An Analytic Approach to the Structure and Composition of General Learning Problems

Zachary Cranko

March 2021

A thesis submitted for the degree of
Doctor of Philosophy of
The Australian National University.

This thesis is the result of original research, and has not been submitted for a postgraduate degree at any other university or institution. Following is a list of publications I contributed to over the course of my studies.

- Nock, R., Cranko, Z., Menon, A. K., Qu, L., and Williamson, R. C. "f-GANs in an Information Geometric Nutshell". *Advances in Neural Information Processing Systems 30.* Long Beach, CA, USA: Curran Associates, Inc., 2017, pp. 456–464
- Cranko, Z. and Nock, R. "Boosted Density Estimation Remastered". *Proceedings of the 36th International Conference on Machine Learning.* Vol. 97. Long Beach, CA, USA: Proceedings of machine learning research, June 9–15, 2019, pp. 1416–1425
- Cranko, Z., Menon, A., Nock, R., Ong, C. S., Shi, Z., and Walder, C. "Monge Blunts Bayes: Hardness Results for Adversarial Training". *Proceedings of the 36th International Conference on Machine Learning.* Vol. 97. Long Beach, CA, USA: Proceedings of machine learning research, June 9–15, 2019, pp. 1406–1415
- Husein, H., Balle, B., Cranko, Z., and Nock, R. "Local Differential Privacy for Sampling". *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics.* Palermo, Italy: Proceedings of machine learning research, June 3–5, 2020

The research presented in this thesis is the result of collaboration with Prof. Robert C. Williamson and Prof. Richard Nock. Approximately 90% of the work presented is my own. Much of the material in Chapter 5 previously appeared in the second listed publication above.

Zachary Cranko
The Australian National University
January 2021

Acknowledgements

I am hopeful that the following pages hold some academic value, but as their author, they have value to me as a memento of the years I spent as a PhD candidate. If you, the reader, are a PhD candidate, you should know that not everyone's doctoral experience is the same, but I found mine quite challenging. It is for that reason that I am indebted to several individuals, without whom I would not have even made it to the halfway mark.

As one of their final PhD students at the Australian National University, I am particularly grateful for my two supervisors, Prof. Robert C. Williamson and Prof. Richard Nock. From the moment I first stepped into Bob's office in the September of 2015, until my conclusory seminar in January of 2020, I received only patience, kindness, and support from Bob, and that is as much as any PhD candidate could hope for. Similarly, Richard's buoyant personality and sense of humor has been an essential conterveiling factor in keeping me sane and engaged.

In addition, my thanks goes to my two examiners Nicholas Vitayas and Tilmann Gneiting for their enthusiasm and helpful comments, and to a selection of my colleagues and mentors, both proximate and distant, who were kind enough to share their thoughts and wisdom with me: Erik Davis and Brendan Pawlowski; Christfried Webers, Qinian Jin, Stephen Roberts; Aditya Menon, Xinhua Zhang, and Christian Walder. Finally I could not have made it through my studies, both undergraduate and postgraduate, without the love and support of my parents Lynne and Craig, my grandmother Ma Ruth, and my girlfriend Carol.

Abstract

Gowers presents, in his 2000 essay "The Two Cultures of Mathematics", two kinds of mathematicians he calls the theory-builders and problem-solvers. Of course both kinds of research are important; theory building may directly lead to solutions to problems, and by studying individual problems one uncovers the general structures of problems themselves. However, referencing a remark of Atiyah [9], Gowers observes that because so much research is produced, the results that can be "organised coherently and explained economically" will be the ones that last. Unlike mathematics, the field of machine learning abounds in problem-solvers — this is wonderful as it leads to a large number of problems being solved — but it is with regard to the point of Gowers that we are motivated to develop an appropriately general analytic framework to study machine learning problems themselves.

To do this we first locate and develop the appropriate analytic objects to study. Chapter 2 recalls some concepts and definitions from the theory of topological vector spaces. In particular, the families of radiant and co-radiant sets and dualities. In Chapter 4 we will need generalisations of a variety of existing results on these families, and these are presented in Chapter 3.

Classically a machine learning problem involves four quantities: an outcome space, a family of predictions (or model),[1] a loss function, and a probability distribution. If the loss function is sufficiently general we can combine it with the set of predictions to form a set of real functions, which under very general assumptions, turns out to be closed, convex, and in

[1] In the sequel we use the terms *prediction* and *model* interchangeably since the distinction is largely semantic.

particular, co-radiant. With the machinery of the previous two chapters in place, in Chapter 4 we lay out the foundations for an analytic theory of the classical machine learning problem, including a general analysis of link functions, by which we may rewrite almost any loss function as a scoring rule; a discussion of scoring rules and their properisation; and using the co-radiant results from Chapter 3 in particular, a theory of prediction aggregation.

Chapters 5 and 6 develop results inspired by and related to adversarial learning. Chapter 5 develops a theory of boosted density estimation with strong convergence guarantees, where density updates are computed by training a classifier, and Chapter 6 uses the theory of optimal transport to formulate a robust Bayes minimisation problem, in which we develop a universal theory of regularisation and deliver new strong results for the problem of adversarial learning.

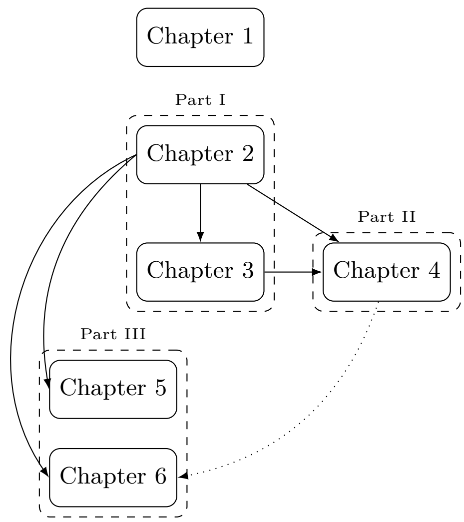


Figure 1: Dependencies among chapters.

Contents

Chapter 1
Introduction

It is a necessity that, when developing theory, one begins at the bottom, working upwards to ensure that various desiderata are satisfied before building on preliminary results. However, in direct opposition, for the purposes of motivating the decisions when building a theory, it is more helpful to structure the results in the opposite way, motivating the theory by what it achieves in those subsequent chapters. These competing demands of motivation and rigour leave the author with a kind of chicken and egg dilemma. By the inclusion of this introduction, we hope to provide the reader with some explanation for the direction of the subsequent chapters, by introducing the following chapters in a non-linear fashion.

In Chapter 4 we seek to develop the basis for a general theory of classical machine learning problems. The most basic way to think of such a problem is the prediction of some kind of distribution over an *outcome space*. The goodness of this prediction is evaluated with a *scoring rule* by calculating the expected loss or penalty incurred, using a loss function. The prediction itself may be a distribution, in which case the loss function is a called a scoring rule, or it may be something more abstract (like a function). It turns out that the natural convex structure of scoring rules lets us write any kind of prediction as one of a family of distributions. It is in this way that almost any kind of *model class* can be reparameterised using a scoring rule like this, using a *link function*; the existence of which is guaranteed through the theory of convex duality (Section 4.3.2). This is one of many results we can achieve by developing our general theory of learning problems using convex analysis as a foundation. Some others we cover are ways to ensure scoring rules yield accurate predictions via properisation (Section 4.3), and how to combine scoring rules and compute their associated link functions (Section 4.4.2). A key group of objects of study in the theory of convex analysis are the convex sets, and so we establish some basic properties of a certain set, associated with a learning problem called the *superprediction set* (Section 4.1.1). These sets have many interesting properties. For example, when the scoring rule is continuous, these sets are convex precisely when the scoring rule is proper (Section 4.2.2). More generally these sets are often members of a kind of unbounded set family, known as the *co-radiant* sets, and it is for this reason that we are motivated to study these sets.

The theory of co-radiant sets is often studied in the convex analysis literature together with their (often bounded) counterpart, the radiant sets. Every co-radiant set is the complement of a *radiant* set, and vice versa. The results mentioned above that are constructed using the superprediction sets require developing a theory for manipulating the co-radiant sets, and for completeness we show the companion results for the radiant sets (Chapter 3). As part of this algebra we have a number of formulas for support and gauge functions (Section 3.5), but to compute these we need to construct a theory of duality (Section 3.4), and in turn to produce the theory of duality we need some simpler formulas for some other *gauge* (Section 3.3) and *support functions* (Section 3.1). Some of these results however require a somewhat lengthy investigation of the topology of these sets (Section 3.2) and, in particular, results associated with their *asymptotic cones*, which are an object to simplify the analysis of unbounded sets. Even though most machine learning problems can be represented in a Banach space, in the convex analysis literature, results of the sort in Chapter 3 are typically proven in a locally convex, Hausdorff topological vector space. For our results here to represent a strict generalisation of other related works, we need to be intimately familiar with the mechanics of these spaces.

The setting of a locally convex, Hausdorff topological space is one of the most general vector spaces in mathematics; it is endowed with the minimal structure necessary for the majority of essential results in analysis and convex analysis to hold (in particular, the Hahn–Banach separation theorem, and the Bourbaki–Alaoglu theorem), and we provide an introduction to these spaces and key results on them in Chapter 2. To manipulate sets in these spaces in Chapter 3 we have certain results that require the sum of two sets is closed. Unfortunately for us, we cannot be sure that either or both of the sets will be bounded, and so we need results on the closure of the sum of sets (Section 2.4.1). To be able to do study the sublinear functions on these spaces (in particular the gauge and support functions) we need to be able to calculate and ensure certain behaviours of their subdifferentials (Sections 2.4.2 and 2.4.3). Chapter 2 begins with an introduction to all of the basic concepts we will need for these results, along with some basic properties.

The theory developed in Chapter 4, while interesting and rich in its foundations, is not general enough to include some more exotic kinds of learning problems that we introduce by way of example. In a binary classification task, the Radon–Nikodym derivative can be related to the Bayes-optimal classifier. Of course being able to compute the Radon–Nikodym derivative between an initial guess and the true distribution makes performing density estimation a triviality. It then should not be too surprising that by making an initial guess at the true distribution and learning a classifier, that we can learn some information about the Radon–Nikodym derivative. Using this observation, in Chapter 5 we show how the how a sequence of binary classifiers can be used to construct density estimates. And by making weak assumptions about the performance of these classifiers, we can derive strong convergence guarantees for density estimation.

When performing risk minimisation, instead of fitting a single distribution, one might instead look at a neighbourhood of distributions called an uncertainty set — that way if it turns out the data that one has access to were not completely representitive of the true distribution, the penalty of the misspecification is not too severe. This kind of risk is called a *robust Bayes risk*. Parallelly, there has been interest in regularisation for ensuring performance against so-called adversarial examples. In Chapter 6 we develop a general theory of regularisation that explains both of these phenomena. Using the transportation cost, from the theory of optimal transport, we formalise a notion of an uncertainty set. It's then shown with equality when the worst case risk over the uncertainty set is equal to the Lipschitz regularised risk, and in the other cases we prove a tight upper bound result. As an application we show how adversarial learning may be located within the robust Bayes framework.

Finally each chapter following Chapter 2 is bookmarked with a brief introduction and conclusion to summarise the intervening material.

In some ways the task of developing theory can be reflected upon as a shifting of onus. In a more problem-driven approach to machine learning, the onus is on the reader to understand a series of problems, and to be familiar with the panoply of idiosyncratic solutions for each. A theory of machine learning problems, on the other hand, shifts the onus of understanding the commonality of problems to the theoretician. It is in this way that when a new problem arises, or a new question is asked, that we have an array of tools at our disposal to analyse and compare a new engineering challenge with the ones we have already thoroughly understood.

Part I

Nonsmooth Analysis

Chapter 2
Technical Preliminaries

Some special sets are $[k] \stackrel{\text{def}}{=} \{1, \dots, k\}$ with $k \in \mathbb{N}$, $\bar{\mathbb{R}} \stackrel{\text{def}}{=} [-\infty, +\infty]$, $\mathbb{R}_{\geq 0} \stackrel{\text{def}}{=} [0, \infty)$, and $\mathbb{R}_{>0} \stackrel{\text{def}}{=} (0, \infty)$. For set T and a function $f : T \to \bar{\mathbb{R}}$, the set of points on which it achieves its infimum is $\operatorname{arginf}_{t \in T} f(t) \stackrel{\text{def}}{=} \{t \in T \mid f(t) = \inf f(T)\}$, with $\operatorname{argsup}_{t \in T} f(t)$ defined similarly. We use the standard conventions $\inf \emptyset = +\infty$ and $\sup \emptyset = -\infty$. The Iverson bracket is $[\![\cdot]\!]$, which takes the value 1 when its argument is a true proposition and 0 otherwise. All vector spaces are implicitly over the real numbers.

2.1 Topological spaces and their measures

Let X be a topological space, its Borel sigma algebra is $\mathscr{B}(X)$ and the collection of Borel probability measures is $\mathfrak{P}(X)$. A subset of X is called G_δ if it is the countable intersection of open sets. A net $(x_i)_{i \in I} \subseteq$ is a function from a directed set I to X. When X is Hausdorff and (x_i) converges, we use $\lim_{i \in I} x_i$ to denote its limit. The Dirac measure at $x \in X$ is $\delta_x\, A \stackrel{\text{def}}{=} [\![x \in A]\!]$ for $A \subseteq X$.

When Y is another topological spaces, the vector space of Borel measurable functions $X \to Y$ are collected in the set $\mathscr{L}_0(X, Y)$. When (X, Σ, λ) is a measurable space, for $p \geq 1$ there is the semi norm

$$\forall_{p \geq 1} \forall_{f \in \mathscr{L}_0(X, \bar{\mathbb{R}})} :\ |f|_p \stackrel{\text{def}}{=} \left(\int |f(x)|^p \lambda(\mathrm{d}x) \right)^{\frac{1}{p}},$$

and $|f|_\infty \overset{\text{def}}{=} \operatorname{ess\,sup}_{x\in X} f(x)$ for $f \in \mathscr{L}_0(X, \bar{\mathbb{R}})$. The Lebesgue spaces are

$$\forall_{p\in[1,\infty]} : \mathscr{L}_p(X,\lambda) \overset{\text{def}}{=} \{f \in \mathscr{L}_0(X,\mathbb{R}) \mid |f|_p < \infty\},$$

with the usual quotient of equivalence under the seminorm. The continuous functions $X \to Y$ are collected in $\mathrm{C}(X,Y)$, and the subcollection of bounded continuous functions is $\mathrm{C_b}(X,Y)$. When Y is the set of real numbers these are abbreviated $\mathscr{L}_0(X)$, $\mathrm{C}(X)$, and $\mathrm{C_b}(X)$.

2.2 Topological vector spaces

Throughout L is a Hausdorff locally convex topological vector space over the reals. The set of continuous linear functions $L \to \mathbb{R}$ is the topological dual, L^*, and these are connected via the duality pairing $\langle \cdot\,,\cdot\rangle : L \times L^* \to \mathbb{R}$. The weakest topology on L that generates L^* is $\sigma(L, L^*)$ and the strongest topology that generates L^* is $\tau(L, L^*)$ and coincides with the initial topology when L is metrisable. Closure operations for sets $A \subseteq L^*$ with $\sigma(L^*, L)$ are denoted $\operatorname{cl}^* A$ and $\overline{A}^*$. The following operations are standard:

$$A + b \overset{\text{def}}{=} \{a + b \in L \mid a \in A\} \qquad c \cdot A \overset{\text{def}}{=} \{ca \mid a \in A\}$$
$$A + B \overset{\text{def}}{=} \bigcup_{b\in B} A + b \qquad I \cdot A \overset{\text{def}}{=} \bigcup_{c\in I} c \cdot A,$$

for $b \in L$, $c \in \mathbb{R}$, $A, B \subseteq L$, $I \subseteq \mathbb{R}$.

A *set-valued mapping* between sets L and M, denoted $F : L \rightrightarrows M$, maps elements of L to subsets of M. By convention its *domain* is the set of points in L where it is nonempty, $\operatorname{dom} F \overset{\text{def}}{=} \{x \in L \mid F(x) \neq \emptyset\}$, and its *graph* is

$$\operatorname{gr} F \overset{\text{def}}{=} \{(x, y) \in L \times M \mid x \in \operatorname{dom} F,\ y \in F(x)\}.$$

If $G : L \rightrightarrows M$ is another set-valued map, then $F \cap G$ is the mapping with $(F \cap G)(x) \overset{\text{def}}{=} F(x) \cap G(x)$ for all $x \in L$. A *selection* of F is a function $f : \operatorname{dom} F \to M$ with $f(x) \in F(x)$ for all $x \in \operatorname{dom} F$, or equivalently, $\operatorname{gr} f \subseteq \operatorname{gr} F$.

Let $f : L \to \bar{\mathbb{R}}$. The *Fenchel conjugate* of f is the function $f^* : L^* \to \bar{\mathbb{R}}$

with

$$f^*(x^*) \stackrel{\text{def}}{=} \sup_{x \in L} \Big(\langle x, x^* \rangle - f(x) \Big).$$

The *lower-semicontinuous closure* of f, denoted by $\overline{f}$, is the greatest lower-semicontinuous minorant of f. The *upper-semicontinuous closure* of f, denoted by $\underline{f}$, is the least upper-semicontinuous majorant of f, and satisfies $\underline{f} = -\overline{(-f)}$. Its *$\epsilon$-subdifferential* is $\partial_\epsilon f : L \rightrightarrows L^*$

$$\partial_\epsilon f(x) \stackrel{\text{def}}{=} \{x^* \in L^* \mid \forall_{y \in L} : \ \langle y - x, x^* \rangle - \epsilon \leq f(y) - f(x)\},$$

where $\epsilon \geq 0$, $x \in L$. The *Moreau–Rockafellar subdifferential* is $\partial \stackrel{\text{def}}{=} \partial_0$, and satisfies $\partial f = \bigcap_{\epsilon > 0} \partial_\epsilon f$. Its *domain* is the set $\operatorname{dom} f \stackrel{\text{def}}{=} \{x \in L \mid f(x) \in \mathbb{R}\}$. Its *epigraph* and *sublevel sets* are

$$\begin{aligned} \operatorname{epi} f &\stackrel{\text{def}}{=} \{(x, t) \in \operatorname{dom}(f) \times \mathbb{R} \mid f(x) \leq t\}, \\ \operatorname{lev}_{\leq c} f &\stackrel{\text{def}}{=} \{x \in L \mid f(x) \leq c\}, \end{aligned}$$

where $c \in \mathbb{R}$. The sets $\operatorname{lev}_{<c} f$, $\operatorname{lev}_{\geq c} f$, and $\operatorname{lev}_{>c} f$, are defined analogously. We let $\widehat{\partial} f \stackrel{\text{def}}{=} -\partial(-f)$. This set-valued map is sometimes called the concave subdifferential or superdifferential [106, p. 308].

If $f(cx) = cf(x)$ for all $x \in L$ and $c \in \mathbb{R}_{>0}$, then f is *positively homogeneous* (or 1-homogeneous). If $f(x + y) \leq f(x) + f(y)$ for all $x, y \in L$, then f is *subadditive*. If f is both subadditive and positively homogeneous it is *sublinear*. Alternatively, f is sublinear if and only if it is positively homogeneous and convex.

Let A be a subset of L. The topological closure is $\operatorname{cl} A$ or $\overline{A}$, the *convex hull* is $\operatorname{co} A$ and the closure of the convex hull is $\overline{\operatorname{co}}\, A \stackrel{\text{def}}{=} \operatorname{cl}(\operatorname{co} A)$. If

$$\forall_{c>0} : \ c \cdot A \subseteq A \quad \text{and} \quad A + A \subseteq A$$

then A is called a *cone*. If A is a cone then A is *pointed* if $0 \in A$. The *conic hull* is $\operatorname{pos}(A) \stackrel{\text{def}}{=} (0, \infty) \cdot A$, and its closure is likewise $\overline{\operatorname{pos}}\, A$. We associate to A the following sets:

$$
\begin{aligned}
A^{\circ} &\stackrel{\text{def}}{=} \{x^* \in L^* \mid \forall_{a \in A} :\ \langle a, x^* \rangle \leq 1\}, \\
A^{\triangledown} &\stackrel{\text{def}}{=} \{x^* \in L^* \mid \forall_{a \in A} :\ \langle a, x^* \rangle \geq 1\}, \\
A^{+} &\stackrel{\text{def}}{=} \{x^* \in L^* \mid \forall_{a \in A} :\ \langle a, x^* \rangle \geq 0\}, \\
A^{-} &\stackrel{\text{def}}{=} \{x^* \in L^* \mid \forall_{a \in A} :\ \langle a, x^* \rangle \leq 0\}.
\end{aligned}
\tag{2.1}
$$

These are called the *polar*, *anti-polar*, *dual cone*, and *negative dual cone* of A respectively. The *barrier cone* of A is the set

$$
\mathrm{bc}(A) \stackrel{\text{def}}{=} \{x^* \in L^* \mid \forall_{a \in A} :\ \langle a, x^* \rangle < \infty\}. \tag{2.2}
$$

When A is convex, its *normal cone* is a mapping $\mathrm{N}_A : L \rightrightarrows L^*$ defined by

$$
\forall_{x \in A} :\ \mathrm{N}_A(x) \stackrel{\text{def}}{=} \{x^* \in L^* \mid \forall_{a \in A} :\ \langle a - x, x^* \rangle \leq 0\}, \tag{2.3}
$$

and by convention $\mathrm{N}_A(x)$ is empty for $x \notin A$.

2.2.1 Ordered vector spaces

When there is an order relation $\geq$ on L that is compatible with the algebraic vector space structure,

$$
\forall_{x,y,z \in L} \forall_{t > 0} :\ x \geq y \implies tx + z \geq ty + z, \tag{2.4}
$$

the pair $(L, \geq)$ is called an *ordered vector space*. The *positive cone* is the set $L_{\geq 0} \stackrel{\text{def}}{=} \{x \in L \mid x \geq 0\}$, so that

$$
\forall_{x,y \in L} :\ x \geq y \iff x - y \in L_{\geq 0}. \tag{2.5}
$$

The relation $\geq$ is reflexive and transitive if and only if $L_{\geq 0}$ is a convex cone. Equivalently if K is the cone of positive vectors in L, we refer the order relation associated to K (via (2.5)) by $\geq_K$ and the ordered vector space by (L, K). Let $P \subseteq L^*$. Then P induces an order $\geq_{P+}$ on L defined by

$$
\forall_{u,v \in L} :\ u \geq_{P+} v \iff \forall_{x^* \in P} :\ \langle u, x^* \rangle \geq \langle v, x^* \rangle, \tag{2.6}
$$

and the positive cone $L_{\geq 0}$ satisfies $L_{\geq 0} = P^+$, justifying the notation $\geq_{P^+}$. To see this, observe

$$v \in L_{\geq 0} \iff v \geq_{P^+} 0 \overset{(2.6)}{\iff} \forall_{x^* \in P} : \langle v, x^* \rangle \geq 0 \iff v \in P^+.$$

Proposition 2.1. *Suppose $L \subseteq \mathscr{L}_0(\Omega, \mathbb{R})$ with a topology so that $\mathfrak{P}(\Omega) \subseteq L^*$. The ordering induced by $\mathfrak{P}(\Omega)$ is the usual pointwise ordering.*

Proof. Let $u, v \in L$ satisfy $u(\omega) \geq v(\omega)$ for all $\omega \in \Omega$ then immediately $u \geq_P v$. Next assume $u, v \in L$ satisfies $u \geq_P v$. Then for every Dirac measure δ_ω there is $\langle u, \delta_\omega \rangle \geq \langle v, \delta_\omega \rangle$, or equivalently $u(\omega) \geq v(\omega)$. ■

Remark 2.2. The inclusion condition of Proposition 2.1 is trivially verified in the case where Ω is finite. When Ω is uncountable, more care is needed to ensure the action of the Dirac measures are continuous with the topology on L, such as requiring L consist of a set of bounded, measurable functions with the sup norm.

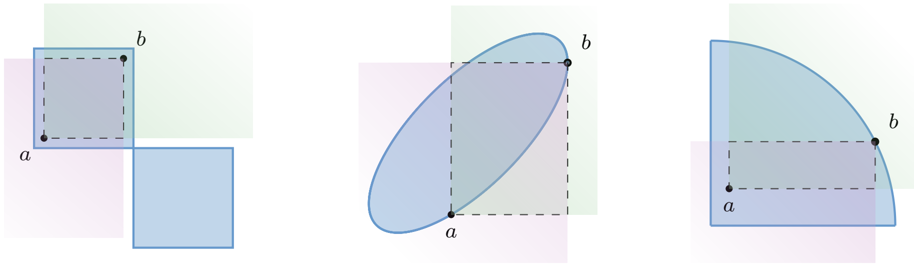
A K-*order interval* joining $a, b \in L$, is the set

$$[a, b]_K \overset{\text{def}}{=} \{x \in L \mid a \leq_K x \leq_K b\}.$$

A set $A \subseteq L$ is said to be K-*full* (or simply full when the order is unambiguous) if $[a, b]_K \subseteq L$ for all $a, b \in A$. The order interval admits a convenient formula

$$[a, b]_K = (a + K) \cap (b - K) \quad \text{and} \quad A = (A + K) \cap (A - L),$$

both of which are simple to derive from (2.5) when A is full. By a full subset of $\mathbb{R}^n$, unless otherwise noted, we mean it is full with respect to the pointwise order. That is, $\mathbb{R}^k_{\geq 0}$-full. Every order interval is a (possibly empty) convex set. The full hull of a convex set is convex. The order interval and the relationship between fullness and convexity is illustrated in Figure 2.1. Finally we say that an *order unit* of K is some point $e \in K$ so that for any $x \in K$ there exists $c > 0$ with $ce \geq x$. The order units of a cone are precisely the points of its relative interior [3, Lem. 1.7].



(a) Not convex, full (b) Convex, not full (c) Convex, full

Figure 2.1: Pictured are three sets and the order interval $[a,b]_{\mathbb{R}^2_{\geq 0}}$ (dashed region) joining two points a, b belonging to each set (blue). The two shaded regions are the sets $a + \mathbb{R}^2_{\geq 0}$ (green) and $b - \mathbb{R}^2_{\geq 0}$ (purple).

Topologies on ordered vector spaces

In a topological vector space $(L, \mathscr{T})$, the vector space operations are assumed compatible with the topology. Analogously, there is a convention in which the order relation is can be compatible with the topology. A convex, proper cone $K \subseteq L$ is said to be $\mathscr{T}$*-normal* if the topology on L has a base at zero consisting of K-full sets [3]. A cone is *weakly normal* ($\sigma(L, L^*)$-normal) if it is normal for $\sigma(L, L^*)$ (consequentially every normal cone is $\sigma(L, L^*)$-normal) [cf. 3, Thm. 2.26, Lem. 2.28]. When L is finite dimensional, every closed cone is normal [3, Lem. 3.1].

Similarly given a cone $K \subseteq L$, the K*-order topology* on L is denoted $\tau_{\geq}(K)$, which is the strongest locally convex topology on K on which every K-order interval is bounded. If $K \subseteq L$ is a cone then K is $\mathscr{T}$-normal if and only if $\mathscr{T} \subseteq \tau_{\geq}(K)$ [3, Lem. 6.27].

2.2.2 Asymptotic cones

The following is standard [cf. 14, 78, 100, 105, 110, 122, 147–149]. The *asymptotic cone* of $A \subseteq L$, is the set

$$A_\infty \stackrel{\text{def}}{=} \left\{ a \in L \;\middle|\; \exists_{(t_i)_{i\in I} \subseteq \mathbb{R}_{>0}} \exists_{(a_i)_{i \in I} \subseteq A} :\; t_i \to 0,\; t_i a_i \to a \right\}, \tag{2.7}$$

denoted A_∞ [37]. It has been used extensively to study the asymptotic properties of unbounded sets. If

$$A_\infty = \left\{ a \in L \;\middle|\; \forall_{(t_i)_{i\in I} \subseteq \mathbb{R}_{>0}} \exists_{(a_i)_{i \in I} \subseteq A} :\; t_i \to 0,\; t_i a_i \to a \right\},$$

then A is said to be *asymptotically regular*. For a scalar $c \in \mathbb{R}$ and an interval $I \subseteq \mathbb{R}_{\geq 0}$ define

$$c \star A \stackrel{\text{def}}{=} \begin{cases} \{ca \mid a \in A\} & c \neq 0 \\ A_\infty & c = 0. \end{cases} \quad \text{and} \quad I \star A \stackrel{\text{def}}{=} \bigcup_{c \in I} c \star A. \tag{2.8}$$

With this convention if A is closed, so is $[0,1) \star A$. When A is bounded $0 \star A = \{0\}$ as usual. We collect some standard results on asymptotic cones.

2.2.3 Extended real arithmetic

Since the extended real numbers are not a group in the algebraic sense, certain conventions turn out to be convenient depending on the purpose. Below, the operations $\cdot_{\mathsf{e}}$ and $+_{\mathsf{e}}$ are common when working with convex functions, the subscript coming from epigraph; and the operations $\cdot_{\mathsf{h}}$ and $+_{\mathsf{h}}$ are common when working with concave function, the subscript coming from hypograph.[1]

We adopt the same conventions as Ward [137] and Zălinescu [146, 150], namely the operations:

$$\begin{aligned} 0 \cdot_{\mathsf{e}} (+\infty) &\stackrel{\text{def}}{=} (+\infty) \cdot_{\mathsf{e}} 0 \stackrel{\text{def}}{=} +\infty, & 0 \cdot_{\mathsf{e}} (-\infty) &\stackrel{\text{def}}{=} (-\infty) \cdot_{\mathsf{e}} 0 \stackrel{\text{def}}{=} 0, \\ 0 \cdot_{\mathsf{h}} (+\infty) &\stackrel{\text{def}}{=} (+\infty) \cdot_{\mathsf{h}} 0 \stackrel{\text{def}}{=} 0, & 0 \cdot_{\mathsf{h}} (-\infty) &\stackrel{\text{def}}{=} (-\infty) \cdot_{\mathsf{h}} 0 \stackrel{\text{def}}{=} -\infty; \end{aligned}$$

and

$$\begin{aligned} (-\infty) +_{\mathsf{e}} (+\infty) &\stackrel{\text{def}}{=} (+\infty) +_{\mathsf{e}} (-\infty) \stackrel{\text{def}}{=} +\infty, \\ (-\infty) +_{\mathsf{h}} (+\infty) &\stackrel{\text{def}}{=} (+\infty) +_{\mathsf{h}} (-\infty) \stackrel{\text{def}}{=} -\infty; \end{aligned}$$

with $\cdot_{\mathsf{e}}$, $\cdot_{\mathsf{h}}$ (resp. $+_{\mathsf{e}}$, $+_{\mathsf{h}}$) agreeing with usual scalar multiplication (resp. scalar addition) in all other situations. The operations $-_{\mathsf{e}}$ and $-_{\mathsf{h}}$ are defed similarly, with $a -_{\mathsf{e}} b \stackrel{\text{def}}{=} a +_{\mathsf{e}} (-b)$ and $a -_{\mathsf{h}} b \stackrel{\text{def}}{=} a +_{\mathsf{h}} (-b)$ for $a, b \in \bar{\mathbb{R}}$.

Remark 2.3. Under these conventions, for a convex function $f : L \to \bar{\mathbb{R}}$, there

[1] Weidner [138] provides an excellent further discussion on the problems of arithmetic with the extended reals.

is [cf. 137, p. 522]

$$\forall_{x \in \operatorname{dom} \partial f} : \ \partial(0 \cdot_{\mathsf{e}} f)(x) = \mathrm{N}_{\operatorname{dom} f}(x) = (\partial f(x))_{\infty} = 0 \star \partial f(x),$$

with our asymptotic set multiplication convention (2.8).[2]

2.3 Minkowski duality

The following summarises a set of well-known results on the operations in (2.1) [2, 99, 100, 149, 150]. For a nonempty set $A \subseteq L$ there are the following polar and bipolar results:

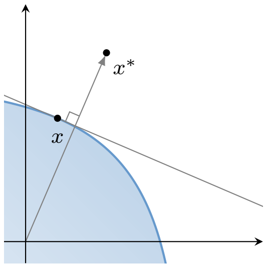
$$\begin{aligned} A^{\circ} &= (\overline{\mathrm{co}}((0,1] \star A))^{\circ}, & A^{\circ\circ} &= \overline{\mathrm{co}}((0,1] \star A), \\ A^{\triangledown} &= (\overline{\mathrm{co}}([1,\infty) \star A))^{\triangledown}, & A^{\triangledown\triangledown} &= \overline{\mathrm{co}}([1,\infty) \star A), \\ A^{+} &= (\overline{\mathrm{co}}((0,\infty) \star A))^{+}, & A^{++} &= \overline{\mathrm{co}}((0,\infty) \star A), \\ A^{-} &= (\overline{\mathrm{co}}((0,\infty) \star A))^{-}, & A^{--} &= \overline{\mathrm{co}}((0,\infty) \star A). \end{aligned} \tag{2.9}$$

When A is a cone $A^{-} = A^{\circ}$ and $A^{+} = A^{\triangledown}$. This motivates the introduction of some classes of sets. The set A is called *radiant* if $(0,1] \star A \subseteq A$, and *co-radiant* if $[1,\infty) \star A \subseteq A$. If A is radiant then it is *star-shaped* if $0 \in A$, and *co-star-shaped* if A is co-radiant with $0 \notin A$ [cf. 97, 109, 111, 122, 144, 145]. The co-radiant and co-star-shaped sets are so named because they are the complements of radiant and co-radiant sets respectively. Clearly if A is radiant (resp. co-radiant) then so is A° (resp. $A^{\triangledown}$).

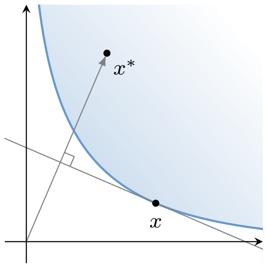
It's well known that radiant sets, convex sets, and cones are asymptotically regular [e.g. 106, Thm 8.2, 14, Prop. 2.15, 147, Prop. 2.1, 110, §5]. And the convention (2.8) is common when working with star-shaped or co-star-shaped sets [cf. 100, 108, 117, 122].

---

[2]The operator $f \mapsto \partial_{\infty} f \stackrel{\text{def}}{=} (\partial f(\,\cdot\,))_{\infty}$ is also known as the *asymptotic subdifferential* [99, p. 235].



(a) Support function



(b) Co-support function

Figure 2.2: (a) A radiant set $A \subseteq \mathbb{R}^2$ (extending infinitely south west) together with two points satisfying $x \in \partial\sigma_A(x^*)$. (b) A co-radiant set $B \subseteq \mathbb{R}^2$ (extending infinitely north east) together with two points satisfying $x \in \widehat{\partial}\zeta_B(x^*)$.

2.3.1 The support and gauge

To a set $A \subseteq L$ we associate the functions $\sigma_A, \zeta_A : L^* \to \bar{\mathbb{R}}$ and $\mu_A, \nu_A : L \to \bar{\mathbb{R}}$, with [cf. 100, 150]

$$\begin{gathered} \sigma_A(x^*) \stackrel{\text{def}}{=} \sup_{s\in A}\langle s, x^*\rangle, \quad \zeta_A(x^*) \stackrel{\text{def}}{=} \inf_{s\in A}\langle s, x^*\rangle; \\ \mu_A(x) \stackrel{\text{def}}{=} \inf\{c \geq 0 \mid x \in c \star A\}, \quad \nu_A(x) \stackrel{\text{def}}{=} \sup\{c \geq 0 \mid x \in c \star A\}, \end{gathered} \tag{2.10}$$

where σ_A is the familiar *support* and the *co-support*, ζ_A is easily identified with $-\sigma_{-S}$. The function μ_A is related to the (Minkowski) *gauge* of A, and ν_A has likewise been related to what has been called Minkowski co-gauge. For every set A the functions μ_A and ν_A are positively homogeneous. When A is convex μ_A and $-\nu_A$ are convex. The subdifferentials of the support and co-support functions are illustrated in Figure 2.2.

Remark 2.4. The exact definition and convention we use comes from Penot and Zălinescu [100] who conduct a thorough study comparing (2.10) to their more classical counterparts. Suffice to say that when A (resp. B) is closed radiant (resp. co-radiant) the function μ_A (resp. ν_B) is equal to the Minkowski gauge (resp. co-gauge). This is summarised in Proposition 2.6 below.

We also define the *indicator*, $\iota_A(x)$, which is ∞ when $x \in A$ and 0 otherwise. The significance of the barrier cone (2.2) is clear in light of

(2.10):

$$\operatorname{bc}(A) = \operatorname{dom}\sigma_A \quad \text{and} \quad -\operatorname{bc}(A) = \operatorname{dom}\zeta_A\,.$$

The normal cone (2.3) allows to invert some important subdifferentials.

Lemma 2.5. *Let $A \subseteq L^*$. Then for all $x \in A$,*

(i) $\mathrm{N}_{\overline{\mathrm{co}}\,A}(x) = (\partial\sigma_A)^{-1}(x)$,

(ii) $-\mathrm{N}_{\overline{\mathrm{co}}\,A}(x) = (\widehat{\partial}\zeta_A)^{-1}(x)$.

Proof. (i) follows from the Young–Fenchel relation [99, Thm. 3.47] observing that $\sigma_A^* = \iota_A^{**} = \iota_{\overline{\mathrm{co}}\,A}$. To invert the co-support (concave) subdifferential note that $\zeta_A(x) = \sigma_{-A}(-x)$ and $\partial(\sigma_{-A}(-\,\cdot\,)) = -\partial\sigma_{-A}$, thus $(-\partial\sigma_{-A})^{-1}(x) = \mathrm{N}_{-\overline{\mathrm{co}}\,A}(-x)$. Let $x \in A$. There is $x^* \in \mathrm{N}_{-A}(-x)$ if and only if

$$\forall_{a\in A} : \Big[\langle -a + x, x^*\rangle \leq 0 \iff \langle a - x, -x^*\rangle \leq 0\Big] \iff x^* \in -\mathrm{N}_A(x).$$

This shows that $\mathrm{N}_{-A}(-x) = -\mathrm{N}_A(x)$ and (ii) follows. ■

The following proposition collects results that are immediate to derive or appear directly in Penot and Zălinescu [100, Props. 2.3, 2.4] and Rubinov [110, §2.9].

Proposition 2.6. *Let $A, B \subseteq L$, $\lambda \in \mathbb{R}_{\geq 0}$. Then*

(i) $\operatorname{dom}\mu_A = [0,\infty) \star A$,

(ii) $\operatorname{lev}_{=0}\mu_A = A_\infty$,

(iii) $\operatorname{lev}_{>0}\mu_A = \operatorname{pos} A \setminus A_\infty$,

(iv) $\mu_A = \mu_{(0,1]\star A}$, $\overline{\mu_A} = \mu_{\mathrm{cl}\,A}$,

(v) $[0,\lambda] \star A \subseteq \operatorname{lev}_{\leq\lambda}\mu_A \subseteq [0,\lambda] \star \overline{A}$,

(vi) $\operatorname{lev}_{<1}\mu_A \subseteq [0,1] \star A \subseteq \operatorname{lev}_{\leq 1}\mu_A$,

(vii) $A \subseteq B$ *if and only if* $\mu_B \leq \mu_A$;

and

(viii) $\operatorname{dom}\nu_A = [0,\infty) \star A$,

(ix) $\operatorname{lev}_{=0}\nu_A = A_\infty \setminus \operatorname{pos} A$,

(x) $\operatorname{lev}_{>0}\nu_A = \operatorname{pos} A$,

(xi) $\nu_A = \nu_{[1,\infty)\star A}$, $\underline{\nu_A} = \nu_{\mathrm{cl}\,A}$,

(xii) $[\lambda,\infty)\star A \subseteq \operatorname{lev}_{\geq\lambda}\nu_A \subseteq [\lambda,\infty)\star\overline{A}$,

(*xiii*) $\operatorname{lev}_{>1} \nu_A \subseteq [1,\infty) \star A \subseteq \operatorname{lev}_{\geq 1} \nu_A$;

(*xiv*) $A \subseteq B$ *if and only if* $\nu_A \leq \nu_B$.

The polarity operations (2.9) induce a duality between the support/co-support and gauge/co-gauge functions (2.10) which is known as *Minkowski duality* [75, 110]. The following is standard or follows immediately from the bipolar theorem (2.9), [110, Prop. 7.27, 100, Lem. 4.1].[3] For a nonempty $A \subseteq L$ there is

$$\begin{aligned} (\mu_A)^* &= \iota_{A^\circ}, & \sigma_{A^\circ} &= \overline{\mu_A}, & \mu_{A^\circ} &= \sigma_{(0,1]\star A}, & A^\circ &= \partial \mu_A(0); \\ (-\nu_A)^* &= \iota_{-A^\triangledown}, & \zeta_{A^\triangledown} &= \underline{\nu_A}, & \nu_{A^\triangledown} &= \zeta_{[1,\infty)\star A}, & A^\triangledown &= \widehat{\partial} \nu_A(0). \end{aligned} \tag{2.11}$$

When A is closed radiant, $\sigma_A = \mu_{A^\circ}$ and $\sigma_{A^\circ} = \mu_A$; and when A is closed co-radiant, $\zeta_A = \nu_{A^\triangledown}$ and $\zeta_{A^\triangledown} = \nu_A$. These identities are summarised in Figure 2.3. If $(L, |\cdot|)$ is a normed space then $|\cdot| = \mu_B$ where B is the closed unit ball in L, or equivalently $|\cdot| = \sigma_{B^\circ}$, and B° coincides with the unit ball in the dual space L^*.

Proposition 2.7 (Minkowski Duality)**.** *Assume $A, B \subseteq L$ are nonempty and closed, with A radiant and B co-radiant. Then:*

(*i*) *for* $(x, x^*) \in \operatorname{pos}(A \times A^\circ) \setminus (A_\infty \times A^-)$

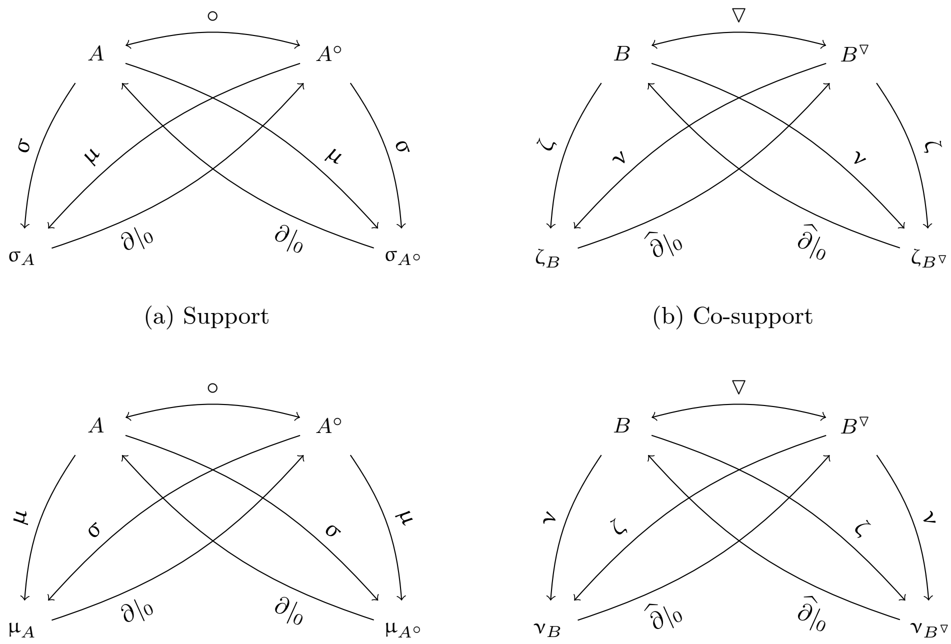
$$\begin{aligned} \frac{x}{\sigma_{A^\circ}(x)} \in \partial \sigma_A(x^*) &\iff \frac{x^*}{\sigma_A(x^*)} \in \partial \sigma_{A^\circ}(x) \\ &\iff \sigma_{A^\circ}(x)\, \sigma_A(x^*) = \langle x, x^* \rangle; \end{aligned}$$

(*ii*) *for* $(x, x^*) \in \operatorname{pos}(B \times B^\triangledown)$

$$\begin{aligned} \frac{x}{\zeta_{B^\triangledown}(x)} \in \widehat{\partial} \zeta_B(x^*) &\iff \frac{x^*}{\zeta_B(x^*)} \in \widehat{\partial} \zeta_{B^\triangledown}(x) \\ &\iff \zeta_{B^\circ}(x)\, \zeta_B(x^*) = \langle x, x^* \rangle. \end{aligned}$$

Proof. Since A is closed and radiant $\sigma_{A^\circ} = \mu_A$ and $\operatorname{lev}_{>0} \sigma_{A^\circ} = \operatorname{lev}_{>0} \mu_A = \operatorname{pos} A \setminus A_\infty$ (Prop. 2.6 (iii)). Similarly $\operatorname{lev}_{>0} \sigma_A = \operatorname{lev}_{>0} \mu_{A^\circ} = \operatorname{pos}(A^\circ) \setminus$

[3] Compare our definition of the co-gauge with that of Barbara and Crouzeix [15] in light of (2.11). In essence, the co-gauge of Barbara and Crouzeix is the upper semicontinuous closure of the co-gauge as defined here. For further discussion on the differences here see Penot and Zălinescu [100] and Zaffaroni [145, §5].



(a) Support (b) Co-support

(c) Gauge (d) Co-gauge

Figure 2.3: Summary of the polar relationships in (2.11) when $A, B \subseteq L$ are closed convex, with A radiant and B co-radiant.

$(A^\circ)_\infty$. Using the fact that the polar of a radiant set is closed and radiant, the formula for the asymptotic cone of a closed radiant set gives

$$(A^\circ)_\infty \overset{\text{P 2.9 (i)}}{=} \bigcap_{\epsilon>0} \epsilon \star A^\circ = \left(\bigcup_{\epsilon>0} \frac{1}{\epsilon} \star A\right)^\circ = (\operatorname{pos} A)^\circ = (\operatorname{pos} A)^- = A^-.$$

Therefore $(x, x^*) \in \operatorname{pos}(A \times A^\circ) \setminus (A_\infty \times A^-)$ if and only if $\sigma_{A^\triangledown}(x) > 0$ and $\sigma_A(x^*) > 0$. By a similar argument, because B is closed and co-radiant $\zeta_{B^\triangledown} = \nu_B$ and $\operatorname{lev}_{>0} \zeta_{B^\triangledown} = \operatorname{lev}_{>0} \nu_B = \operatorname{pos} B$ (Prop. 2.6 (x)). Likewise $\operatorname{lev}_{>0} \zeta_B = \operatorname{pos}(B^\triangledown)$. Therefore $(x, x^*) \in \operatorname{pos}(B \times B^\triangledown)$ if and only if $\zeta_{B^\triangledown}(x) > 0$ and $\zeta_B(x^*) > 0$.

Assume $\frac{x}{\sigma_{A^\circ}(x)} \in \partial\sigma_A(x^*)$. Then

$$\sigma_A(x^*) = \left\langle \frac{x}{\sigma_{A^\circ}(x)}, x^* \right\rangle \quad \text{and} \quad \sigma_{A^\circ}(x) = \left\langle x, \frac{x^*}{\sigma_A(x^*)} \right\rangle.$$

This shows $\frac{x^*}{\sigma_A(x^*)} \in \partial\sigma_{A^\circ}(x)$. By symmetry there is the necessary condition, and an identical argument, mutatis mutandis, yields the corresponding co-support result. ■

Barbara and Crouzeix state a similar result to Proposition 2.7 in a reflexive Banach space [15, Thm. 3.1]. However generalising the Minkowski duality theory to a locally convex space is straight forward (as illustrated by the proof of Proposition 2.7) and simplifies the exposition of the analysis in Section 4.1. The definitions of the co-support and co-gauge we chose also greatly simplifies the sufficient conditions compared with Barbara and Crouzeix, whose definition of the co-gauge corresponds approximately to taking the upper semicontinuous closure of the co-gauge [cf. 100, Prop. 2.3, 2.4, 146, Prop. 1(iii)].

Recall a subset of a topological vector space $A \subseteq L$ is said to be *bounded* if for every neighbourhood of zero $V \in \mathcal{N}(0)$ there is $t_0 > 0$ so that $A \subseteq tV$ for all $t \geq t_0$. A set is $\sigma(L, L^*)$-bounded precisely when it is bounded.

Proposition 2.8. *Let $A, B \subseteq L$ be nonempty. Then*

(*i*) $A_\infty = \bigcap_{\epsilon>0} \overline{(0, \epsilon] \star A}$

(*ii*) A_∞ *is a closed cone,* $A_\infty = (\operatorname{cl} A)_\infty$*, and there is always* $A_\infty \subseteq \overline{\operatorname{pos}} A$

(*iii*) *if* $A \subseteq B$*, then* $A_\infty \subseteq B_\infty$*, if* A *is convex then so is* A_∞

(*iv*) $\{v \in L \mid A + \mathbb{R}_{>0} \star v \subseteq A\} \subseteq A_\infty \subseteq \{v \in L \mid A + \mathbb{R}_{>0} \star v \subseteq \overline{\operatorname{co}} A\}$

(*v*) $A_\infty = \{0\}$ *if* A *is bounded*

(*vi*) $\operatorname{bc}(A)^- = (\operatorname{co} A)_\infty$ *and* $\overline{\operatorname{bc}}(A) = (\operatorname{co} A)_\infty^- \stackrel{\text{def}}{=} ((\operatorname{co} A)_\infty)^-$.

(*vii*) *Let* $A_i \subseteq L$ *for* $i \in I$,

- *i.)* $(\bigcap_{i\in I} A_i)_\infty \subseteq \bigcap_{i\in I} (A_i)_\infty$*, when* $\bigcap_{i\in I} A_i \neq \emptyset$*, with equality when each* A_i *is convex*
- *ii.)* $(\bigcup_{i\in I} A_i)_\infty \supseteq \bigcup_{i\in I} (A_i)_\infty$

(*viii*) *If* A *is asymptotically regular, then* $(A + B)_\infty \supseteq A_\infty + B_\infty$.

These are mostly standard results, and only use the definition (2.7). We provide either references or direct proofs.

Proof. (i): This is well known to the extent that sometimes $\bigcap_{\epsilon>0} \overline{(0, \epsilon] \star A}$ is used for the definition of A_∞ [16, Rem. 1.56, 147, Prop. 1(i)].

(ii): Let $a \in A_\infty$. Then there exists $(a_i)_{i\in I} \subseteq A$ and a convergent net $(t_i)_{i\in I} \subseteq \mathbb{R}_{>0}$ with $t_i \to 0$ so that $t_i a_i \to a$. Thus $a \in \overline{\operatorname{pos}} A$. The other claims are straight-forward.

(iii): Immediate.

(iv): Let $v \in L$ satisfy $A + \mathbb{R}_{\geq 0} \star v \subseteq A$. Then for all $t > 0$, and all $a \in A$ there is $a + tv \in A$. Take $a_i \overset{\text{def}}{=} a + \frac{1}{t_i} v$ for a net $(t_i)_{i\in I} \subseteq \mathbb{R}_{>0}$ with $t_i \to 0$, and $(a_i)_{i\in I} \subseteq A$. Then $\lim_{i\in I} t_i a_i = v$, and thus $v \in A_\infty$. Now assume $v \in A_\infty$. Then there exists $(t_i)_{i\in I} \subseteq \mathbb{R}_{>0}$ with $t_i \to 0$ and $(v_i) \subseteq A$ with $(t_i v_i)_{i\in I} \to v$. Choose any $a \in A$ and let $a_i \overset{\text{def}}{=} (1 - t_i)a + t_i v_i$. Then for a cofinal subnet $(a_j)_{j\in J} \subseteq \operatorname{co} A$ and $\lim_{j\in J} a_j = a + v$. Thus $a + v \in \overline{\operatorname{co}}\, A$.

(v): Let $a \in A_\infty$. Then there exists $(a_i)_{i\in I} \subseteq A$ and a convergent net $(t_i)_{i\in I} \subseteq \mathbb{R}_{>0}$ with $t_i \to 0$ so that $t_i a_i \to a$. If A is bounded then for every neighbourhood of zero $V \in \mathcal{N}(0)$ there exists $t_V > 0$ so that $A \subseteq tV$ for all $t \geq t_V$. Pick an arbitrary $V \in \mathcal{N}(0)$. As $t_i \to 0$ eventually $1/t_i \geq t_V$ for i in some cofinal $I_V \subseteq I$. Thus $t_i a_i \in t_i \cdot A \subseteq V$ for all $i \in I_V$. This shows that for every $V \in \mathcal{N}(0)$,there is a subnet $(t_i a_i)_{i\in I_V}$ that lies entirely within V. Then because $\bigcap_{V\in\mathcal{N}(0)} V = \{0\}$ we have $t_i a_i \to 0$.

(vi): The first claim is well-known [see e.g. 136, p. 142, 14, Thm. 2.2.1, 61, p. 868, 65, Prop. 2.2.4], the second follows from the bipolar theorem.

(vii): From (i) there is

$$\Big(\bigcap_{i\in I} A_i\Big)_\infty = \overline{\bigcap_{\epsilon>0}(0,\epsilon] \cdot \bigcap_{i\in I} A_i} \quad \text{and} \quad \Big(\bigcup_{i\in I} A_i\Big)_\infty = \overline{\bigcap_{\epsilon>0}(0,\epsilon] \cdot \bigcup_{i\in I} A_i}.$$

Therefore

$$\Big(\bigcap_{i\in I} A_i\Big)_\infty = \overline{\bigcap_{\epsilon>0} \bigcup_{t\in(0,\epsilon]} \bigcap_{i\in I} t \star A_i} \subseteq \bigcap_{i\in I} \overline{\bigcap_{\epsilon>0} \bigcup_{t\in(0,\epsilon]} t \star A_i} = \bigcap_{i\in I} (A_i)_\infty,$$

and

$$\Big(\bigcup_{i\in I} A_i\Big)_\infty = \overline{\bigcap_{\epsilon>0} \bigcup_{i\in I} (0,\epsilon] \star A_i} \supseteq \overline{\bigcup_{i\in I} \bigcap_{\epsilon>0} (0,\epsilon] \star A_i} = \overline{\bigcup_{i\in I} (A_i)_\infty} \supseteq \bigcup_{i\in I} (A_i)_\infty.$$

The equality result is standard, and uses the asymptotic regularity of convex sets [e.g. 14, Prop. 2.1.9].

(viii): [148, p. 215, 147, Prop. 2.1, 122, Thm. 3.4] ■

Under certain assumptions, the asymptotic cones have nice representations.

Proposition 2.9. *Let $A, B, C, K \subseteq L$ with A radiant, B co-radiant, C convex, and K a cone. Then* (i) $A_\infty = \bigcap_{\epsilon>0} \epsilon \cdot \overline{A}$, (ii) $B_\infty = \overline{\operatorname{pos}} B$, (iii) $C_\infty = \bigcap_{\epsilon>0} \epsilon \cdot (\overline{C} - x)$ *for any* $x \in C$, *and* (iv) $K_\infty = \overline{K}$.

These are all fairly standard results. We provide either references or proofs.

Proof. (i): Proved by Shveidel [122, Thm. 2.2] in $\mathbb{R}^n$, and the proof in a topological vector space is the same using nets in place of sequences.

(ii): Pick an arbitrary $b \in B$. When B is co-radiant, for all $t \in (0,1]$ there is $\frac{1}{t} b \in B$. Take a convergent net $(t_i)_{i\in I} \subseteq \mathbb{R}_{>0}$ with $t_i \to 0$. Let $b_i \stackrel{\text{def}}{=} \frac{1}{t_i} b$ whenever $t_i \leq 1$ and otherwise b. Then $(b_i)_{i\in I} \subseteq B$ and $t_i b_i \to b$. This shows $b \in B_\infty$ and $B \subseteq B_\infty$. Since B_∞ is a closed cone $\overline{\operatorname{pos}} B \subseteq B_\infty$ and there is always the reverse inclusion (Prop. 2.8 (ii)).

(iii): Direct consequence of the asymptotic regularity of convex sets [14, p. 27].

(iv): A cone is co-radiant, and the claim follows by an application of (ii). ■

2.4 Some useful results

Throughout the subsequent chapters, there are several results which require some obscure, or unknown lemmas. Since these are not specific to our particular applications, we collect them here.

2.4.1 Closure of the sum

The normality of a cone rules out certain pathologies that may otherwise interfere with analysis on noncompact sets.

Lemma 2.10 is inspired by a result due to Choquet [28, Cor. 16].

Lemma 2.10. *Assume $K \subseteq L$ is a normal cone. Let $(a_i)_{i\in I} \subseteq K$, $(b_i)_{i\in I} \subseteq K$. If $(a_i + b_i)_{i\in I}$ converges weakly, then so do $(a_i)_{i\in I}$ and $(b_i)_{i\in I}$.*

Proof. If either (a_i) and (b_i) fail to converge weakly then there exist functions $a^*, b^* \in L^* \setminus \{0\}$ so that $\langle a_i, a^*\rangle \to \infty$ or $\langle b_i, b^*\rangle \to \infty$. (If (a_i) converges, take a^* to be any element of L^* or likewise for (b_i).) We may assume (by passing to a cofinite subnet, or replacing a^* or b^* with $-a^*$ or $-b^*$, if necessary) that $\langle a_i, a^*\rangle \geq 0$ and $\langle b_i, b^*\rangle \geq 0$ for all $i \in I$. Let $c_i \overset{\text{def}}{=} a_i + b_i$. By hypothesis c_i converges weakly.

Since K is weakly normal and L is locally convex, the dual cone, K^+, is *generating* [3, Lem. 2.29]. That is $L^* = K^+ - K^+$. Equivalently

$$\forall_{x^*\in L^*} \exists_{x^*_+\in K^+} :\; x^* + x^*_+ \in K^+. \tag{2.12}$$

For a^* and b^*, let a^*_+ and b^*_+ be the two vectors that each respectively satisfy (2.12), and $c^* \overset{\text{def}}{=} a^* + a^*_+ + b^* + b^*_+$. Then $c^* \in K^+$ and

$$\begin{aligned}
\langle c_i, c^*\rangle &= \langle a_i + b_i, c^*\rangle \\
&= \langle a_i, a^* + a^*_+ + b^* + b^*_+\rangle + \langle b_i, a^* + a^*_+ + b^* + b^*_+\rangle \\
&= \langle a_i, a^*\rangle + \langle a_i, a^*_+ + b^* + b^*_+\rangle + \langle b_i, b^*\rangle + \langle b_i, a^* + a^*_+ + b^*_+\rangle
\end{aligned}$$

If at least one of (a_i) or (b_i) fails to converge converge weakly,

$$\langle a_i, a^*_+\rangle + \langle b_i, b^*_+\rangle \to \infty,$$

while

$$\forall_{i\in I} :\; \langle a_i, a^*_+ + b^* + b^*_+\rangle + \langle b_i, a^* + a^*_+ + b^*_+\rangle \geq 0,$$

because $a^*_+ + b^* + b^*_+ \in K^+$ and $a^* + a^*_+ + b^*_+ \in K^+$. This would be suboptimal, since we assumed (c_i) converges weakly, yielding a contradiction. ∎

Lemma 2.10 yields a number of immediate corollaries for the closure of the sum of two close sets.

Corollary 2.11. *Assume $K \subseteq L$ is a normal cone. If $A, B \subseteq K$ are $\sigma(L, L^*)$-closed, then so is $A + B$.*

Proof. Let $(c_i)_{i\in I} \subseteq C \stackrel{\text{def}}{=} A+B$ be a convergent net with limit c. Then there are nets $(a_i)_{i\in I} \subseteq A$ and $(b_i)_{i\in I} \subseteq B$. Lem. 2.10 proves that both (a_i) and (b_i) are $\sigma(L, L^*)$-convergent, and so let a and b be their respective limits in this topology. Continuity of addition and the fact that $\sigma(L, L^*)$ is Hausdorff implies that $c = a + b$. Since A and B are $\sigma(L, L^*)$-weakly closed $a \in A$ and $b \in B$. This shows that $c \in C$ and $A + B$ is $\sigma(L, L^*)$-closed. ■

Noting that the closure is equal to the weak closure for convex subsets of separated locally convex spaces [e.g. 112, Thm. 3.12] we obtain the following corollary.

Corollary 2.12. *Assume $K \subseteq L$ is a normal cone. If $A, B \subseteq K$ are closed convex, then so is $A + B$.*

The closure of the sum has been shown with a variety of assumptions [cf. 14, 28], the most common (and restrictive) one being that one of A or B is compact, however this is not sufficient for our purposes.

We call a cone $K \subseteq L$ *proper* when $K \cap (-K) \subseteq \{0\}$ and K is convex.

Corollary 2.13. *Assume L is finite dimensional, and $K \subseteq$ is a closed proper cone. If $A, B \subseteq K$ are closed, then $A + B$ is closed.*

Proof. Immediately the cone K is normal [by 3, Lem. 3.1]. Then Cor. 2.11 shows $A + B$ is closed since L is finite dimensional. ■

2.4.2 Measurable selections

For this section equip the topological space $(L, \mathscr{L})$ with a sigma algebra Σ and assume $(M, \mathscr{M})$ is another topological space. For a set-valued map $F : L \rightrightarrows M$, the *upper inverse* and *lower inverse* at $A \subseteq M$ are

$$F^{\mathsf{up}}(A) \stackrel{\text{def}}{=} \{x \in L \mid F(x) \subseteq A\} \quad \text{and} \quad F^{\mathsf{lw}}(A) \stackrel{\text{def}}{=} \{x \in L\} F(x) \cap A \neq \emptyset.$$

It's convenient to observe $F^{-1}(a) = F^{\mathsf{lw}}(\{a\})$ for all $a \in \operatorname{dom} F$.

We say F is *upper hemicontinuous*[4] at $x \in L$ if for every open $U \in \mathscr{M}$ with $F(x) \subseteq U$, $F^{\mathsf{up}}(U)$ is a neighbourhood of x, and upper hemicontinuous

[4] A word of warning: upper hemicontinuous mappings are called variously: uppers semicontinuous, outer continuous, and outward semicontinuous [cf. 7, 13, 26, 99]. We adopt the terminology and definitions of Aliprantis and Border [2] and Aubin [10].

on $M \subseteq L$ if it is upper hemicontinuous at every $x \in M$. We say F is *closed* if $\mathrm{gr}(F)$ is closed in the product topology $\mathscr{L} \otimes \mathscr{M}$.

If $F^{\mathsf{lw}}(V) \in \Sigma$ for every open $V \in \mathscr{M}$ then F is $(\Sigma, \mathscr{M})$-*weakly measurable*. If $F^{\mathsf{lw}}(V) \in \Sigma$ for every closed $V \in \mathscr{M}$ then F is $(\Sigma, \mathscr{M})$-*measurable*. Finally F is upper hemicontinuous precisely when F^{lw} takes closed sets to closed sets [2, Lems. 17.4]. To summarise, when Σ is the Borel sigma algebra on L, $\mathscr{B}(L)$, if F is upper hemicontinuous then F^{lw} is a closed mapping and F is $(\mathscr{B}(L), \mathscr{M})$-measurable. When it does not cause confusion, for brevity we write $(\mathscr{L}, \mathscr{M})$-measurable to mean $(\mathscr{B}(L), \mathscr{M})$-measurable.

Lemma 2.14 (Moreau [89, 10, Prop. 8])**.** *Suppose that a convex function $f : L \to \bar{\mathbb{R}}$ is $\tau(L, L^*)$-continuous on an open subset U. Then the mapping $\partial f : L \rightrightarrows L^*$ is upper hemicontinuous on U when L^* is supplied with the $\sigma(L^*, L)$ topology.*

The Kuratowski and Ryll-Nardzewski [74] selection theorem is the main tool we use to construct measurable selections in Lemmas 2.16 and 2.18.

Lemma 2.15 (Kuratowski and Ryll-Nardzewski [74, 2, Thm. 18.13])**.** *A weakly measurable correspondence with nonempty closed values from a measurable space into a Polish space admits a measurable selection.*

Lemma 2.16. *Assume L is a separable Fréchet space. Let $f : L^* \to \bar{\mathbb{R}}$ be a lower semicontinuous convex function, $\sigma(L^*, L)$-continuous on an open set $U \subseteq L^*$. Then $\partial f : L^* \rightrightarrows L$ has a $(\sigma(L^*, L), \tau(L, L^*))$-measurable selection on U.*

Proof. By assumption f is $\sigma(L^*, L)$-continuous on the $\sigma(L^*, L)$-open set U, therefore ∂f is

- nonempty and closed on U [via 10, Prop. 7, p. 107], and
- $\sigma(L^*, L)$-upper hemicontinuous on U via Lem. 2.14, as the $\tau(L^*, L)$ is stronger than $\sigma(L^*, L)$ topology.

Since ∂f is upper hemicontinuous it is $(\sigma(L^*, L), \tau(L, L^*))$-measurable. Every measurable set-valued mapping into a metrisable space is weakly measurable [2, Lem. 18.2]. Since L is a Fréchet space and separable it is also a Polish space, and the claim follows from Lem. 2.15. ■

Remark 2.17. In practice it is not hard to find an open set U to satisfy Lemma 2.16. For a Banach space L, a convex $f : L \to \bar{\mathbb{R}}$ is continuous on $\operatorname{int}(\operatorname{dom} f)$ if either 1.) f is lower semicontinuous, or 2.) L is finite dimensional [99, Props. 3.3, 3.4].

Lemma 2.18. *Assume L is a separable Fréchet space whose dual is $\sigma(L^*, L)$-separable. Let $A \subseteq L$ be convex. Then the intersection mapping $\mathrm{N}_A \cap P : L \rightrightarrows P$ has a $(\tau(L, L^*), \sigma(L^*, L))$-measurable selection, for any $\sigma(L^*, L)$-compact $P \subseteq L^*$.*

Proof. Assume L is equipped with its $\tau(L, L^*)$ topology and L^* is equipped with the $\sigma(L^*, L)$ topology.

Let $(x_i)_{i\in I} \subseteq L$ and $(y_i)_{i\in I} \subseteq L^*$ satisfy $(x_i, y_i) \in \operatorname{gr}(\mathrm{N}_A)$ for all $i \in I$ and converge in $\tau(L, L^*) \otimes \sigma(L^*, L)$ with limit (x, y). Then for every $a \in A$ and $i \in I$ there is $0 \geq \langle x_i - a, y_i\rangle$ and $\lim_{i\in I}\langle x_i - a, y_i\rangle = \langle x - a, y\rangle \leq 0$ and $(x, y) \in \operatorname{gr} \mathrm{N}_A$. Thus $\operatorname{gr} \mathrm{N}_A$ is closed and N_A is closed.

The set P is compact by hypothesis and so P is a trivially upper hemicontinuous as a set-valued map $L \rightrightarrows L^*$. The intersection of a closed map with a closed, compact-valued upper hemicontinuous map produces an upper hemicontinuous map [2, Thm 17.25]. Therefore $\mathrm{N}_A \cap P$ is upper hemicontinuous and weakly measurable (reusing the measurability argument from the proof of Lem. 2.16). The subspace topology on P is metrisable from the Banach–Alaoglu–Bourbaki theorem [22, Thm. 3.1.4] and compact, whence $(P, \sigma(L^*, L))$ is a Polish space, and the claim follows from Lem. 2.15. ■

Remark 2.19. When L is a Banach space, separability of L^* implies separability of L [38, Prop. 3.6.14].

2.4.3 The subdifferential of a supremum

In Lemma 2.20 we use a proof by contradiction. Zălinescu [149, Thm. 2.4.14(iii)] provides a constructive proof assuming, additionally, that f is convex.

Lemma 2.20. *Let $f : L \to \bar{\mathbb{R}}$ be positively homogeneous. Then $\partial f(0) = \partial_\epsilon f(0)$ for all $\epsilon \geq 0$.*

Proof. We will show $\partial_\epsilon f(0) = \partial f(0)$ for all $\epsilon > 0$ using a proof by contradiction. Fix $\epsilon > 0$. Suppose $x^* \in \partial_\epsilon f(0) \setminus \partial f(0)$. There exists

$y \in L$ with $f(y) < \langle y, x^* \rangle \leq \epsilon - \nu_m(y)$ whence $y \in \operatorname{dom} f$ and therefore $0 < \langle y, x^* \rangle + f(y) \leq \epsilon$. Let $c \overset{\text{def}}{=} (\epsilon + 1)(\langle y, x^* \rangle + f(y))^{-1} > 0$. Then

$$\langle cy, x^* \rangle + f(cy) = c(\langle y, x^* \rangle + f(y)) = \epsilon + 1 > \epsilon,$$

a contradiction. This shows that $x^* \notin \partial_\epsilon f(0)$. Thus $\partial_\epsilon f(0) = \partial f(0)$. ■

Lemma 2.21 (Hantoute, López, and Zălinescu [61, Cor. 9])**.** *Let $(f_t)_{t\in T}$ be a nonempty arbitrary family of lower semicontinuous convex functions $L \to \bar{\mathbb{R}}$, and set $f \overset{\text{def}}{=} \sup_{t\in T} f_t$. Then f is closed and for all $z \in$ and $\alpha \geq 0$ there is*

$$\partial f(z) = \bigcap_{L\in\mathcal{F}(z)} \bigcap_{\epsilon>0} \operatorname{cl}^* \left(\operatorname{co} \left(\bigcup_{t\in T_\epsilon(z)} \partial_{\alpha\epsilon} f_t(z) \right) + \mathrm{N}_{L\cap \operatorname{dom} f}(z) \right),$$

where $\mathcal{F}(z)$ is the collection of finite-dimension linear subspaces of L containing z, and $T_\epsilon(z) \overset{\text{def}}{=} \{t \in T \mid f_t(z) \geq f(z) - \epsilon\}$.

Remark 2.22. Observe that, with the notation of Lemma 2.21, there is is always

$$\forall_{z\in L}: \ \partial f(z) \supseteq \bigcup_{t\in T_0(z)} \partial f_t(z).$$

Lemma 2.23. *Let $(f_t)_{t\in T}$ be a nonempty arbitrary family of sublinear lower semicontinuous convex functions $L \to \bar{\mathbb{R}}$, and set $f \overset{\text{def}}{=} \sup_{t\in T} f_t$. Then f is lower semicontinuous and sublinear and*

$$\partial f(0) = \overline{\operatorname{co}} \bigcup_{t\in T} \partial f_t(0).$$

Proof. Before we can apply Lem. 2.21 we first need to compute some terms. Since f_t is sublinear, there is $f_t(0) = 0$ for every $t \in T$ and

$$\forall_{\epsilon>0}: \ T_\epsilon(0) = \{t \in T \mid f_t(0) \geq f(0) - \epsilon\} = \{t \in T \mid 0 \geq -\epsilon\} = T.$$

Define the orthogonal complement $A^\perp \overset{\text{def}}{=} \{x^* \in L^* \mid \forall_{a\in A}: \ \langle a, x^* \rangle = 0\}$ of a set $A \subseteq L$ [cf. 149, p. 7, 61, p. 866]. Let $\mathscr{N}_{X^*}(0)$ denote set of convex neighbourhoods of 0 in X^*. Let S be a linear subspace of L satisfying $S^\perp \subseteq V$

for some $V \in \mathcal{N}_{X^*}(0)$. Then

$$\begin{aligned}
\mathrm{N}_{S\cap \operatorname{dom} f}(0) &= (S \cap \operatorname{dom} f)^- \\
&= S^\perp + (\operatorname{dom} f)^- \\
&= S^\perp + \left(\bigcap_{t\in T} \partial f_t(0)\right)_\infty \\
&\subseteq V + \left(\operatorname{co} \bigcap_{t\in T} \partial f_t(0)\right)_\infty .
\end{aligned} \tag{2.13}$$

Lem. 2.21 yields

$$\begin{aligned}
\partial f(0) &\overset{\mathrm{L}\,2.20}{=} \bigcap_{S\in\mathcal{F}(0)} \mathrm{cl}^* \left(\operatorname{co} \bigcup_{t\in T} \partial f_t(0) + \mathrm{N}_{S\cap \operatorname{dom} f}(0)\right) \\
&\overset{(2.13)}{\subseteq} \bigcap_{V\in\mathcal{N}_{X^*}(0)} \mathrm{cl}^* \left(\operatorname{co} \bigcup_{t\in T} \partial f_t(0) + V + \left(\operatorname{co} \bigcap_{t\in T} \partial f_t(0)\right)_\infty\right) \\
&\overset{\mathrm{P}\,2.8\,(\mathrm{iv})}{=} \bigcap_{V\in\mathcal{N}_{X^*}(0)} \mathrm{cl}^* \left(\operatorname{co} \bigcup_{t\in T} \partial f_t(0) + V\right) \\
&= \overline{\operatorname{co}} \bigcup_{t\in T} \partial f_t(0).
\end{aligned}$$

It's easy to see that reverse inclusion always holds, completing the proof. ■

Chapter 3

Operations on the Families of Radiant and Co-radiant Sets

Throughout this chapter, L is a locally convex Hausdorff topological vector space over the reals. To a set $M \subseteq \mathbb{R}^k$ we associate the two operations

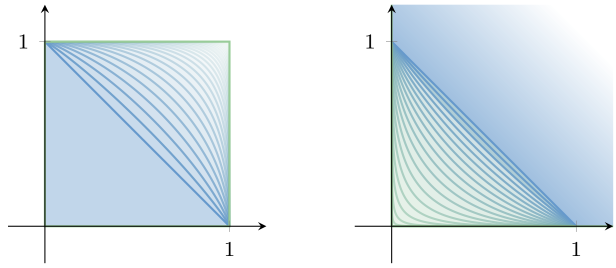
$$\oplus_M, \square_M : \overbrace{2^L \times \cdots \times 2^L}^{k \text{ times}} \to 2^L$$

where, for a sequence of sets $A_1, \ldots, A_k \subseteq L$,

$$\oplus_M(A_1, \ldots, A_k) \stackrel{\text{def}}{=} \bigcup_{m \in M} \sum_{i \in [k]} m_i \star A_i, \tag{3.1}$$

$$\square_M(A_1, \ldots, A_k) \stackrel{\text{def}}{=} \bigcup_{m \in M} \bigcap_{i \in [k]} m_i \star A_i. \tag{3.2}$$

These are called the *M-sum* and *dual M-sum* respectively. For each choice of the set M, they encompass a wide range of operations, most of which have been studied independently in the binary setting, that is, where $k = 2$ [44, 45, 117]. With the exception of Gardner, Hug, and Weil [49], analysis has largely been limited to one several common choices for M, focusing exclusively on subsets $A_i \subseteq \mathbb{R}^k$ for $i \in [k]$ that are often compact and containing the origin. The assumptions made by these previous approaches (summarised in Table 3.1) will prove too restrictive for Chapter 4, and so our goal here is to



(a) $(B_p)_{p\in[1,\infty]}$ (b) $(C_p)_{p\in[1,\infty]}$

Figure 3.2: Illustration of the families $(B_p)_{p\in[1,\infty]}$, $(C_p)_{p\in[1,\infty]}$, when $k=2$. Observe that both families are full and convex, and the sets $(B_p)_{p\in[1,\infty]}$ are star-shaped, whereas $(B_p)_{p\in[1,\infty]}$ are co-star-shaped. In (b) the sets each extend infinitely north-east.

extend a variety of existing results for application to our setting.

Define the following subsets of $\mathbb{R}^k$,

$$\begin{gathered} B_p \overset{\text{def}}{=} \{x \in \mathbb{R}^k_{\geq 0} \mid |x|_p \leq 1\}, \quad C_p \overset{\text{def}}{=} \{x \in \mathbb{R}^k_{\geq 0} \mid |x|_{\frac{p}{p-1}} \geq 1\}, \\ I_1 \overset{\text{def}}{=} \{x \in \mathbb{R}^k_{\geq 0} \mid |x|_1 = 1\}. \end{gathered} \tag{3.3}$$

The set $\{1\}^k \subseteq \mathbb{R}^k_{\geq 0}$ denotes the singleton consisting of a single vector with every element equal to 1. The *harmonic sum* [100] of two sets $A, B \subseteq L$ is

$$A \diamond B \overset{\text{def}}{=} \left(\bigcup_{t\in(0,1)} t \cdot A \cap (1-t) \cdot B \right) \cup A_\infty \cup B_\infty.$$

Remark 3.1. We have $B_1 = [0,1] \cdot_{\mathsf{e}} I_1$ and $C_1 = [1,\infty) \cdot_{\mathsf{e}} I_1$. Consequentially Propositions 2.6 (iv) and 2.6 (xi) show $\mu_{B_1} = \mu_{I_1}$ and $\nu_{C_1} = \nu_{I_1}$.

The vast majority of previous results apply to radiant sets $(A_i)_{i\in[k]}$ with a radiant set M (Table 3.1). Consequentially, the family $(C_p)_{p\in[1,\infty]}$ makes no real appearance in most of the literature mentioned. However since $(C_p)_{p\in[1,\infty]}$ consists of co-radiant sets, it should be no surprise that it is quite relevant when the sets $(A_i)_{i\in[k]}$ are co-radiant. The families $(B_p)_{p\in[1,\infty]}$ and $(C_p)_{p\in[1,\infty]}$ are illustrated in Figure 3.2, and some some special cases of $\oplus_M$ and $\Box_M$ are listed in Table 3.3 using these sets.

Remark 3.2. While there is some inconsistency in the operation for the scalar–set multiplication in (3.1) and (3.2) (Penot and Zălinescu [100] use our

(a) Related works studying the operation $\oplus_M$.

	L	M	k	$A_1, \ldots, A_k$	Functionals
Seeger [117]	locally convex, Hausdorff	$(B_p)_{p\in[1,\infty]}$	2	convex, containing 0	σ
Gardner, Hug, and Weil [49]	$\mathbb{R}^n$	various: compact, convex, radiant, 1-unconditional	k	compact, convex	σ
This work	locally convex, Hausdorff	various: full, convex, radiant, co-radiant	k	various: convex, radiant, co-radiant	σ, μ, ν, ν

(b) Related works studying with the operation $\square_M$.

	L	M	k	$A_1, \ldots, A_k$	Functionals
Firey [43, 44]	$\mathbb{R}^n$	$(B_p)_{p\in[1,\infty]}$	2	compact, convex, containing 0	$\hat{\mu}$
Seeger [117]	locally convex, Hausdorff	$(B_p)_{p\in[1,\infty]}$	2	convex, containing 0	σ
Penot and Zălinescu [100]	locally convex, Hausdorff	I_1	2	various: convex, radiant, co-star-shaped	σ, μ, ζ, ν
This work	locally convex, Hausdorff	various: full, radiant, co-radiant	k	various: convex, radiant, co-radiant	σ, μ, ζ, ν

Table 3.1: A list of related works dealing with the operations $\oplus_M$ and $\square_M$ together with the setting and assumptions. Our results are designed jointly to unify and subsume these works with strict generalisation. The symbol $\hat{\mu}$ denotes the Minkowski gauge.

	M	$\oplus_M(A_1,\dots,A_k)$		M	$\Box_M(A_1,\dots,A_k)$
Minkowski sum	$\{1\}^k$	$A_1+\cdots+A_k$	Intersection	$\{1\}^k$	$A_1\cap\cdots\cap A_k$
Convex hull	I_1	$\mathrm{co}(A_1\cup\cdots\cup A_k)$	Harmonic sum	I_1	$A_1\diamond\cdots\diamond A_k$
Direct sum	B_p	-	Inverse sum	B_p	-

Table 3.3: Some different operations obtained from $\oplus_M$ and $\Box_M$ by choosing different sets M, when the sets $(A_i)_{i\in[k]}$ are bounded [117, Prop. 2.2].

asymptotic multiplication convention (2.8), it is considered by Seeger [117] and Firey [43, 44] and Gardner, Hug, and Weil [49] use classical set–scalar multiplication — in the case where the sets $(A_i)_{i\in[k]}$ are all bounded, for example, when they are compact, the asymptotic multiplication and conventional multiplication coincide. This is a consequence of Proposition 2.8 (v).

The goal of this chapter is to complete the diagram in Figure 2.3 for sets in the image of the operations $\oplus_M$ and $\Box_M$ with respect to the two polarities $\circ,\triangledown$, the two support functions σ,ζ, and the two gauge functions μ,ν. The following support, co-support results are proven in Section 3.1:

$$\oplus_M(A_1,\dots,A_k) \xrightarrow[\text{Theorem 3.4}]{\sigma_{(\,\cdot\,)}} \underset{(3.4)}{\sigma_M(\sigma_{A_1},\dots,\sigma_{A_k})}$$

$$\oplus_M(A_1,\dots,A_k) \xrightarrow[\text{Theorem 3.4}]{\zeta_{(\,\cdot\,)}} \underset{(3.5)}{\zeta_M(\zeta_{A_1},\dots,\zeta_{A_k})}.$$

The companion gauge and co-gauge results are proven next in Section 3.3 after establishing some topological properties of the M-sum and dual M-sum in Section 3.2:

$$\Box_M(A_1,\dots,A_k) \xrightarrow[\text{Theorem 3.19}]{\mu_{(\,\cdot\,)}} \underset{(3.15)}{\mu_M(\mu_{A_1},\dots,\mu_{A_k})}$$

$$\Box_M(A_1,\dots,A_k) \xrightarrow[\text{Theorem 3.20}]{\nu_{(\,\cdot\,)}} \underset{(3.16)}{\nu_M(\nu_{A_1},\dots,\nu_{A_k})}.$$

In Section 3.4 we compute the polarity operations to link the M-sum to the dual M-sum:

$$\Box_M(A_1,\ldots,A_k) \xrightarrow[\text{Theorem 3.26 (iii)}]{(\,\cdot\,)^\circ} \oplus_{M^\circ}(A_1^\circ,\ldots,A_k^\circ)$$
$$\Box_M(A_1,\ldots,A_k) \xrightarrow[\text{Theorem 3.26 (iv)}]{(\,\cdot\,)^\triangledown} \oplus_{M^\triangledown}(A_1^\triangledown,\ldots,A_k^\triangledown),$$

and from the previous polarity results:

$$\oplus_M(A_1,\ldots,A_k) \xrightarrow[(2.9)]{(\,\cdot\,)^\circ} \Box_{M^\circ}(A_1^\circ,\ldots,A_k^\circ)$$
$$\oplus_M(A_1,\ldots,A_k) \xrightarrow[(2.9)]{(\,\cdot\,)^\triangledown} \Box_{M^\triangledown}(A_1^\triangledown,\ldots,A_k^\triangledown).$$

Where, as indicated by the equation references, the second row of arrows follows from the bipolar theorem. In Section 3.5, using the polarity results of the previous section, we compute the gauge and co-gauge functions of the M-sum and dual M-sum:

$$\oplus_M(A_1,\ldots,A_k) \xrightarrow[\text{Theorem 3.28}]{\mu_{(\,\cdot\,)}} \qquad (3.25)$$
$$\Box_M(A_1,\ldots,A_k) \xrightarrow[\text{Theorem 3.28}]{\sigma_{(\,\cdot\,)}} \qquad (3.26),$$

and

$$\oplus_M(A_1,\ldots,A_k) \xrightarrow[\text{Theorem 3.29}]{\nu_{(\,\cdot\,)}} \qquad (3.27)$$
$$\Box_M(A_1,\ldots,A_k) \xrightarrow[\text{Theorem 3.29}]{\zeta_{(\,\cdot\,)}} \qquad (3.28).$$

We conclude with a discussion of some related results in Section 3.6.

3.1 Support functions

For $M \subseteq \mathbb{R}^k$ and $f_1, \dots f_k : L \to \bar{\mathbb{R}}$ let $\sigma_M(f_1, \dots, f_k), \zeta_M(f_1, \dots, f_k) : L \to \bar{\mathbb{R}}$, where

$$\sigma_M(f_1, \dots, f_k) \stackrel{\text{def}}{=} \sup_{m \in M} (m_1 \cdot_{\mathsf{e}} f_1 +_{\mathsf{e}} \cdots +_{\mathsf{e}} m_k \cdot_{\mathsf{e}} f_k), \tag{3.4}$$

$$\zeta_M(f_1, \dots, f_k) \stackrel{\text{def}}{=} \inf_{m \in M} (m_1 \cdot_{\mathsf{h}} f_1 +_{\mathsf{h}} \cdots +_{\mathsf{h}} m_k \cdot_{\mathsf{h}} f_k). \tag{3.5}$$

Therefore

$$\sigma_M(f_1, \dots, f_k) = -\zeta_M(-f_1, \dots, -f_k). \tag{3.6}$$

Occasionally it will be convenient to write (3.4) and (3.5) using a summation symbol, in this case we assume the summation is with respect to the respective addition conventions in (3.4) and (3.5).

Proposition 3.3. *Let $A \subseteq L$ and $m \geq 0$. Then $\sigma_{m \star A} = m \cdot_{\mathsf{e}} \sigma_A$.*

Proof. Since $\sigma_A = \sigma_{\overline{\text{co}}\, A}$, it is without loss of generality that we assume A is convex. We have $\sigma_{m \star A} = m \cdot_{\mathsf{e}} \sigma_A$ when $m \neq 0$. When $m = 0$, $m \star A = A_\infty$ and $\sigma_{m \star A} = \iota_{A_\infty^-}$. It is always the case that $\text{bc}(A) \subseteq A_\infty^-$ (Prop. 2.8(vi)), therefore $\sigma_{m \star A}$ and $m\, \sigma_A$ differ only on the set $L^* \setminus A_\infty^- \ni x^*$, when $\sigma_{m \star A}(x^*) = \infty$ and $m \cdot_{\mathsf{e}} \sigma_A(x^*) = 0$. However $L^* \setminus A_\infty^- \subseteq L^* \setminus \text{bc}(A)$. Therefore $\sigma_A(x^*) = \infty$ for all $x \in L^* \setminus A_\infty^-$. It follows that $\sigma_{m \star A} = m \cdot_{\mathsf{e}} \sigma_A$. ∎

Theorem 3.4. *Let $M \subseteq \mathbb{R}^k_{\geq 0}$ and $A_i \subseteq L$ for $i \in [k]$. Let $A \stackrel{\text{def}}{=} \oplus_M(A_1, \dots, A_m)$. Then*

$$\sigma_M(\sigma_{A_1}, \dots, \sigma_{A_k}) = \sigma_A \quad \textit{and} \quad \zeta_M(\zeta_{A_1}, \dots, \zeta_{A_k}) = \zeta_A\,.$$

Proof. Let $C_{m,i} \stackrel{\text{def}}{=} m_i \star A_i$, $B_m \stackrel{\text{def}}{=} \sum_{i \in [k]} C_{m,i}$ for $m \in M$, $i \in [k]$. Then $A = \bigcup_{m \in M} B_m = \bigcup_{m \in M} \sum_{i \in [k]} C_{m,i}$, and from the usual calculus of support functions [12, p. 31], $\sigma_A = \sup_{m \in M} \sigma_{B_m} = \sup_{m \in M} \sum_{i \in [k]} \sigma_{C_{m,i}}$ and using Prop. 3.3, $\sigma_A = \sup_{m \in M} \sum_{i \in [k]} m_i \cdot_{\mathsf{e}} \sigma_{A_i}$. The claim about ζ_A follows from (3.6), replacing M with $-M$ and A_i with $-A_i$ for $i \in [k]$. ∎

Corollary 3.5. *Let $M \subseteq \mathbb{R}^k_{\geq 0}$ and $A_i \subseteq L$ for $i \in [k]$. Then*

$$\mathrm{bc}(\oplus_M(A_1, \dots, A_m)) = \bigcap_{i \in [k]} \mathrm{bc}(A_i),$$

and $\oplus_M(A_1, \dots, A_k)$ is bounded if and only if each of $M, A_1, \dots, A_k$ are bounded.

3.2 Topological properties

In order to establish similar results to Theorem 3.4 for the gauge and co-gauge function, we first need some results about the topology of the dual M-sum.

3.2.1 Closure

We start by giving some mild conditions under which $\oplus_M$ and $\square_M$ are closed. Proposition 3.6 is simple to derive from Corollary 3.5.

Proposition 3.6. *Suppose $M \subseteq \mathbb{R}^k$ and each of $A_i \subseteq L$, $i \in [k]$ are bounded (resp. $\sigma(L, L^*)$-compact). Then $\oplus_M(A_1, \dots, A_k)$ is bounded (resp. $\sigma(L, L^*)$-compact).*

Proof. If each A_i for $i \in [k]$ is bounded, then Cor. 3.5 implies $\oplus_M(A_1, \dots, A_k)$ is bounded.

Let $(x_i)_{i \in I} \subseteq \oplus_M(A_1, \dots, A_k)$ be a $\sigma(L, L^*)$-convergent net with limit $\bar{x}$. Then there is a net $(m_i)_{i \in I} \subseteq M$ with $x_i \in \sum_{j \in [k]} m_{ij} \star A_j$. Since M is closed and bounded (m_i) may be assumed (possibly by passing to a subnet) to converge, so let $\bar{m}$ be its limit. Then there are nets $(a_{ij})_{i \in I} \subseteq A_j$ for each $j \in [k]$ so that $x_i = \sum_{j \in [k]} m_{ij} a_{ij}$ for every $i \in I$. Since each A_j is $\sigma(L, L^*)$-compact, the nets $(a_{ij})_{i \in I}$ may be assumed to converge with limits $\bar{a}_j$ for each $j \in [k]$. Thus $\bar{x} = \sum_{j \in [k]} \bar{m}_j \bar{a}_j$. This shows $\bar{x} \in \oplus_M(A_1, \dots, A_k)$. ■

There is another result, similar to Proposition 3.6, when the sets $(A_i)_{i \in [k]}$ are unbounded.

Theorem 3.7. *Suppose $L_{\geq 0} \subseteq L$ is a normal cone, and $M \subseteq \mathbb{R}^k_{\geq 0}$ is closed convex, containing an order unit of $\mathbb{R}^k_{\geq 0}$. Suppose M and each of $A_i \subseteq L_{\geq 0}$ for $i \in [k]$ are $\sigma(L, L^*)$-closed. Then $\oplus_M(A_1, \dots, A_k)$ is $\sigma(L, L^*)$-closed.*

Proof. Suppose $(x_i)_{i\in I} \subseteq \oplus_M(A_1,\dots,A_k)$ is a $\sigma(L,L^*)$-convergent net with limit $\bar{x}$. Then there is a net $(m_i)_{i\in I} \subseteq M$ with $x_i \in \sum_{j\in[k]} m_{ij} \star A_j$ for every $i \in I$. Since M is convex and contains an order unit of $\mathbb{R}^k_{\geq 0}$, e, we may assume $m_i \in \mathbb{R}^k_{>0}$ for every $i \in I$. To see this, observe that we can construct another sequence

$$(\epsilon_i e + (1-\epsilon_i)m_i)_{i\in I} \subseteq M \cap \mathbb{R}^k_{>0} \quad \text{with} \quad \epsilon_i e + (1-\epsilon_i)m_i \to \bar{m}, \tag{3.7}$$

where $(\epsilon_i)_{i\in I} \subseteq (0,1)$ is chosen arbitrarily to satisfy $\epsilon_i \to 0$. Because $m_{ij} \star A_j \subseteq L_{\geq 0}$ for every $(i,j) \in I \times [k]$ (via Prop. 2.8(iii)) from (3.7) it follows that there are nets $(a_{ij})_{i\in I} \subseteq A_j \subseteq L_{\geq 0}$ for each $j \in [k]$ so that $x_i = \sum_{j\in[k]} m_{ij} a_{ij}$ for every $i \in I$.

First assume (m_i) converges with limit $\bar{m} \in M$. Since (x_i) and (m_i) converge, Lem. 2.10 shows that so do each of $(a_{ij})_{i\in I}$ for $j \in [k]$. Let $\bar{a}_j$ be the $\sigma(L,L^*)$-limit of $(a_{ij})_{i\in I}$ for $j \in [k]$. Thus $\bar{x} = \sum_{j\in[k]} \bar{m}_j \bar{a}_j$ and $\bar{x} \in \oplus_M(A_1,\dots,A_k)$.

Next assume (m_i) does not converge. We will see this leads to a contradiction. Let $|\cdot|$ be any norm on $\mathbb{R}^n$. Then we define the nets

$$\forall_{i\in I}: \ t_i \overset{\text{def}}{=} |m_i| \quad \text{and} \quad n_i \overset{\text{def}}{=} \frac{1}{t_i} m_i.$$

Then $m_{ij} = t_i n_{ij}$ for $(i,j) \in I \times [k]$. Since (n_i) is bounded, we may assume without loss of generality that it converges with limit $\bar{n}$. Since (m_i) does not converge, we have $t_i \to \infty$. Another application of Lem. 2.10 shows that the nets $(a_{ij})_{i\in I} \subseteq A_j$ converge, with $\sigma(L,L^*)$-limits $\bar{a}_j$ for $j \in [k]$. Thus $\sum_{j\in[k]} \langle n_{ij} a_{ij}, x^* \rangle$ converges in $\sigma(L,L^*)$, whence there exists $x^* \in L^*$ with

$$\langle x_i, x^* \rangle = t_i \sum_{j\in[k]} \langle n_{ij} a_{ij}, x^* \rangle \quad \text{and} \quad t_i \sum_{j\in[k]} \langle n_{ij} a_{ij}, x^* \rangle \to \infty.$$

This contradicts the assumption that (x_i) converges in $\sigma(L,L^*)$, and completes the proof. ∎

Remark 3.8. To our knowledge Theorem 3.7 is the first $\oplus_M$ closure result for unbounded M and unbounded sets $(A_i)_{i\in[k]}$. Seeger [117, Prop. 4.3] proves closure for $k = 2$ when one of the sets $(A_i)_{i\in[k]}$ is bounded. Instead we use

Lemma 2.10 to ensure closure by requiring the sets are all subsets of a normal cone.

Theorem 3.7 will be used to verify the closure of the scoring rule operation we develop in Section 4.4. The use of the $\sigma(L, L^*)$ topology is without loss of generality when the sets $(A_i)_{i\in[k]}$ are convex, since the closure of a convex set in the original topology coincides with the $\sigma(L, L^*)$-closure [112, Thm. 3.12].

Proposition 3.9. *Suppose $M \subseteq \mathbb{R}^k_{\geq 0}$ is closed convex, containing an order unit of $\mathbb{R}^k_{\geq 0}$, and $A_i \subseteq L_{\geq 0}$ for $i \in [k]$ are closed.*

(i) If M is bounded then $\square_M(A_1, \dots, A_k)$ is closed.

(ii) If $A_i \subseteq L_{\geq 0}$ for $i \in [k]$ are bounded and $\sigma(L, L^*)$-compact then $\square_M(A_1, \dots, A_k)$ is $\sigma(L, L^*)$-closed.

Proof. Suppose $(x_i)_{i\in I} \subseteq \oplus_M(A_1, \dots, A_k)$ is a convergent net with limit x. Then by the same argument as Thm. 3.7, in particular (3.7), there are nets $(m_i)_{i\in I} \subseteq M$, $(a_{ij})_{i\in I} \subseteq A_j$ for each $j \in [k]$ so that $x_i = m_{ij}a_{ij}$ for every $(i, j) \in I \times [k]$.

(i): Since M is compact, without loss of generality (m_i) may be assumed to converge in M. Let its limit be $\bar{m} \in M$. Then because (x_i) converges and $x_i = m_{ij}a_{ij}$ for all $i \in I$, necessarily the nets $(a_{ij})_{i\in I}$ converge for $j \in [k]$, let $\bar{a}_j$ be the limit of $(a_{ij})_{i\in I}$ for $j \in [k]$. The sets A_j for $j \in [k]$ are closed, thus $\bar{a}_j \in A_j$ for $j \in [k]$. Thus $x = \bar{m}_j\bar{a}_j \in \square_M(A_1, \dots, A_k)$ and $\square_M(A_1, \dots, A_k)$ is closed.

(ii): Suppose $(x_i)_{i\in I} \subseteq \oplus_M(A_1, \dots, A_k)$ is a $\sigma(L, L^*)$-convergent net with $\sigma(L, L^*)$-limit x. Since A_j is $\sigma(L, L^*)$-compact for all jnk, it is without loss of generality to assume $(a_{ij})_{i\in I}$ converge for $j \in [k]$. Let $\bar{a}_j$ be the limit of $(a_{ij})_{i\in I}$ for $j \in [k]$. Then since (x_i) converges and $x_i = m_{ij}a_{ij}$ for all $(i, j) \in I \times [k]$, the net $(m_i)_{i\in I}$ converges. Let its limit be $\bar{m}$. Because M is closed, $\bar{m} \in M$. Thus $x = \bar{m}_j\bar{a}_j \in \square_M(A_1, \dots, A_k)$ for $j \in [k]$ and $\square_M(A_1, \dots, A_k)$ is $\sigma(L, L^*)$-closed. ∎

Proposition 3.9 (i) essentially uses a straight forward limit argument [cf. 117, Prop. 4.2, 100, Lem. 3.1(b)].

3.2.2 Convexity

We now show that both of the operations $\oplus_M$ and $\Box_M$ preserve convexity when M is convex. Gardner, Hug, and Weil provide similar result for $\oplus_M$ with respect to the domain of compact, convex sets in a finite dimensional space [49, Thm. 6.1], and our proof strategy is essentially the same, with some added care to respect our scalar–set multiplication convention.

Lemma 3.10. *Suppose $(S_i)_{i\in I}$ and $(T_j)_{j\in I}$, are arbitrary families of subsets of L. Then $\bigcap_{i\in I} S_i + \bigcap_{j\in I} T_j \subseteq \bigcap_{i\in I}(S_i + T_i)$.*

Theorem 3.11. *Suppose $M \subseteq \mathbb{R}^k$ and $A_i \subseteq L$, $i \in [k]$, are convex. Then both $\oplus_M(A_1, \ldots, A_k)$ and $\Box_M(A_1, \ldots, A_k)$ are convex.*

Proof of Lemma 3.10. 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 $\bigcap_{i\in I}(S_i + T_i)$. (Lem. 3.10) ■

Proof of Theorem 3.11. Fix arbitrary $x, y \in \oplus_M(A_1, \ldots, A_k)$. Then there are $m, n \in M$, such that

$$x \in \sum_{i\in[k]} m_i \star A_i \quad \text{and} \quad y \in \sum_{j\in[k]} n_j \star A_j. \tag{3.8}$$

To show $\oplus_M(A_1, \ldots, A_k)$ is a convex set, we need to show $tx + (1-t)y \in \oplus_M(A_1, \ldots, A_k)$ for all $t \in (0, 1)$. By virtue of (3.8),

$$\begin{aligned} \forall_{t\in(0,1)}:\ tx + (1-t)y &\in t \sum_{i\in[k]} m_i \star A_i + (1-t) \sum_{j\in[k]} n_j \star A_j \\ &= \sum_{i\in[k]} (tm_i \star A_i + (1-t)n_i \star A_i). \end{aligned} \tag{3.9}$$

We have

$$\forall_{i\in[k]}:\ tm_i \star A_i + (1-t)n_i \star A_i = (tm_i + (1-t)n_i) \star A_i, \tag{3.10}$$

and thus

$$\sum_{i\in[k]} \Big(tm_i \star A_i + (1-t)n_i \star A_i) = \sum_{i\in[k]} (tm_i + (1-t)n_i\Big) \star A_i. \tag{3.11}$$

Finally, convexity of M guarantees $tm + (1-t)n \in M$, and therefore

$$\begin{aligned}\forall_{t\in(0,1)}:\ tx + (1-t)y &\overset{(3.9)}{\in} \sum_{i\in[k]} \Big(tm_i \star A_i + (1-t)n_i \star A_i\Big) \\ &\overset{(3.11)}{=} \sum_{i\in[k]} (tm_i + (1-t)n_i) \star A_i \\ &\subseteq \bigcup_{m\in M} \sum_{i\in[k]} m_i \star A_i, \end{aligned} \tag{3.12}$$

which concludes the proof that $\oplus_M(A_1, \ldots, A_k)$ is convex.

The proof that $\square_M(A_1, \ldots, A_k)$ is convex is very similar. Let $x, y \in \square_M(A_1, \ldots, A_k)$. Then there exists $m, n \in M$ such that $x \in \bigcap_{i\in[k]} m_i \star A_i$ and $y \in \bigcap_{j\in[k]} n_j \star A_j$. Therefore

$$\begin{aligned}\forall_{t\in(0,1)}:\ tx + (1-t)y \quad &\in \quad t \bigcap_{i\in[k]} m_i \star A_i + (1-t) \bigcap_{j\in[k]} n_j \star A_j \\ &= \bigcap_{i\in[k]} tm_i \star A_i) + \bigcap_{j\in[k]} (1-t)n_j \star A_j \\ &\overset{\text{L3.10}}{\subseteq} \bigcap_{i\in[k]} \Big(tm_i \star A_i + (1-t)n_i \star A_i\Big). \end{aligned} \tag{3.13}$$

From (3.10)

$$\bigcap_{i\in[k]} (tm_i \star A_i + (1-t)n_i \star A_i) = \bigcap_{i\in[k]} (tm_i + (1-t)n_i) \star A_i. \tag{3.14}$$

Again the convexity of M guarantees the presence of $tm + (1-t)n \in M$, and mirroring (3.12)

$$\begin{aligned}\forall_{t\in(0,1)}:\ tx + (1-t)y &\overset{(3.13)}{\in} \bigcap_{i\in[k]} \Big(tm_i \star A_i + (1-t)n_i \star A_i\Big) \\ &\overset{(3.14)}{=} \bigcap_{i\in[k]} (tm_i + (1-t)n_i) \star A_i \\ &\subseteq \bigcup_{m\in M} \bigcap_{i\in[k]} m_i \star A_i, \end{aligned}$$

which concludes the proof that $\square_M(A_1, \ldots, A_k)$ is convex. (Thm. 3.11) ■

3.2.3 Radiant and co-radiant properties

Just like our results in the previous section, both of the operations $\oplus_M$ and $\square_M$ preserve radiance (resp. co-radiance) when M is radiant (resp. co-radiant). Propositions 3.12 and 3.13 are essentially immediate, but they are important to have when characterising the asymptotic cones of sets in the image of $\square_M$.

Proposition 3.12. *Suppose $M \subseteq \mathbb{R}^k$ and $A_i \subseteq L$, $i \in [k]$ are convex. Then $\oplus_M(A_1, \ldots, A_k)$ is radiant (resp. co-radiant) if*

(i) M is radiant (resp. co-radiant), or

(ii) the sets A_i are radiant (resp. co-radiant).

Proof. We use the convexity of each of the A_i to distribute the scalar c over the summation

$$\begin{aligned}(0,1] \cdot \oplus_M(A_1, \ldots, A_k) &= \bigcup_{c\in(0,1]} \bigcup_{m\in M} c \cdot \sum_{i\in[k]} m_i \star A_i \\ &= \bigcup_{c\in(0,1]} \bigcup_{m\in M} \sum_{i\in[k]} cm_i \star A_i.\end{aligned}$$

If M is radiant then the set $(0,1]$ gets absorbed into M, if each of A_i for $i \in [k]$ is each radiant it gets absorbed into each of them, proving radiance. Likewise for the set $[1, \infty)$, to prove co-radiance. ∎

There are similar properties for $\square_M$ absent the convexity assumption, and the proof is exactly the same.

Proposition 3.13. *Let $M \subseteq \mathbb{R}^k$ and $A_i \subseteq L$ for $i \in [k]$. Then $\square_M(A_1, \ldots, A_k)$ is radiant (resp. co-radiant) if*

(i) M is radiant (resp. co-radiant), or

(ii) the sets A_i for $i \in [k]$ are radiant (resp. co-radiant).

3.2.4 Asymptotic properties

The behaviour of the gauge and co-gauge functions (2.10) is contingent on the asymptotic behaviour of the associated sets. Consequentially we need some basic results on the asymptotic cones for $\square_M$ to complete the main theorems in Section 3.3.

Lemma 3.14. *Suppose $M \subseteq \mathbb{R}^k_{\geq 0}$ and $A_i \subseteq L$ for $i \in [k]$. Then*

(i) $\square_M(A_1, \ldots, A_k)_\infty \supseteq \bigcap_{i\in[k]}(A_i)_\infty$ every A_i for $i \in [k]$ is radiant or every A_i for $i \in [k]$ is convex, and

(ii) $\square_M(A_1, \ldots, A_k)_\infty \subseteq \bigcap_{i\in[k]}(A_i)_\infty$ if M is bounded or each A_i for $i \in [k]$ is co-radiant.

Proof. Let $A \overset{\text{def}}{=} \square_M(A_1, \ldots, A_k)_\infty$.

Assume the A_i are each radiant: When the sets A_i are each radiant A is radiant (Prop. 3.13) and we can apply Prop. 2.9 (i) to calculate

$$\begin{aligned} A_\infty &\overset{\text{P 2.9 (i)}}{=} \bigcap_{\epsilon>0} \bigcup_{m\in M} \bigcap_{i\in[k]} \epsilon m_i \star \overline{A}_i \\ &\supseteq \bigcup_{m\in M} \bigcap_{i\in[k]} \bigcap_{\epsilon>0} m_i\epsilon \star \overline{A}_i. \\ &\overset{\text{P 2.9 (i)}}{=} \bigcap_{i\in[k]} (A_i)_\infty. \end{aligned}$$

Assume the A_i are each convex: Then

$$\begin{aligned} A_\infty &\overset{\text{P 2.8 (vii)}}{\supseteq} \bigcup_{m\in M} \left(\bigcap_{i\in[k]} m_i \star A_i \right)_\infty \\ &\overset{\text{P 2.8 (vii)}}{=} \bigcup_{m\in M} \bigcap_{i\in[k]} (m_i \star A_i)_\infty \\ &= \bigcap_{i\in[k]} (A_i)_\infty. \end{aligned}$$

Assume M is bounded: Let $x \in A_\infty$. Then there are nets $(x_i)_{i\in I} \subseteq A$ and $(t_i)_{i\in I} \subseteq \mathbb{R}_{>0}$ with $t_i \to 0$ so that $x = \lim_{i\in I} t_i x_i$. Therefore there is a net $(m_i)_{i\in I} \subseteq M$ with $x_i \in \bigcap_{j\in[k]} m_{ij} \star A_j$. If for any $j \in [k]$ there is $m_{ij} = 0$ for all i in a cofinal subset of I, then $x \in (A_j)_\infty$ as desired. So let us assume this is not the case, that is $(m_i) \subseteq \mathbb{R}^k_{>0}$. Then there are nets $(a_{ij})_{i\in I} \subseteq A_j$

with $x_i = m_{ij} a_{ij}$ for each $j \in [k]$ so that

$$\forall_{j\in[k]} : \lim_{i\in I} t_i x_i = \lim_{i \in I} t_i m_{ij} a_{ij}.$$

If M is bounded then we may assume (m_i) converges, by passing to a convergent subnet if necessary. Then $t_i m_{ij} \to 0$ and $t_i m_{ij} a_{ij} \to x$ for all $j \in [k]$. This shows $x \in \bigcap_{i\in[k]} (A_i)_\infty$ and $A_\infty \subseteq \bigcap_{i\in[k]} (A_i)_\infty$.

Assume the A_i are each co-radiant: When the sets A_i are each co-radiant, A is co-radiant (Prop. 3.13), and hence

$$A_\infty \overset{\text{P 2.9(ii)}}{=} \overline{\bigcup_{\epsilon>0} \bigcup_{m\in M} \bigcap_{i\in[k]} \epsilon m_i \star A_i} \subseteq \overline{\bigcup_{m\in M} \bigcap_{i\in[k]} m_i \cdot \bigcup_{\epsilon>0} \epsilon \star A_i} = \overline{\bigcap_{i\in[k]} \operatorname{pos} A_i}.$$

Then using Prop. 2.9 (ii) we have $A_\infty \subseteq \bigcap_{i\in[k]} (A_i)_\infty$. ∎

There are two immediate corollaries from Lemma 3.14.

Corollary 3.15. *Let $M \subseteq \mathbb{R}^k_{\geq 0}$ and $A_i \subseteq L$ for $i \in [k]$. If either 1. both M and the sets A_i are bounded for $i \in [k]$, or 2. the sets A_i are convex and co-radiant for $i \in [k]$, then*

$$\square_M(A_1, \ldots, A_k)_\infty = \bigcap_{i\in[k]} (A_i)_\infty.$$

Corollary 3.16. *Let $M \subseteq \mathbb{R}^k_{\geq 0}$ and $A_i \subseteq L$ $i \in [k]$. Let $A \overset{\text{def}}{=} \square_M(A_1, \ldots, A_k)$. (i) If each A_i for $i \in [k]$ is convex or radiant, then $\overline{\operatorname{bc}} A \subseteq \sum_{i\in[k]} \overline{\operatorname{bc}} A_i$. (ii) If each A_i for $i \in [k]$ is co-radiant or M is bounded, then $\overline{\operatorname{bc}} A \supseteq \sum_{i\in[k]} \overline{\operatorname{bc}} A_i$.*

Proof of Corollary 3.15. The second claim is immediate and so we only prove the first. Lem. 3.14 (ii) shows $\square_M(A_1, \ldots, A_k)_\infty \subseteq \bigcap_{i\in[k]} (A_i)_\infty$ when M is bounded. Since each of the sets $(A_i)_{i\in[k]}$ are bounded $\bigcap_{i\in[k]} (A_i)_\infty = \{0\}$ (from Prop. 2.8 (v)). There is always $\{0\} \subseteq \square_M(A_1, \ldots, A_k)_\infty$, which gives equality. (Cor. 3.15) ∎

Proof of Corollary 3.16. Since the asymptotic cone always contains 0, we have [cf. 149, p. 7]

$$\left(\bigcap_{i\in[k]} (A_i)_\infty \right)^- = \sum_{i\in[k]} (A_i)_\infty^- \overset{\text{P 2.8(vi)}}{=} \sum_{i\in[k]} \overline{\operatorname{bc}} A_i.$$

Using Prop. 2.8 (vi) $\overline{\mathrm{bc}}\, A = (A)_\infty^-$. Since each A_i is convex for $i \in [k]$ we have $A_\infty \supseteq \bigcap_{i\in[k]}(A_i)_\infty$ (via Lem. 3.14 (i)) and $A_\infty \subseteq \bigcap_{i\in[k]}(A_i)_\infty$ if the sets A_i for $i \in [k]$ are co-radiant (Lem. 3.14 (ii)). (Cor. 3.16) ■

3.3 Gauge functions

Mirroring the approach of Section 3.1, for $M \subseteq \mathbb{R}^k_{\geq 0}$ and $f_1, \dots f_k : X \to \bar{\mathbb{R}}$. Let $\mu_M(f_1, \dots, f_k), \nu_M(f_1, \dots, f_k) : L \to \bar{\mathbb{R}}$ be defined by

$$\mu_M(f_1, \dots, f_k)(x) \stackrel{\text{def}}{=} \mu_M \circ (f_1, \dots, f_k)(x) \tag{3.15}$$

$$\nu_M(f_1, \dots, f_k)(x) \stackrel{\text{def}}{=} \nu_M \circ (f_1, \dots, f_k)(x), \tag{3.16}$$

for $x \in \bigcap_{i\in[k]} \operatorname{dom} f_i$. For $x \in L \setminus \bigcap_{i\in[k]} \operatorname{dom} f_i$ we define $\mu_M(f_1, \dots, f_k)(x) \stackrel{\text{def}}{=} \infty$ and $\nu_M(f_1, \dots, f_k)(x) \stackrel{\text{def}}{=} -\infty$. This is a convention that is adopted by Ward [137] in a similar setting to ours.

To demonstrate Theorem 3.4 we needed relatively fewer assumptions compared with their gauge counterparts: Theorems 3.19 and 3.20 (which are proved below). To some degree this is a product of the powerful definition of the support function, which is always convex, and always lower semicontinuous. By comparison the extended gauge function we have defined can fail to be both convex and lower semicontinuous. In order to develop a dual theory for the gauge functions we have already had to appeal to a substantial amount of mathematical machinery to take care of the corner and asymptotic cases — the additional assumptions present in Theorems 3.19 and 3.20 reflect a compromise of mathematical convenience and analytic power. Indeed, these same conditions are again equally beneficial in Section 3.4 when it comes to proving Theorem 3.26, which unifies the operations $\oplus_M$ and $\square_M$ for the convex radiant and convex co-radiant sets. However, before we can prove Theorems 3.19 and 3.20, we need some preparatory lemmas.

Lemma 3.17. *Suppose $M \subseteq \mathbb{R}^k_{\geq 0}$ is convex and contains an order unit of $\mathbb{R}^k_{\geq 0}$. Then if $x \in \mathbb{R}^k$, $\gamma \in \mathbb{R}$ satisfy*

(i) $\mu_M(x) < \gamma$*, then there is* $m \in M$ *with* $x < \gamma m$*, when* M *is bounded;*

(ii) $\nu_M(x) > \gamma$*, then* $x = 0$ *or there is* $m \in M$ *and* $\nu_M(x) \geq \beta > \gamma$ *with* $x > \beta m$*, when* $M_\infty \setminus \operatorname{pos} M = \{0\}$

Proof. By hypothesis M contains an order unit, m_ϵ of $\mathbb{R}^k_{\geq 0}$. Every order unit of a cone is an element of the relative interior [3, Lem. 1.7], and the relative interior of $\mathbb{R}^k_{\geq 0}$ coincides with its topological interior. Thus $m_\epsilon > 0$.

(i): Suppose $\mu_M(x) < \gamma$. Then there exists $t \in [\mu_M(x), \gamma)$ with $x \in t \star M$. If $t = 0$ then $x = 0$ (because M is bounded). Therefore m_ϵ satisfies $x = tm_\epsilon < \gamma m_\epsilon$. Next, if $t > 0$ then there is $a \in M$ with $x = ta$. Immediately $x \leq ta$. Taking the interior point m_ϵ, let $m_x \overset{\text{def}}{=} \frac{t}{\gamma}a + \left(1 - \frac{t}{\gamma}\right)m_\epsilon$. Then $m_x \in M$ by convexity, and $\gamma m_x = ta + (\gamma - t)m_\epsilon$, therefore $x < \gamma m_x$.

(ii): Let $\nu_M(x) > \gamma$, then there is $t \in (\gamma, \nu_M(x)] \cap \mathbb{R}_{\geq 0}$ with $x \in t \star M$. If $t = 0$ then $x \in M_\infty \setminus \operatorname{pos} M = \{0\}$. Take any $m \in M$ and $x \geq tm$. Next, if $t > 0$ there is $a \in M$ with $x = ta$. Immediately $x \geq ta$ and $x > ta - \lambda m_\epsilon$ for all $\lambda > 0$. Let $m_x \overset{\text{def}}{=} \frac{t}{\beta}a + (1 - \frac{t}{\beta})m_\epsilon$, then $m_x \in M$ by convexity, and for all $0 < \beta < t$ there is $\beta m_x = ta - (t - \beta)m_\epsilon \in \beta \star M$ and $x > \beta m_x$ when $\beta < t$. In particular for $\beta \in (\max(0, \gamma), t)$. ∎

Lemma 3.18. *Let $A \subseteq L$, $m \in \mathbb{R}_{\geq 0}$. Then*

(i) $\mu_{m\star A}(x) \leq 1$ *implies* $\mu_A(x) \leq m$

(ii) $\nu_{m\star A}(x) \geq 1$ *implies* $\nu_A(x) \geq m$

Proof. (i): For every $\gamma > \mu_A(x) = m$ there is $t \in [m, \gamma)$ with $x \in t \star A$. If $t = 0$ then $x \in A_\infty$ and $m = 0$, thus $x \in m \star A$. If $t > 0$ then $x \notin A_\infty$, $m > 0$, and $x \in t \star A$, whence $\frac{m}{t}x \in m \star A$. Suppose $m > 0$. Then $\mu_{m\star A}(x) = \mu_A(x/m)$, thus $\mu_A(x) \leq m$. If $m = 0$ then $\mu_{m\star A} = \iota_{A_\infty}$. By assumption $\mu_{m\star A}(x) \leq 1 < \infty$ and so $x \in (A)_\infty$. From Prop. 2.6 (xii), we have $\mu_A(x) = m$.

(ii): Suppose $m > 0$. Then $\nu_{m\star A}(x) = \nu_A(x/m)$ and $\nu_A(x) \geq m$. If $m = 0$ then $\nu_{m\star A}(x) \geq 1$ implies $x \in A_\infty$ and $\nu_A(x) \geq 0 = m$. ∎

For a vector space L and a cone $K \subseteq L$ let $\mathcal{M}_0(K)$ denote the collection of subsets of L which are convex, K-full, bounded, contain both 0 and an order unit of K. Let $\mathcal{M}_\infty(K)$ denote the collection of of subsets M of K which are closed, convex, containing an order unit and have $\operatorname{pos} M = K \setminus \{0\}$.

Theorem 3.19. *Suppose $M \in \mathcal{M}_0(\mathbb{R}^k_{\geq 0})$, $A_i \subseteq L$ for $i \in [k]$. Let $A \stackrel{\text{def}}{=} \square_M(A_1, \ldots, A_k)$. Then $\mu_M(\mu_{A_1}, \ldots, \mu_{A_k}) \geq \mu_A$, and*

$$\mu_A = \mu_M(\mu_{A_1}, \ldots, \mu_{A_k}) \iff A_\infty \supseteq \bigcap_{i\in[k]} (A_i)_\infty.$$

In particular, when the sets $(A_i)_{i\in[k]}$ are all radiant or convex, $A_\infty \supseteq \bigcap_{i\in[k]}(A_i)_\infty$ and $\mu_A = \mu_M(\mu_{A_1}, \ldots, \mu_{A_k})$.

Theorem 3.20. *Suppose $M \in \mathcal{M}_\infty(\mathbb{R}^k_{\geq 0})$, $A_i \subseteq L$ for $i \in [k]$. Let $A \stackrel{\text{def}}{=} \square_M(A_1, \ldots, A_k)$ and assume $A_\infty \subseteq \bigcap_{i\in[k]}(A_i)_\infty$. Then $\nu_M(\nu_{A_1}, \ldots, \nu_{A_k}) \geq \nu_A$ and*

$$\nu_A = \nu_M(\nu_{A_1}, \ldots, \nu_{A_k}) \iff [0, \infty) \star A \supseteq \bigcap_{i\in[k]} (A_i)_\infty.$$

In particular, when the sets $(A_i)_{i\in[k]}$ are each co-radiant, $A_\infty \subseteq \bigcap_{i\in[k]}(A_i)_\infty$ and $\nu_A = \nu_M(\nu_{A_1}, \ldots, \nu_{A_k})$ if and only if $A_\infty \supseteq \bigcap_{i\in[k]}(A_i)_\infty$. When, additionally, the sets $(A_i)_{i\in[k]}$ are each convex, $A_\infty = \bigcap_{i\in[k]}(A_i)_\infty$ and $\nu_A = \nu_M(\nu_{A_1}, \ldots, \nu_{A_k})$.

Proof of Theorem 3.19. Let $x \in L$ and $\gamma \in \mathbb{R}_{>0}$ satisfy $\gamma > \mu_M(y)$, where $y \stackrel{\text{def}}{=} (\mu_{A_1}, \ldots, \mu_{A_k})(x)$. Since M is convex and contains an order unit of $\mathbb{R}^n_{\geq 0}$, Lem. 3.17 (i) shows there is $m \in M$ with $y < \gamma m$ pointwise. Therefore

$$\forall_{i\in[k]} :\ \mu_{A_i}(x) < \gamma m_i \stackrel{\text{P 2.6 (vi)}}{\Longrightarrow} x \in [0, \gamma] \star m_i \star A_i$$

and

$$x \in \bigcup_{m\in M} \bigcap_{i\in[k]} [0, \gamma] \star m_i \star A_i \subseteq [0, \gamma] \cdot \bigcup_{m\in M} \bigcap_{i\in[k]} m_i \star A_i \implies \mu_A(x) \leq \gamma.$$

We have shown

$$\begin{aligned} \forall_{\gamma\in\mathbb{R}_{>0}} \forall_{x\in L} :\ & \Big[\gamma > \mu_M(\mu_{A_1}, \ldots, \mu_{A_k})(x) \implies \gamma \geq \mu_A(x)\Big] \\ & \implies \mu_M(\mu_{A_1}, \ldots, \mu_{A_k}) \geq \mu_A\,. \end{aligned} \tag{3.17}$$

Assume $\mu_M(\mu_{A_1}, \ldots, \mu_{A_k}) = \mu_A$: Pick $x \in L$ and let $y \stackrel{\text{def}}{=} (\mu_{A_1}, \ldots, \mu_{A_k})(x)$.

We have $\mu_M(y) = 0$ precisely when $u \in M_\infty$ (Prop. 2.6 (ii)) and because M is bounded $M_\infty = \{0\}$. Therefore $\mu_M^{-1}(0) = 0$. If we suppose $x \in \bigcap_{i\in[k]}(A_i)_\infty$ then $y = 0$, and $\mu_M(y) = \mu_A(x) = 0$. This shows $x \in A_\infty$ and $\bigcap_{i\in[k]}(A_i)_\infty \subseteq A_\infty$.

Assume $A_\infty \supseteq \bigcap_{i\in[k]}(A_i)_\infty$: From Lem. 3.14 (ii) $A_\infty \subseteq \bigcap_{i\in[k]}(A_i)_\infty$ whenever M is bounded. Thus $A_\infty = \bigcap_{i\in[k]} A_i$ by hypothesis. Suppose $x \in L$ and $\gamma \in \mathbb{R}_{>0}$ satisfy $\gamma > \mu_A(x)$. Then $\lambda \in [\mu_A(x), \gamma)$ with $x \in \lambda \star A$. If $\lambda = 0$, then

$$x \in A_\infty = \bigcap_{i\in[k]} (A_i)_\infty \overset{\text{P 2.6 (ii)}}{\iff} \forall_{i\in[k]} :\ \mu_{A_i}(x) = 0,$$

and immediately $y = 0$. Thus $\mu_M(y) = 0 \leq \lambda$. If $\lambda > 0$, then there exists $m \in M$ with $x/\lambda \in \bigcap_{i\in[k]} m_i \star A_i$. Thus for each $i \in [k]$ we have

$$\mu_{m_i\star A_i}(x/\lambda) \leq 1 \overset{\text{L 3.18 (i)}}{\Longrightarrow} \mu_{A_i}(x) \leq \lambda m_i,$$

and in particular, $\mu_{A_i}(x) = \lambda m_i$ whenever $m_i = 0$ for $i \in [k]$. Thus $y \leq \lambda m_i$. By hypothesis M is full and contains 0, likewise $\lambda \star M$ is full and contains 0, whence $y \in [0, \lambda m]_{\mathbb{R}^k_{\geq 0}} \subseteq \lambda \star M$. Therefore $\mu_M(y) \leq \lambda$. We have shown

$$\begin{aligned}\forall_{\gamma\in\mathbb{R}_{>0}}\forall_{x\in L} :\ &\Big[\gamma > \mu_A(x) \implies \gamma \geq \mu_M(\mu_{A_1}, \ldots, \mu_{A_k})(x)\Big]\\ &\implies \mu_A \geq \mu_M(\mu_{A_1}, \ldots, \mu_{A_k}),\end{aligned}$$

which together with (3.17) gives $\mu_A = \mu_M(\mu_{A_1}, \ldots, \mu_{A_k})$.

The final claim follows from Lem. 3.14 (i). (Thm. 3.19) ■

Proof of Theorem 3.20. The style of proof is similar to the proof of Thm. 3.19, however we need some different constructions to accomplish each step. First, because $\operatorname{pos} M = \mathbb{R}^k_{\geq 0} \setminus \{0\}$ there is $M_\infty \subseteq \overline{\operatorname{pos}} M = \mathbb{R}^k_{\geq 0}$ (from Prop. 2.8 (ii)) and $M_\infty \setminus \operatorname{pos} M \subseteq \mathbb{R}^k_{\geq 0} \setminus \left(\mathbb{R}^k_{\geq 0} \setminus \{0\}\right) = \{0\}$.

Assume $A_\infty \subseteq \bigcap_{i\in[k]}(A_i)_\infty$: Suppose $x \in L$ and $\gamma \in \mathbb{R}$ satisfy $\nu_A(x) > \gamma$. Then $\lambda \in (\gamma, \nu_A(x)] \cap \mathbb{R}_{\geq 0}$ with $x \in \lambda \star A$. If $\lambda = 0$, then $\nu_A(x) = 0$ and

$$x \in A_\infty \subseteq \bigcap_{i\in[k]} (A_i)_\infty \overset{\text{P 2.6 (ix)}}{\Longrightarrow} \forall_{i\in[k]} :\ \nu_{A_i}(x) \geq 0,$$

and immediately $y \in \mathbb{R}^k_{\geq 0}$. Thus $\mu_M(y) \geq \lambda > \gamma$. Next, if $\lambda > 0$ then there exists $m \in M$ with $x/\lambda \in \bigcap_{i\in[k]} m_i \star A_i$. Thus for each $i \in [k]$ we have

$$\nu_{m_i\star A_i}(x/\lambda) \geq 1 \overset{\text{L3.18 (ii)}}{\Longrightarrow} \nu_{A_i}(x) \geq \lambda m_i .$$

Therefore $y \in \lambda M + \mathbb{R}^k_{\geq 0}$. Since M is closed convex, the set $[1,\infty)\star M$ is closed convex and co-radiant thus $([1,\infty)\star M)_\infty = \overline{\operatorname{pos}}([1,\infty)\star M) = \mathbb{R}^k_{\geq 0}$ (from Prop. 2.9 (ii)) thus $\lambda M + \mathbb{R}^k_{\geq 0} \subseteq [\lambda,\infty)\star M$ (from Prop. 2.8 (iv)) whence $\nu_M(y) \geq \lambda$. We have shown

$$\begin{aligned}\forall_{\gamma\in\mathbb{R}}\forall_{x\in L} :\ & \Big[\nu_A(x) > \gamma \implies \nu_M(\nu_{A_1},\dots,\nu_{A_k})(x) \geq \gamma\Big] \\ & \implies \nu_M(\nu_{A_1},\dots,\nu_{A_k}) \geq \nu_A\,.\end{aligned}$$

Assume $\bigcap_{i\in[k]}(A_i)_\infty \subseteq [0,\infty)\star A$: Suppose $x \in L$ and $\gamma \in \mathbb{R}$ satisfy $\nu_M(y) > \gamma$, where $y \overset{\text{def}}{=} (\nu_{A_1},\dots,\nu_{A_k})(x)$. Then $\lambda \in (\gamma, \nu_A(x)] \cap \mathbb{R}_{\geq 0}$ with $y \in \lambda \star M$. If $\lambda = 0$ then $y = 0$. Therefore $x \in \bigcap_{i\in[k]}(A_i)_\infty \subseteq [0,\infty)\star A$ and $0 = \nu_M(y) \leq \nu_A(x)$. Now assume $\lambda > 0$. Then there exists $m \in M$ and $\beta > \gamma$ with $y > \beta m$. Therefore

$$\forall_{i\in[k]} :\ \nu_{A_i}(x) > \beta m_i \overset{\text{P 2.6 (xiii)}}{\Longrightarrow} x \in [\beta,\infty)\star m_i \star A_i$$

and

$$\begin{aligned} x \in \bigcup_{m\in M}\bigcap_{i\in[k]}[\beta,\infty)\star m_i\star A_i &= [\beta,\infty)\cdot \bigcup_{m\in M}\bigcap_{i\in[k]} m_i\star A_i \\ &\implies \nu_A(x) \geq \gamma. \end{aligned}$$

Therefore

$$\begin{aligned}\forall_{\gamma\in\mathbb{R}}\forall_{x\in L} :\ & \Big[\gamma < \nu_M(\nu_{A_1},\dots,\nu_{A_k})(x) \implies \gamma < \nu_A(x)\Big] \\ & \implies \nu_M(\nu_{A_1},\dots,\nu_{A_k}) \leq \nu_A\,.\end{aligned}$$

Assume $\nu_M(\nu_{A_1},\dots,\nu_{A_k}) = \nu_A$: Choose $x \in \bigcap_{i\in[k]}(A_i)_\infty$. Then $\nu_{A_i}(x) \geq 0$ for all $i \in [k]$, and $(\nu_{A_1},\dots,\nu_{A_k})(x) \in \mathbb{R}^k_{\geq 0} = [0,\infty)\star M$ and $\nu_M(\nu_{A_1},\dots,\nu_{A_k})(x) <$

∞. This shows

$$\bigcap_{i\in[k]} (A_i)_\infty \subseteq \operatorname{dom} \nu_M(\nu_{A_1}, \ldots, \nu_{A_k}) = \operatorname{dom} \nu_A = [0, \infty) \star A,$$

thus $\bigcap_{i\in[k]}(A_i)_\infty \subseteq [0, \infty) \star A$.

In the final claim we apply Lem. 3.14 (ii) to show $A_\infty \subseteq \bigcap_{i\in[k]}(A_i)_\infty$. The set A is also co-radiant by Prop. 3.13 and therefore $\operatorname{pos} A \subseteq A_\infty$ by Prop. 2.9 (ii), whence $[0, \infty) \star A = A_\infty$. (Thm. 3.20) ■

3.4 Polarity

Up until now the analysis of $\oplus_M$ and $\square_M$ has been completely separate, with results on $\oplus_M$ limited to support functions and results on $\square_M$ limited to gauge functions. However we are about to see that the two are connected via the duality relations introduced in Section 2.3. Let $(L, \geq)$ be an ordered vector space.

A function $f : L \to \mathbb{R}$, is *isotone* on $T \subseteq L$ if $f(x) \geq f(y)$ whenever $x, y \in T$ and $x \geq y$ [cf. 137, Def. 2.1]. As usual, when $L = \mathbb{R}^k$ the ordering is assumed to be pointwise. It is important for our proof of Theorem 3.26 that the gauge and co-gauge functions are isotone, fortunately the conditions on M we used in Section 3.3 for Theorems 3.19 and 3.20 are both sufficient and (up to closure and/or convexity) necessary to ensure this property.

Proposition 3.21. *Assume $M \subseteq L_{\geq 0}$ is bounded, contains 0, and an order unit.*

(*i*) *If M is full, then μ_M is finite and isotone on $L_{\geq 0}$.*

(*ii*) *If M is closed then μ_M is isotone only if M is full.*

Proposition 3.22. *Assume $M \subseteq L_{\geq 0}$ is convex.*

(*i*) *If $\operatorname{pos} M = L_{\geq 0} \setminus \{0\}$, then ν_M is finite and isotone on $L_{\geq 0}$.*

(*ii*) *If M is closed, then ν_M is isotone only if $M_\infty \supseteq L_{\geq 0}$.*

Proof of Proposition 3.21. (i): Let e be the order unit of $L_{\geq 0}$ contained in M. Pick an arbitrary $x \in L_{\geq 0}$. Then for some $c > 0$ there is $ce \geq x$. Because

M is full, $x \in L_{\geq 0} \cap (ce - L_{\geq 0}) \subseteq c \star M$. Thus $L_{\geq 0} \subseteq \operatorname{pos} M$, and because M is bounded $M_\infty = \{0\}$ and $L_{\geq 0} \subseteq [0, \infty) \star M = \operatorname{dom} \mu_M$.

Suppose $x, y \in L_{\geq 0}$ with $x \geq y$ and assume M is full. Choose $\gamma \in \mathbb{R}$ with $\gamma > \mu_M(x)$. Then $\lambda \in [\mu_M(x), \gamma)$ with $x \in \lambda \star M$. If $\lambda = 0$ then $x \in M_\infty = \{0\}$ and $x \geq y$ means $y = x = 0$, in which case $\mu_M(y) = \mu_M(x)$. Next if $\lambda > 0$ then $x \in \lambda \star M$. Since M is full, containing 0, so is $\lambda \star M$. $x \geq y \geq 0$ means that $y \in [0, x]_{L_{\geq 0}}$. Since $x, 0 \in \lambda \star M$, $[0, x]_{L_{\geq 0}} \subseteq \lambda \star M$. Thus $\mu_M(y) \leq \lambda$. We have proven $\mu_M(x) \geq \mu_M(y)$.

(ii): Assume M is also not full. Then $x, y \in M$ and a point $z \in [x, y]_{L_{\geq 0}}$ with $z \notin M$, in particular, $z \leq y$. If M is also closed then $\mu_M(z) > 1$. Since $y \in M$ there is $\mu_M(y) \leq 1 < \mu(z)$ and μ_M is not isotone. (Prop. 3.21) ■

Proof of Proposition 3.22. (i): Since $\operatorname{pos} M = L_{\geq 0} \setminus \{0\}$ there is $M_\infty \subseteq \overline{\operatorname{pos}} M = L_{\geq 0}$ and $[0, \infty) \star M = L_{\geq 0} = \operatorname{dom} \nu_M$.

Suppose $x, y \in L_{\geq 0}$ with $x \geq y$ and assume M is closed convex. Choose $\gamma \in \mathbb{R}$ with $\gamma < \nu_M(y)$. Then $\lambda \in [\mu_M(y), \gamma) \cap \mathbb{R}_{\geq 0}$ with $y \in \lambda \star M$. If $0 = \lambda \geq \nu_M(y)$ then $\nu(y) \geq 0$. If $\lambda > 0$ then because $x \geq y$ there is $x \in y + L_{\geq 0}$. By hypothesis M is convex with $M_\infty = L_{\geq 0}$, whence $\lambda \star M + M_\infty \subseteq \lambda \star M$ and $y + L_{\geq 0} \subseteq [\lambda, \infty) \star M$ and $\nu_M(x) \geq \lambda$. Thus $\nu_M(x) \geq \nu_M(y)$. Thus ν_M is isotone.

(ii): Assume $M_\infty \subset L_{\geq 0}$. Then some $v \in L_{\geq 0}$ with $v \notin M_\infty$. Since M is closed convex M_∞ is the largest set of points that satisfies $M + M_\infty \subseteq M$ and so $m + v \notin M$ for all $m \in M$ and so $\nu_M(m + v) < 1$. Pick an arbitrary $m \in M$. Since $v \in L_{\geq 0}$ there is $v \geq 0$ and $m + v \geq m$. Since M is assumed closed and $m \in M$ there is $\nu_M(m) \geq 1 > \nu_M(m + v)$, and ν_M is not isotone. (Prop. 3.22) ■

Remark 3.23. With regards to conditions on M of Proposition 3.22 (equivalently Theorem 3.20), observe that when M is closed co-radiant, $M_\infty = \overline{\operatorname{pos}} M$ (via Proposition 2.9 (ii)), so that ν_M is isotone if and only if $\operatorname{pos} M = L_{\geq 0} \setminus \{0\}$. Equivalently, ν_M is isotone if and only if M is closed co-star-shaped and $M_\infty = L_{\geq 0}$.

Corollary 3.24. *Assume $M \subseteq \mathbb{R}^k_{\geq 0}$ and $A_i \subseteq L$ for $i \in [k]$ are convex.*

(*i*) *If M is full, then $\mu_M(\mu_{A_1}, \ldots, \mu_{A_m})$ is convex.*

(*ii*) *If* $\operatorname{pos} M = \mathbb{R}^k_{\geq 0} \setminus \{0\}$, *then* $\nu_M(\nu_{A_1}, \ldots, \nu_{A_m})$ *is concave.*

Before we can proceed, we need a version of the subdifferential chain rule easy to deduce using our notation and asymptotic multiplication (2.8) from Ward [137] (viz. Remark 2.3).

Lemma 3.25. *Suppose* $f_1, \ldots, f_k : L \to \bar{\mathbb{R}}$ *are each convex and finite at* $z \in L$. *Let* $f \overset{\text{def}}{=} (f_1, \ldots, f_k)$ *and assume* $F : \mathbb{R}^k \to \bar{\mathbb{R}}$ *is convex, finite and isotone on the set* $R \overset{\text{def}}{=} \left\{ y \in \mathbb{R}^k \;\middle|\; \exists_{x \in L} : \; f(x) \leq y \right\}$ *and finite at* $f(z)$. *Then if* $R \cap \operatorname{int}(\operatorname{dom} F) \neq \emptyset$

$$\partial(F \circ f)(z) = \left\{ \partial \left(\sum_{i \in [k]} m_i \cdot_{\mathsf{e}} f_i \right)(z) \;\middle|\; m \in \partial F(f(z)) \right\}$$

and $\oplus_{\partial F(f(z))}(\partial f_1(z), \ldots, \partial f_k(z)) \subseteq \partial(F \circ f)(z)$. *In particular, when the* f_i *are additionally positively homogeneous*

$$\partial(F \circ f)(0) = \overline{\oplus_{\partial F(f(0))}(\partial f_1(0), \ldots, \partial f_k(0))}^{*}. \tag{3.18}$$

Proof. Most of the above is proven by Ward [137, Thm. 2.6]. The closure result (3.18) is because $\partial(f_1 + \cdots + f_k)(0) = \overline{\partial f_1(0) + \cdots + \partial f_k(0)}^{*}$ for lower semicontinuous sublinear functions $(f_i)_{i \in [k]}$ [146, Prop. 2, 149, Thm. 2.4.14(viii)]. Ward [137] observes $\partial(c \cdot_{\mathsf{e}} f) = c \star \partial f$ with our asymptotic convention (2.8). It is well known that the subdifferential of proper convex functions is $\sigma(L^*, L)$-compact [2, Thm. 7.13] and the $\sigma(L^*, L)$-closure of the M-sum follows. ∎

Theorem 3.26. *Suppose* $M \subseteq \mathbb{R}^k_{\geq 0}$, $A_i \subseteq L$ *for* $i \in [k]$. *Then*

(*i*) $\oplus_M(A_1, \ldots, A_k)^{\circ} \subseteq \square_{M^{\circ}}(A_1^{\circ}, \ldots, A_k^{\circ})$, *and*

(*ii*) $\oplus_M(A_1, \ldots, A_k)^{\triangledown} \subseteq \square_{M^{\triangledown}}(A_1^{\triangledown}, \ldots, A_k^{\triangledown})$.

Now assume the sets A_i *for* $i \in [k]$ *are each closed and convex.*

(*iii*) *If* $M \in \mathcal{M}_0(\mathbb{R}^k_{\geq 0})$, *then* $\square_M(A_1, \ldots, A_k)^{\circ} = \overline{\oplus_{M^{\circ}}(A_1^{\circ}, \ldots, A_k^{\circ})}^{*}$.

(*iv*) *If* $M \in \mathcal{M}_\infty(\mathbb{R}^k_{\geq 0})$, *then* $\square_M(A_1, \ldots, A_k)^{\triangledown} = \overline{\oplus_{M^{\triangledown}}(A_1^{\triangledown}, \ldots, A_k^{\triangledown})}^{*}$.

Proof of Theorem 3.26. (i): Let $A \stackrel{\text{def}}{=} \oplus_M(A_1, \dots, A_k)$. Choose $x^* \in A^\circ$. Then for all $a \in A$

$$1 \geq \langle a, x^* \rangle \iff 1 \geq \sigma_A(x^*) \stackrel{\text{T3.4}}{\geq} \sigma_M(\sigma_{A_1}, \dots, \sigma_{A_k})(x^*). \tag{3.19}$$

Thus (3.19) shows that there is $m \in M^\circ$ with $(\sigma_{A_1}, \dots, \sigma_{A_k})(x^*) = m$. When $m_i > 0$ there is

$$\sigma_{A_1}(x^*) = m_i \implies \forall_{a \in A} : \frac{1}{m_i}\langle a, x^* \rangle \leq 1 \iff \forall_{a \in \frac{1}{m_i} \star A} : \langle a, x^* \rangle \leq 1$$

and thus $x \in m_i \star A_i^\circ$ for each ink. Next suppose there is $i \in [k]$ with $m_i = 0$. Then

$$\sigma_{A_i}(x^*) = 0 \iff \mu_{A_i^\circ}(x^*) = 0 \implies x \in (A_i^\circ)_\infty = m_i \star A_i^\circ \tag{3.20}$$

This shows $x^* \in \bigcup_{m \in M^\circ} \bigcap_{i \in [k]} m_i \star A_i^\circ = \square_{M^\circ}(A_1^\circ, \dots, A_k^\circ)$. We obtain the same result for $A^\triangledown$ by reversing some inequalities, and observing

$$\zeta_{A_i}(x^*) = 0 \iff \nu_{A_i^\triangledown}(x^*) = 0 \iff x^* \in (A_i^\triangledown)_\infty = m_i \star A_i^\triangledown,$$

in place of (3.20).

(iii): Let $B \stackrel{\text{def}}{=} \square_M(A_1, \dots, A_k)$. Then $\mu_B = \mu_M(\mu_{A_1}, \dots, \mu_{A_k})$ because the sets $(A_i)_{i \in [k]}$ are assumed convex and M satisfies the conditions of Thm. 3.20. The mapping μ_M is isotonic and convex under the assumptions on M (Prop. 3.21). Since each A_i is closed convex, μ_{A_i} is convex and lower semicontinuous for $i \in [k]$. Therefore we can apply Lem. 3.25 to calculate

$$\begin{aligned}
B^\circ &\stackrel{(2.11)}{=} \partial \mu_B(0) \\
&\stackrel{\text{T3.19}}{=} \partial(\mu_M(\mu_{A_1}, \dots, \mu_{A_m}))(0) \\
&\stackrel{\text{L3.25}}{=} \overline{\oplus_{\partial \mu_M(0)}(\partial \mu_{A_1}(0), \dots, \partial \mu_{A_k}(0))}^* \\
&\stackrel{(2.11)}{=} \overline{\oplus_{M^\circ}(A_1^\circ, \dots, A_k^\circ)}^*.
\end{aligned}$$

(iv): Let $B \stackrel{\text{def}}{=} \square_M(A_1, \dots, A_k)$. When M is bounded or $B_\infty \supseteq \bigcap_{i \in [k]}(A_i)_\infty$, we have $\nu_B = \nu_M(\nu_{A_1}, \dots, \nu_{A_k})$ because the sets $(A_i)_{i \in [k]}$ are assumed convex and M satisfies the conditions of Thm. 3.20. Since M is closed by hypothesis, $-\nu_M = -\zeta_{M^\triangledown}$, and

$$\begin{aligned} -\nu_M(\nu_{A_1}, \ldots, \nu_{A_k}) &= \sigma_{M^\triangledown}(-\nu_{A_1}, \ldots, -\nu_{A_k}) \\ &= \sup_{m \in M^\triangledown} \sum_{i \in [k]} m_i \cdot_{\mathsf{e}} (-\nu_{A_i}). \end{aligned} \tag{3.21}$$

Therefore using Thm. 3.20 and Lem. 2.23

$$\begin{aligned} -B^\triangledown &\overset{(2.11)}{=} \partial(-\nu_B)(0) \\ &\overset{\text{T}3.20}{=} \partial(-\nu_M(\nu_{A_1}, \ldots, \nu_{A_k}))(0) \\ &\overset{(3.21)}{=} \partial\left(\sup_{m \in M^\triangledown} \sum_{i \in [k]} m_i \cdot_{\mathsf{e}} (-\nu_{A_i})\right)(0) \\ &\overset{\text{L}2.23}{=} \overline{\text{co}} \bigcup_{m \in T_\epsilon(0)} \partial\left(\sum_{i \in [k]} m_i \cdot_{\mathsf{e}} (-\nu_{A_i})\right)(0) \end{aligned} \tag{3.22}$$

Since $\sum_{i \in [k]} m_i \cdot_{\mathsf{e}} (-\nu_{A_i})$ is positively homogeneous for each $m \in M$, Lem. 2.20 yields

$$\partial_\epsilon\left(\sum_{i \in [k]} m_i \cdot_{\mathsf{e}} (-\nu_{A_i})\right)(0) = \partial\left(\sum_{i \in [k]} m_i \cdot_{\mathsf{e}} (-\nu_{A_i})\right)(0), \tag{3.23}$$

for all $m \in M$ and $\epsilon \geq 0$. Next, like in the proof of Lem. 3.25, for each $m \in M$ there is [via 149, Thm. 2.4.14(viii)]

$$\begin{aligned} \partial\left(\sum_{i \in [k]} m_i \cdot_{\mathsf{e}} (-\nu_{A_i})\right)(0) &= \overline{\sum_{i \in [k]} m_i \star \partial_0(-\nu_{A_i})(0)}^* \\ &= -\overline{\sum_{i \in [k]} m_i \star A_i^\triangledown}^*. \end{aligned} \tag{3.24}$$

Therefore

$$\begin{aligned}
B^{\triangledown} &\overset{(3.22)}{=} -\overline{\mathrm{co}} \bigcup_{m \in M^{\triangledown}} \left(\sum_{i\in[k]} m_i \cdot_{\mathsf{e}} (-\nu_{A_i}) \right)(0) \\
&\overset{(3.23)}{=} \overline{\mathrm{co}} \bigcup_{m \in M^{\triangledown}} \partial \left(\sum_{i\in[k]} m_i \cdot_{\mathsf{e}} (-\nu_{A_i}) \right)(0) \\
&\overset{(3.24)}{=} \overline{\mathrm{co}} \bigcup_{m \in M^{\triangledown}} \overline{\sum_{i\in[k]} m_i \star A_i^{\triangledown}}^{*} \\
&= \overline{\oplus_{M^{\triangledown}} (A_1^{\triangledown}, \ldots, A_k^{\triangledown})}^{*},
\end{aligned}$$

as claimed, and the proof is complete. (Thm. 3.26) ■

Remark 3.27. It is also possible to prove Theorem 3.26 (iii) using a similar supremum subdifferential approach as is used in the proof of Theorem 3.26 (iv). However the converse is not true. That is, the Ward [137] chain rule is not powerful enough for the proof of Theorem 3.26 (iv), because the co-gauge $-\nu_M$ is non isotonic under our assumptions, except when M corresponds to the harmonic sum (cf. Proposition 3.22). This is the reason the proof of Theorem 3.26 (iv) is much more complicated than the proof of Theorem 3.26 (iii) (and much more complicated than the proof of the analogous specialised result of Penot and Zălinescu [100, Prop. 4.3]).

3.5 Further support and gauge results

Using Theorem 3.26 we can now complete the plan in the roadmap at the start of this chapter, and compute the support and co-support functions of sets in the image of $\square_M$ and gauge and co-gauge functions sets in the image of $\oplus_M$.

Theorem 3.28. *Let $M \in \mathcal{M}_0(\mathbb{R}^k_{\geq 0})$, $A_i \subseteq L$ each closed convex for $i \in [k]$. Then for each $x \in L$*

$$\sigma_{\square_M (A_1,\ldots,A_k)}(x) = \inf \left\{ \sup_{m\in M} \sum_{i\in[k]} m_i \cdot_{\mathsf{e}} \sigma_{A_i}(x_i) \;\middle|\; x = \sum_{i\in[k]} x_i \right\}, \qquad (3.25)$$

and

$$\mu_{\oplus_M(A_1,\dots,A_k)}(x) = \inf\left\{ \mu_M(\mu_{A_1}(x_1),\dots,\mu_{A_k}(x_k)) \;\middle|\; x = \sum_{i\in[k]} x_i \right\}, \quad (3.26)$$

where in (3.26) *the infimum is over all sequences* $(x_i)_{i\in[k]} \subseteq L$ *with* $x_i \in \operatorname{dom}\mu_{A_i}$ *for* $i \in [k]$.

Theorem 3.29. *Let* $M \in \mathcal{M}_\infty(\mathbb{R}^k_{\geq 0})$, $A_i \subseteq L$ *each closed convex for* $i \in [k]$ *and additionally either* M *is bounded or* $A_\infty \supseteq \bigcap_{i\in[k]}(A_i)_\infty$. *Then for each* $x \in L$

$$\zeta_{\square_M(A_1,\dots,A_k)}(x) = \sup\left\{ \inf_{m\in M} \sum_{i\in[k]} m_i \cdot_{\mathsf{h}} \zeta_{A_i}(x_i) \;\middle|\; x = \sum_{i\in[k]} x_i \right\}, \quad (3.27)$$

and

$$\nu_{\oplus_M(A_1,\dots,A_k)}(x) = \sup\left\{ \nu_M(\nu_{A_1}(x_1),\dots,\nu_{A_k}(x_k)) \;\middle|\; x = \sum_{i\in[k]} x_i \right\}, \quad (3.28)$$

where in (3.28) *the supremum is over all sequences* $(x_i)_{i\in[k]} \subseteq L$ *with* $x_i \in \operatorname{dom}\nu_{A_i}$ *for* $i \in [k]$.

Proof of Theorem 3.28. Define the sets

$$\mathit{\Lambda}_x \overset{\text{def}}{=} \left\{ \lambda \geq 0 \;\middle|\; x \in \lambda \star \overline{\oplus_{M^\circ}(A_1^\circ,\dots,A_k^\circ)} \right\},$$

and

$$\mathit{\Gamma}_x \overset{\text{def}}{=} \left\{ \sup_{m\in M} \sum_{i\in[k]} m_i \cdot_{\mathsf{e}} \sigma_{A_i}(x_i) \;\middle|\; \exists_{(x_i)_{i\in[k]}\subseteq L} :\; x = \sum_{i\in[k]} x_i \right\}.$$

From Thm. 3.26 (iii)

$$\sigma_{\square_M(A_1,\dots,A_k)}(x) = \mu_{\square_M(A_1,\dots,A_k)^\circ} \overset{\text{T 3.26 (iii)}}{=} \overline{\mu_{\oplus_{M^\circ}(A_1^\circ,\dots,A_k^\circ)}} = \inf \mathit{\Lambda}_x.$$

For every $\lambda \in \mathit{\Lambda}_x$ there is $x \in \lambda \cdot \oplus_{M^\circ}(A_1^\circ,\dots,A_k^\circ)$ and

$$\exists_{m\in M^\circ} :\; x \in \sum_{i\in[k]} \lambda m_i \star A_i^\circ \iff \exists_{m\in\lambda\star M^\circ} \forall_{i\in[k]} \exists_{x_i\in m_i\star A_i^\circ} :\; x = \sum_{i\in[k]} x_i.$$

The condition $m \in \lambda \star M^\circ$ implies $\mu_{M^\circ}(m) = \sigma_M(m) \leq \lambda$. Similarly there exists a sequence $(x_i)_{i\in[k]} \subseteq L$ with $x = \sum_{i\in[k]} x_i$ and $x_i \in m_i \star A_i^\circ$. Thus $\mu_{A_I^\circ}(x_i) = \sigma_{A_i}(x_i) \leq m_i$ for each $i \in [k]$. Since M is assumed full, containing 0, there is $y \in \lambda \star M^\circ$ where $y \stackrel{\text{def}}{=} (\sigma_{A_1}, \ldots, \sigma_{A_k})(x)$, thus $\lambda \geq \mu_{M^\circ}(y) = \sup_{m\in M} \sum_{i\in[k]} m_i \cdot_{\mathsf{e}} y_i$ and $\sup_{m\in M} \sum_{i\in[k]} m_i \cdot_{\mathsf{e}} y_i \in \Gamma_x$ by construction. This shows for every $\lambda \in \Lambda_x$ there exists $\gamma \in \Gamma_x$ with $\gamma \leq \lambda$. Therefore $\inf \Lambda_x \geq \inf \Gamma_x$.

Let $\gamma \in \Gamma_x$. Then there is $(x_i)_{i\in[k]} \subseteq L$ with $\sum_{i\in[k]} x_i = x$ and $\gamma = \sup_{m\in M} \sum_{i\in[k]} m_i \cdot_{\mathsf{e}} \sigma_{A_i}(x_i)$. Let $y \stackrel{\text{def}}{=} (\sigma_{A_1}(x_1), \ldots, \sigma_{A_k}(x_k))$. Then

$$\gamma = \sup_{m\in M} \sum_{i\in[k]} m_i \cdot_{\mathsf{e}} \sigma_{A_i}(x_i) = \mu_{M^\circ}(y) \implies y \in \gamma \star M^\circ,$$

because M° is closed. Let $m \in \gamma \star M^\circ$ satisfy $m = y$. Then for each $i \in [k]$

$$m_i = \sigma_{A_i}(x_i) = \mu_{A_i^\circ}(x_i) \implies x_i \in m_i \star A_i^\circ,$$

again because each A_i° is closed for $i \in [k]$. It follows that $x = \sum_{i\in[k]} x_i \in \sum_{i\in[k]} m_i \star A_i^\circ$, and $\sum_{i\in[k]} m_i \star A_i^\circ \subseteq \gamma \cdot \oplus_{M^\circ}(A_1^\circ, \ldots, A_k^\circ)$. This shows that $\Gamma_x \subseteq \Lambda_x$ and $\inf \Gamma_x \geq \inf \Lambda_x$ and completes the proof. (Thm. 3.28) ■

Proof of Theorem 3.29. The proof is similar to Thm. 3.28, however the first half is different owing to the varying conditions on M and so we show it in full. Firstly note

$$\begin{aligned}
(M^\triangledown)_\infty &= \left(\bigcap_{m\in M} \operatorname{lev}_{\geq 1}\langle \cdot\,, m\rangle \right)_\infty \\
&= \bigcap_{m\in M} (\operatorname{lev}_{\geq 1}\langle \cdot\,, m\rangle)_\infty \\
&= \bigcap_{m\in M} \operatorname{lev}_{\geq 0}\langle \cdot\,, m\rangle \\
&= M^+.
\end{aligned}$$

This is because the sets $\operatorname{lev}_{\geq 1}\langle \cdot\,, m\rangle$ for $m \in M$ are each convex, thus Prop. 2.8 (vii) lets us pass the asymptotic cone over the union. The next equality follows because $\operatorname{lev}_{\geq 1}\langle \cdot\,, m\rangle$ for $m \in M$ are each co-radiant and Prop. 2.9 (ii) gives $(\operatorname{lev}_{\geq 1}\langle \cdot\,, m\rangle)_\infty = \overline{\operatorname{pos}} \operatorname{lev}_{\geq 1}\langle \cdot\,, m\rangle = \operatorname{lev}_{\geq 0}\langle \cdot\,, m\rangle$. Finally

$M^+ = (\overline{\operatorname{pos}} M)^+ = \mathbb{R}^k_{\geq 0}$ because $\operatorname{pos} M = \mathbb{R}^k_{\geq 0} \setminus \{0\}$ by hypothesis, as it is assumed $M \in \mathcal{M}_\infty(\mathbb{R}^k_{\geq 0})$. Therefore $(M^\triangledown)_\infty = \mathbb{R}^k_{\geq 0}$.

Define the sets

$$\mathit{\Lambda}_x \overset{\text{def}}{=} \left\{\lambda \geq 0 \,\middle|\, x \in \lambda \star \overline{\oplus_{M^\triangledown}(A_1^\triangledown, \dots, A_k^\triangledown)}\right\},$$

and

$$\mathit{\Gamma}_x \overset{\text{def}}{=} \left\{ \inf_{m \in M} \sum_{i \in [k]} m_i \cdot_{\mathsf{e}} \zeta_{A_i}(x_i) \,\middle|\, \exists_{(x_i)_{i\in[k]} \subseteq L} : x = \sum_{i \in [k]} x_i \right\}.$$

From Thm. 3.26 (iv)

$$\zeta_{\square_M(A_1,\dots,A_k)}(x) = \nu_{\square_M(A_1,\dots,A_k)^\triangledown} \overset{\text{T 3.26 (iv)}}{=} \underline{\nu_{\oplus_{M^\triangledown}(A_1^\triangledown,\dots,A_k^\triangledown)}} = \sup \mathit{\Lambda}_x.$$

For every $\lambda \in \mathit{\Lambda}_x$ there is $x \in \lambda \cdot \oplus_{M^\triangledown}(A_1^\triangledown, \dots, A_k^\triangledown)$ and

$$\exists_{m \in M^\triangledown} : x \in \sum_{i\in[k]} \lambda m_i \star A_i^\triangledown \iff \exists_{m \in \lambda \star M^\triangledown} \forall_{i \in [k]} \exists_{x_i \in m_i \star A_i^\triangledown} : x = \sum_{i \in [k]} x_i.$$

The condition $m \in \lambda \star M^\triangledown$ implies $\nu_{M^\triangledown}(m) = \zeta_M(m) \geq \lambda$. Similarly there exists a sequence $(x_i)_{i\in[k]} \subseteq L$ with $x = \sum_{i\in[k]} x_i$ and $x_i \in m_i \star A_i^\triangledown$. Thus $\nu_{A_i^\triangledown}(x_i) = \zeta_{A_i}(x_i) \geq m_i$ for each $i \in [k]$. This shows that $y \geq m$, where $y \overset{\text{def}}{=} (\sigma_{A_1}, \dots, \sigma_{A_k})(x)$, thus $y \in m + \mathbb{R}^k_{\geq 0} \subseteq \lambda \star M^\triangledown + \mathbb{R}^k_{\geq 0}$. Since $M^\triangledown$ is closed convex and $(M^\triangledown)_\infty = \mathbb{R}^k_{\geq 0}$ it follows that $m + \mathbb{R}^k_{\geq 0} \subseteq \lambda \star M$ from Prop. 2.8 (iv). Whence $\lambda \leq \nu_M(y) = \inf_{m\in M} \sum_{i\in[k]} m_i \cdot_{\mathsf{e}} \zeta_{A_i}(x_i)$. This shows for every $\lambda \in \mathit{\Lambda}_x$ there exists $\gamma \in \mathit{\Gamma}_x$ with $\gamma \geq \lambda$. Therefore $\sup \mathit{\Lambda}_x \leq \sup \mathit{\Gamma}_x$.

The rest of the proof now proceeds like Thm. 3.28. (Thm. 3.29) ■

Remark 3.30. It is also possible to prove Theorems 3.28 and 3.28 using the more classical infimal convolution result for the support function of an intersection [viz. 12, p. 34, also 65, Thm. 3.3.2]. The advantage here is to show that the assumptions we have already employed (Theorems 3.19, 3.20 and 3.26) are sufficient, whereas the more classical approach would introduce other assumptions and/or function closures. The necessity of the same conditions used in Section 3.3 in the proofs of Propositions 3.21 and 3.22 is evidence for a deeper structure beyond strict mathematical convenience.

3.6 Related results and conclusion

Penot and Zălinescu [100] study a special case of the dual M-sum called the harmonic sum [p. 30], where $k = 2$ and M corresponds to $I_1 \subseteq \mathbb{R}^2$ (3.3), which has some unique properties from our standpoint. Namely

$$\mu_{I_1}(x) = \mu_{(0,1]\star I_1}(x) = x_1 + x_2 \quad \text{and} \quad \nu_{I_1}(x) = \nu_{[1,\infty)\star I_1}(x) = x_1 + x_2.$$

for all $x \in \mathbb{R}^2_{\geq 0}$. It is also not difficult to verify $(0,1] \star I_1 \in \mathcal{M}_0(\mathbb{R}^2_{\geq 0})$ and $[1,\infty) \star I_1 \in \mathcal{M}_\infty(\mathbb{R}^2_{\geq 0})$. Moreover

$$\mu_{A \diamond B} = \mu_{\oplus_{(0,1]\star I_1}(A,B)} \quad \text{and} \quad \nu_{A \diamond B} = \nu_{\oplus_{[1,\infty)\star I_1}(A,B)}\,.$$

There is $(A \cap B)_\infty \supseteq (\oplus_{I_1}(A,B))_\infty$ (via Lemma 3.14(ii)), and so we obtain the following corollaries from Theorems 3.19 and 3.20.

Corollary 3.31 (Penot and Zălinescu [100, Prop. 3.5])**.** *Let $A, B \subseteq L$. Then*

$$\mu_A +_{\mathsf{e}} \mu_B = \mu_{A \diamond B} \iff (A \diamond B)_\infty = A_\infty \cap B_\infty.$$

Corollary 3.32 (Penot and Zălinescu [100, Prop. 3.7])**.** *Let $A, B \subseteq L$. Then*

$$\nu_A +_{\mathsf{h}} \nu_B = \nu_{A \diamond B} \iff [0,\infty) \star (A \diamond B)_\infty \supseteq A_\infty \cap B_\infty.$$

In the previous sections we have shown general results for two families of two operations $\oplus_M$ and $\Box_M$ operating on the families of radiant and co-radiant subsets of a space L. What differs from our approach in this chapter versus the others listed in Table 3.1 is that we axiomitise the admissible sets M. This axiomitisation may not be minimal, however in Chapter 4 we will encounter a family of sets compatible with the family $\mathcal{M}_\infty(L_{\geq 0})$ for some cone $L_{\geq 0} \subseteq L$.

Part II

Convex Decision Theory

Chapter 4
Convex Decision Theory

The modern theory of probability was formalised by Kolmogorov in his seminal 1933 treatise *Grundbegriffe der Wahrscheinlichkeitsrechnung* [71, 72]. Kolmogorov's axiomitisation introduced the measure theoretic framework of Émile Borel to a tradition that began in the early eighteenth century with Jacob Bernoulli and Abraham de Moivre of using mathematics to model uncertainty in the natural world [118, 119]. Ever since, the concept of a probability distribution has remained the platonic object of study for decision theory, statistics, and (more recently) machine learning. *Probability elicitation* [81, 113] is a game in which a forecaster, having private information about the natural world, is encouraged to make that information public by revealing a forecast in the form of a probability distribution. The forecaster then receives a reward as determined by a pay-off function known as a *scoring rule* [23, 51, 59, 81, 113].[1]

The model of probability elicitation is a natural foundation on which to build a theory of machine learning problems, whereby a risk minimisation, in the sense of Vapnik [133], is reduced fundamentally to eliciting a probability distribution. In a general risk minimisation the forecaster is replaced by a dataset, and the probabilistic forecast takes the form of a statistical model[2],

[1] Jose, Nau, and Winkler [68] provide a detailed discussion comparing families of common utility functions from the economics literature and families of scoring rules from the forecasting literature.

[2] Recall we use the terms *prediction* and *model* interchangeably since the distinction is largely semantic.

and instead of the maximisation of a reward it is more common to consider the minimisation of a punishment. However in a large variety of cases, this alternate representation (namely replacing the distribution by some kind of model) has been shown to be merely an alternate formulation. Following Masnadi-shirazi and Vasconcelos [80] there has been a steady stream of papers abstracting the probability elicitation framework to more and more more general classes of machine learning problems starting with classical binary classification problem [viz. 24, 80, 102] and multiclass classification [141]. The so-called proper–composite representation, coined by Reid and Williamson [102], has proven an important tool for the analysis and design of statistical properties in machine learning models [30, 40, 69, 85, 94].

Having decided upon the probability elicitation framework, our next choice is that of a mathematical structure in which to conduct analysis. In pursuit of the goal of studying the underlying structures common to a variety of machine learning problems, our setting will an ordered topological vector space, with the order inherited from a family of probability distributions. This is the minimal structure needed to, in Section 4.1, define a general risk minimisation with a loss function. In Section 4.2 we define the scoring rules, which are a particular kind of loss function. In Section 4.3 we prove new results on properisation and the proper–composite representation, thus locating a large number of machine learning problems within the probability elicitation framework. Having demonstrated the importance of scoring rules, in Section 4.4 we show how to use the results of Chapter 3 to generate a new family of operations on these scoring rules.

4.1 Loss functions

Let V, P, Ω be arbitrary topological spaces. Unless otherwise noted, we assume L is a vector space of functions $L \subseteq \bar{\mathbb{R}}^{\Omega}$ together with a locally convex, Hausdorff topology. We assume that there is some $P \subseteq \mathfrak{P}(\Omega) \cap L^*$ that induces an ordering on L (as in (2.6)). That is, (L, P^+) forms an ordered topological vector space.

The sets V and P are called *model classes* and Ω is the *outcome space*. The set P may be thought of as a set of distributions we care to distinguish between. Some examples of choices of P are listed in Table 4.1. A *loss* *function* is an operator $\ell : V \to L$. The quantity $\ell(v,\omega) \overset{\text{def}}{=} \ell(v)(\omega)$ is to be interpreted as the penalty when predicting $v \in V$ upon observing the outcome $\omega \in \Omega$. The *ℓ-risk* of v under μ is $\operatorname{risk}_\ell(v,\mu) \overset{\text{def}}{=} \langle \ell(v), \mu \rangle$. Classically, a machine learning problem may be posed as the minimisation of a risk function over an outcome space with respect to a model class [viz. 133]:

$$\underset{v \in V}{\text{minimise}} \quad \operatorname{risk}_\ell(v,\mu). \tag{B}$$

The value of the smallest risk over V, $\operatorname{risk}_\ell(\mu) \overset{\text{def}}{=} \inf_{v\in V} \operatorname{risk}_\ell(v,\mu)$, is called the *Bayes risk*. We omit the qualifying ℓ and μ terms in describing these quantities when they are unambiguous.

Description	P
Differing in mean	$\forall_{\mu,\nu \in P} : \int \omega \nu(\mathrm{d}\omega) \neq \int \omega \mu(\mathrm{d}\omega)$
Compact support	$\exists_{\Omega_0 \subseteq \Omega} \forall_{\mu \in P} :\ \Omega_0$ compact and $\mu(\Omega_0) = 1$
Absolutely continuous (4.6)	$\exists_{\pi \in \mathfrak{P}(\Omega)} : \left\{ f\,\mathrm{d}\pi \;\middle|\; f \in \mathscr{L}_\alpha(\Omega,\pi),\ \int f\,\mathrm{d}\pi = 1,\ f \geq 0 \right\}$

Table 4.1: Example choices for the set P.

(a) Common Asplund spaces.

Ω	L	Asplund
finite	$\mathscr{L}_0(\Omega, \mathbb{R})$	yes
measurable space	$\mathscr{L}_p(\Omega, \lambda)$ (†)	yes
Hausdorff, compact, scattered	$\mathrm{C}(\Omega)$ (*)	yes

(b) Common normal order cones.

L	topology	P	P^+
$\mathscr{L}_p(\Omega, \lambda)$ (†)	$\lvert\cdot\rvert_p$	$\{f\,\mathrm{d}\lambda \in \mathfrak{P}(\Omega) \mid f \in \mathscr{L}_q(\Omega,\lambda)\}$	normal
$\mathrm{C}(\Omega)$	$\lvert\cdot\rvert_\infty$	$\mathfrak{P}(\Omega)$	normal

$*$ Yost [143, Prop. 12] shows that $\mathrm{C}(\Omega)$ is Asplund and only if Ω is Hausdorff, compact, and scattered.
$\dagger$ It is assumed that $1 \leq p < \infty$, with p, q Hölder conjugates, $1/p + 1/q = 1$.

Table 4.2: Example choices for the outcome space Ω, L, P, together with their properties, where λ is a positive measure on Ω equipped with a sigma algebra.

4.1.1 The superprediction set

For a loss function $\ell : V \to L$ the (generalised) *superprediction set* [27, 35, 140, 141] is

$$\operatorname{sp}(\ell) \stackrel{\text{def}}{=} \{x \in L \mid \exists_{v \in V} : \; x \geq_{P^+} \ell(v)\},$$

and its closure is denoted $\overline{\operatorname{sp}}(\ell) \stackrel{\text{def}}{=} \operatorname{cl}(\operatorname{sp} \ell)$. The geometry of the the superprediction set is deeply related to properties of the underlying decision problem, in particular properness [140]; classification calibration and consistency [17, 127]; and mixability [40, 69, 85]. We start by giving some general properties of the superprediction set and its relationship to (B) before analysing the case where ℓ is a scoring rule, a special kind of loss function, in Section 4.2.

Proposition 4.1. *Let $\ell : V \to L$. Then*

(i) $\operatorname{sp}(\ell)$ *is full*

(ii) $\sigma_{\operatorname{sp}(\ell)} = \sigma_{\ell(V)} +_{\mathsf{e}} \iota_{L^-_{\geq 0}}$ *and* $\zeta_{\operatorname{sp}(\ell)} = \zeta_{\ell(V)} -_{\mathsf{h}} \iota_{L^+_{\geq 0}}$.

Proof. (i): From the definition of the order interval

$$[a, b]_{L_{\geq 0}} = a + [0, b - a]_{L_{\geq 0}} = a + L_{\geq 0} \cap (b - a - L_{\geq 0}) \subseteq a + L_{\geq 0}. \tag{4.1}$$

Choose $a, b \in \operatorname{sp}(\ell)$ with $b \geq a$ (so that $[a, b]_{L_{\geq 0}}$ is nonempty). From (4.1) we know $[a, b]_{L_{\geq 0}} \subseteq a + L_{\geq 0}$. Since $a \in \operatorname{sp}(\ell)$, there exists $a_+ \in L_{\geq 0}$ and $a_\ell \in \ell(V)$ so that $a = a_\ell + a_+$ and $a + L_{\geq 0} = a_\ell + a_+ + L_{\geq 0} = a_\ell + L_{\geq 0} \subseteq \ell(V) + L_{\geq 0}$.

(ii): With the usual calculus of support functions [12, p. 31]:

$$\sigma_{\operatorname{sp}(\ell)}(-x^*) = \sigma_{\ell(V) + L_{\geq 0}}(-x^*) = \begin{cases} \sigma_{\ell(V)}(-x^*) & -x^* \in L^-_{\geq 0} \\ \infty & -x^* \notin L^-_{\geq 0} \end{cases}$$

thus $\zeta_{\operatorname{sp}(\ell)} = \zeta_{\ell(V)} -_{\mathsf{h}} \iota_{L^+_{\geq 0}}$. ■

Corollary 4.2. *Let $\ell : V \to L$. Then* risk_ℓ *and* $\zeta_{\operatorname{sp}(\ell)}$ *agree on* $-\operatorname{bc}(\operatorname{sp}(\ell))$.

There is a natural way in which the superprediction set may be connected to the co-radiant sets. When the loss functions are bounded mappings in the topology on L this connection can be sharply characterised using tools of Chapter 2. In particular, this is the case when $\ell : V \to L$ takes values in the positive cone $L_{\geq 0}$. The assumption that $\ell(V) \subseteq L_{\geq 0}$ is generally not onerous since, identifying $L_{\geq 0}$ with P^{++}, we have

$$\ell(V) \subseteq L_{\geq 0} \iff \forall_{\mu \in P} \forall_{v \in V} : \operatorname{risk}_\ell(v, \mu) \geq 0.$$

We will say a loss function $\ell : V \to L$ is *co-radiant* if $\operatorname{sp}(\ell)$ is co-radiant.

Theorem 4.3. *Let $\ell : V \to L_{\geq 0}$. Then*

(i) $\operatorname{sp}(\ell)$ *is co-radiant (co-star-shaped if $0 \notin \ell(V)$); and*

(ii) if $L_{\geq 0}$ is $\sigma(L, L^*)$-normal and $\ell(V)$ is $\sigma(L, L^*)$-closed, then $\operatorname{sp}(\ell)$ *is $\sigma(L, L^*)$-closed.*

Proof. (i): Choose $x \in L_{\geq 0}$. Then

$$\forall_{t>1} : \left(1 - \frac{1}{t}\right)x \in L_{\geq 0} \iff x - \frac{1}{t}x \in L_{\geq 0} \iff tx \geq x, \tag{4.2}$$

where in the final biconditional we used the linearity of the order relation (2.4) to multiply across t. Since ℓ is assumed to map into $L_{\geq 0}$ we have $\operatorname{sp}(\ell) \subseteq L_{\geq 0}$. Let $x \in \operatorname{sp}(\ell)$. By assumption there is $v \in V$ with $x \geq \ell(v)$ and

$$\forall_{t>1} : tx \overset{(4.2)}{\geq} x \geq \ell(v).$$

Therefore $[1, \infty) \cdot \operatorname{sp}(\ell) \subseteq \operatorname{sp}(\ell)$. If $0 \notin \ell(V)$ then there is $0 \notin \operatorname{sp}(\ell)$ and $\operatorname{sp}(\ell)$ is co-star-shaped.

(ii): The result follows from Cor. 2.11 applied to $\ell(V) + L_{\geq 0}$. ■

Corollary 4.4. *If $\ell : V \to L$ is bounded then ℓ is co-radiant if and only if $\ell(V) \subseteq L_{\geq 0}$.*

Proof. The sufficient condition is proven in Thm. 4.3 (i). For the necessary condition assume ℓ is bounded. Then $\ell(V)$ is bounded (thus $\ell(V)_\infty = \{0\}$, via Prop. 2.8 (v)) and $\operatorname{sp}(\ell)_\infty = \left(L_{\geq 0}\right)_\infty = L_{\geq 0}$ from Prop. 2.8 (ii). If $\operatorname{sp}(\ell)$ is co-radiant, then $\operatorname{sp}(\ell) \subseteq \operatorname{sp}(\ell)_\infty = L_{\geq 0}$ from Prop. 2.9 (ii). ■

4.1.2 Subdifferentiability

In the next section (Section 4.2) we will study a class of loss functions for which there is a very natural condition to guarantee the subdifferentiability of the superprediction set co-support function, however it will be convenient (particularly in Section 4.3 and Section 4.3.2) to verify that the assumption of subdifferentiability is not onerous. In Theorem 4.5 we see under mild conditions that the assumption $\widehat{\partial}\zeta_{\mathrm{sp}(\ell)}(\mu) \neq \emptyset$ is equivalent to assuming (B) has a minimiser at μ.

Theorem 4.5. *Let $\ell : V \to L$. There is*

$$\left\{\mu \in L^* \;\middle|\; \operatorname*{arginf}_{v \in V} \operatorname{risk}_\ell(v, \mu) \neq \emptyset\right\} \subseteq \operatorname{dom} \widehat{\partial}\zeta_{\ell(V)}, \tag{4.3}$$

with equality when $L_{\geq 0}$ is normal and $\ell(V) \subseteq L_{\geq 0}$.

Proof. (4.3): Suppose $\mu \in \{\mu' \in L^* \mid \operatorname{arginf}_{v \in V} \operatorname{risk}_\ell(v, \mu') \neq \emptyset\}$. Then there is $v \in \operatorname{arginf}_{v' \in V} \operatorname{risk}_\ell(v, \mu)$ with $\langle \ell(v), \mu\rangle = \operatorname{risk}_\ell(\mu) < \infty$. It follows from Cor. 4.2 that $\langle \ell(v), \mu\rangle = \zeta_{\ell(V)}(\mu)$, and $\ell(v) \in \widehat{\partial}\zeta_{\ell(V)}(\mu)$.

Assume $L_{\geq 0}$ is normal and $\ell(V) \subseteq L_{\geq 0}$: Suppose $\mu \in \operatorname{dom} \widehat{\partial}\zeta_{\ell(V)}$. There exists $x \in \widehat{\partial}\zeta_{\ell(V)}(\mu) \subseteq \operatorname{bd}(\operatorname{co} \ell(V))$ with $\langle x, \mu\rangle = \zeta_{\ell(V)}(\mu)$, consequentially for $k \in [n]$ there are nets

$$(x_{ik})_{i \in I} \subseteq \ell(V) \subseteq L_{\geq 0}, \quad \text{and} \quad (t_{ik})_{i \in I} \subseteq [0, 1]$$

with

$$\forall_{i \in I} : \sum_{k \in [n]} t_{ik} = 1 \quad \text{and} \quad \sum_{k \in [n]} t_{ik} x_{ik} \rightharpoonup x.$$

Without loss of generality assume $(t_{ik})_{i \in I}$ converges for $k \in [n]$. Because $L_{\geq 0}$ is normal, Lem. 2.10 shows that $(t_{ik} x_{ik})_{i \in I}$, for every $k \in [n]$, converge in $\sigma(L, L^*)$. Let $t_k x_k \in \overline{\operatorname{co}}\, \ell(V)$ be the $\sigma(L, L^*)$-limit of $(t_{ik} x_{ik})_{i \in I}$ for $k \in [n]$. It follows that $\langle x, \mu\rangle = \sum_{k \in [n]} t_k \langle x_k, \mu\rangle$.

To see that $\langle x_{k'}, \mu\rangle = \langle x, \mu\rangle$ for every $k' \in [n]$, suppose that there is $k \in [n]$ where $\langle x_k, \mu\rangle > \langle x, \mu\rangle$. This produces a contradiction in the optimality of x in the co-support function minimisation, since we would obtain

$$\langle x, \mu \rangle = \sum_{j \in [n]} t_j \langle x_j, \mu \rangle > \sum_{j \in [n] \setminus \{k\}} u_j \langle x_j, \mu \rangle,$$

where $u_j \stackrel{\text{def}}{=} t_j \big(\sum_{j' \in [n] \setminus \{k\}} t_{j'}\big)^{-1}$ for $j \in [n] \setminus \{k\}$. Similarly if there is $k \in [n]$ with $\langle x_k, \mu \rangle < \langle x, \mu \rangle$, this also produces a similar contradiction directly. Consequentially there exists $k \in [n]$ with

$$x_k \in \operatorname*{arginf}_{x' \in \ell(V)} \langle x', \mu \rangle \implies \operatorname*{arginf}_{v \in V} \langle \ell(v), \mu \rangle \neq \emptyset,$$

which shows

$$\mu \in \left\{ \mu \in L^* \;\middle|\; \operatorname*{arginf}_{v \in V} \operatorname{risk}_\ell(v, \mu) \neq \emptyset \right\},$$

and proves equality in (4.3). ■

In Theorem 4.5 we connected the statistical notion of the existence of a minimiser for a particular distribution and the purely mathematical concept of the domain of the co-support function subdifferential. In Proposition 4.6 we leverage some well-known results in convex analysis to yield some new insights into the existence of a minimiser in (B).

Proposition 4.6. *Let* $\ell : V \to L_{\geq 0}$*. Then* $-\overline{\operatorname{bc}}(\operatorname{sp} \ell) = L_{\geq 0}^+$*. In particular*

(i) $\operatorname{int}(L_{\geq 0}^+) \subseteq \operatorname{dom} \widehat{\partial} \zeta_{\operatorname{sp}(\ell)}$*, and*

(ii) $\operatorname{dom} \widehat{\partial} \zeta_{\operatorname{sp}(\ell)}$*, is dense in* $L_{\geq 0}^+$ *when* L *is a smooth Banach space.*[3]

Proof. Since ℓ takes values in $L_{\geq 0}$ there is

$$L_{\geq 0} = \big(L_{\geq 0}\big)_\infty \supseteq \operatorname{sp}(\ell)_\infty = \big(\ell(V) + L_{\geq 0}\big)_\infty = \big(L_{\geq 0}\big)_\infty = L_{\geq 0}, \tag{4.4}$$

which shows $\operatorname{sp}(\ell)_\infty = L_{\geq 0}$. Hence $\overline{\operatorname{bc}}(\operatorname{sp} \ell) = \operatorname{sp}(\ell)_\infty^- = L_{\geq 0}^-$, and so $-\overline{\operatorname{bc}}(\operatorname{sp} \ell) = L_{\geq 0}^+$. A lower semicontinuous convex function on a Banach space is always continuous on the interior of its domain [99, Prop. 3.3] and its subdifferential is nonempty at points of continuity [99, Thm. 3.25]. Finally,

[3] A Banach space is said to be a smooth when its norm is differentiable on the unit sphere [16, p. 34].

the Ekeland–Lebourg theorem [99, Thm 4.65] shows that the domain of the subdifferential of a lower semicontinuous convex function is dense in its domain for a smooth Banach space. ■

4.2 Scoring rules

A *scoring rule* is a particular, classical, kind of loss function for which the set of predictions is a subset of distributions on the outcome space [81, 113].[4] That is, $V = P$ in the notation of Section 4.1, and $\mathfrak{s} : P \to L$. A scoring rule $\mathfrak{s}$ is said to be *P-proper* [51, 62, 81, 113] if

$$\forall_{\mu \neq \nu \in P} : \ \langle \mathfrak{s}(\mu), \mu \rangle \leq \langle \mathfrak{s}(\nu), \mu \rangle, \tag{4.5}$$

and *strictly P-proper* if (4.5) holds with strict inequality.[5]

In continuous spaces it is a common practice assume (Ω, π) is a measurable space and choose some

$$P \subseteq P_\pi^\alpha \stackrel{\text{def}}{=} \left\{ f \,\mathrm{d}\pi \;\middle|\; f \in \mathscr{L}_\alpha(\Omega, \pi), \ \int f \,\mathrm{d}\pi = 1, \ f \geq 0 \right\}, \tag{4.6}$$

and $P_\pi \stackrel{\text{def}}{=} P_\pi^0$. So that one may work instead with a set of density functions [viz. 36, 51, 62, 125]. This construction makes it easy to ensure P is a subset of some space L^* that satisfies certain desirable technical conditions like separability and reflexivity. Another motivation for this relaxation is that many continuous space scoring rules are only defined on a set of densities — most notably the logarithmic scoring rule [52].

Unlike other most of the other approaches mentioned we have made no assumption on the convexity of P. The induced ordering $\geq$, however, is the same whether one takes P or $\operatorname{co} P$ (or $\overline{\operatorname{co}}\, P$), this is because $P^+ = (\overline{\operatorname{co}}\, P)^+$ (see (2.9)). It should be unsurprising, then, to learn that it is without loss of generality that one may assume P is convex. This is a point we touch on

---

[4] Typically in the machine learning literature the name “scoring rule” is used interchangeably with “loss function” [24, 127, 141], but it is useful conceptually for our purposes to draw a distinction. This convention is consistent with Grünwald and Dawid [59] and others [23, 34, 51].

[5] In the statistics and decision theory literature [viz. 23, 51, 59, 62, 81, 113] the quantity $-\operatorname{risk}_{\mathfrak{s}}(v, \mu)$ is called the *expected score* under μ when predicting v and is often notated $S(v, \mu)$ for some $v \in V$ and $\mu \in P$.

(a) General scoring rules.

Name	Symbol	$\mathfrak{s}(\mu)(\,\cdot\,)$	Proper*	P
Brier score	$\mathfrak{s}_{\mathrm{B}r}$	$1 - 2\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\omega) + \int \big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\omega)\big)^2 \pi(\mathrm{d}\omega)$	S.P.	P_π
Pseudospherical	$\mathfrak{s}_\alpha$	$-\frac{\mathrm{d}\mu}{\mathrm{d}\nu}^{\alpha-1}\big(\int \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\omega)^\alpha \pi(\mathrm{d}\omega)\big)^{-\beta}$ (†)	S.P.	P_π^α
Logarithmic‡	$\mathfrak{s}_1$	$-\log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(\omega)$	S.P.	$\operatorname{rint} P_\pi$

(b) Discrete outcome space scoring rules.

Name	Symbol	$\mathfrak{s}(\mu)(i)$	Proper*	P
Zero–one		$[\![i \notin \operatorname{argmax}_{j\in[k]} \mu_j]\!]$	P.	$\mathfrak{P}([k])$
Brier score	$\mathfrak{s}_{\mathrm{B}r}$	$1 - 2\mu_i + \sum_{j\in[k]}(\mu_j)^2$	S.P.	$\mathfrak{P}([k])$
Pseudospherical	$\mathfrak{s}_\alpha$	$-\mu_i^{\alpha-1}\big(\sum_{j\in[k]}\mu_j^\alpha\big)^{-\beta}$ (†)	S.P.	$\mathfrak{P}([k])$
Logarithmic‡	$\mathfrak{s}_1$	$-\log(\mu_i)$	S.P.	$\operatorname{rint}\mathfrak{P}([k])$

∗ Scoring rules are characterised as either proper (P.) or strictly proper (S.P.) with respect to the corresponding set P in the adjacent column.
† It is assumed that $\alpha > 1$, and α, β are Hölder conjugates, $1/\alpha + 1/\beta = 1$.
‡ The logarithmic score is obtained from the pseudospherical score in the limit as $\alpha \to 1$.

Table 4.3: A selection of common proper scoring rules over the measured outcome space (Ω, π), most of which are collected by Gneiting and Raftery [51]. The set P_π is defined in (4.6). When Ω is finite, $\Omega \simeq [k]$, it is common to take π as the counting measure, $\pi A \stackrel{\text{def}}{=} |A|$ for $A \subseteq \Omega$. Whence $P_\pi = \mathfrak{P}([k])$, and we obtain the formulations in (b). We use the shorthand $\mu_i \stackrel{\text{def}}{=} \mu\{i\}$ for $i \in [k]$, $\mu \in \mathfrak{P}([k])$.

again in Section 4.3. Some common scoring rules for discrete and general topological spaces Ω are listed in Table 4.3.

4.2.1 The selection representation

There is a very convenient relationship between proper scoring rules and the subdifferential which will form the basis of many results in this chapter.

Theorem 4.7. *Let $\mathfrak{s} : P \to L$ be a scoring rule. Then $\mathfrak{s}$ is*

(*i*) *P-proper if and only if $\mathfrak{s}(\mu) \in \widehat{\partial}\zeta_{\mathfrak{s}(P)}(\mu)$ for every $\mu \in P$, and*

(*ii*) *strictly P-proper if and only if $\mathfrak{s}$ is injective and $\zeta_{\mathfrak{s}(P)}$ is Gâteaux differentiable on P.*

Proof. (i): Assume $\mathfrak{s}$ is P-proper. Then for $\mu, \nu \in P$ there is $\langle \mathfrak{s}(\mu), \mu\rangle \le \langle \mathfrak{s}(\nu), \mu\rangle$ and

$$\forall_{\mu\in P} : \ \langle \mathfrak{s}(\mu), \mu\rangle = \inf_{\nu\in P}\langle \mathfrak{s}(\nu), \mu\rangle = \inf_{v\in \mathfrak{s}(P)}\langle v, \mu\rangle = \zeta_{\mathfrak{s}(P)}(\mu). \qquad (4.7)$$

Then, for every $\mu, \nu \in P$

$$\begin{aligned}\langle \mathfrak{s}(\mu), \mu\rangle &\leq \langle \mathfrak{s}(\nu), \mu\rangle + (\langle \mathfrak{s}(\nu), \nu\rangle - \langle \mathfrak{s}(\nu), \nu\rangle)\\ &\overset{(4.7)}{\Longleftrightarrow} \langle \mathfrak{s}(\nu), \nu - \mu\rangle \leq \zeta_{\mathfrak{s}(P)}(\nu) - \zeta_{\mathfrak{s}(P)}(\mu).\end{aligned}$$

This shows $\mathfrak{s}(\nu) \in \widehat{\partial}\zeta_{\mathfrak{s}(P)}(\nu)$. Now assume $\mathfrak{s}(\mu) \in \widehat{\partial}\zeta_{\mathfrak{s}(P)}(\mu)$ for every $\mu \in P$. Then $\mathfrak{s}(\mu) \in \operatorname{arginf}_{v\in \mathfrak{s}(P)}\langle v, \mu\rangle$, which implies the converse claim.

(ii): The subdifferential of a Gâteaux differentiable convex function is precisely the singleton of the gradient [149, Thm 2.4.4]. Thus

$$\forall_{\mu\in P} : \operatorname*{arginf}_{v\in \mathfrak{s}(P)}\langle v, \mu\rangle = \widehat{\partial}\zeta_{\mathfrak{s}(P)}(\mu) = \{\mathfrak{s}(\mu)\}. \tag{4.8}$$

It follows from (4.8) that if $\mathfrak{s}$ is injective $\operatorname{arginf}_{\nu\in P}\langle \mathfrak{s}(\nu), \mu\rangle = \{\mu\}$ for all $\mu \in P$ and $\mathfrak{s}$ is strictly P-proper. To complete the proof observe that any strictly proper scoring rule must be injective or else a contradiction is obtained in (4.5). ■

Since a strictly proper scoring rule is automatically proper, it follows from Theorem 4.7(i) that $\mathfrak{s} : P \to L$ is strictly P-proper if and only if it is injective and

$$\forall_{\mu\in P} : \widehat{\partial}\zeta_{\mathfrak{s}(P)}(\mu) = \{\mathfrak{s}(\mu)\}.$$

A version of Theorem 4.7 was first stated (without proof) by McCarthy [81] for the case of a strictly proper scoring rule, and it since has been noted by several authors [34, 36, 51, 62, 141]. However most of the works cited do not make full use of the subdifferential selection representation (Theorem 4.7) in the same way as we will in the subsequent sections. We obtain immediately the following straight-forward corollaries, which appear to be new.

Corollary 4.8. *Assume $\mathfrak{s} : P \to L$ is P-proper (resp. strictly P-proper). Then $P \subseteq -\operatorname{bc} \mathfrak{s}(P)$ (resp. $P \subseteq -\operatorname{int}(\operatorname{bc} \mathfrak{s}(P))$).*

Proof. From Thm. 4.7 we have $\operatorname{dom} \widehat{\partial}\zeta_{\mathfrak{s}(P)} \supseteq P$. There is always $\operatorname{dom} \widehat{\partial}\zeta_{\mathfrak{s}(P)} \subseteq \operatorname{dom} \zeta_{\mathfrak{s}(P)}$. If $\mathfrak{s}$ is strictly P-proper then Thm. 4.7 shows $\zeta_{\mathfrak{s}(P)}$ is differentiable on P, therefore $\zeta_{\mathfrak{s}(P)}$ is differentiable on an open neighbourhood of P (possibly equal to P itself), and $P \subseteq \operatorname{int}(\operatorname{dom} \zeta_{\mathfrak{s}(P)}(\mu))$. ■

If the subdifferential of a convex function is a singleton on an open where that function is continuous, it is differentiable on that open set [99, Prop. 3.4, Cor. 3.26].

Corollary 4.9. *When L is a normed space, the Bayes risk of every P-proper scoring rule is continuous on a neighbourhood of P.*

Proof. For a strictly P-proper scoring rule $\mathfrak{s} : P \to L$. Because $P \subseteq -\operatorname{int}(\operatorname{bc} \mathfrak{s}(P))$ (from Cor. 4.8), and the fact that support function is always lower semicontinuous, it follows [via 99, Prop. 3.4] that $\zeta_{\mathfrak{s}(P)}(\mu)$ is continuous on a neighbourhood of P. ■

A Banach space L is called an *Asplund space* [8] if every continuous convex function, defined on an open convex subset $M \subseteq L$ is Fréchet differentiable on a G_δ set D, that is dense in M. Since continuous convex functions in an Asplund space are differentiable on a dense subset of their domains, a great number of P-proper scoring rules are almost strictly proper in these spaces.

Corollary 4.10. *Suppose L is an Asplund space and $\mathfrak{s} : P \to L$ is P-proper and injective, with a Bayes risk that's finite on a neighbourhood of P. Then there is a dense subset $P_\delta \subseteq P$ for which $\mathfrak{s}$ is strictly P_δ-proper.*

Proof. By assumption $\mathfrak{s}$ is finite on a neighbourhood U of P. Since L is an asplund space there is a G_δ dense subset $D \subseteq U$ on which $\zeta_{\mathfrak{s}(P)}$ is differentiable. Define $P_\delta \stackrel{\text{def}}{=} D \cap P$. Then $\widehat{\partial} \zeta_{\mathfrak{s}(P)}$ is a singleton on P_δ [2, Thm. 7.17], and Thm. 4.7 shows $\mathfrak{s}$ is strictly P_δ-proper. ■

4.2.2 Properness and convexity

Theorem 4.3 suggests some basic strategies to establish whether the superprediction set of a loss function is closed.Theorem 4.13 aids in this endeavour by showing there is a strong relationship between the properness of a continuous scoring rule and the topology of its superprediction set, both in terms of convexity and closure.

Lemma 4.11 (Hahn–Banach [12, Thm. 2, p. 27])**.** *Let A be a subset of a Hausdorff locally convex vector space L. Then*

$$\overline{\operatorname{co}}\, A = \{x \in L \mid \forall_{x^* \in L^*} : \langle x, x^* \rangle \leq \sigma_A(x^*)\}.$$

Lemma 4.12 (Fan [41, Thm. 5])**.** *Let P be a compact set in a topological vector space. Let f be a real-valued function defined on $P \times P$ so that*

1. *$y \mapsto f(x,y)$ is lower semicontinuous for all $x \in P$,*

2. *$x \mapsto f(x,y)$ is quasi-concave for all $y \in P$.*

Then

$$\min_{y\in P}\sup_{x\in P} f(x,y) \le \sup_{x\in P} f(x,x).$$

Theorem 4.13. *Equip L and L^* with topologies so that $P \subseteq L^*$ is compact and $\mathfrak{s} : P \to L$ is continuous. If $\mathfrak{s}$ is P-proper, then $\operatorname{sp}(\mathfrak{s})$ is closed and convex.*

Proof. From the ordering assumption on L, there is $x \in \operatorname{sp}(\mathfrak{s})$ precisely when

$$\begin{aligned}
\exists_{\nu\in P} :\ x \ge_{P^+} \mathfrak{s}(\nu) \iff& \exists_{\nu\in P}\forall_{\mu\in P} :\ \langle x,\mu\rangle \ge \langle \mathfrak{s}(\nu),\mu\rangle \\
\iff& \exists_{\nu\in P} :\ \sup_{\mu\in P}\langle \mathfrak{s}(\nu)-x,\mu\rangle \le 0 \\
\iff& \min_{\nu\in P}\sup_{\mu\in P}\langle \mathfrak{s}(\nu)-x,\mu\rangle \le 0. \qquad (4.9)
\end{aligned}$$

Since $\zeta_{\operatorname{sp}(\mathfrak{s})} = \zeta_{\mathfrak{s}(P)} -_{\mathsf{h}} \iota_{P^{++}}$ (from Prop. 4.1 (ii)), for every $x \in L$ we have

$$\begin{aligned}
\inf_{x^*\in L^*}\Big(\langle x,x^*\rangle -_{\mathsf{e}} \zeta_{\operatorname{sp}(\mathfrak{s})}(x^*)\Big) &= \inf_{x^*\in L^*}\Big(\langle x,x^*\rangle +_{\mathsf{e}} \iota_{P^{++}}(x^*) -_{\mathsf{e}} \inf_{\mu\in P}\langle \mathfrak{s}(\mu),x^*\rangle\Big) \\
&= \inf_{x^*\in P^{++}}\Big(\langle x,x^*\rangle -_{\mathsf{e}} \inf_{\mu\in P}\langle \mathfrak{s}(\mu),x^*\rangle\Big). \qquad (4.10)
\end{aligned}$$

When $x \in \overline{\operatorname{co}}(\operatorname{sp}\mathfrak{s})$, Lem. 4.11 yields

$$\begin{aligned}
\Big[\forall_{\mu\in P^{++}} :\ \langle x,\mu\rangle \ge \inf_{\nu\in P}\langle \mathfrak{s}(\nu),\mu\rangle\Big] \overset{(4.10)}{\iff}& 0 \le \inf_{\mu\in P^{++}}\sup_{\nu\in P}\langle x-\mathfrak{s}(\nu),\mu\rangle \\
\implies& 0 \le \inf_{\mu\in P}\sup_{\nu\in P}\langle x-\mathfrak{s}(\nu),\mu\rangle \\
\iff& 0 \ge \sup_{\mu\in P}\inf_{\nu\in P}\langle \mathfrak{s}(\nu)-x,\mu\rangle, \qquad (4.11)
\end{aligned}$$

where in the second line we exploited $P \subseteq P^{++}$.

For $x \in L$ let $f_x(\mu,\nu) \overset{\text{def}}{=} \langle \mathfrak{s}(\nu)-x,\mu\rangle$. Since $\mathfrak{s}$ is continuous, $f_x(\,\cdot\,,\nu)$ is continuous for all $\nu \in P$. The Fan minimax inequality (Lem. 4.12) applied to f_x gives

$$\sup_{\mu\in P}\langle \mathfrak{s}(\mu) - x, \mu\rangle \geq \min_{\mu\in P}\sup_{\nu\in P}\langle \mathfrak{s}(\nu) - x, \mu\rangle. \tag{4.12}$$

If $\mathfrak{s}$ is P-proper then $\inf_{\nu\in P}\langle \mathfrak{s}(\nu) - x, \mu\rangle = \langle \mathfrak{s}(\mu) - x, \mu\rangle$, and (4.12) becomes

$$\sup_{\mu\in P}\inf_{\nu\in P}\langle \mathfrak{s}(\nu) - x, \mu\rangle \geq \min_{\mu\in P}\sup_{\nu\in P}\langle \mathfrak{s}(\nu) - x, \mu\rangle. \tag{4.13}$$

Fix $x \in \overline{\mathrm{co}}(\mathrm{sp}\ \mathfrak{s})$ and assume $\mathfrak{s}$ is P-proper. Then

$$\begin{aligned} 0 \overset{(4.11)}{\geq} \sup_{\mu\in P}\inf_{\nu\in P}\langle \mathfrak{s}(\nu) - x, \mu\rangle &\overset{(4.13)}{\geq} \min_{\mu\in P}\sup_{\nu\in P}\langle \mathfrak{s}(\nu) - x, \mu\rangle. \\ &\overset{(4.9)}{\iff} x \in \mathrm{sp}(\mathfrak{s}). \end{aligned}$$

This shows that $\overline{\mathrm{co}}(\mathrm{sp}(\mathfrak{s})) \subseteq \mathrm{sp}(\mathfrak{s})$. The reverse inclusion is immediate. ■

Results similar to Theorem 4.13 have been claimed or proved by other authors under a variety of stricter assumptions. It is usually the case that P is assumed convex or the entirety of $\mathfrak{P}(\Omega)$, and Ω is assumed finite [23, 35, 51, 62, 85, 125, 141]. The setting of Dawid [35] is the closest to ours, and provides a brief proof sketch for the case of continuous Ω [35, Lem. 3]. As we have already seen when P is $\mathfrak{P}(\Omega)$, the induced inequality is indeed pointwise (Proposition 2.1) which allows Theorem 4.13 to verify the existing results mentioned.

4.2.3 Dual characterisations

As we have seen, the properness of a scoring rule can be characterised in terms of a selection property of a subdifferential of a convex function (Theorem 4.7). Since the subdifferential of a convex function can be inverted, Proposition 4.14 provides the following new characterisation of properness.

Lemma 2.5, together with Theorem 4.7 yields a dual characterisation of properness.

Proposition 4.14. *Let $\mathfrak{s} : V \to L$. Then $\mathfrak{s}$ is P-proper if and only if for* *all* $\mu \in P$

$$\mu \in -\mathrm{N}_{\overline{\mathrm{co}}(\mathrm{sp}\,\mathscr{s})}(\mathscr{s}(\mu)).$$

Proof. From Thm. 4.7 $\mathscr{s}$ is P-proper if and only if $\mathscr{s}(\mu) \in \widehat{\partial}\zeta_{\mathrm{sp}(\mathscr{s})}(\mu)$, for all $\mu \in P$, which by Lem. 2.5(ii) is equivalent to $\mu \in -\mathrm{N}_{\overline{\mathrm{co}}(\mathrm{sp}\,\mathscr{s})}(\mathscr{s}(\mu))$ for all $\mu \in P$. ■

For the *co-star-shaped* scoring rules there is an additional characterisation, using the co-radiant Minkowski duality from Section 2.3.

Theorem 4.15. *Assume* $0 \notin P$ *and* $\mathscr{s} : P \to L_{\geq 0}$ *is co-star-shaped. Then* $\mathscr{s}$ *is* P*-proper if and only if for all* $\mu \in P$

$$\frac{\mu}{\zeta_{\mathrm{sp}(\mathscr{s})}(\mu)} \in \widehat{\partial}\nu_{\overline{\mathrm{sp}}(\mathscr{s})}(\mathscr{s}(\mu)). \tag{4.14}$$

Proof. Since $\mathscr{s}$ is co-radiant $\overline{\mathrm{sp}}(\mathscr{s})$ is closed, co-radiant. Assume $\mathscr{s}$ is P-proper and fix $\mu \in P$. Then Thm. 4.7 implies

$$\zeta_{\mathrm{sp}(\mathscr{s})}(\mu) = \langle \mathscr{s}(\mu), \mu \rangle. \tag{4.15}$$

Since $\mathscr{s}(\mu) \in \mathrm{pos}(\mathrm{sp}\,\mathscr{s}) \subseteq L_{\geq 0} \setminus \{0\}$ and $\mu \in P \subseteq L_{\geq 0}^{+} \setminus \{0\}$, (4.15) implies $\langle \mathscr{s}(\mu), \mu \rangle = \zeta_{\mathrm{sp}(\mathscr{s})}(\mu) = \nu_{\overline{\mathrm{sp}}(\mathscr{s})^{\triangledown}}(\mu) > 0$. It follows that $\mu \in \mathrm{pos}(\mathrm{sp}(\ell)^{\triangledown})$. Thus Prop. 2.7(ii) implies (4.14).

Next assume (4.14) holds and fix $\mu \in P$. Then

$$\frac{\mu}{\zeta_{\mathrm{sp}(\mathscr{s})}(\mu)} \in \widehat{\partial}\nu_{\overline{\mathrm{sp}}(\mathscr{s})}(\mathscr{s}(\mu)) \implies \frac{\langle \mathscr{s}(\mu), \mu \rangle}{\zeta_{\mathrm{sp}(\mathscr{s})}(\mu)} = \nu_{\overline{\mathrm{sp}}(\mathscr{s})}(\mathscr{s}(\mu)). \tag{4.16}$$

Since $\mathscr{s}(\mu) \in \mathrm{sp}(\mathscr{s})$ we have $\nu_{\overline{\mathrm{sp}}(\mathscr{s})}(\mathscr{s}(\mu)) \leq 1$ and (4.16) implies

$$\langle \mathscr{s}(\mu), \mu \rangle \leq \zeta_{\mathrm{sp}(\mathscr{s})}(\mu). \tag{4.17}$$

The definition of the co-support function means that (4.17) must be an equality. Thus $\mathscr{s}(\mu) \in \widehat{\partial}\zeta_{\mathrm{sp}(\mathscr{s})}(\mu)$ for all $\mu \in P$ and Thm. 4.7 implies $\mathscr{s}$ is P-proper. ■

4.3 Bayes acts, properisations, and link functions

Let $\ell : V \to L$ be a loss function. Fix $\mu \in P$. If there is some $v_\mu \in V$ for which $\langle \ell(v_\mu), \mu\rangle \leq \langle \ell(v), \mu\rangle$ for all $v \in V$ then Grünwald and Dawid [59] call v_μ the *Bayes act* for μ [see also 23, 35]. The Bayes act allows two interesting constructions: One may take an arbitrary scoring rule and reparameterise it to obtain a proper scoring rule. Brehmer and Gneiting [23] call this procedure *properisation*. Even more generally, one may take an arbitrary loss function and reparameterise it so that it can be described using a proper scoring rule and what Reid and Williamson [102, 103] call a *canonical link function*.

The subdifferential characterisation of properness in Theorem 4.7 allows us to apply the theory developed in Section 2.4.2 to generate several existence results for these two applications.

4.3.1 Properisation

We say a scoring rule $\mathscr{s}_{\mathrm{P}} : P \to L$ is a properisation of $\mathscr{s}$ if

$$\forall_{\mu \in P} : \ \mathscr{s}_{\mathrm{P}}(\mu) = \mathscr{s}(v(\mu))$$

where the mapping $\mu \mapsto v(\mu)$ satisfies $v(\mu) \in \operatorname{arginf}_{\nu \in P}\langle \mathscr{s}(\nu), \mu\rangle$ for all $\mu \in P$. Any properisation is automatically a proper scoring rule [23, Thm. 1].

The theory we have established already in Sections 2.4.2 and 4.2 allows us to state an extremely general properisation result. In Theorem 4.16 and Corollary 4.17 measurability refers to measurability with respect to the $\tau(L, L^*)$- and $\sigma(L^*, L)$-Borel sigma algebras.

Theorem 4.16. *Assume L is a Banach space with separable dual. Assume P is $\sigma(L^*, L)$-Borel measurable. Let $\mathscr{s} : P \to L$ be a scoring rule with a Bayes risk function that is finite on a neighbourhood of P. Then there is a P-proper, measurable scoring rule $\mathscr{s}_{\mathrm{P}} : P \to L$ with the same risk function as $\mathscr{s}$, and $\mathscr{s}_{\mathrm{P}}$ is a properisation of $\mathscr{s}$ on a dense subset $P_\delta \subseteq P$.*

Proof. For simplicity of notation let $F \stackrel{\text{def}}{=} -\partial\sigma_{-\mathscr{s}(P)}$. If risk_ℓ is finite on a neighbourhood U of P then it is continuous on $U \supseteq P$ [99, Prop. 3.3] and $P \subseteq \operatorname{dom} F$ [via 99, Thm. 3.25]. Since L has a separable dual it is an Asplund space [99, Thm. 3.97]. Convex functions that are continuous on an open set U of any Asplund space are always differentiable on a dense G_δ subset $U_\delta \subseteq U$. Any dense subset of U is also dense in P, whence there exists the dense subset $P_\delta \stackrel{\text{def}}{=} U_\delta \cap P$ on which F is single-valued [via 2, Thm. 7.17]. Using Lem. 2.16 we observe that F has a measurable selection on a neighbourhood of P. We denote its restriction to P by $\mathfrak{d}_{\mathrm{P}} : P \to L$. Since P is $\sigma(L^*, L)$-Borel measurable, the restriction is measurable.

Since $\mathfrak{d}_{\mathrm{P}}$ selects F, it is automatically a P-proper scoring rule (Thm. 4.7). Any Bayes act properisation of $\mathfrak{d}$ is necessarily a selection, and therefore agrees with $\mathfrak{d}_{\mathrm{P}}$ on P_δ. Finally because the $\mathfrak{d}$ risk function is 1-homogeneous and $\mathfrak{d}_{\mathrm{P}}$ selects its subdifferential,

$$\forall_{\mu \in P} : \ \operatorname{risk}_{\mathfrak{d}_{\mathrm{P}}}(\mu) = \operatorname{risk}_{\mathfrak{d}}(\mu),$$

[via 149, Thm. 2.4.14(iii)]. That is, the $\mathfrak{d}$ and $\mathfrak{d}_{\mathrm{P}}$ Bayes risk functions agree. ∎

As we have already mentioned in Section 4.2.2, it is a common assumption (implicit or explicit) to assume the set P is convex. The duality correspondence already ensures the order induced on L via P is the same as the order induced by $\overline{\mathrm{co}}\, P$. Similarly, under the fairly mild conditions of Theorem 4.16 we can use the same approach to measurably extend a scoring rule defined on a nonconvex P. This lends credence to the P convexity assumption.

Corollary 4.17. *Assume all the assumptions of Theorem 4.16 are met, and additionally assume that P is $\sigma(L^*, L)$-closed, $\mathfrak{d} : P \to L$ is measurable and P-proper. Then $\mathfrak{d}$ has measurable extension to $\overline{\mathrm{co}}(P)$ that is $\overline{\mathrm{co}}(P)$-proper.*

Proof. The corollary follows from the proof of Thm. 4.16 observing that $P \subseteq \operatorname{int}(\operatorname{dom} F)$ implies $\overline{\mathrm{co}}\, P \subseteq \operatorname{int}(\operatorname{dom} F)$ because the subdifferential domain is convex. The extension $\mathfrak{d}_{\mathrm{ext}}$ may now be constructed using

$$\forall_{\mu \in \overline{\mathrm{co}}\, P} : \ \mathfrak{d}_{\mathrm{ext}}(\mu) \stackrel{\text{def}}{=} \begin{cases} \mathfrak{d}(\mu) & \mu \in P \\ \mathfrak{d}_{\mathrm{P}}(\mu) & \mu \in \overline{\mathrm{co}}\, P \setminus P, \end{cases}$$

where $\mathfrak{d}_{\mathrm{P}}(\mu)$ is as in Thm. 4.16. Properness follows from Thm. 4.7, and measurability follows from the measurability of $\overline{\mathrm{co}}\, P \setminus P$. ∎

4.3.2 Link functions

The idea of the *link function* dates to Nelder and Wedderburn [92, see also 82, §2] who introduced it as part of the definition of a generalised linear model, wherein the link function connects a prediction with the parameters of an exponential family distribution. In this sense link functions are a mapping from a set of predictions V to a set of probability distributions P.

The idea has since been resurrected by Reid and Williamson [103] for binary classification problems, and Williamson, Vernet, and Reid [141] for multiclass classification. The setting of multiclass classification corresponds to a conditional density estimation problem over a discrete topological outcome space, and will be the subject of Section 4.3.3. However, first we build upon several ideas from Williamson, Vernet, and Reid [141] in two directions of generality. Firstly from a discrete to a general topological outcome space, and secondly from differentiable to suitably finite Bayes risk functions.

As with Theorem 4.16, measurability in Theorem 4.18 is proven with respect to $\tau(L, L^*)$- and $\sigma(L^*, L)$-Borel sigma algebras.

Theorem 4.18. *Assume L is a Banach space with separable dual, in which $P \subseteq L^*$ is $\sigma(L^*, L)$-compact. Let $\ell : V \to L$ be a Borel loss function with a Bayes risk function that's finite (resp. differentiable) on a $\sigma(L^*, L)$-neighbourhood of P. Then there is a Borel function $\tau : \overline{\mathrm{co}}\,\ell(V) \to P$, a P-proper (resp. strictly P-proper) Borel scoring rule $\mathfrak{s} : P \to L$ and a $\sigma(L^*, L)$-dense subset $P_\delta \subseteq P$ (resp. $P_\delta = P$) so that*

$$\forall_{\mu \in P} : \inf_{v \in V} \mathrm{risk}_\ell(v, \mu) = \inf_{v \in V} \mathrm{risk}_{\mathfrak{s} \circ \tau \circ \ell}(v, \mu)$$

and if ℓ is injective

$$\forall_{\mu \in P_\delta} : \operatorname*{arg\,inf}_{v \in V} \mathrm{risk}_\ell(v, \mu) = \operatorname*{arg\,inf}_{v \in V} \mathrm{risk}_{\mathfrak{s} \circ \tau \circ \ell}(v, \mu).$$

Proof. For simplicity of notation let $F \stackrel{\text{def}}{=} -\partial \sigma_{-\ell(V)}$. If we equip L^* with $\sigma(L^*, L)$ then $L = (L^*, \sigma(L^*, L))^*$ [2, Thm. 5.93]. Since L has a separable dual (by assumption) it is separable itself [38, Prop. 3.6.14]. Since $L = (L^*, \sigma(L^*, L))^*$ is separable, L^* is an Asplund space [99, Thm. 3.97].

By an identical argument to the proof of Thm. 4.16 (observing that P is $\sigma(L^*, L)$-Borel measurable) F has a measurable selection on a neighbourhood of P which is a P-proper scoring rule when restricted to P (Lem. 2.16) which we denote $\mathfrak{s} : P \to L$. By construction (cf. the lower inverse in Section 2.4.2) we have $\operatorname{dom}(F^{-1} \cap P) = F(P)$. From Lem. 2.18 the map $F^{-1} \cap P = -\mathrm{N}_{\overline{\mathrm{co}}\,\ell(V)} \cap P$ has a measurable selection which we denote $\tau : F(P) \to P$ (the equality is due to Lem. 2.5).

Since risk_ℓ is differentiable on U_δ, it follows that $\mathfrak{s}$ is invertible on P_δ with U_δ-inverse τ [cf. 99, Cor. 3.26]. Pick $\mu \in P_\delta$. If ℓ is an injection then

$$v_\mu \in \operatorname*{arg\,inf}_{v \in V} \operatorname{risk}_\ell(v, \mu) \iff \ell(v_\mu) \in F(\mu),$$

because F is single-valued at μ. Next because $\mathfrak{s}$ selects F we have $\mathfrak{s}(\mu) = \ell(v_\mu)$, and $\tau(\ell(v_\mu)) = \mu$. Therefore

$$\forall_{\mu \in P_\delta} : \operatorname*{arg\,inf}_{v \in V} \operatorname{risk}_\ell(v, \mu) = \operatorname*{arg\,inf}_{v \in V} \operatorname{risk}_{\mathfrak{s} \circ \tau \circ \ell}(v, \mu).$$

If the Bayes risk function is differentiable on U then $\mathfrak{s}$ is strictly proper by Thm. 4.7 and P_δ can be taken to be P. ■

Theorem 4.18 shows that a great many risk minimisation problems may be reparameterised in such a way that they can be expressed in terms of the minimisation of a scoring rule risk over a family of distributions. In particular, there is a natural way the set of predictions V can mapped into an a set of distributions P. The mathematics underpinning this surprising relationship is just the duality between measures and functions, combined with the natural concavity of the function $\zeta_{\ell(V)}$. Moreover it argues for the necessity and generality of proper scoring rules in describing (B).

4.3.3 Decomposable risk minimisation

Until now we have assumed $(L, L_{\geq 0})$ is a vector space of functions $\Omega \to \bar{\mathbb{R}}$. If the outcome space is decomposable for some topological spaces X, Y, that is $\Omega = X \times Y$, then the risk minimisation problem (B) is called *regression* when Y is continuous, and *classification* when Y is discrete [133]. When Ω has such a structure, we refer to (B) as the *decomposable risk minimisation* problem. One possible approach to analyse the the decomposable risk minimisation problem would be to replace L by a set of functions $X \times Y \to \bar{\mathbb{R}}$. The question then is what ordering is natural to impose on this space. Similarly to Section 4.1 we could specify a positive cone P^+ via a family of measures $P \subseteq \mathfrak{P}(X \times Y)$. This approach, however, would not allow us to exploit the intrinsic structure of the decomposable problem, and yield similar results to Sections 4.1 and 4.2. Instead, we assume the structure of the preceding section on on the space Y, and specialise our investigation to the *decomposable loss functions* (defined below).

For this section we assume Unless otherwise noted, we assume L is a vector space of functions $L \subseteq \bar{\mathbb{R}}^Y$ together with a locally convex, Hausdorff topology, and there is a subset $P \subseteq L^*$ so that (L, P^+) is an ordered topological vector space. Due to the added structure in this setting we refine some of the notions from Section 4.1. The following conventions end up simplifying the notation that follows. A loss function is a Borel mapping $\ell : V \times X \to L$ and we let $\ell(v)(x,y) \stackrel{\text{def}}{=} \ell(v,x)(y)$ so that $\ell(v) \in \mathscr{L}_0(X \times Y)$ for all $v \in V$. The *evaluation operator* at $x \in X$ is

$$\mathrm{ev}_x : \mathscr{L}_0(X \times Y) \to \mathscr{L}_0(Y) \quad \text{with} \quad \forall_{f \in \mathscr{L}_0(X\times Y)} : \ \mathrm{ev}_x f \stackrel{\text{def}}{=} f(x, \,\cdot\,).$$

This construction ensures for all $x \in X$ that $\mathrm{ev}_x \ell(v)$ is a function in L. Though it is possible to define scoring rules directly on $\mathfrak{P}(X \times Y)$, we will consider scoring rules as mapping $P \to L$, that is, just as we had in Section 4.2 (with Ω replaced by Y).

The evaluation operator generates the *pull-back order* in $\mathscr{L}_0(X \times Y)$ via

$$\forall_{x\in X} : \ (\mathrm{ev}_x)^{-1}(L_{\geq 0}) = \{f \in \mathscr{L}_0(X \times Y) \mid \mathrm{ev}_x f \in L_{\geq 0}\}$$

and so the positive cone in $\mathscr{L}_0(X \times Y)$ is

$$(\textstyle\bigcup_{x\in X} \mathrm{ev}_x)^{-1}(L_{\geq 0}) = \{f \in \mathscr{L}_0(X \times Y) \mid \forall_{x\in X}\forall_{\mu\in P} : \ \langle \mathrm{ev}_x f, \mu\rangle \geq 0\}.$$

In particular, observing that the adjoint of the evaluation operator is the Dirac product,[6]

$$\begin{aligned}(\textstyle\bigcup_{x\in X} \mathrm{ev}_x)^{-1}(L_{\geq 0}) &= \{f \in \mathscr{L}_0(X \times Y) \mid \forall_{x\in X}\forall_{\mu\in P} : \ \langle f, \delta_x \times \mu\rangle \geq 0\} \\ &= (\delta_X \times P)^+,\end{aligned}$$

[6] That is $\langle \mathrm{ev}_x f, \mu\rangle = \langle f, \delta_x \times \mu\rangle$ for all $x \in X$, $\mu \in L^*$.

where $\delta_X \times P \stackrel{\text{def}}{=} \bigcup_{x \in X}\{\delta_x \times \mu \mid \mu \in P\}$. Then $(\mathscr{L}_0(X \times Y), (\delta_X \times P)^+)$ is an ordered vector space, it is with respect to this ordering that we define $\operatorname{sp}(\ell)$:

$$\operatorname{sp}(\ell) \stackrel{\text{def}}{=} \Big\{ x \in \mathscr{L}_0(X \times Y) \;\Big|\; \exists_{v \in V} : \; x \geq_{(\delta_X \times P)^+} \ell(v) \Big\}.$$

We now assume V is a collection of functions $v : X \to Z$. A decomposable loss function, ℓ, is defined using a mapping $g : Z \to L$ so that

$$\forall_{v \in V} \forall_{(x,y) \in X \times Y} : \; \ell(v)(x,y) = g(v(x))(y). \tag{4.18}$$

In practice many conditional prediction problems are specified using a loss function of the form (4.18). With the pull-back order on $\mathscr{L}_0(X \times Y)$ there is a close relationship between $\operatorname{sp}(\ell)$ and $\operatorname{sp}(g)$.

Proposition 4.19. *Suppose $\ell : V \to L$ is a decomposable rule loss function for $g : Z \to L$. Then*

$$\operatorname{sp}(h) \subseteq \bigcup_{x \in X} \operatorname{ev}_x(\operatorname{sp} \ell),$$

with equality if $Z = V(X) \stackrel{\text{def}}{=} \bigcup_{v \in V}\{v(x) \in Z \mid x \in X\}$.

Proof. Choose any $x \in X$ and $f \in \operatorname{sp}(\ell)$. It follows that there exists $v \in V$ with $f \geq_{(\delta_X \times P)^+} \ell(v)$. Because ev_x is a positive operator, that is, $\operatorname{ev}_x((\delta_X \times P)^+) \subseteq (P^+)$ for all $x \in X$,

$$\begin{aligned} f \geq_{(\delta_X \times P)^+} \ell(v) \implies & \qquad\quad \forall_{x \in X} : \; \operatorname{ev}_x f \geq_{P^+} \operatorname{ev}_x \ell(v) \\ & \implies \forall_{x \in X} \exists_{z \in V(X)} : \; \operatorname{ev}_x f \geq_{P^+} g(z), \end{aligned}$$

where $V(X) \stackrel{\text{def}}{=} \bigcup_{v \in V} v(X)$. This shows $\bigcup_{x \in X} \operatorname{ev}_x \operatorname{sp}(\ell) \subseteq \operatorname{sp}(g)$.

Now assume $V(X) = Z$ and choose $f \in \operatorname{sp}(g)$. It follows that

$$\exists_{z \in Z} : \; f \geq_{P^+} g(z) \implies \exists_{v_f \in V} \exists_{x_f \in X} : \; f \geq_{P^+} g(v_f(x_f)) = \operatorname{ev}_{x_f} \ell(v_f).$$

Let $h(x,y) \stackrel{\text{def}}{=} \max\{\ell(v_f)(x,y), f(x)\}$. Then $\operatorname{ev}_{x_f} h = f$ and $h \geq_{(\delta_X \times P)^+} \ell(v)$, which shows $f \in \bigcup_{x \in X} \operatorname{ev}_x \operatorname{sp}(\ell)$. Thus $\operatorname{sp}(g) \subseteq \bigcup_{x \in X} \operatorname{ev}_x \operatorname{sp}(\ell)$. ∎

Corollary 4.20. *Let $\ell : V \times X \to L$ is a decomposable rule loss function for*

$g : Z \to L$. Then $\zeta_{\mathrm{sp}(g)} \geq \inf_{x \in X} \zeta_{\mathrm{sp}(\ell)}(\delta_x \times \cdot\,)$, with equality if $V(X) = Z$.

Proof. For all $\mu \in L^*$, from Prop. 4.19

$$\begin{aligned}
\zeta_{\mathrm{sp}(g)}(\mu) &= \inf_{f \in \mathrm{sp}(g)} \langle f, \mu \rangle \\
&\geq \inf_{h \in \bigcup_{x \in X} \mathrm{ev}_x \mathrm{sp}(\ell)} \langle h, \mu \rangle \\
&= \inf_{x \in X} \inf_{h \in \mathrm{ev}_x \mathrm{sp}(\ell)} \langle h, \mu \rangle \\
&= \inf_{x \in X} \inf_{h \in \mathrm{sp}(\ell)} \langle h, \delta_x \times \mu \rangle \\
&= \inf_{x \in X} \zeta_{\mathrm{sp}(\ell)}(\delta_x \times \mu),
\end{aligned}$$

with equality if $V(X) = Z$. ■

4.4 Scoring rule aggregation

It is interesting that in spite of the generality of the notion of a proper scoring rule, one typically encounters only a handful of concrete examples in the literature [e.g. 24, 51].[7] Consequentially, choosing a scoring rule for a statistical model itself similarly may present its own problems with some theorists recommending instead using a combination of scoring rules [84]. We have seen in Chapter 3 that there is a rich structure in the family of co-radiant sets with the family operations $\oplus_M$ and $\Box_M$. It is our hope that by introducing these operations to the family of proper scoring rules, that we may contribute simultaneously each of these problems.

In Sections 4.1 and 4.2 we saw that a large number of proper scoring rules have an analytically simple representation in terms of the superprediction set, which is convex and co-radiant. By combining the results of Chapter 3 with Sections 4.1 and 4.2 we develop a simple composition operation for the scoring rules which preserves properness. The rich set of polarity results from Sections 3.4 and 4.2.3 then lets us calculate the corresponding link functions.

[7] Most of these are listed in Table 4.3.

4.4.1 Superprediction sets

Before we can proceed, it is helpful to verify that a large number of scoring rules have superprediction sets satisfying the conditions of the theorems and corollaries in Chapter 3.

Proposition 4.21. *Let* $\mathfrak{s} : P \to P^+ \setminus \{0\}$ *be* $\sigma(L^*, L)$*-continuous and* P*-proper, where* P *is* $\sigma(L^*, L)$*-compact. Then* $\operatorname{sp}(\mathfrak{s}) \in \mathcal{M}_\infty(P^+)$.[8]

Proof. Closed and convex: Thm. 4.13 shows that $\operatorname{sp}(\mathfrak{s})$ is closed and convex.

$\operatorname{pos}(\operatorname{sp} \mathfrak{s}) = P^+ \setminus \{0\}$:

As part of the proof of Prop. 4.6 we calculated $P^+ = \operatorname{sp}(\mathfrak{s})_\infty$. Therefore

$$P^+ \setminus \{0\} \overset{(4.4)}{=} \operatorname{sp}(\mathfrak{s})^+_\infty \setminus \{0\} \overset{\text{P 2.9 (ii)}}{=} \overline{\operatorname{pos}}(\operatorname{sp} \mathfrak{s}) \setminus \{0\}. \tag{4.19}$$

We will now show that $\overline{\operatorname{pos}}(\operatorname{sp} \mathfrak{s}) \setminus \{0\} = \operatorname{pos}(\operatorname{sp} \mathfrak{s})$. Take a $\sigma(L, L^*)$-convergent net $(x_i)_{i \in I} \subseteq \operatorname{pos}(\operatorname{sp} \mathfrak{s})$ with limit $x \neq 0$. There are nets $(t_i)_{i\in I} \subseteq \mathbb{R}_{>0}$ and $(l_i)_{i \in I} \subseteq \operatorname{sp}(\mathfrak{s})$ with $x_i = t_i l_i$ for all $i \in I$. If either (t_i) or (l_i) fail to converge, $t_i\langle l_i, x^*\rangle \to \infty$ for any $x^* \in \operatorname{sp}(\mathfrak{s})^+$, and so both nets must converge. Let t and l be their limits, with (l_i) converging in $\sigma(L, L^*)$. If (t_i) converges at 0 then $x = 0l = 0$ which contradicts the assumption $x \neq 0$. If (t_i) converges at some $t > 0$, then $t_i l_i \rightharpoonup tl$. Because $\operatorname{sp}(\mathfrak{s})$ is closed convex, it is $\sigma(L, L^*)$-closed.

This shows $x \in \operatorname{pos}(\operatorname{sp} \mathfrak{s})$ and $\operatorname{pos}(\operatorname{sp} \mathfrak{s})$ is $\sigma(L, L^*)$-closed. Because $\operatorname{sp}(\mathfrak{s})$ is convex, $\operatorname{pos}(\operatorname{sp} \mathfrak{s})$ is convex and therefore it is strongly closed. It follows from (4.19) that $\overline{\operatorname{pos}}(\operatorname{sp} \mathfrak{s}) = P^+ \setminus \{0\}$.

Containing an order unit: From Prop. 2.9 (ii) $(\operatorname{sp}(\mathfrak{s}))_\infty = P^+$, and by assumption $\operatorname{sp}(\mathfrak{s}) \subseteq P^+$. Taking any order unit $e \in P^+$, and $x \in \operatorname{sp}(\mathfrak{s})$ we obtain from Prop. 2.8 (iv) that $e + x \in \operatorname{sp}(\mathfrak{s})$ and $e + x$ is an order unit of P^+. ■

By the set $\mathbb{R}^{[k]}_{\geq 0}$ we mean the collection of functions $[k] \to \mathbb{R}_{\geq 0}$, which is isomorphic, as a vector space, to $\mathbb{R}^k_{\geq 0}$.

[8] Recall from Section 3.3 that $\mathcal{M}_\infty(K)$ denotes the collection of subsets M of the cone K which are closed, convex, containing a K-order unit and have $\operatorname{pos} M = K \setminus \{0\}$.

Corollary 4.22. *Let* $\mathfrak{s} : \mathfrak{P}([k]) \to \mathbb{R}_{\geq 0}^{[k]}$ *be continuous and proper. Then* $\operatorname{sp}(\mathfrak{s}) \in \mathcal{M}_\infty(\mathbb{R}_{\geq 0}^{[k]})$.

Let $\mathfrak{s}_1, \ldots, \mathfrak{s}_k : P \to L_{\geq 0}$ be a sequence of continuous P-proper scoring rules, and let $\mathfrak{s}_0 : \mathfrak{P}([k]) \to \mathbb{R}_{\geq 0}^k$ be a $\mathfrak{P}([k])$-proper scoring rule. Then Theorem 4.13 and Proposition 4.1 show that $\operatorname{sp}(\mathfrak{s}_i)$ is closed, convex and co-radiant for $i \in [k] \cup \{0\}$.[9] It follows from Theorem 3.11, Proposition 3.12, and Corollary 4.22 that

$$\oplus_{\operatorname{sp}(\mathfrak{s}_0)}(\operatorname{sp}(\mathfrak{s}_1), \ldots, \operatorname{sp}(\mathfrak{s}_k)) \quad \text{and} \quad \oplus_{\operatorname{sp}(\mathfrak{s}_0)^\triangledown}(\operatorname{sp}(\mathfrak{s}_1)^\triangledown, \ldots, \operatorname{sp}(\mathfrak{s}_k)^\triangledown),$$

are both convex and co-radiant. In this section will find a proper scoring rule $\mathfrak{s}_\oplus : P \to L$ and link function $\mathfrak{s}_\square : P \to L$ so that

$$\operatorname{sp}(\mathfrak{s}_\oplus) = \oplus_{\operatorname{sp}(\mathfrak{s}_0)}(\operatorname{sp}(\mathfrak{s}_1), \ldots, \operatorname{sp}(\mathfrak{s}_k))$$

and

$$\forall_{\mu \in \mathrm{dom}} : \ \tau(\mu) \in \widehat{\partial}\, \zeta_{\operatorname{sp}(\mathfrak{s}_\oplus)}(\mathfrak{s}_\oplus(\mu)).$$

4.4.2 M-sums of scoring rules

Using Theorem 3.4, Corollaries 3.5 and 4.2 for all $\mu \in -\bigcap_{i \in [k]} \operatorname{bc}(\operatorname{sp} \mathfrak{s}_i)$

$$\operatorname{risk}_{\mathfrak{s}_\oplus}(\mu) \overset{\text{C4.2}}{=} \zeta_{\operatorname{sp}(\mathfrak{s}_\oplus)} \overset{\text{T3.4}}{=} \inf_{m \in \operatorname{sp}(\mathfrak{s}_0)} \sum_{i \in [k]} m_i \cdot_{\mathsf{h}} \zeta_{\operatorname{sp}(\mathfrak{s}_i)}(\mu). \tag{4.20}$$

Since $\mathfrak{s}_0$, as a selection of $\widehat{\partial}\zeta_{\operatorname{sp}(\mathfrak{s}_0)}$, is defined on $\mathfrak{P}([k])$, we need to normalise the vector $(\zeta_{\operatorname{sp}(\mathfrak{s}_1)}(\mu), \ldots, \zeta_{\operatorname{sp}(\mathfrak{s}_k)}(\mu))$ so that it lies in this set. Observe

$$\forall_{c > 0} \forall_{\mu \in \mathfrak{P}(\Omega)} : \ \widehat{\partial}\, \zeta_{\operatorname{sp}(\mathfrak{s}_k)}(c\mu) = \widehat{\partial}\, \zeta_{\operatorname{sp}(\mathfrak{s}_k)}(\mu).$$

Therefore we define

$$s|_\mu \overset{\text{def}}{=} (\zeta_{\operatorname{sp}(\mathfrak{s}_1)}, \ldots, \zeta_{\operatorname{sp}(\mathfrak{s}_k)})(\mu) \in \mathbb{R}_{\geq 0}^k,$$

[9] Recall we use the pointwise-ordering on $\mathbb{R}^k$ to define $\operatorname{sp}(\mathfrak{s}_0)$.

and

$$\tilde{s}|_\mu \overset{\text{def}}{=} \frac{1}{\upmu_{\mathfrak{P}([k])}(s|_\mu)} \star s|_\mu \in \mathfrak{P}([k]). \tag{4.21}$$

The gauge $\upmu_{\mathfrak{P}([k])}(s|_\mu)$ ensures that $\tilde{s}|_\mu$ lies in $\mathfrak{P}([k])$ for every $\mu \in P$. Then using Lemma 2.21 to subdifferentiate (4.20) we have

$$\begin{aligned}\widehat{\partial}\zeta_{\mathrm{sp}(\mathfrak{s}_\oplus)}(\mu) &\overset{\text{R2.22}}{\supseteq} \bigcup_{m \in \widehat{\partial}\zeta_{\mathrm{sp}(\mathfrak{s}_0)}(\mathfrak{s}|_\mu)} \sum_{i\in[k]} m_i \star \widehat{\partial}\zeta_{\mathrm{sp}(\mathfrak{s}_i)}(\mu) \\ &\overset{\text{T4.7}}{\ni} \sum_{i\in[k]} \mathfrak{s}_0(\tilde{s}|_\mu)(i) \cdot \mathfrak{s}_i(\mu).\end{aligned}$$

Let us now define

$$\mathfrak{s}_\oplus : P \to L \quad \text{with} \quad \mathfrak{s}_\oplus(\mu) \overset{\text{def}}{=} \sum_{i\in[k]} \mathfrak{s}_0(\tilde{s}|_\mu)(i) \cdot \mathfrak{s}_i(\mu).$$

Since $\mathfrak{s}_\oplus$ enjoys the subdifferential representation it is automatically P-proper (Theorem 4.7). Next, because $\zeta_{\mathrm{sp}(\mathfrak{s}_\oplus)} = \zeta_{\oplus_{\mathrm{sp}(\mathfrak{s}_0)}(\mathrm{sp}(\mathfrak{s}_1),\ldots,\mathrm{sp}(\mathfrak{s}_k))}$, taking the subdifferential at 0 shows $\overline{\mathrm{co}}(\mathrm{sp}\ \mathfrak{s}_\oplus) = \overline{\mathrm{co}}(\oplus_{\mathrm{sp}(\mathfrak{s}_0)}(\mathrm{sp}(\mathfrak{s}_1),\ldots,\mathrm{sp}(\mathfrak{s}_k)))$ and Theorems 3.7 and 3.11 yield

$$\mathrm{sp}(\mathfrak{s}_\oplus) = \oplus_{\mathrm{sp}(\mathfrak{s}_0)}(\mathrm{sp}(\mathfrak{s}_1),\ldots,\mathrm{sp}(\mathfrak{s}_k)). \tag{4.22}$$

4.4.3 Dual M-sum scoring rules

We use essentially the same approach as Section 4.4.2 to compute the scoring rule $\mathfrak{s}_\square$. However to apply Theorem 3.29 we need to show a sufficient condition for the asymptotic cone of $\square_{\mathrm{sp}(\mathfrak{s}_0)}(\mathrm{sp}(\mathfrak{s}_1),\ldots,\mathrm{sp}(\mathfrak{s}_k))$. Since $\mathrm{sp}(\mathfrak{s}_i)$ is convex for each $i \in [k]$, Lemma 3.14 shows

$$\Big(\square_{\mathrm{sp}(\mathfrak{s}_0)}(\mathrm{sp}(\mathfrak{s}_1),\ldots,\mathrm{sp}(\mathfrak{s}_k))\Big)_\infty \overset{\text{L3.14(i)}}{\supseteq} \bigcap_{i\in[k]} (\mathrm{sp}\ \mathfrak{s}_i)_\infty.$$

Similar to (4.21) ait will simplify things to introduce some notation. Let $\mu_{[k]}$ denote a sequence $(\mu_i)_{i\in[k]} \subseteq L^*$, so that $\mu_{[k]} \in (L^*)^k$ and

$$s|_{\mu_{[k]}} \overset{\text{def}}{=} (\zeta_{\mathrm{sp}(\mathfrak{s}_1)}(\mu_1),\ldots,\zeta_{\mathrm{sp}(\mathfrak{s}_k)}(\mu_k)),$$

and

$$\tilde{s}|_{\mu_{[k]}} \stackrel{\text{def}}{=} \frac{1}{\mu_{\mathfrak{P}([k])}(s|_{\mu_{[k]}})} \star s|_{\mu_{[k]}}. \tag{4.23}$$

Using Theorem 3.29 and Corollary 3.16, and our notation in (4.23), for all

$$\mu \in \operatorname{int} \sum_{i\in[k]} \operatorname{bc}(\operatorname{sp} \ell_i) \stackrel{\text{C3.16}}{\subseteq} \operatorname{bc}(\operatorname{sp} \ell_\square)$$

there is

$$\begin{aligned} \operatorname{risk}_{\ell_\square}(\mu) &\stackrel{\text{T3.29}}{=} \sup\left\{ \inf_{m\in\operatorname{sp}(\ell_0)} \sum_{i\in[k]} m_i \cdot_{\mathsf{h}} \zeta_{\operatorname{sp}(\ell_i)}(\mu_i) \;\middle|\; \mu = \sum_{i\in[k]} \mu_i \right\} \\ &= \sup\left\{ \sum_{i\in[k]} \ell_0(s|_{\mu_{[k]}})(i) \cdot_{\mathsf{h}} \zeta_{\operatorname{sp}(\ell_i)}(\mu_i) \;\middle|\; \mu = \sum_{i\in[k]} \mu_i \right\}. \end{aligned}$$

Next let $T(\mu)$ denote the set

$$\left\{ (\mu_i)_{i\in[n]} \subseteq L^* \;\middle|\; \mu = \sum_{i\in[k]} \mu_i,\ \operatorname{risk}_{\ell_\oplus}(\mu) = \sum_{i\in[k]} \ell_0(s|_{\mu_{[k]}})(i) \cdot_{\mathsf{h}} \zeta_{\operatorname{sp}(\ell_i)}(\mu_i) \right\}.$$

Then, again using Lemma 2.21, we have

$$\widehat{\partial}\zeta_{\operatorname{sp}(\ell_\square)}(\mu) \supseteq \left\{ \sum_{i\in[k]} \ell_0(\tilde{s}|_{\mu_{[k]}})(i) \cdot \ell_i(\mu_i) \;\middle|\; (\mu_1,\ldots,\mu_k) \in T(\mu) \right\}.$$

It is harder to get an exact form for $\ell_\square$ that parallels $\ell_\oplus$ in (4.22) and ensures a result like (3.16). To do so one would need to construct a selection of $\mu \mapsto T(\mu)$, however with a selection of this sort, a similar subdifferential argument to (4.22) would yield the same superprediction set equality.

4.5 Conclusion

Many machine learning problems are not framed in terms of probability elicitation, but rather as a risk minimisation over some class of functions. To free ourselves of the constraints of probability elicitation we introduced the link functions, grounded in the duality of convex sets, which provides a means by which we can generalise the probability elicitation framework to an arbitrary set of predictions in a consistent manner. In many of our theorems we have made no assumption of differentiability or smoothness, and have instead exploited the natural concavity of the risk functional to supply these properties. However, when the stronger assumption of differentiability is satisfied, we recover the stronger existing results in the literature that have been provided in a finite dimensional setting. By studying machine learning problems in the abstract we are forced to consider the shared underlying structures between problems. An example of the simplicity obtained through abstraction is encapsulated very nicely in the study of link functions, wherein the seemingly complicated idea of probabilistic inference just reduces to finding the inverse of a studied, well-behaved monotone operator.

In convex analysis, Hörmander's theorem provides a bridge between sublinear functions and closed radiant sets via the support function. It is perhaps surprising that such a connection exists, and that there is a dual calculus for sublinear functions and their corresponding sets. We have generalised this calculus not only to the family of sublinear operations on a set of support functions, but with a set of superlinear operations on a set of co-support functions — the less common concave counterparts. Similarly, with the introduction of the generalised superprediction set, we can transform a machine learning problem into a member of a family of sets, the calculus of which we then inherit.

Contrary to our approach here, much of modern machine learning research starts with a particular problem that one seeks to solve, whether this is to build a classifier for a particular domain, or to estimate some quantity of interest. In the preceding sections we have built a theoretical framework that goes in the opposite direction; beginning with a fundamental quantity of interest (the probability distribution) and the simplest means of its discovery (a proper scoring rule). In order to endow our theory with a rich analytic structure, we observed and employed deep connections to convex analysis, nonsmooth analysis, and the theory of co-radiant sets, all of which arise from our basic premise of probability elicitation with a proper scoring rule.

Part III

Adversarial Learning

Chapter 5

Boosted Density Estimation

In the emerging area of *Generative Adversarial Networks (GANs)* [53] a binary classifier, called a *discriminator*, is used learn a highly efficient sampler for a data distribution, combining what would traditionally be two steps — first learning the density function from a family of densities, then fine-tuning a sampler — into one. Interest in this field has sparked a series of formal inquiries and generalisations describing GANs in terms of (among other things) divergence minimisation [5, 96]. Using a similar framework to Nowozin, Cseke, and Tomioka [96], Grover and Ermon [58] make a preliminary analysis of an algorithm that takes a series of iteratively trained discriminators to estimate a density function[1]. The cost of this approach, insofar as we have been able to devise, is that one forgoes learning an efficient sampler (as with a GAN), and must make do with classical sampling techniques to sample from the learned density. We leave the issue of efficient sampling from these density as an open problem, and instead focus on analysing the densities learned with formal convergence guarantees under reasonable assumptions (Table 5.2). Previous formal results have established a range of guarantees, from qualitative convergence [58], to geometric convergence rates [129], with numerous results in between.

In learning a density function iteratively, most previous approaches [e.g. 60, 77, 86, 107, 129, 130] have investigated a single update rule, not unlike

[1] Grover and Ermon [58] call this procedure "multiplicative discriminative boosting".

	Updates	Rate (Big-Oh)		Assumptions
Dai *et al.* [33]	particles	$\mathrm{KL}(\mu_\star, \mu_0)$	[33, Thm. 6]	smoothness, Lipschitz, measure concentration, etc.
Tolstikhin *et al.* [129]	density	$\log \mathrm{JS}(\mu_\star, \mu_0)$	[129, Cor. 3]	updates close to optimal
Grover and Ermon [58]	density	none		none
This work	binary classifiers	$\log \mathrm{KL}(\mu_\star, \mu_0)$	Theorem 5.9	essentially bounded classifiers, weak learning, weak dominance

Table 5.1: Summary of related works and results in terms of the nature of the updates, the best quoted rate of convergence, and the structural assumptions.

	I_φ	$\varphi(t)$	$\varphi^*(t^*)$	$\varphi'(t)$	$(\varphi^* \circ \varphi')(t)$
Kullback–Liebler	KL	$t \log t$	$\exp(t^* - 1)$	$\log t + 1$	t
Reverse KL	rKL	$\log(t)$	$-\log(-t^*) - 1$	$-1/t$	$\log t - 1$
Hellinger	-	$(\sqrt{t} - 1)^2$	$3(t^* - 1)^{-1} - 1$	$1 - 1/t$	$\sqrt{t} - 1$
Pearson	χ^2	$(t-1)^2$	$t^*(4 + t^*)/4$	$2(t-1)$	$t^2 - 1$
GAN	GAN	$t \log t - (t+1)\log(t+1)$	$-\log(1 - \exp(t^*))$	$\log(\frac{t}{t+1})$	$\log(\frac{1}{t+1})$

Table 5.2: Some common φ-divergences and their variational components.

Frank–Wolfe optimisation [46], where a sequence (x_t), an initial point y_0 and a set of numbers $(\alpha_t) \subseteq [0, 1]$ is chosen satisfying

$$y_t = \psi(\alpha_t x_t + (1 - \alpha_t) y_{t-1}), \tag{5.1}$$

for some function ψ, so that an objective function (usually a divergence) is minimised along (y_t). Grover and Ermon [58] is a recent exception to (5.1) wherein alternative choices are explored. Few works in this area are accompanied by convergence proofs, and even fewer provide convergence rates [60, 77, 107, 129, 130].

To establish convergence and/or bound the convergence rate all approaches necessarily make structural assumptions or approximations on the parameters involved in (5.1). These assumptions can be on the (local) variation of the divergence [60, 91, 130], the true distribution or the quality of the updates [33, 58, 60, 77], the step size [86, 129], the previous history of updates [33, 107], and so on. Often in order to produce the best geometric convergence bounds, the update is usually required required to be close to the optimal one [129, Cor. 2, 3]. Table 5.1 compares the best results of the leading three to our approach. We give for each of them the updates aggregated, the assumptions on which rely the results and the *rate* to come close to a fixed value of Kullback–Liebler divergence (Jensen-Shannon divergence, for Tolstikhin *et al.* [129]), which is just the order of the number of iterations necessary, hiding the other dependences for simplicity.

However, it must be kept in mind that for many of these works [viz. 129] the primary objective is to develop an efficient black box sampler for $\mu_\star$, in particular for large dimensions. Our objective however is to focus on furtive lack of formal results on the densities and convergence, and deferring the problem of learning an efficient sampler.

5.1 From discriminators to densities

Throughout this chapter, (Ω, μ_0) is a Borel space with $\mu_0 \in \mathfrak{P}(\Omega)$. We denote the μ_0 absolutely continuous probability measures by $\mathfrak{P}(\Omega, \mu_0) \stackrel{\text{def}}{=} \{\mu \in \mathfrak{P}(\Omega) \mid \mu \ll \mu_0\}$ and the target distribution will be denoted $\mu_\star \in \mathfrak{P}(\Omega, \mu_0)$. For a distributions $\mu, \nu \in \mathfrak{P}(\Omega)$, the *Radon–Nikodym deriva-*

tive for ν is the function $\mathrm{d}\nu/\mathrm{d}\mu \in \mathscr{L}_0(\Omega, \mu)$ that satisfies $\nu A = \int_A \cdot \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\,\mathrm{d}\mu$ for all $A \in \mathscr{B}(\Omega)$. For $\mu \in \mathfrak{P}(\Omega)$ and $f \in \mathscr{L}_0(\Omega)$ the expectation operator is $\mathrm{E}_\mu f \overset{\text{def}}{=} \int f\,\mathrm{d}\mu$.

An important tool of ours are the φ-*divergences* of information theory [1, 32, 104]. For $\mu, \nu \in \mathfrak{P}(\Omega)$ with $\mu \ll \nu$, the φ-divergence of ν from μ is

$$\mathrm{I}_\varphi(\mu, \nu) \overset{\text{def}}{=} \int \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu,$$

where it always assumed that $\varphi \in \mathscr{L}_0(\mathbb{R}, \mathbb{R}_{\geq 0})$ is convex and lower semicontinuous, and often additionally the normalisation condition $\varphi(1) = 0$. Every φ-divergence has a *variational representation* via the Fenchel conjugate [viz. 93, also 104]:

$$\begin{aligned}
\mathrm{I}_\varphi(\mu, \nu) &= \int \varphi^{**}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\nu \\
&= \int \sup_{t \in \mathbb{R}}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\omega) - \varphi^*(\omega)\right)\nu(\mathrm{d}\omega) \\
&= \sup_{f \in \mathscr{L}_0(\Omega, \mathbb{R}_{\geq 0})}\left(\int f\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu - \varphi^* \circ f\,\mathrm{d}\nu\right) \\
&= \sup_{f \in \mathscr{L}_0(\Omega, \mathbb{R}_{\geq 0})}\Big(\mathrm{E}_\mu[f] + \mathrm{E}_\nu[-\varphi^* \circ f]\Big).
\end{aligned} \tag{5.2}$$

The variational representation of a φ-divergence has been leveraged by Nowozin, Cseke, and Tomioka [96] to show the equivalence between the GAN saddle point objective of Goodfellow *et al.* [53] and the minimisation of φ-divergence.

When φ is differentiable, it is a common result that the supremum in (5.2) is attained for $\varphi' \circ \mathrm{d}\mu/\mathrm{d}\nu$ [53, 96], so that we may reparameterise (5.2) to obtain the following minimisation problem

$$\underset{d \in \mathscr{L}_0(\Omega, \mathbb{R}_{\geq 0})}{\text{minimise}} \quad \mathrm{E}_\mu[-\varphi' \circ d] + \mathrm{E}_\nu[\varphi^* \circ \varphi' \circ d]. \tag{5.3}$$

Remark 5.1. The reparameterised problem (5.3) shows that φ' serves as a canonical choice for the so-called link function of Nowozin, Cseke, and Tomioka [96].

The objective in (5.3) is easily identified with the expectation of the loss function

$$\begin{aligned} \ell : \mathscr{L}_0(\Omega, \mathbb{R}_{>0}) &\to \mathscr{L}_0(\Omega \times [2], \mathbb{R}) \quad \text{where} \\ \forall_{d \in \mathscr{L}_0(\Omega, \mathbb{R}_{>0})} : \ \ell(d)(\omega, y) &\stackrel{\text{def}}{=} \begin{cases} (-\varphi' \circ d)(\omega) & y = 1 \\ (\varphi^* \circ \varphi' \circ d)(\omega) & y = 2, \end{cases} \end{aligned}$$

under the joint distribution

$$\pi(\mathrm{d}\omega, \mathrm{d}y) \stackrel{\text{def}}{=} \frac{1}{2}\Big(\mu(\mathrm{d}\omega)\,\delta_1(\mathrm{d}y) + \nu(\mathrm{d}\omega)\,\delta_2(\mathrm{d}y)\Big).$$

That is, a classical binary classification problem [17, 24, 95, 102–104], where the task is to classify samples with the labels $\{1, 2\}$. In fact, several common binary classification loss functions can seen to be special cases of (5.3) as evidenced by Table 5.2, wherein we define the Kullback–Liebler divergence, which will be most useful in Sections 5.2 and 5.3.

With a smoothness assumption on φ we can replace the set $\mathscr{L}_0(\Omega, \mathbb{R}_{\geq 0})$ with $\mathscr{L}_0(\Omega, \mathbb{R}_{>0})$ in (5.3). We can further reparameterise the set $\mathscr{L}_0(\Omega, \mathbb{R}_{>0})$ with any bijection to the set $\mathscr{L}_0(\Omega, \mathbb{R})$. The exponential function has several useful properties and so this is the one we use. In the sections that follow, for every $t \in \mathbb{N}$ and $d_t \in \mathscr{L}_0(\Omega, \mathbb{R}_{>0})$ we let $c_t \stackrel{\text{def}}{=} \log \circ d_t$, or equivalently for every $c_t \in \mathscr{L}_0(\Omega, \mathbb{R})$ we let $d_t \stackrel{\text{def}}{=} \exp \circ c_t$. The notation reflects that d_t refers to a *density ratio* and c_t a *binary classifier*. With the exponential function and the GAN divergence (Table 5.2) we obtain the usual logistic sigmoid in (5.3), that is

$$\underset{c_t \in \mathscr{L}_0(\Omega, \mathbb{R})}{\text{minimise}} \quad \mathrm{E}_\mu \log(1 + \exp(-c_t)) + \mathrm{E}_\nu \log(1 + \exp(c_t)).$$

The analysis in Section 5.2 proceeds using density ratios, whereas Section 5.3 makes use of binary classifiers.

5.2 Boosted density estimation

We will study a sequence $(\mu_t) \subseteq \mathfrak{P}(\Omega, \mu_0)$, defined for a sequence of functions $(d_t) \subseteq \mathscr{L}_1(\Omega, \mathbb{R}_{>0})$, and a sequence of real numbers $(\alpha_t) \subseteq [0, 1]$ that satisfies

$$\begin{aligned} \mu_t &= \frac{1}{z_t}\tilde{\mu}_t, \quad \text{where} \quad z_t \stackrel{\text{def}}{=} \int \mathrm{d}\tilde{\mu}_t, \\ \text{and} \quad \tilde{\mu}_t(\mathrm{d}\omega) &= d_t^{\alpha_t}(\omega)\cdot\mu_{t-1}(\mathrm{d}\omega). \end{aligned} \tag{5.4}$$

For each $t \in \mathbb{N}$ the *error term* is the function $\epsilon_t \in \mathscr{L}_0(\Omega, \mathbb{R}_{>0})$ satisfying

$$\forall_{\omega\in\Omega}: \; d_t(\omega) = \epsilon_t(\omega)\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{t-1}}(\omega).$$

It measures the optimality of the update d_t in the sense that if ϵ_t is a constant function, choosing $\alpha_t = 1$ means that $\mu_t = \mu_\star$ (via (5.4)). The goal of the analysis will be to develop conditions on the sequences (d_t) and (α_t) to ensure $\mathrm{KL}(\mu_\star, \mu_t)$ converges at 0 with vigour.

Proposition 5.2. *The normalisation factors can be written recursively with* $z_t = z_{t-1} \cdot \int d_t^{\alpha_t}\,\mathrm{d}\mu_{t-1}$.

Proof. We just need to write

$$\begin{aligned} \frac{z_t}{z_{t-1}} &= \frac{1}{z_{t-1}}\int \mathrm{d}\tilde{\mu}_t \\ &= \frac{1}{z_{t-1}}\int d_t^{\alpha_t}\,\mathrm{d}\tilde{\mu}_{t-1} \\ &= \int d_t^{\alpha_t}\,\mathrm{d}\mu_{t-1} \\ &= \int d_t^{\alpha_t}\,\mathrm{d}\mu_{t-1}, \end{aligned} \tag{5.5}$$

thus $z_t = z_{t-1} \cdot \int d_t^{\alpha_t}\,\mathrm{d}\mu_{t-1}$. ∎

Proposition 5.3 (Cranko and Nock [31])**.** *The distribution μ_t is an exponential family distribution with natural parameter $(\alpha_1, \ldots, \alpha_t)$ and sufficient statistic $(c_1(x), \ldots, c_t(x))$.*

The connection between the sufficient statistics of an exponential family and deep learning (that is, when (c_i) is a sequence of neural network classifiers) has also been made elsewhere [viz. 94].

Lemma 5.4. *For any $\alpha_t \in [0,1]$ and $\epsilon_t \in \mathscr{L}_0(\Omega, \mathbb{R}_{\geq 0})$ we have:*

$$\exp\Big(\mathrm{E}_{\mu_{t-1}}\log\epsilon_t - \mathrm{rKL}(\mu_\star, \mu_{t-1})\Big)^{\alpha_t} \leq \frac{z_t}{z_{t-1}} \leq (\mathrm{E}_{\mu_\star}\epsilon_t)^{\alpha_t}.$$

Theorem 5.5. *For $\alpha_t \in [0,1]$, $d_t \in \mathscr{L}_0(\Omega, \mathbb{R}_{>0})$, there is*

$$\mathrm{KL}(\mu_\star, \mu_t) \leq (1-\alpha_t)\,\mathrm{KL}(\mu_\star, \mu_{t-1}) + \alpha_t(\log \mathrm{E}_{\mu_\star} \epsilon_t - \mathrm{E}_{\mu_\star} \log \epsilon_t), \quad (5.6)$$

where $\epsilon_t \stackrel{\text{def}}{=} (\mathrm{d}\mu_\star / \mathrm{d}\mu_{t-1})^{-1} d_t$.

Remark 5.6. Grover and Ermon [58, Thm. 2] assume a uniform error term, $\epsilon_t \equiv c$ for some $c > 0$. In this case Theorem 5.5 yields geometric convergence

$$\forall_{\alpha_t \in [0,1]} : \mathrm{KL}(\mu_\star, \mu_t) \leq (1-\alpha_t)\,\mathrm{KL}(\mu_\star, \mu_{t-1}).$$

This result is significantly stronger than Grover and Ermon [58, Thm. 2], who just show the non-increase of the KL divergence. If, in addition to achieving uniform error, we let $\alpha_t = 1$, then (5.6) guarantees $\mu_t = \mu_\star$.

Proof of Lemma 5.4. Since $\alpha_t \in [0,1]$, by Jensen's inequality it follows that

$$\mathrm{E}_{\mu_{t-1}} d_t^{\alpha_t} \leq (\mathrm{E}_{\mu_{t-1}} d_t)^{\alpha_t} = \left(\int \frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{t-1}} \cdot \epsilon_t \,\mathrm{d}\mu_{t-1}\right)^{\alpha_t} = (\mathrm{E}_{\mu_\star} \epsilon_t)^{\alpha_t}. \quad (5.7)$$

The upper bound on z_t/z_{t-1} follows:

$$\frac{z_t}{z_{t-1}} \stackrel{(5.5)}{=} \mathrm{E}_{\mu_{t-1}} d_t^{\alpha_t} \stackrel{(5.7)}{\leq} (\mathrm{E}_{\mu_\star} \epsilon_t)^{\alpha_t}.$$

For the lower bound on z_t/z_{t-1}, note that

$$\begin{aligned} \log\left(\frac{z_t}{z_{t-1}}\right) &\stackrel{(5.5)}{=} \log \mathrm{E}_{\mu_{t-1}} d_t^{\alpha} \\ &\geq \alpha_t \,\mathrm{E}_{\mu_{t-1}} \log d_t \\ &= \alpha_t \,\mathrm{E}_{\mu_{t-1}} \left[\log \epsilon_t + \log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{t-1}}\right)\right], \end{aligned}$$

which implies the lemma. (Lem. 5.4) ■

Proof of Theorem 5.5. First note that

$$\mathrm{d}\mu_t = \frac{1}{z_t}\,\mathrm{d}\tilde{\mu}_t = \frac{1}{z_t} d_t^{\alpha_t}\,\mathrm{d}\tilde{\mu}_{t-1} = \frac{z_{t-1}}{z_t} d_t^{\alpha_t}\,\mathrm{d}\mu_{t-1}. \quad (5.8)$$

Now consider the following two identities:

$$-\alpha_t \log \mathrm{E}_{\mu_\star} \epsilon_t \leq \log\left(\frac{z_{t-1}}{z_t}\right), \tag{5.9}$$

which follows from Lem. 5.4, and

$$\begin{aligned}
&\int \left(\log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{t-1}}\right) - \alpha_t \log d_t\right) \mathrm{d}\mu_\star \qquad (5.10)\\
&\qquad = \int \left(\log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{t-1}}\right) - \alpha_t \log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{t-1}}\right) - \alpha_t \log \epsilon_t\right) \mathrm{d}\mu_\star\\
&\qquad = (1-\alpha_t) \int \log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{t-1}}\right) \mathrm{d}\mu_\star - \alpha_t \int \log \epsilon_t \, \mathrm{d}\mu_\star\\
&\qquad = (1-\alpha_t)\, \mathrm{KL}(\mu_\star, \mu_{t-1}) - \alpha_t \, \mathrm{E}_{\mu_\star} \log \epsilon_t.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{KL}(\mu_\star, \mu_t) &= \int \log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_t}\right) \mathrm{d}\mu_\star\\
&\overset{(5.8)}{=} \int \left(\log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{t-1}}\right) - \log\left(\frac{z_{t-1}}{z_t} d_t^{\alpha_t}\right)\right) \mathrm{d}\mu_\star\\
&= \underbrace{\int \left(\log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_{t-1}}\right) - \alpha_t \log d_t\right) \mathrm{d}\mu_\star}_{(5.10)} - \underbrace{\log\left(\frac{z_{t-1}}{z_t}\right)}_{(5.9)}\\
&\leq (1-\alpha_t)\, \mathrm{KL}(\mu_\star, \mu_{t-1}) + \alpha_t (\log \mathrm{E}_{\mu_\star} \epsilon_t - \mathrm{E}_{\mu_\star} \log \epsilon_t),
\end{aligned}$$

as claimed. (Thm. 5.5) ■

We can express the update (5.6) in a way that more closely resembles Frank–Wolfe update (5.1). Since ϵ_t takes on positive values, we can identify it with a density ratio involving a nonnegative measure as follows

$$\tilde{\rho}_t(\mathrm{d}x) \overset{\text{def}}{=} \epsilon_t(x) \cdot \mu_\star(\mathrm{d}x) \quad \text{and} \quad \rho_t \overset{\text{def}}{=} \frac{1}{\int \mathrm{d}\tilde{\rho}_t} \cdot \tilde{\rho}_t.$$

Introducing $\tilde{\rho}_t$ allows us to lend some interpretation to Theorem 5.5 in terms of the probability measure ρ_t. Letting $m_t \overset{\text{def}}{=} \mathrm{d}\mu_t / \mathrm{d}\mu_0$, $r_t \overset{\text{def}}{=} \mathrm{d}\rho_t / \mathrm{d}\mu_0$, then

$$m_t \propto d_t^{\alpha_t} m_{t-1} = \left(\frac{p}{m_{t-1}} \epsilon_t\right)^{\alpha_t} m_{t-1} = \tilde{r}_t^{\alpha_t} m_{t-1}^{1-\alpha_t}.$$

Or equivalently,

$$m_t = \psi(\alpha_t \log r_t + (1 - \alpha_t) m_{t-1}),$$

where $\psi(f) \stackrel{\text{def}}{=} \exp(f(\,\cdot\,) - \int \log(f) d\mu_0)$ for $f \in \mathscr{L}_0(\Omega)$. This shows the manner in which (5.4) is a special case of the general Frank–Wolfe form (5.1), with updates $x_t \stackrel{\text{def}}{=} \log r_t$, and initial point $y_0 \stackrel{\text{def}}{=} \mu_0$.

Corollary 5.7. *If ρ_t satisfies*

$$\mathrm{KL}(\mu_\star, \rho_t) \leq \gamma\, \mathrm{KL}(\mu_\star, \mu_{t-1}), \tag{5.11}$$

for some $\gamma \in [0, 1]$, then for any $\alpha_t \in [0, 1]$

$$\mathrm{KL}(\mu_\star, \mu_t) \leq (1 - \alpha_t(1 - \gamma))\, \mathrm{KL}(\mu_\star, \mu_{t-1}). \tag{5.12}$$

Proof. We first show

$$\mathrm{KL}(\mu_\star, \mu_t) \leq (1 - \alpha_t)\, \mathrm{KL}(\mu_\star, \mu_{t-1}) + \alpha_t\, \mathrm{KL}(\mu_\star, \rho_t). \tag{5.13}$$

By definition $\epsilon_t = \mathrm{d}\rho_t / \mathrm{d}\mu_\star$. From Thm. 5.5, the rightmost term in (5.6) reduces as follows

$$\begin{aligned}
\log \mathrm{E}_{\mu_\star} \epsilon_t - \mathrm{E}_{\mu_\star} \log \epsilon_t &= \log \int \frac{\mathrm{d}\tilde{\rho}_t}{\mathrm{d}\mu_\star}\, \mathrm{d}\mu_\star - \int \log\left(\frac{\mathrm{d}\tilde{\rho}_t}{\mathrm{d}\mu_\star}\right) \mathrm{d}\mu_\star \\
&= \log \int \mathrm{d}\tilde{\rho}_t + \int \log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\tilde{\rho}_t}\right) \mathrm{d}\mu_\star \\
&= \int \left(\log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\tilde{\rho}_t}\right) + \log \int \mathrm{d}\tilde{\rho}_t\right) \mathrm{d}\mu_\star \\
&= \int \log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\tilde{\rho}_t} \cdot \int \mathrm{d}\tilde{\rho}_t\right) \mathrm{d}\mu_\star \\
&= \int \log\left(\frac{\mathrm{d}\mu_\star}{\frac{1}{\int \mathrm{d}\tilde{\rho}_t}\, \mathrm{d}\tilde{\rho}_t}\right) \mathrm{d}\mu_\star \\
&= \mathrm{KL}(\mu_\star, \rho_t),
\end{aligned}$$

which shows (5.13). The proof of (5.12) is then immediate. ■

We obtain the same convergence rate as Tolstikhin *et al.* [129, Cor. 2] (geometric) for a boosted distribution μ_t which is not a convex mixture, which, to our knowledge, is a new result. Corollary 5.7 is restricted to the KL divergence, however, we do not need the technical domination assumption of Tolstikhin *et al.* [129, Cor. 2]. From the standpoint of weak versus strong learning, Tolstikhin *et al.* [129, Cor. 2] require a condition similar to (5.11), that is, the iterate ρ_t has to be close enough to $\mu_\star$. It is the objective of the following sections to relax this requirement to something akin to the weak updates common in a boosting scheme.

5.3 Convergence under weak assumptions

In the previous section we have established two preliminary convergence results (Remark 5.6, Corollary 5.7) that equal the state of the art and/or rely on similarly strong assumptions. We now show how to relax these in favour of placing some weak conditions on the binary classifiers learnt in (5.2).

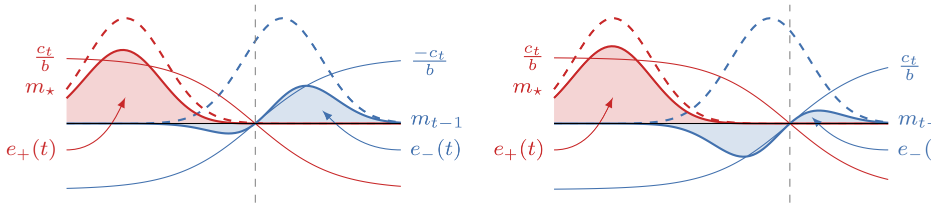
Define the two *expected edges of* c_t [cf. 95]:

$$e_-(t) \stackrel{\text{def}}{=} \frac{1}{b}\,\mathrm{E}_{\mu_{t-1}}[-c_t] \quad \text{and} \quad e_+(t) \stackrel{\text{def}}{=} \frac{1}{b}\,\mathrm{E}_{\mu_\star}[c_t],$$

where $b \geq \operatorname{ess\,sup}|c_t|$ for all $t \in \mathbb{N}$, and the essential supremum is with respect to μ_0. Classical boosting results rely on assumption on such edges for different kinds of c_t [47, 115, 116]. We also assume $b < \infty$ and $|c_t| > 0$ for all $t \in \mathbb{N}$. That is, the classifiers have bounded and nonzero confidence. By construction $e_-(t), e_+(t) \in [-1, 1]$ for every $t \in \mathbb{N}$. The difference of sign of c_t is due to the decision rule for a binary classifier, whereby $c_t(\omega) \geq 0$ reflects that c_t classifies $\omega \in \Omega$ as originating from $\mu_\star$ rather than μ_{t-1}, and vice versa for $-c_t(\omega)$.

Assumption WL$_T$ (Weak learning)**.** *For $T \in \mathbb{N}$ there exist $\gamma_+, \gamma_- > 0$ so that $e_+(t) \geq \gamma_+$ and $e_-(t) \geq \gamma_-$ for all $t \leq T$.*

The weak learning assumption is in effect a separation condition of $\mu_\star$ and μ_{t-1}. That is, the decision boundary associated with c_t correctly divides most of the mass of $\mu_\star$ and most of the mass of μ_{t-1}. This is illustrated in Figure 5.3. Note that if μ_{t-1} has converged to $\mu_\star$, the weak learning assumption cannot hold. This is reasonable since as $\mu_{t-1} \to \mu_\star$ it becomes harder to build a classifier to tell them apart. We note that classical boosting



(a) WL_T is not violated, $e_+(t), e_-(t) > 0$ (b) WL_T is violated, $e_-(t) < 0$

Figure 5.3: Illustration of WL_T in one dimension with a classifier c_t and its decision rule (indicated by the dashed grey line). The red (resp. blue) area is the area under the $c_t/b \star p$ (resp. $-c_t/b \star m_{t-1}$) line (where $m_\star, m_{t-1}$ are corresponding density functions of $\mu_\star$ and μ_{t-1}), that is, $e_+(t)$ (resp. $e_-(t)$).

would rely on a single inequality for the weak learning assumption (involving the two edges) [116] instead of two as in WL_T. The difference is, however, superficial as we can show that both assumptions are equivalent. A boosting algorithm would ensure, for any given error $\varrho > 0$, that there exists a number of iterations T for which we do have $\mathrm{KL}(\mu_\star, \mu_T) \leq \varrho$, where T is required to be polynomial in all relevant parameters, in particular $1/\gamma_+, 1/\gamma_-, b, \mathrm{KL}(\mu_\star, \mu_0)$. Notice that we have to put $\mathrm{KL}(\mu_\star, \mu_0)$ in the complexity requirement since it can be arbitrarily large compared to the other parameters.

Theorem 5.8 (Cranko and Nock [31])**.** *Assume there is WL_T, where the sequence $(\alpha_t)_{t\leq T}$ satisfies*

$$\alpha_t = \min\left\{1, \frac{1}{2b}\log\left(\frac{1+e_-(t)}{1-e_-(t)}\right)\right\}.$$

Then $\mathrm{KL}(\mu_\star, \mu_T) \leq \varrho$ *when*

$$T \geq 2 \cdot \frac{\mathrm{KL}(\mu_\star, \mu_0) - \varrho}{\gamma_+\gamma_-}.$$

The question naturally arises as to whether faster convergence is possible. Define

$$e(t) \stackrel{\text{def}}{=} \frac{1}{b} \cdot \mathrm{E}_{\mu_\star} \log \epsilon_t,$$

the normalised expected log-density estimation error. Then we have $e_+(t) = \frac{1}{b} \cdot \mathrm{KL}(\mu_\star, \mu_{t-1}) + e(t)$, so controlling $e_+(t)$ does not give substantial leverage on $\mathrm{KL}(\mu_\star, \mu_t)$ because of the unknown $e(t)$. Therefore we can show that that an additional weak assumption on $e(t)$ (not unlike boundedness condition on the log-density ratio of Li and Barron [77, Thm. 1]) is all that is needed with WL_T, to obtain convergence rates that compete with Tolstikhin *et al.* [129, Lem. 2] but using much weaker assumptions.

Assumption $\mathbf{WD}_T$ (Weak dominance)**.** *For $T \in \mathbb{N}$ there exists $\Gamma_\epsilon > 0$ so that $e(t) \geq -\Gamma_\epsilon$ for all $t \leq T$.*

Under WL_T and WD_T we are able to obtain a geometric convergence rate.

Theorem 5.9 (Cranko and Nock [31])**.** *If WL_T and WD_T hold, then*

$$\mathrm{KL}(\mu_\star, \mu_T) \leq \left(1 - \frac{\gamma_+}{2(1+\Gamma_\epsilon)} \min\left\{2, \frac{\gamma_-}{b}\right\}\right)^T \cdot \mathrm{KL}(\mu_\star, \mu_0).$$

Note that the bound obtained in Theorem 5.9 is, in fact, logarithmic in $\mathrm{KL}(\mu_\star, \mu_0)$, that is, we have $\mathrm{KL}(\mu_\star, \mu_T) \leq \varrho$ when

$$T \geq \frac{2(1+\Gamma_\epsilon)}{\gamma_+ \min\{2, \gamma_-/b\}} \log\left(\frac{\mathrm{KL}(\mu_\star, \mu_0)}{\varrho}\right).$$

The proofs of Theorems 5.8 and 5.9 are due to Prof. Richard Nock and are quite lengthy. These can be found in full in the original work this chapter was based upon [31].

5.4 Conclusion

The prospect of learning a density iteratively with a boosting-like procedure has recently been met with significant attention. However, the success of these approaches hinge on the existence of oracles satisfying very strong assumptions. By contrast, the task of learning a binary classifier iteratively is well understood and backed by a large amount of research. By leveraging this understanding for the seemingly disparate application of density estimation, we are able to improve upon other state-of-the-art guarantees. Finally, since the work on which this chapter was published [31], in a follow-up, Husein *et al.* [67] have shown how density estimation of the form we analyse here can be adapted to yield strong differential privacy properties.

Chapter 6
Robust Bayes, Regularisation, and Adversarial Learning

When learning a statistical model, it is rare that one has complete access to the distribution. More often it is the case that one approximates the risk minimisation by an empirical risk, using sequence of samples from the distribution. In practice this can be problematic, particularly when the curse of dimensionality is in full force, to: *1.*) know with certainty that one has enough samples, and *2.*) guarantee good performance away from the data. Both of these two problems can, in effect, be cast as problems of ensuring generalisation. A remedy for both of these problems has been proposed in the form of a modification to the risk minimisation framework, wherein we integrate a certain amount of distrust of the distribution. This distrust results in a certification of worst case performance if it turns out later that the distribution was specified imprecisely, improving generalisation.

To make this concept of distrust concrete, in the notation of Chapter 4, for a loss function $\ell : V \to \mathscr{L}_0(\Omega, \bar{\mathbb{R}})$ we replace the classical risk minimisation (B) [on p. 63] with

$$\underset{v \in V}{\text{minimise}} \quad \sup_{\nu \in B} \operatorname{risk}_\ell(v, \nu), \tag{rB}$$

where $B \subseteq \mathfrak{P}(\Omega)$ is called the *uncertainty set* and (rB) is called the *B-robust Bayes risk* [59, §4, 18, 134]. The problem (rB) is an example of a machine learning problem that is incompatible with the risk minimisation, and therefore the probability elicitation framework in general. However, we shall see that for a class of loss functions $\ell : V \to L$, and a particular uncertainty set, $\mathrm{B}_c(\mu, r)$ (containing $\mu \in \mathfrak{P}(\Omega)$ and depending on $r \geq 0$ and $c \in \mathscr{L}_0(\Omega, \bar{\mathbb{R}}_{\geq 0})$), there is a function $\mathrm{lip}_c : L \to \bar{\mathbb{R}}_{\geq 0}$ so that the regularised objective

$$\underset{v \in V}{\text{minimise}} \quad \mathrm{risk}_\ell(v, \mu) + r\,\mathrm{lip}_c(\ell(v)), \tag{Reg}$$

has the same minimisers as the $\mathrm{B}_c(\mu, r)$-robust Bayes risk.

There are two reasons we are interested in finding a relationship between (rB) and (Reg). There is independent interest in the objective function in (Reg), particularly when C corresponds to the least Lipschitz constant of $\ell(v)$ measured with respect to some metric on L. The applications for Lipschitz regularisation are as disparate as generative adversarial networks [5, 87], generalisation [42, 55, 142], and adversarial learning [4, 29, 30, 131] among others [56, 114]. Building a model that is robust to a particular uncertainty set is very intuitive and tractable. However, the left hand side of (Reg) involves an optimisation over a subset of an infinite dimensional space,[1] by comparison, (Reg) is often much easier to work with in practice. For these reasons then it is always interesting to note when a robust Bayes problem admits an equivalent formulation of (rB) in the form of (Reg), or vice versa.

It happens that for the applications mentioned above, the relevant uncertainty set is parameterised by the *transportation cost*. In Section 6.1 we state the major definitions to define the transportation cost and its associated uncertainty set, the *transportation cost ball*. In Section 6.2 we begin with a series of technical lemmas before proving we are able to prove our major result, Theorem 6.5. This result connects (rB) and (Reg) with new generality and tightness guarantees, applying to a class of models broad enough to include nonconvex models, such as deep neural networks. In Section 6.3, we introduce the previously mentioned problem of adversarial learning, and give a new generalised result showing equality with the transportation-cost-parameterised uncertainty set from Sections 6.1 and 6.2. This completes the loop for the problem of adversarial learning and suggests new ways in which

[1] Except for when $\mathrm{B}_c(\mu, r)$ is chosen in a particularly trivial way.

robustness can be learnt for a broad class of models, discussion of which is postponed to the conclusion, Section 6.4.

6.1 Preliminaries

For the remainder of this chapter we let $\bar{\mathbb{R}} \overset{\text{def}}{=} (-\infty, \infty]$. Unless otherwise specified, X, Y, Ω are topological outcome spaces. Often X will be used when there is some linear structure, compatible with the topology, so that $X \times Y$ may be interpreted as the classical outcome space for classification problems [cf. 133]. For a measure $\mu \in \mathfrak{P}(X)$ its push-forward by $f \in \mathscr{L}_0(X, Y)$ is $f_\#\mu \in \mathfrak{P}(Y)$, where $f_\#\mu A \overset{\text{def}}{=} \mu(f^{-1}(A))$ for all Borel $A \subseteq Y$. When (Ω, d) is a metric space, the closed ball of radius $r \geq 0$, centred at $x \in X$ is denoted $\mathrm{B}_d(x, r) \overset{\text{def}}{=} \{y \in X \mid d(x, y) \leq r\}$.

For two measures $\mu, \nu \in \mathfrak{P}(\Omega)$ the set of (μ, ν)-*couplings* is

$$\Pi(\mu, \nu) \overset{\text{def}}{=} \left\{\pi \in \mathfrak{P}(\Omega \times \Omega) \;\middle|\; \mu = \int \pi(\,\cdot\,, \mathrm{d}\omega),\ \nu = \int \pi(\mathrm{d}\omega,\,\cdot\,)\right\}.$$

For a Borel *coupling function* $c : \Omega \times \Omega \to \bar{\mathbb{R}}$ the c-*transportation cost* of $\mu, \nu \in \mathfrak{P}(\Omega)$ is

$$\mathrm{cost}_c(\mu, \nu) \overset{\text{def}}{=} \inf_{\pi \in \Pi(\mu,\nu)} \int c \,\mathrm{d}\pi, \tag{6.1}$$

and the c-transportation cost ball of radius $r \geq 0$ centred at $\mu \in \mathfrak{P}(\Omega)$ is

$$\mathrm{B}_c(\mu, r) \overset{\text{def}}{=} \{\nu \in \mathfrak{P}(\Omega) \mid \mathrm{cost}_c(\mu, \nu) \leq r\}, \tag{6.2}$$

and serves as our *uncertainty set*. When (Ω, d) is a Polish space, the d-transportation cost is called the *Wasserstein distance*. When d is bounded, cost_d completely metrises the $\sigma(\mathfrak{P}(\Omega), \mathrm{C_b}(\Omega))$-topology on $\mathfrak{P}(\Omega)$ [see 135, Cor. 6.13].

A coupling function $c : X \times X \to \bar{\mathbb{R}}$ has an associated conjugacy operation with

$$f^c(x) \overset{\text{def}}{=} \sup_{y \in X}\Big(f(y) - c(x, y)\Big), \tag{6.3}$$

for any function $f : X \to \bar{\mathbb{R}}$. Coupling functions and their conjugates have many applications in the theory of generalised convexity and polarities, including those we have already encountered in Chapters 2 and 3 [cf. 39, 90, 97, 98, 135]. We define the *least c-Lipschitz constant* [cf. 30] of a function $f : X \to \bar{\mathbb{R}}$:

$$\operatorname{lip}_c(f) \stackrel{\text{def}}{=} \inf\{\lambda \geq 0 \mid \forall_{x,y \in X} : f(x) - f(y) \leq \lambda c(x, y)\}, \tag{6.4}$$

so that when (X, d) is a metric space $\operatorname{lip}_d(f)$ agrees with the usual Lipschitz notion. When $c : X \to \bar{\mathbb{R}}$, for example when c is a norm, we take $c(x, y) \stackrel{\text{def}}{=} c(x - y)$ for all $x, y \in X$ in (6.1), (6.2), (6.3), and (6.4).

For a function $f : X \to \bar{\mathbb{R}}$ there is another function $\overline{\operatorname{co}} f : X \to \bar{\mathbb{R}}$, called the *convex envelope* of f, satisfying $\operatorname{epi}(\overline{\operatorname{co}} f) = \overline{\operatorname{co}}(\operatorname{epi} f)$. It is the greatest closed convex function that minorises f. The quantity $\rho(f) \stackrel{\text{def}}{=} \sup_{x \in X}(f(x) - \overline{\operatorname{co}} f(x))$ was first suggested by Aubin and Ekeland [11] to quantify the lack of convexity of a function f, and has since shown to be of considerable interest for, among other things, bounding the duality gap in nonconvex optimisation [cf. 6, 70, 76, 132].

Let $\Delta^n(x) \stackrel{\text{def}}{=} \{\alpha \in \mathbb{R}^n_{\geq 0} \mid \sum_{i \in [n]} \alpha_i = 1\}$. When $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is minorised by an affine function, $\operatorname{epi}(\overline{\operatorname{co}} f) = \overline{\operatorname{co}}(\operatorname{epi} f)$ means that [cf. 66, Prop. 1.5.4] for all $x \in \mathbb{R}^n$

$$\overline{\operatorname{co}} f(x) = \inf\left\{ \sum_{i \in [n+1]} \alpha_i f(x_i) \;\middle|\; \alpha \in \Delta^{n+1},\ (x_i)_{i \in [n+1]} \subseteq \mathbb{R}^n,\ x = \sum_{i \in [n+1]} \alpha_i x_i \right\}.$$

Consequentially there is the common expression

$$\rho(f) = \sup\left\{ f\left(\sum_{i \in [n+1]} \alpha_i x_i \right) - \sum_{i \in [n+1]} \alpha_i f(x_i) \;\middle|\; \alpha \in \Delta^{n+1},\ (x_i)_{i \in [n+1]} \subseteq \mathbb{R}^n \right\}.$$

For simplicity of notation in the subsequent sections, for a loss function $\ell : V \to L$, we identify ℓ at a particular model $v \in V$, with the function $f : \Omega \to \bar{\mathbb{R}}$, so that $\ell(v) = f$.

6.2 Robust learning

Duality results like Lemma 6.1 have been the basis of a number of recent theoretical efforts in the theory of adversarial learning [20, 48, 120, 123], the results of Blanchet and Murthy [21] being the most general to date.

Lemma 6.1 (Blanchet and Murthy [21, Thm. 1])**.** *Assume Ω is a Polish space and fix $\mu \in \mathfrak{P}(\Omega)$. Let $c : \Omega \times \Omega \to \bar{\mathbb{R}}_{\geq 0}$ be lower semicontinuous with $c(\omega, \omega) = 0$ for all $\omega \in \Omega$, and $f \in \mathscr{L}_1(\Omega, \mu)$ is upper semicontinuous. Then for all $r \geq 0$ there is*

$$\sup_{\nu \in \mathrm{B}_c(\mu, r)} \int f \, \mathrm{d}\nu = \inf_{\lambda \geq 0} \Big(\lambda r + \int f^{\lambda c} \, \mathrm{d}\mu \Big). \tag{6.5}$$

The necessity for such duality results like Lemma 6.1 is because while the supremum on the left hand side of (6.5) is over a (usually) infinite dimensional space, the right hand side only involves only a finite dimensional optimisation. The generalised conjugate in (6.5) also hides an optimisation, but when the outcome space Ω is finite dimensional, this too is a finite dimensional problem.

The following lemma is sometimes stated a consequence of, or in the proof of, the McShane–Whitney extension theorem [83, 139], but it is immediate to observe.

Lemma 6.2. *Let X be a set. Assume $c : X \times X \to \bar{\mathbb{R}}_{\geq 0}$ satisfies $c(x, x) = 0$ for all $x \in X$, $f : X \to \mathbb{R}$. Then*

$$1 \geq \operatorname{lip}_c(f) \iff \forall_{y \in X} : \; f(y) = \sup_{x \in X} \Big(f(x) - c(x, y) \Big).$$

Proof. Suppose $1 \geq \operatorname{lip}_c(f)$. Fix $y_0 \in X$. Then

$$\forall_{x \in X} : \; f(x) - c(x, y_0) \leq f(y_0),$$

with equality when $x = y_0$. Next suppose

$$\forall_{y \in X} : \; f(y) = \sup_{x \in X} \Big(f(x) - c(x, y) \Big),$$

then

$$\begin{aligned} \forall_{x,y \in X} : \; f(y) \geq f(x) - c(x, y) &\iff \forall_{x,y \in X} : \; f(x) - f(y) \leq c(x, y) \\ &\iff 1 \geq \operatorname{lip}_c(f), \end{aligned}$$

as claimed. ∎

Lemma 6.3. *Assume X is a vector space. Suppose $c : X \to \bar{\mathbb{R}}_{\geq 0}$ satisfies $c(0) = 0$, and $f : X \to \mathbb{R}$ is convex. Then*

$$1 \geq \operatorname{lip}_c(f) \iff \forall_{\epsilon \geq 0} : \ \partial_\epsilon f(X) \subseteq \partial_\epsilon c(0).$$

Proof. Suppose $1 \geq \operatorname{lip}_c(f)$. Then $f(x) - f(y) \leq c(x - y)$ for all $x, y \in X$. Fix $\epsilon \geq 0$, $x \in X$ and suppose $x^* \in \partial_\epsilon f(x)$. Then

$$\begin{aligned} &\forall_{y \in X} : \ \langle y - x, x^* \rangle - \epsilon \leq f(y) - f(x) \leq c(y - x) \\ &\qquad \iff \forall_{y \in X} : \ \langle y, x^* \rangle - \epsilon \leq f(y + x) - f(x) \leq c(y) - c(0), \end{aligned}$$

because $c(0) = 0$. This shows $x^* \in \partial_\epsilon c(0)$.

Next assume $\partial_\epsilon f(x) \subseteq \partial_\epsilon c(0)$ for all $\epsilon \geq 0$ and $x \in X$. Because f is not extended-real valued, it is continuous on all of X [via 149, Cor. 2.2.10], $\partial f(x)$ is nonempty for all $x \in X$ [via 149, Thm. 2.4.9]. Fix an arbitrary $x \in X$. Then $\emptyset \neq \partial f(x) \subseteq \partial c(0)$, and

$$\begin{aligned} &\exists_{x^* \in \partial f(x)} \forall_{y \in X} : \ f(x) - f(y) \leq \langle x - y, x^* \rangle \\ &\qquad \implies \forall_{y \in X} : \ f(x) - f(y) \leq \langle x - y, x^* \rangle \leq c(x - y), \end{aligned} \tag{6.6}$$

where the implication is because $x^* \in \partial c(0)$ and $c(0) = 0$. Since the choice of x in (6.6) was arbitrary, the proof is complete. ∎

Lemma 6.4. *Assume X is a locally convex Hausdorff topological vector space. Suppose $c : X \to \bar{\mathbb{R}}$ is closed sublinear, and $f : X \to \mathbb{R}$ is closed convex. Then there is*

$$\forall_{y \in X} : \ \sup_{x \in X} \Big(f(x) - c(x - y) \Big) = \begin{cases} f(y) & 1 \geq \operatorname{lip}_c(f) \\ \infty & \textit{otherwise.} \end{cases}$$

Proof. Fix an arbitrary $y_0 \in X$. From Lem. 6.3 we know

$$1 \geq \operatorname{lip}_c(f) \iff \forall_{\epsilon \geq 0} : \ \partial_\epsilon f(X) \subseteq \partial_\epsilon c(0).$$

Assume $\partial_\epsilon f(X) \subseteq \partial_\epsilon c(0)$ for all $\epsilon \geq 0$: Consequentially $\partial_\epsilon f(y_0) \subseteq \partial_\epsilon c(0) = \partial_\epsilon c(\,\cdot\, - y_0)(y_0)$ for every $\epsilon \geq 0$. From the usual difference-convex global

ϵ-subdifferential condition [64, Thm. 4.4] it follows that

$$\inf_{x \in X} \Big(c(x - y_0) - f(x) \Big) = \underbrace{c(y_0 - y_0)}_{0} - f(y_0) = -f(y_0),$$

where we note that $c(y_0 - y_0) = c(0) = 0$ because c is sublinear.

Assume $\partial_\epsilon f(X) \not\subseteq \partial_\epsilon c(0)$ *for some* $\epsilon \geq 0$: By hypothesis there exists $\epsilon_0 \geq 0$, $x_0 \in X$, and $x_0^* \in X^*$ with

$$x_0^* \in \partial_{\epsilon_0} f(x_0) \quad \text{and} \quad x_0^* \notin \partial_{\epsilon_0} c(0).$$

Using the Toland [128] duality formula [viz. 63, Cor. 2.3] and the usual calculus rules for the Fenchel conjugate [e.g. 149, Thm. 2.3.1] we have

$$\begin{aligned} \inf_{x \in X} \Big(c(x - y_0) - f(x) \Big) &= \inf_{x^* \in X^*} \Big(f^*(x^*) - (c(\,\cdot\, - y_0))^*(x^*) \Big) \\ &= \inf_{x^* \in X^*} \Big(f^*(x^*) - c^*(x^*) + \langle y_0, x^* \rangle \Big) \\ &\leq f^*(x_0^*) - c^*(x_0^*) + \langle y_0, x_0^* \rangle \\ &\leq \epsilon_0 + \langle x_0, x_0^* \rangle - f(x_0) - c^*(x_0^*) + \langle y_0, x_0^* \rangle \\ &= \underbrace{\epsilon_0 + \langle x_0 + y_0, x_0^* \rangle - f(x_0)}_{<\infty} - c^*(x_0^*), \end{aligned} \tag{6.7}$$

where the second inequality is because $x_0^* \in \partial_{\epsilon_0} f(x_0)$.

We have assumed $x_0^* \notin \partial_\epsilon c(0) \supseteq \partial c(0)$. Because c is sublinear, $c^* = \iota_{\partial c(0)}$ [149, Thm. 2.4.14 (i)], and therefore $c^*(x_0^*) = \infty$. Then (6.7) yields

$$\inf_{x \in X} \Big(c(x - y_0) - f(x) \Big) \leq -\infty,$$

which completes the proof. ■

Theorem 6.5 subsumes many existing results [48, Cor. 2 (iv), 29, §3.2, 123, various, 120, Thm. 14] with a great deal more generality, applying to a very broad family of models, loss functions, and outcome spaces.

Theorem 6.5. *Assume* X *is a separable Fréchet space and fix* $\mu \in \mathfrak{P}(X)$. *Suppose* $c : X \to \bar{\mathbb{R}}_{\geq 0}$ *is closed sublinear, and* $f \in \mathscr{L}_1(X, \mu)$ *is upper semicontinuous with* $\operatorname{lip}_c(f) < \infty$. *Then for all* $r \geq 0$, *there is a number*

$\varDelta(f,\mu,r,c) \geq 0$ *so that*

$$\sup_{\nu \in \mathrm{B}_c(\mu,r)} \int f \,\mathrm{d}\nu + \varDelta(f,\mu,r,c) = \int f \,\mathrm{d}\mu + r \operatorname{lip}_c(f). \tag{6.8}$$

Moreover

$$0 \leq \varDelta(f,\mu,r,c) \leq r \operatorname{lip}_c(f) - \Big[r \operatorname{lip}_c(\overline{\operatorname{co}}\, f) - \int (f - \overline{\operatorname{co}}\, f) \,\mathrm{d}\mu \Big]_+, \tag{6.9}$$

where $[\,\cdot\,]_+ \overset{\text{def}}{=} \max\{\,\cdot\,, 0\}$, *so that when* f *is closed convex* $\varDelta(f,\mu,r,c) = 0$.

Observing that $\varDelta(f,\mu,r,c) \geq 0$, the equality (6.8) yields the upper bound

$$\sup_{\nu \in \mathrm{B}_c(\mu,r)} \int f \,\mathrm{d}\nu \leq \int f \,\mathrm{d}\mu + r \operatorname{lip}_c(f). \tag{6.10}$$

By controlling $\varDelta(f,\mu,r,c)$ we are able to guarantee that the regularised risk in (Reg) is a good surrogate for the robust risk. The number $\varDelta(f,\mu,r,c)$ itself is quite hard to measure (since it would require computing the robust risk directly), which is why we upper bound it in (6.9). Proposition 6.6 shows the slackness bound (6.9) is tight for a large family of distributions after observing

$$\forall_{f \in \mathscr{L}_0(X,\bar{\mathbb{R}})} \forall_{\mu \in \mathfrak{P}(X)} : \int (f - \overline{\operatorname{co}}\, f) \,\mathrm{d}\mu \leq \uprho(f).$$

Which yields

$$\begin{aligned} r \operatorname{lip}_c(f) - \Big[r \operatorname{lip}_c(\overline{\operatorname{co}}\, f) - \int (f - \overline{\operatorname{co}}\, f) \,\mathrm{d}\mu \Big]_+ \\ \leq r \operatorname{lip}_c(f) - \Big[r \operatorname{lip}_c(\overline{\operatorname{co}}\, f) - \uprho(f) \Big]_+, \end{aligned}$$

for all $f \in \mathscr{L}_0(X,\bar{\mathbb{R}})$, $\mu \in \mathfrak{P}(X)$, and $r \geq 0$.

Proposition 6.6. *Let* X *be a separable Fréchet space with* $X_0 \subseteq X$. *Suppose* $c : X \to \bar{\mathbb{R}}_{\geq 0}$ *is closed sublinear, and* $f \in \bigcap_{\mu \in \mathfrak{P}(X_0)} \mathscr{L}_1(X,\mu)$ *is upper semicontinuous, has* $\operatorname{lip}_c(f) < \infty$, *and attains its maximum on* X_0. *Then*

$$\forall_{r \geq 0} : \sup_{\mu \in \mathfrak{P}(X_0)} \varDelta(f,\mu,r,c) = r \operatorname{lip}_c(f) - \Big[r \operatorname{lip}_c(\overline{\operatorname{co}}\, f) - \uprho(f) \Big]_+.$$

Remark 6.7. In particular, for any compact subset of a Fréchet space X_0 (such as the set of n-dimensional images, $X_0 = [0,1]^n \subseteq \mathbb{R}^n$) the bound (6.8) is tight with respect to the set $\mathfrak{P}(X_0)$ for any upper semicontinuous $f \in \bigcap_{\mu\in\mathfrak{P}(X_0)} \mathscr{L}_1(X,\mu)$. Since the behaviour of f away from X_0 is not important, the c-Lipschitz constant in (6.8) need only be computed here. To do so one may replace c with $\tilde{c}$, where $\tilde{c}(x) = c(x)$ for $x \in X_0$ and $\tilde{c}(x) = \infty$ for $x \in X \setminus X_0$, and observe $\operatorname{lip}_{\tilde{c}}(f) \leq \operatorname{lip}_c(f)$, because $\tilde{c} \geq c$.

The extension of Theorem 6.5 for robust classification in the absence of label noise is straight-forward:

Corollary 6.8. *Assume X is a separable Fréchet space and Y is a topological space. Fix $\mu \in \mathfrak{P}(X\times Y)$. Assume $c : (X\times Y)\times(X\times Y) \to \bar{\mathbb{R}}_{\geq 0}$ satisfies*

$$c((x,y),(x',y')) = \begin{cases} c_0(x-x') & y = y' \\ \infty & y \neq y', \end{cases}$$

where $c_0 : X \to \bar{\mathbb{R}}_{\geq 0}$ is closed sublinear, and $f \in \mathscr{L}_1(X\times Y,\mu)$ is upper semicontinuous with $\operatorname{lip}_c(f) < \infty$. Then for all $r \geq 0$ there is (6.8) *and* (6.9)*, where the closed convex hull is interpreted as $\overline{\operatorname{co}}(f)(x,y) \overset{\text{def}}{=} \overline{\operatorname{co}}(f(\,\cdot\,,y))(x)$.*

It is the first time to our knowledge that the slackness in (6.9) has been characterised tightly. Clearly from Theorem 6.5 the upper bound (6.10) is tight for closed convex functions, but Proposition 6.6 shows it is also tight for a large family of nonconvex functions and measures — particularly the upper semi-continuous loss functions on a compact set, with the collection of probability distributions supported on that set.

Proof of Theorem 6.5. (6.8): Since c is assumed sublinear, it is positively homogeneous and there is $c(x,x) = c(x-x) = c(0) = 0$ for all $x \in X$. Therefore we can apply Lem. 6.1 and Lem. 6.2 to obtain

$$\begin{aligned} \sup_{\nu\in\mathrm{B}_c(\mu,r)} \int f \,\mathrm{d}\nu &\overset{\text{L6.1}}{=} \inf_{\lambda\geq 0}\left(r\lambda + \int f^{\lambda c}\,\mathrm{d}\mu\right) \\ &\leq \inf_{\lambda\geq \operatorname{lip}_c(f)}\left(r\lambda + \int f^{\lambda c}\,\mathrm{d}\mu\right) \\ &\overset{\text{L6.2}}{=} r\operatorname{lip}_c(f) + \int f\,\mathrm{d}\mu, \end{aligned} \tag{6.11}$$

and therefore $\Delta(f,\mu,r,c) \geq 0$.

(6.9): Thus observing that $\overline{\mathrm{co}}\, f \leq f$, from Lem. 6.4 we find for all $x \in X$

$$\begin{aligned}
&\sup_{\lambda\in[0,\infty)}\Big(f(x) - f^{\lambda c}(x) - r\lambda\Big)\\
&\quad= \sup_{\lambda\in[0,\infty)}\Big(f(x) - \sup_{y\in X}\Big(f(y) - \lambda c(x-y)\Big) - r\lambda\Big)\\
&\quad= \sup_{\lambda\in[0,\infty)}\inf_{y\in X}\Big(f(x) - f(y) + \lambda c(x-y) - r\lambda\Big)\\
&\quad\leq \sup_{\lambda\in[0,\infty)}\inf_{y\in X}\Big(f(x) - \overline{\mathrm{co}}\, f(y) + \lambda c(x-y) - \lambda r\Big)\\
&\quad\overset{\mathrm{L}6.4}{=} \sup_{\lambda\in[0,\infty)}\begin{cases} f(x) - \overline{\mathrm{co}}\, f(x) - \lambda r & \mathrm{lip}_c(\overline{\mathrm{co}}\, f) \leq \lambda \\ -\infty & \mathrm{lip}_c(\overline{\mathrm{co}}\, f) > \lambda \end{cases}\\
&\quad= f(x) - \overline{\mathrm{co}}\, f(x) - r\,\mathrm{lip}_c(\overline{\mathrm{co}}\, f).
\end{aligned}\tag{6.12}$$

Similarly, for all $x \in X$ there is

$$\begin{aligned}
\sup_{\lambda\in[0,\infty)}\Big(f(x) - f^{\lambda c}(x) - r\lambda\Big) &\leq \sup_{\lambda\in[0,\infty)}\Big(f(x) - f^{\lambda c}(x)\Big) + \sup_{\lambda\in[0,\infty)}\Big(-r\lambda\Big)\\
&= \sup_{\lambda\in[0,\infty)}\Big(f(x) - f^{\lambda c}(x)\Big)\\
&= \sup_{\lambda\in[0,\infty)}\inf_{y\in X}\Big(f(x) - f(y) + \lambda c(x-y)\Big)\\
&\leq \inf_{y\in X}\sup_{\lambda\in[0,\infty)}\Big(f(x) - f(y) + \lambda c(x-y)\Big)\\
&= \inf_{y\in X}\begin{cases}\infty & c(x-y) > 0\\ 0 & c(x-y) = 0\end{cases}\\
&= 0.
\end{aligned}\tag{6.13}$$

Together, (6.12) and (6.13) show

$$\begin{aligned}
&\int \sup_{\lambda\in[0,\infty)}\Big(f - f^{\lambda c} - r\lambda\Big)\,\mathrm{d}\mu\\
&\qquad\qquad \leq \min\Big\{\int (f - \overline{\mathrm{co}}\, f)\,\mathrm{d}\mu - r\,\mathrm{lip}_c(\overline{\mathrm{co}}\, f), 0\Big\}.
\end{aligned}\tag{6.14}$$

Then

$$\begin{aligned}
\mathit{\Delta}(f,\mu,r,c) &= \left(r\,\mathrm{lip}_c(f) + \int f\,\mathrm{d}\mu\right) - \sup_{\nu\in \mathrm{B}_c(\mu,r)} \int f\,\mathrm{d}\nu \\
&\overset{(6.11)}{=} \left(r\,\mathrm{lip}_c(f) + \int f\,\mathrm{d}\mu\right) - \inf_{\lambda\in[0,\infty)} \left(r\lambda - \int f^{\lambda c}\,\mathrm{d}\mu\right) \\
&= r\,\mathrm{lip}_c(f) + \sup_{\lambda\in[0,\infty)} \int \left(f - f^{\lambda c} - \lambda r\right) \mathrm{d}\mu \\
&\leq r\,\mathrm{lip}_c(f) + \int \sup_{\lambda\in[0,\infty)} \left(f - f^{\lambda c} - \lambda r\right) \mathrm{d}\mu \\
&\overset{(6.14)}{\leq} r\,\mathrm{lip}_c(f) + \min\left\{\int (f - \overline{\mathrm{co}}\, f)\,\mathrm{d}\mu - r\,\mathrm{lip}_c(\overline{\mathrm{co}}\, f), 0\right\},
\end{aligned}$$

which implies (6.9). (Thm. 6.5) ■

Proof of Proposition 6.6. Let $x_0 \in X_0$ be a point at which $f(x_0) = \sup f(X_0)$. Then $\mathrm{cost}_c(\delta_{x_0}, \delta_{x_0}) = 0 \leq r$, and $\sup_{\nu\in\mathrm{B}_c(\delta_{x_0},r)} \int f\,\mathrm{d}\nu = f(x_0)$. Therefore

$$\mathit{\Delta}(f, \delta_{x_0}, r, c) = r\,\mathrm{lip}_c(f) + f(x_0) - f(x_0) = r\,\mathrm{lip}_c(f). \qquad (6.15)$$

And so we have

$$\begin{aligned}
r\,\mathrm{lip}_c(f) &\overset{(6.15)}{\leq} \sup_{\mu\in\mathfrak{P}(X_0)} \mathit{\Delta}(f,\mu,r,c) \\
&\overset{\mathrm{T6.5}}{\leq} r\,\mathrm{lip}_c(f) - \max\left\{r\,\mathrm{lip}_c(\overline{\mathrm{co}}\, f) - \rho(f), 0\right\} \\
&\leq r\,\mathrm{lip}_c(f),
\end{aligned}$$

which implies the claim. (Prop. 6.6) ■

6.3 Adversarial learning

Szegedy *et al.* [126] observe that deep neural networks, trained for image classification using empirical risk minimisation, exhibit a curious behaviour whereby an image, $x \in \mathbb{R}^n$, and a small, imperceptible amount of noise, $\epsilon_x \in \mathbb{R}^n$, may found so that the network classifies x and $x + \epsilon_x$ differently. Imagining that the troublesome noise vector is sought by an adversary seeking to defeat the classifier, such pairs have come to be known as *adversarial examples* [54, 73, 88].

Let X be a linear space and Y a topological space. Fix $\mu \in \mathfrak{P}(X \times Y)$, $r \geq 0$, and let d be a metric on X. The following objective has been proposed [viz. 25, 29, 79, 121] as a means of learning classifiers that are robust to adversarial examples

$$\int \sup_{\epsilon \in \mathrm{B}_d(0,r)} f(x+\epsilon, y)\mu(\mathrm{d}x \times \mathrm{d}y) = \int \sup_{\tilde{\omega} \in \mathrm{B}_{\tilde{d}}(\omega, r)} f(\tilde{\omega})\mu(\mathrm{d}\omega), \quad (6.16)$$

where $f : X \times Y \to \bar{\mathbb{R}}$ is the loss of some classifier, and in the equality we extend d to a metric on $\Omega \stackrel{\text{def}}{=} X \times Y$ with

$$\tilde{d}((x,y),(x',y')) \stackrel{\text{def}}{=} \begin{cases} d(x,x') & y = y' \\ \infty & y \neq y'. \end{cases}$$

The goal of this section is to prove a strong result linking (6.16) to the distributionally robust risk in (rB). We begin with Proposition 6.9 which verifies (6.16) is well defined. We then have a technical lemma before the main result, Theorem 6.11, is proven.

For a Borel measure $\mu \in \mathfrak{P}(\Omega)$, the completion of $\mathscr{B}(\Omega)$ with respect to μ is denoted $\mathscr{B}_\mu(\Omega)$. The *universal sigma algebra* on Ω is $\mathscr{U}(\Omega) \stackrel{\text{def}}{=} \bigcap_{\mu \in \mathfrak{P}(\Omega)} \mathscr{B}_\mu(\Omega)$. We say a function $f : X \to Y$ is *universally measurable* if for every open $U \subseteq Y$ there is $f^{-1}(U) \in \mathscr{U}(X)$. Universally measurable functions can be integrated under a Borel measure because for $\mu \in \mathfrak{P}(X)$, $f : X \to \bar{\mathbb{R}}$ is universally measurable if and only if there is a unique Borel $f_\mu : X \to \bar{\mathbb{R}}$ with $f(x) = f_\mu(x)$ for μ-almost every $x \in X$ [19, Lem. 7.27], and so we let $\int f \,\mathrm{d}\mu \stackrel{\text{def}}{=} \int f_\mu \,\mathrm{d}\mu$. The *push forward* of the measure $\mu \in \mathfrak{P}(X)$ by a measurable function $f : X \to Y$ is the measure $f_\#\mu \in \mathfrak{P}(Y)$ with $f_\#\mu(\mathrm{d}y) \stackrel{\text{def}}{=} \mu f^{-1}(\mathrm{d}y)$.

Proposition 6.9. *If $f : \Omega \to \bar{\mathbb{R}}$, $g : \Omega \to \mathbb{R}_{\geq 0}$, and $c : \Omega \times \Omega \to \bar{\mathbb{R}}_{\geq 0}$ are Borel, then the function $\omega \mapsto \sup_{\omega' \in \mathrm{B}_c(\omega, g(\omega))} f(\omega')$ is universally measurable.*

Proof. Let $T(\omega_1, \omega_2) \stackrel{\text{def}}{=} \iota_{\mathrm{B}_c(\omega_1, c(\omega_1))}(\omega_2)$ and fix $\omega_1 \in \Omega$. Since $\mathrm{B}_c(\omega_1, r)$ is closed for every $r \geq 0$, the level sets

$$\forall_{u \in \mathbb{R}} : \ \operatorname{lev}_{>u} T(\omega_1, \,\cdot\,) = \begin{cases} \Omega \setminus \mathrm{B}_d(\omega_1, g(\omega_1)) & u \geq 0 \\ \Omega & u < 0, \end{cases}$$

are all Borel, therefore $T(\omega_1, \cdot\,)$ is Borel for every $\omega_1 \in \Omega$.

Let $c_{\omega_2}(\omega_1) \overset{\text{def}}{=} c(\omega_1, \omega_2)$, fix $\omega_2 \in \Omega$ and consider

$$\begin{aligned}\operatorname{lev}_{=0} T(\,\cdot\,, \omega_2) &= \{\omega_1 \in \Omega \mid c(\omega_1, \omega_2) \le g(\omega_1)\} \\ &= \{\omega_1 \in \Omega \mid c_{\omega_2}(\omega_1) \le g(\omega_1)\} \\ &= \{\omega_1 \in \Omega \mid 0 \le g(\omega_1) - c_{\omega_2}(\omega_1)\} \\ &= \operatorname{lev}_{\ge 0}(g(\omega_1) - c_{\omega_2}(\omega_1)).\end{aligned}$$

Since g and c are Borel, so is the set $\operatorname{lev}_{=0} T(\,\cdot\,, \omega_2)$. By a similar argument, it's clear the set $\operatorname{lev}_{>0} T(\,\cdot\,, \omega_2)$ is Borel too. This shows that T is a Borel function. Then for all $u \in \mathbb{R}$, using the concave convention $\infty - \infty \overset{\text{def}}{=} -\infty$, we have

$$\begin{aligned}&\operatorname{lev}_{>u}\left(\sup_{\omega' \in \mathrm{B}_c(\,\cdot\,, g(\,\cdot\,))} f(\omega')\right) \\ &\qquad = \operatorname{lev}_{>u}\left(\sup_{\omega' \in \Omega} \left(f(\omega') - T(\,\cdot\,, \omega')\right)\right) \\ &\qquad = \operatorname{proj}_1\{(\omega_1, \omega_2) \in \Omega \times \Omega \mid f(\omega_2) - T(\omega_1, \omega_2) > u\}, \quad (6.17)\end{aligned}$$

where $\operatorname{proj}_1(\omega_1, \omega_2) \overset{\text{def}}{=} \omega_1$. Since f and T are Borel, the argument of the projection in (6.17) is Borel too. The projection of a Borel set is universally measurable [19, Prop. 7.39, Cor. 7.42.1], therefore $\omega \mapsto \sup_{\omega' \in \mathrm{B}_c(\omega, g(\omega))} f(\omega')$ is universally measurable. ∎

Lemma 6.10 will be used to show an equality result in Theorem 6.11.

Lemma 6.10. *Assume* (Ω, c) *is a compact Polish space and* $\mu \in \mathfrak{P}(\Omega)$ *is non-atomic. For* $r > 0$ *and* $\nu^\star \in \mathrm{B}_c(\mu, r)$ *there is a sequence* $(f_i)_{i \in \mathbb{N}} \subseteq A_\mu(r) \overset{\text{def}}{=} \{f \in \mathscr{L}_0(\Omega, \Omega) \mid \int c\,\mathrm{d}(\mathrm{id}, f)_\# \mu \le r\}$ *with* $(f_i)_\# \mu$ *converging at* $\nu^\star$ *in* $\sigma(\mathfrak{P}(\Omega), \mathrm{C}(\Omega))$.

Proof. Let $P(\mu, \nu) \overset{\text{def}}{=} \{f \in \mathscr{L}_0(X, X) \mid f_\# \mu = \nu\}$. Since μ is non-atomic and c is continuous we have [via 101, Thm. B]

$$\forall_{\nu \in \mathfrak{P}(\Omega)} : \inf_{f \in P(\mu, \nu)} \int c\,\mathrm{d}(\mathrm{id}, f)_\# \mu = \operatorname{cost}_c(\mu, \nu).$$

Let $r^\star \overset{\text{def}}{=} \operatorname{cost}_c(\mu, \nu^\star)$, obviously $r^\star \le r$. Assume $r^\star > 0$, otherwise the lemma is trivial. Fix a sequence $(\epsilon_k)_{k\in\mathbb{N}} \subseteq (0, r^\star)$ with $\epsilon_k \to 0$. For $u \geq 0$ let $\nu(u) \stackrel{\text{def}}{=} \mu + u(\nu^\star - \mu)$. Then

$$\operatorname{cost}_c(\mu, \nu(0)) = 0 \quad \text{and} \quad \operatorname{cost}_c(\mu, \nu(1)) = r^\star,$$

and because cost_c metrises the $\sigma(\mathfrak{P}(\Omega), \mathrm{C}(\Omega))$-topology on $\mathfrak{P}(\Omega)$ [135, Cor. 13], the mapping $u \mapsto \operatorname{cost}_c(\mu, \nu(u))$ is $\sigma(\mathfrak{P}(\Omega), \mathrm{C}(\Omega))$-continuous. Then by the intermediate value theorem for every $k \in \mathbb{N}$ there is some $u_k > 0$ with $\operatorname{cost}_c(\mu, \nu(u_k)) = r^\star - \epsilon_k$, forming a sequence $(u_k)_{k\in\mathbb{N}} \subseteq [0, 1]$. Then for every k there is a sequence $(f_{jk})_{j\in\mathbb{N}} \subseteq P(\mu, \nu(u_k))$ so that $(f_{jk})_{\#}\mu \stackrel{*}{\rightharpoonup} \nu(k)$ and

$$\begin{aligned}\lim_{j\in\mathbb{N}} \int c\, \mathrm{d}(\mathrm{id}, f_{jk})_{\#}\mu &= \inf_{f\in P(\mu,\nu(k))} \int c\, \mathrm{d}(\mathrm{id}, f_k)_{\#}\mu \\ &= \operatorname{cost}_c(\mu, \nu(k)) \\ &= r^\star - \epsilon_k.\end{aligned}$$

Therefore for every $k \in \mathbb{N}$ there exists $j_k \geq 0$ so that for every $j \geq j_k$

$$\int c\, \mathrm{d}(\mathrm{id}, f_{jk})_{\#}\mu \leq r^\star. \tag{6.18}$$

Let us pass directly to this subsequence of $(f_{jk})_{j\in\mathbb{N}}$ for every $k \in \mathbb{N}$ so that (6.18) holds for all $j, k \in \mathbb{N}$. Next by construction we have $\nu(u_k) \to \nu^\star$. Therefore $(f_{jk})_{j,k\in N}$ has a subsequence in k so that $(f_{jk})_{\#}\mu \stackrel{*}{\rightharpoonup} \nu^\star$. By ensuring (6.18) is satisfied, the sequences $(f_{jk})_{j\in\mathbb{N}} \subseteq A_\mu(r)$ for every $k \in \mathbb{N}$. ■

We can now prove our main result for this section.

Theorem 6.11. *Assume (X, c) is a separable Banach space. Fix $\mu \in \mathfrak{P}(X)$ and for $r \geq 0$ let*

$$R_\mu(r) \stackrel{\text{def}}{=} \left\{ g \in \mathscr{L}_0(X, \mathbb{R}_{\geq 0}) \;\middle|\; \int g\, \mathrm{d}\mu \leq r \right\}.$$

Then for $f \in \mathscr{L}_0(\Omega, \bar{\mathbb{R}})$ and $r \geq 0$ there is

$$\sup_{g\in R_\mu(r)} \int \mu(\mathrm{d}\omega) \sup_{\omega'\in \mathrm{B}_c(\omega, g(\omega))} f(\omega') \leq \sup_{\nu\in \mathrm{B}_c(\mu,r)} \int f\, \mathrm{d}\nu. \tag{6.19}$$

Furthermore if μ is non-atomically concentrated on a compact subset of X, on which f is continuous with the subspace topology, then (6.19) *holds as an equality.*

Remark 6.12. It's easy to see that the left side of (6.19) upper bounds (6.16) by observing the constant function $g_r \equiv r$ is included in the supremum over $R_\mu(r)$.

Theorem 6.11 generalises and subsumes a number of existing results [48, Cor. 2 (iv), 124, Prop. 3.1, 124, Prop. 3.1, 120, Thm. 12] to relate the adversarial risk minimisation (6.16) to the distributionally robust risk in Theorem 6.5. The previous results mentioned are all are formulated with respect to an empirical distribution, that is, an average of Dirac masses. Of course any finite set is compact, and so these empirical distributions satisfy the concentration assumption.

Proof of Theorem 6.11. When $r = 0$, the set $R_\mu(r)$ consists of the set of functions g which are 0 μ-almost everywhere, in which case $\mathrm{B}_c(x, g(x)) = \{0\}$ for μ-almost all $x \in X$. Thus the left hand side of (6.19) is equal to $\int f(x)\mu(\mathrm{d}x)$. Since c is a norm, $c(0) = 0$, and by a similar argument there is equality with the right hand side. We now complete the proof for the cases where $r > 0$.

(6.19): For $g \in R_\mu(r)$, let $\Gamma_g : X \rightrightarrows X$ denote the set-valued mapping with $\Gamma_g(x) \overset{\text{def}}{=} \mathrm{B}_c(x, g(x))$. Let $\mathscr{L}_0(X, \Gamma_g)$ denote the set of Borel $a : X \to X$ so that $a(x) \in \Gamma_g(x)$ for μ-almost all $x \in X$. Let $A_\mu(r) \overset{\text{def}}{=} \bigcup_{g \in \mathbb{R}_\mu(r)} \mathscr{L}_0(X, \Gamma_g)$. Clearly for every $a \in A_\mu(r)$ there is

$$r \geq \int c(x, a(x))\,\mathrm{d}\mu = \int c\,\mathrm{d}(\mathrm{id}, a)_{\#}\mu,$$

which shows $\{a_{\#}\mu \mid a \in A_\mu(r)\} \subseteq \mathrm{B}_c(\mu, r)$. Then if there is equality in (6.20), we have

$$\begin{aligned}
\sup_{g \in R_\mu(r)} \int \sup_{x' \in \Gamma_g(x)} f(x) &= \sup_{g \in R_\mu(r)} \sup_{a \in \mathscr{L}_0(X, \Gamma_g)} \int f\,\mathrm{d}a_{\#}\mu \qquad (6.20)\\
&= \sup_{a \in A_\mu(r)} \int f\,\mathrm{d}a_{\#}\mu\\
&\leq \sup_{\nu \in \mathrm{B}_c(\mu, r)} \int f\,\mathrm{d}\nu,
\end{aligned}$$

which proves the inequality (6.19).

(6.20): To complete the proof we will now justify the exchange of integration and supremum. The set $\mathscr{L}_0(X, \mathit{\Gamma}_g)$ is trivially decomposable [50, see the remark at the bottom of p. 323, Def. 2.1]. By assumption f is Borel measurable. Since f is measurable, any decomposable subset of $\mathscr{L}_0(X, X)$ is f-decomposable [50, Prop. 5.3] and f-linked [50, Prop. 3.7 (i)]. Giner [50, Thm. 6.1 (c)] therefore allows us to exchange integration and supremum in (6.20).

Equality in (6.19): Under the additional assumptions there exists $\nu^\star \in \mathfrak{P}(\mathit{\Omega})$ with [via 21, Prop. 2]

$$\int f \,\mathrm{d}\nu^\star = \sup_{\nu \in \mathrm{B}_c(\mu, r)} \int f \,\mathrm{d}\nu.$$

The compact subset where μ is concentrated and non-atomic is a Polish space with the Banach metric. Therefore using Lem. 6.10 there is a sequence $(f_i)_{i\in\mathbb{N}} \subseteq A_\mu(r)$ so that

$$\lim_{i\in\mathbb{N}} \int f_i \,\mathrm{d}\mu = \int f \,\mathrm{d}\nu^\star = \sup_{\nu \in \mathrm{B}_c(\mu, r)} \int f \,\mathrm{d}\nu,$$

proving equality in (6.19). (Thm. 6.11) ■

6.4 Conclusion

Risk minimisation can fail to be optimal when there is some misspecification of the distribution, such as when working with its empirical counterpart. Therefore we must turn to other techniques in order to ensure stability when learning a model. The robust Bayes framework provides a systematic approach to these problems, however it leaves open the choice as to which uncertainty set is most appropriate. We avoid this question by showing that the popular Lipschitz regularisation corresponds to robust Bayes using a transportation-cost-based uncertainty set. To further justify this choice of uncertainty set we have seen that there are strong connections linking the transportation cost uncertainty set to phenomenon of adversarial examples.

To do this we have borrowed tools from the nonconvex optimisation literature. In particular the closed convex envelope appears to be of somewhat novel application in this area. By its introduction we have been able to maintain tractability while making minimal assumptions about the model class or loss function so that this theory can be applied to popular exotic model classes such as deep neural networks.

Symbols

A^+	The dual cone of the set A.
$A^\perp$	The orthogonal space of the set A.
A^-	The negative of the dual cone of the set A.
$\bar{\mathbb{R}}$	The set $[-\infty, +\infty]$.
$[k]$	The set $\{1, 2, \ldots, k\}$.
A_∞	The asymptotic cone of the set A.
$\mathscr{B}(L)$	The collection of Borel subsets of L.
$\mathrm{bc}(A)$	The barrier cone of the set A.
$\mathrm{C}(X)$	The set of real, continuous functions on X.
$\mathrm{cl}(A)$, $\overline{A}$	The closure of the set A.
$\mathrm{co}(A)$	The convex hull of the set A.
$\mathrm{cl}^*(A)$, $\overline{A}^*$	The weak closure of the set A.
$\mathrm{B}_d(\mu, r)$	The d-metric ball of radius r centered at μ.
$\mathrm{cost}_c(\mu, \nu)$	The c-transportation cost of transporting the mass of μ to ν.
$\widehat{\partial} f$	The Moreau–Rockerfellar superdifferential of the function f.
$\partial_\epsilon f$	The approximate or ϵ-subdifferential of the function f.
∂f	The Moreau–Rockerfellar subdifferential of the function f.
δ_x	The Dirac measure at x.
$\mathrm{dom}\, f$	The domain of the mapping f.
$\mathrm{epi}(f)$	The epigraph of the function f.
ev_x	The evaluation operator with $\mathrm{ev}_x(f)(y) \stackrel{\text{def}}{=} f(x, y)$ where $f \in \mathscr{L}_0(X \times Y)$.
$A \diamond B$	The harmonic sum of A and B.
ι_A	The indicator function of the set A.
$\mathscr{L}_p(\Omega, \lambda)$	The Lebesgue space of λ-measurable functions $f : \Omega \to \mathbb{R}$ for which $\lvert f\rvert_p < \infty$.

$\mathscr{L}_0(X,Y)$	The set of Borel mappings $X \to Y$.
$\mathrm{lev}_{\leq c} f$	The c lower level set of the function f.
$\mathrm{lip}_c(f)$	The least c-Lipschitz constant of the function f.
$\mathcal{M}_0(L)$	The collection of subsets of L which are convex, $L_{\geq 0}$-full, bounded, contain both 0 and an order unit of $L_{\geq 0}$.
$\mathcal{M}_\infty(K)$	The collection of of subsets M of the cone K which are closed, convex, containing an order unit of K and have $\operatorname{pos} M = K \setminus \{0\}$.
$\oplus_M(A_1, \ldots, A_m)$	The M-sum of the sets $A_1, \ldots, A_k$.
$\square_M(A_1, \ldots, A_m)$	The dual M-sum of the sets $A_1, \ldots, A_k$.
μ_A	The gauge of the set A.
N_A	The normal cone of the set A.
$\mathscr{N}(x)$	The neighbourhood filter at x.
ν_A	The co-gauge of the set A.
$\mathfrak{P}(L)$	The set of probability measures on L.
$L_{\geq 0}$	The positive cone in the ordered vector space $(L, \geq)$.
$\Pi(\mu, \nu)$	The set of couplings joining μ to ν..
A°	The polar of the set A.
$\mathrm{pos}(A)$	The conic hull of the set A.
proj_1	The operator sending $(x_1, x_2) \mapsto x_1$.
$\rho(f)$	The lack of convexity of the function f.
σ_A	The support of the set A.
$\sigma(L, L^*)$	The weakest topology on L that generates L^*.
$\mathrm{sp}(\ell)$	The superprediction set of the loss function ℓ.
$\tau(L, L^*)$	The strongest topology on L that generates L^*, more commonly known as the Mackey topology.
$\tau_{\geq}(K)$	The order topology on L generated by the cone K.
$\mathscr{U}(X)$	The universal sigma algebra on X.
$A^\triangledown$	The antipolar of the set A.
ζ_A	The co-support of the set A.

Index

Bibliography

[1] Ali, S. M. and Silvey, S. D. "A General Class of Coefficients of Divergence of One Distribution from Another". *Journal of the Royal Statistical Society.* Series B (Methodological) 28.1 (1966), pp. 131–142. DOI: 10.1111/j.2517-6161.1966.tb00626.x.

[2] Aliprantis, C. D. and Border, K. *Infinite Dimensional Analysis.* 3rd ed. Berlin, Germany: Springer-Verlag, 2006. DOI: 10.1007/3-540-29587-9.

[3] Aliprantis, C. D. and Tourky, R. *Cones and Duality.* Graduate Studies in Mathematics. Providence, RI, USA: American Mathematical Society, 2007.

[4] Anil, C., Lucas, J., and Grosse, R. "Sorting out Lipschitz Function Approximation". *Proceedings of the 36th International Conference on Machine Learning.* Vol. 97. Proceedings of Machine Learning Research. Long Beach, CA, USA: PMLR, June 9–15, 2019, pp. 291–301.

[5] Arjovsky, M., Chintala, S., and Bottou, L. "Wasserstein Generative Adversarial Networks". *Proceedings of the 34th International Conference on Machine Learning.* Vol. 70. Proceedings of Machine Learning Research. Sydney, Australia: PMLR, Aug. 6–11, 2017, pp. 214–223.

[6] Askari, A., d'Aspremont, A., and El Ghaoui, L. "Naive Feature Selection: Sparsity in Naive Bayes" (May 23, 2019). arXiv: 1905.09884 [cs, stat].

[7] Asplund, E. and Rockafellar, R. T. "Gradients of Convex Functions". *Transactions of the American Mathematical Society* 139 (1969), pp. 443–443. DOI: 10.1090/s0002-9947-1969-0240621-x.

[8] Asplund, E. "Fréchet Differentiability of Convex Functions". *Acta Mathematica* 121 (1968), pp. 31–47. DOI: 10.1007/bf02391908.

[9] Atiyah, M. F. “How Research Is Carried Out”. *Bulletin of the Institute of Mathematics and its Applications* 10 (1974), pp. 232–234.

[10] Aubin, J.-P. *Mathematical Methods of Game and Economic Theory*. Burlington: Elsevier Science, 2014.

[11] Aubin, J.-P. and Ekeland, I. “Estimates of the Duality Gap in Nonconvex Optimization”. *Mathematics of Operations Research* 1.3 (1976), pp. 225–245. DOI: 10.1287/moor.1.3.225. JSTOR: 3689565.

[12] ——, *Applied Nonlinear Analysis*. Pure and Applied Mathematics. Mineola, USA: Wiley-Interscience, 1984.

[13] Aubin, J.-P. and Frankowska, H. *Set-Valued Analysis*. Modern Birkhauser Classics. Boston, MA, USA: Birkhäuser, 2009.

[14] Auslender, A. and Teboulle, M. *Asymptotic Cones and Functions in Optimization and Variational Inequalities*. 1st ed. Springer Monographs in Mathematics. New York, NY, USA: Springer-Verlag, 2003. DOI: 10.1007/b97594.

[15] Barbara, A. and Crouzeix, J.-P. “Concave Gauge Functions and Applications”. *Zeitschrift für Operations Research* 40.1 (Mar. 1, 1994), pp. 43–74. DOI: 10.1007/bf01414029.

[16] Barbu, V. and Precupanu, T. *Convexity and Optimization in Banach Spaces*. 4th ed. Springer Monographs in Mathematics. Dordrecht: Springer Netherlands, 2012. DOI: 10.1007/978-94-007-2247-7.

[17] Bartlett, P. L., Jordan, M. I., and Mcauliffe, J. D. “Large Margin Classifiers: Convex Loss, Low Noise, and Convergence Rates”. *Advances in Neural Information Processing Systems 16*. MIT Press, 2004, pp. 1173–1180.

[18] Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. Springer Series in Statistics. New York, NY, USA: Springer-Verlag, 1993.

[19] Bertsekas, D. P. and Shreve, S. E. *Stochastic Optimal Control: The Discrete Time Case*. Mathematics in Science and Engineering v. 139. New York, NY, USA: Academic Press, 1978.

[20] Blanchet, J., Kang, Y., and Murthy, K. “Robust Wasserstein Profile Inference and Applications to Machine Learning”. *Journal of Applied Probability* 56.3 (2019), pp. 830–857. DOI: 10.1017/jpr.2019.49.

[21] Blanchet, J. and Murthy, K. “Quantifying Distributional Model Risk via Optimal Transport”. *Mathematics of Operations Research* 44.2 (May 2019), pp. 565–600. DOI: 10.1287/moor.2018.0936.

[22] Bogachev, V. and Smolyanov, O. *Topological Vector Spaces and Their Applications*. Springer Monographs in Mathematics. Cham, Switzerland: Springer International Publishing, 2017. DOI: 10.1007/978-3-319-57117-1.

[23] Brehmer, J. R. and Gneiting, T. "Properization: Constructing Proper Scoring Rules via Bayes Acts". *Annals of the Institute of Statistical Mathematics* (Feb. 22, 2019). DOI: 10.1007/s10463-019-00705-7.

[24] Buja, A., Stuetzle, W., and Shen, Y. *Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications*. Nov. 3, 2005.

[25] Carlini, N. and Wagner, D. "Towards Evaluating the Robustness of Neural Networks". *IEEE Symposium on Security and Privacy*. San Jose, CA, USA: IEEE, May 2017, pp. 39–57. DOI: 10.1109/sp.2017.49.

[26] Chakrabarty, A. K., Shunmugaraj, P., and Zălinescu, C. "Continuity Properties for the Subdifferential and ϵ-Subdifferential of a Convex Function and Its Conjugate". *Journal of Convex Analysis* 14.3 (2007), pp. 479–514.

[27] Chernov, A. and Vovk, V. "Prediction with Expert Evaluators' Advice". *Algorithmic Learning Theory* (2009), pp. 8–22. DOI: 10.1007/978-3-642-04414-4_6.

[28] Choquet, G. "Les Cônes Convexes Faiblement Complets Dans l'analyse". *Proceedings of the International Congress of Mathematicians*. Stockholm, Sweden, 1962, pp. 317–330.

[29] Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. "Parseval Networks: Improving Robustness to Adversarial Examples". *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Sydney, Australia: Proceedings of machine learning research, Aug. 6–11, 2017, pp. 854–863.

[30] Cranko, Z., Menon, A., Nock, R., Ong, C. S., Shi, Z., and Walder, C. "Monge Blunts Bayes: Hardness Results for Adversarial Training". *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Long Beach, CA, USA: Proceedings of machine learning research, June 9–15, 2019, pp. 1406–1415.

[31] Cranko, Z. and Nock, R. "Boosted Density Estimation Remastered". *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Long Beach, CA, USA: Proceedings of machine learning research, June 9–15, 2019, pp. 1416–1425.

[32] Csiszár, I. "Information-Type Measures of Difference of Probability Distributions and Indirect Observations". *Studia Scientiarum Mathematicarum Hungarica* 2 (1967), pp. 299–318.

[33] Dai, B., He, N., Dai, H., and Song, L. “Provable Bayesian Inference via Particle Mirror Descent”. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Vol. 51. Cadiz, Spain: Proceedings of machine learning research, May 9–11, 2016, pp. 985–994.

[34] Dawid, A. P. *Coherent Measures of Discrepancy, Uncertainty and Dependence, with Applications to Bayesian Predictive Experimental Design*. 139. London, UK: Department of Statistics, University College London, 1998.

[35] ——, “The Geometry of Proper Scoring Rules”. *Annals of the Institute of Statistical Mathematics* 59.1 (Mar. 1, 2007), pp. 77–93. DOI: 10.1007/s10463-006-0099-8.

[36] Dawid, A. P. and Musio, M. “Theory and Applications of Proper Scoring Rules”. *METRON* 72.2 (Aug. 2014), pp. 169–183. DOI: 10.1007/s40300-014-0039-y.

[37] Dedieu, J.-P. “Cône Asymptote d’un Ensemble Non Convexe”. *Comptes Rendus Hebdomadaires des Séances de l’Académie des Sciences. Séries A et B* 285.7 (1977), A501–A503.

[38] Denkowski, Z., Migórski, S., and Papageorgiou, N. S. *An Introduction to Nonlinear Analysis: Theory*. Boston, MA, USA: Springer, 2003. DOI: 10.1007/978-1-4419-9158-4.

[39] Dolecki, S. and Kurcyusz, S. “On Φ-Convexity in Extremal Problems”. *SIAM Journal on Control and Optimization* 16.2 (Mar. 1978), pp. 277–300. DOI: 10.1137/0316018.

[40] Erven, T. van, Reid, M. D., and Williamson, R. C. “Mixability Is Bayes Risk Curvature Relative to Log Loss”. *Journal of Machine Learning Research* 13 (May 2012), pp. 1639–1663.

[41] Fan, K. “Some Properties of Convex Sets Related to Fixed Point Theorems”. *Mathematische Annalen* 266.4 (Dec. 1, 1984), pp. 519–537. DOI: 10.1007/bf01458545.

[42] Farnia, F., Zhang, J., and Tse, D. “Generalizable Adversarial Training via Spectral Normalization”. International Conference on Learning Representations. New Orleans, LA, United States, May 6, 2019.

[43] Firey, W. J. “Polar Means of Convex Bodies and a Dual to the Brunn–Minkowski Theorem”. *Canadian Journal of Mathematics* 13 (1961), pp. 444–453. DOI: 10.4153/cjm-1961-037-0.

[44] ——, “p-Means of Convex Bodies”. *Mathematica Scandinavica* 10 (1962), pp. 17–24. DOI: 10.7146/math.scand.a-10510.

[45] ——, “Some Means of Convex Bodies”. *Transactions of the American Mathematical Society* 129 (1967), pp. 181–217. DOI: 10.2307/1994373.

[46] Frank, M. and Wolfe, P. "An Algorithm for Quadratic Programming". *Naval Research Logistics Quarterly* 3.1-2 (Mar. 1956), pp. 95–110. DOI: `10.1002/nav.3800030109`.

[47] Freund, Y. and Schapire, R. E. "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting". *Journal of Computer and System Sciences* 55.1 (Aug. 1997), pp. 119–139. DOI: `10.1006/jcss.1997.1504`.

[48] Gao, R. and Kleywegt, A. J. "Distributionally Robust Stochastic Optimization with Wasserstein Distance" (July 16, 2016). arXiv: `1604.02199 [math]`.

[49] Gardner, R. J., Hug, D., and Weil, W. "Operations between Sets in Geometry". *Journal of the European Mathematical Society* 15.6 (2013), pp. 2297–2352. DOI: `10.4171/jems/422`.

[50] Giner, E. "Necessary and Sufficient Conditions for the Interchange between Infimum and the Symbol of Integration". *Set-Valued and Variational Analysis* 17.4 (2009), pp. 321–357. DOI: `10.1007/s11228-009-0119-y`.

[51] Gneiting, T. and Raftery, A. E. "Strictly Proper Scoring Rules, Prediction, and Estimation". *Journal of the American Statistical Association* 102.477 (Mar. 1, 2007), pp. 359–378. DOI: `10.1198/016214506000001437`.

[52] Good, I. J. "Rational Decisions". *Breakthroughs in Statistics: Foundations and Basic Theory*. New York, NY, USA: Springer, 1992, pp. 365–377. DOI: `10.1007/978-1-4612-0919-5_24`.

[53] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. "Generative Adversarial Nets". *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 2672–2680.

[54] Goodfellow, I., Shlens, J., and Szegedy, C. "Explaining and Harnessing Adversarial Examples". *International Conference on Learning Representations* (2015).

[55] Gouk, H., Frank, E., Pfahringer, B., and Cree, M. J. "Regularisation of Neural Networks by Enforcing Lipschitz Continuity" (Sept. 14, 2018). arXiv: `1804.04368 [cs, stat]`.

[56] Gouk, H., Pfahringer, B., Frank, E., and Cree, M. J. "MaxGain: Regularisation of Neural Networks by Constraining Activation Magnitudes". *Machine Learning and Knowledge Discovery in Databases*. Cham, Switzerland: Springer International Publishing, 2019, pp. 541–556.

[57] Gowers, W. T. "The Two Cultures of Mathematics". *Mathematics: frontiers and perspectives* (2000), pp. 65–78.

[58] Grover, A. and Ermon, S. "Boosted Generative Models". *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New Orleans, LA, USA: AAAI Press, Feb. 2–7, 2018, pp. 3077–3084.

[59] Grünwald, P. D. and Dawid, A. P. "Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory". *The Annals of Statistics* 32.4 (Aug. 2004), pp. 1367–1433. DOI: 10.1214/009053604000000553.

[60] Guo, F., Wang, X., Fan, K., Broderick, T., and Dunson, D. B. "Boosting Variational Inference" (Mar. 1, 2017). arXiv: 1611.05559 [cs, stat].

[61] Hantoute, A., López, M. A., and Zălinescu, C. "Subdifferential Calculus Rules in Convex Analysis: A Unifying Approach via Pointwise Supremum Functions". *SIAM Journal on Optimization* 19.2 (Jan. 2008), pp. 863–882. DOI: 10.1137/070700413.

[62] Hendrickson, A. D. and Buehler, R. J. "Proper Scores for Probability Forecasters". *The Annals of Mathematical Statistics* 42.6 (1971), pp. 1916–1921. DOI: 10.1214/aoms/1177693057. JSTOR: 2240117.

[63] Hiriart-Urruty, J.-B. "A General Formula on the Conjugate of the Difference of Functions". *Canadian Mathematical Bulletin* 29.4 (Dec. 1, 1986), pp. 482–485. DOI: 10.4153/cmb-1986-076-7.

[64] Hiriart-Urruty, J.-B. "From Convex Optimization to Nonconvex Optimization. Necessary and Sufficient Conditions for Global Optimality". *Nonsmooth Optimization and Related Topics*. Boston, MA, USA: Springer, 1989, pp. 219–239. DOI: 10.1007/978-1-4757-6019-4_13.

[65] Hiriart-Urruty, J.-B. and Lemaréchal, C. *Fundamentals of Convex Analysis*. 2nd ed. Grundlehren Text Editions. Berlin, Germany: Springer-Verlag, 2004.

[66] ——, *Convex Analysis and Minimization Algorithms II*. Berlin, Germany: Springer-Verlag, 2010.

[67] Husein, H., Balle, B., Cranko, Z., and Nock, R. "Local Differential Privacy for Sampling". *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*. Palermo, Italy: Proceedings of machine learning research, June 3–5, 2020.

[68] Jose, V. R. R., Nau, R. F., and Winkler, R. L. "Scoring Rules, Generalized Entropy, and Utility Maximization". *Operations Research* 56.5 (Oct. 2008), pp. 1146–1157. DOI: 10.1287/opre.1070.0498.

[69] Kamalaruban, P., Williamson, R. C., and Zhang, X. "Exp-Concavity of Proper Composite Losses". *Proceedings of the 28th Conference on Learning Theory*. Proceedings of Machine Learning Research. Paris, France: PMLR, June 6, 2015, pp. 1035–1065.

[70] Kerdreux, T., Colin, I., and d'Aspremont, A. "An Approximate Shapley-Folkman Theorem" (July 1, 2019). arXiv: `1712.08559 [math]`.

[71] Kolmogorov, A. N. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin, Germany: Springer, 1933. DOI: `10.1007/978-3-642-49888-6`.

[72] Kolmogorov, A. N., Bharucha-Reid, A. T., and Morrison, N. *Foundations of the Theory of Probability*. 2nd ed. Muneola, New York: Dover Publications, Inc, 2018.

[73] Kurakin, A., Goodfellow, I., and Bengio, S. "Adversarial Examples in the Physical World" (Feb. 10, 2017). arXiv: `1607.02533 [cs, stat]`.

[74] Kuratowski, K. and Ryll-Nardzewski, C. "A General Theorem on Selectors". *Bulletin L'Académie Polonaise des Science, Série des Sciences Mathématiques, Astronomiques et Physiques* 13.6 (1965), pp. 397–403.

[75] Kutateladze, S. S. and Rubinov, A. M. "Minkowski Duality, and Its Applications". *Russian Mathematical Surveys* 27.3 (June 30, 1972), pp. 137–191. DOI: `10.1070/rm1972v027n03abeh001380`.

[76] Lemaréchal, C. and Renaud, A. "A Geometric Study of Duality Gaps, with Applications". *Mathematical Programming* 90.3 (May 2001), pp. 399–427. DOI: `10.1007/pl00011429`.

[77] Li, J. Q. and Barron, A. R. "Mixture Density Estimation". *Advances in Neural Information Processing Systems 12*. MIT Press, 2000, pp. 279–285.

[78] Luc, D. T. *Theory of Vector Optimization*. Lecture Notes in Economics and Mathematical Systems 319. Berlin, Germany: Springer-Verlag, 1989.

[79] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. "Towards Deep Learning Models Resistant to Adversarial Attacks". *International Conference on Learning Representations*. 2018.

[80] Masnadi-shirazi, H. and Vasconcelos, N. "On the Design of Loss Functions for Classification: Theory, Robustness to Outliers, and SavageBoost". *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc., 2009, pp. 1049–1056.

[81] McCarthy, J. "Measures of the Value of Information". *Proceedings of the National Academy of Sciences of the United States of America* 42.9 (1956), p. 654. DOI: `10.1073/pnas.42.9.654`.

[82] McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. 2nd ed. Monographs on Statistics and Applied Probability 37. Boca Raton: Chapman & Hall/CRC, 1998.

[83] McShane, E. J. “Extension of Range of Functions”. *Bulletin of the American Mathematical Society* 40.12 (1934), pp. 837–842. DOI: 10.1090/s0002-9904-1934-05978-0.

[84] Merkle, E. C. and Steyvers, M. “Choosing a Strictly Proper Scoring Rule”. *Decision Analysis* 10.4 (Dec. 2013), pp. 292–304. DOI: 10.1287/deca.2013.0280.

[85] Mhammedi, Z. and Williamson, R. C. “Constant Regret, Generalized Mixability, and Mirror Descent”. *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 7419–7428.

[86] Miller, A. C., Foti, N. J., and Adams, R. P. “Variational Boosting: Iteratively Refining Posterior Approximations”. *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. Sydney, Australia: PMLR, Aug. 6–11, 2017, pp. 2420–2429.

[87] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. “Spectral Normalization for Generative Adversarial Networks”. *International Conference on Learning Representations*. 2018.

[88] Moosavi Dezfooli, S. M., Fawzi, A., Fawzi, O., and Frossard, P. “Universal Adversarial Perturbations”. *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA, Ieee, 2017, p. 9. DOI: 10.1109/Cvpr.2017.17.

[89] Moreau, J.-J. “Semi-Continuité Du Sous-Gradient d’une Fonctionnelle”. *Comptes Rendus Hebdomadaires des Séances de l’Académie des Sciences* 260 (1965), pp. 1067–1070.

[90] ——, “Inf-Convolution, Sous-Additivité, Convexité Des Fonctions Numériques”. *Journal de Mathématiques Pures et Appliquées* 49 (1970), pp. 109–154.

[91] Naito, K. and Eguchi, S. “Density Estimation with Minimization of U-Divergence”. *Machine Learning* 90.1 (Jan. 2013), pp. 29–57. DOI: 10.1007/s10994-012-5298-3.

[92] Nelder, J. A. and Wedderburn, R. W. M. “Generalized Linear Models”. *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), pp. 370–384. DOI: 10.2307/2344614. JSTOR: 2344614.

[93] Nguyen, X., Wainwright, M. J., and Jordan, M. I. “Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization”. *IEEE Transactions on Information Theory* 56.11 (Nov. 2010), pp. 5847–5861. DOI: 10.1109/tit.2010.2068870.

[94] Nock, R., Cranko, Z., Menon, A. K., Qu, L., and Williamson, R. C. "f-GANs in an Information Geometric Nutshell". *Advances in Neural Information Processing Systems 30*. Long Beach, CA, USA: Curran Associates, Inc., 2017, pp. 456–464.

[95] Nock, R. and Nielsen, F. "On the Efficient Minimization of Classification Calibrated Surrogates". *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc., 2009, pp. 1201–1208.

[96] Nowozin, S., Cseke, B., and Tomioka, R. "f-GAN: Training Generative Neural Samplers Using Variational Divergence Minimization". *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016, pp. 271–279.

[97] Penot, J.-P. "Duality for Radiant and Shady Programs". *Acta Mathematica Vietnamica* 22.2 (1997), pp. 541–566.

[98] ——, "Unilateral Analysis and Duality". *Essays and Surveys in Global Optimization*. New York, USA: Springer-Verlag, 2005, pp. 1–37. DOI: `10.1007/0-387-25570-2_1`.

[99] ——, *Calculus without Derivatives*. Graduate Texts in Mathematics 266. New York, NY, USA: Springer, 2013.

[100] Penot, J.-P. and Zălinescu, C. "Harmonic Sum and Duality". *Journal of Convex Analysis* 7.1 (2000), pp. 95–114.

[101] Pratelli, A. "On the Equality between Monge's Infimum and Kantorovich's Minimum in Optimal Mass Transportation". *Annales de l'Institut Henri Poincare (B) Probability and Statistics* 43.1 (Jan. 2007), pp. 1–13. DOI: `10.1016/j.anihpb.2005.12.001`.

[102] Reid, M. D. and Williamson, R. C. "Composite Binary Losses". *Journal of Machine Learning Research* 11 (Sep 2010), pp. 2387–2422.

[103] ——, "Convexity of Proper Composite Binary Losses". *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. PMLR, Mar. 31, 2010, pp. 637–644.

[104] ——, "Information, Divergence and Risk for Binary Experiments". *Journal of Machine Learning Research* 12 (July 2011), pp. 731–817.

[105] Rockafellar, R. T. "Level Sets and Continuity of Conjugate Convex Functions". *Transactions of the American Mathematical Society* 123 (1966), pp. 46–63. DOI: `10.2307/1994612`.

[106] ——, *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton, USA: Princeton University Press, 1997.

[107] Rosset, S. and Segal, E. "Boosting Density Estimation". *Advances in Neural Information Processing Systems 15*. MIT Press, 2003, pp. 657–664.

[108] Rubinov, A. M. “Antihomogeneous Conjugacy Operators in Convex Analysis”. *Journal of Convex Analysis* 2.1-2 (1995), pp. 291–307.

[109] ——, “Radiant Sets and Their Gauges”. *Quasidifferentiability and Related Topics*. Boston, MA, USA: Springer, 2000, pp. 235–261. DOI: 10.1007/978-1-4757-3137-8_10.

[110] ——, *Abstract Convexity and Global Optimization*. New York, NY, USA: Springer, 2011.

[111] Rubinov, A. M. and Yagubov, A. A. “The Space of Star-Shaped Sets and Its Applications in Nonsmooth Optimization”. *Quasidifferential Calculus*. Berlin, Germany: Springer, 1986, pp. 176–202. DOI: 10.1007/BFb0121146.

[112] Rudin, W. *Functional Analysis*. 2nd ed. International Series in Pure and Applied Mathematics. Boston, MA, USA: McGraw-Hill, 1991.

[113] Savage, L. J. “Elicitation of Personal Probabilities and Expectations”. *Journal of the American Statistical Association* 66 (1971), pp. 783–801. DOI: 10.1080/01621459.1971.10482346.

[114] Scaman, K. and Virmaux, A. “Lipschitz Regularity of Deep Neural Networks: Analysis and Efficient Estimation”. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Montréal, Canada: Curran Associates Inc., 2018, pp. 3839–3848.

[115] Schapire, R. E. “The Strength of Weak Learnability”. *Machine Learning* 5.2 (June 1, 1990), pp. 197–227. DOI: 10.1007/bf00116037.

[116] Schapire, R. E. and Singer, Y. “Improved Boosting Algorithms Using Confidence-Rated Predictions”. *Machine Learning* 37.3 (1999), pp. 297–336. DOI: 10.1023/a:1007614523901.

[117] Seeger, A. “Direct and Inverse Addition in Convex Analysis and Applications”. *Journal of Mathematical Analysis and Applications* 148.2 (May 15, 1990), pp. 317–349. DOI: 10.1016/0022-247x(90)90004-y.

[118] Shafer, G. and Vovk, V. “The Sources of Kolmogorov’s Grundbegriffe”. *Statistical Science* 21.1 (Feb. 2006), pp. 70–98. DOI: 10.1214/088342305000000467.

[119] ——, *The Origins and Legacy of Kolmogorov’s Grundbegriffe*. 2018.

[120] Shafieezadeh-Abadeh, S., Kuhn, D., and Esfahani, P. M. “Regularization via Mass Transportation”. *Journal of Machine Learning Research* 20.103 (2019), pp. 1–68.

[121] Shaham, U., Yamada, Y., and Negahban, S. “Understanding Adversarial Training: Increasing Local Stability of Supervised Models through Robust Optimization”. *Neurocomputing* 307 (Sept. 2018), pp. 195–204. DOI: 10.1016/j.neucom.2018.04.027.

[122] Shveidel, A. P. "Recession Cones of Star-Shaped and Co-Star-Shaped Sets". *Optimization and Related Topics*. Boston, MA, USA: Springer, 2001, pp. 403–414. DOI: 10.1007/978-1-4757-6099-6_19.

[123] Sinha, A., Namkoong, H., and Duchi, J. "Certifiable Distributional Robustness with Principled Adversarial Training". *International Conference on Learning Representations*. 2018.

[124] Staib, M. and Jegelka, S. "Distributionally Robust Deep Learning as a Generalization of Adversarial Training". *31st Conference on Neural Information Processing Systems*. Long Beach, CA, USA, 2017.

[125] Steinwart, I., Pasin, C., Williamson, R. C., and Zhang, S. "Elicitation and Identification of Properties". *Proceedings of The 27th Conference on Learning Theory*. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, May 29, 2014, pp. 482–526.

[126] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. "Intriguing Properties of Neural Networks". *International conference on learning representations* (2014).

[127] Tewari, A. and Bartlett, P. L. "On the Consistency of Multiclass Classification Methods". *Learning Theory*. Vol. 3559. Berlin, Germany: Springer, 2005, pp. 143–157. DOI: 10.1007/11503415_10.

[128] Toland, J. F. "A Duality Principle for Non-Convex Optimisation and the Calculus of Variations". *Archive for Rational Mechanics and Analysis* 71.1 (May 1979), pp. 41–61. DOI: 10.1007/bf00250669.

[129] Tolstikhin, I. O., Gelly, S., Bousquet, O., Simon-Gabriel, C.-J., and Schölkopf, B. "AdaGAN: Boosting Generative Models". *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 5424–5433.

[130] Tong Zhang. "Sequential Greedy Approximation for Certain Convex Optimization Problems". *IEEE Transactions on Information Theory* 49.3 (Mar. 2003), pp. 682–691. DOI: 10.1109/tit.2002.808136.

[131] Tsuzuku, Y., Sato, I., and Sugiyama, M. "Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks". *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 6541–6550.

[132] Udell, M. and Boyd, S. "Bounding Duality Gap for Separable Problems with Linear Constraints". *Computational Optimization and Applications* 64.2 (June 1, 2016), pp. 355–378. DOI: 10.1007/s10589-015-9819-4.

[133] Vapnik, V. N. *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2000.

[134] Vidakovic, B. "Γ-Minimax: A Paradigm for Conservative Robust Bayesians". *Robust Bayesian Analysis*. Vol. 152. New York, NY, USA: Springer, 2000, pp. 241–259. DOI: 10.1007/978-1-4612-1306-2_13.

[135] Villani, C. *Optimal Transport: Old and New*. Grundlehren Der Mathematischen Wissenschaften 338. Berlin, Germany: Springer-Verlag, 2009.

[136] Volle, M. "On the Subdifferential of an Upper Envelope of Convex Functions". *Acta Mathematica Vietnamica* 19.2 (1994), pp. 137–148.

[137] Ward, D. "Chain Rules for Nonsmooth Functions". *Journal of Mathematical Analysis and Applications* 158.2 (1991), pp. 519–538. DOI: 10.1016/0022-247x(91)90254-w.

[138] Weidner, P. "Extended Real-Valued Functions—a Unified Approach". *Investigación Operacional* 39.3 (2018), pp. 303–325.

[139] Whitney, H. "Analytic Extensions of Differentiable Functions Defined in Closed Sets". *Transactions of the American Mathematical Society* 36.1 (1934), pp. 63–89. DOI: 10.1090/s0002-9947-1934-1501735-3.

[140] Williamson, R. C. "The Geometry of Losses". *Proceedings of the 27th Conference on Learning Theory*. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, May 29, 2014, pp. 1078–1108.

[141] Williamson, R. C., Vernet, E., and Reid, M. D. "Composite Multiclass Losses". *Journal of Machine Learning Research* 17.222 (2016), pp. 1–52.

[142] Yoshida, Y. and Miyato, T. "Spectral Norm Regularization for Improving the Generalizability of Deep Learning" (May 31, 2017). arXiv: 1705.10941 [cs, stat].

[143] Yost, D. "Asplund Spaces for Beginners". *Acta Universitatis Carolinae. Mathematica et Physica* (1993), pp. 159–177.

[144] Zaffaroni, A. "Convex Coradiant Sets with a Continuous Concave Cogauge". *Journal of Convex Analysis* 15.2 (2008), pp. 325–343.

[145] ——, "Convex Radiant Costarshaped Sets and the Least Sublinear Gauge". *Journal of Convex Analysis* 20.2 (2013), pp. 307–328.

[146] Zălinescu, C. "On an Abstract Control Problem". *Numerical Functional Analysis and Optimization* 2.6 (1980), pp. 531–542. DOI: 10.1080/01630568008816074.

[147] ——, "Stability for a Class of Nonlinear Optimization Problems and Applications". *Nonsmooth Optimization and Related Topics*. Boston, MA, USA: Springer, 1989, pp. 437–458. DOI: 10.1007/978-1-4757-6019-4_26.

[148] ——, "Recession Cones and Asymptotically Compact Sets". *Journal of Optimization Theory and Applications* 77.1 (Apr. 1, 1993), pp. 209–220. DOI: 10.1007/bf00940787.

[149] ——, *Convex Analysis in General Vector Spaces*. River Edge, NJ, USA: World Scientific, 2002.

[150] ——, "Duality Results Involving Functions Associated to Nonempty Subsets of Locally Convex Spaces". *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A, Matemáticas* 103 (Sept. 1, 2009), pp. 219–234. DOI: 10.1007/bf03191905.