} = \rho_{\mathrm{B}_\alpha} + (1 + n^{-1/\alpha}) \| \cdot \|_1 + \rho_{\mathbb{R}^n_{\geq 0}}.$$



(a) The family $(\rho_{S_\alpha})_{\alpha\in[1,\infty]}$, coloured from blue to red.



(b) The loss functions $\ell_{\|\cdot\|_1}$, $\ell_{\|\cdot\|_2}$ and $\ell_{\|\cdot\|_\infty}$ acting on the vector $p = (7/10, 3/10) \in \Delta$.

FIGURE 15: The family $\left(\ell_{\|\cdot\|_\alpha}\right)_{\alpha\in[1,\infty]}$ smoothly varies between $\ell_{0/1} + 1/2 \cdot 1_n$ and the constant loss function.

Using (7) and the fact that B_α is symmetric we have

$$\rho_{\mathrm{B}_\alpha} = -\sigma_{-\mathrm{B}_\alpha} = -\sigma_{\mathrm{B}_\alpha} = -\beta_{\mathrm{B}_\gamma} = -\|\cdot\|_\gamma,$$

where γ is the Hölder conjugate of α, that is $1/\alpha + 1/\gamma = 1$. Therefore

$$\rho_{S_\alpha} = (1 + n^{-1/\alpha})\|\cdot\|_1 - \|\cdot\|_{\frac{\alpha}{\alpha-1}} + \rho_{\mathbb{R}^n_{\geq 0}}.$$

The family $(\rho_{S_\alpha})_{\alpha\in[1,\infty]}$ is plotted in Figure 15a. We can compute the subdifferential of ρ_{S_α} directly:

$$\forall p \in X_+ \setminus \{0\},\ \partial\rho_{S_\alpha}(p) = (1 + n^{-1/\alpha})\,\partial\|\cdot\|_1(p) - \frac{p}{\|p\|_{\frac{\alpha}{\alpha-1}}},$$

giving us a closed form expression for $\ell_{\|\cdot\|_\alpha}$:

$$\forall p \in X_+ \setminus \{0\},\ \forall y \in \mathscr{Y},\ \ell_{\|\cdot\|_\alpha}(p, y) := 1 + n^{-1/\alpha} - \frac{p_y}{\|p\|_{\frac{\alpha}{\alpha-1}}}.$$

Some special values of α include:

$$\forall p \in X_+ \setminus \{0\},\ \forall y \in \mathscr{Y},\ \ell_{\|\cdot\|_1}(p, y) = \ell_{0/1}(p, y) + \frac{1}{2} \text{ and } \ell_{\|\cdot\|_\infty}(p, y) = \ell_{-\infty}(p; y) = 1,$$

where $\ell_{0/1}$ is misclassification loss (38) and $\ell_{-\infty}$ is the *constant loss* (which we derived in a completely different way in §4.1). The various intermediaries like $\ell_{\|\cdot\|_2}$ smoothly interpolate between these two extremes as illustrated in Figure 15b, where it is clear that the condition (59) is equivalent to the simpler geometric property $1_n \in \bigcap_{\alpha\in[1,\infty]} \mathrm{bd}(S_\alpha)$. Finally we note the family $\left(\ell_{\|\cdot\|_\alpha}\right)_{\alpha\in[1,\infty]}$ is clearly bounded.

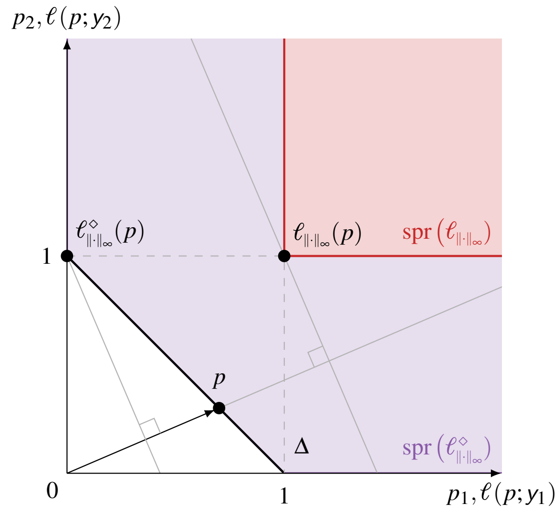


FIGURE 16: Illustration of $\operatorname{co}\mathrm{spr}(\ell_{0/1}) = \mathrm{spr}(\ell^{\diamond}_{\|\cdot\|_\infty})$ and that $\langle \ell^{\diamond}_{\|\cdot\|_1}(p), p\rangle = \min_{y\in\mathscr{Y}} p_y$.

Example 5 We can now derive the antipolar result $\ell^{\diamond}_{0/1} = \ell_{-\infty} = \ell_{\|\cdot\|_\infty}$ from §4.1 using a simpler geometrical argument: Consider the set $\mathrm{spr}(\ell_{-\infty})$ which has the property that $\rho_{\mathrm{spr}(\ell_{-\infty})} = \|\cdot\|_1$ over $\mathbb{R}^n_{\geq 0}$. And so we have

$$\mathrm{spr}(\ell_{-\infty})^{\diamond} = \mathrm{lev}_{\geq 1}\,\rho_{\mathrm{spr}(\ell_{-\infty})} = \mathrm{lev}_{\geq 1}(\|\cdot\|_1 + \rho_{\mathbb{R}^n_{\geq 0}}) = (\mathrm{lev}_{\geq 1}\|\cdot\|_1)\cap \mathbb{R}^n_{\geq 0}.$$

These sets are illustrated in Figure 16.

6. Combining Given Proper Losses to Form New Ones

Much machine learning practice works with a small family of loss functions for the pragmatic reason that they are familiar, available, and have explicit formulas. The above development shows there is an enormous range of possible proper losses one could use, but offers no concrete way of constructing them (with explicit formulas); if one had an analytic description of a desired superprediction set, then it is clear how to cosntruct the loss function. But such a premise seems implausible. In this section, we develop a straightforward way of constructing a larger usable set of loss functions by finding ways of combining existing proper losses in a manner that guarantees the result is also a proper loss, and which provides explicit formulas for the resulting loss function and associated conditional Bayes risk (concave support function).

In §5 we observed the power of defining loss functions ℓ by directly building their superprediction sets $\mathrm{spr}(\ell)$. We also saw that the family $\mathcal{P}(X_+)$ is a cone, and is therefore closed under the family of operations

$$\forall T \subseteq X_+,\ \forall \alpha > 0,\ \mathcal{P}(X_+) \ni S \mapsto \alpha S + T \in \mathcal{P}(X_+). \tag{60}$$

It's natural then to consider what other operations have a closure property analogous to (60) for the family $\mathcal{P}(X_+)$. With the relationship between proper losses and $S \in \mathcal{P}(X_+)$, this amounts to asking whether one can combine multiple proper losses non-additively and still be guaranteed that the result is a proper loss; when working directly with ℓ, it is not obvious how to ensure the resulting loss is proper[17]. The convex analysis literature has largely studied the closely related family $\mathcal{R}(\mathcal{X})$ (Seeger, 1990; Seeger and Volle, 1995; Gardner et al., 2013; Gardner and Kiderlen, 2018), with some results for the family $\mathcal{S}(\mathcal{X})$ (Barbara and Crouzeix, 1994; Penot, 1997; Penot and Zălinescu, 2000).

We will present a general family of operations, called M-sums and dual M-sums, (Gardner et al., 2013; Mesikepp, 2016) which provide a general means by which to create new proper losses from two or more given proper losses. These M-sums provide the opportunity to smoothly interpolate between several proper losses in a variety of ways (beyond merely taking the sum)[18].

In §6.2 we introduce our approach, and then successively demonstrate the preservation of the convexity of the superprediction sets (§6.3), their closure (§6.4), and orientation (§6.5). We then introduce the functional analog of our combination rules (which serve to combine conditional Bayes risks) (§6.6), examine their properties in terms of support functions (§6.7), and the effect of polar operations (§6.8). The argument is summarised and tied together in §6.9.

The *epimultiplication* operation is

$$\mathbb{R} \times 2^{\mathcal{X}} \ni (\alpha, S) \mapsto \alpha \star S := \begin{cases} \alpha S & \alpha \neq 0, \\ \operatorname{rec}(S) & \alpha = 0. \end{cases}$$

Fix $M \subseteq \mathbb{R}^m$. Then the *M-sum* and *dual M-sum* operations are defined as

$$\left(2^{\mathcal{X}}\right)^m \ni (A_1, \ldots, A_m) \mapsto \oplus_M(A_1, \ldots, A_m) := \bigcup_{\mu \in M} \sum_{i \in [m]} \mu_i \star A_i,$$

and

$$\left(2^{\mathcal{X}}\right)^m \ni (A_1, \ldots, A_m) \mapsto \oplus_M^*(A_1, \ldots, A_m) := \bigcup_{\mu \in M} \bigcap_{i \in [m]} \mu_i \star A_i.$$

6.1 M-Composition of Losses

Throughout this section let $\ell_1, \ldots, \ell_m$ be a sequence of proper loss functions, paired with their conditional Bayes risk functions $\underline{L}_1, \ldots, \underline{L}_m \colon X_+ \to \overline{\mathbb{R}}$. Let $m \colon \mathbb{R}^m_{>0} \to \mathbb{R}^m_{\geq 0}$ be a proper loss function with the associated conditional Bayes risk $\underline{M} \colon \mathbb{R}^m_{\geq 0} \to \overline{\mathbb{R}}$. We introduce the two functions

$$\oplus_{\underline{M}}(\underline{L}_1, \ldots, \underline{L}_m) := p \mapsto \underline{M}(\underline{L}_1(p), \ldots, \underline{L}_m(p))$$

and

$$\oplus_{\underline{M}}^*(\underline{L}_1, \ldots, \underline{L}_m) := p \mapsto \sup\{\underline{M}(\underline{L}_1(a_1), \ldots, \underline{L}_m(a_m)) \mid a_1 + \cdots + a_m = p\},$$

which we call the *functional M-sum* and *dual functional M-sum* respectively.

17. This question is obviously analogous to the question of "aggregation" in economics; see for example (Shephard, 1970, Chapter 6). In our case, restricting consideration to proper losses makes the problem situation simpler, and a rather more comprehensive answer can be given.

18. In applying the existing theory of M-sums we have needed to extend it in two ways: we have developed the concave version (which combines shady sets rather than radiant ones), and we have developed a comprehensive duality theory. These results may be of interest in their own right. They extend and generalise a range of results in the literature on the combination of convex bodies, including (Artstein-Avidan and Rubinstein, 2017; Penot and Zălinescu, 2000; Seeger and Volle, 1995; Volle, 1998; Luc and Volle, 1997; Penot and Zălinescu, 2001; Barbara and Crouzeix, 1994; Pallaschke and Urbanski, 2013; Milman and Rotem, 2017a; Slomka, 2011; Milman and Rotem, 2017b).

The functional M-sum encompasses a wide range of operations on losses. For example using Corollary 56 we can see the sum of losses $\jmath$ and $\hbar$ can be written as an M-sum using the constant loss $\ell_{-\infty}$ with $\underline{N} := \rho_{\mathrm{spr}(\ell_{-\infty})}$:

$$\partial \oplus_{\underline{N}}(\rho_{\mathrm{spr}(\jmath)}, \rho_{\mathrm{spr}(\hbar)}) \ni \jmath + \hbar.$$

Note that $\jmath$ defined in this way is not guaranteed to retain the pseudo-inverse property of the antipolar $\ell^{\diamond}$ in the sense of Proposition 29.

6.2 Constructing Superprediction Sets with M-Sums

Our approach here is organised as follows: First we show some general sufficient conditions for the operations $\oplus_M$ and $\oplus_M^*$ (introduced in §6) to map $\mathcal{P}(X_+)$ to $\mathcal{P}(X_+)$. Since we are ultimately interested in the support functions of these sets, in §6.6 we compute the (convex and concave) support functions of sets in the images of $\oplus_M$ and $\oplus_M^*$. The choices of name and notation for the dual M-sum are not accidental, in §6.8 we characterise the duality relationship of $\oplus_M$ and $\oplus_M^*$ in terms of the polar and antipolar and present closure results for the families $\mathcal{R}(\mathcal{X})$ and $\mathcal{S}(X_+)$ (Theorem 61 and Theorem 63), which allows us to compute the gauge and antigauge functions of sets in the images of $\oplus_M$ and $\oplus_M^*$ (Corollary 65).

We first seek to establish closure (in the algebraic sense) of the family $\mathcal{P}(X_+)$ with the operations $\oplus_M$ and $\oplus_M^*$. In order to show this requires a number of theorems, which culminate in the dénouement Corollary 51. For ease of exposition these results are summarised in Table 5.

It is necessary to introduce the *Panlevé–Kuratowski* notion of convergence for sequences of sets (Rockafellar and Wets, 2004). Define the following classes of subsets:

$$\mathcal{N} := \{N \subseteq \mathbb{N} \mid \mathbb{N} \setminus N \text{ is finite}\} \text{ and } \mathcal{N}^{\#} := 2^{\mathbb{N}}.$$

Let $(S_n)_{n \in \mathbb{N}}$ with $S_n \subseteq \mathcal{X}$ be a sequence of sets. Then the *inner and outer limit* are

$$\liminf_{n\to\infty} S_n := \{x \in \mathcal{X} \mid \forall N \in \mathcal{N},\ \forall n \in N,\ \exists x_n \in S_n,\ x_n \to x\}$$

and

$$\limsup_{n\to\infty} S_n := \left\{x \in \mathcal{X} \;\middle|\; \forall N \in \mathcal{N}^{\#},\ \forall n \in N,\ \exists x_n \in S_n,\ x_n \to x\right\}.$$

If $\liminf_{k\to\infty} C_k = \limsup_{k\to\infty} C_k$ then we say $(C_k)_{k\in\mathbb{N}}$ converges with limit $\lim_{k\to\infty} C_k$. As one might hope, since $\mathcal{N} \subseteq \mathcal{N}^{\#}$ it follows that $\liminf_{n\to\infty} S_n \subseteq \limsup_{n\to\infty} S_n$.

Proposition 42 *Let $S \in \mathcal{K}(\mathcal{X})$, $(\mu_k)_{k\in\mathbb{N}} \to \mu$ with $\mu_k \in \mathbb{R}_{>0}$ for all $k \in \mathbb{N}$. Then $\mu \star S = \lim_{k\to\infty} \mu_k \cdot S$.*

Proof The only interesting case is when $\mu = 0$, which is immediate from Lemma 13. ■

6.3 Convexity

The convexity of superprediction sets plays an essential role in our theory, and thus if we wish to combine multiple proper losses by combining their superprediction sets, we need to ensure the resulting set is guaranteed convex. We first need some auxilliary lemmas.

Lemma 43 *Let $S \in \mathcal{K}(\mathcal{X})$, and $\alpha, \beta \in [0,\infty)$. Then*

1. $\mathrm{rec}(\alpha \star S) = \mathrm{rec}(S)$;
2. $\alpha \star S + \beta \star S = (\alpha+\beta) \star S$.

Proof 1. Let $\alpha = 0$. Then $\alpha \star S = \mathrm{rec}(S)$. Since $\mathrm{rec}(S)$ is a closed cone, it is easily verified (Proposition 14) that $\mathrm{rec}(\mathrm{rec}(S)) = \mathrm{rec}(S)$. For $\alpha > 0$ we have $\mathrm{rec}(\alpha \star S) = \mathrm{rec}(S)$. This is an immediate consequence of Proposition 1(3).

Turning now to claim 2 there are three cases: neither α nor β is zero, only one of α or β is zero, or both α and β are zero. The only interesting case is the second. Let $\alpha \neq 0, \beta = 0$ and we have

$$\alpha \star S + \beta \star S = \alpha \star S + \mathrm{rec}(S) = \alpha \star S = (\alpha+\beta) \star S.$$

The second equality follows from Proposition 1(2) since $\mathrm{rec}(S) = \mathrm{rec}(\alpha \star S)$. ■

Lemma 44 *Let I be an arbitrary index set. For families of subsets of $\mathcal{X}$, $(S_i)_{i\in I}$ and $(T_j)_{j\in I}$ we have $\bigcap_{i\in I} S_i + \bigcap_{j\in I} T_j \subseteq \bigcap_{i\in I}(S_i + T_i)$.*

Proof Let $x \in \bigcap_{i\in I} S_i + \bigcap_{j\in I} T_j$. Then $x = s + r$ for some points s, r where s is in every S_i, and r is in every T_j. Thus $x \in S_i + T_j$ for all $i, j \in I$, including the pairs (S_i, T_j) with $j = i$. Consequently x is in the intersection $\bigcap_{i\in I}(S_i + T_i)$. ■

The main result of this subsection is the following.

Theorem 45 *Let $M \in \mathcal{K}(\mathbb{R}^m)$, and $A_i \in \mathcal{K}(\mathcal{X})$ for $i \in [m]$. Then the sets $\oplus_M(A_1,\ldots,A_m)$ and $\oplus^*_M(A_1,\ldots,A_m)$ are convex.*

Proof Fix arbitrary $x, y \in \oplus_M(A_1,\ldots,A_m)$. Then there are $\mu, \nu \in M$, such that

$$x \in \sum_{i\in[m]} \mu_i \star A_i \text{ and } y \in \sum_{j=1}^{m} \nu_j \star A_j. \tag{61}$$

To show $\oplus_M(A_1,\ldots,A_m)$ is a convex set, we need to show $tx + (1-t)y \in \oplus_M(A_1,\ldots,A_m)$ for all $t \in (0,1)$. By virtue of (61), $\forall t \in (0,1)$,

$$\begin{aligned} tx + (1-t)y &\in t \sum_{i\in[m]} \mu_i \star A_i + (1-t) \sum_{j=1}^{m} \nu_j \star A_j \\ &= \sum_{i\in[m]} \left(t\mu_i \star A_i + (1-t)\nu_i \star A_i\right). \end{aligned} \tag{62}$$

Applying Lemma 43 with $S = A_i$, $\alpha = t\mu_i$ and $\beta = (1-t)\nu_i$ implies

$$\forall i \in [m],\ t\mu_i \star A_i + (1-t)\nu_i \star A_i = (t\mu_i + (1-t)\nu_i) \star A_i, \tag{63}$$

and thus

$$\sum_{i\in[m]} (t\mu_i \star A_i + (1-t)\nu_i \star A_i) = \sum_{i\in[m]} (t\mu_i + (1-t)\nu_i) \star A_i. \tag{64}$$

Finally, convexity of M guarantees $t\mu + (1-t)\nu \in M$, and therefore $\forall t \in (0,1)$,

$$\begin{aligned} tx + (1-t)y &\overset{(62)}{\in} \sum_{i\in[m]} (t\mu_i \star A_i + (1-t)\nu_i \star A_i) \\ &\overset{(64)}{=} \sum_{i\in[m]} (t\mu_i + (1-t)\nu_i) \star A_i \\ &\subseteq \bigcup_{\mu\in M} \sum_{i\in[m]} \mu_i \star A_i, \end{aligned} \tag{65}$$

which concludes the proof that $\oplus_M(A_1,\dots,A_m)$ is convex.

The proof that $\oplus^*_M(A_1,\dots,A_m)$ is convex is similar. Let $x, y \in \oplus^*_M(A_1,\dots,A_m)$. Then there exists $\mu, \nu \in M$ such that $x \in \bigcap_{i\in[m]} \mu_i \star A_i$ and $y \in \bigcap_{j=1}^m \nu_j \star A_j$. Therefore $\forall t \in (0,1)$,

$$\begin{aligned} tx + (1-t)y &\in t\left(\bigcap_{i\in[m]} \mu_i \star A_i\right) + (1-t)\left(\bigcap_{j\in[m]} \nu_j \star A_j\right) \\ &= \left(\bigcap_{i\in[m]} t\mu_i \star A_i\right) + \left(\bigcap_{j\in[m]} (1-t)\nu_j \star A_j\right) \\ &\overset{44}{\subseteq} \bigcap_{i\in[m]} (t\mu_i \star A_i + (1-t)\nu_i \star A_i). \end{aligned} \tag{66}$$

From (63), $\forall t \in (0,1)$,

$$\bigcap_{i\in[m]} (t\mu_i \star A_i + (1-t)\nu_i \star A_i) = \bigcap_{i\in[m]} (t\mu_i + (1-t)\nu_i) \star A_i. \tag{67}$$

Again the convexity of M guarantees that $t\mu + (1-t)\nu \in M$, and mirroring (65), $\forall t \in (0,1)$,

$$\begin{aligned} tx + (1-t)y &\overset{(66)}{\in} \bigcap_{i\in[m]} (t\mu_i \star A_i + (1-t)\nu_i \star A_i) \\ &\overset{(67)}{=} \bigcap_{i\in[m]} (t\mu_i + (1-t)\nu_i) \star A_i \\ &\subseteq \bigcup_{\mu\in M} \bigcap_{i\in[m]} \mu_i \star A_i, \end{aligned}$$

which concludes the proof that $\oplus^*_M(A_1,\dots,A_m)$ is convex. ■

Proposition 46 *Let $M \in \mathcal{K}(\mathbb{R}^m_{\geq 0} \setminus \{0\})$, and $A_i \in \mathcal{K}(X_+ \setminus \{0\})$ for $i \in [m]$. Then*

$$\oplus_M(A_1,\dots,A_m) \subseteq X_+ \setminus \{0\} \;\; \text{and} \;\; \oplus^*_M(A_1,\dots,A_m) \subseteq X_+ \setminus \{0\}.$$

Proof Since $X_+ \setminus \{0\}$ is a cone, it is closed under addition and positive multiplication. The set M does not contain $0 \in \mathbb{R}^m$ and since the A_i are all subsets of $X_+ \setminus \{0\}$, for all $\mu \in M$ the following inclusions are immediate:

$$\mu_1 \star A_1 + \cdots + \mu_m \star A_m \subseteq X_+ \setminus \{0\} \;\text{ and }\; \mu_1 \star A_1 \cap \cdots \cap \mu_m \star A_m \subseteq X_+ \setminus \{0\}.$$

■

6.4 Closure

Superprediction sets are closed by construction, so we also need to ensure that our combination rules preserve closure. First we need the following lemma.

Lemma 47 *Let $A_i \in \mathcal{K}(X_+)$ for $i \in [m]$. Let $(\mu^k)_{k\in\mathbb{N}} \to \mu$ with $\mu^k \in \mathbb{R}^m_{\geq 0}$. Then*

1. $\lim_{k\to\infty}(\mu_1^k A_1 + \cdots + \mu_m^k A_m) = \mu_1 \star A_1 + \cdots + \mu_m \star A_m$,
2. $\lim_{k\to\infty}(\mu_1^k A_1 \cap \cdots \cap \mu_m^k A_m) = \mu_1 \star A_1 \cap \cdots \cap \mu_m \star A_m$.

Proof Define the set $C := \mu_1 A_1 + \cdots + \mu_m A_m$ and the sequence of sets $C_k := \mu_1^k A_1 + \cdots + \mu_m^k A_m$. Likewise the set $D := \mu_1 A_1 \cap \cdots \cap \mu_m A_m$ and the sequence of sets $D_k := \mu_1^k A_1 \cap \cdots \cap \mu_m^k A_m$.

Take an arbitrary convergent sequence $(x_k) \to x$ such that $x_k \in C_k$. Then $x_k = \sum_{i\in[m]} \mu_i^k a_i^k$ with $a_i^k \in A_i$ for $i \in [m]$ and each $k \in \mathbb{N}$. All the sequences $\mu_i^k a_i^k$ have convergent subsequences (Lemma 13) so it is without loss of generality to assume that they are convergent (by passing to subsequence if necessary), and it follows that $\lim_{k_i\to\infty} \mu_i^{k_i} a_i^{k_i}$ exists for each $i \in [m]$ and

$$x = \lim_{k_1\to\infty} \mu_1^{k_1} a_1^{k_1} + \cdots + \lim_{k_m\to\infty} \mu_m^{k_m} a_m^{k_m} \overset{\text{P42}}{\implies} x \in \mu_1 \star A_1 + \cdots + \mu_m \star A_m,$$

since $\mu^k \in \mathbb{R}^n_{\geq 0}$ for all $k \in \mathbb{N}$, and we have proven (1).

Again take an arbitrary sequence $(x_k) \to x$ such that $x_k \in D_k$. Then $x_k = \mu_i^k a_i^k$ for some sequences $\mu_i^k a_i^k \in \mu_j^k A_j$ for all $i, j \in [m]$ and all $n \in \mathbb{N}$. Applying Proposition 42 completes the proof of claim 2. ∎

Our main result for this subsection is:

Theorem 48 *Let $M \in \mathcal{K}(\mathbb{R}^m_{\geq 0} \setminus \{0\})$, $A_i \in \mathcal{K}(X_+ \setminus \{0\})$ for $i \in [m]$. Then $\oplus_M(A_1, \ldots, A_m)$ and $\oplus^*_M(A_1, \ldots, A_m)$ are both closed.*

Proof Take an arbitrary sequence $(x_k) \to x$ such that $x_k \in \oplus_M(A_1, \ldots, A_m)$. Then there exists a sequence $(\mu^k)_{k\in\mathbb{N}}$ with $\mu_k \in M$ so that $x_k \in \sum_{i\in[m]} \mu_i^k \star A_i$ for all $k \in \mathbb{N}$. Assume the sequence $(\mu^k)_{k\in\mathbb{N}}$ is bounded. Then without loss of generality we may assume it is convergent (by passing to a subsequence if necessary) with limit μ. Since M is closed, $\mu \in M$. It follows that

$$\begin{aligned} x &= \lim_{k\to\infty} x_k \in \lim_{k\to\infty} (\mu_1^k \star A_1 + \cdots + \mu_m^k \star A_m) \\ &\overset{\text{L47(1)}}{=} \mu_1 \star A_1 + \cdots + \mu_m \star A_m \subseteq \oplus_M(A_1, \ldots, A_m). \end{aligned}$$

A proof by contradiction shows that the sequence $(\mu^k)_{k\in\mathbb{N}}$ is bounded. Assume $(\mu^k)_{k\in\mathbb{N}}$ is unbounded. Then we can write $\mu^k = \nu^k \cdot \|\mu^k\|$ where $\nu^k = \mu^k/\|\mu^k\|$ for each $k \in \mathbb{N}$. Thus $(\nu^k)_{k\in\mathbb{N}}$ is bounded and we may assume it is convergent with limit ν. Therefore

$$\begin{aligned} x_k \in \mu_1^k \star A_1 + \cdots + \mu_m^k \star A_m &\iff \frac{x_k}{\|\mu_k\|} \in \nu_1^k \star A_1 + \cdots + \nu_m^k \star A_m \\ \implies \lim_{k\to\infty} \frac{x_k}{\|\mu_k\|} \in \lim_{k\to\infty} \nu_1^k \star A_1 + \cdots + \nu_m^k \star A_m &\overset{\text{L47(1)}}{\iff} 0 \in \nu_1 \star A_1 + \cdots + \nu_m \star A_m, \end{aligned}$$

which contradicts Proposition 46 (taking $M = \{\nu\}$). Thus $\oplus_M(A_1, \ldots, A_m)$ is closed.

Again take an arbitrary sequence $(x_k) \to x$ such that $x_k \in \oplus_M^*(A_1, \ldots, A_m)$. Then $x_k = \mu_i^k a_i^k$ for some sequences $\mu_i^k a_i^k \in \mu_j^k A_j$ for all $i, j \in [m]$ and all $n \in \mathbb{N}$. Clearly the sequences $(\mu_i^k)_{k\in\mathbb{N}}$ must be bounded for all $i \in [m]$. And so, as before, without loss of generality we assume it has limit $\mu \in M$. Applying Proposition 42 completes the proof that $\oplus_M^*(A_1, \ldots, A_m)$ is closed. ■

Corollary 49 *Let $M \in \mathcal{K}(\mathbb{R}^m)$, $A_i \in \mathcal{K}(\mathbb{R}^m)$ for $i \in [m]$. Assume M and each A_i for $i \in [m]$ are compact. Then both $\oplus_M(A_1, \ldots, A_m)$ and $\oplus_M^*(A_1, \ldots, A_m)$ are closed.*

Proof The above result follows by an almost identical proof to Theorem 48, however since M and each A_i for $i \in [m]$ are closed and bounded, this rules out the above pathologies when it comes to building bounded sequences $(a_i^k)_{k\in\mathbb{N}}$ with $a_i^k \in A_i$ for all $i \in [m]$. ■

6.5 Orientation

Superprediction sets recess to the positive orthant. Thus we also need to ensure this property is preserved under our combination rules. This is captured by the main result of this subsection:

Theorem 50 *Let $M \subseteq \mathbb{R}^m_{\geq 0}$, $A_i \in \mathcal{P}(X_+)$ for $i \in [m]$. Then*

1. $\operatorname{rec}(\oplus_M(A_1, \ldots, A_m)) = X_+$, *and*
2. $\operatorname{rec}(\oplus_M^*(A_1, \ldots, A_m)) = X_+$.

Proof Let $\mu \in M$, $B_\mu := \sum_{i\in[m]} \mu_i \star A_i$, $C_\mu := \bigcap_{i\in[m]} \mu_i \star A_i$. Then

$$\forall \mu \in M,\ \operatorname{rec}(B_\mu) \overset{\text{P1(8)}}{\supseteq} \sum_{i\in[m]} \operatorname{rec}(\mu_i \star A_i) \overset{\text{L43(1)}}{=} \sum_{i\in[m]} \operatorname{rec}(A_i) \overset{\text{P1(2)}}{=} X_+. \tag{68}$$

For the equivalent result for C_μ by Lemma 20, $\bigcap_{i\in[m]} A_i \neq \varnothing$, Proposition 1(6) implies

$$\forall \mu \in M,\ \operatorname{rec}(C_\mu) \overset{\text{P1(6)}}{=} \bigcap_{i\in[m]} \operatorname{rec}(\mu_i \star A_i) \overset{\text{L43(1)}}{=} \bigcap_{i\in[m]} \operatorname{rec}(A_i) = X_+. \tag{69}$$

It follows that

$$\operatorname{rec}(\oplus_M(A_1, \ldots, A_m)) = \operatorname{rec}(\bigcup_{\mu\in M} B_\mu) \overset{\text{P1(7)}}{\supseteq} \bigcup_{\mu\in M} \operatorname{rec}(B_\mu) \overset{(68)}{\supseteq} X_+, \tag{70}$$

and

$$\operatorname{rec}(\oplus_M^*(A_1, \ldots, A_m)) = \operatorname{rec}(\bigcup_{\mu\in M} C_\mu) \overset{\text{P1(7)}}{\supseteq} \bigcup_{\mu\in M} \operatorname{rec}(C_\mu) \overset{(69)}{=} X_+. \tag{71}$$

The reverse inclusion is shown by contradiction. Suppose the inclusions in (70) and (71) are all strict. Then there exists

$$d \in \operatorname{rec}(\oplus_M(A_1, \ldots, A_m)) \text{ where } d \notin X_+. \tag{72}$$

	$(A_i)_{i\in[m]}$	M	$\oplus_M(A_1,\ldots,A_m)$	$\oplus^*_M(A_1,\ldots,A_m)$
Theorem 45	$A_i \in \mathcal{K}(\mathcal{X})$	$M \in \mathcal{K}(\mathbb{R}^m)$	convex	convex
Theorem 48	$A_i \in \mathcal{K}(X_+ \backslash\{0\})$	$M \in \mathcal{K}(\mathbb{R}^m_{\geq 0}\backslash\{0\})$	closed	closed
Theorem 50	$A_i \in \mathcal{P}(X_+)$	$M \in \mathbb{R}^m_{\geq 0}$	X_+-oriented	X_+-oriented
Proposition 46	$A_i \in \mathcal{P}(X_+)$	$M \in \mathbb{R}^m_{\geq 0}\setminus\{0\}$	subset of $X_+\setminus\{0\}$	subset of $X_+\setminus\{0\}$
Corollary 51	$A_i \in \mathcal{P}(X_+)$	$M \in \mathcal{P}(\mathbb{R}^m_{\geq 0})$	in $\mathcal{P}(X_+)$	in $\mathcal{P}(X_+)$

TABLE 5: Summary of M-sum structure results.

From Proposition 1(3) this means there are sequences $(t_k)_{k\in\mathbb{N}} \searrow 0$, $t_k \in (0,1]$, and $(x_k)_{k\in\mathbb{N}}$, $x_k \in \operatorname{rec}(\oplus_M(A_1,\ldots,A_m))$ such that $\lim_{k\to\infty} t_k x_k = d$. From the definition of $\oplus_M(A_1,\ldots,A_m)$ there must be a sequence $(\mu^k)_{k\in\mathbb{N}} \subseteq M$ such that

$$x_k \in \mu^k_1 \star A_1 + \cdots + \mu^k_m \star A_m.$$

By assumption we have $A_i \subseteq \operatorname{rec}(A_i) = X_+$ for $i \in [m]$, which implies $\mu^k_i \star A_i \subseteq X_+$ for $i \in [m]$. Since X_+ is a cone we have

$$t_k x_k \in \sum_{i\in[m]} t_k \mu^k_i \star A_i \subseteq X_+.$$

By hypothesis X_+ is closed, and therefore contains $\lim_{k\to\infty} t_k x_k$, giving us a contradiction in (72). Thus equality holds throughout in (70). By an identical argument, *mutatis mutandis*, applied to $\oplus^*_M$ we have equality throughout in (71), proving claim 2. ■

The collection of results amassed in this section is summarised in Table 5, collectively they imply:

Corollary 51 *Let $M \subseteq \mathcal{P}(\mathbb{R}^m_{\geq 0})$. Then $\oplus_M$ maps from $\mathcal{P}(X_+)^m$ to $\mathcal{P}(X_+)$, and $\oplus^*_M$ maps from $\mathcal{P}(X_+)^m$ to $\mathcal{P}(X_+)$.*

With Corollary 51 we see that the family of superprediction sets of proper losses $\mathcal{P}(\mathbb{R}^n_{\geq 0})$ is closed under the the $\oplus_M$ and $\oplus^*_M$ operations. Corollary 51 is illustrated in Figure 17.

6.6 The Functional M-Sum

While the superprediction sets are our starting point, to be able to derive proper loss functions we work with the support function of these sets. The combination rules for the sets have an analog for their corresponding support functions. We first introduce the functional M-sum and in the following subsection justify the overloading of the naming and notation.

Let $f_1,\ldots,f_m \colon \mathcal{X} \to \mathbb{R}$ be convex functions. For convex $g \colon \mathbb{R}^m \to \overline{\mathbb{R}}$ define the *convex functional M-sum*

$$\mathcal{X} \ni x \mapsto \oplus_g(f_1,\ldots,f_m)(x) := g(f_1(x),\ldots,f_m(x))$$

and the *dual convex functional M-sum*

$$\mathcal{X} \ni x \mapsto \oplus^*_g(f_1,\ldots,f_m)(x) := \inf\{g(f_1(a_1),\ldots,f_m(a_m)) \mid a_1+\cdots+a_m = x\}.$$

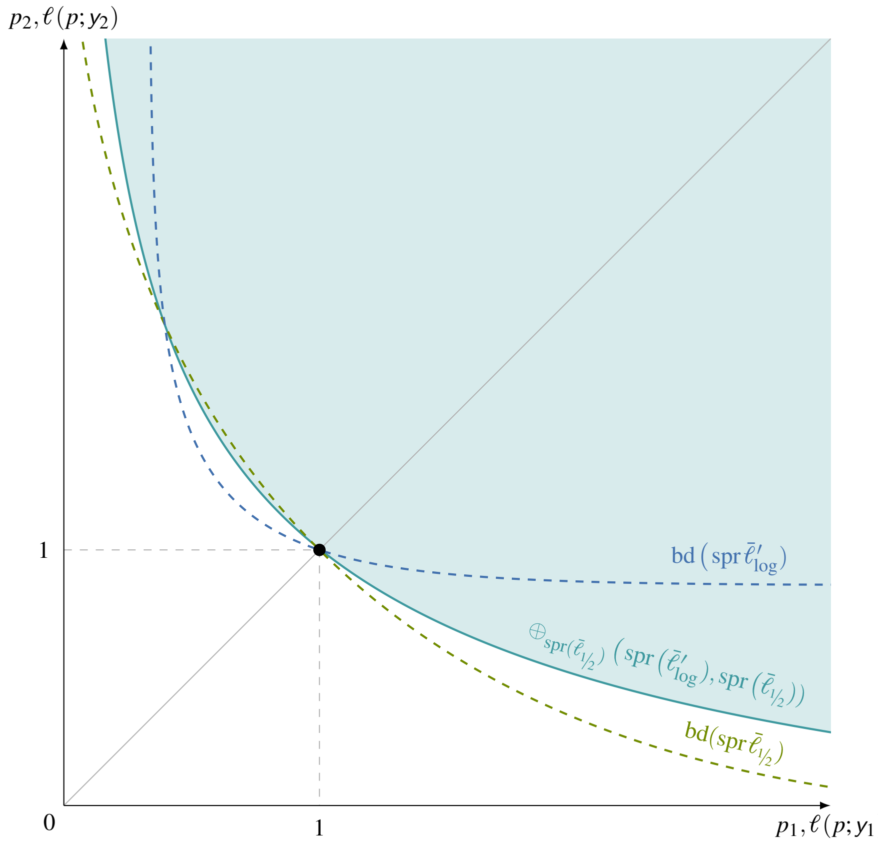


FIGURE 17: Illustration of Corollary 51. We take the normalised log loss $\bar{\ell}_{\log}$ (§5.1) and shift the maximum of its conditional Bayes risk (see §5.2) to $(3/4, 1/4)$ to generate the loss function $\bar{\ell}'_{\log}$. We then show the resulting renormalised M-sum of this loss with the normalised concave norm loss $\bar{\ell}_{1/2}$, and $M := \mathrm{spr}(\bar{\ell}_{1/2})$.

If $f_1, \ldots, f_m$ and g are concave functions the above two notations are overloaded with the *concave functional M-sum* and *dual concave functional M-sum*:

$$\mathscr{X} \ni x \mapsto \oplus_g(f_1, \ldots, f_m)(x) := g(f_1(x), \ldots, f_m(x))$$

and

$$\mathscr{X} \ni x \mapsto \oplus^*_g(f_1, \ldots, f_m)(x) := \sup\{g(f_1(a_1), \ldots, f_m(a_m)) \mid a_1 + \cdots + a_m = x\}.$$

The overload of notation can be defended since we will be either dealing with convex or concave functions but not combinations of the two.

6.7 Support Functions

We now justify the overload of the name M-sum for both the set operation and the functional operation, via the following theorem.

Theorem 52 *Let $A_i \subseteq \mathscr{X}$ for $i \in [m]$ and either*

1. $M \subseteq \mathbb{R}^m_{\geq 0}$ *with* $\mathrm{ri}(M) \neq \varnothing$, *or*

2. $M \subseteq \mathbb{R}^m$, *and* A_i *for* $i \in [m]$ *are each bounded.*

Then $\oplus_{\sigma_M}(\sigma_{A_1}, \ldots, \sigma_{A_m}) = \sigma_{\oplus_M(A_1,\ldots,A_m)}$.

Proof From the definition of the convex support function:

$$\forall x^* \in \mathscr{X}^*,\ \sigma_{\oplus_M(A_1,\ldots,A_m)}(x^*) = \sup_{s \in \oplus_M(A_1,\ldots,A_m)} \langle x^*, s\rangle = \sup_{\mu \in M} \sup_{s \in \sum_{i\in[m]} \mu_i \star A_i} \langle x^*, s\rangle. \tag{73}$$

To simplify analysis it is useful to be able to replace the epimultiplication in (73) with ordinary scalar multiplication.

Assume claim 1. Proposition 42 shows epimultiplication is continuous in the Painleve–Kuratowski sense with respect to sequences of positive scalars, this means that (73) implies the existence of a sequence $(\mu^n)_{n\in\mathbb{N}}$ with $\mu^n \in \mathrm{ri}(M)$ such that

$$\lim_{n\to\infty} \sup_{s \in \sum_{i\in[m]} \mu_i^n \star A_i} \langle x^*, s\rangle = \lim_{n\to\infty} \sup_{s \in \sum_{i\in[m]} \mu_i^n \cdot A_i} \langle x^*, s\rangle = \sup_{\mu\in M} \sup_{s \in \sum_{i\in[m]} \mu_i \star A_i} \langle x^*, s\rangle,$$

where the first equality follows since $\mathrm{ri}(M) \subseteq \mathrm{ri}(\mathbb{R}^m_{\geq 0}) = \mathbb{R}^m_{>0}$. Therefore replacing M by $\mathrm{ri}(M)$ in the supremum has no effect. Thus the expression in (73) simplifies, giving $\forall x^* \in \mathscr{X}^*$,

$$\begin{aligned}
&\sup_{\mu\in M} \sup_{s \in \sum_{i\in[m]} \mu_i \star A_i} \langle x^*, s\rangle \\
&= \sup_{\mu\in \mathrm{ri}(M)} \sup_{s \in \sum_{i\in[m]} \mu_i \cdot A_i} \langle x^*, s\rangle \\
&= \sup_{\mu\in \mathrm{ri}(M)} \sup_{a_1\in A_1} \left(\sup_{a_2 \in A_2} \left(\cdots \sup_{a_m\in A_m} \langle x^*, \mu_1 a_1 + \cdots + \mu_m a_m\rangle \right)\right) \\
&= \sup_{\mu\in \mathrm{ri}(M)} \sup_{a_1\in A_1} \cdots \sup_{a_m\in A_m} \Big(\langle x^*, \mu_1 a_1\rangle + \cdots + \langle x^*, \mu_m a_m\rangle \Big) \\
&= \sup_{\mu\in \mathrm{ri}(M)} \left(\sup_{a_1\in A_1} \mu_1 \langle x^*, a_1\rangle + \cdots + \sup_{a_m\in A_m} \mu_m \langle x^*, a_m\rangle \right) \\
&= \sup_{\mu\in \mathrm{ri}(M)} \left(\mu_1 \sigma_{A_1}(x^*) + \cdots + \mu_m \sigma_{A_m}(x^*)\right) \\
&= \sup_{\mu\in \mathrm{ri}(M)} \langle (\sigma_{A_1}(x^*), \ldots, \sigma_{A_m}(x^*)), \mu\rangle \\
&= \sigma_{\mathrm{ri}(M)}(\sigma_{A_1}(x^*), \ldots, \sigma_{A_m}(x^*)) \\
&\overset{\text{L7}}{=} \sigma_M(\sigma_{A_1}(x^*), \ldots, \sigma_{A_m}(x^*)),
\end{aligned} \tag{74}$$

as desired.

Now assume claim 2. Since each of the sets A_i are bounded, Proposition 1(1) implies that epimultiplication reduces to scalar multiplication. By a similar argument to (74) *mutatis mutandis* (we no longer need to replace M by $\mathrm{ri}(M)$) we are able to derive the same identity. ■

The relationship between convex and concave support functions (7) yields the following corollary:

Corollary 53 *Let $A_i \subseteq \mathscr{X}$ for $i \in [m]$ and either*

1. *$M \subseteq \mathbb{R}^m_{\geq 0}$ with $\mathrm{ri}(M) \neq \varnothing$, or*
2. *$M \subseteq \mathbb{R}^m$, and A_i for $i \in [m]$ are each bounded.*

Then $\oplus_{\rho_M}(\rho_{A_1}, \ldots, \rho_{A_m}) = \rho_{\oplus_M(A_1,\ldots,A_m)}$.

Corollary 54 *Let $\ell_1, \ldots, \ell_m$ be a sequence of proper loss functions with conditional Bayes risks $\underline{L}_1, \ldots, \underline{L}_m$. Let $\underline{M} \colon \mathbb{R}^m_{\geq 0} \to \overline{\mathbb{R}}$ be a conditional Bayes risk function. Let*

$$\jmath \in \partial \oplus_{\underline{M}}(\underline{L}_1, \ldots, \underline{L}_m) \text{ and } \hbar \in \partial \oplus^*_{\underline{M}}(\underline{L}_1, \ldots, \underline{L}_m).$$

Then $\jmath$ and $\hbar$ are proper loss functions.

This corollary shows how we can create new proper losses from old via the M-sums and dual M-sums. The following is a restatement in the terminology of M-sums of an old result which we require subsequently.

Proposition 55 (Hiriart-Urruty and Lemaréchal 2001, Theorem D.4.3.1)
Let $f_1, \ldots, f_m \colon \mathbb{R}^n \to \mathbb{R}$ be convex. Let $g \colon \mathbb{R}^m \to \mathbb{R}$ be convex and increasing component-wise in the sense that $x \succeq_{\mathbb{R}^m_{\geq 0}} y$ implies $g(x) \geq g(y)$. Define $F \coloneqq x \mapsto (f_1(x), \ldots, f_m(x))$. Then

$$\forall x \in \mathbb{R}^n,\ \partial(g \circ F)(x) = \oplus_{\partial g(F(x))}(\partial f_1(x), \ldots, \partial f_m(x)).$$

Corollary 56 *Let $(\underline{L}_i)_{i\in[m]}$, and $\underline{M}$ be as defined above, and let $\ell_i = \partial \underline{L}_i$ for $i \in [m]$ and $m = \partial \underline{M}$. Then*

$$\forall p \in \Delta,\ \underbrace{\begin{pmatrix} \ell_1(p, y_1) & \ldots & \ell_m(p, y_1) \\ \vdots & \ddots & \vdots \\ \ell_1(p, y_n) & \ldots & \ell_m(p, y_n) \end{pmatrix}}_{n\times m} \underbrace{m(\underline{L}_1(p), \ldots, \underline{L}_m(p))}_{m\times 1} \in \partial \oplus_{\underline{M}}(\underline{L}_1, \ldots, \underline{L}_m)(p) \subseteq \mathbb{R}^n.$$

Lemma 57 *Let $f_1, \ldots, f_m \colon \mathscr{X} \to \mathbb{R}$ be convex (resp. concave) functions then $\oplus^*_{\sigma_M}(f_1, \ldots, f_m)$ is convex (resp. $\oplus^*_{\rho_M}(f_1, \ldots, f_m)$ is concave).*

Proof Let $f_1, \ldots, f_m \colon \mathscr{X} \to \mathbb{R}$ be convex. Fix arbitrary $(a_i)_{i\in[m]}$ and $(b_i)_{i\in[m]}$ with $a_i, b_i \in \mathrm{dom}(f_i)$ for $i \in [m]$, and pick an arbitrary $t \in (0,1)$. Then

$$\begin{aligned} \forall i \in [m], \qquad & f_i(ta_i + (1-t)b_i) \ \leq\ t f_i(a_i) + (1-t) f_i(b_i) \\ \implies \quad \sup_{\mu\in M} \sum_{i\in[m]} \mu_i f_i(ta_i + (1-t)b_i) \ &\leq\ t \sup_{\mu\in M} \sum_{i\in[m]} \mu_i f_i(a_i) + (1-t) \sup_{\nu\in M} \sum_{j=1}^m \nu_j f_j(b_j). \end{aligned}$$

Since $(a_i)_{i\in[m]}$ and $(b_i)_{i\in[m]}$ are arbitrary, taking the infimum over both sides yields

$$\inf_{\substack{a_1+\cdots+a_m=x \\ b_1+\cdots+b_m=y}} \left(\sup_{\mu\in M} \sum_{i\in[m]} \mu_i f_i(ta_i + (1-t)b_i) \right)$$

$$
\begin{aligned}
&\leq \inf_{\substack{a_1+\cdots+a_m=x\\ b_1+\cdots+b_m=y}} \left(t \sup_{\mu\in M} \sum_{i\in[m]} \mu_i f_i(a_i) + (1-t) \sup_{\nu\in M} \sum_{j=1}^{m} \nu_j f_j(b_j) \right) \\
&= t \left(\inf_{a_1+\cdots+a_m=x} \sup_{\mu\in M} \sum_{i\in[m]} \mu_i f_i(a_i) \right) + (1-t) \left(\inf_{b_1+\cdots+b_m=y} \sup_{\nu\in M} \sum_{j=1}^{m} \nu_j f_j(b_j) \right) \\
&= t \oplus^*_{\sigma_M}(f_1,\ldots,f_m)(x) + (1-t) \oplus^*_{\sigma_M}(f_1,\ldots,f_m)(y). \qquad (75)
\end{aligned}
$$

To complete the proof

$$
\begin{aligned}
&\inf_{\substack{a_1+\cdots+a_m=x\\ b_1+\cdots+b_m=y}} \left(\sup_{\mu\in M} \sum_{i\in[m]} \mu_i f_i(ta_i+(1-t)b_i) \right) \\
&= \inf \left\{ \sup_{\mu\in M} \sum_{i\in[m]} \mu_i f_i(ta_i+(1-t)b_i) \;\middle|\; \begin{array}{l} \forall i\in[m],\ a_i,b_i\in\operatorname{dom}(f_i) \text{ and} \\ \sum_{i\in[m]} a_i = x \text{ and } \sum_{i\in[m]} b_i = y \end{array} \right\} \\
&= \inf \left\{ \sup_{\mu\in M} \sum_{i\in[m]} \mu_i f_i(c_i) \;\middle|\; \forall i\in[m],\ c_i\in\operatorname{dom}(f_i) \text{ and } \sum_{i\in[m]} c_i = tx+(1-t)y \right\} \\
&= \oplus^*_{\sigma_M}(f_1,\ldots,f_m)(tx+(1-t)y), \qquad (76)
\end{aligned}
$$

where the second equality follows since $\operatorname{dom}(f_i)$ is convex for each $i\in[m]$. Since (76) $\leq$ (75), we have that $\oplus^*_{\sigma_M}(f_1,\ldots,f_m)$ is convex. By an identical argument, *mutatis mutandis*, the concave result for the concave functional M-sum follows. ■

Lemma 58 *Let $S\in\mathcal{K}(\mathscr{X})$, and $\mu\in\mathbb{R}_{\geq 0}$. Then $x\mapsto \sup_{x^*\in\operatorname{dom}(\sigma_S)} \left(\langle x^*,x\rangle - \mu\sigma_S(x^*)\right) = (\sigma_{\mu\star A})^*$.*

Proof Let $\mu>0$. Then $\mu\sigma_S = \sigma_{\mu A} = \sigma_{\mu\star A}$ (Hiriart-Urruty and Lemaréchal, 2001, Theorem C.3.3.2) and the result follows by conjugation. Now assume $\mu=0$. Then from Lemma 9, $\overline{\operatorname{dom}(\sigma_S)} = -(\operatorname{rec} S)^*$ and $\forall x\in\mathscr{X}$,

$$
\begin{aligned}
\sup_{x^*\in\operatorname{dom}(\sigma_S)} \left(\langle x^*,x\rangle - \mu\sigma_S(x^*)\right) &= \sup_{x^*\in-(\operatorname{rec} S)^*} \langle x^*,x\rangle \\
&= \sup_{x^*\in(\operatorname{rec} S)^*} -\langle x^*,x\rangle \\
&= \begin{cases} 0 & x\in\operatorname{rec}(S) \\ \infty & \text{otherwise,} \end{cases} \qquad (77)
\end{aligned}
$$

where the final equality follows from the definition of the dual cone (2). Noticing (77) is the indicator of the set $\operatorname{rec}(S)$, it therefore is also the convex conjugate of the support function σ_S (Rockafellar, 1970, Theorem 13.2, p. 114). ■

Theorem 59 *Let $M\in\mathcal{R}(\mathbb{R}^m)$ be compact, $N\in\mathcal{S}(\mathbb{R}^m)$, and $A_i\in\mathcal{K}(\mathscr{X})$ for $i\in[m]$ with $\cap_{i\in[m]}A_i\neq\varnothing$. Then*

1. $\sigma_{\oplus^*_M(A_1,\ldots,A_m)} = \overline{\oplus^*_{\sigma_M}(\sigma_{A_1},\ldots,\sigma_{A_m})}$*, and*
2. $\rho_{\oplus^*_N(A_1,\ldots,A_m)} = \overline{\oplus^*_{\rho_N}(\rho_{A_1},\ldots,\rho_{A_m})}$.

Proof 1. Let $D := \mathrm{dom}(\sigma_{A_1}) \times \cdots \times \mathrm{dom}(\sigma_{A_m}) \subseteq (\mathscr{X}^*)^m$ and $a := (a_1, \ldots, a_m) \in D$. The Legendre-Fenchel conjugate (11) of the right-hand side is $\forall x \in \mathscr{X}$,

$$
\begin{aligned}
&\left(\oplus^*_{\sigma_M}(\sigma_{A_1}, \ldots, \sigma_{A_m})\right)^*(x) \\
&\qquad = \sup_{x^* \in \mathscr{X}^*} \left(\langle x^*, x\rangle - \inf_{a_1+\cdots+a_m = x^*} \sigma_M((\sigma_{A_1}(a_1), \ldots, \sigma_{A_m}(a_m))) \right) \\
&\qquad = \sup_{x^* \in \mathscr{X}^*} \left(\sup_{a_1+\cdots+a_m = x^*} \langle x^*, x\rangle - \sigma_M((\sigma_{A_1}(a_1), \ldots, \sigma_{A_m}(a_m))) \right) \\
&\qquad = \sup_{a_1,\ldots,a_m \in \mathscr{X}^*} (\langle a_1, x\rangle + \cdots + \langle a_m, x\rangle - \sigma_M((\sigma_{A_1}(a_1), \ldots, \sigma_{A_m}(a_m)))) \\
&\qquad = \sup_{a \in D} (\langle a_1, x\rangle + \cdots + \langle a_m, x\rangle - \sigma_M((\sigma_{A_1}(a_1), \ldots, \sigma_{A_m}(a_m)))) \\
&\qquad = \sup_{a \in D} \left(\langle a_1, x\rangle + \cdots + \langle a_m, x\rangle - \sup_{\mu \in M} \langle (\sigma_{A_1}(a_1), \ldots, \sigma_{A_m}(a_m)), \mu\rangle \right) \\
&\qquad = \sup_{a \in D} \inf_{\mu \in M} (\langle a_1, x\rangle + \cdots + \langle a_m, x\rangle - \langle (\sigma_{A_1}(a_1), \ldots, \sigma_{A_m}(a_m)), \mu\rangle) \\
&\qquad = \sup_{a \in D} \inf_{\mu \in M} L_x(a, \mu),
\end{aligned} \tag{78}
$$

where for $x \in \mathscr{X}$ we have $L_x \colon D \times M \to \mathbb{R}$ with

$$\forall a \in D,\ \forall \mu \in M,\ L_x(a, \mu) := \langle a_1, x\rangle + \cdots + \langle a_m, x\rangle - \langle (\sigma_{A_1}(a_1), \ldots, \sigma_{A_m}(a_m)), \mu\rangle.$$

Immediately $L_x(a, \mu)$ is concave and upper semi-continuous in a, and convex and lower semi-continuous in μ. Both D and M are convex and M is compact, and so we can apply Sion's minimax theorem (1958) and write

$$
\begin{aligned}
\sup_{a \in D} \inf_{\mu \in M} L_x(a, \mu) &= \inf_{\mu \in M} \sup_{a \in D} L_x(a, \mu) \\
&= \inf_{\mu \in M} \sum_{i \in [m]} \sup_{x^* \in \mathrm{dom}(\sigma_{A_i})} (\langle x^*, x\rangle - \mu_i \sigma_{A_i}(x^*)) \\
&= \inf_{\mu \in M} \sum_{i \in [m]} r_{A_i, \mu_i}(x),
\end{aligned} \tag{79}
$$

where $r_{A_i,\mu_i}(x) := \sup_{x^* \in \mathrm{dom}\, \sigma_{A_i}} (\langle x^*, x\rangle - \mu_i \sigma_{A_i}(x^*))$. Examining the functions r_{A_i,μ_i} with Lemma 58, we see that $r_{A_i,\mu_i} = \sigma^*_{\mu_i \star A_i}$ and in summary we have $\forall x \in \mathscr{X}$,

$$
\begin{aligned}
\left(\oplus^*_{\sigma_M}(\sigma_{A_1}, \ldots, \sigma_{A_m})\right)^*(x) &\overset{(78)}{=} \sup_{a \in D} \inf_{\mu \in M} L_x(a, \mu) \\
&\overset{(79)}{=} \inf_{\mu \in M} \sum_{i \in [m]} r_{A_i, \mu_i}(x) \\
&\overset{\text{L58}}{=} \inf_{\mu \in M} (\sigma^*_{\mu_1 \star A_1}(x) + \cdots + \sigma^*_{\mu_m \star A_m}(x)) \\
&= \inf_{\mu \in M} (\sigma_{\mu_1 \star A_1} \,\square\, \cdots \,\square\, \sigma_{\mu_m \star A_m})^*(x),
\end{aligned}
$$

where $\square$ denotes infimal convolution (4). We now apply the biconjugate theorem to both sides to obtain $\forall x^* \in \mathscr{X}^*$,

$$
\begin{aligned}
\left(\oplus^*_{\sigma_M}(\sigma_{A_1},\ldots,\sigma_{A_m})\right)^{**}(x^*) &= \sup_{x\in\mathscr{X}}\left(\langle x^*,x\rangle - \inf_{\mu\in M}(\sigma_{\mu_1\star A_1}\,\square\cdots\square\,\sigma_{\mu_m\star A_m})^*(x)\right)\\
&= \sup_{\mu\in M}\sup_{x\in\mathscr{X}}\left(\langle x^*,x\rangle - (\sigma_{\mu_1\star A_1}\,\square\cdots\square\,\sigma_{\mu_m\star A_m})^*(x)\right)\\
&= \sup_{\mu\in M}\overline{(\sigma_{\mu_1\star A_1}\,\square\cdots\square\,\sigma_{\mu_m\star A_m})}(x^*)\\
&= \sup_{\mu\in M}\sigma_{\mu_1\star A_1\cap\cdots\cap\mu_m\star A_m}(x^*)\\
&= \sigma_{\bigcup_{\mu\in M}\mu_1\star A_1\cap\cdots\cap\mu_m\star A_m}(x^*)\\
&= \sigma_{\oplus^*_M(A_1,\ldots,A_m)}(x^*),
\end{aligned}
$$

where the final two equalities are due to Hiriart-Urruty and Lemaréchal (2001, Theorem C.3.3.2). From Lemma 57 we have $\left(\oplus^*_{\sigma_M}(\sigma_{A_1},\ldots,\sigma_{A_m})\right)^{**} = \overline{\oplus^*_{\sigma_M}(\sigma_{A_1},\ldots,\sigma_{A_m})}$, which completes the proof of claim 1.

We now turn our attention to claim 2. Let $E := \mathrm{dom}(\rho_{A_1})\times\cdots\times\mathrm{dom}(\rho_{A_m}) \subseteq (\mathscr{X}^*)^m$ and $a := (a_1,\ldots,a_m)\in E$. The concave conjugate (12) of the right-hand side is $\forall x\in\mathscr{X}$,

$$
\begin{aligned}
&\left(\oplus^*_{\rho_N}(\rho_{A_1},\ldots,\rho_{A_m})\right)_*(x)\\
&\quad= \inf_{x^*\in\mathscr{X}^*}\left(\langle x^*,x\rangle - \sup_{a_1+\cdots+a_m=x^*}\rho_N((\rho_{A_1}(a_1),\ldots,\rho_{A_m}(a_m)))\right)\\
&\quad= \inf_{a\in E}\left(\langle a_1,x\rangle+\cdots+\langle a_m,x\rangle - \inf_{\mu\in N}\langle(\rho_{A_1}(a_1),\ldots,\rho_{A_m}(a_m)),\mu\rangle\right)\\
&\quad= \inf_{a\in E}\sup_{\mu\in M}\left(\langle a_1,x\rangle+\cdots+\langle a_m,x\rangle - \langle(\rho_{A_1}(a_1),\ldots,\rho_{A_m}(a_m)),\mu\rangle\right)\\
&\quad= \inf_{a\in E}\sup_{\mu\in N}E_x(a,\mu),
\end{aligned}
$$

where for $x\in\mathscr{X}$ we have $E_x\colon E\times N\to\mathbb{R}$ with

$$\forall a\in E,\ \forall\mu\in M,\ E_x(a,\mu) := \langle a_1,x\rangle+\cdots+\langle a_m,x\rangle - \langle(\rho_{A_1}(a_1),\ldots,\rho_{A_m}(a_m)),\mu\rangle.$$

We immediately have that $E_x(a,\mu)$ is convex and lower semi-continuous in a, and concave and upper semi-continuous in μ. In order to apply Lemma 36 we need to find certain sets $E'\subseteq E$ and $N'\subseteq N$ such that we satisfy (52). From the definition of E we have $0\in E$. From the 1-homogeneity of the functions ρ_{A_m} we know $E_x(0,\mu)=0$ for all $\mu\in N$. Therefore $\forall\mu\in N$,

$$\inf_{a\in E}E_x(a,\mu)\le E_x(0,\mu) \iff \inf_{a\in E}E_x(a,\mu)\le\inf_{b\in\{0\}}E_x(b,\mu)=0. \tag{80}$$

Let $N' := N\cap \mathrm{B}_{\beta_N(0)+1}\subseteq N$, where $\mathrm{B}_{\beta_N(0)+1}$ is the norm ball of radius $\beta_N(0)+1$. Then $N'\subseteq N$ is compact, and $E' := \{0\}\subseteq E$ is convex and nonempty. From (80) we have

$$\inf_{a\in E}\sup_{\mu\in N}E_x(a,\mu)\le 0 = \inf_{a\in E'}\sup_{\mu\in N'}E_x(a,\mu),$$

and therefore we satisfy (52) and we can apply Lemma 36.

From here we proceed with an argument that parallels the proof for claim 1, *mutatis mutandis* (infima and suprema exchanged, convex conjugates replaced with concave conjugates, inf convolution replaced with sup convolution) and we find $\rho_{\oplus_N^*(A_1,\dots,A_m)} = \overline{\oplus_{\rho_N}^*(\rho_{A_1},\dots,\rho_{A_m})}$, completing the proof of claim 2. ■

6.8 The *M*-Sum Polar

The polar (antipolar) operation plays an important role in our theory, and thus it is natural to ask how polarity interacts with the M-sum operations. We first need the following proposition.

Proposition 60 *Let $A_i \subseteq \mathcal{X}$ for $i \in I$ be an arbitrary collection of sets with index set I, let $R \in \mathcal{R}(\mathcal{X})$ and $S \in \mathcal{S}(\mathcal{X})$. Then*

1. $(\bigcup_{i\in I} A_i)^\circ = \bigcap_{i\in I}(A_i^\circ)$, *and* $(\bigcup_{i\in I} A_i)^\diamond = \bigcap_{i\in I}(A_i^\diamond)$;
2. $R = \operatorname{lev}_{\leq 1}(\gamma_R)$, *and* $S = \operatorname{lev}_{\geq 1}(\beta_S)$.

Proof The polar identity in claim 1 is a standard result in functional analysis (Aliprantis and Border, 2006, Corollary 5.104, p. 218). The proof for the antipolar is identical modulo the reversal of some inequalities. Claim 2 is immediate from the definition of star-shaped and co-star-shaped sets in §2.7. ■

We now show that the polarity operation preserves the radiant and shady nature of the sets.

Theorem 61 *Let $A_i \in \mathcal{R}(\mathcal{X})$ and $B_i \in \mathcal{S}(X_+)$ for $i \in [m]$, $M \in \mathcal{R}(\mathbb{R}^m)$ compact, $N \in \mathcal{S}(\mathbb{R}^m_{\geq 0})$. Then*

1. $\oplus_M(A_1,\dots,A_m) \in \mathcal{R}(\mathcal{X})$, *and* $\oplus_M^*(A_1,\dots,A_m) \in \mathcal{R}(\mathcal{X})$;
2. $\oplus_N(B_1,\dots,B_m) \in \mathcal{S}(\mathcal{X})$, *and* $\oplus_N^*(B_1,\dots,B_m) \in \mathcal{S}(\mathcal{X})$.

Proof We take the first case:

$$\begin{aligned}(0,1]\cdot\oplus_M(A_1,\dots,A_m) &= (0,1]\cdot\bigcup_{\mu\in M}\sum_{i\in[m]}\mu_i\star A_i\\ &= \bigcup_{\mu\in M}\sum_{i\in[m]}(0,1]\cdot\mu_i\star A_i\\ &= \bigcup_{\mu\in M}\sum_{i\in[m]}\mu_i\star(0,1]\cdot A_i\\ &= \bigcup_{\mu\in M}\sum_{i\in[m]}\mu_i\star A_i,\end{aligned}$$

and so it is clear that $\oplus_M(A_1,\dots,A_m)$ is star-shaped. By the same argument, *mutatis mutandis* $\oplus_M^*(A_1,\dots,A_m)$ is star-shaped, and $\oplus_N(B_1,\dots,B_m)$ and $\oplus_N^*(B_1,\dots,B_m)$ are co-star-shaped.

Theorem 45 guarantees convexity, and Theorem 48 guarantees closure. Proposition 46 guarantees the exclusion of the origin in the shady case, and so all that is left is to show $0 \in \operatorname{int}(\oplus_M(A_1,\dots,A_m))$, and $0 \in \operatorname{int}(\oplus_M^*(A_1,\dots,A_m))$.

By hypothesis $A_i \in \mathcal{R}(\mathcal{X})$ for each $i \in [m]$ and so $0 \in \operatorname{int}(A_i)$. By definition there exists $(r_i)_{i\in[m+1]}$ with $r_i \in \mathbb{R}_{>0}$ such that

$$\forall i \in [m],\ \mathrm{B}_{r_i} \subseteq A_i \subseteq \mathcal{X} \text{ and } \mathrm{B}_{r_{m+1}} \subseteq M \subseteq \mathbb{R}^m.$$

Let $r := \min_{i\in[m]} r_i$, $\mu' \in \mathbb{R}^m_{>0} \cap \mathrm{B}_{r_{m+1}} \subseteq M$, and $\mu'_{\min} := \min_{i\in[m]} \mu'_i > 0$. Then

$$\oplus_M(A_1,\ldots,A_m) = \bigcup_{\mu\in M} \sum_{i\in[m]} \mu_i \star A_i \supseteq \bigcup_{\mu\in M} \sum_{i\in[m]} \mu_i \star \mathrm{B}_r \supseteq \sum_{i\in[m]} \mu'_i \mathrm{B}_r \supseteq \mathrm{B}_{r\cdot\mu'_{\min}},$$

and

$$\oplus^*_M(A_1,\ldots,A_m) = \bigcup_{\mu\in M} \bigcap_{i\in[m]} \mu_i \star A_i \supseteq \bigcup_{\mu\in M} \bigcap_{i\in[m]} \mu_i \star \mathrm{B}_r \supseteq \bigcap_{i\in[m]} \mu'_i \cdot \mathrm{B}_r = \mathrm{B}_{r\cdot\mu'_{\min}},$$

completing the proof. ■

Lemma 62 *Let $R \in \mathcal{R}(\mathcal{X})$, $S \in \mathcal{S}(\mathcal{X})$ with $S \subseteq \operatorname{rec}(S)$, $\alpha \geq 0$. Then*

1. $\operatorname{lev}_{\leq\alpha} \gamma_R = \alpha \star R$, *and*
2. $\operatorname{lev}_{\geq\alpha} \beta_S = \alpha \star S$.

Proof Suppose $\alpha > 0$ and let $x^* \in \operatorname{lev}_{\leq\alpha}(\gamma_R)$, then exploiting the 1-homogeneity of γ_R:

$$\gamma_R(x^*) \leq \alpha \iff \frac{1}{\alpha}\gamma_R(x^*) \leq 1 \iff \gamma_{\alpha R}(x^*) \leq 1.$$

Thus applying Proposition 60(2) we find $\operatorname{lev}_{\leq\alpha} \gamma_R = \operatorname{lev}_{\leq 1}(\gamma_{\alpha R}) = \alpha R$. Penot and Zălinescu (2000, Proposition 2.3) prove $\operatorname{lev}_{=0}(\gamma_R) = \operatorname{rec}(R)$. Since gauge functions are nonnegative (9), this shows claim 1.

Suppose $\alpha > 0$ and let $x^* \in \operatorname{lev}_{\geq\alpha}(\beta_S)$, then appealing to the 1-homogeneity of β_S:

$$\beta_S(x^*) \geq \alpha \iff \frac{1}{\alpha}\beta_S(x^*) \geq 1 \iff \beta_{\alpha S}(x^*) \geq 1.$$

Thus applying Proposition 60(2) we find $\operatorname{lev}_{\geq\alpha}(\beta_S) = \operatorname{lev}_{\geq 1}(\beta_{\alpha S}) = \alpha S$. Now suppose $\alpha = 0$. Penot and Zălinescu (2000, Proposition 2.4) prove $\operatorname{lev}_{=0}(\beta_S) = \operatorname{rec}(S) \setminus \operatorname{cone}(S)$, and $\operatorname{lev}_{>0}(\beta_S) = \operatorname{cone}(S)$, giving $\operatorname{lev}_{\geq 0}(\beta_S) = \operatorname{rec}(S)$. This shows claim 2 and completes the proof. ■

The following general duality result extends a range of classical results, as well as results in (Seeger, 1990); it demonstrates the appealling fact that the polar of an M-sum of $(A_i)_i$ is the dual (polar-M)-sum of the polars of $(A_i)_i$; and similarly for antipolars.

Theorem 63 *Let $A_i \in \mathcal{R}(\mathcal{X})$ and $B_i \in \mathcal{S}(\mathcal{X})$ with $B_i \subseteq \operatorname{rec}(B_i)$ for $i \in [m]$, $M \in \mathcal{R}(\mathbb{R}^m)$, and $N \in \mathcal{S}(\mathbb{R}^m_{\geq 0})$. Assume M and A_i for $i \in [m]$ are compact. Then*

1. $\oplus_M(A_1,\ldots,A_m)^\circ = \oplus^*_{M^\circ}(A_1^\circ,\ldots,A_m^\circ)$, *and*
2. $\oplus_N(B_1,\ldots,B_m)^\diamond = \oplus^*_{N^\diamond}(B_1^\diamond,\ldots,B_m^\diamond)$.

Proof Calculating the polar of the left hand side of claim 1 we have

$$
\begin{aligned}
(\oplus_M(A_1,\dots,A_m))^\circ &\overset{(13)}{=} \operatorname{lev}_{\le 1}\left(\sigma_{\oplus_M(A_1,\dots,A_m)}\right)\\
&\overset{\text{T52}}{=} \operatorname{lev}_{\le 1}\left(\oplus_M(\sigma_{A_1},\dots,\sigma_{A_m})\right)\\
&= \operatorname{lev}_{\le 1}\left(x^* \mapsto \sigma_M((\sigma_{A_1}(x^*),\dots,\sigma_{A_m}(x^*)))\right)\\
&\overset{(15)}{=} \operatorname{lev}_{\le 1}\left(x^* \mapsto \gamma_{M^\circ}((\gamma_{A_1^\circ}(x^*),\dots,\gamma_{A_m^\circ}(x^*)))\right)\\
&\overset{\text{P60(2)}}{=} \left\{x^* \in \mathscr{X}^* \,|\, (\gamma_{A_1^\circ}(x^*),\dots,\gamma_{A_m^\circ}(x^*)) \in M^\circ\right\}\\
&= \bigcup_{\mu^*\in M^\circ} \left\{x^* \in \mathscr{X}^* \,\middle|\, (\gamma_{A_1^\circ}(x^*),\dots,\gamma_{A_m^\circ}(x^*)) = \mu^*\right\}\\
&= \bigcup_{\mu^*\in M^\circ} \left\{x^* \in \mathscr{X}^* \,\middle|\, \forall i \in [m],\ \gamma_{A_i^\circ}(x^*) = \mu_i^*\right\}. \qquad (81)
\end{aligned}
$$

Observe that M° is star-shaped and closed, thus $0 \in M^\circ$ and $[0,1]\cdot M^\circ = M^\circ$. Hence

$$
\begin{aligned}
\left(\oplus_M(A_1,\dots,A_m)\right)^\circ &\overset{(81)}{=} \bigcup_{\mu^*\in M^\circ} \left\{x^* \in \mathscr{X}^* \,\middle|\, \forall i \in [m],\ \gamma_{A_i^\circ}(x^*) = \mu_i^*\right\}\\
&= \bigcup_{\mu^*\in [0,1]\cdot M^\circ} \left\{x^* \in \mathscr{X}^* \,\middle|\, \forall i \in [m],\ \gamma_{A_i^\circ}(x^*) = \mu_i^*\right\}\\
&= \bigcup_{\mu^*\in M^\circ} \bigcup_{\lambda\in[0,1]} \left\{x^* \in \mathscr{X}^* \,\middle|\, \forall i \in [m],\ \gamma_{A_i^\circ}(x^*) = \lambda\mu_i^*\right\}\\
&= \bigcup_{\mu^*\in M^\circ} \left\{x^* \in \mathscr{X}^* \,\middle|\, \forall i \in [m],\ \gamma_{A_i^\circ}(x^*) \le \mu_i^*\right\}\\
&= \bigcup_{\mu^*\in M^\circ} \bigcap_{i\in[m]} \operatorname{lev}_{\le \mu_i^*}(\gamma_{A_i^\circ})\\
&\overset{\text{L62(1)}}{=} \bigcup_{\mu^*\in M^\circ} \bigcap_{i\in[m]} \mu_i^* \star A_i^\circ\\
&= \oplus^*_{M^\circ}(A_1^\circ,\dots,A_m^\circ), \qquad (82)
\end{aligned}
$$

which completes the proof of claim 1. The proof of claim 2 proceeds much like the proof of claim 1. Calculating the left hand side of claim 2, by a similar argument to (82), *mutatis mutandis*

$$
\begin{aligned}
(\oplus_N(B_1,\dots,B_m))^\diamond &\overset{(13)}{=} \operatorname{lev}_{\ge 1}\left(\rho_{\oplus_N(B_1,\dots,B_m)}\right)\\
&\overset{\text{C53}}{=} \operatorname{lev}_{\ge 1}\left(\oplus_N(\sigma_{B_1},\dots,\sigma_{B_m})\right)\\
&= \operatorname{lev}_{\ge 1}\left(x^* \mapsto \rho_N((\rho_{B_1}(x^*),\dots,\rho_{B_m}(x^*)))\right)\\
&\overset{(15)}{=} \operatorname{lev}_{\ge 1}\left(x^* \mapsto \beta_{N^\diamond}((\beta_{B_1^\diamond}(x^*),\dots,\beta_{B_m^\diamond}(x^*)))\right)\\
&\overset{\text{P60(2)}}{=} \left\{x^* \in \mathscr{X}^* \,\middle|\, (\beta_{B_1^\diamond}(x^*),\dots,\beta_{B_m^\diamond}(x^*)) \in N^\diamond\right\}\\
&= \bigcup_{\mu^*\in N^\diamond} \left\{x^* \in \mathscr{X}^* \,\middle|\, (\beta_{B_1^\diamond}(x^*),\dots,\beta_{B_m^\diamond}(x^*)) = \mu^*\right\}\\
&= \bigcup_{\mu^*\in N^\diamond} \left\{x^* \in \mathscr{X}^* \,\middle|\, \forall i \in [m],\ \beta_{B_i^\diamond}(x^*) = \mu_i^*\right\}. \qquad (83)
\end{aligned}
$$

Observe that $N^\diamond$ is co-star-shaped, thus $[1,\infty)\cdot N^\diamond = N^\diamond$. Hence

$$
\begin{aligned}
\big(\oplus_N(B_1,\ldots,B_m)\big)^\diamond &\overset{(83)}{=} \bigcup_{\mu^*\in N^\diamond} \big\{x^*\in\mathcal{X}^* \,\big|\, \forall i\in[m],\ \beta_{B_i^\diamond}(x^*)=\mu_i^*\big\} \\
&= \bigcup_{\mu^*\in[1,\infty)\cdot N^\diamond} \big\{x^*\in\mathcal{X}^* \,\big|\, \forall i\in[m],\ \beta_{B_i^\diamond}(x^*)=\mu_i^*\big\} \\
&= \bigcup_{\mu^*\in N^\diamond}\bigcup_{\lambda\in[1,\infty)} \big\{x^*\in\mathcal{X}^* \,\big|\, \forall i\in[m],\ \beta_{B_i^\diamond}(x^*)=\lambda\mu_i^*\big\} \\
&= \bigcup_{\mu^*\in N^\diamond} \big\{x^*\in\mathcal{X}^* \,\big|\, \forall i\in[m],\ \beta_{B_i^\diamond}(x^*)\geq\mu_i^*\big\} \\
&= \bigcup_{\mu^*\in N^\diamond}\bigcap_{i\in[m]} \operatorname{lev}_{\geq\mu_i^*}(\beta_{B_i^\diamond}) \\
&\overset{\text{L62(2)}}{=} \bigcup_{\mu^*\in N^\diamond}\bigcap_{i\in[m]} \mu_i^*\star B_i^\diamond \\
&= \oplus^*_{N^\diamond}(B_1^\diamond,\ldots,B_m^\diamond),
\end{aligned}
$$

which completes the proof of claim 2. ■

We can take the polars of both sides of the above theorem to obtain the result that the polar of the dual polar M-sum of the polars of $(A_i)_i$ is the M-sum of $(A_i)_i$:

Corollary 64 *Let $A_i\in\mathcal{R}(\mathcal{X})$ and $B_i\in\mathcal{S}(\mathcal{X})$ with $B_i\subseteq\operatorname{rec}(B_i)$ for $i\in[m]$, $M\in\mathcal{R}(\mathbb{R}^m)$, and $N\in\mathcal{S}(\mathbb{R}^m_{\geq 0})$. Assume M and A_i for $i\in[m]$ are compact. Then*

1. $\oplus_M(A_1,\ldots,A_m)=(\oplus^*_{M^\circ}(A_1^\circ,\ldots,A_m^\circ))^\circ$, *and*
2. $\oplus_N(B_1,\ldots,B_m)=(\oplus^*_{N^\diamond}(B_1^\diamond,\ldots,B_m^\diamond))^\diamond$.

Proof From Theorem 61 we know $\oplus_M(A_1,\ldots,A_m)$ is radiant and $\oplus_N(B_1,\ldots,B_m)$ is shady. The bipolar theorem (14) applied to Theorem 63 gives

$$\overline{\oplus_M(A_1,\ldots,A_m)}=(\oplus^*_{M^\circ}(A_1^\circ,\ldots,A_m^\circ))^\circ \quad\text{and}\quad \overline{\oplus_N(B_1,\ldots,B_m)}=(\oplus^*_{N^\diamond}(B_1^\diamond,\ldots,B_m^\diamond))^\diamond.$$

Theorem 48 and Corollary 49 makes the explicit closure operations redundant. ■

The above duality result can also be expressed in terms of gauges:

Corollary 65 *Let $A_i\in\mathcal{R}(\mathcal{X})$ and $B_i\in\mathcal{S}(\mathcal{X})$ with $B_i\subseteq\operatorname{rec}(B_i)$ for $i\in[m]$, $M\in\mathcal{R}(\mathbb{R}^m)$, and $N\in\mathcal{S}(\mathbb{R}^m_{\geq 0})$. Assume M and A_i for $i\in[m]$ are compact. Then*

1. $\gamma_{\oplus_M(A_1,\ldots,A_m)}=\overline{\oplus^*_{\gamma_M}(\gamma_{A_1},\ldots,\gamma_{A_m})}$, *and*
2. $\beta_{\oplus_N(B_1,\ldots,B_m)}=\overline{\oplus^*_{\beta_N}(\beta_{B_1},\ldots,\beta_{B_m})}$.

M	$\oplus_M(A_1,\ldots,A_m)$	Operation	$\oplus_{\sigma_M}(\sigma_{A_1},\ldots,\sigma_{A_m})$
$\{1_m\}$	$A_1+\cdots+A_m$	Minkowski sum	$\sum_{i=1}^m \sigma_{A_i}$
$\Delta([m])$	$\mathrm{co}\,(A_1\cup\cdots\cup A_m)$	Convex union	$\bigvee_{i=1}^m \sigma_{A_i}$
M	$\oplus_M^*(A_1,\ldots,A_m)$		$\oplus_{\sigma_M}^*(\sigma_{A_1},\ldots,\sigma_{A_m})$
$\{1_m\}$	$A_1\cap\cdots\cap A_m$	Intersection	$\sigma_{A_1}\,\square\cdots\square\,\sigma_{A_m}$
$\Delta([m])$	$A_1\lozenge\cdots\lozenge A_m$	Inverse sum	$\inf_{a_1+\cdots a_m=x}\bigvee_{i=1}^m \sigma_{A_i}$

TABLE 6: Examples of *convex* M-sums and dual M-sums. The inverse sum is discussed in (Rockafellar, 1970, page 21).

Proof The two identities are derived as follows:

$$\begin{aligned}\gamma_{\oplus_M(A_1,\ldots,A_m)} &\overset{(15)}{=} \sigma_{(\oplus_M(A_1,\ldots,A_m))^\circ} \\ &\overset{\text{T63(1)}}{=} \sigma_{\oplus_M^*(A_1^\circ,\ldots,A_m^\circ)} \\ &\overset{\text{T59}}{=} \overline{\oplus_{\rho_{M^\circ}}^*(\sigma_{A_1^\circ},\ldots,\sigma_{A_m^\circ})} \\ &\overset{(15)}{=} \overline{\oplus_{\gamma_M}^*(\gamma_{A_1},\ldots,\gamma_{A_m})},\end{aligned} \qquad \begin{aligned}\beta_{\oplus_M(B_1,\ldots,B_m)} &\overset{(15)}{=} \rho_{(\oplus_N(B_1,\ldots,B_m))^\diamond} \\ &\overset{\text{T63(2)}}{=} \rho_{\oplus_{N^\diamond}^*(B_1^\diamond,\ldots,B_m^\diamond)} \\ &\overset{\text{T59}}{=} \overline{\oplus_{\rho_{N^\diamond}}^*(\rho_{B_1^\diamond},\ldots,\rho_{B_m^\diamond})} \\ &\overset{(15)}{=} \overline{\oplus_{\beta_N}^*(\beta_{B_1},\ldots,\beta_{B_m})}.\end{aligned}$$

■

Finally, the duality result implies the following result expressed in terms of proper loss functions.

Corollary 66 *Let $\ell_1,\ldots,\ell_m$ be a sequence of proper loss functions with conditional Bayes risks $\underline{L}_1,\ldots,\underline{L}_m$. Let $\underline{M}:\mathbb{R}^m_{\geq 0}\to\overline{\mathbb{R}}$ be another conditional Bayes risk function. Then*

$$\ell\in\partial\oplus_{\underline{M}}(\underline{L}_1,\ldots,\underline{L}_m) \iff \ell^\diamond\in\partial\,\overline{\oplus_{\underline{M}^\diamond}^*(\underline{L}_1^\diamond,\ldots,\underline{L}_m^\diamond)}.$$

6.9 Conclusion on M-sums and New Losses from Old

The above development shows a general and powerful way of combining existing proper losses into new ones. An intriguing and satisfying feature of the results is that, in essence, the way you combine several proper losses into a new one, is to combine them using yet another proper loss! The M-sum operations are thus a very natural means to combine multiple existing proper loss functions to create a new proper loss function. Observe that by suitable scaling the component proper losses, one can smoothly interpolate between them in a variety of ways (depending upon the choice of M).

To provide some intuition, several classical examples of *convex* M-sums are presented in Table 6. Confer (Seeger, 1990) who uses different notation. See also (Mesikepp, 2016) and (Gardner et al., 2013) and (Kusraev and Kutateladze, 1995, Chapter 1).

7. Conclusion

We have presented a geometric theory of proper losses, whereby we take an unbounded convex set with particular properties as the starting point, and derive the proper loss as the (sub)-gradient of the support function of the set. The new perspective shows the natural duality between a loss and the "antipolar loss" newly introduced here, which is of value in understanding Vovk's aggregating algorithm (Kamalaruban et al., 2015). It also shows how one can generally combine multiple proper losses to create new proper losses. In developing that combinatorial theory, we extended a number of results on M-sums, and in particular proved an elegant and general duality theorem. The theory shows a deep but simple relationship between loss functions and concave gauges, the concave analog of a norm. Such concave gauges arise naturally in economics, but until now they have not been explicitly utilised within machine learning. We have presented the theory for finite outcomes (n-class probability estimation). Many of the results extend to a more general setting (Cranko, 2021).

The geometry of losses complements existing geometric approaches to machine learning, which have focussed on the geometry of the data distribution (through its likelihood function) (Amari, 2016), the geometry of the prior (Mahony and Williamson, 2001), and the geometry of the model class (Lee et al., 1998; Mendelson and Williamson, 2002).

The geometric theory developed here enables a general and insightful perspective relating Bayes risks and measures of information extending the results in (Reid and Williamson, 2011) and (Garcia-Garcia and Williamson, 2012). By developing measures of information in terms of convex sets (directly related to the superprediction sets used in the present paper) one can extend and refine the famous data processing theorem of information theory (Williamson and Cranko, 2022).

Given the foundational role loss functions play in a wide range of machine learning problems, it seems reasonable to suppose that the theory presented here will lead to further insights. One concrete direction for future work is to relate the geometry of the loss function developed here to the geometry of hypotheses classes and thus illuminate the interaction between losses and hypothesis classes that controls the speed of convergence in learning problems (van Erven et al., 2015). Ideally one would have a theory that simultaneously incorporated the loss function ℓ, the hypothesis class $\mathscr{F}$, the distribution of the data P, and one's prior knowledge into some overall geometric structure.

8. Acknowledgements

This work was inspired by conversations between RW and Mark Reid during 2009–2012. An early version of some of the results appeared in COLT2014. RW's contribution was funded in part by NICTA and the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy — EXC number 2064/1 — Project number 390727645. Thanks to extraordinarily assiduous referees for catching many typos and glitches in an earlier version.

References

Charalambos D. Aliprantis and Kim Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer Science & Business Media, 2006.

Claudi Alsina, Justyna Sikorska, and M. Santos Tomás. *Norm Derivatives and Characterizations of Inner Product Spaces*. World Scientific, 2010.

Shun-ichi Amari. *Information Geometry and its Applications*. Springer, 2016.

Alexandre Y. Aravkin, James V. Burke, Dmitriy Drusvyatskiy, Michael P. Friedlander, and Kellie J. MacPhee. Foundations of gauge and perspective duality. *SIAM Journal on Optimization*, 28(3): 2406–2434, 2018.

Shiri Artstein-Avidan and Yanir A. Rubinstein. Differential analysis of polarity: Polar Hamilton-Jacobi, conservation laws, and Monge Ampère equations. *Journal d'Analyse Mathématique*, 132 (1):133–156, 2017.

Edward H. Askwith. *A Course of Pure Geometry*. Cambridge University Press, 1917.

Alfred Auslender and Marc Teboulle. *Asymptotic cones and functions in optimization and variational inequalities*. Springer, 2003.

Abdessamad Barbara and Jean-Pierre Crouzeix. Concave gauge functions and applications. *ZOR — Mathematical Methods of Operations Research*, 40:43–74, 1994.

Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.

Valeriĭ N. Berestovskĭ and Victor M. Gichev. Metrized semigroups. *Journal of Mathematical Sciences*, 119(1):10–29, 2004.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985.

Marcel Berger. *Geometry Revealed: A Jacob's Ladder to Modern Higher Geometry*. Springer, 2010.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, January 1950.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, University of Pennsylvania, November 2005.

Shiing-Shen Chern. Finsler geometry is just Riemannian geometry without the quadratic restriction. *Notices of the American Mathematical Society*, 43(9):959–963, 1996.

Gustave Choquet. Les cônes convexes faiblement complets dans l'analyse. *Proceedings of the International Congress of Mathematicians, Stockholm*, pages 317–330, 1962.

Charles W. Cobb and Paul H. Douglas. A theory of production. *The American Economic Review*, 18 (1):139–165, March 1928.

Richard Cornes. *Duality and Modern Economics*. Cambridge University Press, 1992.

Zac Cranko. *An analytic approach to the structure and composition of General Learning Problems*. PhD thesis, The Australian National University, 2021. URL `https://openresearch-repository.anu.edu.au/bitstream/1885/219338/1/Cranko%20Thesis%202021.pdf`.

A. Philip Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, March 2007.

Ricky Der and Daniel Lee. Large-margin classification in Banach spaces. In *Artificial Intelligence and Statistics*, pages 91–98. PMLR, 2007.

Michel Marie Deza and Elena Deza. *Encyclopedia of Distances*. Springer, 2009.

Jean Dieudonné. *History of Functional Analysis*. North-Holland Publishing Company, 1981.

Rolf Färe and Daniel Primont. The unification of Ronald W. Shephard's duality theory. *Journal of Economics (Zeitschrift für Nationalökonomie)*, 60(2):199–207, 1994.

Rolf Färe and Daniel Primont. *Multi-output Production and Duality: Theory and Applications*. Kluwer Academic Publishers, 1995.

Michael P. Friedlander, Ives Macedo, and Ting Kei Pong. Gauge optimization, duality, and applications. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.

Dario Garcia-Garcia and Robert C. Williamson. Divergences and Risks for Multiclass Experiments. In *Conference on Learning Theory (JMLR: W&CP)*, volume 23, pages 28.1–28.20, 2012.

Richard J. Gardner and Markus Kiderlen. Operations between functions. *Communications in Analysis and Geometry*, 26(4):787–855, 2018.

Richard J. Gardner, Daniel Hug, and Wolfgang Weil. Operations between sets in geometry. *Journal of the European Mathematical Society*, 15:2297–2352, 2013.

John Robilliard Giles. Classes of semi-inner-product spaces. *Transactions of the American Mathematical Society*, 129(3):436–446, 1967.

Armand Gissler and Tim Hoheisel. A note on the K-epigraph. *Optimization*, (online preview):1–35, 2022.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and its Applications*, 1:125–151, 2014.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007.

Irving John Good. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114, 1952.

Peter D. Grünwald and A. Phillip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

Chung Ha. A noncompact minimax theorem. *Pacific Journal of Mathematics*, 97(1):115–117, 1981.

David J. Hand and Veronica Vinciotti. Local Versus Global Models for Classification Problems: Fitting Models Where it Matters. *The American Statistician*, 57(2):124–131, 2003.

Giora Hanoch. Symmetric duality and polar production functions. In Melvyn Fuss and Daniel McFadden, editors, *Production Economics: A Dual Approach to Theory and Applications*, volume 1, pages 111–132. North-Holland, 1978.

Georg Hasenkamp and Jürgen Schrader. Dual polar price and quantity aggregation. *Zeitschrift für Nationalökonomie*, 38(3-4):305–322, 1978.

Eduard Helly. Über Systeme linear Gleichungen mit unendlich vielen Unbekannten. *Monatshefte für Mathematik und Physik*, 31(1):60–91, 1921.

Arlo D. Hendrickson and Robert J. Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42(6):1916–1921, 1971.

Ralf Herbrich and Robert C. Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3(Sep):175–212, 2002.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer, Berlin, 2001.

Ákos G. Horváth, Zsolt Lángi, and Margarita Spirova. Semi-inner products and the concept of semi-polarity. *Results in Mathematics*, 71(1-2):127–144, 2017.

Hendrik S. Houthhakker. A note on self-dual preferences. *Econometrica*, 33(4):797–801, October 1965.

Stephen E. Jacobsen. On Shephard's duality theorem. *Journal of Economic Theory*, 4:458–464, 1972.

Johannes Jahn. *Vector optimization: Theory, Applications, and Extensions*. Springer, 2nd edition, 2011.

Dale W. Jorgenson. Foreword. In *Ronald W. Shephard: Cost and Production Functions (Reprint of the First Edition)*, pages iii–iv. Spinger-Verlag, 1981.

Yuri Kalnishkan and Michael V. Vyugin. Mixability and the existence of weak complexities. In *The 15th Annual Conference on Computational Learning Theory (COLT 2002)*, volume 2375 of *Lecture Notes in Artificial Intelligence*, pages 105–120. Springer-Verlag, 2002.

Yuri Kalnishkan, Volodya Vovk, and Michael V. Vyugin. Loss functions, complexities, and the Legendre transformation. *Theoretical Computer Science*, 313:195–207, 2004.

Parameswaran Kamalaruban, Robert C. Williamson, and Xinhua Zhang. Exp-concavity of proper composite losses. In *Conference on Learning Theory (JMLR: W&CP)*, volume 40, pages 1–31, 2015.

Anatoly G. Kusraev and Semen Samsonovich Kutateladze. *Subdifferentials: theory and applications*. Kluwer, 1995.

Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

Dinh The Luc and Michel Volle. Levels sets infimal convolution and level addition. *Journal of optimization theory and applications*, 94:695–714, 1997.

Günter Lumer. Semi-inner-product spaces. *Transactions of the American Mathematical Society*, 100 (1):29–43, 1961.

Helmut Lütkepohl. *Handbook of Matrices*. John Wiley and Sons, 1996.

Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics (revised edition)*. John Wiley & Sons, 1999.

Robert E. Mahony and Robert C. Williamson. Prior knowledge and preferential structures in gradient descent learning algorithms. *Journal of Machine Learning Research*, 1:311–355, 2001.

Horst Martini and Konrad J. Swanepoel. Antinorms and Radon curves. *Aequationes Mathematicae*, 72(1-2):110–138, 2006.

Horst Martini, Konrad J. Swanepoel, and Gunter Weiß. The geometry of Minkowski spaces — a survey. Part I. *Expositiones Mathematicae*, 19:97–142, 2001.

Andreu Mas-Collel, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

Daniel McFadden. Cost, revenue, and profit functions. In Melvyn Fuss and Daniel McFadden, editors, *Production Economics: A Dual Approach to Theory and Applications*, volume 1, pages 1–110. North-Holland, 1978.

Shahar Mendelson and Robert C. Williamson. Agnostic learning nonconvex function classes. In *International Conference on Computational Learning Theory*, pages 1–13. Springer, 2002.

Karl Menger. Untersuchungen über allgemeine Metrik. *Mathematische Annalen*, 100:75–163, 1928.

Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. *FAT*18, Proceedings of Machine Learning Research*, 81:1–43, 2018a.

Aditya Krishna Menon and Robert C. Williamson. A loss framework for calibrated anomaly detection. In *Proceedings of the 32nd international conference on neural information processing systems*, pages 1494–1504, 2018b.

Jorma Kaarlo Merikoski. On c-norms and c-antinorms on cones. *Linear Algebra and its Applications*, 150:315–329, 1991.

Tim Mesikepp. M-addition. *Journal of Mathematical Analysis and Applications*, 443:146–177, 2016.

Zakaria Mhammedi and Robert C. Williamson. Constant regret, generalized mixability, and mirror descent. In *Neural Information Processing Systems*, 2018.

Vitali Milman and Liran Rotem. Non-standard constructions in convex geometry: geometric means of convex bodies. In *Convexity and concentration*, pages 361–390. Springer, 2017a.

Vitali Milman and Liran Rotem. "Irrational" constructions in convex geometry. *St. Petersburg Mathematics Journal*, 29(1):165–175, 2017b.

Hermann Minkowski. *Geometrie der Zahlen*. B.G. Teubner, 1896.

Boris S. Mordukhovich and Yongheng Shao. On nonconvex subdifferential calculus in Banach spaces. *Journal of Convex Analysis*, 2(1/2):211–227, 1995.

Maria Moszyńska. *Selected Topics in Convex Geometry*. Birkhäuser, 2006.

Maria Moszyńska and Wolf-Dieter Richter. Reverse triangle inequality: Antinorms and semi-antinorms. *Studia Scientiarum Mathematicarum Hungarica*, 49(1):120–138, 2012.

Armando Cabrera Pacheco and Robert C. Williamson. The geometry of mixability. *Transactions on Machine Learning Research*, 2023.

Diethard Ernst Pallaschke and Ryszard Urbanski. *Pairs of compact convex sets: fractional arithmetic with convex sets*. Springer Science & Business Media, 2013.

Teemu Pennanen. Graph-convex mappings and K-convex functions. *Journal of Convex Analysis*, 6 (2):235–266, 1999.

Jean-Paul Penot. Duality for radiant and shady programs. *Acta Mathematica Vietnamica*, 22(2): 541–566, 1997.

Jean-Paul Penot. The bearing of duality on microeconomics. In *Advances in Mathematical Economics*, pages 113–139. Springer, 2005.

Jean-Paul Penot. *Calculus Without Derivatives*. Springer, 2012.

Jean-Paul Penot and Constantin Zălinescu. Approximation of functions and sets. In *Approximation, optimization and mathematical economics*, pages 255–274. Springer, 2001.

Jean-Paul Penot and Constantin Zălinescu. Harmonic sum and duality. *Journal of Convex Analysis*, 7(1):95–113, 2000.

Albrecht Pietsch. *History of Banach Spaces and Linear Operators*. Birkhäuser, 2007.

John R. Platt. Strong inference. *Science*, 146(3642):347–353, October 1962.

Friedrich Pukelsheim. On information functions and their polars. *Journal of Optimization Theory and Applications*, 41(4):533–546, 1983.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

R. Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

R. Tyrrell Rockafellar. *Monotone Processes of Convex and Concave Type*, volume 77 of *Memoirs of the American Mathematical Society*. American Mathematical Society, 1967.

R. Tyrrell Rockafellar and Roger J-B. Wets. *Variational Analysis*. Springer-Verlag, 2004.

Gian-Carlo Rota. *Indiscrete Thoughts*. Birkhäuser, 1997.

Hanno Rund. *The differential geometry of Finsler spaces*. Springer-Verlag, 1959.

Paul A. Samuelson. Using full duality to show that simultaneously additive direct and indirect utilities implies unitary price elasticity of demand. *Econometrica*, 33(4):781–796, October 1965.

Ryuzo Sato. Self-dual preferences. *Econometrica*, 44(5):1017–1032, September 1976.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Rolf Schneider. *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge University Press, 2014. Second expanded edition.

Alberto Seeger. Direct and inverse addition in convex analysis and applications. *Journal of Mathematical Analysis and Applications*, 148:317–349, 1990.

Alberto Seeger. Second derivatives of a convex function and of its Legendre-Fenchel transformate. *SIAM Journal on Optimization*, 2(3):405–424, 1992.

Alberto Seeger and Michel Volle. On a convolution operation obtained by adding level sets: classical and new results. *Recherche opérationnelle/Operations Research*, 29(2):131–154, 1995.

Ronald W. Shephard. *Cost and Production Functions*. Princeton University Press, 1953.

Ronald W. Shephard. *Theory of Cost and Production Functions*. Princeton University Press, 1970.

Emir H. Shuford Jr., Arthur Albert, and H. Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958.

Boaz A. Slomka. On duality and endomorphisms of lattices of closed convex sets. *Advances in Geometry*, 11:225–239, 2011.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Anthony C. Thompson. *Minkowski Geometry*. Cambridge University Press, 1996.

Tim van Erven, Mark D. Reid, and Robert C. Williamson. Mixability is Bayes risk curvature relative to log loss. *Journal of Machine Learning Research*, 13:1639–1663, 2012.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

Hal R. Varian. *Microeconomic Analysis*. W.W. Norton and Company, 1978.

Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17(223):1–52, 2016.

Michel Volle. Duality for the level sum of quasiconvex functions and applications. *ESAIM: Control, Optimisation and calculus of variations*, 3:329–343, 1998.

Carl-Axel S. Staël von Holstein. A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology*, 9:360–364, 1970.

Volodya Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT)*, pages 371–383, 1990.

Volodya Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 51–60. ACM, 1995.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Bo Waggoner. Linear functions to the extended reals. *arXiv:2102.09552*, 2021.

Abraham Wald. *Statistical Decision Functions*. John Wiley & Sons, 1950.

Robert C. Williamson. Loss functions. In Bernhard Schölkopf, Zhiyuan Luo, and Vladimir Vovk, editors, *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 71–80. Springer, 2013.

Robert C. Williamson. The geometry of losses. In *Proceedings of The 27th Conference on Learning Theory*, pages 1078–1108, 2014.

Robert C. Williamson and Zac Cranko. Information processing equalities and the information–risk bridge. arXiv:2207.11987, 2022.

Haizhang Zhang, Yuesheng Xu, and Jun Zhang. Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research*, 10(12), 2009.

Fedor Zhdanov. *Theory and Applications of Competitive Prediction*. PhD thesis, Department of Computer Science, Royal Holloway, University of London, 2011.

Constantin Zălinescu. *Convex analysis in general vector spaces*. World Scientific, 2002.

Constantin Zălinescu. On the differentiability of the support function. *Journal of Global Optimization*, 57(3):719–731, 2013.

Constantin Zălinescu. Relations between the convexity of a set and the differentiability of its support function. *Optimization*, 65(3):651–670, 2016.