JMLR: Workshop and Conference Proceedings vol 35 (2014) 1–31

The Geometry of Losses

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

Abstract

Loss functions are central to machine learning because they are the means by which the quality of a prediction is evaluated. Any loss that is not proper, or can not be transformed to be proper via a link function is inadmissible. All admissible losses for n-class problems can be obtained in terms of a convex body in $\mathbb{R}^n$. We show this explicitly and show how some existing results simplify when viewed from this perspective. This allows the development of a rich algebra of losses induced by binary operations on convex bodies (that return a convex body). Furthermore it allows us to define an "inverse loss" which provides a universal "substitution function" for the Aggregating Algorithm. In doing so we show a formal connection between proper losses and norms.

Keywords: convex bodies, support functions, gauges, polars, proper losses, distorted probabilities, inverse losses, entropies, norms, Bregman divergences, aggregating algorithm, substitution functions.

1. Introduction

Loss functions are central to machine learning because they are the means by which the quality of a prediction is evaluated. The choice of loss matters, especially when there is modelling error, which is of course the usual situation (Buja et al., 2005; Hummel and McAfee, 2013; Merkle and Steyvers, 2013; Hand, 1994; Hand and Vinciotti, 2003; Gneiting, 2011; Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014).

We will restrict ourselves to finite dimensional action spaces (suitable for multiclass probability estimation or classification). The construction of multiclass losses is complex; see the summary of previous work in (Vernet et al., 2012; Williamson, 2013). There is a simple admissibility argument (Vernet et al., 2012) that shows that there is no point in using losses that are neither proper nor proper composite (the composition of a proper loss with an invertible link function). This is why there has been a focus of work on proper losses which are losses for probability estimation that are minimised when predicting the true underlying probability (a formal definition is provided later).

This paper presents a new way to develop proper losses. Rather than starting with a loss function and hence defining the superprediction set (a convex body in $\mathbb{R}^n$), we start with the convex body and derive the loss. This trivial change of perspective opens up several interesting avenues. It allows us to connect to a range of results in the theory of convex bodies (Schneider, 2014) and Minkowski geometry (Thompson, 1996). It shows that rather than defining the Bayes risk (or entropy) on the simplex Δ^n, it is slicker to define it on the positive orthant $\mathbb{R}^n_+$ but require positive one-homogeneity. This then implies that the induced proper loss is also defined on $\mathbb{R}^n_+$ but is zero-homogeneous. One can handle the non-smooth or non-differentiable case as easily as the smooth case. It leads naturally to the idea of an "inverse loss" which is what one needs to form a substitution function for the Aggregating Algorithm (Vovk, 2001, 1995, 1990). We also will show, by extending known results

- $0 \notin S \subseteq \mathbb{R}^n$, S convex $\Rightarrow \ell_S = \hat{\partial}\hat{\sigma}_S$ — the supergradient of the concave support function of S induces a loss ℓ_S (with S being its superprediction set) and which is proper, 0-homogeneous and defined on the entire positive orthant $\ell_S : \mathbb{R}^n_+ \to \mathbb{R}^n_+$.
- Bayes risk $\underline{L} = \hat{\sigma}_S$ is 1-homogeneous and so $\underline{L}\colon \mathbb{R}^n_+ \to \mathbb{R}$ (rather than $\underline{L}\colon \Delta^n \to \mathbb{R}$)
- Consequently Bregman divergences simplify: $B_{-\underline{L}}(x, y) = \langle x, \hat{\partial}\underline{L}(X) - \hat{\partial}\underline{L}(y)\rangle$.
- Inverse loss of ℓ_S is induced by the concave polar of S: $\ell_S^{-1} = \hat{\partial}\hat{\sigma}_{S^{\circledast}}$. It satisfies $\ell_S \circ \ell_S^{-1} \circ \ell_S = \ell_S$. It provides a generic substitution function for the aggregating algorithm.
- Algebra on convex bodies forms a basis for algebra of losses; generalisations of existing binary operations on convex bodies — see (17), (18) and (19).

Figure 1: Summary of key contributions of the paper.

on generalised direct and inverse Minkowski addition of convex sets, how to induce a wide range of binary operations on proper losses. This gives insight into how to design and construct such losses.

Since proper losses are such a foundational concept it is hardly surprising there have been several attempts to develop a geometric understanding of them. Early attempts include (McCarthy, 1956) who showed a scoring rule is proper if and only if it is the derivative of a positively homogeneous convex function. Our perspective simply involves recognising that such a function must be the support function of a convex body, and using that body as the starting point. Since then, there have been many works studying the geometry of proper losses including (Staël von Holstein, 1970; Murphy and Staël von Holstein, 1975; Staël von Holstein and Murphy, 1978). Further historical references can be found in (Reid and Williamson, 2011).

More recently Dawid (2007); Dawid and Lauritzen (2006) have drawn connections between differential geometry and proper losses (scoring rules). In particular, Dawid (2007) presented the relationship between a proper loss and the superprediction set (defined below). However he did not develop this idea along the lines of the present paper. Finally we merely mention the work by Ruberry (2013) who has a variant of the normal perspective, but again does not fully exploit the viewpoint adopted in this paper.

The rest of the paper is organised as follows. In §2 we introduce the basic mathematical machinery we will use and set notation. In §3 we introduce the notion of a gauge and the polar of a convex body. Section 4 formally introduces proper losses and shows their connection to convex bodies. Section 5 introduces the idea of an inverse loss and shows its relationship to the polar of the superprediction set of the loss in question. Section 6 studies binary operations on convex bodies and their implication for combining and designing proper losses. In doing so we present a generalisation of an earlier result by Seeger (1990) that may be of independent interest. Section 7 concludes. Appendices A–F derive some additional properties and illustrate the general theory with examples (l_p loss, Brier loss, cost-sensitive misclassification loss, and the boosting loss which is shown to correspond to a Cobb-Douglas style support function and is in fact self-inverse). Detailed proofs are in Appendix G. The gist of the paper is summarised in Figure 1; however perhaps the main value of the paper is the apparently new viewpoint.

2. Preliminaries

We introduce some standard machinery from the theory of convex sets and functions; see (Hiriart-Urruty and Lemaréchal, 2001; Rockafellar, 1970; Schneider, 2014). The concave cases of some of these results are in (Pukelsheim, 1983; Barbara and Crouzeix, 1994).

Let $\mathbb{R}_- := (-\infty, 0]$, $\mathbb{R}_+ := [0, \infty)$, and let $\mathcal{K}^n$ denote the class of convex subsets of $\mathbb{R}^n$. If $S, T \subset \mathbb{R}^n$, $x \in \mathbb{R}^n$, $\alpha > 0$, then $\alpha S := \{\alpha s \colon s \in S\}$, $S + x := \{s + x \colon s \in S\}$, and the *Minkowski sum* $S + T := \{s + t \colon s \in S,\ t \in T\}$. If $V \subset \mathbb{R}_+$, then $V \cdot S := \{vs \colon v \in V,\ s \in S\}$.

A convex set $C \subset \mathbb{R}^n$ *recedes in the direction* $y \in \mathbb{R}^n$ if $x + \lambda y \in C$, $\forall \lambda \geq 0$, $\forall x \in C$. The *recession cone* $0^+C := \{y \in \mathbb{R}^n \colon C$ recedes in direction $y\}$. Let $H_c(s) := \{y \in \mathbb{R}^n \colon \langle s, y \rangle = c\}$ denote the *hyperplane with normal s and offset c*. Given a set $C \subset \mathbb{R}^n$, we denote the *boundary* $\operatorname{bd} C$, the *interior* $\operatorname{int} C$, the *closure* $\operatorname{cl} C$, and *convex hull* $\operatorname{co} C$. The convex hull of a set C is equal to the intersection of all the supporting half-spaces.

If $f \colon \mathbb{R}^n \to \mathbb{R}$ its *epigraph* and *hypograph* are

$$\operatorname{epi} f := \{(x, y) \in \mathbb{R}^{n+1} \colon f(x) \leq y\} \text{ and } \operatorname{hypo} f := \{(x, y) \in \mathbb{R}^{n+1} \colon f(x) \geq y\}$$

and the *below*, *level* and *above sets* are $\operatorname{lev}_{\leq \alpha}(f) := \{x \in \operatorname{dom} f \colon f(x) \leq \alpha\}$, $\operatorname{lev}_{= \alpha}(f) := \{x \in \operatorname{dom} f \colon f(x) = \alpha\}$, and $\operatorname{lev}_{\geq \alpha}(f) := \{x \in \operatorname{dom} f \colon f(x) \geq \alpha\}$.

When working with losses, concave functions naturally arise. Many of the results we use are developed for convex functions. Although in some cases this is merely a matter of a sign flip, in the case of polars and gauges it is more subtle. Thus to avoid confusion via overloading the notation too much, we will annotate those symbols using $\hat{}$ and $\check{}$ for the concave and convex cases respectively. Suppose $\phi \colon \mathbb{R}^n \to \mathbb{R}$ is convex. Its *subdifferential* is

$$\check{\partial}\phi(x) := \{x^* \colon \phi(x) + \langle x^*, y - x \rangle \leq \phi(y)\ \forall y \in \mathbb{R}^n\}.$$

Suppose $\phi \colon \mathbb{R}^n \to \mathbb{R}$ is concave. Its *superdifferential* is

$$\hat{\partial}\phi(x) := \{x^* \colon \phi(x) + \langle x^*, y - x \rangle \geq \phi(y)\ \forall y \in \mathbb{R}^n\}.$$

A function $f \colon \mathbb{R}^n \to \mathbb{R}$ is *1-homogeneous* (resp. *0-homogeneous*) if for all $\alpha > 0$ $f(\alpha x) = \alpha f(x)$ (resp. $f(\alpha x) = f(x)$) for all $x \in \mathbb{R}^n$. Euler's theorem for homogeneous functions is typically stated for differentiable functions. We will make use of the following special case for 1-homogeneous functions which holds for subdifferentials. The proof is straight-forward and included in Appendix G. Here ∂ can stand for either $\hat{\partial}$ or $\check{\partial}$.

Proposition 1 *Suppose $f \colon \mathbb{R}^n \to \mathbb{R}$ is 1-homogeneous. Then ∂f is 0-homogeneous.*

A set C is of *negative type* (resp. *positive type*) if $C \in \mathcal{K}^n$, C is closed, $0 \in \operatorname{int} C$ and $0^+C = \mathbb{R}^n_-$ (resp. $0 \notin \operatorname{int} C$ and $0^+C = \mathbb{R}^n_+$). The class of sets of negative (resp. positive type) is denoted $\check{\mathcal{K}}$ (resp. $\hat{\mathcal{K}}$). The *support function* of a set C is a central object in convex analysis:

$$\check{\sigma}_C(x) := \sup_{y \in C} \langle x, y \rangle.$$

If $C = \bigcup_{i \in I} C_i$, then $\check{\sigma}_C(x) = \sup_{i \in I} \check{\sigma}_{C_i}(x)$. This support function corresponds naturally to proper gains; when working with proper losses we will use

$$\hat{\sigma}_C(x) := \inf_{y \in C} \langle x, y \rangle.$$

If $C = \bigcup_{i\in I} C_i$, then $\hat{\sigma}_C(x) = \inf_{i\in I} \hat{\sigma}_{C_i}(x)$. The support functions are closed, one-homogeneous and $\check{\sigma}_C$ is convex and $\hat{\sigma}_C$ is concave.

The *support plane* $H_C^+(u) := \{x \in \mathbb{R}^n \colon \langle x, u\rangle = \sigma_C(u)\}$ and the *supporting halfspace* for C of negative type $\check{H}_C := \{x \in \mathbb{R}^n \colon \langle x, u\rangle \leq \check{\sigma}_C(u)\}$ or for C of positive type $\hat{H}_C := \{x \in \mathbb{R}^n \colon \langle x, u\rangle \geq \hat{\sigma}_C(u)\}$. The *support set* (Schneider, 2014, Section 1.7) $F_C(x) := \check{H}_C(x) \cap C = \check{\partial}\check{\sigma}_C(x)$. It also holds that $F_C(x) := \hat{H}_C(x) \cap C = \hat{\partial}\hat{\sigma}_C(x)$.

The *indicator function* of negative type of a set C is the convex function $\check{\imath}_C(x) = 0$ for $x \in C$ and $+\infty$ otherwise. The positive type variant is the concave function $\hat{\imath}_C(x) = 0$ if $x \in C$ and $-\infty$ otherwise.

If $f \colon \mathbb{R}^n \to \mathbb{R}$ is convex its (convex)-conjugate is $f^{\check{*}}(x^*) := \sup_{x\in \operatorname{dom} f} \langle x, x^*\rangle - f(x)$. If g is concave, its concave conjugate is $g^{\hat{*}}(x^*) := -(-g)^{\check{*}}(x^*)$.

3. Gauges and Polars

The theory of gauges (Minkowski functionals) and polars has been traditionally developed for convex sets of negative type; see for example (Hiriart-Urruty and Lemaréchal, 2001; Rockafellar, 1970; Schneider, 2014; Thompson, 1996). The theory of gauges for sets of positive type is less well known; see (Rockafellar, 1967; Pukelsheim, 1983; Barbara and Crouzeix, 1994). Concave gauges have been used in statistics in a manner similar to that which we will use them (Pukelsheim, 1983) and in economics (e.g. Hasenkamp and Schrader (1978); see also the further references at the end of section 5). We will largely follow the development (but not the notation) of Penot and Zălinescu (2000) where proofs of the following results can be found.

A set $A \subset \mathbb{R}^n$ is *star-shaped* if $0 \in A$ and $(0, 1] \cdot A \subset A$. It is *shady* if $[1, \infty) \cdot A \subset A$. The *convex-gauge* of closed star-shaped set $C \subset \mathbb{R}^n$ is defined by

$$\check{\gamma}_C(x) := \inf\{\mu \geq 0 \colon x \in \mu C\}.$$

The *concave-gauge* of $C \in \mathcal{K}$ is defined by

$$\hat{\gamma}_C(x) := \sup\{\mu \geq 0 \colon x \in \mu C\}.$$

The infimum (resp. supremum) in these definitions is attained if they are finite and C is *centrally closed* (for each $x \in \mathbb{R}^n$, $(t_n) \to 1$, if $t_n x \in C$ for each n, then $x \in C$).

The *convex-polar* of $C \in \mathcal{K}^n$ is

$$C^{\varovee} := \{x \in \mathbb{R}^n \colon \langle x, y\rangle \leq 1 \ \forall y \in C\} = \{x \in \mathbb{R}^n \colon \check{\sigma}_C(x) \leq 1\}.$$

If C is of negative type, then $\check{\gamma}_C = \check{\sigma}_{C^{\varovee}}$ and $\check{\gamma}_{C^{\varovee}} = \check{\sigma}_C$. Thus convex-gauges are 1-homogeneous and convex and monotonic with respect to set inclusion: $A \subseteq B \Rightarrow \check{\sigma}_A \leq \check{\sigma}_B$ and since $A \subseteq B \Rightarrow B^{\varovee} \subseteq A^{\varovee}$ we have $A \subseteq B \Rightarrow \check{\gamma}_A \geq \check{\gamma}_B$.

The *concave-polar* (sometimes called the *anti-polar*) of $C \in \mathcal{K}^n$ is

$$C^{\varowedge} := \{x \in \mathbb{R}^n \colon \langle x, y\rangle \geq 1 \ \forall y \in C\} = \{x \in \mathbb{R}^n \colon \hat{\sigma}_C(x) \geq 1\}.$$

If C is of positive type, then $\hat{\gamma}_C = \hat{\sigma}_{C^{\varowedge}}$ and $\hat{\gamma}_{C^{\varowedge}} = \hat{\sigma}_C$. Thus concave-gauges are 1-homogeneous and concave.

Closed convex-gauges are in 1:1 correspondence with closed star-shaped sets. (We rely on (Penot and Zălinescu, 2000, Propositions 2.2 and 2.3(d)) for this to hold using our definition of $\check{\gamma}_C$ which corresponds to their α_C.) The correspondence is given by

$$k(x) = \check{\gamma}_C(x), \quad C = \operatorname{lev}_{\leq 1} k.$$

Closed concave gauges are in 1:1 correspondence with closed convex shady sets not containing 0. The correspondence is given by

$$k(x) = \hat{\gamma}_C(x), \quad C = \operatorname{lev}_{\geq 1} k.$$

If k is a convex-gauge (and thus a non-negative, 1-homogeneous convex function with $k(0) = 0$) then the *convex-polar* of k is defined as

$$k^{\ovee}(y) := \inf\{\mu \geq 0 \colon \langle x, y\rangle \leq \mu k(x) \ \forall x\}.$$

If k is finite everywhere except the origin, one can instead write

$$k^{\ovee}(y) = \sup_{x \neq 0} \frac{\langle x, y\rangle}{k(x)}.$$

The notation $k^{\ovee}$ is justified since under the assumptions above, $(\check{\gamma}_C)^{\ovee} = \check{\gamma}_{C^{\ovee}}$.

If k is a concave-gauge (and thus non-negative, 1-homogeneous concave function with $k(0) = 0$) then the *concave-polar* of k is defined as

$$k^{\owedge}(y) := \sup\{\mu \geq 0 \colon \langle x, y\rangle \leq \mu k(x) \ \forall x\}. \tag{1}$$

If k is finite everywhere except the origin, one can instead write

$$k^{\owedge}(y) = \inf_{x \neq 0} \frac{\langle x, y\rangle}{k(x)}. \tag{2}$$

The notation $k^{\owedge}$ is justified since under the assumptions above, $(\hat{\gamma}_C)^{\owedge} = \hat{\gamma}_{C^{\owedge}}$.

The following dual representation of a concave gauge will be used later. If $\hat{\gamma}$ is a concave-gauge which is finite everywhere and positive except at the origin, then since $\hat{\gamma}$ is 1-homogeneous,

$$\hat{\gamma}^{\ovee}(y) = \sup_{x \neq 0} \frac{\langle x, y\rangle}{\hat{\gamma}(x)} = \sup\{\langle x, y\rangle \colon \hat{\gamma}(x) = 1\}. \tag{3}$$

The recession cones of $C^{\ovee}$ or $C^{\owedge}$ (assuming the latter nonempty) are

$$\begin{aligned} 0^+\left(C^{\ovee}\right) &= C^- := \{y \in \mathbb{R}^n \colon \langle x, y\rangle \leq 0 \ \forall x \in C\} \\ 0^+\left(C^{\owedge}\right) &= C^+ := \{y \in \mathbb{R}^n \colon \langle x, y\rangle \geq 0 \ \forall x \in C\}. \end{aligned}$$

There are also relationships between polars and conjugates. For $C \in \check{\mathfrak{K}}^n$, $(\check{\gamma}_C)^{\check{*}} = \check{\iota}_{C^{\ovee}}$ and $C^{\ovee} = \check{\partial}\check{\gamma}_C(0)$; for $C \in \hat{\mathfrak{K}}^n$, $\hat{\gamma}_C^{\hat{*}} = \hat{\iota}_{C^{\owedge}}$ and $C^{\owedge} = \hat{\partial}\hat{\gamma}_C(0)$.

4. Proper Losses

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. We will consider loss functions as functions that map from the n-simplex to a n-vector — $\ell\colon \Delta^n \to \mathbb{R}^n_+$. The partial functions $\ell_1(p), \ldots, \ell_n(p)$ are called *partial losses*. The *conditional risk* is defined via

$$L\colon \Delta^n \times \Delta^n \ni (p,q) \mapsto L(p,q) = \mathbb{E}_{\mathsf{Y}\sim p}\ell_{\mathsf{Y}}(q) = p' \cdot \ell(q) = \sum_{i=1}^{n} p_i \ell_i(q) \in \mathbb{R}_+.$$

A natural requirement to impose upon ℓ is that it is *proper* (Hendrickson and Buehler, 1971), which means that $L(p,p) \leq L(p,q)$ for all $p, q \in \Delta^n$. (It is *strictly proper* if the inequality is strict when $p \neq q$.) The *conditional Bayes risk* $\underline{L}\colon \Delta^n \ni p \mapsto \inf_{q\in\Delta^n} L(p,q)$ is always concave. If ℓ is proper, $\underline{L}(p) = L(p,p) = p' \cdot \ell(p)$. The full risk $\mathbb{L}(q) = \mathbb{E}_{\mathsf{X}}\mathbb{E}_{\mathsf{Y}|\mathsf{X}}\ell_{\mathsf{Y}}(q(\mathsf{X}))$. One can understand the effect of choice of loss in terms of the conditional perspective (which allows one to ignore the distribution of X which is typically unknown; see (Steinwart and Christmann, 2008; Reid and Williamson, 2011) for a discussion of this conditional perspective. Examples of proper losses include 0-1 loss (not strictly proper), squared loss and log loss (both strictly proper). Instead of losses, one can work with gains, for example $g(p) = -\ell(p)$; see Appendix A for more details on the conversion between losses and gains.

The *superprediction* set

$$S_\ell := \{x \in \mathbb{R}^n \colon \exists y \in \operatorname{dom} \ell,\ x \geq \ell(y)\},$$

where inequality is componentwise. Similarly for gains, the *infraprediction set* is

$$I_g := \{x \in \mathbb{R}^n \colon \exists y \in \operatorname{dom} g,\ x \leq g(y)\}.$$

Every proper loss has a superprediction set which is a convex set S with recession cone $0^+S = \mathbb{R}^n_+$ (Vernet et al., 2012). This motivates the key viewpoint of the present paper: *start with the set of positive type S and derive the loss (and other quantities) from it.*

Suppose then that $S \in \hat{\mathfrak{K}}$ is of positive type and has concave support function $\hat{\sigma}_S$. Let

$$\ell := F_S = \hat{\partial}\hat{\sigma}_S. \tag{4}$$

Note ℓ so defined may be set valued in which case proper means proper for each selection.

Proposition 2 *If S is of positive type, then ℓ defined by (4) is a proper loss.*

Proof Since $\ell_S(x) = F_S(x)$ is the support set of S, it follows that for all $z \in S$, $\langle z, x\rangle \geq \hat{\sigma}_S(x) = \langle x, y\rangle\ \ \forall y \in \hat{F}_S(x)$. Hence $\langle x, \ell_S(x)\rangle \leq \langle x, \ell_S(z)\rangle$ for all $z \in \mathbb{R}^n$ and so ℓ_S is proper. ■

We see that for $\ell = \ell_S$, $\underline{L}_\ell(x) = \hat{\sigma}_S(x) = \langle x, \hat{\partial}\hat{\sigma}_S(x)\rangle$. Since convex, (resp. concave) support functions provide for a bijection between closed convex sets and 1-homogeneous proper convex (resp. concave) functions (Hiriart-Urruty and Lemaréchal, 2001; Schneider, 2014), it is clear that we can either start with ℓ and construct S_ℓ; or start with S and construct ℓ_S.

The relationship between ℓ and $\underline{L}$ (the concave support function of the superprediction set) is usually credited to Savage (1971) and is intimately related to Bregman divergences. Given a convex function on a convex set X, $\phi\colon X \to \mathbb{R}$, the *Bregman divergence* between $x, y \in X$ is defined to be

$$B_\phi(x,y) := \phi(x) - \phi(y) - \langle x - y, \check{\partial}\phi(y)\rangle \tag{5}$$

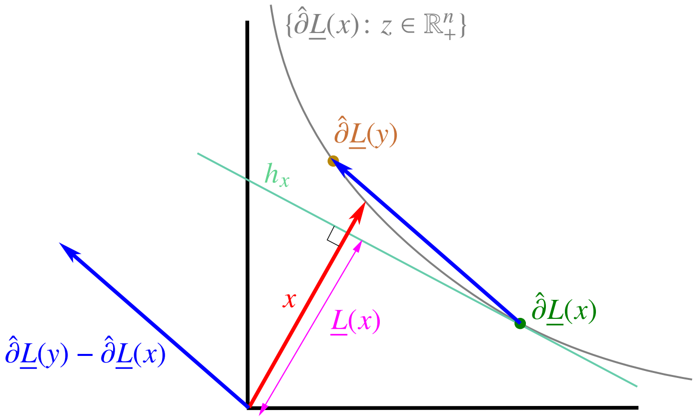
$\{\hat{\partial}\underline{L}(x) \colon z \in \mathbb{R}^n_+\}$
$\hat{\partial}\underline{L}(y)$
h_x
x
$\underline{L}(x)$
$\hat{\partial}\underline{L}(x)$
$\hat{\partial}\underline{L}(y) - \hat{\partial}\underline{L}(x)$

Figure 2: Geometrical intepretation of Bregman divergence $\langle x, \hat{\partial}\underline{L}(y) - \hat{\partial}\underline{L}(x)\rangle$ when the concave function $\underline{L}$ is 1-homogeneous and thus a concave support function of some convex set of positive type (whose lower left boundary is shown in grey).

It is known that the *regret* $L(p,q) - \underline{L}(p)$ is a Bregman divergence with $\phi = -\underline{L}$. Using (4) we have

$$L_S(x,y) = \langle x, \hat{\partial}\hat{\sigma}_S(y)\rangle$$

and thus the general form of the Bregman divergence simplifies:

$$\begin{aligned} B_{-\underline{L}}(x,y) &= -\underline{L}(x) + \underline{L}(y) + \langle x - y, \hat{\partial}\underline{L}(y)\rangle \\ &= \langle x, \hat{\partial}\underline{L}(y) - \hat{\partial}\underline{L}(x)\rangle, \end{aligned} \tag{6}$$

where we have used the fact that $\langle y, \hat{\partial}\underline{L}(y)\rangle = \underline{L}(y)$ since $\underline{L} = \hat{\sigma}_S$ and $\hat{\partial}\hat{\sigma}_S(y) = F_S(y)$.

The simpler form (6) provides a simpler geometrical interpretation of the Bregman divergence as the inner product of the vectors x and $(\hat{\partial}\underline{L}(y) - \hat{\partial}\underline{L}(x))$; see Figure 2. Obviously as $y \to x$, $(\hat{\partial}\underline{L}(y) - \hat{\partial}\underline{L}(x))$ becomes orthogonal to x and thus $B_{-\underline{L}}(x,y)$ approaches 0.

Observe that since $\hat{\sigma}_{\mathrm{co}\,S} = \hat{\sigma}_S$, there is no additional flexibility obtained in starting with non-convex S; although it is sometimes convenient to *parameterise* proper losses in this manner. For example if $S_{0-1} := \{e_1, \ldots, e_n\}$ then the zero-one loss $\ell_{0-1} = \hat{\partial}\hat{\sigma}_{S_{0-1}}$.

The construction of proper losses from a set S allows one to translate a range of existing results from the geometry of convex sets into the terminology of proper losses. We provide some examples below, some of which we will use later.

Since support functions are additive under Minkowski addition (Schneider, 2014) and subdifferentials of the sums of convex functions are the sums of the subdifferentials (Hiriart-Urruty and Lemaréchal, 2001), we have

$$S = S_1 + S_2 \Rightarrow \hat{\sigma}_S = \hat{\sigma}_{S_1} + \hat{\sigma}_{S_2} \text{ and } F_S = F_{S_1} + F_{S_2}.$$

Thus $\ell_{S_1+S_2} = \ell_{S_1} + \ell_{S_2}$. A special case of this is where $S_2 = \{s\}$ whence $\hat{\sigma}_{S_2}(x) = \inf_{y\in S_2}\langle x, y\rangle = \langle x, s\rangle$. Note that $S_1 + \{s\} = \{t + s \colon t \in S_1\}$ is the translaton of S_1 by s.

Since $\underline{L}$ is a support function and thus 1-homogeneous, ℓ_S is (via Proposition 1) 0-homogeneous. Analogously we have, for $\alpha \geq 0$, $\underline{L}_{\alpha S} = \alpha \underline{L}_S$, so $\ell_{\alpha S} = \alpha \ell_S$. If $S_1 \subseteq S_2$, then $\underline{L}_{S_1} \geq \underline{L}_{S_2}$.

We can also write

$$S_\ell = \ell(\mathbb{R}^n_+) + \mathbb{R}^n_+ \ \text{ and } \ I_g = g(\mathbb{R}^n_+) + \mathbb{R}^n_-.$$

Since $\hat{\iota}_S = (\hat{\sigma}_S)^{\hat{*}}$ we can also write $S_\ell = \operatorname{lev}_{\leq 0} \underline{L}^{\hat{*}}$, although we do not make use of the latter in the present paper.

We also have (Schneider, 2014, Corollary 1.7.3) that $\underline{L}_S$ is differentiable at $x \neq 0$ if and only if $F_S(x)$ contains only one element z. In this case $z = \operatorname{grad} \underline{L}(x)$.

There is an elegant bridge between metrics on superprediction (or infraprediction) sets and their corresponding support functions. Let

$$d_H(S_1, S_2) := \max\{ \sup_{x_1 \in S_1} \inf_{x_2 \in S_2} \|x_1 - x_2\|_2, \sup_{x_2 \in S_2} \inf_{x_1 \in S_1} \|x_1 - x_2\|_2 \}$$

denote the *Hausdorff distance* between two sets S_1 and S_2. Then translating the result of (Hiriart-Urruty and Lemaréchal, 2001, page 155) to sets of positive type, if S_1 and S_2 are super prediction sets $d_H(S_1, S_2) = \max_{\|x\|_2 \leq 1} |\hat{\sigma}_{S_1}(x) - \hat{\sigma}_{S_2}(x)|$. This allows translation between convergence of a series of superprediction sets to convergence of the corresponding concave support functions; see (Hiriart-Urruty and Lemaréchal, 2001, page 156).

5. Polars and Inverse Losses

Suppose $\phi\colon \mathcal{X} \rightrightarrows \mathcal{Y}$ is a set-valued map (Aubin and Frankowska, 1990). Its *inverse* $\phi^{-1}\colon \mathcal{Y} \rightrightarrows \mathcal{X}$ is

$$\phi^{-1}(y) := \{x \in \mathcal{X} \colon y \in \phi(x)\}. \tag{7}$$

A loss $\ell\colon \mathbb{R}^n_+ \to \mathbb{R}^n_+$ maps a prediction v to a loss vector x. Given the loss vector x, how can one recover v? This problem arises naturally in finding a "substitution function" for Vovk's Aggregating Algorithm (Vovk, 2001). Apparently one seeks an "inverse loss" ℓ^{-1} such that $\ell^{-1}(x)$"="v. If the loss $\ell = \ell_S = \hat{\partial}\hat{\sigma}_S$, for some $S \in \hat{\mathcal{K}}^n$, then it is 0-homogeneous. Thus, an arbitrary $x \in \mathbb{R}^n_+$ is unlikely to be equal to $\ell(v)$ for some v. However, by exploiting the 0-homogeneity of proper losses when defined in terms of subdifferentials of support functions of convex sets of positive type, we will see there is a very natural and geometrically satisfying way to "invert" such as loss.

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 3.2.7) regarding subdifferentials of Legendre-Fenchel conjugates: $x \in \check{\partial}\phi(y) \Leftrightarrow y \in \check{\partial}\phi^*(x)$. We express the result for the concave case (sets of positive type) because that is what we need for losses; an analogous result holds for convex gauges, sets of negative type and subdifferentials.

Proposition 3 *Suppose $S \in \hat{\mathcal{K}}$. For all $s, d \in \mathbb{R}^n$,*

$$\frac{d}{\hat{\gamma}_S(d)} \in \hat{\partial}\hat{\gamma}_{S^{\circledcirc}}(s) \quad \Leftrightarrow \quad \frac{s}{\hat{\gamma}_{S^{\circledcirc}}(s)} \in \hat{\partial}\hat{\gamma}_S(d) \quad \Leftrightarrow \quad \hat{\gamma}_S(d)\,\hat{\gamma}_{S^{\circledcirc}}(s) = \langle s, d \rangle. \tag{8}$$

Barbara and Crouzeix (1994) provide a sketch of a proof. However since it is central to what follows we present a complete proof in Appendix G.

The above theorem has the following consequence. Since $\hat{\gamma}_S(d) = \hat{\sigma}_{S^\circledast}(d)$ and $\hat{\gamma}_{S^\circledast}(s) = \hat{\sigma}_S(s)$, (8) implies

$$\frac{d}{\hat{\gamma}_S(d)} \in \hat{\partial}\sigma_S(s) = \ell_S(s) \;\Leftrightarrow\; \frac{s}{\hat{\gamma}_{S^\circledast}(s)} \in \hat{\partial}\sigma_{S^\circledast}(d) = \ell_{S^\circledast}(d) \;\Leftrightarrow\; \langle s, d\rangle = \hat{\gamma}_S(d)\,\hat{\gamma}_{S^\circledast}(s). \tag{9}$$

Since ℓ_S is 0-homogeneous, the inverse loss, in the sense of (7), satisfies $y \in \ell_S^{-1}(x)$ if and only if $x \in \ell_S(y)$. From (9) we see that given $d \in \mathbb{R}_+^n$, $s = \ell_{S^\circledast}(d)$ means that for some constant $c > 0$, $\frac{d}{c} \in \ell_S(s)$. (Although it does not matter for what follows, the value of c can be determined from $\hat{\gamma}_{S^\circledast}(s) = \sup\{\mu \geq 0\colon\ s \in \mu S\}$ and hence c is such that $\langle s, \frac{d}{c}\rangle = \hat{\gamma}_S(d)\,\hat{\gamma}_{S^\circledast}(s)$ and so $c = \frac{\hat{\gamma}_S(d)\,\hat{\gamma}_{S^\circledast}(s)}{\langle s,d\rangle}$). The constant c does not matter because ℓ_S is 0-homogeneous. Hence $\ell_S(\alpha\ell_{S^\circledast}(d)) = \ell(\ell_{S^\circledast}(d))$ for all $\alpha > 0$. It is important to realise that the inverse is only up to a positive scaling (since the losses are 0-homogeneous). Thus while one is *not* guaranteed that $d = \ell_S^{-1}(\ell_S(d))$ for all $d \in \mathbb{R}_+^n$, one *is* guaranteed that for all $d \in \mathbb{R}_+^n$,

$$\ell_S(\ell_S^{-1}(\ell_S(d))) = \ell_S(d).$$

Thus the "inverse loss" can be seen to be a special case of the Drazin inverse or pseudo-inverse (Drazin, 1958), which is an abstraction of the notion of pseudo-inverse in linear algebra. The above argument is illustrated in Figure 3. We have thus shown:

Corollary 4 *Given a set S of positive type and hence a proper loss ℓ_S, the inverse loss $\ell_S^{-1} = \ell_{S^\circledast}$.*

There are two ways the inverse loss can be computed — one can either compute the polar of S, or the gauge of S since $\ell_S^{-1} = \hat{\partial}\hat{\sigma}_{S^\circledast} = \hat{\partial}\hat{\gamma}_S$. We illustrate these concepts explicitly with the ℓ_p family in Appendix D.

The inverse loss provides a substitution function for the Aggregating Algorithm (Vovk, 2001, 1995, 1990). The substitution function needs to map an arbitrary superprediction $d \in S_\ell$ to a prediction s such that $x = \ell(s)$ dominates d in the sense that $x \leq d$ (where the inequality is meant pointwise). Determination of 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).

The notion of the inverse loss and its relationship to the concave polar of the superprediction set provides conceptual clarity. Furthermore, at least in some cases one can determine the inverse in explicit form: see equation 21 in Appendix D, as well as the other examples in Appendices E and F.

Figure 3 should be compared with that in (Shephard, 1953, page 23) which was the inspiration for this argument. This has become known as Shephard's duality theorem in the economics literature (Shephard, 1953; Jacobsen, 1972; McFadden, 1978; Hanoch, 1978; Cornes, 1992; Färe and Primont, 1994, 1995; Penot, 2005; Zălinescu, 2013) and appears in standard microeconomics texts (Varian, 1978).

6. Binary operations on superprediction sets and losses

Given superprediction sets (sets of positive type) as the starting point, it is clear that any binary operations on sets of positive type that return a set of positive type will have corresponding operations on the associated proper losses. Thus it is of interest to develop as rich a family of such binary operations on sets of positive type.

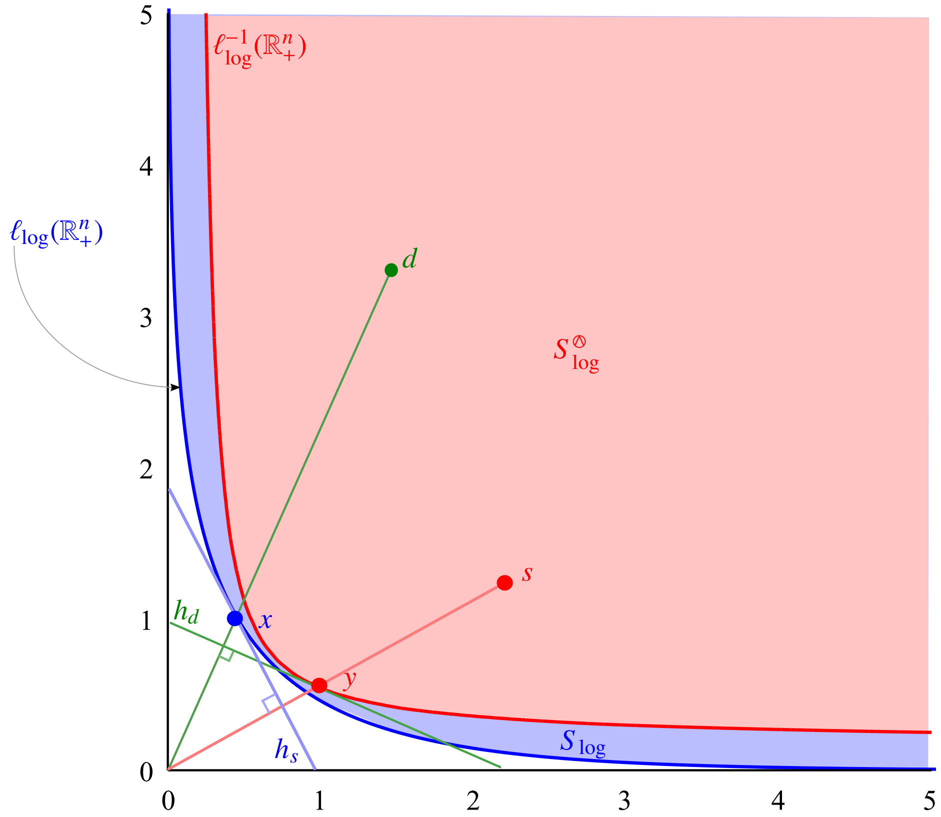

Figure 3: Illustration of polar / inverse loss using $\ell = \ell_{\log}$ with $n = 2$. Shown are the loss curves $\ell_{\log}(\mathbb{R}^n_+)$ and the corresponding superprediction set $S_{\log}$ which extends infinitely "north-east" (both in blue), and $\ell^{-1}_{\log}(\mathbb{R}^n_+)$ and corresponding superprediction set $S^{\oslash}_{\log}$ (the concave polar of $S_{\log}$), both in red. Given a point $d \in \mathbb{R}^n_+$ it is desired to find $s \in \mathbb{R}^n$ such that $\ell(s) = x$, where x is the point of intersection of the line segment $[0, d]$ with the curve $\ell(\mathbb{R}^n_+)$. Evaluating $\ell^{-1}(d)$ gives the point y; observe the hyperplane h_d with normal vector d supports $\ell^{-1}_{\log}(\mathbb{R}^n_+)$ at the point y because of properness. Any positive scaling $s = \alpha y$, for $\alpha > 0$, would also suffice. It can be seen that h_s (with normal vector s) supports $\ell_{\log}(\mathbb{R}^n_+)$ at x (again because of properness). As argued in the text, by construction we are guaranteed that $\ell_{\log}(s) = x$. This justifies calling the polar loss $\hat{\partial}\hat{\sigma}_{S^{\oslash}_{\log}}$ the "inverse loss" $\ell^{-1}_{\log}$. The curves were drawn by plotting $\ell_{\log}(x) = \hat{\partial}\hat{\sigma}_{S_{\log}}(x)$ for $x \in \Delta^2$ (exploiting zero-homogeneity) and the level set $\{z \in \mathbb{R}^2_+ : \hat{\sigma}_{S_{\log}}(x) = 1\}$, relying on the definition of the concave polar of the set $S_{\log}$.

We will now consider a class of binary operations on superprediction sets that seem particularly useful for this purpose. They generalise the simple operations of Minkowski sum or intersection. There are explicit formulas for the support functions of the results of these operations in terms of the support functions of the sets. Furthermore there are expressions for the polars of binary operations of sets in terms of the polars of the sets. Since support functions and polars are central to the relationship with proper losses, these operations seem very appropriate for developing an algebra of losses.

We generalise the results of Seeger (1990). Rather than a family parameterised by $p \in [1, \infty]$ corresponding to the classical l_p norms), we work with arbitrary norms (gauges) on $\mathbb{R}^2$ which we parameterise by the convex set C which corresponds to the unit ball with respect to the corresponding norm (every convex gauge induces a norm). This has the additional benefit of simplifying some of the proofs. Seeger's results are for convex sets of negative type. In order to make the present results directly comparable to his we stick with this convention. They can be directly applied to gains. We present the results for subsets of $\mathbb{R}^n$ since that is all we need, but they would actually hold (like Seeger's results) for subsets of arbitrary locally convex topological vector spaces.

Let $A, B \subset \mathbb{R}^n$ be sets of negative type (closed, convex, and containing the origin). Let $C \subset \mathbb{R}^2_+$ also be of negative type. By analogy with the operation of epimultiplication of functions, for a set S and $\alpha \geq 0$ define

$$\alpha \star S := \begin{cases} \alpha S, & \alpha > 0 \\ 0^+ S, & \alpha = 0 \end{cases} . \tag{10}$$

Following Seeger (1990) we define

$$A \mathbin{\check{\oplus}_C} B := \bigcup_{\lambda \in C^{\circledvee} \cap \mathbb{R}^2_+} \lambda_1 A + \lambda_2 B \tag{11}$$

$$A \mathbin{\check{\bar{\oplus}}_C} B := \bigcup_{\lambda \in C^{\circledvee} \cap \mathbb{R}^2_+} \lambda_1 \star A + \lambda_2 \star B \tag{12}$$

$$A \mathbin{\check{\square}_C} B := \bigcup_{\lambda \in C^{\circledvee} \cap \mathbb{R}^2_+} \lambda_1 A \cap \lambda_2 B \tag{13}$$

$$A \mathbin{\check{\bar{\square}}_C} B := \bigcup_{\lambda \in C^{\circledvee} \cap \mathbb{R}^2_+} \lambda_1 \star A \cap \lambda_2 \star B. \tag{14}$$

Seeger (1990) studied the special cases where $C = C_p = \{x \colon \|x\|_p \leq 1\}$ and $p \in [1, \infty]$. We will write $\check{\oplus}_{C_p}$ as $\check{\oplus}_p$ and $\check{\square}_{C_p}$ as $\check{\square}_p$. Special cases pointed out by Seeger include

$$\begin{array}{lcll} A \mathbin{\check{\oplus}_1} B & = & A + B & \text{Minkowski sum} \\ A \mathbin{\check{\oplus}_\infty} B & = & \operatorname{co}(A \cup B) & \text{convex hull of union} \\ A \mathbin{\check{\square}_1} B & = & A \cap B & \text{intersection} \\ A \mathbin{\check{\square}_\infty} B & = & A \sharp B & \text{inverse sum.} \end{array}$$

The history of the inverse sum operation is summarised by Seeger (1990). He expresses the operations somewhat differently, but they are seen to be the same as follows (we present the argument for $\check{\oplus}_C$; the same argument holds with the obvious variation for $\check{\square}_C$). We can write $C = \operatorname{lev}_{\leq 1} \check{\gamma}_C$, thus

$$A \mathbin{\check{\oplus}_C} B = \bigcup \{\lambda_1 A + \lambda_2 B \colon \lambda \geq 0, \check{\gamma}_{C^{\circledvee}}(\lambda) \leq 1\} = \bigcup \{\lambda_1 A + \lambda_2 B \colon \lambda \geq 0, \check{\gamma}_{C^{\circledvee}}(\lambda) = 1\},$$

which corresponds to Seeger's definition when $C = C_p$, recalling the standard result that $B_p^{\circledvee} = B_q$ where $\frac{1}{p} + \frac{1}{q} = 1$. The operations $\check{\oplus}_C$ and $\check{\square}_C$ are of interest because of the following closure result. Seeger (1990, Theorem 2.3) proves that if A and B are convex and $S \subset \mathbb{R}^2$ is convex then so is $\bigcup_{\lambda \in S} (\lambda_1 A + \lambda_2 B)$. If $0 \in A$ and $0 \in B$ then $0 \in A \cap B$. Thus it is clear that $0 \in A \mathbin{\check{\oplus}_C} B$ and $0 \in A \mathbin{\check{\square}_C} B$. Thus we have:

Proposition 5 *If $A, B \subset \mathbb{R}^n$ are sets of negative type and $C \subset \mathbb{R}^2$ is of negative type then $A \mathbin{\check{\oplus}_C} B$ and $A \mathbin{\check{\square}_C} B$ are also of negative type.*

The operations $A\ \check{\Box}_C\ B$ and $A\ \check{\bar{\Box}}_C\ B$ are identical when A and B are bounded since in that case $0 \star A = 0 \star B = \{0\}$. However for unbounded sets they differ. We also need the following generalisation of (Seeger, 1990, Proposition 4.1). The proof is the same as Seeger's and is omitted.

Proposition 6 *if $C \subset \mathbb{R}^2$ is of negative type and $A, B \subset \mathbb{R}^n$ are of negative type then*

$$A\ \check{\Box}_C\ B \subset A\ \check{\bar{\Box}}_C\ B \subset \overline{A\ \check{\Box}_C\ B}.$$

We now consider some binary operations on functions. The inherent notation overloading in the following is justified later in Theorem 9. If $x \in \mathbb{R}^n$, then x' denotes its transpose.

Definition 7 *Suppose $C \subset \mathbb{R}^2$ is of negative type and suppose $f, g \colon \mathbb{R}^n \to [0, \infty]$. The* direct and inverse sum of type C of f and g *are respectively*

$$(f\ \check{\oplus}_C\ g)(x^*) := \check{\gamma}_C((f(x^*), g(x^*))') \tag{15}$$

$$(f\ \check{\Box}_C\ g)(x^*) := \inf_{x_1^*+x_2^*=x^*} \check{\gamma}_C((f(x_1^*), g(x_2^*))'). \tag{16}$$

As with the set operations, for $p \in [1, \infty]$ we abbreviate $\check{\oplus}_{C_p}$ by $\check{\oplus}_p$ and $\check{\Box}_{C_p}$ by $\check{\Box}_p$. Special cases of these operations are

$$\begin{array}{lcll}
(f\ \check{\oplus}_1\ g)(x^*) & = & f(x^*) + g(x^*) & \text{sum} \\
(f\ \check{\oplus}_\infty\ g)(x^*) & = & f(x^*) \vee g(x^*) & \text{maximum} \\
(f\ \check{\Box}_1\ g)(x^*) & = & \inf\limits_{x_1^*+x_2^*=x^*} (f(x_1^*) + g(x_2^*)) & \text{infimal convolution} \\
(f\ \check{\Box}_\infty\ g)(x^*) & = & \inf\limits_{x_1^*+x_2^*=x^*} (f(x_1^*) \vee g(x_2^*)) & \text{inf-max convolution.}
\end{array}$$

The first three are standard; the last corresponds to the addition of the level sets of the two functions f and g, and has been studied in more detail by Seeger and Volle (1995).

These operations on functions preserve convexity as is shown in the following generalisation of (Seeger, 1990, Theorem 3.2); the proof of the second part is essentially the same as in (Seeger, 1990) modulo minor details. The proof is in Appendix G.

Proposition 8 *Suppose $C \subset \mathbb{R}^2$ is of negative type and $f, g \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ are convex. Then $f\ \check{\oplus}_C\ g$ and $f\ \check{\Box}_C\ g$ are also convex.*

The justification for the overloading of notation $\check{\oplus}_C$ and $\check{\Box}_C$ to refer to operations on both sets and functions is provided by the following proposition. It is a generalisation of (Seeger, 1990, Theorems 5.1 and 5.2). It explains the relevance of these operations on infraprediction sets, as their effect can be equivalently calculated using the corresponding operation on the support functions. The proof is in Appendix G.

Proposition 9 *Suppose $C \subset \mathbb{R}^2$ is of negative type and $A, B \subset \mathbb{R}^n$ are of negative type. Then*

$$\check{\sigma}_{A \check{\oplus}_C B} = \check{\sigma}_A\ \check{\oplus}_C\ \check{\sigma}_B \tag{17}$$

$$\check{\sigma}_{A \check{\Box}_C B} = \overline{\check{\sigma}_A\ \check{\Box}_C\ \check{\sigma}_B}. \tag{18}$$

We now consider the polars of the direct and inverse sum operations. Since the polar of a infraprediction set corresponds to the inverse gain, the following result shows how these inverses behave under the generalised sum operations. The form of the right-hand side (with the appearance of $C^{\circledvee}$) is appealing. The proof is in Appendix G.

Proposition 10 *Suppose $C \subset \mathbb{R}^2_+$ is of positive type and $A, B \subset \mathbb{R}^n$ are of positive type. Then*

$$(A \mathbin{\check{\oplus}_C} B)^{\circledvee} = \overline{A^{\circledvee} \mathbin{\check{\square}_{C^{\circledvee}}} B^{\circledvee}}. \tag{19}$$

7. Conclusions

We showed how one could start with a convex body and hence derive the theory of proper losses (or gains). As well as the aesthetic attraction, there are some concrete advantages of doing so. It shows that the natural way to define the Bayes risks is as 1-homogeneous functions which means that the associated losses (or gains) are 0-homogeneous. The theory of polar duality (long used in production economics) then provides a general and elegant way to define an inverse loss which can be used as a universal substitution function in the aggregating algorithm. Furthermore, operations on convex bodies have corresponding operations on Bayes risks. We have spelt out some of these connections explicitly. Appendices B, D, E, F illustrate the general theory with specific examples: cost-sensitive misclassification loss, l_p losses (the family is closed under inverses), Brier loss, and the self-inverse "Boosting loss" which is induced by a Cobb-Douglas style concave support function.

The viewpoint of losses as 0-homogenous functions allows us to think of losses differently: losses are simply "distorted probabilities". The "inverse loss" can undo the distortion. This notion of distorted probabilities seems different to that used in the insurance and risk literature where it is the cumulative distribution (or more particularly the survival function) that is distorted (Reesor and McLeish, 2002; Pflug and Römisch, 2007; Furman and Zitikis, 2009; Chateauneuf, 1996).

The (0- and 1-) positive homogeneity of ℓ and $\underline{L}$ means one can equally define ℓ on the probability simplex (as is traditionally done). Interestingly, one can achieve the same results if instead ℓ is defined on the Euclidean unit ball, which corresponds to working with the square roots of probabilities. Such a viewpoint allows connections to be drawn to the theory of kernels on probability distributions (Hein and Bousquet, 2004).

The framework can also be used to derive surrogate regret bounds (for the general multiclass case) using the theory of decomposition of convex bodies (Schneider, 2014, Section 3.2) and the bridge to f-divergences (Garcia-Garcia and Williamson, 2012) can be recovered by parametrising f-divergences also in terms of convex sets. These further results will appear in an extended version of the present paper.

The viewpoint of the paper also shows the strong connection between losses (or more precisely gains) and norms. A gain corresponds to a set of negative type. If that set is symmetric about the origin then it is the unit ball of a norm. The symmetry is not relevant when the argument is always in the positive orthant. It is an interesting future direction to see what can be exploited by this connection with norms, and the theory of asymmetric metric spaces (Zaustinsky, 1959).

Acknowledgements

This work was supported by the ARC and NICTA, which is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. Thanks to the Xinhua Zhang and the referees for helpful comments.

References

Jean-Pierre Aubin and Hélène Frankowska. *Set-Valued Analysis*. Birkhäuser, 1990.

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

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.

Alain Chateauneuf. Decomposable capacities, distorted probabilities and concave capacities. *Mathematical Social Sciences*, 31:19–37, 1996.

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.

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

A. Philip Dawid and Steffen L. Lauritzen. The geometry of decision theory. In *Proceedings of the 2nd International Symposium on Information Geometry and its Applications*, pages 22–28, 2006.

Michael P. Drazin. Pseudo-inverses in associative rings and semigroups. *The American Mathematical Monthly*, 65(7):506–514, August 1958.

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.

Edward Furman and Ričardas Zitikis. Weighted pricing functionals with applications to insurance: An overview. *North American Actuarial Journal*, 13(4):483–496, 2009.

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.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

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.

David J. Hand. Deconstructing Statistical Questions. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(3):317–356, 1994.

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.

Matthias Hein and Olivier Bousquet. Hilbertian metrics and positive definite kernels on probability measures. Technical Report 126, Max-Planck-Institut für biologische Kybernetik, July 2004.

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

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

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

Patrick Hummel and R. Preston McAfee. Loss functions for predicted click-through rates in auctions for online advertising. Preprint, Google Inc., October 2013.

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

John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42:654–655, 1956.

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.

Edgar C. Merkle and Mark Steyvers. Choosing a strictly proper scoring rule. Preprint, Department of Psychological Sciences, University of Missouri, Columbia, Missouri, June 2013.

Walter Meyer and David C. Kay. A convexity structure admits but one real linearization of dimension greater than one. *Journal of the London Mathematical Society*, 7:124–130, 1973. Series 2.

Allan H. Murphy and Carl-Axel S. Staël von Holstein. A geometrical framework for the ranked probability score. *Monthly Weather Review*, 103(1):16–20, 1975.

Jean-Paul Penot. The bearing of duality on microeconomics. In *Advances in Mathematical Economics*, pages 113–139. Springer, 2005.

Jean-Paul Penot and Constantin Zălinescu. Harmonic sum and duality. *Journal of Convex Analysis*, 7(1):95–113, 2000.

Georg Ch. Pflug and Werner Römisch. *Modeling, Measuring and Managing Risk.* World Scientific, 2007.

Friedrich Pukelsheim. On information functions and their polars. *Journal of Optimization Theory and Applications*, 41(4):533–546, 1983.

R. Mark Reesor and Don L. McLeish. Risk, entropy, and the transformation of distributions. Bank of Canada Working Paper 2002-11, April 2002.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, March 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*. 1967.

Michael Edward Ruberry. *Prediction Markets: Theory and Applications*. PhD thesis, School of Engineering and Applied Sciences, Harvard University, November 2013.

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 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.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8:171–176, 1958.

Carl-Axel S. Staël von Holstein. *Assessment and evaluation of subjective probability distributions*. PhD thesis, The Economic Research Institute, Stockholm School of Economics, September 1970.

Carl-Axel S. Staël von Holstein and Allan H. Murphy. The family of quadratic scoring rules. *Monthly Weather Review*, 106(7):917–924, July 1978.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.

Anthony C. Thompson. *Minkowski Geometry*. Cambridge University Press, 1996.

Hal R. Varian. *Microeconomic Analysis*. W.W. Norton and Company, 1978.

Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. Submitted to *Journal of Machine Learning Research*, 42 pages., June 2012. URL http://users.cecs.anu.edu.au/~williams/papers/P189.pdf.

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.

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.

Eugene M. Zaustinsky. Spaces with non-symmetric distance. *Memoirs of the American Mathematical Society*, 34:1–91, 1959.

Fedor Zhdanov. *Theory and Applications of Competitive Prediction*. PhD thesis, Department of Computer Science, Royal Holloway, University of London, 2011.

Constantin Zălinescu. Relations between the convexity of a set and the differentiability of its support function. arXiv:1301.0810v1, January 2013.

Appendix A. Losses and Gains

We will now present some example proper losses and gains, which will serve to illustrate the perspective presented in the body of the paper.

Rather than seperately working out losses and gains, one can easily convert from one to the other (not uniquely). In order to convert, one needs to map sets of positive type to negative type (and vice versa). The simplest transformation is $S = -I = \{-x\colon x \in I\}$. However in order to respect the common conventions for losses (e.g. non-negativity), and more technically to ensure polars continue to exist (see section 5) we will require the conversion be done in a manner which ensures that $0 \in I$ and $0 \notin S$.

Given an infraprediction set $I \subseteq \mathbb{R}^n$ with associated gain $g_I = \check{\partial}\check{\sigma}_I$ and some $c \in \mathbb{R}^n$ observe that by setting $S = c - I$ one can derive the loss ℓ_S as follows:

$$\begin{aligned}\ell_S(x) = \hat{\partial}\hat{\sigma}_S(x) &= \{x^*\colon \hat{\sigma}_S(x) + \langle x^*, y-x\rangle \geq \hat{\sigma}_S(y),\ \forall y \in \mathbb{R}^n\}\\ &= \{x^*\colon \langle x, c\rangle - \check{\sigma}_I(x) + \langle x^*, y-x\rangle \geq \langle y, c\rangle - \check{\sigma}_I(y),\ \forall y \in \mathbb{R}^n\}\\ &= \{x^*\colon\ -\check{\sigma}_I(x) + \langle x^* - c, y-x\rangle \geq -\check{\sigma}_I(y),\ \forall y \in \mathbb{R}^n\}\end{aligned}$$

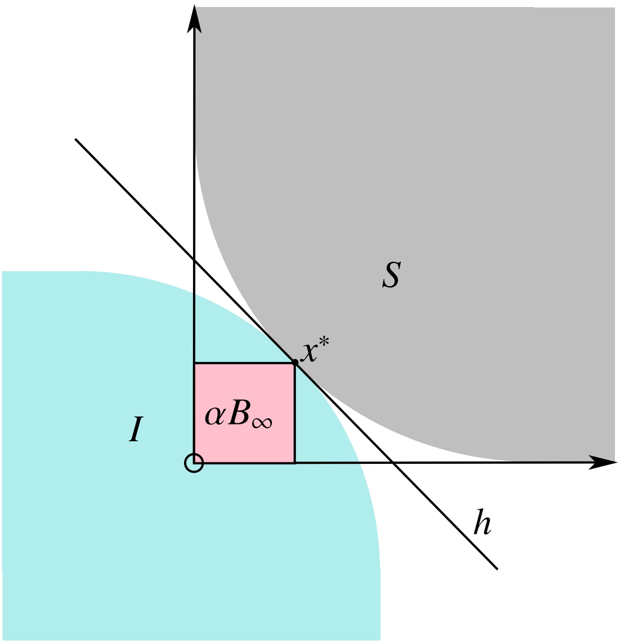

Figure 4: Canonical translation from losses to gains; superprediction sets to infraprediction sets.

Let $z^* = x^* - c$ so $z^* + c = x^*$

$$\begin{aligned}
&= \{z^* + c \colon -\check{\sigma}_I(x) + \langle z^*, y - x\rangle \geq -\check{\sigma}_I(y),\ \forall y \in \mathbb{R}^n\} \\
&= c + \{z^* \colon \check{\sigma}_I(x) + \langle -z^*, y - x\rangle \leq \check{\sigma}_I(y),\ \forall y \in \mathbb{R}^n\} \\
&= c - \{z^* \colon \check{\sigma}_I(x) + \langle z^*, y - x\rangle \leq \check{\sigma}_I(y),\ \forall y \in \mathbb{R}^n\} \\
&= c - \check{\partial}\check{\sigma}_I(x)
\end{aligned}$$

and thus

$$\ell_S(x) = c - g_I(x) \tag{20}$$

as one would expect.

We now present a canonical choice of c to choose when mapping S to I via $S = c - I$ which illustrates another property of super- (infra-) prediction sets. Since the concave support function $\hat{\sigma}_S$ is indeed concave, it has a maximum when restricted suitably. It is natural to consider $\arg\max_{p \in \Delta^n} \hat{\sigma}_S(p)$. Since S is convex and $\langle x, p\rangle$ is linear, by the minimax theorem we have

$$V := \max_{p \in \Delta^n} \hat{\sigma}_S(p) = \max_{p \in \Delta^n} \min_{x \in S} \langle x, p\rangle = \min_{x \in S} \max_{p \in \Delta^n} \langle x, p\rangle.$$

Since $S \subseteq \mathbb{R}^n_+$, $\arg\max_{p \in \Delta^n} \langle x, p\rangle = e_i$, where $i = \arg\max_j x_j$ and thus $V = \min_{x \in S} \|x\|_\infty$. Thus $p^* = \arg\max_{p \in \Delta^n} \hat{\sigma}_S(p)$ is the (normalised onto Δ^n) normal to the hyperplane that supports S at the point of intersection with the corner of the l_∞ ball — see Figure 4. The same argument holds, *mutatis mutandis*, for infraprediction sets. Hence a natural conversion from S to I is, given $x^* = \ell(p^*)$, $T_{x^*} \colon \mathbb{R}^n \ni x \mapsto 2x^* - x \in \mathbb{R}^n$.

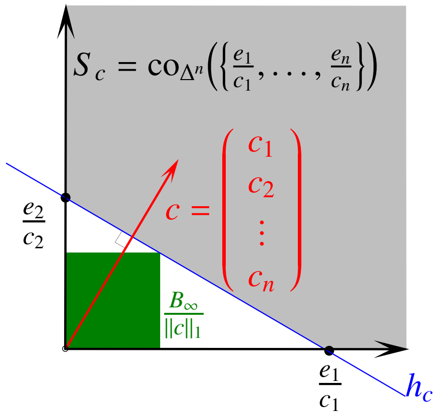
$S_c = \mathrm{co}_{\Delta^n}\left(\left\{\frac{e_1}{c_1}, \ldots, \frac{e_n}{c_n}\right\}\right)$
$c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$
$\frac{e_2}{c_2}$
$\frac{B_\infty}{\|c\|_1}$
$\frac{e_1}{c_1}$
h_c

Figure 5: Superprediction set for cost-sensitive misclassification loss.

An alternative transformation is attractive for fair bounded losses. A proper loss $\ell_S : \mathbb{R}^n_+ \to \mathbb{R}^n_+$ is *fair* if $\underline{L}_S(e_i) = 0$ for all $i \in [n]$ and is *bounded* if $\|\ell_S(z)\|_\infty \leq \infty$ for all $z \in \mathbb{R}^n_+$. In this case set $\alpha := \arg\max_{z\in\mathbb{R}^n_+} \|\ell_S(x)\|_\infty$ and $x^* := \alpha\mathbf{1}$. Then the transformation $T\colon x \mapsto x^* - x$ ensures not only is $0 \in I = T(S)$ but that $T(\ell(\mathbb{R}^n_+)) \subseteq \mathbb{R}^n_+$ and thus I can be considered the unit ball of a norm. Analogous to the development in section 5 for losses, the corresponding Bayes risk $\overline{G}_I = \breve{\sigma}_I$ and gain $g = \breve{\partial}\overline{G}$ can be related to the gauge of the norm. Thus one can see the intimate relationship between gains and norms. Solely because gains are only defined on $\mathbb{R}^n_+$ there is no need for the symmetry constraint. However, to correspond to a norm the gain (or loss) needs to be bounded.

Appendix B. Cost-sensitive misclassification loss

We now consider the cost-sensitive misclassification loss. Let $c \in (0,\infty)^n$ and consider the set $T_c := \{t_1, \ldots, t_n\} := \left\{\frac{e_1}{c_1}, \ldots, \frac{e_n}{c_n}\right\}$. The Δ^n-convex hull (intersection of supporting halfspaces with normal vectors in Δ^n) is

$$S_c := \mathrm{co}_{\Delta^n} T_c = \bigcap_{H \text{ supports } T_c} H = H_c \cap \mathbb{R}^n_+$$

where $H_c := \{x \in \mathbb{R}^n : \langle c, x\rangle - 1 \geq 0\}$. Thus S_c is a convex set of positive type and hence a superprediction set. As can be seen $\hat{\sigma}_{S_c}(x)$ attains its maximum over $x \in \Delta^n$ at c since c is the normal vector to the hyperplane $h_c = \{x \in \mathbb{R}^n : \langle c, x\rangle - 1 = 0\}$. See Figure 5. The corresponding loss can be found by observing that T_c will be supported by a hyperplane with normal vector $z \in \mathbb{R}^n_+$ at t_{i^*} when $i^* \in \arg\min_{i\in[n]}\langle z, t_i\rangle$ and hence

$$\ell_c(z) = \hat{\partial}\hat{\sigma}_{S_c}(z) = \left\{ t_{i^*} : i^* \in \arg\min_{i\in[n]} \langle z, t_i\rangle \right\}.$$

Furthermore if z is such that h_z supports T_c at t_{i^*}, then $\langle z, t_{i^*}\rangle - \hat{\sigma}_{S_c}(z) = 0$ and thus $\hat{\sigma}_{S_c}(z) = \langle z, t_{i^*}\rangle$ and thus $\hat{\sigma}_{S_c}$ is piecewise linear. The maximum occurs at c at which point all components of $z = \ell_c(c)$ are equal (the intersection with the corner of the l_∞ ball). Suppose then that $z = \alpha\mathbf{1}$. We have (noting the earlier definition of h_c) that $\langle x, \alpha\mathbf{1}\rangle - 1 = 0$ and thus $\alpha = \frac{1}{\|c\|_1}$ so $\ell_c(c) = \mathbf{1}/\|c\|_1$.

Appendix C. Shifting the maximum of $\underline{L}$

The previous example suggests the following question. Given an arbitrary superprediction set S with associated proper loss ℓ_S, how might one modify ℓ_S in order that the maximum of $\underline{L}_S$ occurs at a pre-specified point? The motivation for this question is that this is a canonical way of making a given loss asymmetrical in a desired way. If one is given only ℓ or $\underline{L}$ this is not obvious. However the answer is trivial in terms of S. As we have seen the maximum of $\hat{\sigma}_S(x)$ occurs for the value of x normal to the hyperplane that supports S at the point it intersects the corner of a scaled l_∞ ball. Thus suppose that $p^* := \arg\max_{p\in\Delta^n} \hat{\sigma}_S(p)$ and we desire the maximum to occur at $p' \neq p^*$. Then it suffices to translate the set S by $(\ell_S(p') - \ell_S(p^*))$. That is, let $S' = S - (\ell_S(p') - \ell_S(p^*))$. Then $\operatorname{bd} S'$ intersects the corner of the same l_∞ ball and hence $\hat{\sigma}_{S'}$ (and $\ell_{S'}$ has the desired property). By the additivity properties of support functions, we can of course write

$$\hat{\sigma}_{S'}(x) = \hat{\sigma}_S(x) - \langle \ell_S(p') - \ell_S(p^*), x\rangle.$$

and

$$\ell_{S'}(x) = \ell_S(x) - (\ell_S(p') - \ell_S(p^*)).$$

While this argument answers the question, it seems like cheating. One may instead want a more powerful result; given an arbitrary *fair* proper loss ℓ_S with $p^* = \arg\max_{p\in\Delta^n} \hat{\sigma}_S(p)$, transform the loss to a new *fair* proper loss $\ell_{S'}$ with $\arg\max_{p\in\Delta^n} \hat{\sigma}_{S'}(p) = p'$ in a manner that preserves "something" about the shape of S. However this is too much to ask and is in fact impossible using any fixed non-affine mapping T that would work for all superprediction sets. This effectively follows from a result due to Meyer and Kay (1973) that the only maps that map arbitrary convex sets to convex sets are affine.

Appendix D. ℓ_p losses

In general the calculation of polars of superprediction sets, or equivalently the inverse loss may be difficult to achieve in closed form. However, analogous to the case of convex gauges, there is a parametric family which has an attractive self-closure property with respect to taking polars. Following Barbara and Crouzeix (1994), we define $\hat{\gamma}_p \colon \mathbb{R}^n_+ \to \mathbb{R}$ as follows for $p \in [-\infty, 0) \cup (0, 1]$.

$$\hat{\gamma}_p(x) := \begin{cases} \left(\sum_{i=1}^n x_i^p\right)^{\frac{1}{p}}, & x \in \mathbb{R}^n_+, \\ -\infty & \text{otherwise} \end{cases} \qquad p \in (0,1],$$

$$\hat{\gamma}_p(x) := \begin{cases} \left(\sum_{i=1}^n x_i^p\right)^{\frac{1}{p}}, & x \in \operatorname{int} \mathbb{R}^n_+ \\ 0 & x \in \operatorname{bd} \mathbb{R}^n_+ \\ -\infty & \text{otherwise} \end{cases} \qquad p \in (-\infty, 0),$$

$$\hat{\gamma}_{-\infty}(x) := \begin{cases} \bigwedge_{i=1}^n x_i, & x \in \mathbb{R}^n_+ \\ -\infty, & \text{otherwise.} \end{cases}$$

Barbara and Crouzeix (1994) show that for all $p \in [-\infty, 0) \cup (0, 1]$, $\hat{\gamma}_p$ is indeed a concave gauge and furthermore if q satisfies $\frac{1}{p} + \frac{1}{q} = 1$ then

$$\hat{\gamma}_p^{\circledwedge} = \hat{\gamma}_q. \tag{21}$$

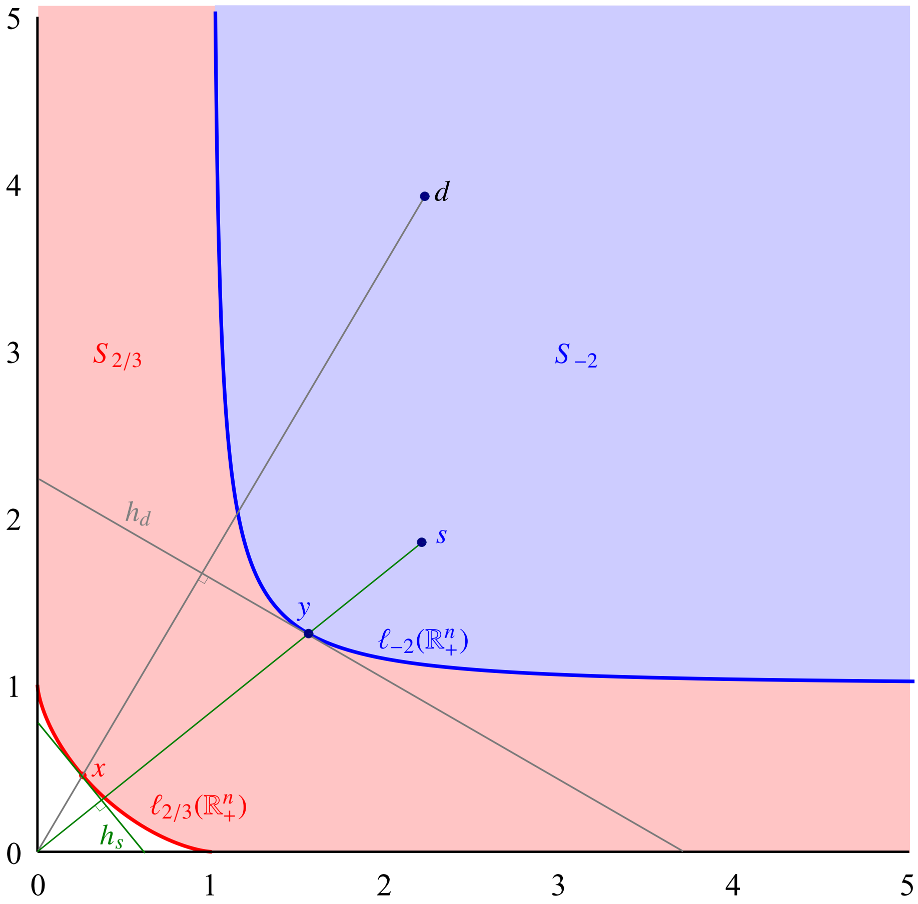

Figure 6: The family of ℓ_p losses is closed under inverse (polars). The figure shows that $\ell_{2/3}^{-1} = \ell_{-2}$.

If $p \in (0,1]$ then $q \in [-\infty, 0)$, and if $p \in [-\infty, 0)$ then $q \in (0,1]$. Thus no gauge $\hat{\gamma}_p$ is self-polar.

The family of concave gauges $\hat{\gamma}_p$ can be used to define a family of proper losses on n outcomes Since $\hat{\sigma}_S = \hat{\gamma}_{S^{\oslash}} = \hat{\gamma}_S^{\oslash}$, and the polar is given via (21), we will determine the loss function

$$\begin{aligned}\ell_p(x) &:= \hat{\partial}\sigma_p(x), \quad x \in \operatorname{int}\mathbb{R}^n_+\\ &= \hat{\partial}\gamma_q(x)\\ &=: (y_1, \ldots, y_n)'.\end{aligned}$$

When $q \neq -\infty$ (so $p \neq 1$), we have

$$y_i = \frac{\partial}{\partial x_i}\left(\sum_{j=1}^n x_j^q\right)^{\frac{1}{q}} = \left(\sum_{j=1}^n x_j^q\right)^{\frac{1}{q}-1} x_i^{q-1} = \frac{x_i^{\frac{1}{p-1}}}{\left(\sum_{j=1}^n x_j^{\frac{p}{p-1}}\right)^{\frac{-1}{p}}}.$$

When $q = -\infty$ (so $p = 1$), we have $\ell_p(x) = \hat{\partial}\left(\bigwedge_{i=1}^n x_i\right)$. But $x \mapsto \bigwedge_i x_i$ is the support function of $\{e_1, \ldots, e_n\}$ or its convex hull $\overline{\mathbb{R}^+_n \setminus \Delta^n}$ and we see ℓ_1 is just 0-1 loss.

The closure under inverse of the family $\{\ell_p \colon p \in [-\infty, 0) \cup (0,1]\}$ is illustrated in Figure 6, and their variation in shape of the corresponding superprediction sets S_p is illustrated in Figure 7.

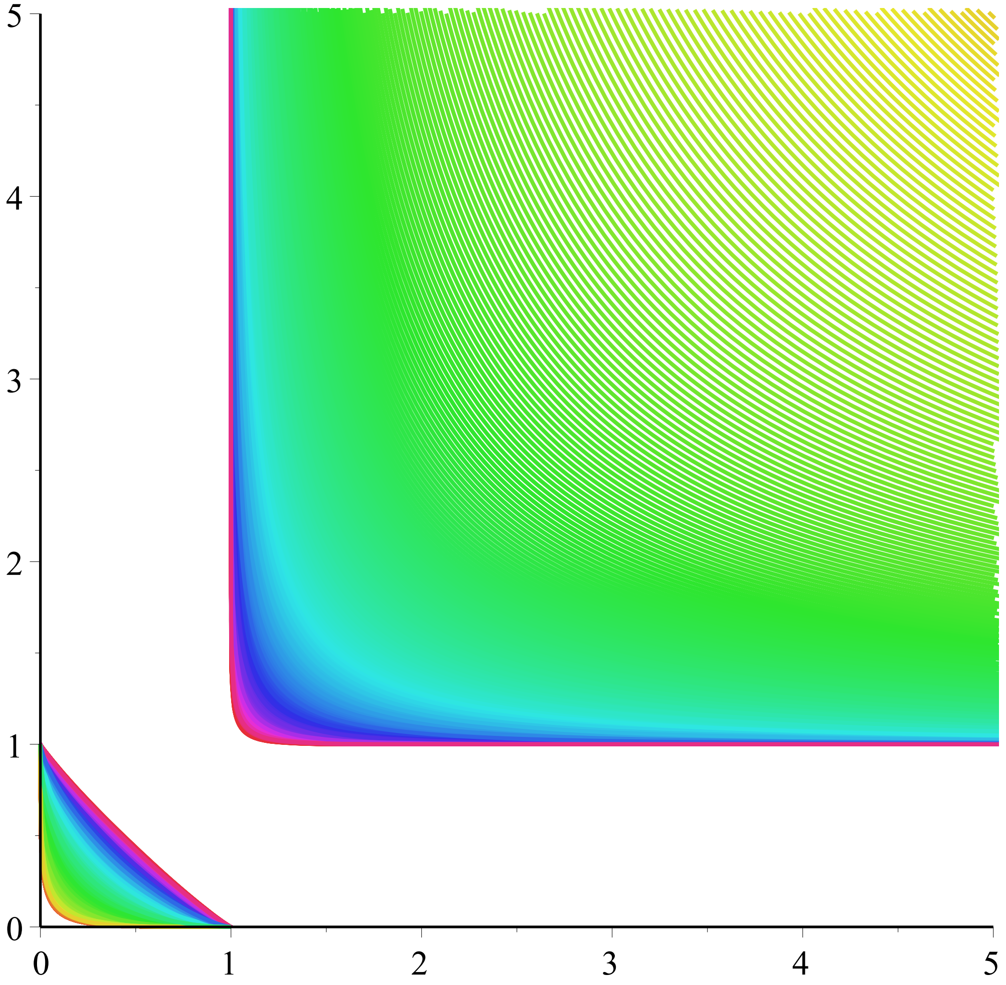

Figure 7: Illustration of the range of ℓ_p losses. The upper part shows $\ell_p(\mathbb{R}^2_+)$ for $p \in [-10, -4/10]$ and the lower part the corresponding polar (inverse) loss $\ell_q(\mathbb{R}^2_+)$, the colors matching and $q \in [4/14, 10/11]$.

Appendix E. Brier Loss

Consider the spherical proper gain. Define the infraprediction set $I_{\text{sph}} = \mathbb{R}^n_+ \cap B^n$, where $B^n = \{x\colon \|x\|_2 \leq 1\}$ is the unit ball for the 2-norm. Obviously, for $x \in \mathbb{R}^n_+$, with $\|x\|_2 = 1$, we have $\check{\sigma}_{I_{\text{sph}}}(x) = 1$ and thus (due to 1-homogeneity), $\check{\sigma}_{I_{\text{sph}}}(x) = \|x\|_2$ for all $x \in \mathbb{R}^n_+$. Hence the gain $g_{\text{sph}}(x) = \check{\partial}\|x\|_2 = \frac{x}{\|x\|_2}$, which as expected is 0-homogeneous.

One might expect that g_{sph} would correspond to the Brier score (upon inversion); however that turns out not to be the case. Set $c = \frac{2}{n}\mathbf{1}$ and $S = c - I$ so via (20) $\ell(x) = c = \frac{x}{\|x\|_2}$. Consider $n = 2$ and restrict ℓ to Δ^2 to get

$$\ell((p, 1-p)') = \mathbf{1} - ((p, 1-p)')\frac{1}{(p^2 + (1-p)^2)^{1/2}}$$

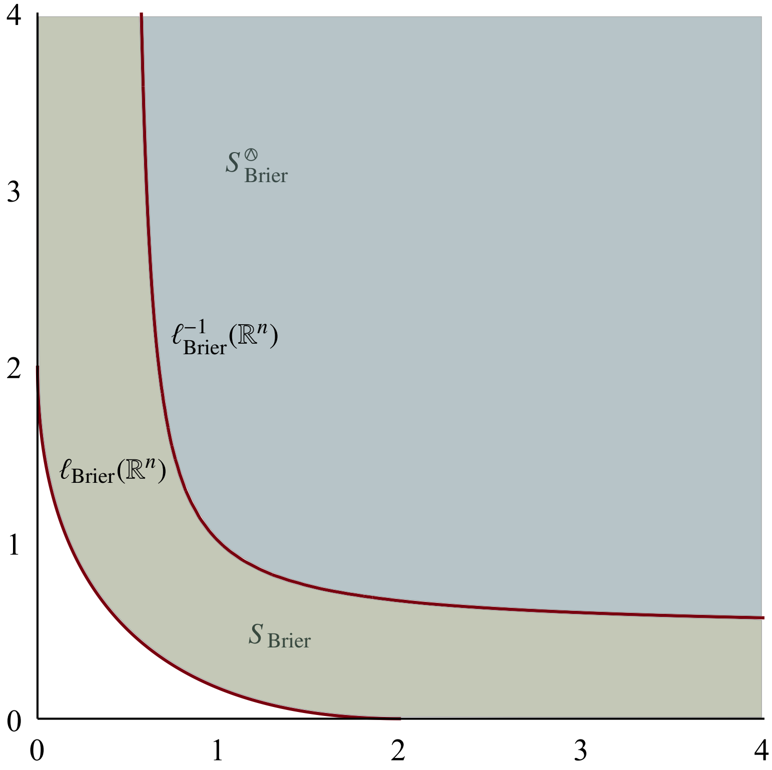

Figure 8: Brier loss and its inverse for $n = 2$.

and thus for $p \in [0, 1]$ the Bayes risk $\underline{L}(p) = \langle (p, 1-p)', \ell((p, 1-p)') \rangle = 1 - (p^2 + (1-p)^2)^{1/2}$. The weight function (Reid and Williamson, 2011) associated with the binary proper loss is

$$w(p) = -\underline{L}''(p) = \frac{1}{(2p^2 - 2p + 1)^{3/2}},$$

which differs from the weight function for the Brier score which is identically 1.

We will now determine the inverse loss for the Brier loss. This is easily plotted as with other losses because $S^{\owedge}_{\text{Brier}} = \operatorname{lev}_{\geq 1} \hat{\sigma}_{S_{\text{Brier}}}$; see Figure 8.

The Brier score is usually defined for $p \in \Delta^n$ in terms of its Bayes risk $\underline{L}_{\text{Brier}}(p) = 1 - \sum_{i=1}^n p_i^2$. For our purposes we need to work with the 1-homogeneous extension:

$$\underline{L}_{\text{Brier}} = \hat{\sigma}_{S_{\text{Brier}}} \colon \mathbb{R}^n \ni x \mapsto \|x\|_1 \left(1 - \sum_{i=1}^n \left(\frac{x_i}{\|x\|_1} \right)^2 \right) = \|x\|_1 - \frac{\|x\|_2^2}{\|x\|_1}.$$

Thus

$$\begin{aligned}\hat{\sigma}_{S^{\circledast}_{\text{Brier}}}(y) &= \inf_{x\neq 0} \frac{\langle x, y\rangle}{\hat{\sigma}_{S_{\text{Brier}}}(x)} \\ &= \inf_{x\neq 0} \frac{\langle x, y\rangle}{\|x\|_1 - \frac{\|x\|_2^2}{\|x\|_1}}.\end{aligned} \tag{22}$$

When $n = 2$ this is solvable explicitly. Since we know $\hat{\sigma}_{S^{\circledast}_{\text{Brier}}}$ must be 1-homogeneous it suffices to evaluate it on the simplex Δ^2 and then 1-homogeneously extend it. Parametrising an element of Δ^2 as $(p, 1-p)'$ we obtain

$$\hat{\sigma}_{S^{\circledast}_{\text{Brier}}}(p) = \inf_{x\neq 0} \frac{x_1 p + x_2(1-p)}{x_1 + x_2 - \frac{x_1^2+x_2^2}{x_1+x_2}}$$

This can be solved directly resulting in

$$\hat{\sigma}_{S^{\circledast}_{\text{Brier}}}(p) = \frac{1}{2}\frac{(2p-1)^2\sqrt{p(1-p)}}{2p^2 + \sqrt{p(1-p)} - 2p}$$

It does not seem possible to find a closed form for ℓ_{Brier}^{-1} when $n > 2$. However the objective function in (22) can be seen to be quasi-convex in x (since $\hat{\sigma}_S(x)$ is concave in x and thus $1/\hat{\sigma}_S(x)$ is quasi-convex) and thus is amenable to numerical solution.

Appendix F. Cobb-Douglas Functions and Boosting Loss

As a final example consider the parametrised concave function

$$\psi_a(x) := \begin{cases} \left(\prod_{i=1}^n x_i^{a_i}\right)^{1/\|a\|_1} & x \in \mathbb{R}^n_+ \\ -\infty & \text{otherwise,}\end{cases} \tag{23}$$

where $a = (a_1, \ldots, a_n)' \in (0, \infty)^n$. Barbara and Crouzeix (1994, page 52) show that ψ_a is "self-dual" in the sense that for all $x \in \mathbb{R}^n$

$$\psi_a^{\circledast}(x) = \frac{\|a\|_1}{\psi_a(a)}\psi_a(x). \tag{24}$$

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).

We illustrate the self-duality with a simple example. Set $n = 2$ and $a_1 = a_2 = 1$ and thus $\psi_a(x) = \sqrt{x_1 x_2}$. Taking this is as the concave support function for a loss ℓ_ψ, the polar is immediately obtainable from (24). One has

$$\ell_\psi(\mathbb{R}^2_+) = \operatorname{lev}_{=1} S_\psi \ \text{ and } \ \ell_\psi^{-1}(\mathbb{R}^2_+) = \operatorname{lev}_{=1/2} S_\psi.$$

The form of ℓ_ψ is of interest. Differentiating $\hat{\sigma}_\psi(x) = \psi(x)$ one obtains the partial losses

$$\ell_{\psi,1}(x) = \frac{1}{2}\frac{x_2}{\sqrt{x_1 x_2}} \ \text{ and } \ \ell_{\psi,2}(x) = \frac{1}{2}\frac{x_1}{\sqrt{x_1 x_2}}, \tag{25}$$

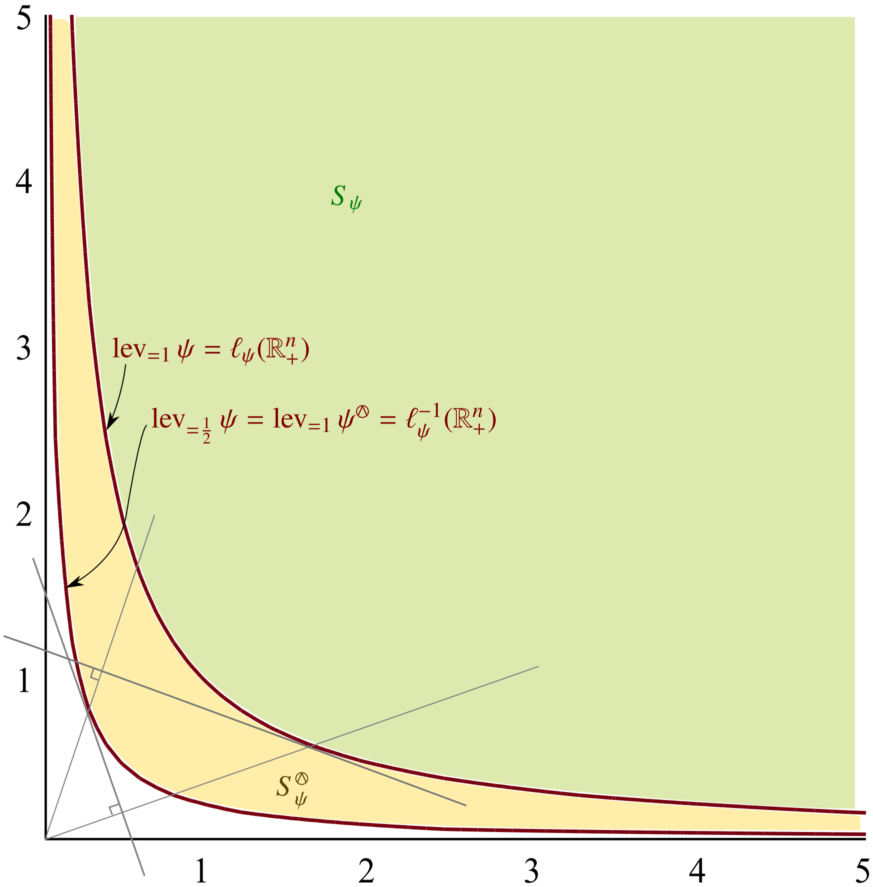

Figure 9: Illustration of the self-dual nature of the Cobb-Douglas loss when $a_1 = a_2 = 1$. The inverse loss can be found by taking the level set at level $\frac{1}{2}$. See also Figure 10.

and hence restricting the loss to Δ^2 we obtain

$$\ell_{\psi,1}(p) = \frac{1}{2}\sqrt{\frac{1-p}{p}} \quad \text{and} \quad \ell_{\psi,2}(p) = \frac{1}{2}\sqrt{\frac{p}{1-p}},$$

which can be recognised as the "boosting loss" — see Buja et al. (2005). This loss has weight function

$$w_\psi(p) = -\psi_a''(p) = \frac{1}{4(p(1-p))^{3/2}}.$$

The superprediction sets associated with the loss ℓ_ψ and its inverse are illustrated in Figures 9 and 10, which also shows the self-dual nature of the loss.

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" — the fact that for all $x \in \mathbb{R}^n_+$, $\ell_\psi(\ell_\psi(\ell_\psi(x)))) = \ell_\psi(x)$, a fact one can verify directly by using the formulae in (25).

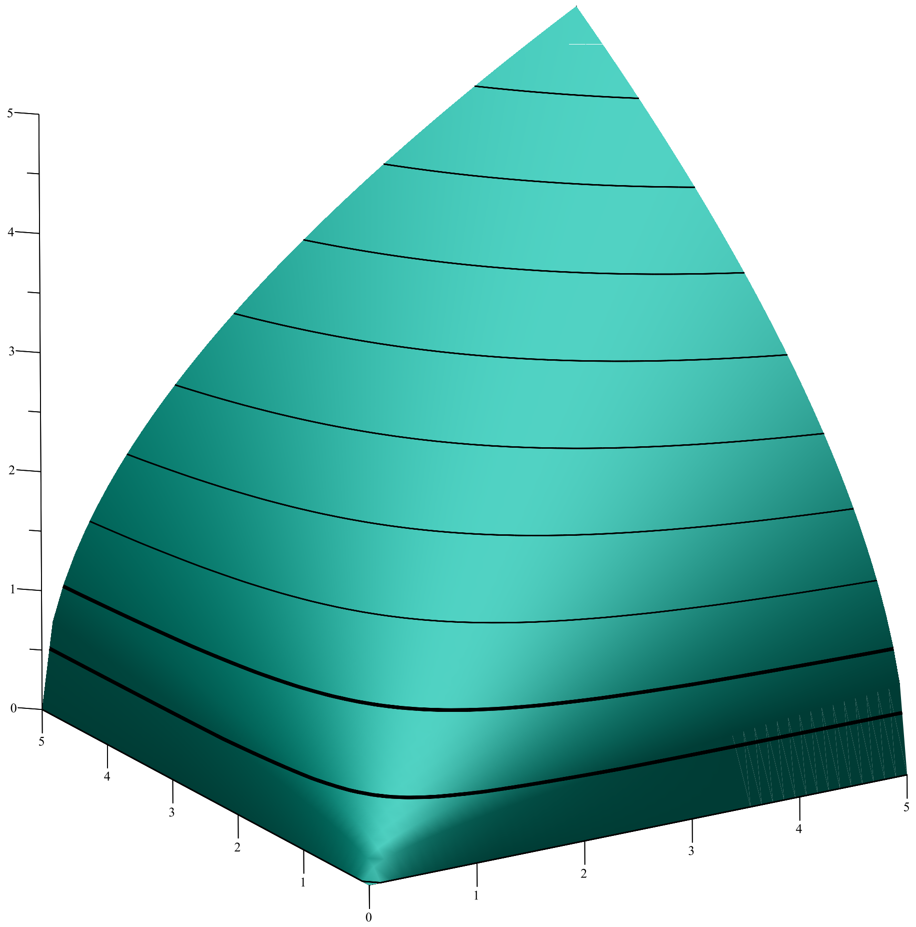

Figure 10: Graph of Cobb-Douglas support function (with $a_1 = a_2 = 1$) $\psi\colon \mathbb{R}^2_+ \ni (x, y) \mapsto \sqrt{xy}$ with the contours at height $\frac{1}{2}$ and 1 (thicker lines) corresponding to the inverse loss and loss curves respectively of Figure 9.

Appendix G. Proofs

Proof (Proposition 1) By definition, $\check{\partial} f(x) = \{s \in \mathbb{R}^n\colon f(y) \geq f(x) + \langle s, y - x\rangle, \ \forall y \in \mathbb{R}^n\}$. Thus for $\alpha > 0$,

$$\begin{aligned}\check{\partial} f(\alpha x) &= \{s\colon f(y) \geq f(\alpha x) + \langle s, y - \alpha x\rangle, \ \forall y \in \mathbb{R}^n\} \\ &= \{s\colon f(\alpha y') \geq f(\alpha x) + \langle s, \alpha y' - \alpha x\rangle, \ \forall y' \in \mathbb{R}^n\} \\ &= \{s\colon \alpha f(y') \geq \alpha f(x) + \alpha\langle s, y' - x\rangle, \ \forall y' \in \mathbb{R}^n\} = \check{\partial} f(x).\end{aligned}$$

■

Proof (Proposition 3) We first prove

$$s \in \hat{\partial}\hat{\gamma}_C(d) \Leftrightarrow \left[\hat{\gamma}_C(d) = \langle s, d\rangle \text{ and } s \in C^{\owedge}\right]. \tag{26}$$

($\Rightarrow$) Suppose $d \in \mathbb{R}^n$. If $s \in \hat{\partial}\hat{\gamma}_C(d)$ then

$$\forall y \in \mathbb{R}^n,\ \hat{\gamma}_C(y) \leq \hat{\gamma}_C(d) + \langle s, y - d\rangle.$$

Hence (setting $y = 0$ and then $y = 2d$, and exploiting the 1-homogeneity of $\hat{\gamma}_C$), we have

$$\begin{aligned} \hat{\gamma}_C(0) = 0 &\leq \hat{\gamma}_C(d) - \langle s, d\rangle \qquad (27)\\ \hat{\gamma}_C(2d) = 2\hat{\gamma}_C(d) &\leq \hat{\gamma}_C(d) + \langle s, d\rangle \end{aligned}$$

and thus

$$\hat{\gamma}_C(d) \leq \langle s, d\rangle \tag{28}$$

and hence (27) and (28) imply

$$\hat{\gamma}_C(d) = \langle s, d\rangle. \tag{29}$$

Hence

$$\forall y \in \mathbb{R}^n,\ \hat{\gamma}_C(y) \leq \langle s, d\rangle + \langle s, y - d\rangle = \langle s, y\rangle. \tag{30}$$

Observe now that

$$t \in C^{\owedge} \Leftrightarrow [\langle t, y\rangle \geq 1\ \forall y \text{ such that } \hat{\gamma}_C(y) = 1]$$

and so using (30)

$$\begin{aligned} &\Leftrightarrow [\hat{\gamma}_C(y) \geq 1 \Rightarrow \langle t, y\rangle \geq 1]\\ &\Leftrightarrow s = t \end{aligned}$$

and thus $s \in C^{\owedge}$.

($\Leftarrow$) Suppose then that $s \in C^{\owedge}$ and there exists d such that $\hat{\gamma}_C(d) = \langle s, d\rangle$. Then

$$\begin{aligned} 0 &= \hat{\gamma}_C(d) + \langle s, -d\rangle\\ \Rightarrow\ \langle s, y\rangle &= \hat{\gamma}_C(d) + \langle s, y - d\rangle, \forall y \in \mathbb{R}^n\\ \Rightarrow\ \langle s, y\rangle &\leq \hat{\gamma}_C(d) + \langle s, y - d\rangle, \forall y \in \mathbb{R}^n. \qquad (31) \end{aligned}$$

Since $s \in C^{\owedge}$, we have

$$\begin{aligned} &[\forall y \in \mathbb{R}^n,\ \hat{\gamma}_C(y) \geq 1 \Rightarrow \langle s, y\rangle \geq 1]\\ \Rightarrow\ &[\forall y \in \mathbb{R}^n,\ \forall \alpha > 0,\ \hat{\gamma}_C(y) \geq \alpha \Rightarrow \langle s, y\rangle \geq 1]\\ \Rightarrow\ &[\forall y \in \mathbb{R}^n,\ \forall \alpha > 0,\ \langle s, y\rangle < \alpha \Rightarrow \hat{\gamma}_C(y) < \alpha]\\ \Rightarrow\ &[\forall y \in \mathbb{R}^n,\ \forall \alpha > 0,\ \langle s, y\rangle \leq \alpha \Rightarrow \hat{\gamma}_C(y) \leq \alpha], \end{aligned}$$

where we used again the 1-homogeneity of $\hat{\gamma}_C$ and the fact that it is upper semi-continuous (all its above-level sets are closed), which is a consequence of $\hat{\gamma}_C = \hat{\sigma}_{C^{\owedge}}$ and the fact that concave support functions are upper-semicontinuous. Thus with (31), we have

$$\hat{\gamma}_C(y) \leq \hat{\gamma}_C(d) + \langle s, y-d \rangle, \ \forall y \in \mathbb{R}^n$$

which means $s \in \hat{\partial}\hat{\gamma}(d)$.

Equation 26 implies

$$\frac{s}{\hat{\gamma}_{C^{\owedge}}(s)} \in \hat{\partial}\hat{\gamma}_C(d) \quad \Leftrightarrow \quad \left[\hat{\gamma}_C(d)\hat{\gamma}_{C^{\owedge}}(s) = \langle s, d\rangle \text{ and } \frac{s}{\hat{\gamma}_{C^{\owedge}}(s)} \in C^{\owedge}\right]. \tag{32}$$

By swapping the roles of d and s and of C and $C^{\owedge}$, (26) also implies

$$d \in \hat{\partial}\hat{\gamma}_{C^{\owedge}}(s) \quad \Leftrightarrow \quad [\hat{\gamma}_{C^{\owedge}}(s) = \langle s, d\rangle \text{ and } d \in C].$$

Thus

$$\frac{d}{\hat{\gamma}_C(d)} \in \hat{\partial}\hat{\gamma}_{C^{\owedge}}(s) \quad \Leftrightarrow \quad \left[\hat{\gamma}_{C^{\owedge}}(s)\hat{\gamma}_C(d) = \langle s, d\rangle \text{ and } \frac{d}{\hat{\gamma}_C(d)} \in C\right]. \tag{33}$$

Since $C = \{x \colon \hat{\gamma}_C(x) \geq 1\}$,

$$\frac{d}{\hat{\gamma}_C(d)} \in C \quad \Leftrightarrow \quad \frac{\hat{\gamma}_C(d)}{\hat{\gamma}_C(d)} \geq 1.$$

Since the second term is always true we conclude that $\frac{d}{\hat{\gamma}_C(d)} \in C$ for all d. Likewise $\frac{s}{\hat{\gamma}_{C^{\owedge}}(s)} \in C^{\owedge}$ for all s. Hence (32) is equivalent to (33) and we have proved the proposition. ■

Proof (Proposition 8) From (3) we can write (since $(f(z), g(z)) \in \mathbb{R}^2_+$)

$$(f \check{\oplus}_C g)(z) = \sup\{\langle x, (f(z), g(z))'\rangle \colon \check{\gamma}_{C^{\ovee}}(x) = 1\}$$

Since $z \mapsto \langle x, (f(z), g(z))'\rangle$ is convex in z (for each x), $(f \check{\oplus}_C g)(z)$ is the supremum of a family of convex functions and thus convex.

Similarly we can write

$$(f \check{\square}_C g)(z) = \inf_{x_1+x_2=z} \sup\{\langle x, (f(x_1), g(x_2))'\rangle \colon \gamma_{C^{\ovee}}(x) \leq 1, \ x \in \mathbb{R}^n\}.$$

Let $\phi(x_1, x_2; x) = \langle x, (f(x_1), g(x_2))'\rangle$. For each x, $\phi(x_1, x_2; x)$ is jointly convex in x_1 and x_2. Thus $\psi(x_1, x_2) = \sup\{\phi(x_1, x_2; x) \colon \gamma_{C^{\ovee}} \leq 1\}$ shares this property as it is the supremum (over a convex set) of a family of jointly convex functions. Finally observe that $(f \check{\square}_C g)(z)$ is the restriction of $\psi(x_1, x_2)$ to a linear subspace and is thus convex in z. ■

Proof (Proposition 9) We have

$$
\begin{aligned}
\check{\sigma}_{A\check{\oplus}_C B}(x) &= \check{\sigma}_{\bigcup_{\lambda\in C^{\circledvee}\cap\mathbb{R}^2_+}(\lambda_1 A+\lambda_2 B)}(x)\\
&= \sup_{\lambda\in C^{\circledvee}\cap\mathbb{R}^2_+} \check{\sigma}_{\lambda_1 A+\lambda_2 B}(x)\\
&= \sup_{\lambda\in C^{\circledvee}\cap\mathbb{R}^2_+} \lambda_1\check{\sigma}_A(x)+\lambda_2\check{\sigma}_B(x)\\
&= \sup_{\lambda\in C^{\circledvee}\cap\mathbb{R}+^2} \langle\lambda,(\check{\sigma}_A(x),\check{\sigma}_B(x))'\rangle \qquad (34)\\
&= \check{\sigma}_{C^{\circledvee}}((\check{\sigma}_A(x),\check{\sigma}_B(x))') \qquad (35)\\
&= \check{\gamma}_C((\check{\sigma}_A(x),\check{\sigma}_B(x))')\\
&= (\check{\sigma}_A \check{\oplus}_C \check{\sigma}_B)(x)
\end{aligned}
$$

where the step from (34) to (35) is justified since A and B are of negative type, $\check{\sigma}_A(x)$ and $\check{\sigma}_B(x)$ are positive for all x, and hence the supremum over λ will automatically be in $\mathbb{R}^2_+$. This proves (17). In order to prove (18) we observe that

$$
\begin{aligned}
(\check{\sigma}_A \,\check{\square}_C\, \check{\sigma}_B)^{\check{*}}(x) &= \sup_{x^*\in\mathbb{R}^n}\{\langle x,x^*\rangle - \inf_{x_1^*+x_2^*=x^*} \check{\gamma}_{C^{\circledvee}}((\check{\sigma}_A(x_1^*),\check{\sigma}_B(x_2^*))')\}\\
&= \sup_{x^*\in\mathbb{R}^n}\sup_{x_1^*+x_2^*=x^*} \{\langle x,x^*\rangle - \check{\gamma}_{C^{\circledvee}}((\check{\sigma}_A(x_1^*),\check{\sigma}_B(x_2^*))')\}\\
&= \sup_{x_1^*,x_2^*\in\mathbb{R}^n} \{\langle x,x_1^*\rangle + \langle x,x_2^*\rangle - \check{\gamma}_{C^{\circledvee}}((\check{\sigma}_A(x_1^*),\check{\sigma}_B(x_2^*))')\}\\
&= \sup_{x_1^*,x_2^*\in\operatorname{dom}\check{\sigma}_A\times\operatorname{dom}\check{\sigma}_B=:M} \{\langle x,x_1^*\rangle + \langle x,x_2^*\rangle - \check{\gamma}_{C^{\circledvee}}((\check{\sigma}_A(x_1^*),\check{\sigma}_B(x_2^*))')\}
\end{aligned}
$$

but for $(a,b)\in\mathbb{R}^2_+$, $\check{\gamma}_{C^{\circledvee}}((a,b)') = \sup\{\lambda_1 a+\lambda_2 b\colon \lambda\in C\}$ and since $0\in A$, $\check{\sigma}_A(x_1^*) = \sup_{y^*\in A}\langle x^*,y^*\rangle \geq \langle x^*,0\rangle = 0$ (and similarly for $\check{\sigma}_B(x_2^*)$), we can write

$$
(\check{\sigma}_A \,\check{\square}_C\, \check{\sigma}_B)^{\check{*}}(x) = \sup_{(x_1^*,x_2^*)\in M}\inf_{\lambda\in C} L_x((x_1^*,x_2^*),\lambda),
$$

where L_x is a finite function defined on $M\times C$ by

$$
L_x((x_1^*,x_2^*),x) = \langle x,x_1^*\rangle + \langle x,x_2^*\rangle - \lambda_1\check{\sigma}_A(x_1^*) - \lambda_2\check{\sigma}_B(x_2^*).
$$

Since $L_x((x_1^*,x_2^*),\lambda)$ is concave in (x_1^*,x_2^*) and convex in λ, $\operatorname{dom}\check{\sigma}_A$ and $\operatorname{dom}\check{\sigma}_B$ are convex (since $\overline{\operatorname{dom}\check{\sigma}_A} = (0^+A)^{\circledvee}$ (Hiriart-Urruty and Lemaréchal, 2001, p. 140)) so M is convex, and C is convex by assumption, and thus Sion's minimax theorem (Sion, 1958) applies and we can write

$$
\begin{aligned}
(\check{\sigma}_A \,\check{\square}_C\, \check{\sigma}_B)^{\check{*}} &= \inf_{\lambda\in C}\sup_{(x_1^*,x_2^*)\in M} L_x((x_1^*,x_2^*),\lambda)\\
&= \inf_{\lambda\in C}\left(r_{A,\lambda_1}(x)+r_{B,\lambda_2}(x)\right), \qquad (36)
\end{aligned}
$$

where

$$
\begin{aligned}
r_{A,\lambda_1}(x) &:= \sup_{x_1^* \in \operatorname{dom} \check{\sigma}_A} \left(\langle x, x^* \rangle - \lambda_1 \star \check{\sigma}_A(x_1^*) \right) \\
r_{B,\lambda_2}(x) &:= \sup_{x_2^* \in \operatorname{dom} \check{\sigma}_B} \left(\langle x, x^* \rangle - \lambda_2 \star \check{\sigma}_B(x_2^*) \right).
\end{aligned}
\tag{37}
$$

Now when $\lambda_1 = 0$,

$$
r_{A,\lambda_1}(x) = \sup_{x_1^* \in \operatorname{dom} \check{\sigma}_A} \langle x, x^* \rangle = \check{\sigma}_{\operatorname{dom} \check{\sigma}_A}(x) = (\check{\imath}_A 0^+)(x) = \check{\imath}_{0^+ A}(x),
$$

where the last step follows from (Hiriart-Urruty and Lemaréchal, 2001, p. 107). Hence for all $\lambda_1 \geq 0$, $r_{A,\lambda_1}(x) = \check{\imath}_{\lambda_1 \star A}(x)$ and similarly for all $\lambda_2 \geq 0$ $r_{B,\lambda_2}(x) = \check{\imath}_{\lambda_2 \star B}(x)$. Thus taking conjugates of both sides of (36) and using Fenchel's duality theorem, we have

$$
\begin{aligned}
(\check{\sigma}_A \,\check{\Box}_C\, \check{\sigma}_B)^{\check{*}\check{*}}(x^*) &= \overline{(\check{\sigma}_A \,\check{\Box}_C\, \check{\sigma}_B)}(x^*) \\
&= \sup_{\lambda \in C \cap \mathbb{R}^2_+} (r_{A,\lambda_1} + r_{B,\lambda_2})^{\check{*}}(x^*) \\
&= \sup_{\lambda \in C \cap \mathbb{R}^2_+} (\check{\imath}_{\lambda_1 \star A} + \check{\imath}_{\lambda_2 \star B})^{\check{*}}\,(x^*) \\
&= \sup_{\lambda \in C \cap \mathbb{R}^2_+} \left(\check{\sigma}_{\lambda_1 \star A} \,\check{\Box}_1\, \check{\sigma}_{\lambda_2 \star B} \right)(x^*) \\
&= \sup_{\lambda \in C \cap \mathbb{R}^2_+} \check{\sigma}_{\lambda_1 \star A \cap \lambda_2 \star B}(x^*).
\end{aligned}
$$

Now

$$
A \,\check{\bar{\Box}}_C\, B = \bigcup_{\lambda \in C \cap \mathbb{R}^2_+} (\lambda_1 \star A \cap \lambda_2 \star B)
$$

and thus

$$
\check{\sigma}_{A \check{\bar{\Box}} B} = \sup_{\lambda \in C \cap \mathbb{R}^2_+} \sigma_{\lambda_1 \star A \cap \lambda_2 \star B}.
$$

By Proposition 6 we have

$$
\check{\sigma}_{\overline{A \check{\Box}_C B}} = \check{\sigma}_{A \check{\Box}_C B} \leq \check{\sigma}_{A \check{\bar{\Box}}_C B} \leq \check{\sigma}_{\overline{A \check{\Box}_C B}}
$$

and thus $\check{\sigma}_{A \check{\bar{\Box}}_C B} = \check{\sigma}_{A \check{\Box}_C B}$. ■

Proof (Proposition 10) We make use of the classical result that $(\operatorname{co}(A \cup B))^{\ovee} = A^{\ovee} \cap B^{\ovee}$ (which is in fact proved in Seeger (1990) as a special case of the present theorem). Due to associativity of unions and intersections this extends to the polar of arbitrary unions of sets in the obvious manner.

Also note that as we have shown, $A \check{\oplus}_C B$ is convex and hence

$$
\begin{aligned}
(A \check{\oplus}_C B)^{\circledvee} &= \left(\bigcup_{\lambda \in C^{\circledvee} \cap \mathbb{R}^2_+} (\lambda_1 A + \lambda_2 B) \right)^{\circledvee} \\
&= \bigcap_{\lambda \in C^{\circledvee} \cap \mathbb{R}^2_+} (\lambda_1 A + \lambda_2 B)^{\circledvee}.
\end{aligned}
$$

Set $S_1 := \{x \in \mathbb{R}_+ \colon (x, 0)' \in C^{\circledvee}\}$, $S_2 := \{x \in \mathbb{R}_+ \colon (0, x)' \in C^{\circledvee}\}$ and $S_0 = C^{\circledvee} \setminus (S_1 \cup S_2)$. Then

$$
(A \check{\oplus}_C B)^{\circledvee} = \left(\bigcap_{\lambda \in S_0} \overline{((\lambda_1 A)^{\circledvee} \sharp (\lambda_2 B)^{\circledvee})} \right) \cap \left(\bigcap_{\lambda_1 \in S_1} (\lambda_1 A)^{\circledvee} \right) \cap \left(\bigcap_{\lambda_2 \in S_2} (\lambda_2 B)^{\circledvee} \right),
$$

where we have used the result $(A + B)^{\circledvee} = \overline{A^{\circledvee} \sharp B^{\circledvee}}$ (Seeger, 1990). Furthermore, since the intersection of closed sets is closed and for $\epsilon > 0$, $(\epsilon A)^{\circledvee} = \frac{1}{\epsilon} A^{\circledvee}$ we have

$$
\begin{aligned}
(A \check{\oplus}_C B)^{\circledvee} &= \overline{\left(\bigcap_{\lambda \in S_0} \left(\frac{1}{\lambda_1} A^{\circledvee} \right) \sharp \left(\frac{1}{\lambda_2} B^{\circledvee} \right) \right) \cap \left(\bigcap_{\lambda_1 \in S_1} (\lambda_1 A)^{\circledvee} \right) \cap \left(\bigcap_{\lambda_2 \in S_2} (\lambda_2 B)^{\circledvee} \right)} \\
&= \overline{\left(\bigcap_{\lambda \in S_0} \bigcup_{\alpha_1 + \alpha_2 = 1, \alpha_1, \alpha_2 \geq 0} \left(\frac{\alpha_1}{\lambda_1} A^{\circledvee} \cap \frac{\alpha_2}{\lambda_2} B^{\circledvee} \right) \right) \cap \left(\bigcap_{\lambda_1 \in S_1} (\lambda_1 A)^{\circledvee} \right) \cap \left(\bigcap_{\lambda_2 \in S_2} (\lambda_2 B)^{\circledvee} \right)}.
\end{aligned}
$$

For $\lambda \in S_0$, let $\beta_1 := \frac{\alpha_1}{\lambda_1}$ and $\beta_2 := \frac{\alpha_2}{\lambda_2}$, so $\alpha_1 = \beta_1 \lambda_1$, $\alpha_2 = \beta_2 \lambda_2$ and hence

$$
\begin{aligned}
(A \check{\oplus}_C B)^{\circledvee} &= \overline{\left(\bigcap_{\lambda \in S_0} \bigcup_{\beta \colon \langle \beta, \lambda \rangle \leq 1,\, \beta \in \mathbb{R}^2_+} (\beta_1 A^{\circledvee} \cap \beta_2 B^{\circledvee}) \right) \cap \left(\bigcap_{\lambda_1 \in S_1} (\lambda_1 A)^{\circledvee} \right) \cap \left(\bigcap_{\lambda_2 \in S_2} (\lambda_2 B)^{\circledvee} \right)} \\
&= \overline{\left(\bigcup_{\beta \colon \langle \beta, \lambda \rangle \leq 1,, \beta \in \mathbb{R}^2_+ \ \forall \lambda \in S_0} (\beta_1 A^{\circledvee} \cap \beta_2 B^{\circledvee}) \right)} \cap \beta_1^* A^{\circledvee} \cap \beta_2^* B^{\circledvee},
\end{aligned}
$$

where $\beta_i^* = 1/\lambda_i^*$, and $\lambda_1^* = \max\{\lambda_1 \colon (\lambda_1, 0)' \in C^{\circledvee}\}$ and $\lambda_2^* = \max\{\lambda_2 \colon (0, \lambda_2)' \in C^{\circledvee}\}$ and thus β_1^* satisfies $\langle (\beta_1^*, 0)', (\lambda_1, 0)' \rangle \leq 1 \ \forall (\lambda_1, 0)' \in C^{\circledvee}$ and β_2^* satisfies $\langle (0, \beta_2^*)', (0, \lambda_2)' \rangle \leq 1 \ \forall (0, \lambda_2)' \in C^{\circledvee}$. Therefore we can incorporate the second and third terms in the equation above into a more restrictive condition on the union, and so

$$
(A \check{\oplus}_C B)^{\circledvee} = \overline{\bigcup_{\beta \colon \langle \beta, \lambda \rangle \leq 1,\, \beta \in \mathbb{R}^2_+,\ \forall \lambda \in C^{\circledvee}} (\beta_1 A^{\circledvee} \cap \beta_2 B^{\circledvee})},
$$

which by the dual representation of C (3) gives

$$
\begin{aligned}
&= \overline{\bigcup_{\beta \in C \cap \mathbb{R}^2_+} (\beta_1 A^{\circledvee} \cap \beta_2 B^{\circledvee})} \\
&= \overline{A^{\circledvee} \check{\Box}_{C^{\circledvee}} B^{\circledvee}}.
\end{aligned}
$$

■