Designing Consistent and Convex Surrogates for General Prediction Tasks

by

Jessica Finocchiaro

B.S., Florida Southern College, 2017

M.S., University of Colorado Boulder, 2020

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2022

Committee Members:

Rafael Frongillo, Chair

Bo Waggoner

Claire Monteleoni

Stephen Becker

Robert Williamson

Finocchiaro, Jessica (Ph.D., Computer Science)

Designing Consistent and Convex Surrogates for General Prediction Tasks

Thesis directed by Dr. Rafael Frongillo

Supervised machine learning algorithms are often predicated on the minimization of loss functions which measure error of a given prediction against a ground truth label. The choice of loss function to minimize corresponds to a summary statistic of the underlying data distribution that is learned in this process. Historically, loss function design has often been ad-hoc, and often results in losses that are not actually statistically *consistent* with respect to the target prediction task. This work focuses on the design of losses that are simultaneously convex, consistent with respect to a target prediction task, and efficient in the dimension of the prediction space. We provide frameworks to construct such losses in both discrete prediction and continuous estimation settings, as well as tools to lower bound the prediction dimension for certain classes of consistent convex losses. We apply our results throughout to understand prediction tasks such as high-confidence classification, top-k prediction, variance estimation, conditional value at risk, and ratios of expectations.

Dedication

To Stephen, Madison, Leonardo, and Anthony.

Acknowledgements

I would absolutely not be the pseron I am today and in a position to defend my dissertation without the support of so many people. First, I would like to thank Raf for taking a chance on me five years ago and taking me as a student. His feedback and advice was always honest, and always what I needed to hear, even if it wasn't what I wanted to hear in the moment. I've learned about myself as a person and the type of research mentor I want to be from working with him.

I have been fortunate to not only have the academic support of Raf, but also the support of my committee members, Bo, Claire, Stephen, and Bob, as well as our wonderful department staff, including Rajshree and Chantel. I would also be remiss if I forgot to mention the work in this dissertation benefitted from fedback from Arpit Agarwal, Krisztina Dearborn, Peter Bartlett, Nishant Mehta, Adam Blonairz, and Hari Ramaswamy. Much of the academic freedom and time to pursue this work was afforded to me by the NSF graduate research felolwship, for which I am very grateful as well.

I also have had the unique opportunity to collaboration and friendships I've begun with folks in Mechanism Design for Social Good, BEAAMO, and my broader research community: Rediet, Kira, Lily, Faidra, Manish, Duncan, Logan, Edwin, Ana, Keziah, Jakob, Ali, Deni, Chara, Alex, Marc, Roland, Gourab, and Stratis, among many others.

At CU, I've made and commiserated with so many from the CS department and graduate student association, club soccer, CAPS, and a D&D campaign that started from a machine learning group project: Rachel, Kara, Alyssia, Stephen, Maggie, Taisa, Angela, Christine, Emily, Adam, Jack, Juraj, Chris, Vinitha, Carli, Kristen, Annika, Diana, and Shawn.

Moving to Colorado from Florida, I left many friends, who, despite the distance, have been some of my grounding anchors over the past few years: Maria, Abby, Heather, Becca, Natalie, Lauren, and Austin. I suppose I owe my Florida-turned-Colorado adventure buddy, Ace, a thanks as well: he has made this past year full of adventures and chaos, he has always been there for me and for encouraged me to push past what I thought my limits were as a runner and a friend.

Finally, I cannot express how much the love and support from my family has meant over the years. I love you all more than words can say; regardless, I will try. First, I need to thank Madison for everything this past year. In a difficult year, she has been a kick ass mother, yet still the most loving and hospitable sister-in-law I could ask for. She, Leo, and Tony have never failed to bring a smile to my face, and I am a better person and aunt for this past year of living with them. My mom, dad, and brother have been there for me since day one. I am the person that I am today because of these three, and I'm truly at a loss for words to express my thanks. And of course, my extended family is too large to name everyone else here, but I'm grateful for all of you.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Decisions are made every day that make forecasts about future events: should someone in Boulder wear a coat on their commute to work? Is the answer to this multiple choice question 'C'? What web pages should we return for the search query "How to write a dissertation"? How tall will a 10 month old child grow to be? Only rarely are these decisions made with absolute certainty. Machine learning algorithms are often designed to make predictions answering these questions based on some labeled examples that "teach" the algorithm about different patterns that help answer the question at hand. As we design algorithms that operate under with the vast amounts of uncertainty in their decisions and recommendations, it is becoming increasingly important to think critically about the predictions we make.

In supervised machine learning, we teach machine learning algorithms by feeding them a large set of labeled training examples. The algorithm then makes predictions about unlabeled future inputs by using some hypothesis function learned by minimizing a *loss function*, or error punishment, over the training set. If everything else works out perfectly, minimizing this loss should teach us about some summary statistic of the data distribution. While some summary statistics are well-studied, such as binary classification and least-squares regressions, others have emerged such as variance estimation, high-confidence classifiers, top-k classification, and others, with no obvious method for constructing a corresponding loss function for the statistic.

This dissertation moves towards a general framework for being able to design a "nice" loss function that appropriately corresponds to a given prediction task. Particularly, we study loss functions satisfying three desiderata: they should be convex, consistent, and efficient. In Algorithms 101, we aim construct algorithms that are simultaneously correct and efficient. This same idea applies, where correctness corresponds to consistency (§ 2.3), and efficiency comes from the notion of prediction dimension (§ 2.4). Throughout this dissertation, we typically hold convexity and consistency as firm requirements, and at times construct lower bounds on the efficiency of a loss.

1.1 Motivating examples

1.1.1 Leo's morning commute

Suppose our protagonist Leo wants bring an umbrella to work if it rains that afternoon, but does not want the extra burden of carrying the umbrella if there is no rain. As a Computer Scientist (and therefore not a meteorologist), Leo does not know much about the probability of rain today. Thankfully, Leo does know a meteorologist, Madison, who happens to be an expert in forecasting rain probabilities. When he asks what the probability of rain is for tomorrow, however, Madison may not want to be honest with him.[1] Because of this, Leo decides to pay Madison with respect to her accuracy. If she forecasts rain chance $p\%$ and it rains, Leo decides he will pay her $\$p$ dollars, and if there is no rain, he will pay her $\$(100 - p)$. Under this scoring rule, we will see that Madison is actually incentivized to exaggerate her beliefs, so Leo will want to reward her some other way.

These types of forecasting problems have been studied since the early 1950s, dating back to Brier [15]. Brier's intuition was that experts on a topic, here: the weather, have some probabilistic belief about the outcome of a future event. Moreover, Leo can incentivize truthful reporting about the his belief of the expected future outcome, rain or no rain, if Madison's reward for accuracy is maximized by her prediction matching her belief in some sense. Concretely, if Madison believes there is a 70% chance of rain tomorrow, Leo wants to score her so that she maximizes her *expected score* by predicting a 70% chance of rain— not by exaggerating and saying 100% of rain. While we talk about maximizing scores here, one can equivalently consider minimizing losses, as shown

[1] While this scene is enacted by my nibling and sister-in-law, we deviate from reality here, as Madison is a very honest person.

in Figures 1.1 and 1.2. Indeed, throughout this dissertation, we will talk about minimizing loss functions, as this is the norm in machine learning.

1.1.2 Banking

We use *properties*, or summary statistics, as a tool to understand loss function design for a variety of tasks in machine learning. In some contexts, the it may not be feasible to assume an algorithm or person knows the full distribution over outcomes, but they may at least be able to give a good estimate of the summary statistic we are asking about.

For example, a regulating agency might want to learn a bank's financial risk in their investment portfolio. For variety of reasons (finite sample size, granularity, etc.), it is often too expensive for the bank to learn their entire risk distribution on their investments. Suppose instead that the while the bank cannot plausibly estimate their entire risk distrubtion, they can estimate some summary statistics of the distribution, such as variance or conditional value at risk. While the regulating agency may know what measure of estimate of financial risk they want to learn from banks, it is important for banks to be honest, as manipulating their report may enable them to take on excessive risk in their investments. As an auditing tool, the regulating agency decides they want to ask banks for risk, and score them after observing financial outcomes at the end of the month. However, given the summary statistic the regulator seeks to learn (conditional value at risk, variance, etc.), it is not immediately clear how they should score banks for their reports. Thus, it is important to have a framework that starts with a desired statistic to learn, and moves towards a scoring rule or loss function that is consistent with respect to this given statistic.

1.1.3 To predict or not to predict?

Machine learning algorithms are used to make all sorts of predictions: some more high-stakes than the other. While it might not be a huge deal if you lose $1 betting on a soccer game based on an algorithm's recommendation, it would be very costly if someone was denied or delayed medical attention because an algorithm's recommendation did not detect any anomalies in their records.

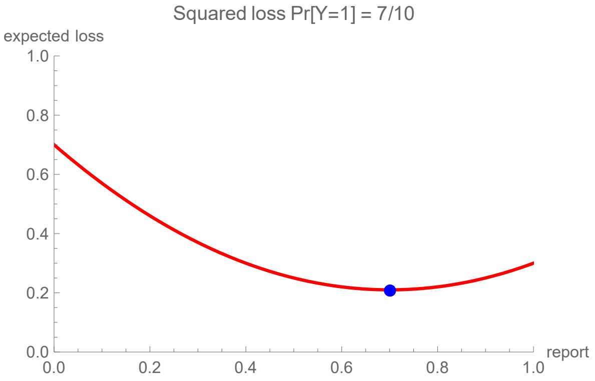


Figure 1.1: Squared loss

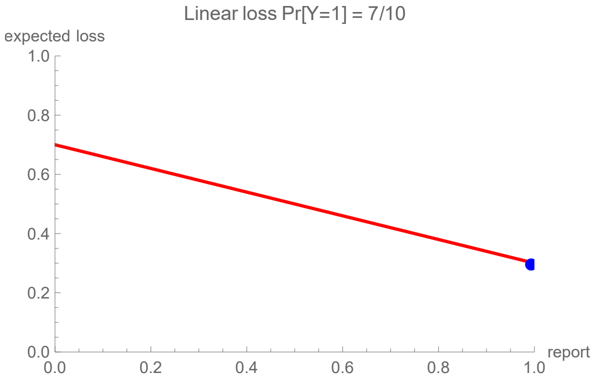


Figure 1.2: Linear loss

Figure 1.3: Minimizing two different expected losses when $\mathbb{P}[Y = 1] = \frac{7}{10}$ on report set $\mathcal{R} = [0, 1]$ and outcome set $\mathcal{Y} = \{0, 1\}$. The blue point is $\arg\min_{r\in\mathcal{R}} \mathbb{E}_{Y\sim p} L(r, Y)$. Squared loss in Figure 1.1 elicits the expected value, while the linear loss in Figure 1.2 elicits the mode. In this model, we are free to choose the report r on the horizontal axis, and ideally want to pick r that minimizes the loss in expectation, regardless of what the distribution is.

Given the well-known racial disparities in medical data, it would be unsurprising for an algorithm detecting cancerous tumors on someone's skin to have high uncertainty about the presence of a tumor [5].

Given a lack of proportional positive (tumor-present) examples on non-white skin tones, suppose our algorithm has 49% confidence that a given scan depicts a cancerous tumor given an image x.

Risk scores Suppose the patient's doctor was given a risk score (49% cancer) for this particular scan to contain a tumor. An overwhelmed, overbooked doctor might not know how to properly interpret such a risk score, and might be too time-constrained to invest the time to look into each scan carefully, so they determine they need some assistance flagging examples that require their attention, but are happy to let the computer filter out some examples that are clearly not tumors.

Binary classifier Since it is most likely the patient does not have a tumor, a thresholded binary classifier would return "no tumor," leading to delays and inaccess to medical care for this patient. This lack of care might quite literally be a matter of life and death, and might have cost the patient more money when more drastic and invasive care is needed later [53].

Learning to defer Suppose instead, the algorithm does not always make decisions. Instead, it only flags the doctor to review cases the algorithm is uncertain about, and classifies either positively or negatively on instances with pretty high certainty. This might lead to more manual review of one group's scans over another's, but yields a tradeoff that allows the doctor more time to focus on other work, while still flagging them to review cases where their expertise can supplement an algorithm.

While learning to defer may optimistically might improve performance of human-assisted algorithms for underserved groups such as Black folks, this is not guaranteed to be the case. Learning to abstain is not necessarily the correct way to design a melanoma-detection algorithm, but this example simply points to the different outcomes that can be attained in algorithm-assisted decision-making. However, we pose this example as an invitation to think critically about the prediction task at hand, and how its design might benefit or harm the end users or subjects of these algorithms.

1.1.4 Impact of loss function choice on predictions

Least-squares regression In least squares regression, we wish to estimate a continuous variable $y \in \mathcal{Y} = \mathbb{R}$, and have our sample of labeled examples $\{(x_i, y_i)\}_{i=1}^m$. As an example, suppose x_i represents an infant's birth date (relative to due date), and y_i their height at birth. In Figure 1.4, we find the linear regressor that minimizes empirical loss on the dataset shown. Typically, in Least Squares Regression, we minimize squared loss $L^{sq}(r, y) = (r - y)^2$, yielding the regressor shown by the blue dotted line. Fixing any input $x \in \mathbb{R}$, there is a conditional distribution over $\mathcal{Y}$ on what the outcomes might be, so what are we supposed to estimate in one real-valued prediction? Since we are minimizing squared loss, let us reason quickly about the minimizer(s) of expected squared loss[2] .

$$
\begin{aligned}
L(u, y) &= (u, y)^2 \\
\mathbb{E}_p L(u, Y) &= u^2 - 2u\mathbb{E}_p[Y] + \mathbb{E}_p[Y^2] \\
r \in \arg\min_u \mathbb{E}_p L(u, Y) \iff \frac{d}{du}|_{u=r} \mathbb{E}_p L(u, Y) &= 0 \\
\frac{d}{du} \mathbb{E}_p L(u, Y) &= \frac{d}{du}\left(u^2 - 2u\mathbb{E}_p[Y] + \mathbb{E}_p[Y^2]\right) \\
&= 2u - 2\mathbb{E}_p[Y] \\
\implies \frac{d}{du}|_{u=r} \mathbb{E}_p L(u, Y) = 0 \iff r &= \mathbb{E}_p[Y]
\end{aligned}
$$

Therefore, our least squares regression teaches us to predict the expected value on the conditional distribution, $\mathbb{E}_p[Y]$, where $p := \mathbb{P}_D[Y \mid X = \hat{x}]$ is the conditional distribution on a given input $\hat{x}$ and D is the data distribution over $\mathcal{X} \times \mathcal{Y}$.

Quartile Regression As we just saw above, fitting a dataset with least squares regression teaches us to predict the expected value over the dataset. However, least-squares is only one of many kinds of regression. Another commonly applied regression is α-quartile regression, in which one wants their regressor $f_L(\hat{x}) \geq \hat{y}$ with probability α. In particular, when $\alpha = 1/2$, we want to learn the median of the conditional data distribution, and can do so by minimizing absolute loss on the dataset. Figure 1.4 shows this regressor in the dashed red line, and it is important to observe

[2] For this derivation, let us assume that p has a finite second moment.

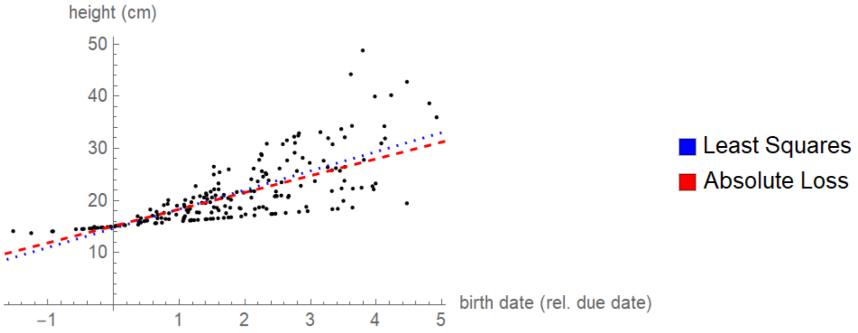


Figure 1.4: Fitting a linear model to estimate $\hat{y} = f_L(\hat{x}) = m_L\hat{x} + b_L$ on new input $\hat{x}$ by minimizing least squares and absolute losses yield different regressors on this dataset; that is, we observe different m_L and b_L depending on the loss L.

that this regressor is less skewed by outliers than the regressor learned from squared loss.

1.2 Contributions of this dissertation

There are four papers which lay the foundation of this dissertation [26, 27, 29, 30]. All four move towards understanding and characterizing convex, consistent surrogates in various settings, introduced later in Table 2.1.

An Embedding Framework for Consistent Polyhedral Surrogates Chapter 3 focuses on discrete prediction tasks for which a target loss matrix is given (Quadrant 1 in Table 2.1), and construct a polyhedral (piecewise linear and convex) surrogate that is consistent with respect to the target loss. We introduce the notion of a polyhedral surrogate *embedding* the loss matrix, and show this notion of embedding implies consistency.

We say that a surrogate embeds a target loss if there is some injection mapping target reports into $\mathbb{R}^d$ such that the target loss for the report and surrogate of its embedding match for all reports and outcomes and a report minimizes the target loss (in expectation) if and only if its embedding minimizes the surrogate. Perhaps surprisingly, we find that this embedding condition is equivalent to matching Bayes risks of the surrogate and target losses (Proposition 5).

In making the case for restricting to polyhedral surrogates, we find that every discrete target loss can be embedded by a polyhedral surrogate and every polyhedral surrogate embeds a target loss (Theorem 1).

It is not obvious that embedding is a sufficient condition for constructing a consistent surrogate. In order for a surrogate to be consistent with respect to a target loss matrix, one needs a link mapping surrogate reports back to target reports so that the surrogate and link pair is consistent. Chapter 3 gives a link construction and proves that it yields a consistent link and surrogate pair, implying that embedding is a sufficient condition for consistency, and even equivalent when restricting to polyhedral surrogates.

Convex Elicitation of Continuous Properties Chapter 4 focuses on continuous estimation tasks properties (Quadrant 4 of Table 2.1) with a finite outcome set. In particular, we focus on identifiable properties — those whose level sets can be described by flats (affine subspaces).

As these properties are identifiable, we know they are elicitable, but ask the question of when they are elicitable *via a convex loss*. It turns out the answer is quite simple: in this case, we show a property is elicitable $\iff$ it is convex elicitable.

For intuition, Steinwart, Pasin, Williamson, and Zhang [83, Theorem 5] show that identifiable properties can be elicited by a loss function comparing reports r to outcomes y of the form

$$L(r, y) := \int_0^r \lambda(x) V(x, y) dx \ ,$$

where V is the function which identifies the property of interest.

The proof relies on constructing the function λ so that $\lambda(r)V(r, y)$ is monotonically increasing for all $y \in \mathcal{Y}$, and therefore the constructed surrogate is convex and elicits the given property. We can do this by constructing a bound using the "most decreasing" and "least increasing" identification functions $V(\cdot, y)$ pointwise for $r \in \mathcal{R}$. Any λ that takes all of its values in this bound will yield a convex surrogate, so for simplicity we propose taking the midpoint of this bound for all $r \in \mathcal{R}$.

Embedding Dimension of Polyhedral Surrogate Losses Chapter 5 further investigates the embedding framework and proposes embedding dimension: a notion of efficiency for polyhedral embeddings. Notably, in one dimension (e.g., $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$), embedding and indirect elicitation are equivalent.

Chapter 5 introduces lower bounds on embedding dimension by reasoning about necessary conditions for polyhedral embeddings: optimality and monotonicity. The chapter focuses on bounds attained by considering optimality conditions, and show the bounds given by optimality conditions are equivalent to a quadratic feasibility program. While this is slightly discouraging (as quadratic feasibility programs are often computationally expensive to solve), the necessity for obtaining embedding dimension bounds via optimality suggests that we can do no better by studying optimality. Tighter bounds might be better found via monotonicity, left for future work.

The quadratic feasibility program given in Chapter 5 yields new bounds on embedding dimension of the multiclass abstain loss of [71, 72]. The current “best” surrogate (in terms of prediction dimension) that is calibrated for this task [71] is an embedding, so if there is a gap between embedding and convex calibration dimension, this suggests that we need new techniques of constructing such surrogates.

Unifying Bounds on Prediction Dimension for Consistent Convex Surrogates Chapter 6 uses heavily the fact that indirect elicitation is necessary for consistency, and proceeds to derive lower bounds on convex consistency dimension via property elicitation.

In this chapter, we observe that any level set of a convex elicitable property must be the union of some flats. This allows us to evaluate the convex consistency dimension (another notion of efficiency) of a target statistic by understanding the highest-dimension flat we can contain in a level set of the statistic which also contains a desired distribution over the outcomes. These bounds are presented in a similar format to those of Ramaswamy and Agarwal [71] in Quadrant 1, though our bounds that are complementary to theirs. In essence, fitting this flat through the level set allows us to observe lower bounds by leveraging the global geometry of the property, while the feasible subspace dimension bounds from Ramaswamy and Agarwal [71] focus on the local structure of the property around p.

Chapter 2

Setting and desiderata of losses

2.1 Setting

Notation Throughout, we let $\mathcal{R}$ denote a report set, and $\mathcal{Y}$ an outcome set of size $n := |\mathcal{Y}|$. Often, n is finite, but this is not always the case.

We let D over $\mathcal{X} \times \mathcal{Y}$ be a joint probability distribution over (possibly vector-valued) random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, and p be a probability distribution over $\Delta_{\mathcal{Y}} := \{p \mid p \text{ is a probability measure on } \mathcal{Y}\}$. If $\mathcal{Y}$ is finite, this is equivalent to $\Delta_{\mathcal{Y}} := \{p \in \mathbb{R}^n_+ \mid \langle \mathbb{1}, p \rangle = 1\}$, where $\mathbb{1}$ is the n-dimensional all-ones vector. Consider $p_y = \mathbb{P}[Y = y]$ to be the probability of outcome y on the distribution over $\mathcal{Y}$. We denote by $\mathcal{P}$ some subset of $\Delta_{\mathcal{Y}}$, and is therefore a set of probability distributions. $\text{relint}(S)$ is the relative interior of the set S; this is most often used to discuss $\text{relint}(\Delta_{\mathcal{Y}}) = \{p \in \Delta_{\mathcal{Y}} : \min_i p_i > 0\}$, only well-defined for finite $\mathcal{Y}$.

We denote a loss function $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ for generic (measurable) loss functions, and use $\ell : \mathcal{R}' \times \mathcal{Y} \to \mathbb{R}$; such a loss ℓ is called a *target loss*. When $\mathcal{Y}$ is finite, we interchangeably write $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ as $[L(r, y)]_{y \in \mathcal{Y}}$ as a vector of loss values over each outcome $y \in \mathcal{Y}$. We also call the loss L *convex* if $L(\cdot, y)$ is convex in its first argument for all $y \in \mathcal{Y}$. Similarly, we call a loss L *polyhedral* if $L(\cdot, y)$ is polyhedral (piecewise linear and convex) in its first argument for all $y \in \mathcal{Y}$. Moreover, we often consider the expected loss $\mathbb{E}_{Y \sim p} L(u, Y)$, and will use the shorthand $\mathbb{E}_p L(u, Y)$ or even $L(u; p)$ for this term. At times, it is mathematically more convenient to consider $\mathbb{E}_p L(u, Y) = \langle L(u), p \rangle$ when $\mathcal{Y}$ is finite. We also discuss the *Bayes risk* of a loss $\underline{L} : \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ such that $\underline{L} : p \mapsto \inf_{u \in \mathcal{R}} \mathbb{E}_p L(u, Y)$.

We write $\mathcal{L}_d$ for the set of (Borel) $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{Y}$-measurable and lower semi-continuous surrogates $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ such that $\mathbb{E}_{Y \sim p} L(u, Y) < \infty$ for all $u \in \mathbb{R}^d, p \in \mathcal{P}$, that are minimizable in that $\arg\min_u \mathbb{E}_p L(u, Y)$ is nonempty for all $p \in \mathcal{P}$. Moreover, $\mathcal{L}_d^{\text{cvx}} \subseteq \mathcal{L}_d$ is the set of convex (in $\mathbb{R}^d$ for every $y \in \mathcal{Y}$) losses in $\mathcal{L}_d$. Set $\mathcal{L} = \cup_{d \in \mathbb{N}} \mathcal{L}_d$, and $\mathcal{L}^{\text{cvx}} = \cup_{d \in \mathbb{N}} \mathcal{L}_d^{\text{cvx}}$. The assumption of minimizability is implicit in previous work, e.g., [6].

The "four quadrants" of problem types In this dissertation, we study *surrogate* loss functions, which are used to solve a related, but not identical, "target" problem of interest. Selecting a hypothesis by minimizing surrogate risk is one of the most widespread techniques in supervised machine learning. There are two main reasons why a surrogate loss is necessary: (I) the target problem is to minimize a loss, the *target loss*, that does not satisfy some desiderata such as continuity or convexity; or (II) the target problem is to estimate some *target statistic* and some associated surrogate loss is required to do so, as in many continuous estimation problems.

Above, we discuss a significant divergence in previous frameworks: constructing a surrogate given a *target loss* versus a *target statistic.* In addition to the two possible targets, we may have one of two domains: a *discrete* (i.e. finite) target prediction space, like a classification problem, or a *continuous* one, like a regression or point estimation problem. We informally refer to the four resulting cases—target loss vs. target statistic, and discrete vs. continuous predictions—as the "four quadrants" of supervised learning problems, shown in Table 2.1.

Despite the ubiquity of surrogate losses, we lack general frameworks to design and analyze consistent surrogates. While machine learning often seeks to design surrogates for (discrete) target losses, which in turn elicit some (discrete) property, many surrogates sought in finance are for target statistics which are known to not be directly elicitable. We provide such a general framework by "translating" problems given a target loss (Quadrants 1 and 3 in Table 2.1) to the settings of Quadrants 2 and 4 in Table 2.1, respectively, in which one is given a target property. In § 2.3, we define and juxtapose three notions of "consistency" to a target problem and justify the sufficiency of studying all through through the lens of property elicitation. One advantage of this approach is that indirect property elicitation can be applied in settings spanning any of the four quadrants in

Table 2.1.

	Target loss	*Target statistic*
Discrete prediction	**Q1**, e.g. classification	**Q2**, e.g. hierarchical classification
Continuous estimation	**Q3**, e.g. least-squares regression	**Q4**, e.g. variance estimation

Table 2.1: The four quadrants of problem types, with an example for each.

The three desiderata: convexity, consistency, and efficiency This dissertation studies the existence and construction of *convex* losses that are *consistent* to a prediction task, and ideally are *efficient* in the process. Each of the desiderata are discussed below.

2.2 Convexity

This dissertation studies the consistency and efficiency tradeoffs of *convex* loss function design. A function $f : \mathbb{R}^d \to \mathbb{R}_+$ is *convex* if, for all $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \ ,$$

demonstrated in Figure 2.1. Moreover, a function f is concave if $-f$ is convex. Affine functions are simultaneously concave and convex. We say a loss function $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ is convex if $L(\cdot, y)$ is convex in $\mathbb{R}^d$ for all $y \in \mathcal{Y}$. Convex functions have only global optima, and in turn many standard optimization algorithms require convexity of the objective function so that once stopping conditions are met (e.g., gradient approaches $\vec{0}$), one can conclude the algorithm has approached a global optimum, and not just stuck at a local optima somewhere far away from the global.

Mathematically, convex functions a few characteristics we highlight here: they are continuous on $\mathbb{R}^d$, so long as d is finite [77, Corollary 10.1.1], and differentiable almost everywhere. We often leverage is the fact that all optima are global: mathematically, we write this as $\vec{0} \in \partial f(x) \iff x$ is an optimal report, where $\partial f(x)$ is the subdifferential of f at x. Without convexity, only the reverse

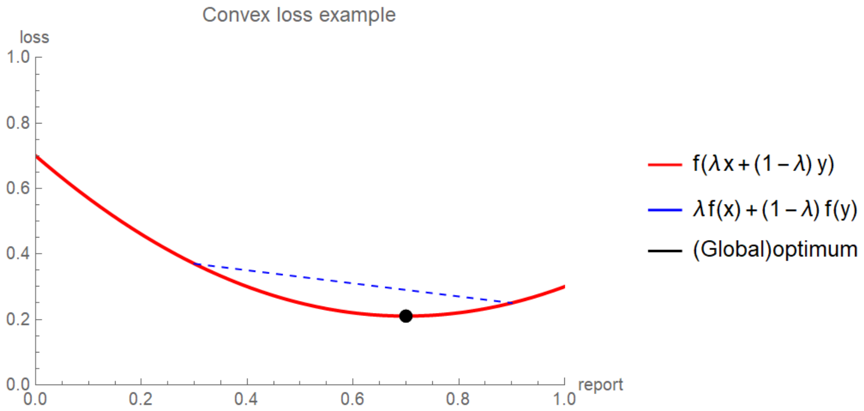


Figure 2.1: Example of a convex function. For any two points x, y, drawing a line connecting x and y does not go below the function on the range $[x, y]$.

implication holds.

2.3 Consistency for a target

In supervised machine learning, we undertake the task of designing losses that are minimized in expectation by answer the questions to which we want to know the answer. We assume that samples are drawn independently and identically distributed over some distribution D over the covariate and label space $\mathcal{X} \times \mathcal{Y}$. Now, supervised algorithms seek to learn a hypothesis $h^* : \mathcal{X} \to \mathcal{R}$ minimizing risk such that

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{(X,Y)\sim D} L(h(X), Y) \ . \tag{2.1}$$

In practice, one rarely knows the true distribution D over $\mathcal{X} \times \mathcal{Y}$, so instead it is common to minimize empirical risk, with the assumption that empirical data is drawn independently and identically distributed (i.i.d.) from D. As we tend toward infinite data samples, we then observe, with high probability, the empirical (sample-based) distribution tends towards the true distribution by the central limit theorem. Through *Empirical Risk Minimization* (ERM; eq. (2.2)), machine learning algorithms learn a hypothesis function to predict a something about the input given some description by minimizing a given loss function over the labeled training data, given

$$h^* \in \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i) \ . \tag{2.2}$$

While these supervised algorithms may not know the true joint distribution D over features $\mathcal{X}$ and labels $\mathcal{Y}$ underlying the data, we can "teach" it on some, say m, labeled training samples $\{(x_i, y_i)\}_{i=1}^m$. Here, x_i is some vector-representation of an instance, and y_i is the label associated with such instance, which approaches the true distribution D with high probability by the Central Limit Theorem.

While ERM of a convex surrogate can be done in polynomial time, target problems are often not in this form. Thus, given a target problem, we want a convex surrogate loss that is *consistent* with respect to the target problem, as this is a prerequisite for empirical risk bounds.

In this dissertation, we are primarily concerned with three notions of consistency for a task. First, we introduce (indirect) property elicitation, which yields formalism around the notion of "target statistics" mentioned above, as well as a special case of elicitation called embeddings. The second, statistical consistency, is what we actually desire in machine learning, but is practically often difficult to with with. The final notion, calibration, is a proxy for consistency that is often easier to work with than studying consistency directly. The literature has extensively studied the conditions under which calibration and consistency are equivalent [11, 71, 84, 96], though the tightness of the relationship between consistency and indirect elicitation is recently formalized [30].

2.3.1 Elicitation

For computational and practical reasons, it is often not plausible to design machine learning algorithms to estimate the entire outcome distribution, and frankly, this often yields more information than needed. Instead, we often try to learn some *summary statistic*, or property, of a data distribution. For intuition, one can think of $p := \mathbb{P}_D[Y \mid X = x]$ being the conditional distribution on outcomes $\mathcal{Y}$ for a given $x \in \mathcal{X}$, where $(X, Y) \sim D$.

Definition 1 (Property, level set)**.** *A property* $\Gamma : \mathcal{P} \to 2^{\mathcal{R}} \setminus \{\emptyset\}$ *is a (set-valued) function that maps probability distributions to reports. The level set* $\Gamma_r = \{p \in \mathcal{P} : r \in \Gamma(p)\}$ *is the set of distributions for which* r *is in the property set.*

We denote $\Gamma : \mathcal{P} \to 2^{\mathcal{R}} \setminus \{\emptyset\}$ by $\Gamma : \mathcal{P} \rightrightarrows \mathcal{R}$, and if $|\Gamma(p)| = 1$ for all $p \in \mathcal{P}$, then we call Γ a *single-valued* property. Moreover, a property is *finite* if $|\mathcal{R}| < \infty$, and we are in a finite-outcome setting if $|\mathcal{Y}| = n < \infty$. If a property is finite, it is assumed that we are in a finite-outcome setting. In general, we use the notation γ to denote a finite property, and Γ to denote a general property.

Definition 2 ((Directly) Elicits)**.** *A loss* $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ *elicits a property* $\Gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ *if*

$$\forall p \in \mathcal{P},\ \Gamma(p) = \arg\min_{u \in \mathcal{R}} \mathbb{E}_{Y \sim p} L(u, Y) \tag{2.3}$$

for $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$*. If such a loss elicits* Γ*, then we say* Γ *is (directly) elicitable.*

Any (minimizable) loss $L \in \mathcal{L}$ elicits *some* property; we denote this property $\Gamma := \mathrm{prop}_{\mathcal{P}}[L]$.

When we say a property is *elicitable*, we generally mean it is directly elicitable. The notion of direct elicitation might be too strict. For example, the variance is very plausibly a summary statistic one might want to learn, but it is not elicitable by any loss $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$. This leads us to discuss a more general notion of *indirect elicitation.* Indirect elicitation generalizes results required for the discussion in § 2.4 regarding *elicitation complexity.*

Definition 3 (Indirectly elicits)**.** *A surrogate loss* $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ *and link* $\psi : \mathbb{R}^d \to \mathcal{R}$ *pair* (L, ψ) *indirectly elicit a property* $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ *if* L *directly elicits a property* $\Gamma : \mathcal{P} \rightrightarrows \mathbb{R}^d$ *such that for all* $u \in \mathbb{R}^d$*, we have* $\Gamma_u \subseteq \gamma_{\psi(u)}$*. Moreover, we say* L *indirectly elicits* γ *if such a link exists.*

One example of an indirectly elicitable property that is not directly elicitable is $\Gamma : p \mapsto (\mathbb{E}_p[Y])^2$. One can simply elicit $\mathbb{E}_p[Y]$ by squared loss[1] and square the result, with $\psi : x \mapsto x^2$. Intuitively, the set of directly elicitable properties is a subset of indirectly elicitable properties, as we can take ψ to be the identity.

[1] Restricting $p \in \mathcal{P}$ to have a finite second moment.

2.3.2 Embeddings

Chapters 3 and 5 of this dissertation focus on a special case of indirect elicitation called *embeddings* for discrete prediction problems in Quadrants 1 and 2 of Table 2.1. When constructing a surrogate for a discrete target loss, there is often a natural way in which one "embeds" their discrete reports into $\mathbb{R}^d$. Perhaps we map all of the discrete reports into the real line according to some pre-determined order. This is very natural when $\mathcal{R} = \{1, 2, 3\}$, for example; however, when $\mathcal{R} = \{\text{red, green, blue}\}$, how to embed these reports is a lot more ambiguous.

The embeddings framework in Definition 5 formalizes this notion of moving from discrete, target prediction spaces into surrogate prediction spaces, and draws a close connection between polyhedral (piecewise linear and convex) surrogates and discrete target tasks, and show that polyhedral embeddings imply calibration, and therefore consistency, of the surrogate with respect to the target task.

Definition 4 (Representative set)**.** *Let $\Gamma : \mathcal{P} \rightrightarrows \mathcal{R}$. We say $\mathcal{S} \subseteq \mathcal{R}$ is a $\mathcal{P}$-representative set for* Γ *if, for all $p \in \mathcal{P}$, we have $\mathcal{S} \cap \text{prop}_{\mathcal{P}}[L] \neq \emptyset$. We further say $\mathcal{S}$ is a minimum representative set if it has the smallest cardinality among all representative sets. Given a minimizable loss $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, we say $\mathcal{S}$ is a (minimum) representative set for L if it is a (minimum) representative set for* $\text{prop}_{\mathcal{P}}[L]$*. If $\mathcal{P} = \Delta_{\mathcal{Y}}$, we simply say $\mathcal{S}$ is a representative set.*

Definition 5 (Embedding)**.** *A $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ such that $L \in \mathcal{L}$ embeds a loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ if there exists a representative set $\mathcal{S}$ for ℓ and an injective embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$ such that (i) for all $r \in \mathcal{S}$ we have $L(\varphi(r)) = \ell(r)$, and (ii) for all $p \in \Delta_{\mathcal{Y}}, r \in \mathcal{S}$ we have*

$$r \in \text{prop}_{\Delta_{\mathcal{Y}}}[\ell](p) \iff \varphi(r) \in \text{prop}_{\Delta_{\mathcal{Y}}}[L](p) \ . \tag{2.4}$$

If $\mathcal{S}$ is a minimal representative set, we say L tightly embeds ℓ.

We show later that a surrogate embeds a target loss if and only if their Bayes risks match, which implies that a polyhedral Bayes risk is necessary for embedding a discrete loss. Moreover, in

Chapter 3, given an embedding, we give a link construction so that (L, ψ) is calibrated with respect to ℓ, as in the setting of Definition 8.

2.3.3 Consistency: the gold standard

In the nicest of settings, theoretical guarantees of algorithm correctness are often understood through the requirement of being "probably approximately correct," or PAC learnable [81, Definition 3.1]. We often generalize PAC learnability to the weaker notion of nonuniform learnability (in order to cope with infinite VC dimension), and even weaker than this notion is consistency. For simplicity, we assume that our hypothesis class $\mathcal{H}$ is sufficiently rich to attain the Bayes risk, and call this the realizable setting. In this realizable setting, consistency allows one to derive guarantees that for a large enough sample, empirical loss of an ERM algorithm approaches the true expected risk with high probability, regardless of the underlying data distribution [81, Section 7.4].

Constructing consistent surrogates then is a minimum requirement if we want to say anything about PAC learnability or excess risk bounds more generally. Consistency at least gives us the weakest guarantees possible that with enough samples, minimizing a surrogate loss on the empirical sample set corresponds to minimizing the target loss or statistic.

As discussed above, notions of consistency have appeared in the literature with respect to target losses, and to target statistics or properties. First, given a target loss ℓ, we say L is consistent if optimizing L and applying a link ψ optimizes ℓ (Definition 6). Second, given a target property γ, such as the α-quantile, we say L is consistent if optimizing L implies approaching, in some sense, the correct statistic $\gamma(D_x)$ of the conditional distributions $D_x = \mathbb{P}[Y|X = x]$ (Definition 7). We then observe that Definition 6 is subsumed by Definition 7, and use this to show consistency implies L indirectly elicits $\text{prop}_{\mathcal{P}}[\ell]$ or γ respectively.

Condition 1 (Covers)**.** *A set $\mathcal{D} \subseteq \Delta(\mathcal{X} \times \mathcal{Y})$ covers a convex set $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ if, for all $p \in \mathcal{P}$, there exists $D \in \mathcal{D}$ and $x \in \mathcal{X}$ such that D has a point mass on x and $p = D_x$.*

Definition 6 (Consistent: loss)**.** *A loss $L \in \mathcal{L}$ and link (L, ψ) are $\mathcal{D}$-consistent for a set $\mathcal{D}$ of* *distributions over $\mathcal{X} \times \mathcal{Y}$ with respect to a target loss ℓ if, for all $D \in \mathcal{D}$ and all sequences of measurable hypothesis functions $\{f_m : \mathcal{X} \to \mathcal{R}\}$,*

$$\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y) \implies \mathbb{E}_D \ell((\psi \circ f_m)(X), Y) \to \inf_f \mathbb{E}_D \ell((\psi \circ f)(X), Y) \ .$$

For a given convex set $\mathcal{P} \subseteq \Delta_\mathcal{Y}$, we simply say (L, ψ) is consistent if it is $\mathcal{D}$-consistent for some $\mathcal{D}$ covering $\mathcal{P}$.

Instead of a target loss ℓ, one may want to learn a target property, i.e. a conditional statistic such as the expected value, variance, or entropy. In this case, following the tradition in the statistics literature on conditional estimation [25, 47, 78], we formalize consistency as converging to the correct conditional estimates of the property. Convergence is measured by functions $\mu(r, p)$ that formalize how close r is to "correct" for conditional distribution p. In particular we should have $\mu(r, p) = 0 \iff r \in \gamma(p)$.

Definition 7 (Consistent: property)**.** *Suppose we are given a loss $L \in \mathcal{L}$, link function $\psi : \mathbb{R}^d \to \mathcal{R}$, and property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$. Moreover, let $\mu : \mathcal{R} \times \mathcal{P} \to \mathbb{R}_+$ be any function satisfying $\mu(r, p) = 0 \iff r \in \gamma(p)$. We say (L, ψ) is $(\mu, \mathcal{D})$-consistent with respect to γ if, for all $D \in \mathcal{D}$ and sequences of measurable functions $\{f_m : \mathcal{X} \to \mathcal{R}\}$,*

$$\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y) \implies \mathbb{E}_X \mu(\psi \circ f_m(X), D_X) \to 0 \ . \tag{2.5}$$

We simply say (L, ψ) is μ-consistent if it is $(\mu, \mathcal{D})$-consistent for some $\mathcal{D}$ covering $\mathcal{P}$. Additionally, we say (L, ψ) is consistent if there is a μ such that (L, ψ) is μ-consistent.

Typical definitions of consistency require $\mathcal{D}$ to be the set of all distributions over $\mathcal{X} \times \mathcal{Y}$, so our condition of covering is much weaker.

Lemma 1 in § 2.3.5 shows that, in fact, one can capture consistency with respect to a target loss as a special case of consistency with respect to a target property. Specifically, given a target loss ℓ, one can take $\gamma = \mathrm{prop}_\mathcal{P}[\ell]$ and define $\mu(r, p) := \mathbb{E}_p \ell(r, Y) - \min_{r'} \mathbb{E}_p \ell(r', Y)$ to be the ℓ-regret of the report r. This observation allows us to translate consistency from Quadrant 1 to Quadrant 2, and from Quadrant 3 to Quadrant 4.

2.3.4 Calibration

In the machine learning literature, Bartlett et al. [11], Zhang [96] originally proposed the notion of *classification calibration*, which was later generalized for more prediction tasks to *calibration* [71, 82, 84], which is equivalent to consistency when one is given a target loss and wants to make discrete predictions, such as in Quadrants 1 and 3 of Table 2.1.

Bartlett, Jordan, and McAuliffe [11, Theorem 1] shows that a convex, margin-based surrogate and the link $\psi(u) = \operatorname{sgn}(u)$ are consistent with respect to a 0-1 loss if and only if they are calibrated with respect to 0-1 loss. Bartlett et al. proceed to give a full characterization of calibrated margin-based convex surrogates (e.g., $L(u, y) = f(uy)$): L is classification calibrated if and only if f is differentiable at 0 and $f'(0) < 0$ [11, Theorem 4].

Of course, this leaves a lot to be desired; what if we have a surrogate that isn't margin-based? What if we have a more general prediction task than binary classification? In the literature, calibration has two common definitions, which are equivalent in Quadrant 1 of Table 2.1, though one definition generalizes to Quadrant 3 [82, Chapter 3]. Studying calibration instead of consistency allows us to move from distributions D over $\mathcal{X} \times \mathcal{Y}$ to distributions p over $\mathcal{Y}$, simplifying the "instance space" we are considering. For intuition, one may think of $p = \mathbb{P}_D[Y \mid X = x]$ as the conditional distribution over labels given a data point $x \in \mathcal{X}$.

In calibration, any surrogate report that is not linked to the optimal target report has surrogate loss strictly greater than the Bayes risk of the surrogate.

Definition 8 (Calibrated: Quadrant 1)**.** *Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ be a discrete target loss eliciting γ. A surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ and link $\psi : \mathbb{R}^d \to \mathcal{R}$ pair (L, ψ) is $\mathcal{P}$-calibrated with respect to ℓ if*

$$\forall p \in \mathcal{P} : \inf_{u \in \mathbb{R}^d : \psi(u) \notin \gamma(p)} \mathbb{E}_p L(u, Y) > \inf_{u \in \mathbb{R}^d} \mathbb{E}_p L(u, Y) \ . \tag{2.6}$$

We simply say L is calibrated if $\mathcal{P} = \Delta_{\mathcal{Y}}$.

Many works characterize calibrated surrogates for specific discrete target losses [11, 62, 84, 96], including the canonical 0-1 loss for binary and multiclass classification. We give another definition of calibration which is a special case of calibration via Steinwart and Christmann [82], and show it is equivalent to Definition 8 in discrete prediction settings (Quadrant 1), but can be applied in continuous estimation settings as well. We use this more general definition of calibration when proving statements about the relationship between consistency, calibration, and indirect elicitation.

For a given $p \in \mathcal{P}$, the (conditional) *regret*, or excess risk, of a loss L is given by $R_L(u,p) := \mathbb{E}_p L(u,Y) - \inf_{u^*} \mathbb{E}_p L(u^*, Y)$.

Definition 9 (Calibrated: Quadrants 1 and 3)**.** *A loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ is $\mathcal{P}$-calibrated with respect to a loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ if there is a link $\psi : \mathbb{R}^d \to \mathcal{R}$ such that, for all distributions $p \in \mathcal{P}$, there exists a function $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$ with ζ continuous at 0^+ and $\zeta(0) = 0$ such that for all $u \in \mathbb{R}^d$, we have*

$$\ell(\psi(u); p) - \underline{\ell}(p) \leq \zeta\left(\mathbb{E}_p L(u,Y) - \underline{L}(p)\right) \ . \tag{2.7}$$

If $\mathcal{P} = \Delta_{\mathcal{Y}}$, we simply say (L, ψ) is calibrated.

Proposition 1. *When $\mathcal{R}$ and $\mathcal{Y}$ are finite, a continuous loss and link (L, ψ) are $\mathcal{P}$-calibrated with respect to a target loss ℓ via Definition 9 if and only if they are $\mathcal{P}$-calibrated via Definition 8.*

2.3.5 Relating calibration, consistency, and indirect elicitation.

Consistency implies indirect elicitation

In what follows, we show consistency implies indirect elicitation, which allows us to apply indirect elicitation to yield state-of-the-art lower bounds on convex consistency dimension, discussed in § 2.4.

Given a target loss ℓ, we can define a statistic γ as the property it elicits. Intuitively, consistency of a surrogate L with respect to ℓ and γ are equivalent, i.e. in both cases estimates should converge to values that minimize ℓ-loss. We formalize this by letting μ be the ℓ-regret, $R_\ell := \mathbb{E}_p \ell(r, Y) - \min_{r'} \mathbb{E}_p \ell(r', Y)$, yielding Lemma 1.

Lemma 1. *Let a convex $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ be given. Given a surrogate loss $L \in \mathcal{L}$, link ψ, and target loss ℓ, set $\mu(r,p) := \mathbb{E}_p \ell(r, Y) - \min_{r'} \mathbb{E}_p \ell(r', Y)$ as the excess risk of ℓ, R_ℓ. Then there is a $\mathcal{D}$ covering*

$\mathcal{P}$ such that (L, ψ) is $\mathcal{D}$-consistent with respect to ℓ if and only if (L, ψ) is $(\mu, \mathcal{D})$-consistent with respect to $\gamma := \mathrm{prop}_{\mathcal{P}}[\ell]$.

Proof. First, observe that $\mu(r, p) = 0 \iff \mathbb{E}_p \ell(r, Y) = \inf_{r' \in \mathcal{R}} \mathbb{E}_p \ell(r', Y) \iff r \in \gamma(p)$. Now suppose (L, ψ) are consistent with respect to ℓ, and take any sequence $\{f_m\}$ of measurable hypotheses. Rewriting the right-hand side of Definition 6,

$$\begin{aligned} &\mathbb{E}_D \ell(\psi \circ f_m(X), Y) \to \inf_f \mathbb{E}_D \ell(\psi \circ f(X), Y) && (2.8) \\ \iff& \mathbb{E}_X R_\ell(\psi \circ f_m(X), D_X) \to 0 \\ \iff& \mathbb{E}_X \mu(\psi \circ f_m(X), D_X) \to 0\ . && (2.9) \end{aligned}$$

Therefore, $\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y)$ implies (2.8) if and only if it implies (2.9). Observe that the assumptions on $\mathcal{L}$ allow us to apply the Fubini-Tonelli Theorem [36, Theorem 2.37], which yields the equivalence of eq. 2.8 to the next line. □

Because each target loss in $\mathcal{L}$ elicits some property, but not all target properties can be elicited by a loss (e.g. the variance), consistency with respect to a property is the strictly broader notion. In a loose sense, Proposition 1 lets us translate problems about target losses to be about the properties these losses elicit. This points to indirect elicitation as a natural necessary condition for consistency, as formalized in Proposition 2.

Proposition 2 ([30, Proposition 1])**.** *For a surrogate $L \in \mathcal{L}$, if the pair (L, ψ) is consistent with respect to a property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ or a loss ℓ eliciting γ, then (L, ψ) indirectly elicits γ.*

In other words, indirect elicitation is a necessary condition for consistency.

While the literature has historically used calibration as a proxy for calibration, this suggests that one can also use indirect property elicitation as a proxy for consistency. Moreover, indirect elicitation is generally more applicable than calibration, which allows us to consider problem settings across the four quadrants: most notably, when when one is given a target statistic rather than target loss. With these results in hand, we proceed through the rest of this dissertation studying indirect property elicitation as a proxy for consistency.

Consistency implies calibration (and is sometimes equivalent) Even with the more general notion of calibration that extends beyond discrete predictions, we still have consistency implying calibration. The close connection between indirect elicitation and consistency was first explored by Agarwal and Agarwal [6]. In particular, calibration of $L \in \mathcal{L}$ with respect to ℓ implies indirect elicitation quite directly: take $u \in \mathbb{R}^d$ and $p \in \Gamma_u$, implying $u \in \Gamma(p)$. By the definition of elicitation, $\mathbb{E}_p L(u, Y) = \inf_{u' \in \mathbb{R}^d} \mathbb{E}_p L(u', Y)$, so we must have $\psi(u) \in \gamma(p)$ from eq. (2.6), as desired.

Proposition 3 ([30, Proposition 4])**.** *If a loss and link (L, ψ) are consistent with respect to a loss ℓ, then they are calibrated with respect to ℓ.*

Moreover, we have calibration implying indirect elicitation.

Lemma 2 ([30, Lemma 6])**.** *If a surrogate and link (L, ψ) with $L \in \mathcal{L}$ are calibrated with respect to a loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$, then L indirectly elicits the property $\gamma := \mathrm{prop}_{\mathcal{P}}[\ell]$.*

Combining the two results, we can observe the result of Proposition 2 another way: *through calibration.*

2.4 Notions of efficiency

For a prediction task, there are infinitely many surrogate losses one might use to learn the task at hand. However, this dissertation focuses on convex surrogates as this yields better accuracy guarantees on the optimization problem itself [14, Chapter 9.1]. Recall that in ERM, we are minimizing the average loss, which is a constant (inverse of sample size) times the sum of convex functions, which in turn is convex: hence, ERM is also convex.

However, the class of convex, consistent surrogates for a prediction task may still be incredibly large. For example, exponential loss, hinge loss, logistic loss, and square loss are all convex and consistent (with the right link) surrogate losses for binary classification. This prompts us to ask: is one of these losses better than the others, in the sense that optimizing a better loss leads to more efficient learning algorithms.

This work studies a few related notion of efficiency, all of which fall under the umbrella category of *prediction dimension*.

Definition 10. *The prediction dimension of a given surrogate $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ is d.*

Since, in ERM on a loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$, we find the a hypothesis $h : \mathcal{X} \to \mathbb{R}^d$, the complexity of ERM is then a function of the prediction dimension d. Thus, we aim to get d as low as possible, without compromising on convexity or consistency, though studying efficiency-consistency tradeoffs poses an interesting area of future work.

Prediction dimension has a few metrics that are studied throughout. First, the notion that we actually care about is *convex consistency dimension.* We will define this formally, and the rest informally for now, though the intuition should follow.

Definition 11 (Convex consistency dimension)**.** *Given target loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ or property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$, its convex consistency dimension* $\mathrm{cons}_{\mathrm{cvx}}(\cdot)$ *is the minimum dimension d such that $\exists L \in \mathcal{L}_d^{\mathrm{cvx}}$ and link ψ such that (L, ψ) is consistent with respect to ℓ or γ.*

Other studied notions of efficiency include convex calibration dimension, which is equivalent to convex consistency dimension for tasks in Quadrant 1. We additionally study (convex) elicitation complexity [39], which is the minimum dimension d where there is a (convex) surrogate $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ and link ψ indirectly eliciting a target statistic γ.

Definition 12 (Convex elicitation complexity)**.** *Given a target property γ, the convex elicitation complexity* $\mathrm{elic}_{\mathrm{cvx}}(\gamma)$ *is the minimum dimension d such that there is a $L \in \mathcal{L}_d^{\mathrm{cvx}}$ indirectly eliciting γ.*

Finally, as one might guess, embedding dimension is the minimum surrogate dimension such that a surrogate embeds the target loss (or property).

These notions are all closely related, though not necessarily equal. For example, we know that embedding is a strict subset of consistency, since the logistic loss does not embed 0-1 loss, but is consistent with respect to 0-1. In this example, their dimensions are the same anyways since hinge (which, like logistic loss, has prediction dimension 1) does embed 0-1 loss, but it is unclear, and perhaps doubtful, that this holds in general.

Chapter 3

An embeddings framework for consistent polyhedral surrogates

Historically, loss function design is often ad-hoc, and often does not statistically correspond to the intended prediction task. We start our adventure into the land of convex and consistent surrogate loss design by first studying prediction problems where there are a finite number of predictions possible, e.g., Quadrants 1 and 3 in Table 2.1. In this setting, this chapter introduces a framework motivated by a particularly natural approach for finding convex surrogates, wherein one "embeds" a discrete loss into $\mathbb{R}^d$ and "convexifies" in between. Specifically, we say a convex surrogate L embeds a discrete target loss ℓ if there is an injective embedding function from the discrete reports to a (finite-dimensional) vector space such that (i) the original loss values are recovered, and (ii) a report is ℓ-optimal if and only if the embedded report is L-optimal. Common examples of this general construction include hinge loss as a surrogate for 0-1 loss and the abstain surrogate [72].

We prove that such an embedding scheme is intimately related to the class of polyhedral (piecewise-linear and convex) loss functions. In particular, every discrete loss is embedded by a polyhedral surrogate. Moreover, such an embedding gives rise to calibrated link function, and is therefore consistent with respect to the target loss.

Theorem 1. *Every discrete loss ℓ is embedded by some polyhedral loss L, and every polyhedral loss L embeds some discrete loss ℓ.*

Theorem 2. *Given any polyhedral loss L, let ℓ be a discrete loss it embeds. There exists a link function ψ such that (L, ψ) is calibrated with respect to ℓ.*

Beyond consistency, we show that any calibrated link gives rise to a linear surrogate regret bound, which allows one to translate generalization bounds from the surrogate to the target [42].

Our proofs give explicit constructions for the surrogate (§ 3.1) and link (§ 3.2) embedding a given discrete loss. Conversely, given an existing polyhedral surrogate, we provide tools to find the discrete losses they embed (Proposition 4), which may or may not be the desired target. In short, if one can identify a finite *representative set* $\mathcal{S}$ of reports for a surrogate L, meaning one that always contains an L-optimal report, then L embeds the loss $L|_{\mathcal{S}}$ (L restricted to $\mathcal{S}$). We illustrate all of these concepts with several examples (§ 3.3).

Underpinning our results are several observations which formalize the idea that polyhedral losses "behave like" discrete losses. For example, polyhedral losses have a finite number of optimal sets (the set of reports which minimize the expected loss for some conditional label distribution). As a result, by selecting a report from each set, one arrives at a finite representative set, which gives an embedding. For the converse, we prove that the conditions of an embedding are equivalent to matching Bayes risks (Proposition 5), and use the fact that discrete losses and polyhedral losses both have polyhedral Bayes risks.

We also provide several observations beyond what is needed to prove our main results, which we view as conceptual contributions (§ 3.4, 3.5). Using tools from property elicitation, we show an equivalence between minumum representative sets and "non-redundancy", wherein no report is dominated by another. We further show that, while the minimum representative set is not always unique, the loss values associated with it are unique, giving rise to a natural "trim" operation on losses. Finally, we show that when restricting to the class of polyhedral surrogates, indirect elicitation is both necessary and sufficient for consistency (Theorem 8).

Taken together, we the contributions of this chapter are both conceptual and practical. We uncover the remarkable structure of polyhedral surrogates, deepening our understanding of the relationship between surrogate and discrete target losses. This structure leads to a powerful new framework to design and analyze surrogate losses. As we illustrate with several examples, this framework has already been applied to solve open questions by designing new surrogates, to uncover the behavior of existing surrogates, and to construct link functions in complex structured problems.

These results are largely based on Finocchiaro et al. [27], accepted to NeurIPS 2019, with the journal version [28] currently in preparation.

Related works. The literature on convex surrogates focuses mainly on smooth surrogate losses [9–11, 21, 23, 65, 75, 89, 95]. In practice, minimizing such surrogates often corresponds to learning the entire underlying data distribution and compute the desired target task with the full distribution in hand. However, Ramaswamy et al. [72, Section 1.2] contend that optimizing nonsmooth losses may enable reduction of the prediction dimension reduction (relative to smooth losses) while maintaining consistency, improving downstream efficiency of the learning algorithm.

We study consistency through the indirect property elicitation, which one may recall is a necessary condition. Agarwal and Agarwal [6] were the first to formally connect property elicitation to consistency, though their results generally do not apply to discrete prediction tasks. The notion of embedding introduced in § 3.0.2 is a special case of indirect property elicitation. While property elicitation has an extensive literature by now [35, 38, 45, 55, 57, 66, 79, 83], these works are mostly concerned with point estimation problems, which is in direct contrast to polyhedral embeddings, whose structure yields a finite set of possible predictions.

3.0.1 Polyhedral losses

In this chapter, we focus on settings where $\mathcal{Y}$ is finite. Most of the surrogate losses we consider will be *polyhedral*, meaning piecewise linear and convex; we therefore briefly recall the relevant definitions. In $\mathbb{R}^d$, a *polyhedral set* or *polyhedron* is the intersection of a finite number of closed halfspaces. A *polytope* is a bounded polyhedral set. A convex function $f : \mathbb{R}^d \to \mathbb{R}$ is *polyhedral* if its epigraph is polyhedral, or equivalently, if it can be written as a pointwise maximum of a finite set of affine functions [77].

Definition 13 (Polyhedral loss)**.** *A loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ is polyhedral if $L(u)_y$ is a polyhedral convex function of u for each $y \in \mathcal{Y}$.*

For example, hinge loss is polyhedral, whereas logistic loss is not. As mentioned in Chapter 1, we focus on the construction of *consistent* polyhedral surrogates through the embeddings framework we now proceed to give. Since we are in Quadrant 1 of Table 2.1, we can evaluate consistency through the lens of calibration (Definition 8) and abstract away the feature space $\mathcal{X}$.

3.0.2 Embedding

We now formalize the sense in which a convex surrogate can *embed* a target loss ℓ. Here one maps each report (prediction) of ℓ to a point in $\mathbb{R}^d$, then constructs a convex loss on $\mathbb{R}^d$ that agrees with ℓ at these points. This approach captures several consistent surrogates in the literature (e.g., [59, 70, 71, 85]; see § 3.3).

An important subtlety is that it is not always necessary to map *all* target reports to $\mathbb{R}^d$. It is often convenient to allow ℓ to have reports that are "redundant" in some sense. (We explore redundancy further in § 3.4; see also Wang and Scott [85].) Because of this redundancy, we will only require an embedding map to be defined on a *representative set*: a set of reports $\mathcal{S}$ such that, for all label distributions, at least one report $r \in \mathcal{S}$ minimizes expected loss.

Definition 14 (Representative set)**.** *Let $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$. We say $\mathcal{S} \subseteq \mathcal{R}$ is representative for Γ if we have $\Gamma(p) \cap \mathcal{S} \neq \emptyset$ for all $p \in \Delta_{\mathcal{Y}}$. We further say $\mathcal{S}$ is a minimum representative set if it has the smallest cardinality among all representative sets. Given a minimizable loss $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, we say $\mathcal{S}$ is a (minimum) representative set for L if it is a (minimum) representative set for* $\mathrm{prop}_{\mathcal{P}}[L]$*.*

Wang and Scott [85] first studies the notion of minimum representative sets under the name *embedding cardinality.*

We now define an embedding. In addition to matching loss values, as described above, we require the original reports to be optimal exactly when the corresponding embedded points are optimal.

Definition 15 (Embedding)**.** *A minimizable loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ embeds a loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ if there exists a representative set $\mathcal{S}$ for ℓ and an injective embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$ such that (i) for all $r \in \mathcal{S}$*

we have $L(\varphi(r)) = \ell(r)$, and (ii) for all $p \in \Delta_\mathcal{Y}, r \in \mathcal{S}$ we have

$$r \in \text{prop}_{\Delta_\mathcal{Y}}[\ell](p) \iff \varphi(r) \in \text{prop}_{\Delta_\mathcal{Y}}[L](p) \ . \tag{3.1}$$

If $\mathcal{S}$ is a minimal representative set, we say L tightly embeds ℓ.

To illustrate the idea of embedding, let us examine hinge loss in detail as a surrogate for 0-1 loss for binary classification. Recall that we have $\mathcal{R} = \mathcal{Y} = \{-1, +1\}$, with $L_{\text{hinge}}(u)_y = (1 - uy)_+$ and $\ell_{\text{0-1}}(r)_y := \mathbf{1}\{r \neq y\}$, typically with link function $\psi(u) = \text{sgn}(u)$. We will see that hinge loss embeds (2 times) 0-1 loss, via the embedding $\varphi(r) = r$. For condition (i), it is straightforward to check that $L_{\text{hinge}}(r)_y = 2\ell_{\text{0-1}}(r)_y$ for all $r, y \in \{-1, 1\}$. For condition (ii), let us compute the property each loss elicits, i.e., the set of optimal reports for each $p \in \Delta_\mathcal{Y}$:

$$\text{prop}_{\Delta_\mathcal{Y}}[\ell_{\text{0-1}}](p) = \begin{cases} 1 & p_1 > 1/2 \\ \{-1, 1\} & p_1 = 1/2 \\ -1 & p_1 < 1/2 \end{cases} \qquad \text{prop}_{\Delta_\mathcal{Y}}[L_{hinge}](p) = \begin{cases} [1, \infty) & p_1 = 1 \\ 1 & p_1 \in (1/2, 1) \\ [-1, 1] & p_1 = 1/2 \\ -1 & p_1 \in (0, 1/2) \\ (-\infty, -1] & p_1 = 0 \end{cases} .$$

In particular, we see that $-1 \in \text{prop}_{\Delta_\mathcal{Y}}[\ell_{\text{0-1}}](p) \iff p_1 \in [0, 1/2] \iff -1 \in \text{prop}_{\Delta_\mathcal{Y}}[L_{\text{hinge}}](p)$, and $1 \in \text{prop}_{\Delta_\mathcal{Y}}[\ell_{\text{0-1}}](p) \iff p_1 \in [1/2, 1] \iff 1 \in \text{prop}_{\Delta_\mathcal{Y}}[L_{\text{hinge}}](p)$. With both conditions of Definition 15 satisfied, we can conclude that L_{hinge} embeds $2\ell_{\text{0-1}}$. By results in § 3.4.2, one could also show that L_{hinge} embeds $2\ell_{\text{0-1}}$ by the fact that their Bayes risks match (Figure 3.5).

In this particular example, it is known $(L_{\text{hinge}}, \text{sgn})$ is calibrated with respect to 0-1 loss. More generally, however, it is not clear whether an arbitrary embedding yields a calibrated link. Indeed, apart from mapping the embedded points back to their original reports, via $\psi(\varphi(r)) = r$, how to map the remaining values is far from obvious. When the surrogate is polyhedral, we give a construction to map the remaining values in § 3.2, showing that embeddings always yield calibration.

While our notion of embedding is sufficient for calibration (and therefore consistency), it is worth noting that it is not *necessary* for these conditions. For example, while logistic loss does not embed 0-1 loss, the surrogate and link for logistic loss are consistent.

3.1 Embeddings and polyhedral losses

In this section, we establish a tight relationship between the technique of embedding and the use of polyhedral (piecewise-linear convex) surrogate losses, showing Theorem 1. We defer the question of when such surrogates are consistent to § 3.2.

A first observation is that if a loss L elicits a property Γ, then L restricted to some representative set $\mathcal{S}$, denoted $L|_{\mathcal{S}}$, elicits Γ restricted to $\mathcal{S}$. As a consequence, restricting to representative sets preserves the Bayes risk. We will use these observations throughout.

Lemma 3. *Let $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ elicit Γ, and let $\mathcal{S} \subseteq \mathcal{R}$ be representative for L. Then $L|_{\mathcal{S}}$ elicits $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{S}$ defined by $\gamma(p) = \Gamma(p) \cap \mathcal{S}$. Moreover, $\underline{L} = \underline{L|_{\mathcal{S}}}$.*

Proof. Let $p \in \Delta_{\mathcal{Y}}$ be fixed throughout. First let $r \in \gamma(p) = \Gamma(p) \cap \mathcal{S}$. Then $r \in \Gamma(p) = \arg\min_{u\in\mathcal{R}}\langle p, L(u)\rangle$, so as $r \in \mathcal{S}$ we have in particular $r \in \arg\min_{u\in\mathcal{S}}\langle p, L(u)\rangle$. For the other direction, suppose $r \in \arg\min_{u\in\mathcal{S}}\langle p, L(u)\rangle$. As $\mathcal{S}$ is representative for L, we must have some $s \in \Gamma(p) \cap \mathcal{S}$. On the one hand, $s \in \Gamma(p) = \arg\min_{u\in\mathcal{R}}\langle p, L(u)\rangle$. On the other, as $s \in \mathcal{S}$, we certainly have $s \in \arg\min_{u\in\mathcal{S}}\langle p, L(u)\rangle$. But now we must have $\langle p, L(r)\rangle = \langle p, L(s)\rangle$, and thus $r \in \arg\min_{u\in\mathcal{R}}\langle p, L(u)\rangle = \Gamma(p)$ as well. We now see $r \in \Gamma(p) \cap \mathcal{S}$. Finally, the equality of the Bayes risks $\min_{u\in\mathcal{R}}\langle p, L(u)\rangle = \min_{u\in\mathcal{S}}\langle p, L(u)\rangle$ follows immediately by the above, as $\emptyset \neq \Gamma(p) \cap \mathcal{S} \subseteq \Gamma(p)$ for all $p \in \Delta_{\mathcal{Y}}$. □

Lemma 3 leads to the following useful tool for finding embeddings.

Proposition 4. *Let a minimizable surrogate loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be given. If L has a finite representative set $\mathcal{S} \subseteq \mathbb{R}^d$, then L embeds the discrete loss $L|_{\mathcal{S}}$.*

Proof. Let $\Gamma = \mathrm{prop}_{\Delta_{\mathcal{Y}}}[L]$ and $\gamma = \mathrm{prop}_{\Delta_{\mathcal{Y}}}[L|_{\mathcal{S}}]$. Define $\varphi : \mathcal{S} \to \mathcal{S}$ to be the identity embedding. Condition (i) of an embedding is trivially satisfied, as $L|_{\mathcal{S}}(u) = L(u)$ for all $u \in \mathcal{S}$. Now let $u \in \mathcal{S}$.

From Lemma 3, for all $p \in \Delta_{\mathcal{Y}}$ we have $u \in \gamma(p) \iff u \in \Gamma(p) \cap \mathcal{S} \iff u \in \Gamma(p)$. We conclude condition (ii) of an embedding. □

We now shift our focus to *polyhedral* (piecewise-linear and convex) surrogates. Our first observation is that while polyhedral surrogates cannot elicit finite properties, in the sense that they have infinitely many possible reports, they do elicit properties with a finite range, meaning a finite set of possible optimal sets. This observation lets us apply results about finite representative sets to understand the structure of polyhedral surrogates and the losses they embed.

Lemma 4. *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral loss; then L is minimizable and elicits a property $\Gamma := \mathrm{prop}_{\Delta_{\mathcal{Y}}}[L]$. Then the range of Γ, given by $\Gamma(\Delta_{\mathcal{Y}}) = \{\Gamma(p) \subseteq \mathbb{R}^d : p \in \Delta_{\mathcal{Y}}\}$, is a finite set of closed polyhedra.*

The full proof is in § 3.7.

Proof sketch. We know that L is minimizable from Rockafellar [77, Corollary 19.3.1] as L is bounded from below. With $\mathcal{Y}$ finite, there are only finitely many supporting sets over $\Delta_{\mathcal{Y}}$. For $p \in \Delta_{\mathcal{Y}}$, the power diagram induced by projecting the epigraph of expected loss onto $\mathbb{R}^d$ is the same for any p of the same support ([28, Lemma 5]). Moreover, we have $\Gamma(p)$ being exactly one of the faces of the projected epigraph since the hyperplane $u \mapsto (u, \langle p, L(u)\rangle)$ supports the epigraph of the expected loss at exactly the property value; moreover, since the loss is polyhedral the supporting hyperplane must support on a face of the epigraph. Since this epigraph has finitely many faces (as it is polyhedral), the range of Γ is then (a subset) of elements of a finitely generated (finite supports) set of finite elements (finite faces). Moreover, each element of $\Gamma(\Delta_{\mathcal{Y}})$ is a closed polyhedron since it corresponds exactly to a closed face of a polyhedral set. □

See § 3.7 for the full proof.

Theorem 3. *Every polyhedral loss L embeds a discrete loss.*

Proof. Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral loss, and $\Gamma = \mathrm{prop}_{\Delta_{\mathcal{Y}}}[L]$. By Lemma 4, $\Gamma(\Delta_{\mathcal{Y}})$ is finite set. For each $U \in \Gamma(\Delta_{\mathcal{Y}})$, select $u_U \in U$, and let $\mathcal{S} = \{u_U : U \in \Gamma(\Delta_{\mathcal{Y}})\}$, which is again finite. For any $p \in \Delta_{\mathcal{Y}}$ then, let $U = \Gamma(p)$. We have $U \in \Gamma(\Delta_{\mathcal{Y}})$ by definition, and thus some $u_U \in \mathcal{S}$; in particular, $u_U \in U = \Gamma(p)$. We conclude that $\mathcal{S}$ is representative for L. Proposition 4 now states that L embeds $L|_{\mathcal{S}}$. □

We now turn to the reverse direction: which discrete losses are embedded by some polyhedral loss? Perhaps surprisingly, we show in Theorem 4 that *every* discrete loss is embeddable. Combining this result with Theorem 3 establishes Theorem 1. Further combining with Theorem 2, proved in the following section, this construction gives a consistent polyhedral surrogate for every discrete target loss.

The proof of Theorem 4 uses a construction via convex conjugate duality which has appeared in several different forms in the literature (e.g. [2, 23, 37]). We then apply a result we will prove in § 3.4: a minimizable surrogate embeds a discrete loss if and only if their Bayes risks match (Proposition 5).

Theorem 4. *Every discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ is embedded by a polyhedral loss.*

Proof. Let $n = |\mathcal{Y}|$, and let $C : \mathbb{R}^n \to \mathbb{R}$ be given by $(-\underline{\ell})^*$, the convex conjugate of $-\underline{\ell}$. From standard results in convex analysis, C is polyhedral as $-\underline{\ell}$ is, and C is finite on all of $\mathbb{R}^{\mathcal{Y}}$ as the domain of $-\underline{\ell}$ is bounded [77, Corollary 13.3.1]. Note that $-\underline{\ell}$ is a closed convex function, as the infimum of affine functions, and thus $(-\underline{\ell})^{**} = -\underline{\ell}$. Define $L : \mathbb{R}^n \to \mathbb{R}^{\mathcal{Y}}$ by $L(u) = C(u)\mathbb{1} - u$, where $\mathbb{1} \in \mathbb{R}^{\mathcal{Y}}$ is the all-ones vector. As C is polyhedral, so is L. We first show that L embeds ℓ, and then establish that the range of L is in fact $\mathbb{R}^{\mathcal{Y}}_+$, as desired.

We compute Bayes risks and apply Proposition 5 to see that L embeds ℓ. Observe that $\underline{\ell}$ is polyhedral as ℓ is discrete. For any $p \in \Delta_{\mathcal{Y}}$, we have

$$\begin{aligned}\underline{L}(p) &= \inf_{u \in \mathbb{R}^n} \langle p, C(u)\mathbb{1} - u \rangle \\ &= \inf_{u \in \mathbb{R}^n} C(u) - \langle p, u \rangle \\ &= -\sup_{u \in \mathbb{R}^n} \langle p, u \rangle - C(u) \\ &= -C^*(p) = -(-\underline{\ell}(p))^{**} = \underline{\ell}(p) \ .\end{aligned}$$

It remains to show $L(u)_y \geq 0$ for all $u \in \mathbb{R}^n$, $y \in \mathcal{Y}$. Letting $\delta_y \in \Delta_\mathcal{Y}$ be the point distribution on outcome $y \in \mathcal{Y}$, we have for all $u \in \mathbb{R}^n$, $L(u)_y \geq \inf_{u' \in \mathbb{R}^n} L(u')_y = \underline{L}(\delta_y) = \underline{\ell}(\delta_y) \geq 0$, where the final inequality follows from the nonnegativity of ℓ. □

While Theorem 4 constructs a consistent surrogate for any discrete loss, in some settings, such as structured prediction and information retrieval, the prediction dimension $d = n := |\mathcal{Y}|$ can be prohibitively large.[1] Recent work [29, 30, 71] yield characterizations for bounding the prediction dimension d for consistent convex surrogates and embeddings.

3.2 Consistency via calibrated links

We have now seen the tight relationship between polyhedral losses and embeddings; in particular, every polyhedral loss embeds some discrete loss. The embedding itself tells us how to link the embedded points back to the discrete reports (map $\varphi(r)$ to r). But it is not clear how to extend this to yield a full link function $\psi : \mathbb{R}^d \to \mathcal{R}$, and whether such a ψ can lead to consistency. In this section, we prove Theorem 2, restated below, which gives a construction to generate calibrated links for *any* polyhedral surrogate.

Theorem 2. *Given any polyhedral loss L, let ℓ be a discrete loss it embeds. There exists a link function ψ such that (L, ψ) is calibrated with respect to ℓ.*

Theorem 2 will follow immediately from Theorems 5 and 6, as discussed below.

Theorem 5 shows that calibration is equivalent to a geometric condition, which we call *separation*, of a link function ψ. Recall that for indirect elicitation, any point $u \in \Gamma(p)$ must link to a report $\psi(u) \in \gamma(p)$. (In terms of losses, u minimizing expected L-loss implies that $\psi(u)$ minimizes expected ℓ-loss, with respect to p.) The idea of separation is that points in the neighborhood of u must also link to to a report in $\gamma(p)$. Furthermore, there must be a uniform lower bound ϵ on the size of any such neighborhood.

[1] One can always reduce to $d = n - 1$ in Theorem 4 via a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^{n-1}$ which is injective on $\Delta_\mathcal{Y}$; redefining the surrogate appropriately, the Bayes risks will still match.

Definition 16 (Separated Link)**.** *Let properties $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathbb{R}^d$ and $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ be given. We say a link $\psi : \mathbb{R}^d \to \mathcal{R}$ is ϵ-separated with respect to Γ and γ if for all $u \in \mathbb{R}^d$ with $\psi(u) \notin \gamma(p)$, we have $d_\infty(u, \Gamma(p)) \geq \epsilon$, where $d_\infty(u, A) \doteq \inf_{a \in A} \|u - a\|_\infty$.*[2] *Similarly, we say ψ is ϵ-separated with respect to L and ℓ if it is ϵ-separated with respect to $\mathrm{prop}_{\Delta_{\mathcal{Y}}}[L]$ and $\mathrm{prop}_{\Delta_{\mathcal{Y}}}[\ell]$.*

Theorem 5. *Let polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, and link $\psi : \mathbb{R}^d \to \mathcal{R}$ be given. Then (L, ψ) is calibrated with respect to ℓ if and only if ψ is ϵ-separated with respect to L and ℓ for some $\epsilon > 0$.*

The proof is deferred to § 3.7.

To prove Theorem 2, it now suffices to show that for any polyhedral L embedding some ℓ, there exists a *separated* link ψ with respect to L and ℓ. This separated link is given by Construction 1 below.

Theorem 6. *Let polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ embed the discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$. Then there exists $\epsilon_0 > 0$ such that, for all $0 < \epsilon \leq \epsilon_0$, Construction 1 yields an ϵ-separated link with respect to L and ℓ.*

The proof is deferred to § 3.7.

To set the stage for Construction 1, we sketch the two main steps in proving Theorem 6: *(a)* showing that one can produce a link ψ such that (L, ψ) indirectly elicits ℓ; *(b)* "thickening" ψ such that it is separated.

For *(a)*, begin by linking each embedding point back to its original report. Now we must determine $\psi(u)$ for non-embedding points. The challenge is that we may have $u \in \Gamma(p) \cap \Gamma(p')$. Because u minimizes expected surrogate loss for both p and p', the link must satisfy $\psi(u) \in \gamma(p) \cap \gamma(p')$. It is not even clear *a priori* that these sets intersect. We use the definition of embedding and elicitation results, discussed in § 3.4, to show that for each such u there exists $r \in \mathcal{R}$ such that

[2] Frongillo and Waggoner [42] define ϵ-separation with a strict inequality $d_\infty(u, \Gamma(p)) > \epsilon$; we adopt a weak inequality as it is more convenient in examples.

$\Gamma_u \subseteq \gamma_r$, i.e. any p satisfying $u \in \Gamma(p)$ also satisfies $r \in \gamma(p)$. This implies that if $u \in \Gamma(p) \cap \Gamma(p')$, then there exists $r \in \gamma(p) \cap \gamma(p')$, so we may safely choose $\psi(u) = r$.

For *(b)*, we show that this link can be "thickened" by some positive ϵ, as described next. Consider an optimal surrogate report set, i.e. set of the form $U = \Gamma(p) = \arg\min_u \langle p, L(u) \rangle$. By indirect elicitation, ψ is already correct on U. Now, we "thicken" U to obtain $U_\epsilon = \{u : \|u - U\| \leq \epsilon\}$. Then we require that all points in U_ϵ are linked to some element of $\gamma(p) = \arg\min_r \langle p, \ell(r) \rangle$. For $\epsilon > 0$, this directly implies separation.

However, it is not clear that this linking is possible because a point u may be in multiple thickened sets U_ϵ, U'_ϵ, etc. Therefore, we need to take each possible collection U, U', etc. and thicken their intersection in an analogous way.

Given $u \in U \cap U' \cap \ldots$, we define a *link envelope* $\Psi(u)$ which encodes the remaining legal choices for $\psi(u)$ after imposing the requirements for each such set U, U', etc. The key claim is that, for small enough $\epsilon > 0$, $\Psi(u)$ is nonempty: at least one legal value for $\psi(u)$ remains. This claim follows from a geometric result that, for all small enough ϵ, a subset of thickenings U_ϵ intersect if and only if the U sets themselves intersect. When they do intersect, indirect elicitation implies that there exists a legal choice of link for the intersection of the thickenings. It is also important that, by Lemma 4, for polyhedral surrogates there are only finitely many sets of the form $U = \Gamma(p)$. This yields a single uniform smallest ϵ such that the key claim is true for all $u \in \mathbb{R}^d$.

Given the above proof sketch, the following construction is relatively straightforward. We initialize the link using the embedding points and optimal report sets, then use Ψ to narrow down to only legal choices; we then pick from $\psi(u)$ from $\Psi(u)$ arbitrarily. Theorem 6 implies that, for all small enough ϵ, the resulting link ψ is well-defined at all points.

Construction 1 (ϵ-thickened link)**.** *Given a polyhedral L that embeds some ℓ, an $\epsilon > 0$, and a norm $\|\cdot\|$, the ϵ-thickened link ψ is constructed as follows. First, define $\mathcal{U} = \{\Gamma(p) : p \in \Delta_\mathcal{Y}\}$. For each $U \in \mathcal{U}$, let $R_U = \{r \in \mathcal{R} : \varphi(r) \in U\}$, the reports whose embedding points are in U. First, initialize the link envelope $\Psi : \mathbb{R}^d \rightrightarrows \mathcal{R}$ by setting $\Psi(u) = \mathcal{R}$ for all u. Then for each $U \in \mathcal{U}$, for all* *points u such that $\inf_{u^* \in U} \|u^* - u\| < \epsilon$, update $\Psi(u) = \Psi(u) \cap R_U$. Finally, define $\psi(u) \in \Psi(u)$, breaking ties arbitrarily. If $\Psi(u)$ became empty, then leave $\psi(u)$ undefined.*

Remarks. Construction 1 is not necessarily computationally efficient as the number of labels n grows. In practice this potential inefficiency is not typically a concern, as the family of losses typically has some closed form expression in terms of n, and thus the construction can proceed at the symbolic level. We illustrate this formulaic approach in § 3.3.1.

Applying the ϵ-thickened link construction additionally enables one to verify the consistency of a proposed link ψ^*. For a given ϵ and norm $\|\cdot\|$, suppose one follows the routine of Construction 1 until the last step in which values for the link ψ are selected. Instead, we can simply test whether the proposed link values are contained in the valid choices, i.e., if $\psi^*(u) \in \Psi(u)$ for all $u \in \mathbb{R}^d$. If so, then the proposed link ψ^* is calibrated.

Regret transfer rates of calibrated polyhedral surrogates. Recall that the goal of surrogate regret minimization is to learn a hypothesis h that minimizes expected surrogate loss, then output hypothesis $\psi \circ h$, which hopefully minimizes expected target loss. Consistency is a minimal requirement: when surrogate regret[3] of h converges to zero, i.e. $\text{Regret}_L(h) \to 0$, so does target regret of $\psi \circ h$, i.e. $\text{Regret}_\ell(\psi \circ h) \to 0$. A natural question is whether *fast* convergence in surrogate regret implies fast convergence in target regret. Frongillo and Waggoner [42] shows that, for polyhedral surrogates, this is always the case.

Theorem 7 ([42], Theorem 1)**.** *Let (L, ψ) be a polyhedral surrogate that is consistent for a discrete loss ℓ. Then there exists $c > 0$ such that, for all hypotheses h, Regret$_\ell(\psi \circ h) \leq c \cdot$ Regret$_L(h)$.*

3.3 Application to Specific Surrogates

Our results give a framework to construct consistent polyhedral surrogates and link functions for any discrete target loss, as well as to verify consistency or inconsistency for specific surrogate and link pairs. Below, we illustrate the power of this framework with specific examples from the literature.

[3] Regret in this context is the difference between the expected loss of a hypothesis and the expected loss of the Bayes optimal hypothesis that minimizes expected loss. We refer the reader to [42] for a formal definition.

To warm up, we study the abstain surrogate given by Ramaswamy et al. [72], and show how to rederive their link function and surrogate regret bounds (§ 3.3.1). We then give three examples of subsequent works that use our framework, in the context of structured binary classification (§ 3.3.2), top-k classification (§ 3.3.4), and multiclass classification (§ 3.3.3). In all cases, our framework illuminates the behavior of inconsistent surrogates by revealing the discrete losses they embed, i.e., the true targets for which they are consistent. In structured binary classification and top-k classification, our framework also gives new consistent surrogates and/or link functions which would likely have been extremely challenging to derive without our framework.

When using our framework to study the (in)consistency of an existing surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, often the first step is determining the loss it embeds. To this end, we suggest the following general approach. First, for each $y \in \mathcal{Y}$, divide $\mathbb{R}^d$ into a finite number of polyhedral regions on which $L(\cdot)_y$ is an affine function. Second, identify the vertices of these polyhedral regions.[4] Third, conclude that the union of these vertices, $\mathcal{S} \subset \mathbb{R}^d$, is a finite representative set for L. Now L embeds $L|_{\mathcal{S}}$ from Proposition 4. From here one can further remove redundant reports until one arrives at a tight embedding if desired. Once the embedded discrete loss is known, the behavior of the surrogate becomes more clear: what problem it is solving, and for which restrictions on label distributions is it consistent for the original problem.

With an embedding in hand, Construction 1 provides a consistent link function. Yet for some target problems, the search for consistent surrogates has been restricted to those accommodating a particular canonical link function, such as k largest coordinates of the surrogate report in top-k classification (§ 3.3.4). Interestingly, our construction is also useful in this situation, where one wishes to verify the consistence of a given proposed link ψ. Recall that Construction 1 produces a set-valued link envelope Ψ, which yields the possible values any ϵ-separated link ψ could map to. If the given ψ is indeed consistent, then it is ϵ-separated for sufficiently small ϵ, so one can always construct such a Ψ and verify that $\psi(u) \in \Psi(u)$ for all $u \in \mathbb{R}^d$. More generally, while such canonical

[4] In some cases, these regions do not have vertices, such as the top-k surrogates which are invariant in the all-ones direction; here one can restrict to a subspace, or otherwise select among equivalent reports.

link functions may be intuitive for a given problem, our results suggest that researchers should consider setting them aside and instead let Construction 1 determine the link. See § 3.3.2 for a somewhat intricate example.

3.3.1 Consistency of abstain surrogate and link construction

Several authors consider a variant of multiclass classification, with the addition of an *abstain* option [10, 20, 24, 64, 72]. For $\alpha \in (0,1)$, Ramaswamy et al. [72] study the loss $\ell^\alpha : [n] \cup \{\bot\} \to \mathbb{R}^{\mathcal{Y}}_+$ defined by $\ell^\alpha(r)_y = 0$ if $r = y$, α if $r = \bot$, and 1 otherwise. The report $\bot$ corresponds to "abstaining" for a constant loss regardless of outcome y. For the case $\alpha = 1/2$, Ramaswamy et al. provide a polyhedral surrogate $L^{1/2}$, which they call the *binary encoded predictions (BEP)* surrogate, and link $\psi^{1/2}$ which are calibrated for $\ell^{1/2}$. Letting $d = \lceil \log_2(n) \rceil$, their surrogate is $L^{1/2} : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ given by

$$L^{1/2}(u)_y = \max_{j \in [d]} \left(1 - \varphi(y)_j u_j\right)_+ \ , \tag{3.2}$$

where $\varphi : [n] \to \{-1,1\}^d$ is an injection.[5] Observe that $L^{1/2}$ is exactly hinge loss when $n = 2$ and thus $d = 1$.

The authors show that the link $\psi^{1/2}$ is calibrated, where

$$\psi^{1/2}(u) = \begin{cases} \bot & \min_{i \in [d]} |u_i| \leq 1/2 \\ \varphi^{-1}(\mathrm{sgn}(u)) & \text{otherwise} \end{cases} \ , \tag{3.3}$$

and they go on to establish linear surrogate regret bounds for $(L^{1/2}, \psi^{1/2})$.

Using our framework, one can show that $L^{1/2}$ embeds (2 times) $\ell^{1/2}$, with the embedding given by φ where we define $\varphi(\bot) = 0 \in \mathbb{R}^d$. (Following the general procedure outlined above, the regions where $L^{1/2}$ is affine all have vertices in the set $\{-1,1\}^d \cup \{0\}$, meaning it is representative, and $L^{1/2}$ restricted to that set is precisely $2\ell^{1/2} \circ \varphi^{-1}$.) Moreover, as we illustrate in Figure 3.1(L), the link $\psi^{1/2}$ proposed by Ramaswamy et al. can be recovered from Construction 1 by choosing the norm $\|\cdot\|_\infty$ and $\epsilon = 1/2$. Hence, our framework could have simplified the process of finding $\psi^{1/2}$,

[5] To translate our notation to that of Ramaswamy et al. [72], take $B = -\varphi$.

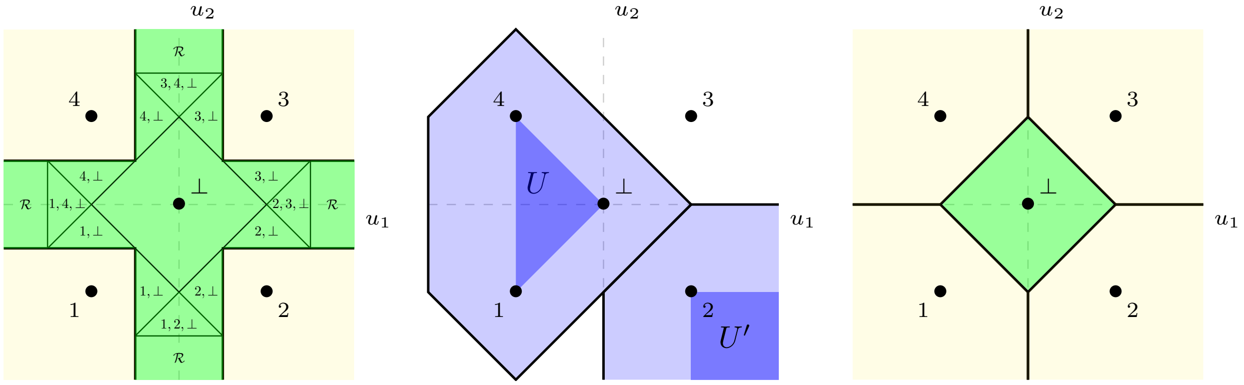


Figure 3.1: Constructing links for the abstain surrogate $L^{1/2}$ with $d = 2$. The embedding is shown in bold labeled by the corresponding reports. (L) The link envelope Ψ resulting from Construction 1 using $\|\cdot\|_\infty$ and $\epsilon = 1/2$, and a possible link ψ which matches eq. (3.3) from [72]. (M) An illustration of the thickened sets from Construction 1 for two sets $U, U' \in \mathcal{U}$, using $\|\cdot\|_1$ and $\epsilon = 1$. (R) The Ψ and ψ from Construction 1 using $\|\cdot\|_1$ and $\epsilon = 1$.

and the corresponding proof of consistency and surrogate regret bounds. To illustrate this point further, we give an alternate link ψ' corresponding to $\|\cdot\|_1$ and $\epsilon = 1$, shown in Figure 3.1(R),

$$\psi'(u) = \begin{cases} \perp & \|u\|_1 \leq 1 \\ \varphi^{-1}(\mathrm{sgn}(u)) & \text{otherwise} \end{cases} . \tag{3.4}$$

Construction 1 gives calibration of $(L^{1/2}, \psi')$ with respect to $\ell^{1/2}$. Aside from its simplicity, one possible advantage of ψ' is that it assigns $\perp$ to much less of the surrogate space $\mathbb{R}^d$.

3.3.2 Lovász hinge and the structured abstain problem

Many structured prediction settings can be thought of as making multiple predictions at once, with a loss function that jointly measures error based on the relationship between these predictions [44, 48, 68]. In the case of k binary predictions, these settings are typically formalized by taking the predictions and outcomes to be $\mathcal{R} = \mathcal{Y} = \{-1, 1\}^k$, with the ith coordinate giving the result for the ith binary prediction. A natural family of losses are those which are functions of the misprediction or disagreement set $\mathrm{dis}(r, y) = \{i \in [k] \mid r_i \neq y_i\}$, meaning we may write $\ell^f(r)_y = f(\mathrm{dis}(r, y))$ for some set function $f : 2^{[k]} \to \mathbb{R}$. For example, Hamming loss is given by $f(S) = |S|$. In an effort to provide a general convex surrogate for these settings when f is a submodular function, Yu and Blaschko [93] introduce the *Lovász hinge* surrogate $L^f : \mathbb{R}^k \to \mathbb{R}^{\mathcal{Y}}_+$ which leverages the well-known convex Lovász extension of submodular functions. While the authors provide theoretical justification and experiments, they leave open whether the Lovász hinge actually is consistent for ℓ^f.

Finocchiaro et al. [32] use our embedding framework to resolve the consistency of L^f, showing that it is inconsistent with respect to ℓ^f outside of the trivial case where f is modular, and thus ℓ^f is a weighted Hamming loss. Moreover, they show that L^f embeds a variant ℓ^f_{abs} of ℓ^f where one is allowed to abstain on a set of indices $A \subseteq [k]$, which they call the *structured abstain problem.* The inclusion of abstain options is natural when observing that the BEP surrogate $L^{1/2}$ from § 3.3.1 coincides with L^f for the function $f(S) = \mathbb{1}\{S \neq \emptyset\}$, so the multiclass abstain problem must be a special case of the Lovász hinge.

To derive ℓ^f_{abs}, the authors show that the set $\mathcal{V} = \{-1, 0, 1\}^k$ is representative for L^f, for any choice of f. From Proposition 4, they conclude that L^f embeds $\ell^f_{\text{abs}} := L^f|_{\mathcal{V}}$. Letting $\text{abs}(v) = \{i \in [k] \mid v_i = 0\}$ denote the "abstain" set, we may write $\ell^f_{\text{abs}} : \mathcal{V} \to \mathbb{R}^{\mathcal{Y}}_+$ as

$$\ell^f_{\text{abs}}(v)_y = f(\text{dis}(v, y) \setminus \text{abs}(v)) + f(\text{dis}(v, y)) \ . \tag{3.5}$$

(Observe that $\text{abs}(v, y) \subseteq \text{dis}(v, y)$, since $y \in \{-1, 1\}^k$.) By Theorem 2, then, the Lovász hinge is consistent with respect to the structured abstain loss ℓ^f_{abs} for some link function.

Actually determining this link function is nontrivial. Simple threshold links like for the BEP surrogate in § 3.3.1 are not always calibrated, thus casting doubt that a trial-and-error approach for finding the link would be successful. Instead, they leverage our thickened link construction (Construction 1) to derive two links ψ^* and $\psi^\diamond$, which have somewhat intricate geometric structure (Figure 3.2). Perhaps surprisingly, by deriving a link envelope $\hat{\Psi}$ which is contained in the envelopes for L^f for all submodular and increasing f, they prove that both (L^f, ψ^*) and $(L^f, \psi^\diamond)$ are simultaneously calibrated with respect to ℓ^f_{abs} for all such f.

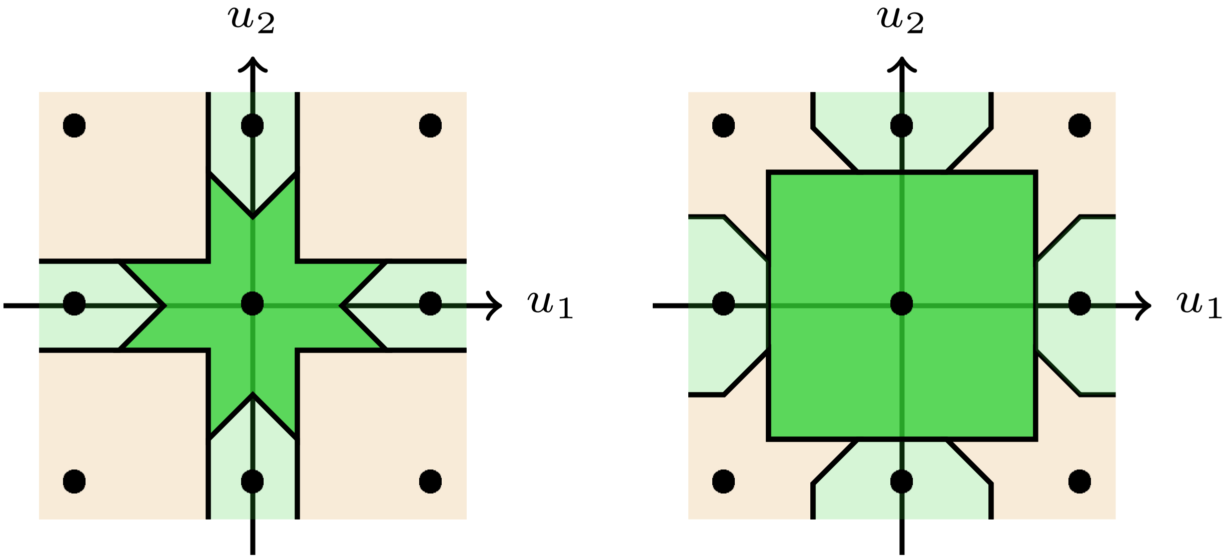


Figure 3.2: Links ψ^* and $\psi^\diamond$ such that (L^f, ψ^*) and $\psi^\diamond$ are calibrated with respect to ℓ^f_{abs} for all suitable f. All points in a region link to the point in $\{-1, 0, 1\}^2$ containing the point. ψ^* has a smaller abstain region than $\psi^\diamond$, and may lead to more predictions being made (as opposed to abstentions).

3.3.3 Embedding ordered partitions via Weston-Watkins hinge

As the hinge loss is one of the most common surrogates for binary support vector machines (SVMs), original extensions to the multiclass setting included a one-vs-all reduction to the binary problem via hinge loss, generating $\binom{n}{2}$ hyperplanes for n labels. Proposing a more efficient solution, Weston et al. [87] give an alternate surrogate for multiclass SVM prediction, defined as follows for predictions $u \in \mathbb{R}^n$,

$$L^{WW}(u)_y = \sum_{i \in \mathcal{Y} : i \neq y} (1 - (u_y - u_i))_+ \ , \tag{3.6}$$

which was later shown to be inconsistent with respect to 0-1 loss [63, 84].

Wang and Scott [85] use the embedding framework to show that the Weston-Watkins hinge embeds the *ordered partition* loss, and in turn, recover the result of inconsistency with respect to 0-1 loss. The report space for this discrete loss can be defined in terms of nested subsets of $[n] := \{1, \ldots, n\}$, as follows.[6]

$$\mathcal{T} = \{(T_0, \ldots, T_s) \mid s \geq 1, \emptyset = T_0 \subsetneq T_1 \subsetneq \ldots \subsetneq T_s = [n]\} \ .$$

The ordered partition target loss $\ell^{\mathcal{OP}} : \mathcal{T} \to \mathbb{R}^{\mathcal{Y}}_+$ embedded by Weston-Watkins hinge is then defined

$$\ell^{\mathcal{OP}}(T)_y = \sum_{i=1}^{s} (|T_i| \cdot \mathbf{1}\{y \notin T_{i-1}\}) - 1 \ .$$

The ordered partition loss can be interpreted as a variation of 0-1 loss incorporating varying confidence in different outcomes: reports are a nested sequence of sets, and the punishment for the outcome y is the cardinality of the first set containing y, plus the cardinality of all earlier sets.

Upon showing that L^{WW} embeds $\ell^{\mathcal{OP}}$, Wang and Scott proceed to use their characterization of $\text{prop}_{\mathcal{P}}[\ell^{\mathcal{OP}}]$ to give sufficient distributions assumptions over labels in $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ such that L^{WW} and the canonical argmax link $\psi^1(u) : u \mapsto r \in \arg\max_y \langle e_y, u \rangle$ are calibrated with respect to 0-1 loss on $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ (i.e., that eq. 2.6 holds for all $p \in \mathcal{P}$).

Sufficient constraints to recover consistency are characterized by comparing $\text{prop}_{\mathcal{P}}[\ell^{\mathcal{OP}}]$ and mode. Figure 3.3 gives the cells $\{p \in \Delta_{\mathcal{Y}} \mid T \in \text{prop}_{\mathcal{P}}[\ell^{\mathcal{OP}}](p)\}$ for each $T \in \mathcal{T}$, outlined in solid

[6] To recover the partition of Wang and Scott [85], one can define $S_i = T_i \setminus T_{i-1}$.

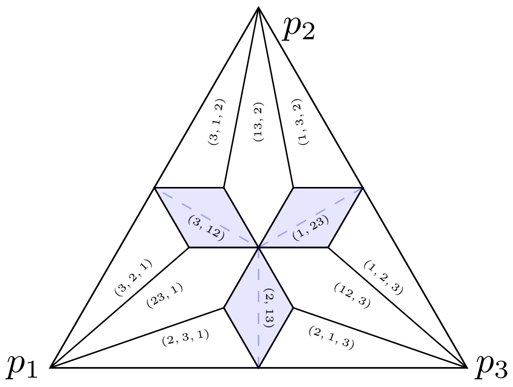


Figure 3.3: Level sets of $\text{prop}_{\mathcal{P}}[\ell^{\mathcal{OP}}]$, the property elicited by the ordered partition loss and embedded by L^{WW}. The level sets of the mode (for which L^{WW} is proposed as a surrogate) are given by the cells formed by the dashed blue lines. The level sets of $\text{prop}_{\mathcal{P}}[\ell^{\mathcal{OP}}]$ whose relative interiors span multiple cells of the mode cannot be properly linked to the mode. Here, this is demonstrated as the report corresponding to the cell has highest partition has more than one element, where in the white cells, the "highest" element of the partition is well-defined.

black. These cells are juxtaposed with the cells $\{p \in \Delta_{\mathcal{Y}} \mid y \in \text{mode}(p)\}$ for each $y \in \mathcal{Y}$, outlined in dashed blue, for which L^{WW} was originally proposed as a surrogate. Since each distribution in a cell of $\text{prop}_{\Delta_{\mathcal{Y}}}[\ell^{\mathcal{OP}}]$ corresponds to the same optimal report T, the choice of where to link that report must be constant. Thus, if a cell of $\text{prop}_{\Delta_{\mathcal{Y}}}[\ell^{\mathcal{OP}}]$ is fully contained in a cell of mode, then the corresponding $\text{prop}_{\Delta_{\mathcal{Y}}}[\ell^{\mathcal{OP}}]$ value can be mapped to the corresponding mode value. Conversely, if the relative interior of a cell of $\text{prop}_{\Delta_{\mathcal{Y}}}[\ell^{\mathcal{OP}}]$ corresponding to T spans multiple mode cells, it becomes unclear to which report the set T should be linked. To recover consistency on $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$, it suffices that $\mathcal{P}$ excludes these cells.

3.3.4 Surrogates for top-k classification

In settings like object recognition and information retrieval, the top-k classification problem arises in which one predicts a set S of k labels, and given the true label y, receives loss $\ell^{\text{top-}k}(S)_y = \mathbb{1}\{y \notin S\}$ [12, 59–61, 73, 74, 92]. In the literature on surrogates for top-k classification, one goal has been to find a surrogate satisfying the following three desiderata: convexity, consistency, and piecewise linear ("hinge-like") structure. Yang and Koyejo [92] show that a number of previously proposed polyhedral losses, i.e., those which are convex and hinge-like, are inconsistent. They further suggest that perhaps no surrogate could satisfy all three properties.

Finocchiaro et al. [31] apply the general approach outlined above to each of the polyhedral surrogates shown to be inconsistent by Yang and Koyejo, and determine the target problems they do solve, i.e., the discrete losses they embed. Each of the examined surrogates embeds a discrete loss which can be viewed as a variant of the top-k problem, allowing the algorithm to express varying levels of "confidence" on the top k labels or report less than k labels. The data distributions for which these optimal reports differ from the optimal top-k reports are shown in Table 3.1 with $n = 4$ and $k \in \{2, 3\}$.

For example, consider one of the surrogates, $L^{(4)}(u)_y = \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]}\right)_+$, where $u_{[i]}$ denotes the i^{th} largest element of $u \in \mathbb{R}^n$ and n is the total number of labels; the authors show that $L^{(4)}$ embeds $\ell^{(4)}(T)_y = \frac{k+1}{k+1-|T|}\mathbb{1}\{y \notin T\}$, where T is a set of at most k labels. These embedded losses may therefore be useful in top-k settings where choosing smaller sets may have some benefit, such as a search engine that can use unused space for advertisements. Using the losses each proposed surrogate embeds, Finocchiaro et al. go on to derive constraints on the label distributions under which the proposed surrogates are actually consistent for top-k classification which subsume previous constraints [92].

Beyond these previously proposed surrogates, Finocchiaro et al. also use our framework to derive the first consistent polyhedral surrogate for $\ell^{\text{top-}k}$,

$$L^k(u)_y = \max\left(u_{[1]}, \max_{m \in \{k+1,\ldots,n\}}\left[1 - \frac{k}{m} + \frac{1}{m}\sum_{i=1}^{m} u_{[i]}\right]\right) - u_y \ . \tag{3.7}$$

That is, they show that indeed a surrogate exists satisfying convexity, consistency, and hinge-like structure. In light of our framework, this fact is unsurprising: Theorems 1 and 2 imply that *every* discrete loss has a consistent polyhedral surrogate. This new surrogate L^k is given directly by the construction from the proof of Theorem 4 and applying Theorem 2 to obtain consistency. While Theorem 2 guarantees the existence of some consistent link function, the authors further ask whether the canonical argmax link function ψ^k, which returns the k largest elements of u, is consistent. They indeed confirm its consistency using our framework, showing that ψ^k is ϵ-separated for L^k and $\ell^{\text{top-}k}$, for any $\epsilon \leq \frac{1}{2n}$ [31, Theorem 4.4].

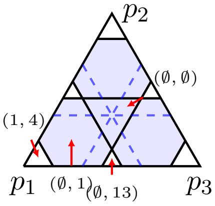
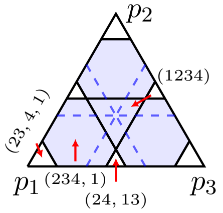
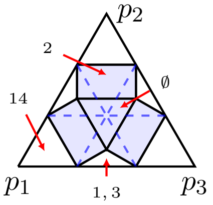
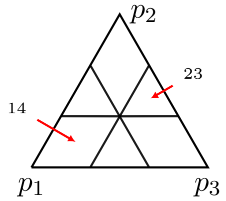
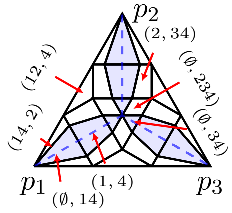
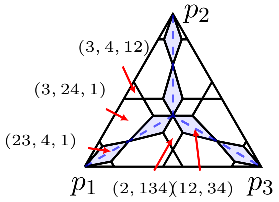
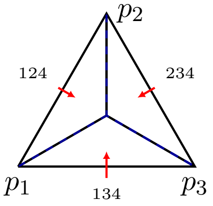
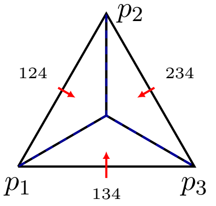
	$\text{prop}_{\Delta_{\mathcal{Y}}}[L^{(2)}]$	$\text{prop}_{\Delta_{\mathcal{Y}}}[L^{(3)}]$	$\text{prop}_{\Delta_{\mathcal{Y}}}[L^{(4)}]$	$\text{prop}_{\Delta_{\mathcal{Y}}}[L^{k}]$
$k = 2$	 p_2 $(\emptyset, \emptyset)$ $(1, 4)$ p_1 $(\emptyset, 1)$ $(\emptyset, 13)$ p_3	 p_2 (1234) $(23, 4, 1)$ p_1 $(234, 1)$ $(24, 13)$ p_3		
$k = 3$	 p_2 $(2, 34)$ $(12, 4)$ $(\emptyset, 234)$ $(14, 2)$ $(\emptyset, 34)$ p_1 $(1, 4)$ $(\emptyset, 14)$ p_3	 p_2 $(3, 4, 12)$ $(3, 24, 1)$ $(23, 4, 1)$ p_1 $(2, 134)$ $(12, 34)$ p_3		

Table 3.1: Visualizations of the properties elicited by the losses (embedded by) $L^{(2)}$, $L^{(3)}$, $L^{(4)}$ studied by Yang and Koyejo, and L^k in eq. (3.7) with $n = 4$ and $k \in \{2, 3\}$, fixing $p_4 = 1/4$. The dashed blue lines form cells whose elements are distributions p corresponding to the same $u \in \text{prop}_{\mathcal{P}}[\ell^{(k)}](p)$ labeling the cell. As the link ψ must be deterministic, in order for $(L^{(k)}, \psi)$ to be consistent with respect to $\ell^{(k)}$, each cell outlined in black is fully contained in exactly one cell from the dashed blue lines. Regions outlined in black that are filled in blue and cross the dashed blue lines suggest where deciding how to construct a link ψ is ambiguous, as the top-k elements of the optimal report u are ambiguous. White regions are therefore where the surrogate and any top-k link are consistent when restricting to data distributions whose conditional distributions are contained here. On the right, L^k shows our proposed surrogate that is consistent for top-k classification, demonstrated by no blue regions. Table from Finocchiaro et al. [31].

3.4 Additional Structure of Embeddings

We have shown in § 3.1 a tight connection between embeddings and polyhedral losses. Here we go beyond polyhedral losses, showing a more general necessary condition for an embedding: a surrogate embeds a discrete loss if and only if it has a polyhedral Bayes risk, or equivalently, a finite representative sets (Lemma 5). This result implies that the embedding condition simplifies to matching Bayes risks (Proposition 5). It also reveals some deeper structure of embeddings, even down to the geometry of the underlying property, and the equivalence of various notions of non-redundant predictions. In particular, we study a natural notion of a "trimed" loss function (Definition 17), and connect this definition to both tight embeddings and non-redundancy from property elicitation (Proposition 6).

3.4.1 Structure of polyhedral Bayes risks

While we have focused on polyhedral losses thus far, many of our results about embeddings extend to losses with polyhedral Bayes risks, a weaker condition. (We say a concave function is polyhedral if its negation is a polyhedral convex function.) To see that every polyhedral loss has a polyhedral Bayes risk, recall that Theorem 3 constructs a finite representative set $\mathcal{S}$ for any polyhedral loss L, and thus $\underline{L} = \underline{L|_{\mathcal{S}}}$ by Lemma 3, which is polyhedral. The condition is strictly weaker: a Bayes risk may be polyhedral even if the loss itself is not. For example, a modified hinge loss $L(r)_y = \max(r^2 - 1, 1 - ry)$ as shown in Figure 3.4, which matches hinge loss on the interval $[-1, 1]$ but is strictly convex outside the interval $[-2, 2]$, still embeds twice 0-1 loss.

We now present the structural result of Lemma 5, which will lay the foundation for the rest of this section. Lemma 5 observes that (minimizable) losses L with polyhedral Bayes risk have finite representative sets, and derives equivalent conditions on the level sets of the property elicited by L and tight embeddings. The proof of Lemma 5 is deferred to § 3.7.4.

Lemma 5. *Let $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ be a minimizable loss with a polyhedral Bayes risk $\underline{L}$. Then L has a finite representative set. Furthermore, letting $\Gamma = \mathrm{prop}_{\Delta_{\mathcal{Y}}}[L]$, there exist finite sets $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$ and*

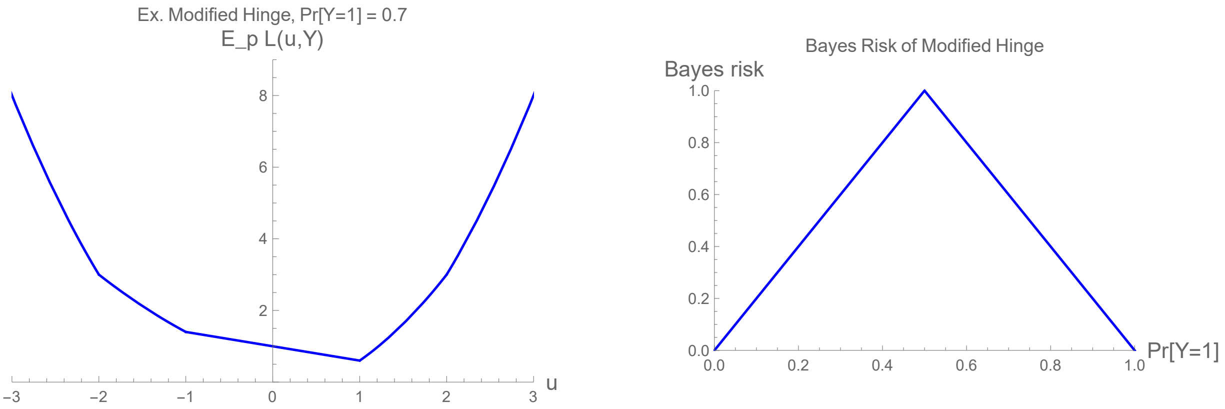


Figure 3.4: (L) Expected modified hinge loss for fixed distribution; (R) Bayes risk of modified hinge still matches the Bayes risk of hinge.

$\Theta = \{\theta_v \subseteq \Delta_{\mathcal{Y}} \mid v \in \mathcal{V}\}$, both uniquely determined by $\underline{L}$ alone, such that

(1) A set $\mathcal{R}' \subseteq \mathcal{R}$ is representative if and only if $\mathcal{V} \subseteq L(\mathcal{R}')$.

(2) A set $\mathcal{R}' \subseteq \mathcal{R}$ is minimum representative if and only if $L(\mathcal{R}') = \mathcal{V}$.

(3) A set $\mathcal{R}' \subseteq \mathcal{R}$ is representative if and only if $\Theta \subseteq \{\Gamma_r \mid r \in \mathcal{R}'\}$.

(4) A set $\mathcal{R}' \subseteq \mathcal{R}$ is minimum representative if and only if $\{\Gamma_r \mid r \in \mathcal{R}'\} = \Theta$.

(5) Every representative set for L contains a minimum representative set for L.

(6) The set of full-dimensional level sets of Γ is exactly Θ.

(7) For any $r \in \mathcal{R}$, there exists $\theta \in \Theta$ such that $\Gamma_r \subseteq \theta$.

(8) L tightly embeds $\ell : \mathcal{R}' \to \mathbb{R}^{\mathcal{Y}}_+$ if and only if ℓ is injective and $\ell(\mathcal{R}') = \mathcal{V}$.

As a finite representative set implies a polyhedral Bayes risk by Lemma 3, Lemma 5 shows that polyhedral Bayes risks are equivalent to having finite representative sets, which in turn gives an embedding by Proposition 4.

Corollary 1. *The following are equivalent for any minimizable loss $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$.*

(1) $\underline{L}$ is polyhedral.

(2) L has a finite representative set.

(3) L embeds a discrete loss.

From Corollary 1, L having a finite representative set is an equivalent condition to L being minimizable and $\underline{L}$ being polyhedral. (Recall that having a finite representative set already implies minimizability.) As it is also a more succinct condition, we will use the former in the sequel. In particular, the implications of Lemma 5 follow whenever L has a finite representative set.

3.4.2 Equivalent condition: matching Bayes risks

Lemma 5 leads to another appealing equivalent condition to our embedding condition in Definition 15: a surrogate embeds a discrete loss if and only if their Bayes risks match, visually demonstrated by Figure 3.5.

Proposition 5. *Let discrete loss ℓ and minimizable loss L be given. Then L embeds ℓ if and only if $\underline{L} = \underline{\ell}$.*

Proof. Define $\Gamma = \mathrm{prop}_{\Delta_\mathcal{Y}}[L]$ and $\gamma = \mathrm{prop}_{\Delta_\mathcal{Y}}[\ell]$.

$\implies$ Suppose L embeds ℓ, so we have some $\mathcal{S} \subseteq \mathcal{R}$ which is representative for ℓ and an embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$; take $\mathcal{U} := \varphi(\mathcal{S})$. Since $\mathcal{S}$ is representative for ℓ, by embedding condition (ii) we have $\{\gamma_s \mid s \in \mathcal{S}\} = \{\Gamma_u \mid u \in \mathcal{U}\}$, so $\mathcal{U}$ is representative for L. By Lemma 3, we have $\underline{\ell} = \underline{\ell|_\mathcal{S}}$ and $\underline{L} = \underline{L|_\mathcal{U}}$. As $L(\varphi(\cdot)) = \ell(\cdot)$ by embedding condition (i), for all $p \in \Delta_\mathcal{Y}$ we have

$$\underline{\ell}(p) = \underline{\ell|_\mathcal{S}}(p) = \min_{r\in\mathcal{S}} \langle p, \ell(r)\rangle = \min_{r\in\mathcal{S}} \langle p, L(\varphi(r))\rangle = \min_{u\in\mathcal{U}} \langle p, L(u)\rangle = \underline{L|_\mathcal{U}}(p) = \underline{L}(p)\ .$$

$\impliedby$ For the reverse implication, assume $\underline{L} = \underline{\ell}$, which are polyhedral functions as ℓ is discrete. From Lemma 5(2), we have some set $\mathcal{V} \subseteq \mathbb{R}^\mathcal{Y}_+$ and minimum representative sets $\mathcal{R}^* \subseteq \mathcal{R}$ and $\mathcal{U}^* \subseteq \mathcal{U}$, for ℓ and L respectively, such that $\ell(\mathcal{R}^*) = \mathcal{V} = L(\mathcal{U}^*)$. As $\mathcal{R}^*$ and $\mathcal{U}^*$ are miniumum, they cannot repeat loss vectors, and thus $|\mathcal{R}^*| = |\ell(\mathcal{R}^*)|$ and $|L(\mathcal{U}^*)| = |\mathcal{U}^*|$. We conclude that $\mathcal{R}^*$ and $\mathcal{U}^*$ are both in bijection with $\mathcal{V}$. The map $\varphi : \mathcal{R}^* \to \mathbb{R}^d$, given by $\varphi(r) = u \in \mathcal{U}^*$ where $\ell(r) = L(u)$, is therefore well-defined. Condition (i) of an embedding is immediate. From Proposition 4, ℓ embeds $\ell|_{\mathcal{R}^*}$ and L embeds $L|_{\mathcal{U}^*}$, both via the identity embedding. Using condition (ii) from both embeddings, for all $p \in \Delta_\mathcal{Y}$ and $r \in \mathcal{R}^*$, we have

$$r \in \gamma(p) \iff r \in \mathrm{prop}_{\Delta_\mathcal{Y}}[\ell|_{\mathcal{R}^*}](p) \iff \varphi(r) \in \mathrm{prop}_{\Delta_\mathcal{Y}}[L|_{\mathcal{U}^*}](p) \iff \varphi(r) \in \mathrm{prop}_{\Delta_\mathcal{Y}}[L](p)\ ,$$

giving condition (ii). □

Previous work from Duchi et al. [23, Proposition 4] realized the significance of matching Bayes risks for calibration with respect to the 0-1 loss. Proposition 5 broadens this general insight to any

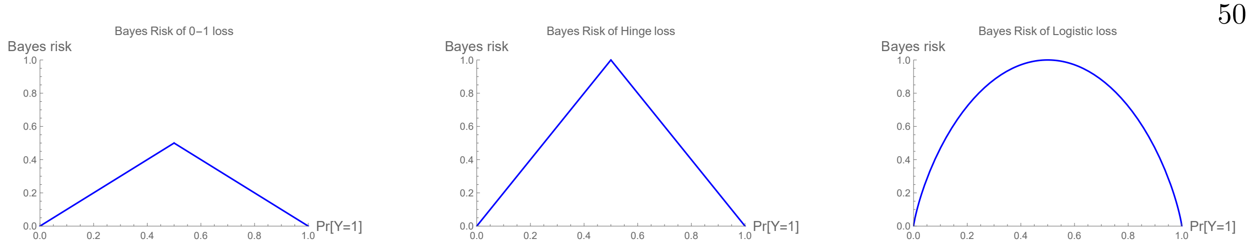


Figure 3.5: Bayes risks $\underline{L} : p \mapsto \inf_u \langle p, L(u) \rangle$ of 0-1, hinge, and logistic losses, respectively, plotted as a function of $p_1 = \mathbb{P}[Y = 1]$. Observe that the Bayes risks of 0-1 and hinge loss are both piecewise lienar and concave, while the Bayes risk of logistic loss is also concave, but not piecewise linear. Proposition 5 states that embedding is equivalent to matching Bayes risks. This confirms that hinge loss (M) embeds 0-1 loss (L), while logistic loss (R) does not.

discrete loss. Moreover, their result relies the Bayes risk of the surrogate being strictly concave, whereas polyhedral Bayes risks are never strictly concave.

3.4.3 Trimming a loss

Central to the structural results in Lemma 5 is the existence of a canonical set of loss vectors $\mathcal{V}$ which match the loss vectors of any minimum representative set. This fact may seem surprising when one considers that losses may have many mimimum representative sets. For example, consider hinge loss with a spurious extra dimension, i.e., $L : \mathbb{R}^2 \to \mathbb{R}^{\mathcal{Y}}$, $L((r_1, r_2))_y = \max(0, 1 - r_1 y)$ for $\mathcal{Y} = \{-1, +1\}$. Here the minimum representative sets are exactly the two-element sets of the form $\{(-1, a), (1, b)\}$ for any $a, b \in \mathbb{R}$. Lemma 5(2) states that, while the minimum representative set is not unique, its loss vectors are.

Motivated by this observation, let us define the "trim" of a loss to be this unique set $\mathcal{V}$ of loss vectors induced by any minimum representative set, which again is well-defined by Lemma 5(2).

Definition 17 (Trim)**.** *Given a loss $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ with a finite representative set, we define* $\mathrm{trim}(L) = \{L(r) \mid r \in \mathcal{R}^*\}$ *given any minimum representative set $\mathcal{R}^*$ for L.*

Using this notion of trimming a loss, we can again recast our embedding condition: a loss embeds another if and only if they have the same trim.

Proposition 6. *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ have a finite representative set, and let $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ be a discrete* *loss. Then L embeds ℓ if and only if* $\operatorname{trim}(L) = \operatorname{trim}(\ell)$*. Furthermore, L tightly embeds ℓ if and only if ℓ is injective and* $\operatorname{trim}(L) = \ell(\mathcal{R})$*.*

Proof. As L has a finite representative set, it is minimizable. Proposition 5 gives L embeds ℓ if and only if $\underline{L} = \underline{\ell}$. If $\underline{L} = \underline{\ell}$, Lemma 5(2) gives $\operatorname{trim}(L) = \operatorname{trim}(\ell)$. For the converse, suppose $\operatorname{trim}(L) = \operatorname{trim}(\ell) =: \mathcal{V}$. Define the discrete loss $\ell_{\text{trim}} : \mathcal{V} \to \mathcal{V}, v \mapsto v$. Then ℓ_{trim} is injective and $\ell_{\text{trim}}(\mathcal{V}) = \mathcal{V}$, so from Lemma 5(8), both L and ℓ tightly embed ℓ_{trim}. We conclude $\underline{L} = \underline{\ell_{\text{trim}}} = \underline{\ell}$ from Proposition 5. The second statement also follows directly from Lemma 5(8). □

3.4.4 Minimum representative sets and non-redundancy

The condition that a representative set be minimum implies that one has identified exactly the "active" reports of a loss, in some sense. We now relate this condition to another natural notion from the property elicitation literature: non-redundancy [37, 55]. Intuitively, a loss is non-redundant if no report is weakly dominated by another report.

Definition 18 (Non-redundancy)**.** *A loss $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ eliciting $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ is redundant if there are reports $r, r' \in \mathcal{R}$ with $r \neq r'$ such that $\Gamma_r \subseteq \Gamma_{r'}$, and non-redundant otherwise.*

From the structural result of Lemma 5, we can see that in fact these two notions are equivalent when L has a polyhedral Bayes risk.

Proposition 7. *Let $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ have a finite representative set $\mathcal{R}'$. Then $\mathcal{R}'$ is a minimum representative set for L if and only if $L|_{\mathcal{R}'}$ is non-redundant.*

Proof. Let $\Gamma = \operatorname{prop}_{\Delta_{\mathcal{Y}}}[L]$. Suppose first that $L|_{\mathcal{R}'}$ is redundant. Then there exist $r, r' \in \mathcal{R}'$ such that $\Gamma_r \subseteq \Gamma_{r'}$. Thus, for all $p \in \Gamma_r$, we have $\{r, r'\} \subseteq \Gamma(p)$. Therefore $\mathcal{R}' \setminus \{r\}$ still a representative set, so $\mathcal{R}'$ is not minimum.

Now suppose $L|_{\mathcal{R}'}$ is non-redundant. As $\mathcal{R}'$ is a representative set, Lemma 5(5) gives some minimum representative set $\mathcal{S} \subseteq \mathcal{R}'$. Suppose we had some $r \in \mathcal{R}' \setminus \mathcal{S}$. Now Lemma 5(4,7) gives some $s \in \mathcal{S}$ such that $\Gamma_r \subseteq \Gamma_s$, which contradicts $L|_{\mathcal{R}'}$ being non-redundant. We conclude $L(\mathcal{S}) = L(\mathcal{R}')$, meaning $\mathcal{R}'$ is a minimum representative set. □

Corollary 2. *Let loss $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ with finite representative set $\mathcal{R}'$ be given. Then L tightly embeds $L|_{\mathcal{R}'}$ if and only if $L|_{\mathcal{R}'}$ is non-redundant.*

In fact, we can show something stronger: the reports in minimum representative sets are precisely those which are not strictly redundant. To formalize this statement, given $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$, let $\mathrm{red}(\Gamma) := \{r \in \mathcal{R} \mid \exists r' \in \mathcal{R},\ \Gamma_r \subsetneq \Gamma_{r'}\}$ be the set of strictly redundant reports. Similarly, for minimizable L, let $\mathrm{red}(L) := \mathrm{red}(\mathrm{prop}_{\mathcal{P}}[L])$.

Proposition 8. *Let $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ have a finite representative set. Let $\mathcal{R}'$ be the union of all minimum representative sets for L. Then $\mathcal{R}' = \mathcal{R} \setminus \mathrm{red}(L)$.*

Proof. Let $\Gamma = \mathrm{prop}_{\Delta_{\mathcal{Y}}}[L]$. Let $\mathcal{S}$ be a minimum representative set for L, and let $s \in \mathcal{S}$. Suppose for a contradiction that $s \in \mathrm{red}(\Gamma)$. Then we have some $r \in \mathcal{R}$ with $\Gamma_s \subsetneq \Gamma_r$. From Lemma 5(4,7) we have some $s' \in \mathcal{S}$ such that $\Gamma_r \subseteq \Gamma_{s'}$. But now $\Gamma_s \subsetneq \Gamma_r \subseteq \Gamma_{s'}$, contradicting $\mathcal{S}$ being minimum representative. Thus $\mathcal{S} \subseteq \mathcal{R} \setminus \mathrm{red}(\Gamma)$.

For the reverse inclusion, let $r \in \mathcal{R} \setminus \mathrm{red}(\Gamma)$. Let $\mathcal{S}$ again be a minimum representative set for L. From Lemma 5(4,7), we have some $s \in \mathcal{S}$ such that $\Gamma_r \subseteq \Gamma_s$. By definition of $\mathrm{red}(L)$, we conclude $\Gamma_r = \Gamma_s$. Now take $\mathcal{S}' = (\mathcal{S} \setminus \{s\}) \cup \{r\}$, that is, the same set of reports with r replacing s. We have $\{\Gamma_s \mid s \in \mathcal{S}\} = \{\Gamma_{s'} \mid s' \in \mathcal{S}'\}$, and thus $\mathcal{S}'$ is a minimum representative for L by Lemma 5(4). As $r \in \mathcal{S}'$, we have $r \in \mathcal{R}'$ and we are done. □

As a corollary, we can state another characterization of trim in terms of redundant reports. The result follows immediately from the definition of trim.

Corollary 3. *Let $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ have a finite representative set. Then $\mathrm{trim}(L) = L(\mathcal{R} \setminus \mathrm{red}(L))$.*

This result motivates the analogous definition for properties, $\mathrm{trim}(\Gamma) := \{\Gamma_r \mid r \in \mathcal{R} \setminus \mathrm{red}(\Gamma)\}$. We leverage this definition next, to study embeddings at the property level.

3.4.5 A property elicitation perspective on trimmed losses

We conclude this section with a similar structural result about the properties embedded by another property. We say a property $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathbb{R}^d$ embeds a finite property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ if condition

(ii) of Definition 15 holds. In other words, Γ embeds γ if we have some representative set $\mathcal{S} \subseteq \mathcal{R}$ for γ and embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$ such that for all $s \in \mathcal{S}$ we have $\gamma_s = \Gamma_{\varphi(s)}$.

Roughly, our result is as follows. First, if Γ embeds γ and γ is non-redundant, the level sets of Γ must all be redundant relative to γ. In other words, Γ is exactly the property γ up to relabelling reports, just with other reports filling in the gaps between the embedded reports of γ. When working with convex losses, these extra reports often arise in the convex hull of the embedded reports. In this sense, we can regard embedding as only a slight departure from direct elicitation: if a loss L elicits Γ which embeds γ, we can almost think of L as eliciting γ itself. Finally, we have an important converse: if Γ has finitely many full-dimensional level sets, or equivalently, if $\mathrm{trim}(\Gamma)$ is finite, then Γ must embed some finite elicitable property with those same level sets. The statements about level sets make use of another corollary of Proposition 6, stated for properties.

Corollary 4. *Let $\Gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ be an elicitable property with a finite representative set. Then $\mathrm{trim}(\Gamma)$ is the set of full-dimensional level sets of Γ.*

Proof. Let L elicit Γ. From Lemma 5(4,6), for any finite minumum representative set $\mathcal{S} \subseteq \mathcal{R}$, the set $\{\Gamma_s \mid s \in \mathcal{S}\}$ is exactly the set of full-dimensional level sets Θ of Γ. From Proposition 7, we have $r \in \mathcal{R} \setminus \mathrm{red}(\Gamma)$ if and only if r is an element of some minimum representative set. As Γ has at least one minimum representative set, we conclude $\mathrm{trim}(\Gamma) = \{\Gamma_r \mid r \in \mathcal{R} \setminus \mathrm{red}(\Gamma)\} = \Theta$. □

Proposition 9. *Let $\Gamma : \Delta_\mathcal{Y} \rightrightarrows \mathbb{R}^d$ be an elicitable property. The following are equivalent:*

(1) Γ embeds a elicitable finite property $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$.

(2) $\mathrm{trim}(\Gamma)$ is a finite set.

(3) There is a finite minimum representative set $\mathcal{U}$ for Γ.

(4) There is a finite set of full-dimensional level sets $\hat{\Theta}$ of Γ, and $\cup \hat{\Theta} = \Delta_\mathcal{Y}$.

Moreover, when any of the above hold, $\mathrm{trim}(\gamma) = \mathrm{trim}(\Gamma) = \{\Gamma_u \mid u \in \mathcal{U}\} = \hat{\Theta}$.

Proof. Let L be a fixed loss eliciting Γ, so that in particular $\underline{L}$ is fixed. By definition of elicitable properties, L is minimizable. In each case, we will show that $\underline{L}$ is polyhedral (or equivalently, that

L has a finite representative set), and thus Lemma 5 will give us the set Θ of full-dimensional level sets of Γ, uniquely determined by $\underline{L}$. We will prove $1 \Rightarrow 2 \Rightarrow 3 \Rightarrow 4 \Rightarrow 1$, and in each case show that the relevant set of level sets is equal to Θ, giving the result.

$1 \Rightarrow 2$: Let $\mathcal{S}$ be the representative set for γ and $\varphi : \mathcal{S} \to \mathbb{R}^d$ the embedding. Since $\mathcal{S}$ is finite, $\varphi(\mathcal{S})$ is a finite representative set for Γ (and L; thus, $\underline{L}$ is polyhedral). Corollary 4 now gives $\mathrm{trim}(\Gamma) = \Theta$, which is finite, showing Case 2.

$2 \Rightarrow 3$: If $\mathrm{trim}(\Gamma)$ is finite, then in particular we have a finite set of reports $\mathcal{S} \subseteq \mathbb{R}^d$ such that $\mathrm{trim}(\Gamma) = \{\Gamma_s \mid s \in \mathcal{S}\}$. As Γ is elicitable, $\mathbb{R}^d$ is representative for Γ. By definition of trim, we have $\Delta_{\mathcal{Y}} = \cup_{r \in \mathbb{R}^d} \Gamma_r = \cup \mathrm{trim}(\Gamma) = \cup_{s \in \mathcal{S}} \Gamma_s$, and therefore $\mathcal{S}$ is representative for Γ and for L. As $\mathcal{S}$ is finite, we have $\underline{L}$ polyhedral. From Lemma 5(5), we have some minimum representative set $\mathcal{U} \subseteq \mathcal{S}$ for L and Γ, implying statement 3. Moreover, Lemma 5(4,6) gives $\{\Gamma_u \mid u \in \mathcal{U}\} = \Theta$.

$3 \Rightarrow 4$: Let $\mathcal{U}$ be a finite minimum representative set for Γ. Then $\underline{L} = \underline{L|_{\mathcal{U}}}$ is polyhedral. Lemma 5(4,6) once again gives $\{\Gamma_u \mid u \in \mathcal{U}\} = \Theta$. We simply let $\hat{\Theta} = \Theta$, giving statement 4 as $\mathcal{U}$ is representative.

$4 \Rightarrow 1$: Let $\mathcal{S} \subseteq \mathcal{R}$ such that $\{\Gamma_s \mid s \in \mathcal{S}\} = \hat{\Theta}$. Then $\mathcal{S}$ is representative for Γ and L, as $\cup \hat{\Theta} = \Delta_{\mathcal{Y}}$. Again, this yields a finite representative set for L. Lemma 3 now states that L embeds $L|_{\mathcal{S}}$, so Γ embeds $\gamma := \Gamma|_{\mathcal{S}}$, giving Case 1. Finally, Corollary 4 gives $\mathrm{trim}(\gamma) = \Theta$. □

As a final observation, recall that a property Γ elicited by a polyhedral loss has a finite range, in the sense that there are only finitely many optimal sets $\Gamma(p)$ for $p \in \Delta_{\mathcal{Y}}$ (Lemma 4). Proposition 9 shows the dual statement: there are only finitely many level sets Γ_u for $u \in \mathbb{R}^d$. In other words, both Γ and Γ^{-1} have a finite range as multivalued maps.

3.5 Polyhedral Indirect Elicitation Implies Consistency

Our last result concerns indirect elicitation as a necessary condition for consistency when restricting to polyhedral losses. Intuitively, a loss L indirectly elicits a property γ if we can compute γ from $\mathrm{prop}_{\mathcal{P}}[L]$. To formalize the condition, we use the notion of a property refining another from Frongillo and Kash [37].

Definition 19 (Refines)**.** *Let* $\Gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ *and* $\Gamma' : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}'$*. Then* Γ *refines* Γ' *if for all* $r \in \mathcal{R}$*, there exists* $r' \in \mathcal{R}'$ *such that* $\Gamma_r \subseteq \Gamma'_{r'}$*.*

Equivalently, Γ refines Γ' if there is some "link" function $\psi : \mathcal{R} \to \mathcal{R}'$ such that $r \in \Gamma(p) \implies \psi(r) \in \Gamma'(p)$ for all $p \in \Delta_\mathcal{Y}$. We will use the fact that refinement is transitive: if Γ refines Γ' and Γ' refines Γ'', then Γ refines Γ''.

Definition 20 (Indirectly elicits)**.** *A loss* L *indirectly elicits a property* γ *if* $\mathrm{prop}_\mathcal{P}[L]$ *refines* γ*.*

It is straightforward to verify that consistency, and therefore calibration, implies indirect elicitation [6, 30, 82]. Indirect elicitation may appear much weaker than calibration, since in particular it does not depend on the loss except through the property it elicits, and thus only depends on the exact minimizers of the loss. Surprisingly, for minimizable polyhedral surrogates, we show the converse: indirect elicitation implies calibration, and therefore consistency.

A useful lemma is that for minimizable polyhedral losses, indirect elicitation must always pass through an embedding. This result holds more generally whenever L has a finite representative set, as in § 3.4.

Lemma 6. *Let* L *be a minimizable polyhedral loss. Then* L *indirectly elicits a property* γ *if and only if* L *tightly embeds a discrete loss* ℓ *such that* $\mathrm{prop}_\mathcal{P}[\ell]$ *refines* γ*.*

Proof. Let $L : \mathbb{R}^d \to \mathbb{R}^\mathcal{Y}_+$ be polyhedral, and $\Gamma = \mathrm{prop}_\mathcal{P}[L]$. Then L tightly embeds a discrete loss from Lemma 5(8). Furthermore, Lemma 5(4,7,8) implies that $\mathrm{prop}_\mathcal{P}[L]$ refines $\mathrm{prop}_\mathcal{P}[\ell]$ for any discrete loss ℓ that L tightly embeds.

We claim that, for any property γ, and any loss ℓ that L tightly embeds, $\mathrm{prop}_\mathcal{P}[L]$ refines γ if and only if $\mathrm{prop}_\mathcal{P}[\ell]$ refines γ. If $\mathrm{prop}_\mathcal{P}[\ell]$ refines γ, then $\mathrm{prop}_\mathcal{P}[L]$ refines γ by transitivity. For the other direction, Lemma 5(4,8) shows that the level sets of $\mathrm{prop}_\mathcal{P}[\ell]$ are contained in the set $\{\Gamma_u \mid u \in \mathbb{R}^d\}$. Thus, if $\mathrm{prop}_\mathcal{P}[L]$ refines γ, then in particular $\mathrm{prop}_\mathcal{P}[\ell]$ refines γ. The result now follows immediately from the claim. □

Theorem 8. *Let L be a minimizable polyhedral loss which indirectly elicits a finite property γ. For any loss ℓ eliciting γ, there exists a link ψ such that (L, ψ) is calibrated (and consistent) with respect to ℓ.*

Proof. Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral loss indirectly eliciting $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$, and let ℓ be a discrete loss eliciting γ. By Lemma 6, L tightly embeds a discrete loss $\ell^{\mathsf{e}} : \mathcal{R}^{\mathsf{e}} \to \mathbb{R}^{\mathcal{Y}}_+$ such that $\gamma^{\mathsf{e}} := \mathrm{prop}_{\Delta_{\mathcal{Y}}}[\ell^{\mathsf{e}}]$ refines γ. From refinement, we can define a function $\psi^{\mathcal{R}} : \mathcal{R}^{\mathsf{e}} \to \mathcal{R}$ such that for all $r \in \mathcal{R}^{\mathsf{e}}$ and $p \in \Delta_{\mathcal{Y}}$ we have $r \in \gamma^{\mathsf{e}}(p) \implies \psi^{\mathcal{R}}(r) \in \gamma(p)$. Finally, Theorem 2 gives a link function $\psi^{\mathsf{e}} : \mathbb{R}^d \to \mathcal{R}^{\mathsf{e}}$ such that (L, ψ^{e}) is calibrated with respect to ℓ^{e}.

Consider $\psi := \psi^{\mathcal{R}} \circ \psi^{\mathsf{e}}$ and fix $p \in \Delta_{\mathcal{Y}}$. For any $u \in \mathbb{R}^d$, if $\psi^{\mathsf{e}}(u) \in \gamma^{\mathsf{e}}(p)$, then $\psi(u) = \psi^{\mathcal{R}}(\psi^{\mathsf{e}}(u)) \in \gamma(p)$ by definition of ψ and $\psi^{\mathcal{R}}$. Contrapositively, $\psi(u) \notin \gamma(p) \implies \psi^{\mathsf{e}}(u) \notin \gamma^{\mathsf{e}}(p)$. Thus, we have

$$\{u \in \mathbb{R}^d \mid \psi(u) \notin \gamma(p)\} \subseteq \{u \in \mathbb{R}^d \mid \psi^{\mathsf{e}}(u) \notin \gamma^{\mathsf{e}}(p)\} \ . \tag{3.8}$$

Combining eq. (3.8) with the fact that (L, ψ^{e}) is calibrated with respect to ℓ^{e}, we have

$$\inf_{u \in \mathbb{R}^d : \psi(u) \notin \gamma(p)} \langle p, L(u) \rangle \geq \inf_{u \in \mathbb{R}^d : \psi^{\mathsf{e}}(u) \notin \gamma^{\mathsf{e}}(p)} \langle p, L(u) \rangle > \inf_{u \in \mathbb{R}^d} \langle p, L(u) \rangle \ ,$$

showing calibration of ψ. Consistency follows as calibration and consistency are equivalent in this setting [71]. □

Theorem 8 gives a somewhat surprising result: despite the fact that indirect elicitation appears to be a somewhat weak necessary condition for consistency in general, the two conditions are equivalent for polyhedral surrogates.

3.6 Chapter conclusion

Several directions for future work remain. We show in Theorem 8 that indirect elicitation is equivalent to consistency when restricting to the class of polyhedral surrogates; we would like to identify other classes of surrogates for which this equivalence holds. It would also be interesting to explore embeddings through the lens of superprediction sets [88]. Finally, it is important for applications to understand the minimum prediction dimension d of a consistent convex surrogate $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ for a given target problem, also called its elicitation complexity. One approach to this question is to first understand the minimum d for which an embedding L exists, a study initiated by Finocchiaro et al. [29], and then relate this dimension to polyhedral, or general convex, elicitation complexity.

3.7 Chapter appendix

3.7.1 Power diagrams

First, we present several definitions from Aurenhammer [8].

Definition 21. *A cell complex in $\mathbb{R}^d$ is a set C of faces (of dimension $0, \dots, d$) which (i) union to $\mathbb{R}^d$, (ii) have pairwise disjoint relative interiors, and (iii) any nonempty intersection of faces F, F' in C is a face of F and F' and an element of C.*

Definition 22. *Given sites $s_1, \dots, s_k \in \mathbb{R}^d$ and weights $w_1, \dots, w_k \geq 0$, the corresponding power diagram is the cell complex given by*

$$\mathrm{cell}(s_i) = \{x \in \mathbb{R}^d : \forall j \in \{1, \dots, k\} \; \|x - s_i\|^2 - w_i \leq \|x - s_j\|^2 - w_j\} \; . \tag{3.9}$$

Definition 23. *A cell complex C in $\mathbb{R}^d$ is affinely equivalent to a (convex) polyhedron $P \subseteq \mathbb{R}^{d+1}$ if C is a (linear) projection of the faces of P.*

Proposition 5, focuses on matching the values of Bayes Risks, while the following result from Aurenhammer [8] allows us to move towards understanding the projection of the Bayes Risk onto the simplex $\Delta_{\mathcal{Y}}$. In particular, one can consider the epigraph of a polyhedral convex function on $\mathbb{R}^d$ and the projection down to $\mathbb{R}^d$; in this case we call the resulting power diagram *induced* by the convex function.

Theorem 9 (Aurenhammer [8])**.** *A cell complex is affinely equivalent to a convex polyhedron if and only if it is a power diagram.*

We extend Theorem 17 to a weighted sum of convex functions, showing that the induced power diagram is the same for any choice of strictly positive weights.

Lemma 7. *Let $f_1, \ldots, f_m : \mathbb{R}^d \to \mathbb{R}$ be polyhedral convex functions. The power diagram induced by $\sum_{i=1}^m p_i f_i$ is the same for all $p \in \overset{\circ}{(\Delta_{\mathcal{Y}})}$.*

Proof. For any polyhedral convex function g with epigraph P, the proof of Aurenhammer [8, Theorem 4] shows that the power diagram induced by g is determined by the facets of P. Let F be a facet of P, and F' its projection down to $\mathbb{R}^d$. It follows that $g|_{F'}$ is affine, and thus g is differentiable on $\overset{\circ}{(F')}$ with constant derivative $d \in \mathbb{R}^d$. Conversely, for any subgradient d' of g, the set of points $\{x \in \mathbb{R}^d : d' \in \partial g(x)\}$ is the projection of a face of P; we conclude that $F = \{(x, g(x)) \in \mathbb{R}^{d+1} : d \in \partial g(x)\}$ and $F' = \{x \in \mathbb{R}^d : d \in \partial g(x)\}$.

Now let $f := \sum_{i=1}^k f_i$ with epigraph P, and $f' := \sum_{i=1}^k p_i f_i$ with epigraph P'. By Rockafellar [77], f, f' are polyhedral. We now show that f is differentiable whenever f' is differentiable:

$$\begin{aligned}
\partial f(x) = \{d\} &\iff \sum_{i=1}^k \partial f_i(x) = \{d\} \\
&\iff \forall i \in \{1, \ldots, k\},\ \partial f_i(x) = \{d_i\} \\
&\iff \forall i \in \{1, \ldots, k\},\ \partial p_i f_i(x) = \{p_i d_i\} \\
&\iff \sum_{i=1}^k \partial p_i f_i(x) = \left\{\sum_{i=1}^k p_i d_i\right\} \\
&\iff \partial f'(x) = \left\{\sum_{i=1}^k p_i d_i\right\} .
\end{aligned}$$

From the above observations, every facet of P is determined by the derivative of f at any point in the interior of its projection, and vice versa. Letting x be such a point in the interior, we now see that the facet of P' containing $(x, f'(x))$ has the same projection, namely $\{x' \in \mathbb{R}^d : \nabla f(x) \in \partial f(x')\} = \{x' \in \mathbb{R}^d : \nabla f'(x) \in \partial f'(x')\}$. Thus, the power diagrams induced by f and f' are the same. The conclusion follows from the observation that the above held for any strictly positive weights p, and f was fixed. □

We now include the full proof of Lemma 4.

Lemma 4. *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral loss; then L is minimizable and elicits a property $\Gamma := \mathrm{prop}_{\Delta_{\mathcal{Y}}}[L]$. Then the range of Γ, given by $\Gamma(\Delta_{\mathcal{Y}}) = \{\Gamma(p) \subseteq \mathbb{R}^d : p \in \Delta_{\mathcal{Y}}\}$, is a finite set of closed polyhedra.*

Proof. First, observe that $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ is finite and bounded from below (by $\vec{0}$), and thus its infimum is finite. Therefore, we can apply Rockafellar [77, Corollary 19.3.1] to conclude that its infimum is attained for all $p \in \Delta_{\mathcal{Y}}$ and is therefore minimizable; thus, elicits a property.

For all p, let $P(p)$ be the epigraph of the convex function $u \mapsto \langle p, L(u) \rangle$. From Lemma 7, we have that the power diagram $D_{\mathcal{Y}}$ induced by the projection of $P(p)$ onto $\mathbb{R}^d$ is the same for any $p \in \mathring{(\Delta_{\mathcal{Y}})}$. Let $\mathcal{F}_{\mathcal{Y}}$ be the set of faces of $D_{\mathcal{Y}}$, which by the above are the set of faces of $P(p)$ projected onto $\mathbb{R}^d$ for any $p \in \mathring{(\Delta_{\mathcal{Y}})}$.

We claim for all $p \in \mathring{(\Delta_{\mathcal{Y}})}$, that $\Gamma(p) \in \mathcal{F}_{\mathcal{Y}}$. To see this, let $u \in \Gamma(p)$, and $u' = (u, \langle p, L(u) \rangle) \in P(p)$. The optimality of u is equivalent to u' being contained in the face F of $P(p)$ exposed by the normal $(0, \dots, 0, -1) \in \mathbb{R}^{d+1}$. Thus, $\Gamma(p) = \arg\min_{u \in \mathbb{R}^d} \langle p, L(u) \rangle$ is a projection of F onto $\mathbb{R}^d$, which is an element of $\mathcal{F}_{\mathcal{Y}}$.

Now for $p \notin \mathring{(\Delta_{\mathcal{Y}})}$, consider $\mathcal{Y}' \subsetneq \mathcal{Y}$, $\mathcal{Y}' \neq \emptyset$. Applying the above argument, we have a similar guarantee: a finite set $\mathcal{F}_{\mathcal{Y}'}$ such that $\Gamma(p) \in \mathcal{F}_{\mathcal{Y}'}$ for all p with support exactly $\mathcal{Y}'$. Taking $\mathcal{F} = \bigcup \{\mathcal{F}_{\mathcal{Y}'} | \mathcal{Y}' \subseteq \mathcal{Y}, \mathcal{Y}' \neq \emptyset\}$, we have for all $p \in \Delta_{\mathcal{Y}}$ that $\Gamma(p) \in \mathcal{F}$, giving $\mathcal{U} \subseteq \mathcal{F}$. As $\mathcal{F}$ is finite, so is $\mathcal{U}$, and the elements of $\mathcal{U}$ are closed polyhedra as faces of $D_{\mathcal{Y}'}$ for some $\mathcal{Y}' \subseteq \mathcal{Y}$. □

3.7.2 Equivalence of separation and calibration for polyhedral surrogates

We recall that Theorem 2 states that, if a polyhedral L embeds a discrete ℓ, then there exists a calibrated link ψ. Theorem 2 is directly implied by the combination of Theorem 5, that calibration is equivalent to separation (Definition 16); and Theorem 6, existence of a separated link. Theorem 5 is proven in this section and Theorem 6 is proven in Appendix 3.7.3.

Throughout we will work with the two *regret* functions: the *surrogate regret* $R_L(u, p) = \langle p, L(u) \rangle - \underline{L}(p)$, and similarly the *target regret* $R_\ell(r, p) = \langle p, \ell(r) \rangle - \underline{\ell}(p)$. In fact, the results in this section can be extended to surrogate regret bounds; see Frongillo and Waggoner [42].

We first show one direction: any calibrated link from a polyhedral surrogate to a discrete target must be ϵ-separated. The proof follows a similar argument to that of Tewari and Bartlett [84, Lemma 6].

Lemma 8. *Let polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, and link $\psi : \mathbb{R}^d \to \mathcal{R}$ be given such that (L, ψ) is calibrated with respect to ℓ. Then there exists $\epsilon > 0$ such that ψ is ϵ-separated with respect to* $\mathrm{prop}_{\mathcal{P}}[L]$ *and* $\mathrm{prop}_{\mathcal{P}}[\ell]$.

Proof. Let $\Gamma := \mathrm{prop}_{\mathcal{P}}[L]$ and $\gamma := \mathrm{prop}_{\mathcal{P}}[\ell]$. Suppose that ψ is not ϵ-separated for any $\epsilon > 0$. Then letting $\epsilon_i := 1/i$ we have sequences $\{p_i\}_i \subset \Delta_{\mathcal{Y}}$ and $\{u_i\}_i \subset \mathbb{R}^d$ such that for all $i \in \mathbb{N}$ we have both $\psi(u_i) \notin \gamma(p_i)$ and $d_\infty(u_i, \Gamma(p_i)) < \epsilon_i$. First, observe that there are only finitely many values for $\gamma(p_i)$ and $\Gamma(p_i)$, as $\mathcal{R}$ is finite and L is polyhedral (from Lemma 4). Thus, there must be some $p \in \Delta_{\mathcal{Y}}$ and some infinite subsequence indexed by $j \in J \subseteq \mathbb{N}$ where for all $j \in J$, we have $\psi(u_j) \notin \gamma(p)$ and $\Gamma(p_j) = \Gamma(p)$.

Next, observe that, as L is polyhedral, the expected loss $\langle p, L(u)\rangle$ is β-Lipschitz in $\|\cdot\|_\infty$ for some $\beta > 0$. Thus, for all $j \in J$, we have

$$\begin{aligned} d_\infty(u_i, \Gamma(p)) < \epsilon_j &\implies \exists u^* \in \Gamma(p)\ \|u_j - u^*\|_\infty < \epsilon_j \\ &\implies |\langle p, L(u_j)\rangle - \langle p, L(u^*)\rangle| < \beta\epsilon_j \\ &\implies |\langle p, L(u_j)\rangle - \underline{L}(p)| < \beta\epsilon_j\ . \end{aligned}$$

Finally, for this p, we have

$$\inf_{u:\psi(u)\notin\gamma(p)} \langle p, L(u)\rangle \le \inf_{j\in J} \langle p, L(u_j)\rangle = \underline{L}(p)\ ,$$

contradicting the calibration of ψ. □

For the other direction, we will make use of Hoffman constants for systems of linear inequalities. See Zalinescu [94] for a modern treatment.

Theorem 10 (Hoffman constant [51]). *Given a matrix $A \in \mathbb{R}^{m\times n}$, there exists some smallest $H(A) \ge 0$, called the Hoffman constant (with respect to $\|\cdot\|_\infty$), such that for all $b \in \mathbb{R}^m$ and all*

$x \in \mathbb{R}^n$,

$$d_\infty(x, S(A,b)) \le H(A)\|(Ax-b)_+\|_\infty \ , \tag{3.10}$$

where $S(A,b) = \{x \in \mathbb{R}^n \mid Ax \le b\}$ and $(u)_+ := \max(u,0)$ component-wise.

Lemma 9. *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral loss with $\Gamma = \mathrm{prop}_{\mathcal{P}}[L]$. Then for any fixed p, there exists some smallest constant $H_{L,p} \ge 0$ such that $d_\infty(u, \Gamma(p)) \le H_{L,p} R_L(u,p)$ for all $u \in \mathbb{R}^d$.*

Proof. Since L is polyhedral, there exist $a_1, \ldots, a_m \in \mathbb{R}^d$ and $c \in \mathbb{R}^m$ such that we may write $\langle p, L(u)\rangle = \max_{1\le j\le m} a_j \cdot u + c_j$. Let $A \in \mathbb{R}^{m\times d}$ be the matrix with rows a_j, and let $b = \underline{L}(p)\mathbb{1} - c$, where $\mathbb{1} \in \mathbb{R}^m$ is the all-ones vector. Then we have

$$\begin{aligned}
S(A,b) &:= \{u \in \mathbb{R}^d \mid Au \le b\} \\
&= \{u \in \mathbb{R}^d \mid Au + c \le \underline{L}(p)\mathbb{1}\} \\
&= \{u \in \mathbb{R}^d \mid \forall i \, (Au+c)_i \le \underline{L}(p)\} \\
&= \{u \in \mathbb{R}^d \mid \max_i \, (Au+c)_i \le \underline{L}(p)\} \\
&= \{u \in \mathbb{R}^d \mid \langle p, L(u)\rangle \le \underline{L}(p)\} \\
&= \Gamma(p) \ .
\end{aligned}$$

Similarly, we have $\max_i \, (Au-b)_i = \langle p, L(u)\rangle - \underline{L}(p) = R_L(u,p) \ge 0$. Thus,

$$\begin{aligned}
\|(Au-b)_+\|_\infty &= \max_i \, ((Au-b)_+)_i \\
&= \max((Au-b)_1, \ldots, (Au-b)_m, 0) \\
&= \max(\max_i \, (Au-b)_i, \, 0) \\
&= \max_i \, (Au-b)_i \\
&= R_L(u,p) \ .
\end{aligned}$$

Now applying Theorem 10, we have

$$\begin{aligned} d_\infty(u, \Gamma(p)) &= d_\infty(u, S(A,b)) \\ &\le H(A) \|(Au - b)_+\|_\infty \\ &= H(A) R_L(u,p) \ . \end{aligned}$$ □

We are now ready to prove Theorem 5 as desired.

Theorem 5. *Let polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, and link $\psi : \mathbb{R}^d \to \mathcal{R}$ be given. Then (L, ψ) is calibrated with respect to ℓ if and only if ψ is ϵ-separated with respect to L and ℓ for some $\epsilon > 0$.*

Proof. Let $\gamma = \text{prop}_{\mathcal{P}}[\ell]$ and $\Gamma = \text{prop}_{\mathcal{P}}[L]$. From Lemma 8, calibration implies ϵ-separation. For the converse, suppose ψ is ϵ-separated with respect to L and ℓ. Fix $p \in \Delta_{\mathcal{Y}}$. To show calibration, it suffices to find a positive lower bound for $R_L(u,p)$ that holds for all $u \in \mathbb{R}^d$ with $\psi(u) \notin \gamma(p)$.

Applying the definition of ϵ-separated and Lemma 9, $\psi(u) \notin \gamma(p)$ implies

$$\epsilon \le d_\infty(u, \Gamma(p)) \le H_{L,p} R_L(u,p) \implies 1 \le \frac{H_{L,p}}{\epsilon} R_L(u,p) \ .$$

Let $C_\ell = \max_{r,p} R_\ell(r,p)$. Then $R_\ell(\psi(u),p) \le C_\ell \le \frac{C_\ell H_{L,p}}{\epsilon} R_L(u,p)$.

If $H_{L,p} = 0$, then for all $u \in \mathbb{R}^d$ we have $R_\ell(\psi(u),p) = 0$, so calibration for this p is trivial. Similarly, if $C_\ell = 0$, then $R_\ell(r,p) = 0$ for all $r \in \mathcal{R}$, so again $R_\ell(\psi(u),p) = 0$ for all $u \in \mathbb{R}^d$.

Now assume $C_\ell > 0$ and $H_{L,p} > 0$. Let $C'_{\ell,p} \doteq \min_{r \notin \gamma(p)} R_\ell(r,p) > 0$. (As we assume $C_\ell > 0$, we must have $\gamma(p) \neq \mathcal{R}$, so the minimum is attained.) Then for all u such that $\psi(u) \notin \gamma(p)$, we have $R_\ell(\psi(u),p) \ge C'_{\ell,p}$. Rearranging, we have

$$\psi(u) \notin \gamma(p) \implies R_L(u,p) \ge \frac{C'_{\ell,p}\epsilon}{C_\ell H_{L,p}} > 0 \ .$$

Thus, $\inf_{u : \psi(u) \notin \gamma(p)} \langle L(u), p \rangle > \underline{L}(p)$. Since the above holds for all $p \in \Delta_{\mathcal{Y}}$, ψ is calibrated. □

3.7.3 Existence of a separated link

In this section, we prove Theorem 6, as discussed at the beginning of § 3.7.2.

We define some notation and assumptions to be used throughout this section. Let some norm $\|\cdot\|$ on finite-dimensional Euclidean space be given. Given a set T and a point u, let $d(T,u) = \inf_{t\in T} \|t-u\|$. Given two sets T, T', let $d(T,T') = \inf_{t\in T, t'\in T'} \|t-t'\|$. Finally, let the "thickening" $B(T,\epsilon)$ be defined as

$$B(T,\epsilon) = \{u \in \mathcal{R}' : d(T,u) < \epsilon\}.$$

Assumption 1. *$\ell : \mathcal{R}\times\mathcal{Y} \to \mathbb{R}^{\mathcal{Y}}_+$ is a loss on a finite report set $\mathcal{R}$, eliciting the property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$. It is embedded by $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^{\mathcal{Y}}_+$, which elicits the property $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathbb{R}^d$. The embedding points are $\{\varphi(r) : r \in \mathcal{R}\}$.*

Given Assumption 1, let $\mathcal{S} \subseteq 2^{\mathcal{R}}$ be defined as $\mathcal{S} = \{\gamma(p) : p \in \Delta_{\mathcal{Y}}\}$. In other words, for each p, we take the set of optimal reports $R = \gamma(p) \subseteq \mathcal{R}$, and we add R to $\mathcal{S}$. Let $\mathcal{U} \subseteq 2^{\mathbb{R}^d}$ be defined as $\mathcal{U} = \{\Gamma(p) : p \in \Delta_{\mathcal{Y}}\}$. For each $U \in \mathcal{U}$, let $R_U = \{r : \varphi(r) \in U\}$.

The next lemma shows that if a subset of $\mathcal{U}$ intersect, then their corresponding report sets intersect as well.

Lemma 10. *Let $\mathcal{U}' \subseteq \mathcal{U}$. If $\cap_{U\in\mathcal{U}'} U \neq \emptyset$ then $\cap_{U\in\mathcal{U}'} R_U \neq \emptyset$.*

Proof. Let $u \in \cap_{U\in\mathcal{U}'} U$. Our first claim is that there exists r such that $\Gamma_u \subseteq \gamma_r$. This follows from Proposition 9, which shows that $\mathrm{trim}(\Gamma) = \{\gamma_r : r \in \mathcal{R}\}$. Each Γ_u is either in $\mathrm{trim}(\Gamma)$ or is contained in some set in $\mathrm{trim}(\Gamma)$, by definition, proving the first claim. Our second claim is that $r \in \cap_{U\in\mathcal{U}'} R_U$, which proves the lemma. To prove the second claim, take any $U \in \mathcal{U}'$. There is some p such that $U = \Gamma(p)$, and we have in particular $p \in \Gamma_u$. By the first claim, $p \in \gamma_r$. By definition of embedding, $p \in \gamma_r \implies \varphi(r) \in \Gamma(p) = U$, so $r \in R_U$. □

Lemma 10 implies that there exists a ψ such that (L,ψ) indirectly elicits ℓ: for each u, let $\mathcal{U}' = \{U \in \mathcal{U} : u \in U\}$ be the optimal sets that contain it; choose r from the nonempty set $\cap_{U\in\mathcal{U}'} R_U$; and set $\psi(u) = r$.

The main problem now is to prove a "thickened" analogue of Lemma 10 that extends this link to points u that are up to ϵ far from an optimal set U. Namely, Lemma 13 will show that if ϵ

is small enough, then the ϵ-thickenings of all $U \in \mathcal{U}'$ intersect if and only if the U sets themselves intersect. Thus, if $u \in \cap_{U \in \mathcal{U}'} B(U, \epsilon)$, then $u \in \cap_{U \in \mathcal{U}'} U$, and Lemma 10 gives some legal target report $\psi(u) = r \in \cap_{U \in \mathcal{U}'} R_U$.

The next few geometric results build to Lemma 13. Then, the main proof will be completed as we have just sketched.

Lemma 11. *Let D be a closed, convex polyhedron in $\mathbb{R}^d$. For any $\epsilon > 0$, there exists an open, convex set D', the intersection of a finite number of open halfspaces, such that*

$$D \subseteq D' \subseteq B(D, \epsilon).$$

Proof. Let S be the standard open ϵ-ball $B(\{\vec{0}\}, \epsilon)$. Note that $B(D, \epsilon) = D + S$ where $+$ is the Minkowski sum. Now let $S' = \{u : \|u\|_1 \leq \delta\}$ be the closed δ ball in L_1 norm. By equivalence of norms in Euclidean space [14, Appendix A.1.4], we can take δ small enough yet positive such that $S' \subseteq S$. By standard results, the Minkowski sum of two closed, convex polyhedra, $D'' = D + S'$ is a closed polyhedron, i.e. the intersection of a finite number of closed halfspaces. (A proof: we can form the higher-dimensional polyhedron $\{(x, y, z) : x \in D, y \in S', z = x + y\}$, then project onto the z coordinates.)

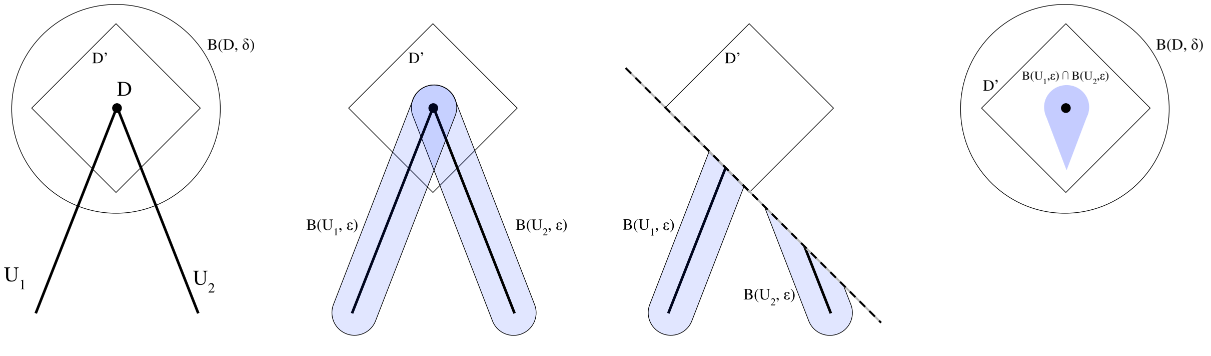
Now, if $T' \subseteq T$, then the Minkowksi sum satisfies $D + T' \subseteq D + T$. In particular, because $\emptyset \subseteq S' \subseteq S$, we have

$$D \subseteq D'' \subseteq B(D, \epsilon).$$

Now let D' be the interior of D'', i.e. if $D'' = \{x : Ax \leq b\}$, then we let $D' = \{x : Ax < b\}$. We retain $D' \subseteq B(D, \epsilon)$. Further, we retain $D \subseteq D'$, because D is contained in the interior of $D'' = D + S'$. (Proof: if $x \in D$, then for some γ, $x + B(\{\vec{0}\}, \gamma) = B(x, \gamma)$ is contained in $D + S'$.) This proves the lemma. □

Lemma 12. *Let $\{U_j : j \in \mathcal{J}\}$ be a finite collection of closed, convex sets with $\cap_{j \in \mathcal{J}} U_j \neq \emptyset$. Let $\delta > 0$ be given. Then there exists $\epsilon_0 > 0$ such that, for all $0 < \epsilon \leq \epsilon_0$, $\cap_j B(U_j, \epsilon) \subseteq B(\cap_j U_j, \delta)$.*

Figure 3.6: Illustration of a special case of the proof of Lemma 12 where there are two sets U_1, U_2 and their intersection D is a point. We build the polyhedron D' inside $B(D, \delta)$. By considering each halfspace that defines D', we then show that for small enough ϵ, $B(U_1, \epsilon)$ and $B(U_2, \epsilon)$ do not intersect outside D'. So the intersection is contained in D', so it is contained in $B(D, \delta)$.



Proof. We induct on $|\mathcal{J}|$. If $|\mathcal{J}| = 1$, set $\epsilon = \delta$. If $|\mathcal{J}| > 1$, let $j \in \mathcal{J}$ be arbitrary, let $U' = \cap_{j' \neq j} U_{j'}$, and let $C(\epsilon) = \cap_{j' \neq j} B(U_{j'}, \epsilon)$. Let $D = U_j \cap U'$. We must show that $B(U_j, \epsilon) \cap C(\epsilon) \subseteq B(D, \delta)$. By Lemma 11, we can enclose D strictly within a polyhedron D', the intersection of a finite number of open halfspaces, which is itself strictly enclosed in $B(D, \delta)$. (For example, if D is a point, then enclose it in a hypercube, which is enclosed in the ball $B(D, \delta)$.) We will prove that, for all small enough ϵ, $B(U_j, \epsilon) \cap C(\epsilon)$ is contained in D'. This implies that it is contained in $B(D, \delta)$.

For each halfspace defining D', consider its complement F, a closed halfspace. We prove that $F \cap B(U_j, \epsilon) \cap C(\epsilon) = \emptyset$. Consider the intersections of F with U and U', call them G and G'. These are closed, convex sets that do not intersect (because D in contained in the complement of F). So G and G' are separated by a nonzero distance, so $B(G, \gamma) \cap B(G', \gamma) = \emptyset$ for all small enough γ. And $B(G, \gamma) = F \cap B(U_j, \gamma)$ while $B(G', \gamma) = F \cap B(U', \gamma)$. This proves that $F \cap B(U_j, \gamma) \cap B(U', \gamma) = \emptyset$. By inductive assumption, $C(\epsilon) \subseteq B(U', \gamma)$ for small enough $\epsilon = \epsilon_F$. So $F \cap B(U_j, \gamma) \cap C(\epsilon) = \emptyset$. We now let ϵ_0 be the minimum over these finitely many ϵ_F (one per halfspace). □

Lemma 13. *Let $\{U_j : j \in \mathcal{J}\}$ be a finite collection of nonempty closed, convex sets with $\cap_{j \in \mathcal{J}} U_j = \emptyset$. Then there exists $\epsilon_0 > 0$ such that, for all $0 < \epsilon \leq \epsilon_0$, $\cap_{j \in \mathcal{J}} B(U_j, \epsilon) = \emptyset$.*

Proof. By induction on the size of the family. Note that the family must have size at least two. Let

U_j be any set in the family and let $U' = \cap_{j' \neq j} U_{j'}$. There are two possibilities.

The first possibility, which includes the base case where the size of the family is two, is the case U' is nonempty. Because U_j and U' are non-intersecting closed convex sets, they are separated by some distance δ. So $B(U_j, \delta/3) \cap B(U', \delta/3) = \emptyset$. By Lemma 12, there exists $\epsilon_0' > 0$ such that $\cap_{j' \neq j} B(U_{j'}, \epsilon) \subseteq B(U', \delta/3)$ for all $0 < \epsilon \leq \epsilon_0'$. Pick $\epsilon_0 = \min\{\epsilon_0', \delta/3\}$. Then for all $0 < \epsilon \leq \epsilon_0$, the intersection of ϵ-thickenings is contained in the $(\delta/3)$-thickening of the intersection, which is disjoint from the $(\delta/3)$-thickening of U_j, which contains the ϵ-thickening of U_j.

The second possibility is that U' is empty. This implies we are not in the base case, as the family must have three or more sets. By inductive assumption, for all small enough ϵ we have $\cap_{j' \neq j} B(U_{j'}, \epsilon) = \emptyset$, which proves this case. □

Corollary 5. *There exists $\epsilon_0 > 0$ such that, for any $0 < \epsilon \leq \epsilon_0$, for any subset $\{U_j : j \in \mathcal{J}\}$ of $\mathcal{U}$, if $\cap_j U_j = \emptyset$, then $\cap_j B(U_j, \epsilon) = \emptyset$.*

Proof. For each subset, Lemma 13 gives an $\epsilon_0 > 0$. We take the minimum over these finitely many subsets of $\mathcal{U}$. □

Theorem 11. *For all small enough ϵ, the epsilon-thickened link ψ (Construction 1) is a well-defined link function from $\mathcal{R}'$ to $\mathcal{R}$, i.e. $\psi(u) \neq \bot$ for all u.*

Proof. Fix a small enough ϵ as promised by Corollary 5. Consider any $u \in \mathcal{R}'$. If u is not in $B(U, \epsilon)$ for any $U \in \mathcal{U}$, then we have $\Psi(u) = \mathcal{R}$, so it is nonempty. Otherwise, let $\{U_j : j \in \mathcal{J}\}$ be the family whose thickenings intersect at u. By Corollary 5, because of our choice of ϵ, the family themselves has nonempty intersection. By Lemma 10, their corresponding report sets $\{R_j : j \in \mathcal{J}\}$ also intersect at some r, so $\Psi(u)$ is nonempty. □

Theorem 6, which we restate here, is now almost immediate.

Theorem 6. *Let polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ embed the discrete loss $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$. Then there exists $\epsilon_0 > 0$ such that, for all $0 < \epsilon \leq \epsilon_0$, Construction 1 yields an ϵ-separated link with respect to L and ℓ.*

Proof. We create ψ using Construction 1 with the L_∞ norm. By Theorem 11, for all small enough ϵ, ψ is well-defined everywhere.

To prove separation, suppose u and p are given such that $d_\infty(u, U) \leq \epsilon$, where $U = \Gamma(p)$. Then in Construction 1, $\psi(u) \in \Psi(u) \subseteq R_U = \{r : \varphi(r) \in U\}$. By definition of embedding, $\varphi(r) \in U = \Gamma(p) \implies r \in \gamma(p)$. So we obtain $\psi(u) \in \gamma(p)$ whenever $d_\infty(u, \Gamma(p)) \leq \epsilon$, which proves ϵ-separation of the link ψ. □

3.7.4 Connecting losses with finite representative sets to the theory of convex polyhedra

3.7.4.1 Definitions and preliminaries

Definition 24 (Closed halfspace)**.** *A closed halfspace is a set of the form* $H^+_{(w,b)} := \{x \in \mathbb{R}^d \mid \langle x, w\rangle \geq b\}$ *for any* $(w, b) \in \mathbb{R}^d \times \mathbb{R}$*.*

Definition 25 (Hyperplane)**.** *A hyperplane is a set of the form* $H_{(w,b)} := \{x \in \mathbb{R}^d \mid \langle x, w\rangle = b\}$ *for any* $(w, b) \in \mathbb{R}^d \times \mathbb{R}$*.*

Observe that $H_{(w,b)} = \partial H^+_{(w,b)}$, meaning the hyperplane $H_{(w,b)}$ is the boundary of $H^+_{(w,b)}$. Thus, for any halfspace H^+, we have that H^+ is one of the two halfspaces corresponding to the hyperplane $\partial H^+ = H$.

Definition 26 (Polyhedron halfspace representation)**.** *A polyhedron P is an intersection of a finite set of closed halfspaces $\mathcal{H}$ presented in the form $P = \cap\mathcal{H}$.*

Observe that by the halfspace representation, a polyhedron need not be bounded.

Definition 27 (Valid, Supports)**.** *A halfspace H^+ is valid for P if $P \subseteq H^+$. A hyperplane H supports the polyhedron P if (i) $P \subseteq H^+$ for a halfspace H^+ with $H = \partial H^+$, and (ii)$H \cap \partial P \neq \emptyset$. Moreover, H supports P at x if $x \in H \cap \partial P$.*

Definition 28 (Face, facet)**.** *Let $P \subseteq \mathbb{R}^d$ be a convex polyhedron. A (non-trivial) face F of the polytope P is any set of the form*

$$F = P \cap H \ ,$$

for a hyperplane H supporting P. The dimension of a face F is the dimension of its affine hull $\dim(F) := \dim(\mathrm{affhull}(F))$*. A face F with* $\dim(F) = \dim(\mathrm{affhull}(P)) - 1$ *is called a facet.*

Observe that P is a trivial face of itself, and cannot be written by the above definition. Throughout, we restrict our focus to non-trivial faces, and omit mentioning non-trivial henceforth.

Claim 1. *A face F of the polyhedron P such that $F = P \cap H$ is nonempty if and only if H is a supporting hyperplane of P.*

It is often useful to understand polyhedra in terms of their halfspace representations and the set of hyperplanes generating facets of P. To find this set, we must first establish when a halfspace representation is irredundant for a given polyhedron.

Definition 29 (Gallier [43])**.** *Let $P = \cap\mathcal{H}$ for a finite set of halfspaces $\mathcal{H}$ be a polyhedron. We say that $\cap\mathcal{H}$ is an irredundant decomposition for P (and $\mathcal{H}$ is irredundant for P) if P cannot be expressed as $P = \cap\mathcal{H}'$ for some $\mathcal{H}'$ such that $|\mathcal{H}'| < |\mathcal{H}|$.*

Gallier [43] shows that every full-dimensional (i.e. $\dim(\mathrm{affhull}(P)) = d$) polyhedron $P \subseteq \mathbb{R}^d$ has a unique and irredundant halfspace representation $\mathcal{H}^*$, and each $H^+ \in \mathcal{H}^*$ generates a facet of P.

Theorem 12. *Given a d-dimensional polyhedron $P \subseteq \mathbb{R}^d$, (i) there is a unique irredundant and finite set of closed halfspaces $\mathcal{H}^*$ such that $P = \cap\mathcal{H}^*$, (ii) $\{H \cap P \mid H^+ \in \mathcal{H}^*, H = \partial H^+\}$ is the set of facets of P, and (iii) for all finite sets of closed halfspaces $\mathcal{H}$ such that $P = \cap\mathcal{H}$, we have $\mathcal{H}^* \subseteq \mathcal{H}$, .*

Proof. Since P is d-dimensional in $\mathbb{R}^d$, it therefore has nonempty interior. We claim that P must have some irredundant representation $P = \cap\mathcal{H}$ for a finite set $\mathcal{H}$. As P has a finite halfspace representation, it must have a smallest halfspace representation $\mathcal{H}^*$. That is, $|\mathcal{H}^*| = \min\{\cap\mathcal{H} \mid P = \cap\mathcal{H}, \mathcal{H} \text{ finite}\}$. As the smallest halfspace representation, $\mathcal{H}^*$ is irredundant; if it was redundant, this would imply there is a smaller representation $\mathcal{H}'$ so that $P = \cap\mathcal{H}'$ and $|\mathcal{H}'| < |\mathcal{H}^*|$, contradicting $\mathcal{H}^*$ as the smallest representation. Gallier [43, Proposition 4.5(i)] then states that the irredundant representation $\mathcal{H}^*$ is unique up to ordering, allowing us to conclude (i). Additionally, (ii) is shown by [43, Proposition 4.5(ii)].

It is just left to show (iii). By (i), we know that each $H^+ \in \mathcal{H}^*$ uniquely determines a facet of P. Moreover, by (ii), define $F := P \cap H$, where $H = \partial H^+$ and $H^+ \in \mathcal{H}^*$, which is a facet of P and of dimension $d-1$. The facet F can then be defined by d affinely independent points (contained in P), whose affine hull is H. As halfspaces are uniquely determined, so is the facet $F = P \cap H$. As polyhedron are uniquely determined by their facets (by Minkowski's uniqueness theorem, cf., [54]), we must have $H^+ \in \mathcal{H}^*$.

□

3.7.4.2 Notation

Within this appendix, we use some self-contained notation. We will later consider losses over a finite set of outcomes $\mathcal{Y}$; to make notation consistent, we use $\mathbb{R}^{\mathcal{Y}}_+$ throughout as shorthand for $\mathbb{R}^{|\mathcal{Y}|}_+$, and let $d := |\mathcal{Y}| + 1$.

Fix a set $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$, and consider the concave function $g_{\mathcal{V}} : x \mapsto \inf_{v\in\mathcal{V}}\langle v, x\rangle - \delta(x \mid \mathbb{R}^{\mathcal{Y}}_+)$. We denote the hypograph of $g_{\mathcal{V}}$ by $\text{hypo}(g_{\mathcal{V}}) = \{(x, c) \mid c \leq g(x)\} \subseteq \mathbb{R}^{\mathcal{Y}}_+ \times \mathbb{R}$.

Given any $v \in \mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$, define $H^+_v := H^+_{(v,-1)} = \{(x, c) \in \mathbb{R}^{\mathcal{Y}}_+ \times \mathbb{R} \mid \langle v, x\rangle = c\}$. Similarly, we denote $H^+_y := H^+_{(e_y,0)}$ for any $y \in \mathcal{Y}$; the latter will help us restrict a constructed polyhedron to the nonnegative orthant. Extending to hyperplanes, we construct $H_v := H_{(v,-1)}$ and observe that $H_v = \partial H^+_v$ for $v \in \mathbb{R}^{\mathcal{Y}}_+$ and define $H_y := H_{(e_y,0)}$ so that $H_y = \partial H^+_y$. Given a polyhedron P, we denote the face $F^P_v := H_v \cap P$. If P is understood from context, we simply denote this face F_v.

Finally, given a set $\mathcal{V} \subseteq \mathbb{R}^d$, we let $\mathcal{H}_{\mathcal{V}} = \{H^+_v \mid v \in \mathcal{V}\}$ denote the set of halfspaces generated by $\mathcal{V}$, $\mathcal{H}_{\mathcal{Y}} = \{H^+_y \mid y \in \mathcal{Y}\}$. If $\mathcal{V}$ and $\mathcal{Y}$ are understood from context, we may denote $\mathcal{H} := \mathcal{H}_{\mathcal{V}} \cup \mathcal{H}_{\mathcal{Y}}$.

3.7.4.3 Finitely generated polyhedron

Throughout, we will work with a (minimizable) function $g_{\mathcal{V}}$ generated by a set $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$ of the following form.

Definition 30. *Given a set $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$, define the function $g_{\mathcal{V}} : \mathbb{R}^{\mathcal{Y}}_+ \to \mathbb{R}_+$ by*

$$g_{\mathcal{V}}(x) = \inf_{v \in \mathcal{V}} \langle x, v \rangle - \delta(x \mid \mathbb{R}^{\mathcal{Y}}_+) \ .$$

We first observe that the region generated by the intersection of the H^+_y halfspaces restricts the hypograph of any $g_{\mathcal{V}}$ to be finite only on the nonnegative orthant.

Lemma 14. $\cap \mathcal{H}_{\mathcal{Y}} = \mathbb{R}^{\mathcal{Y}}_+ \times \mathbb{R}$.

Proof. The result follows if we show $x \in \mathbb{R}^{\mathcal{Y}}_+ \iff (x, c) \in \cap \mathcal{H}_{\mathcal{Y}}$ for all $c \in \mathbb{R}$.

$\implies$ Fix any $c \in \mathbb{R}$. $x \in \mathbb{R}^{\mathcal{Y}}_+ \iff x_y \geq 0$ for all $y \in \mathcal{Y}$. This means that for any $y \in \mathcal{Y}$, $(x, c) \in \{(x, c) \mid x_y \geq 0\} = H^+_y$. As y and c were arbitrary, this shows the forward direction.

$\impliedby$ $(x, c) \in \cap \mathcal{H}_{\mathcal{Y}}$ implies $x_y \geq 0$ for all $y \in \mathcal{Y}$, and therefore $x \in \mathbb{R}^d_+$. □

3.7.4.4 Infinitely generated polyhedra with finite representation

We now contextualize the setting of the previous section. Suppose $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ is a minimizable loss function. For $x \in \mathbb{R}^{\mathcal{Y}}_+$, consider the 1-homogeneous extension of Bayes risk $\underline{L}_+(x) := \inf_{r \in \mathcal{R}} \langle x, L(r) \rangle - \delta(x \mid \mathbb{R}^{\mathcal{Y}}_+)$, which we assume is polyhedral throughout.

We now consider $L(\mathcal{R}) \subseteq \mathbb{R}^{\mathcal{Y}}_+$ and $\mathcal{H}_{L(\mathcal{R})} = \{H^+_v \mid v \in L(\mathcal{R})\}$. Observe that $L(\mathcal{R})$ and $\mathcal{H}_{L(\mathcal{R})}$ may be infinitely generated sets. Now let $\mathcal{H} = \mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{L(\mathcal{R})}$; again, this may be infinitely generated. We find the existence of a finite $\mathcal{V} \subseteq L(\mathcal{R})$ such that $g_{\mathcal{V}} = g_{L(\mathcal{R})}$, and proceed to work with such a $\mathcal{V}$.

Claim 2. *Given a minimizable $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ with polyhedral extended risk $\underline{L}_+$, consider $\mathcal{H} = \mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{L(\mathcal{R})}$. Then* $\mathrm{hypo}(g_{L(\mathcal{R})}) = \cap \mathcal{H}$*. Moreover, there is a finite $\mathcal{V} \subseteq L(\mathcal{R})$ such that $g_{L(\mathcal{R})} = g_{\mathcal{V}}$.*

Proof. Observe that $x \in \mathbb{R}^{\mathcal{Y}}_{+} \iff (x, c) \in \cap \mathcal{H}_{\mathcal{Y}}$. Let $x \in \mathbb{R}^{\mathcal{Y}}_{+}$.

$$\begin{aligned}
(x, c) \in \operatorname{hypo}(g_{L(\mathcal{R})}) &\iff g_{L(\mathcal{R})}(x) \geq c && \text{Definition of hypograph;} \\
&\iff \underline{L}_{+}(x) \geq c && g_{L(\mathcal{R})} = \underline{L}_{+} \text{ for } x \in \mathbb{R}^{\mathcal{Y}}_{+}; \\
&\iff \langle v, x \rangle \geq c \; \forall v \in L(\mathcal{R}) && \text{by def of } \underline{L}_{+} \text{ as the infimum over } v \in L(\mathcal{R}) \text{ of} \\
& && \text{the inner product with } x \text{ and minimizable;} \\
&\iff (x, c) \in H^{+}_{v} \; \forall v \in L(\mathcal{R}) && \text{by definition of each halfspace;} \\
&\iff (x, c) \in \cap \mathcal{H}_{L(\mathcal{R})} && \text{since true for all } v \in L(\mathcal{R}).
\end{aligned}$$

Combining the two equalities (e.g., $\cap \mathcal{H} = (\cap \mathcal{H}_{\mathcal{Y}}) \cap (\cap \mathcal{H}_{L(\mathcal{R})})$), we have $\operatorname{hypo}(g_{L(\mathcal{R})}) = \cap \mathcal{H}$. The existence of a finite $\mathcal{V} \subseteq L(\mathcal{R})$ follows as $\underline{L}_{+}$ polyhedral implies $\operatorname{hypo}(\underline{L}_{+})$ is a polyhedron, which has a finite representation by definition. □

This claim allows us to proceed while considering the finite set $\mathcal{V}$ rather than the full range $L(\mathcal{R})$. We now evaluate the structure of $g_{L(\mathcal{R})} = g_{\mathcal{V}}$ and its hypograph through $\mathcal{V}$.

3.7.4.5 Structure of $g_{\mathcal{V}}$

We will assume $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_{+}$ is finite; if $L(\mathcal{R})$ is finite, then take $\mathcal{V} = L(\mathcal{R})$. Otherwise, take any finite $\mathcal{V} \subseteq L(\mathcal{R})$ as in Claim 2. Now, we can define $\operatorname{hypo}(g_{\mathcal{V}})$ as the intersection of halfspaces generated by $\mathcal{V}$ on the nonnegative orthant.

Claim 3. *Given a finite set $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_{+}$, define $\mathcal{H} = \mathcal{H}_{\mathcal{V}} \cup \mathcal{H}_{\mathcal{Y}}$. Then $\operatorname{hypo}(g_{\mathcal{V}}) = \cap \mathcal{H}$.*

Proof. $(x, c) \in \operatorname{hypo}(g_{\mathcal{V}}) \iff g(x) - c \geq 0 \iff \min_{v \in \mathcal{V}} \langle v, x \rangle - c \geq 0$ and $x \in \mathbb{R}^{\mathcal{Y}}_{+}$, which is true if and only if $\langle v, x \rangle - c \geq 0 \, \forall v \in \mathcal{V}$ and $x_y \geq 0$ for all y. In turn, this statement holds if and only if $(x, c) \in H^{+}_{v}$ for all $v \in \mathcal{V}$ and in H^{+}_{y} for all $y \in \mathcal{Y}$, so $(x, c) \in \cap \mathcal{H}$. □

We proceed with some observations about facets and dimension of $\operatorname{hypo}(g_{\mathcal{V}})$ in order to finite the *smallest* halfspace representation for $g_{L(\mathcal{R})} = g_{\mathcal{V}}$.

Lemma 15. *Given a finite, nonempty set $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_{+}$, $\operatorname{hypo}(g_{\mathcal{V}})$ is d-dimensional.*

Proof. Since $g_{\mathcal{V}}$ is nonnegative on $\mathbb{R}^{\mathcal{Y}}_+$, $\operatorname{hypo}(g_{\mathcal{V}})$ therefore contains $\{(x, c) \mid x \in \mathbb{R}^{\mathcal{Y}}_+, c \leq 0\}$, which is $(|\mathcal{Y}| + 1)$-dimensional. Recall $d = (|\mathcal{Y}| + 1)$. □

Lemma 15 allows us to apply Theorem 12 to observe a unique set of halfspaces $\mathcal{H}^*$ generating $\operatorname{hypo}(g_{\mathcal{V}})$.

Lemma 16. *Given a finite set $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_+$, define $\mathcal{H} = \mathcal{H}_{\mathcal{V}} \cup \mathcal{H}_{\mathcal{Y}}$. There is some unique $\mathcal{H}^* \subseteq \mathcal{H}$ such that $\operatorname{hypo}(g_{\mathcal{V}}) = \cap\mathcal{H}^*$. Moreover, for each $H^+ \in \mathcal{H}^*$ and H such that $H = \partial H^+$, the face $F = \operatorname{hypo}(g_{\mathcal{V}}) \cap H$ is a facet.*

Proof. Since $\operatorname{hypo}(g_{\mathcal{V}})$ is full-dimensional by Lemma 15, this follows immediately from Theorem 12(i) and (iii). □

We now show that the set $\mathcal{H}_{\mathcal{Y}}$ is contained in $\mathcal{H}^*$ so that we can separate the facets generated by $\mathcal{H}^*$ into a partition of vertical and non-vertical facets of $\operatorname{hypo}(g_{\mathcal{V}})$.

Lemma 17. *Given a finite set $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_+$, consider the unique finite set $\mathcal{H}^*$ as given by Lemma 16. $\mathcal{H}_{\mathcal{Y}} \subseteq \mathcal{H}^*$.*

Proof. If there was a $y \in \mathcal{Y}$ such that $H_y^+ = \{(x, c) \mid x_y \geq 0\}$ was not in $\mathcal{H}^*$, then we would either have some $c_1 > 0$ such that $\{(x, c) \mid x_y \geq c_1\} \in \mathcal{H}^*$, or we have a point x such that $x_y < 0$ but $g(x) > -\infty$. The first cannot happen as we take $g_{\mathcal{V}}$ is finite at $x = e_y \in \mathbb{R}^{\mathcal{Y}}_+$ and is concave. Moreover, the second cannot be true by construction of $g_{\mathcal{V}}$ since $\mathcal{V}$ is finite including the $0 - \infty$ indicator on $\mathbb{R}^{\mathcal{Y}}_+$. □

Corollary 6. *Suppose we are given a finite set $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_+$, and consider the unique irredundant set $\mathcal{H}^*$ given by Lemma 16. There is a unique finite set $\mathcal{V}^* \subseteq \mathbb{R}^{\mathcal{Y}}_+$ such that $\mathcal{H}^* = \mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{\mathcal{V}^*}$. Moreover, F_v is a facet of $\operatorname{hypo}(g_{\mathcal{V}})$ for each $v \in \mathcal{V}^*$.*

Proof. Since $\operatorname{hypo}(g_{\mathcal{V}})$ is full-dimensional, the facets of $\operatorname{hypo}(g_{\mathcal{V}})$ are uniquely determined by the hyperplanes H such that $H = \partial H^+$ and $\mathcal{H}^* = \{H^+\}$ by Lemma 16. Any facet must then be some intersection of an $H_y \cap \operatorname{hypo}(g_{\mathcal{V}})$ or $H_v \cap \operatorname{hypo}(g_{\mathcal{V}})$. Consider $\mathcal{H}_{\mathcal{V}^*} := \mathcal{H}^* \setminus \mathcal{H}_{\mathcal{Y}}$, and $\mathcal{V}^*$ the unique set generating $\mathcal{H}_{\mathcal{V}^*}$, since $\mathcal{H}_{\mathcal{Y}} \subseteq \mathcal{H}^*$ by Lemma 17. ($\mathcal{V}^*$ is unique as halfspaces are uniquely determined.) Moreover, $H_v \in \mathcal{H}_{\mathcal{V}^*} \subseteq \mathcal{H}^*$ generates the facet F_v of $\text{hypo}(g_{\mathcal{V}})$ by Lemma 16. □

Corollary 7. *Let $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_+$ be a finite set, and $\mathcal{V}^*$ the unique finite set from Corollary 6 such that $\mathcal{H}^* = \mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{\mathcal{V}^*}$. Then $\mathcal{V}^* \subseteq \mathcal{V}$.*

Proof. $\mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{\mathcal{V}^*} = \mathcal{H}^*$ by Corollary 6, and $\mathcal{H}^* \subseteq \mathcal{H} = \mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{\mathcal{V}}$ by Lemma 17, ergo $\mathcal{V}^* \subseteq \mathcal{V}$. □

Thus, we will introduce our first assumption for the existence of $\mathcal{V}^*$ given a finite $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_+$.

We now show that we can equivalently construct $g_{\mathcal{V}}$ through the unique finite set $\mathcal{V}^*$ instead of the given set of vectors $\mathcal{V}$, and in turn, loss vectors $L(\mathcal{R})$.

Lemma 18. *Given a finite set $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_+$, consider $\mathcal{V}^* \subseteq \mathcal{V}$ as in Corollary 6. Then $g_{\mathcal{V}}(x) = \min_{v \in \mathcal{V}^*} \langle v, x\rangle - \delta(x \mid \mathbb{R}^{\mathcal{Y}}_+) = g_{\mathcal{V}^*}(x)$*

Proof. The result holds if $\text{hypo}(g_{\mathcal{V}}) = \text{hypo}(g_{\mathcal{V}^*})$. By construction, $\text{hypo}(g_{\mathcal{V}}) = \cap(\mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{\mathcal{V}}) = \cap\mathcal{H}^* = \cap(\mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{\mathcal{V}^*}) = \{(x, c) \in \mathbb{R}^{\mathcal{Y}}_+ \times \mathbb{R} \mid \langle v^*, x\rangle \geq c \text{ for all } v^* \in \mathcal{V}^*\}$ where the first equality follows as $\mathcal{H}^* \subseteq \mathcal{H}$. This means $g_{\mathcal{V}}$ can be written as $g_{\mathcal{V}}(x) = \min_{v \in \mathcal{V}^*} \langle v, x\rangle - \delta(x \mid \mathbb{R}^{\mathcal{Y}}_+) = g_{\mathcal{V}^*}(x)$. □

Corollary 8. *Given minimizable $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ such that $\underline{L}_+$ is polyhedral, there exists a (unique) finite $\mathcal{V}^* \subseteq L(\mathcal{R})$ such that $g_{\mathcal{V}} = g_{\mathcal{V}^*}$ and $\text{hypo}(g_{L(\mathcal{R})}) = \text{hypo}(g_{\mathcal{V}^*}) = \cap(\mathcal{H}_{\mathcal{V}^*} \cup \mathcal{H}_{\mathcal{Y}})$ is irredundant.*

This paves the way for our primary assumption for the rest of this appendix.

Assumption 2. *$L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ is a minimizable loss function such that $\underline{L}_+ = g_{L(\mathcal{R})}$ is polyhedral. $\mathcal{V} \subseteq L(\mathcal{R})$ is a finite set such that $g_{\mathcal{V}} = g_{L(\mathcal{R})}$. Finally, $\mathcal{V}^* \subseteq \mathcal{V} \subseteq L(\mathcal{R})$ is the (unique) finite irredundant set such that $g_{L(\mathcal{R})} = g_{\mathcal{V}} = g_{\mathcal{V}^*}$ and $\text{hypo}(g_{L(\mathcal{R})}) = \text{hypo}(g_{\mathcal{V}}) = \text{hypo}(g_{\mathcal{V}^*}) = \cap(\mathcal{H}_{\mathcal{V}^*} \cup \mathcal{H}_{\mathcal{Y}})$, the last of which is irredundant.*

With this assumption in hand, we can show a few more statements involving $\mathcal{V}^*$ and how it relates to $\mathcal{V}$. For intuition, the construction of $\mathcal{V}^*$ will be helpful to consider as a minimum representative set in the proof of Lemma 5.

Claim 4. *Given $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2, there is a finite set $\mathcal{R}^* \subseteq \mathcal{R}$ such that $L(\mathcal{R}^*) = \mathcal{V}^*$ (without duplicates).*

Proof. This follows immediately from the Assumption 2 as $\mathcal{V}^* \subseteq L(\mathcal{R})$. □

Claim 5. *Given $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2, for all $x \in \mathbb{R}^{\mathcal{Y}}_+$, there exists $v^* \in \mathcal{V}^*$ such that H_{v^*} supports* $\text{hypo}(g_{\mathcal{V}})$ *at $(x, g_{\mathcal{V}}(x))$.*

Proof. By Assumption 2, we have $g_{\mathcal{V}}(x) = g_{\mathcal{V}^*}(x) = \inf_{v \in \mathcal{V}^*} \langle v, x \rangle = \min_{v \in \mathcal{V}^*} \langle v, x \rangle$ for $x \in \mathbb{R}^{\mathcal{Y}}_+$. In particular, consider a normal $v^* \in \arg\min_{v \in \mathcal{V}^*} \langle v, x \rangle$; we claim that the hyperplane H_{v^*} such that $H_{v^*} = \partial H^+_{v^*}$ supports $\text{hypo}(g_{\mathcal{V}})$ at $(x, g(x))$. First, $\text{hypo}(g_{\mathcal{V}}) \subseteq H^+_{v^*}$ by definition of $\text{hypo}(g_{\mathcal{V}})$ as the intersection of halfspaces including $H^+_{v^*}$. Thus, it is just left to show that $(x, \langle v^*, x \rangle) \in H_{v^*} \cap \text{hypo}(g_{\mathcal{V}})$. By definition of $g_{\mathcal{V}}$, we have $g_{\mathcal{V}}(x) = g_{\mathcal{V}^*}(x) = \langle v^*, x \rangle$, so $(x, g_{\mathcal{V}}(x)) \in H_{v^*}$. Moreover, $(x, g_{\mathcal{V}}(x)) \in \text{hypo}(g_{\mathcal{V}}) = \{(x, c) \mid g_{\mathcal{V}}(x) \geq c\}$ trivially since $g_{\mathcal{V}}(x) \geq g_{\mathcal{V}}(x)$. □

3.7.4.6 Projecting from $\mathbb{R}^d_+$ to $\mathbb{R}^{\mathcal{Y}}_+$

We now define the projection $\pi : \mathbb{R}^{\mathcal{Y}} \times \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}, (x, c) \mapsto x$. The projected faces generated by $\mathcal{V}^*$ cover the nonnegative orthant.

Corollary 9. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For each $v \in \mathcal{V}^*$, let $F_v = F_v^{\text{hypo}(g_{\mathcal{V}^*})} = H_v \cap \text{hypo}(g_{\mathcal{V}^*})$. Then $\cup_{v \in \mathcal{V}^*} \pi(F_v) = \mathbb{R}^{\mathcal{Y}}_+$.*

Moreover, the projection π preserves dimension of faces.

Claim 6. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For all $v \in L(\mathcal{R})$, define F_v as the face of $g_{\mathcal{V}}$ generated by $v \in \mathbb{R}^{\mathcal{Y}}_+$. Then* $\dim(F_v) = \dim(\pi(F_v))$.

Proof. Recall from Definition 28 that the dimension of a polytope to be the dimension of its affine hull. Suppose we are given $|\mathcal{Y}| + 1$ affinely independent vectors z_i in F_v. We claim their projections $\{\pi(z_i)\}$ are affinely independent. Let $a_1 + \ldots + a_{|\mathcal{Y}|+1} = 0$, such that $\sum_i a_i \pi(z_i) = 0$. We want to conclude that we must have $a_i = 0$ for all i, meaning they are affinely independent.

Observe $z_i = (x_i, \langle v, x_i \rangle)$ for all i; therefore, if $z_i \in F_v$ (e.g., F_v supports $\text{hypo}(g_{\mathcal{V}})$ at $(x, \langle v, x \rangle))$, then we also have $z_i \in H_v$. So $0 = \sum_i a_i \pi(z_i) = \sum_i a_i x_i$. Moreover, the sum $\sum_i a_i z_i = \sum_i a_i (x_i, \langle v, x_i \rangle) = (\sum_i a_i x_i, \langle v, \sum_i a_i x_i \rangle) = (\vec{0}, 0) = \vec{0}$. Thus, since $a_i = 0$ for all i, the set $\{z_i\}$ is affinely independent and the dimensions of the affine hulls are therefore equal. □

Since we preserve the dimension of these projected spaces, we can now study equivalence of projected faces of the hypograph and regions of support of $g_{\mathcal{V}}$ for any $v \in \mathcal{L}(\mathcal{R})$.

Lemma 19. *Given $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2, fix $x \in \mathbb{R}^{\mathcal{Y}}_+$. For any $v \in L(\mathcal{R})$, the following are equivalent:*

(1) $(x, g_{\mathcal{V}}(x)) \in F_v^{\text{hypo}(g_{\mathcal{V}})}$;

(2) $\langle v, x \rangle = g_{\mathcal{V}}(x)$;

(3) $v \in \arg\min_{v' \in \mathcal{V}} \langle v', x \rangle$; and

(4) $x \in \pi(F_v^{\text{hypo}(g_{\mathcal{V}})})$.

Proof. For $v \in L(\mathcal{R})$, define $F_v := F_v^{\text{hypo}(g_{\mathcal{V}})}$.

$$\begin{aligned}
(1) \qquad (x, g(x)) \in F_v &\iff (x, g_{\mathcal{V}}(x)) \in \{(x', c) \in \text{hypo}(g_{\mathcal{V}}) \mid \langle v, x' \rangle = c\} \\
&\iff \langle v, x \rangle = g_{\mathcal{V}}(x) && (2) \\
&\iff \langle v, x \rangle = \min_{v' \in \mathcal{V}} \langle v', x \rangle \\
&\iff v \in \arg\min_{v' \in \mathcal{V}} \langle v', x \rangle \ . && (3)
\end{aligned}$$

This covers $1 \iff 2 \iff 3$.

For $1 \iff 4$, the forward implication follows trivially by applying the definition of the projection π. For the reverse implication, consider some $x \in \pi(F_v)$. There must be a $c \in \mathbb{R}$ so that $(x, c) \in F_v$. Expanding, this is actually saying $(x, c) \in \{(x', c') \in \text{hypo}(g_{\mathcal{V}}) \mid \langle v, x' \rangle = c\}$. In particular, this is true when $c = \langle v, x \rangle$, which defines a face of $\text{hypo}(g_{\mathcal{V}})$ at x if any only if $\langle v, x \rangle = g_{\mathcal{V}}(x)$. Therefore, we have $(x, g_{\mathcal{V}}(x)) \in F_v$. □

Taking $L, \mathcal{V}, \mathcal{V}^*$ as in Assumption 2 and a face of $\text{hypo}(g_{\mathcal{V}})$, $F_{v^*} = H_{v^*} \cap \text{hypo}(g_{\mathcal{V}})$ for $v^* \in \mathcal{V}^*$, the projection π preserves full-dimensionality.

Claim 7. *Given $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2, define $F_v := F_v^{\text{hypo}(g_{\mathcal{V}})} = H_v \cap \text{hypo}(g_{\mathcal{V}})$. For all $v \in \mathcal{V}^*$, $\pi(F_v)$ is full dimensional in $\mathbb{R}^{\mathcal{Y}}_+$.*

Proof. By Corollary 6, F_v is a facet of $\text{hypo}(g_{\mathcal{V}^*}) = \text{hypo}(g_{\mathcal{V}})$ in $\mathbb{R}^d_+$, meaning it is $(d-1)$-dimensional. Moreover, Claim 6 states that the dimension of F_v is preserved for each $v \in \mathcal{V}^*$. Thus, $\dim(F_v) = \dim(\pi(F_v)) = |\mathcal{Y}|$. □

Now we can observe a set of normals $\mathcal{V}'$ generates faces of $g_{\mathcal{V}}$ whose projections cover $\mathbb{R}^{\mathcal{Y}}_+$ if and only if the set contains $\mathcal{V}^*$. This will translate to a set being representative for a loss if and only if it contains a finite minimum representative set (in settings where one exists.)

Claim 8. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For $\mathcal{V}' \subseteq L(\mathcal{R})$, we have $\cup_{v \in \mathcal{V}'} \pi(F_v) = \mathbb{R}^{\mathcal{Y}}_+ \iff \mathcal{V}^* \subseteq \mathcal{V}'$.*

Proof. For any $v \in \mathcal{V}$, define $F_v := F_v^{\text{hypo}(g_{\mathcal{V}})} = H_v \cap \text{hypo}(g_{\mathcal{V}})$.

($\implies$) For contraposition, suppose $\mathcal{V}^* \not\subseteq \mathcal{V}'$. Then $\exists v \in \mathcal{V}^* \setminus \mathcal{V}'$. Observe that $\mathcal{V}^*$ is unique and irredundant (by assumption) and $\pi(F_v^{\text{hypo}(g_{\mathcal{V}})})$ is full-dimensional in $\mathbb{R}^{\mathcal{Y}}_+$ by Claim 7. Moreover, $\pi(F_v) \notin \cup_{v' \in \mathcal{V}'} \pi(F_v)$, which implies $\cup_{v' \in \mathcal{V}'} \pi(F_v) \neq \cup_{v^* \in \mathcal{V}^*} \pi(F_{v^*}) = \mathbb{R}^{\mathcal{Y}}_+$.

($\impliedby$) Since $\mathcal{V}^* \subseteq \mathcal{V}'$, we immediately have $\cup_{v \in \mathcal{V}^*} \pi(F_v) \subseteq \cup_{v' \in \mathcal{V}'} \pi(F_{v'})$. Moreover, $\cup_{v \in \mathcal{V}^*} \pi(F_v) = \mathbb{R}^{\mathcal{Y}}_+$ by Corollary 9, so $\mathbb{R}^{\mathcal{Y}}_+ \subseteq \cup_{v' \in \mathcal{V}'} \pi(F_{v'})$. As $g_{\mathcal{V}}$ is only finite on $\mathbb{R}^{\mathcal{Y}}_+$ by construction, equality follows.

□

We now claim that a set of projected faces $\{F_v^{\text{hypo}(g_{\mathcal{V}})}\}_{v \in \mathcal{V}'}$ for some $\mathcal{V}'$ will cover $\mathbb{R}^{\mathcal{Y}}_+$ if and only if $\mathcal{V}^* \subseteq \mathcal{V}'$. Given $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2, denote $\Lambda_S := \{\pi(F_v) \mid v \in S\}$ as the set of projected facets generated by $\mathcal{V}^*$.

Claim 9. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2, and $\mathcal{R}' \subseteq \mathcal{R}$ with $\mathcal{V}' := L(\mathcal{R}')$. We have $\cup_{v \in \mathcal{V}'} \pi(F_v) = \mathbb{R}^{\mathcal{Y}}_+ \iff \Lambda_{\mathcal{V}^*} \subseteq \Lambda_{\mathcal{V}'}$.*

Proof. $\implies$ The result follows if $\mathcal{V}^* \subseteq \mathcal{V}'$, which follows from the forward implication of Claim 8. Explicitly, for all $v \in \mathcal{V}^*$ we also have $v \in \mathcal{V}'$, so, $\pi(F_v) \in \Lambda_{\mathcal{V}^*} \cap \Lambda_{\mathcal{V}'} = \Lambda_{\mathcal{V}^*}$.

$\Longleftarrow$ If $\{\pi(F_v) \mid v \in \mathcal{V}^*\} \subseteq \{\pi(F_v) \mid v \in L(\mathcal{R}')\}$, then $\cup\{\pi(F_v) \mid v \in \mathcal{V}^*\} \subseteq \cup_{v \in \mathcal{V}'} \pi(F_v)$. By Corollary 9, we have $\cup\{\pi(F_v) \mid v \in \mathcal{V}^*\} = \mathbb{R}^{\mathcal{Y}}_+$, so $\mathbb{R}^{\mathcal{Y}}_+ \subseteq \cup_{v \in \mathcal{V}'} \pi(F_v)$. The other direction of subset inequality following from hypo$(g_{\mathcal{V}})$ being finite only on $\mathbb{R}^{\mathcal{Y}}_+$. □

Claim 10. *Given $L, \mathcal{V}$ satisfying Assumption 2, if some $H_v \in \mathcal{H}$ supports* hypo$(g_{\mathcal{V}})$*, then $F_v := H_v \cap$* hypo$(g_{\mathcal{V}})$ *is a nonempty face of* hypo$(g_{\mathcal{V}})$*, and thus a subset of a facet.*

Proof. Since H_v supports hypo$(g_{\mathcal{V}})$, we know that hypo$(g_{\mathcal{V}}) \subseteq H_v^+$ and F is not empty. Moreover, H_v^+ is valid, and we have $F_v = H_v \cap \text{hypo}(g_{\mathcal{V}})$ is a face of hypo$(g_{\mathcal{V}})$ by definition.

Moreover, the faces of hypo$(g_{\mathcal{V}})$ are all convex polyhedra as hypo$(g_{\mathcal{V}})$ is a polyhedron. Any face of hypo$(g_{\mathcal{V}})$ is must then be a lower-dimensional face of a facet, and therefore a subset. □

Claim 11. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For any $v \in \mathcal{V}$, the face $F_v \subseteq F_{v^*}$ for some $v^* \in \mathcal{V}^*$.*

Proof. As hypo$(g_{\mathcal{V}^*})$ is a polyhedron, each of its faces are convex polyhedra, and is also a face of some facet of hypo$(g_{\mathcal{V}^*})$; these facets are defined by $\mathcal{V}^*$ and $\mathcal{Y}$ via Corollary 6.

It then suffices to show that if $F_v := H_v \cap \text{hypo}(g_{\mathcal{V}})$ is a face of the facet $F_y := H_y \cap \text{hypo}(g_{\mathcal{V}})$ for some $y \in \mathcal{Y}$, then it must also be a face of F_{v^*} for a $v \in \mathcal{V}^*$; it suffices to show F_v is not a facet, and thus $F_v \neq F_y$. Recall from the definition of a face that $F_v = \text{hypo}(g_{\mathcal{V}}) \cap H_v$ and $F_y = \text{hypo}(g_{\mathcal{V}}) \cap H_y$. As facets of a polyhedron are (uniquely) determined by hyperplanes and F_y is a facet, then if F_v is a facet we must have $v \in \mathcal{V}^*$, and constructed $\mathcal{V}^*$ such that $\mathcal{H}_{\mathcal{V}^*} \cap \mathcal{H}_{\mathcal{Y}} = \emptyset$. Thus, $F_v \neq F_y$ by unique determination of facets. If F_v is not a facet of hypo$(g_{\mathcal{V}})$, then $F_v \neq F_y$ is immediate as F_y is a facet of hypo$(g_{\mathcal{V}})$. In both cases, we have $F_v \neq F_y$, and the result follows.

□

Corollary 10. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2, and $F_v := H_v \cap$* hypo$(g_{\mathcal{V}})$ *be a face of* hypo$(g_{\mathcal{V}})$*. For v, v^* such that $F_v \subseteq F_{v^*}$ and $v^* \in \mathcal{V}^*$, $\pi(F_v) \subseteq \pi(F_{v^*})$.*

Now, we can conclude that projected facets generated by $\mathcal{V}$ contain all other projected faces of hypo$(g_{\mathcal{V}})$.

Corollary 11. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For $v \in L(\mathcal{R})$, let $F_v := H_v \cap \mathrm{hypo}(g_{\mathcal{V}})$ be a face of $\mathrm{hypo}(g_{\mathcal{V}})$. For any $v \in \mathcal{V}$, there is a $v^* \in \mathcal{V}^*$ such that $\pi(F_v) \subseteq \pi(F_{v^*})$.*

Proof. This is exactly Claim 11 and Corollary 10 chained together. □

3.7.4.7 Translating to properties: projecting from $\mathbb{R}^{\mathcal{Y}}_+$ to $\Delta_{\mathcal{Y}}$

Let $f_{\mathcal{V}} : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}_+ \cup \{-\infty\}$ be a polyhedral concave function with $\mathrm{dom}(f_{\mathcal{V}}) = \Delta_{\mathcal{Y}}$.

Claim 12. *There is some finite set $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$ such that $f_{\mathcal{V}}(p) = \min_{v \in \mathcal{V}} \langle p, v\rangle - \delta(p \mid \Delta_{\mathcal{Y}})$.*

Proof. We will think of $f_{\mathcal{V}}$ as defined $f_{\mathcal{V}} : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}_+ \cup \{-\infty\}$ with $\mathrm{dom}(f_{\mathcal{V}}) = \Delta_{\mathcal{Y}}$. For $p \in \Delta_{\mathcal{Y}}$, we know $\sum_i p_i = 1$, and can write any inner product $\langle p, b\rangle - \beta = \langle p, b\rangle - \langle p, \beta \mathbb{1}\rangle = \langle p, b - \beta\mathbb{1}\rangle$. If $p \notin \Delta_{\mathcal{Y}}$, then $f_{\mathcal{V}}(p) = -\infty$ and inner products are not used to compute $f_{\mathcal{V}}$. Moreover, since $f_{\mathcal{V}}$ is polyhedral, it is finitely generated [77, Proposition 19.1.2] and can be written

$$\begin{aligned} f_{\mathcal{V}}(p) &= \min(\langle p, b_1\rangle - \beta_1, \dots, \langle p, b_k\rangle - \beta_k) - \delta(p \mid \Delta_{\mathcal{Y}}) \\ &= \min(\langle p, b_1 - \beta_1\mathbb{1}\rangle, \dots, \langle p, b_k - \beta_k\mathbb{1}\rangle) - \delta(p \mid \Delta_{\mathcal{Y}}) \ . \end{aligned}$$

□

This allows us to project $g_{\mathcal{V}}$ from $\mathbb{R}^{\mathcal{Y}}_+$ to the $\Delta_{\mathcal{Y}}$.

Lemma 20. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For all polyhedral concave $f_{\mathcal{V}} : \mathbb{R}^{\mathcal{Y}}_+ \to \mathbb{R}_+ \cup \{-\infty\}$ with $\mathrm{dom}(f_{\mathcal{V}}) = \Delta_{\mathcal{Y}}$, the polyhedral concave function $g_{\mathcal{V}} : \mathbb{R}^{\mathcal{Y}}_+ \to \mathbb{R}_+$ with $\mathrm{dom}(g_{\mathcal{V}}) = \mathbb{R}^{\mathcal{Y}}_+$ matches $f_{\mathcal{V}}(p) = g_{\mathcal{V}}(p)$ for all $p \in \Delta_{\mathcal{Y}}$.*

Given the function $f_{\mathcal{V}}$, we consider $g_{\mathcal{V}}$ to be its extension and L such that $\underline{L}_+ = g_{\mathcal{V}}$ and $\underline{L} = f_{\mathcal{V}}$. Moreover, define the function $\theta(v) = \{p \in \Delta_{\mathcal{Y}} \mid \langle v, p\rangle = f_{\mathcal{V}}(p)\}$ as the level sets of the loss vector $v \in \mathcal{V}$.

Claim 13. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. Then $\underline{L}$ is polyhedral (on the simplex) and $f_{\mathcal{V}} = \underline{L}$. Moreover, $f_{\mathcal{V}}(p) = g_{\mathcal{V}}(p)$ for all $p \in \Delta_{\mathcal{Y}}$.*

Claim 14. *Consider $L, \mathcal{V}, \mathcal{V}^*$ as in Assumption 2. For all $v \in \mathcal{V}$, consider the face F_v of* $\mathrm{hypo}(g_{\mathcal{V}})$*. Then $\theta(v) = \pi(F_v) \cap \Delta_{\mathcal{Y}}$.*

Proof. Fix $p \in \Delta_{\mathcal{Y}}$.

$$\begin{aligned}
p \in \theta(v) &\iff \langle v, p \rangle = f_{\mathcal{V}}(p) && \text{Definition of } \theta \\
&\iff \langle v, p \rangle = \min_{v' \in \mathcal{V}} \langle v', p \rangle && f_{\mathcal{V}} = g_{\mathcal{V}} \text{ on } \Delta_{\mathcal{Y}} \text{ (Cor. 20)} \\
&\iff v \in \arg\min_{v' \in \mathcal{V}} \langle v', p \rangle \\
&\iff p \in \pi(F_v)\ . && \text{Lemma 19}
\end{aligned}$$

□

3.7.4.8 Moving from $\Delta_{\mathcal{Y}}$ to $\mathbb{R}^{\mathcal{Y}}_+$

Now that we have translated from $\mathbb{R}^d_+$ to $\mathbb{R}^{\mathcal{Y}}_+$ in § 3.7.4.6 and from $\mathbb{R}^{\mathcal{Y}}_+$ to $\Delta_{\mathcal{Y}}$ in § 3.7.4.7, we take some final steps to prove Lemma 5 by showing equivalences from $\Delta_{\mathcal{Y}}$ to $\mathbb{R}^{\mathcal{Y}}_+$.

Lemma 21. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For all $r \in \mathcal{R}$ with $v = L(r)$, $\Gamma_r = \theta(v) = \pi(F_v) \cap \Delta_{\mathcal{Y}}$.*

Proof. Let us rewrite

$$\begin{aligned}
\Gamma_r &= \{p \in \Delta_{\mathcal{Y}} \mid r \in \arg\min_{r' \in \mathcal{R}} \langle L(r'), p \rangle\} \\
&= \{p \in \Delta_{\mathcal{Y}} \mid v \in \arg\min_{v' \in \mathcal{V}} \langle v', p \rangle\} \\
&= \{p \in \Delta_{\mathcal{Y}} \mid \langle v, p \rangle = \min_{v' \in \mathcal{V}} \langle v', p \rangle\} \\
&= \{p \in \Delta_{\mathcal{Y}} \mid \langle v, p \rangle = f(p)\} \\
&= \theta(v)\ .
\end{aligned}$$

The rest of the result follows from Claim 14.

□

Lemma 22. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. Then $g_{\mathcal{V}^*}(x) = \min_{v \in \mathcal{V}^*} \langle v, x \rangle$ is (positively) 1-homogeneous.*

Proof. If $x \notin \mathbb{R}^{\mathcal{Y}}_+$, then $cg(x) = -\infty = g(cx)$ for any $c > 0$. If $x \in \mathbb{R}^{\mathcal{Y}}_+$, then we have $g(cx) = \min_{v \in \mathcal{V}^*} \langle v, cx \rangle = c \min_{v \in \mathcal{V}^*} \langle v, x \rangle = cg(x)$ for any $c > 0$ by linearity of the inner product. □

Every minimizable loss L elicits a unique property $\Gamma := \text{prop}_{\mathcal{P}}[L]$, and we can define the *extended level set* $\bar{\Gamma}_r := \{x \in \mathbb{R}^{\mathcal{Y}}_+ \mid \langle L(r), x \rangle = \underline{L}_+(x)\}$.

Lemma 23. *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For any $r \in \mathcal{R}$ and $c > 0$, if $p \in \Gamma_r$, then $cp \in \bar{\Gamma}_r$.*

Proof. Fix $r \in \mathcal{R}$ and $c > 0$. We have

$$\begin{aligned}
p \in \Gamma_r &= \{p' \in \Delta_{\mathcal{Y}} \mid r \in \arg\min_{r' \in \mathcal{R}} \langle L(r'), p' \rangle\} && \text{Definition of level set} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid v \in \arg\min_{v' \in L(\mathcal{R})} \langle v', p' \rangle\} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid \langle v, p' \rangle = \min_{v' \in L(\mathcal{R})} \langle v', p' \rangle\} && L \text{ minimizable} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid \langle v, p' \rangle = g_{\mathcal{V}}(p')\} && \text{Assumption 2: } g_{L(\mathcal{R})} = g_{\mathcal{V}} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid c\langle v, p' \rangle = cg_{\mathcal{V}}(p')\} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid \langle v, cp' \rangle = g_{\mathcal{V}}(cp')\} && \text{Lemma 22} \\
\implies cp &\in \{x \in \mathbb{R}^{\mathcal{Y}}_+ \mid \langle v, x \rangle = g_{\mathcal{V}}(x)\} \\
&= \bar{\Gamma}_r \ .
\end{aligned}$$

□

Lemma 24. *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For any $v \in L(\mathcal{R})$, define the face F_v of* $\text{hypo}(g_{\mathcal{V}})$*. For any $r \in \mathcal{R}$ with $v = L(r)$, $\bar{\Gamma}_r = \pi(F_v)$.*

Proof.

$$\begin{aligned}
\bar{\Gamma}_r &= \{x \in \mathbb{R}^{\mathcal{Y}}_+ \mid \langle L(r), x\rangle = \underline{L}_+(x)\} && \text{Definition of } \bar{\Gamma}_r \\
&= \{x \in \mathbb{R}^{\mathcal{Y}}_+ \mid \langle L(r), x\rangle = g_{\mathcal{V}}(x)\} && \text{Assumption 2: } \underline{L}_+(x) = g_{\mathcal{V}}(x) \\
&= \{x \in \mathbb{R}^{\mathcal{Y}}_+ \mid \langle v, x\rangle = g_{\mathcal{V}}(x)\} && v = L(r) \\
&= \pi(F_v) && \text{Since } F_v = \{(x, g_{\mathcal{V}}(x)) \mid \langle v, x\rangle = g_{\mathcal{V}}(x)\}
\end{aligned}$$

□

Claim 15. *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. For any $v \in L(\mathcal{R})$, denote the face $F_v := H_v \cap \text{hypo}(g_{\mathcal{V}})$. A set $\mathcal{R}' \subseteq \mathcal{R}$ with $\mathcal{V}' := L(\mathcal{R}')$ is representative for L if and only if $\cup_{v\in\mathcal{V}'}\pi(F_v) = \mathbb{R}^{\mathcal{Y}}_+$.*

Proof. $\implies$ This proof follows from three lemmas: first, we observe that g is 1-homogeneous via Lemma 22. Then we extend the notion of a level set Γ_r to the nonnegative orthant $\bar{\Gamma}_r$, and show that any scalar transformation of a distribution in the level set is contained in the same (extended) level set via Lemma 23. Finally, we show the extended level set is exactly the projection $\pi(F_v)$ in Lemma 24. As a corollary, we chain the results to observe $\cup_{r\in\mathcal{R}'}\Gamma_r = \Delta_{\mathcal{Y}} \implies \cup_{r\in\mathcal{R}'}\bar{\Gamma}_r = \mathbb{R}^{\mathcal{Y}}_+ = \cup_{v\in L(\mathcal{R}')}\pi(F_v) = \mathbb{R}^{\mathcal{Y}}_+$, yielding the forward implication.

$\impliedby$ Fix $p \in \Delta_{\mathcal{Y}} \subseteq \mathbb{R}^{\mathcal{Y}}_+$. By the assumption, there is a $v \in \mathcal{V}'$ such that $p \in \pi(F_v)$. By Lemma 21, we have $p \in \pi(F_v) \cap \Delta_{\mathcal{Y}} = \Gamma_r$ for the $r \in \mathcal{R}'$ such that $v = L(r)$. As this is true for all $p \in \Delta_{\mathcal{Y}}$, we have $\mathcal{R}'$ representative.

□

3.7.4.9 Proving Lemma 5

We now proceed with a few final lemmas that ultimately yield the proof of Lemma 5.

Lemma 25. *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. A finite set $\mathcal{R}' \subseteq \mathcal{R}$ with $\mathcal{V}' = L(\mathcal{R}')$ is representative if and only if $\mathcal{V}^* \subseteq \mathcal{V}'$.*

Proof. Chain Claim 15 and Claim 8 to yield the result. □

Lemma 26. *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{+}, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. A finite set $\mathcal{R}' \subseteq \mathcal{R}$ with $\mathcal{V}' = L(\mathcal{R}')$ is representative if and only if $\Theta_{\mathcal{V}^*} \subseteq \{\theta(v) \mid v \in \mathcal{V}'\}$.*

Proof. Chain Claim 15 and Claim 9 to yield the result. □

Define $\Theta_{\mathcal{V}^*} := \{\theta(v) \mid v \in \mathcal{V}^*\}$; it follows that this set is exactly the set of level sets of the property elicited by L. Moreover, let $\mathcal{R}^*$ be the finite set of reports given by Claim 4.

Corollary 12. *Consider $L, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2, and $\mathcal{R}^*$ such that $\mathcal{V}^* = L(\mathcal{R}^*)$ as in Claim 4. Moreover, suppose L elicits Γ. $\Theta_{\mathcal{V}^*} = \{\Gamma_r \mid r \in \mathcal{R}^*\}$.*

Lemma 27. *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{+}, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2. Moreover, let $\Gamma := \mathrm{prop}_{\mathcal{P}}[L]$. $\Theta_{\mathcal{V}^*} = \{\Gamma_r \mid r \in \mathcal{R}, \dim(\Gamma_r) = |\mathcal{Y}| - 1\}$.*

Proof. From Claim 7, we know $\Lambda_{\mathcal{V}^*}$ is exactly the set of full-dimensional level sets in $\mathbb{R}^{\mathcal{Y}}_{+}$. Each element of $\Lambda_{\mathcal{V}^*}$ is $\pi(F_v)$ for some $v \in \mathcal{V}^*$. Take $r \in \mathcal{R}^*$ so that $v = L(r)$. By Lemma 21, we have $\theta(v) = \Gamma_r = \pi(F_v) \cap \Delta_{\mathcal{Y}}$ is full-dimensional relative to the simplex. The result follows. □

Lemma 28. *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{+}, \mathcal{V}, \mathcal{V}^*$ satisfying Assumption 2, and $\mathcal{R}^* \subseteq \mathcal{R}$ the set such that $\mathcal{V}^* = L(\mathcal{R}^*)$ as in Claim 4. Moreover, consider $\Gamma := \mathrm{prop}_{\mathcal{P}}[L]$. For any $r \in \mathcal{R}$, there exists a $v^* \in \mathcal{V}^*$ such that $\Gamma_r \subseteq \theta(v^*)$.*

Proof. Take $v = L(r)$. By Corollary 11, there is a $v^* \in \mathcal{V}^* \subseteq \mathcal{V}$ such that $\pi(F_v) \subseteq \pi(F_{v^*})$. Therefore, $\pi(F_v) \cap \Delta_{\mathcal{Y}} \subseteq \pi(F_{v^*}) \cap \Delta_{\mathcal{Y}}$. We know $\theta(v) = \pi(F_v) \cap \Delta_{\mathcal{Y}}$ and similarly for $\theta(v^*)$ by Lemma 21. The result follows. □

Now this brings us to Lemma 5. The framework in this appendix cues up this proof: any loss L satisfying the assumptions of Lemma 5 has some $g_{L(\mathcal{R})} = \underline{L}_{+}$ as in this section that we can work with.

Lemma 5. *Let $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ be a minimizable loss with a polyhedral Bayes risk $\underline{L}$. Then L has a finite representative set. Furthermore, letting $\Gamma = \mathrm{prop}_{\Delta_{\mathcal{Y}}}[L]$, there exist finite sets $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$ and $\Theta = \{\theta_v \subseteq \Delta_{\mathcal{Y}} \mid v \in \mathcal{V}\}$, both uniquely determined by $\underline{L}$ alone, such that*

(1) A set $\mathcal{R}' \subseteq \mathcal{R}$ is representative if and only if $\mathcal{V} \subseteq L(\mathcal{R}')$.

(2) A set $\mathcal{R}' \subseteq \mathcal{R}$ is minimum representative if and only if $L(\mathcal{R}') = \mathcal{V}$.

(3) A set $\mathcal{R}' \subseteq \mathcal{R}$ is representative if and only if $\Theta \subseteq \{\Gamma_r \mid r \in \mathcal{R}'\}$.

(4) A set $\mathcal{R}' \subseteq \mathcal{R}$ is minimum representative if and only if $\{\Gamma_r \mid r \in \mathcal{R}'\} = \Theta$.

(5) Every representative set for L contains a minimum representative set for L.

(6) The set of full-dimensional level sets of Γ is exactly Θ.

(7) For any $r \in \mathcal{R}$, there exists $\theta \in \Theta$ such that $\Gamma_r \subseteq \theta$.

(8) L tightly embeds $\ell : \mathcal{R}' \to \mathbb{R}^{\mathcal{Y}}_+$ if and only if ℓ is injective and $\ell(\mathcal{R}') = \mathcal{V}$.

Proof. Consider $\underline{L} = f_{\mathcal{V}}$ for a finite set $\mathcal{V}$ by Claim 12. There is a polyhedral concave function $g_{\mathcal{V}}$ on $\mathbb{R}^{\mathcal{Y}}_+$ matching $f_{\mathcal{V}}$ on $\Delta_{\mathcal{Y}}$ by Corollary 20. Moreover, consider $g_{\mathcal{V}} = \underline{L}_+$, and observe that $\underline{L}_+$ matches $\underline{L}$ on $\Delta_{\mathcal{Y}}$ as well. By Corollary 6, we then have a finite set $\mathcal{V}^*$ of smallest cardinality such that $f_{\mathcal{V}} = f_{\mathcal{V}^*}$ and $g_{\mathcal{V}} = g_{\mathcal{V}^*}$. Consider $\mathcal{R}^* \subseteq \mathcal{R}$ such that $\mathcal{V}^* = L(\mathcal{R}^*)$ as in Claim 4. First, observe that $\mathcal{R}^*$ is representative for L as a corollary of Claim 15. Moreover, consider the following set of level sets of $f_{\mathcal{V}^*}$, $\Theta_{\mathcal{V}^*} = \{\theta(v) \mid v \in \mathcal{V}^*\}$.

Now that we have the preliminaries, consider the itemized statements. For $f_{\mathcal{V}^*}$, Lemma 25 is exactly statement (1). This immediately implies statement (2). Moreover, Lemma 26 is exactly statement (3), and again statement (4) immediately follows. Statement (5) is a corollary of the existence of a finite representative set, as shown in Claim 4. Statement (6) is exactly Lemma 27. Statement (7) is exactly Lemma 28. Finally, Statement (8) follows as a corollary of statement (2) and Corollary 2. □

Chapter 4

Convex Elicitation of Continuous Properties

4.1 Introduction

In Table 2.1, we consider four types of problems, depending on the continuity and form of the prediction task at hand: this chapter gives necessary and sufficient conditions on a case within Quadrant 4, in which one is given a (nowhere-locally-constant) continuous statistic they wish to estimate.

A central thread of elicitation literature, weaving between the statistics, economics, and machine learning communities, asks which continuous real-valued properties are elicitable, and which loss functions elicit them. Building on earlier work of Osband [67] and Lambert [55], Steinwart et al. [83] show that a property is elicitable if and only if it is *identifiable*, a concept introduced by Osband which says that the set of distributions sharing the same property value can be described by a set of linear constraints. Moreover, these papers give characterizations of the loss functions eliciting these identifiable properties, showing that every loss can be written as the integral of a positive-weighted identification function.

This chapter studies the convex elicitability of continuous estimation problems in the finite-outcome setting. Surprisingly, we find that, under somewhat mild smoothness assumptions, *every* elicitable real-valued property is convex elicitable (Theorem 14). The proof proceeds by observing that elicitability is equivalent to a condition called *identifiability*, and pinpoints a few key attributes of identification functions. We proceed to solve the following abstract problem: given a set of functions $\mathcal{F} \subseteq \{f : \mathcal{R} \to \mathbb{R}\}$, when does there exist a weight function $\lambda : \mathcal{R} \to \mathbb{R}_+$ making λf

increasing over the (convex) report space $\mathcal{R}$ for all $f \in \mathcal{F}$? We give a constructive solution to this problem under certain conditions, and show that identification functions happen to satisfy these conditions.

Thi chapter is heavily based on the work of Finocchiaro and Frongillo [26], published at NeurIPS 2018.

4.2 Setting and Background

In property elicitation, we aim to learn some distributional property by minimizing a loss function. For continuous properties, a central notion in property elicitation is that of identifiability, where the level sets $\Gamma_r := \{p \in \mathcal{P} \mid r \in \Gamma(p)\}$ can be expressed by an affine constraint. Throughout this chapter, we assume properties Γ are single-valued, meaning $|\Gamma(p)| = 1$ for all $p \in \mathcal{P}$. In this single-valued setting, we omit set-notation and write $\Gamma(p) = r$ instead of $\Gamma(p) = \{r\}$.

Definition 31. *Let an elicitable property $\Gamma : \mathcal{P} \to \mathcal{R}$ be given, where $\mathcal{R} = \Gamma(\mathcal{P})$. A function $V : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ identifies Γ if*

$$\mathbb{E}_{Y \sim p}\left[V(r, Y)\right] = 0 \iff r = \Gamma(p) \ , \tag{4.1}$$

for all $r \in \mathring{\mathcal{R}}$ and $p \in \mathcal{P}$, where $\mathring{\mathcal{R}}$ is the interior of $\mathcal{R}$. In this case we say Γ is identifiable. We say V is oriented if we additionally have $\mathbb{E}_{Y \sim p}\left[V(r, Y)\right] > 0 \iff r > \Gamma(p)$, for all $r \in \mathring{\mathcal{R}}$ and $p \in \mathcal{P}$.

Note that by the terminology of Steinwart et al. [83], an identification function satisfying eq. (4.1) on all of $\mathring{\mathcal{R}}$ is called *strong*, and otherwise it must hold almost everywhere.

We can loosely think of an identification function as a derivative of a loss; if L is differentiable and elicits Γ, then roughly speaking, we expect $\frac{d}{dr}\mathbb{E}_{Y \sim p} L(r, Y) = 0 \iff \Gamma(p) = r$. Thus, modifying the identification function V multiplicatively allows one to change the corresponding loss L while keeping the minimizer (property value) the same.

In this chapter, we assume our properties possess two important qualities: continuity, and being nowhere-locally-constant.

Definition 32 (Lambert [55]). *A continuous property* $\Gamma : \mathcal{P} \to \mathcal{R}$ *is nowhere-locally-constant if there does not exist any open neighborbood* U *of* $\mathcal{P}$ *such that* $\Gamma(p) = r$ *for all* $p \in U$*.*

Intuitively, restricting to nowhere-locally-constant properties is merely to ease bookkeeping, as one could always collapse different report values together afterwards.

It is known that for continuous, nowhere-locally-constant, real-valued properties, identifiability is equivalent to elicitability. In this chapter, we show that under slightly stronger assumptions, identifiability is equivalent to *convex* elicitability.

4.2.1 Relevant prior work

While Savage [79] studied the elicitation of expected values, the literature on the elicitation of general properties began with Osband [67], who gave several important results. One of Osband's observations is that the level sets $\Gamma_r = \{p \in \mathcal{P} \mid \Gamma(p) = r\}$ of an elicitable property Γ must be convex [67, Proposition 2.5]. *Osband's principle* states that (under a mild regularity assumption) every loss function eliciting a given property can be written as the integral of a weighted identification function [67, Theorem 2.1], giving the rough "derivative" connection mentioned above. Osband also gave several other results, such as the separability of loss functions jointly eliciting quantiles.

Independent of Osband, Lambert [55], Lambert and Shoham [56], Lambert et al. [57] provide a geometric characterization of both continuous and finite properties when the set of outcomes $\mathcal{Y}$ is finite. Lambert represents the identification function as a vector and relates finite-valued properties to *power diagrams* in computational geometry. In turn, Lambert rediscovered several results of Osband for the real-valued case, such as convexity of level sets and a one-dimensional version of Osband's principle.

Theorem 13 (Lambert [55, Theorem 5]). *Let* $\Gamma : \mathcal{P} \to \mathbb{R}$ *be a continuous, nowhere-locally-constant property. If the level sets* $\{p \in \mathcal{P} : \Gamma(p) = r\}$ *are convex, then* Γ *is elicitable, and has a continuous, bounded, and oriented identification function. Conversely, if* Γ *is elicitable, its level sets are convex.*

Steinwart et al. [83] extend this result to the case of infinite $\mathcal{Y}$. None of the above-mentioned papers characterize when the loss eliciting a given property is *convex.*

This chapter studies the *direct* elicitibility of continuous properties. While convex losses are well-known for several continuous properties of interest, including the mean and other expected values (squared loss), ratios of expectations (weighted squared loss), and the median and other quantiles (pinball loss), to our knowledge, there were no previous results on the direct convex elicitation of general continuous properties.

4.3 Continuous properties are convex elicitable iff they are elicitable

We will show that, under mild conditions, every elicitable real-valued property is also convex elicitable. Let us first give some intuition why one might suspect this statement to be true. From a geometric perspective, the level sets $\Gamma_r = \{p \in \mathcal{P} \mid \Gamma(p) = r\}$ of continuous elicitable properties are hyperplanes intersected with $\mathcal{P}$. As one varies r, the level sets may be locally parallel, in which case the property is locally a link of a linear property (expected value), or the level sets may not be parallel, in which case the property locally resembles a link of a ratio of expectations. In fact, the second case also covers the first, so we can say that, roughly speaking, every continuous property looks locally like a ratio of expectations. The following proposition states that if the property can actually be written as a finite piecewise ratio of expectations, it is convex elicitable. Hence, taking the limit as one approximates a given property better and better by ratios of expectations, one may suspect that indeed every continuous property is convex elicitable.

Proposition 10. *Continuous piecewise ratio-of-expectation properties are convex elicitable.*

Proof. First, we formalize the statement. Recall that $\mathcal{Y}$ is a finite set. Let $\phi_i : \mathcal{Y} \to \mathbb{R}$ and $\psi_i : \mathcal{Y} \to \mathbb{R}_+$ be arbitrary for $i = 1, \ldots, k$, and let $\gamma_i(p) := \mathbb{E}_{Y \sim p}\phi_i(Y)/\mathbb{E}_{Y \sim p}\psi_i(Y)$. Assume that we have $a_0 < \cdots < a_k$ such that for all $p \in \mathcal{P}$, there is a unique $i \in \{1, \ldots, k\}$ such that either $\gamma_{i-1}(p) \in (a_{i-1}, a_i)$ or $\gamma_{i-1}(p) = \gamma_i(p) = a_{i-1}$. Call this $i(p)$, and by extension $i(r)$ where $r = \gamma_i(p)$ for this i. We will show that $\Gamma(p) := \gamma_{i(p)}(p)$ is convex elicitable with respect to the full probability simplex $\mathcal{P} = \Delta(\mathcal{Y})$.

Observe that by construction, for each $i \in \{1, \ldots, k-1\}$ the level sets for a_i coincide: $S_i = \{p : \Gamma(p) = a_i\} = \{p : \gamma_i(p) = a_i\} = \{p : \gamma_{i-1}(p) = a_i\}$. Moreover, for all such i, these level sets are full-dimensional in $\mathcal{P}$, i.e., there are $(n-2)$-dimensional affine sets which are the intersection of a hyperplane and $\mathcal{P}$. Now let $V_i(r, y) = \psi_i(y)r - \phi_i(y)$, which identifies γ_i, and is strictly increasing in r as $\psi_i(y) > 0$ for all y. We now see that the hyperplane which is the span of S_i in $\mathbb{R}^n$ is orthogonal to the vectors $V_{i-1}(a_i, \cdot) \in \mathbb{R}^n$ and $V_i(a_i, \cdot) \in \mathbb{R}^n$, by the definition of identifiability. We conclude that there is some coefficient α_{i-1} such that $V_{i-1}(a_i, y) = \alpha_{i-1} V_i(a_i, y)$ for all $y \in \mathcal{Y}$. (In fact, $\alpha_{i-1} > 0$, as the coefficient of r must be positive.) We then construct $\beta_{i(r)} = \prod_{j=0}^{i(r)} \alpha_j$ and write the identification as $V(r, y) = \beta_{i(r)} V_{i(r)}(r, y)$. □

Moving now to the formal result, let $\mathcal{I} \subseteq \mathbb{R}$ be an interval. Our main technical ingredient shows, given a collection $\mathcal{F}$ of functions $f : \mathcal{I} \to \mathbb{R}$ satisfying certain conditions, how to construct a multiplier $\lambda : \mathcal{I} \to \mathbb{R}_+$ making λf strictly increasing on $\mathring{\mathcal{I}}$ for all $f \in \mathcal{F}$. In our proof, the family $\mathcal{F}$ will be the set of identification functions $\{V(\cdot, y)\}_{y \in \mathcal{Y}}$, and λ will play the role of the weight function in previous work showing that any loss of the form $L(r, y) = \int_{r_0}^{r} \lambda(x) V(x, y) dx$ elicits Γ [83, Theorem 5]. As $\lambda V(\cdot, y)$ is increasing for all $y \in \mathcal{Y}$, the loss L will be convex.

We give three conditions below which are only mildly stronger than what the literature shows to be true of the desired properties. We begin with our three conditions; the first we will assume, and the second and third we will prove hold for any oriented identification function.

Condition 2. *Every $f : \mathcal{I} \to \mathbb{R} \in \mathcal{F}$ is continuous on $\mathring{\mathcal{I}}$, and continuously differentiable on $\mathring{\mathcal{I}}$ except on a finite set $S_f \subsetneq \mathring{\mathcal{I}}$. When f is differentiable, $\frac{d}{dx} f(x)$ is finite. Additionally, if $x \in \mathring{\mathcal{I}}$ and $f(x) = 0$, then for all z in some open neighborhood U of x, $\frac{d}{dz} f(z) \geq 0$ whenever f is differentiable.*

Condition 3. *Every $f \in \mathcal{F}$ is bounded and has at most one zero $x_f \in \mathring{\mathcal{I}}$ so that if x_f exists, $f(x) < 0$ for $x < x_f$ and $f(x) > 0$ for $x > x_f$. If f does not have a zero on $\mathring{\mathcal{I}}$, then either $f(x) < 0$ or $f(x) > 0$ for all $x \in \mathring{\mathcal{I}}$. For all $x \in \mathring{\mathcal{I}}$, at least one function $f \in \mathcal{F}$ is nonzero at x.*

Condition 4. *For all $f, g \in \mathcal{F}$ and all open subintervals $\mathcal{I}' \subseteq \mathring{\mathcal{I}}$ such that $f > 0 > g$ on $\mathcal{I}'$, the function $\frac{g}{f}$ is strictly increasing on $\mathcal{I}'$.*

Proposition 11. *If $\mathcal{F}$ satisfies Condition 1, 2, and 3, then there exists a function $\lambda : \mathcal{I} \to \mathbb{R}_+$ so that λf is increasing over $\mathring{\mathcal{I}}$ for every $f \in \mathcal{F}$.*

With this tool in hand, we are ready to state our main result, with proof deferred to § 4.7.

Theorem 14. *For $\mathcal{P} = \Delta(\mathcal{Y})$, let $\Gamma : \mathcal{P} \to \mathcal{R}$ be a continuous, nowhere-locally-constant property which is identified by a bounded and oriented $V : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$. If $\mathcal{F} = \{V(\cdot, y)\}_{y \in \mathcal{Y}}$ satisfies Condition 1, then Γ is convex elicitable.*

Recall that this chapter focuses on finite $\mathcal{Y}$. The argument establishing Condition 3 immediately extends to infinite outcome spaces. Beginning with p_0, p_1 being arbitrary distributions, $\Gamma(p_0) \neq \Gamma(p_1)$, one simply observes that $V(\gamma(\alpha), p_0)/V(\gamma(\alpha), p_1) = 1 - 1/\alpha$ by the same logic. The central challenge to extending Theorem 14 therefore lies in the proof of Proposition 11.

Loosely speaking, when combining Theorem 14 with the existing literature, we conclude that every "nice" elicitable property is additionally convex elicitable.

Corollary 13. *Let $\mathcal{P} = \Delta(\mathcal{Y})$ be the probability simplex over n outcomes, and let $\Gamma : \mathcal{P} \to \mathcal{R}$ be a nowhere-locally-constant property with a bounded and nowhere vanishing first derivative, a bounded second derivative, and a differentiable right inverse.*[1] *Then the following are equivalent:*

(1) For all $r \in \mathcal{R}$, the level set $\{p \in \mathcal{P} \mid \Gamma(p) = r\}$ is convex.

(2) Γ is quasi-monotonic.

(3) Γ is identifiable and has a bounded and oriented identification function.

(4) Γ is elicitable.

[1] We may identify $\mathcal{P}$ with $\{v \in \mathbb{R}_+^{|\mathcal{Y}|-1} : \sum_{i=1}^{|\mathcal{Y}|-1} v_i \leq 1\}$ so that the derivatives are well defined. In the proof, for ease of notation, we will still write dot products in $\mathbb{R}^{|\mathcal{Y}|}$.

(5) There exists a non-negative, measurable, locally Lipschitz continuous loss function eliciting Γ*.*

(6) Γ *is convex elicitable.*

Proof. We essentially reduce to a similar result of Steinwart et al. [83, Corollary 9]. First, note that the definition of nowhere-locally-constant from Lambert et al. [57] coincides with the definition of Steinwart et al. [83, Definition 4] in finite dimensions. Second, as our assumptions are stronger than theirs, the equivalence of the first five conditions follows. As convex elicitability implies convex level sets (cf. Lambert [55, Theorem 5], which follows even if L can be infinite on the boundary of $\mathcal{R}$), it then suffices to show that identifiability implies convex elicitability.

By standard arguments, the convexity of the level sets $\{p : \Gamma(p) = r\}$ for $r \in \mathring{\mathcal{R}}$ imply that each level set must be a hyperplane intersected with $\mathcal{P}$. (See e.g. Theorem 1 of [57].) Letting $\hat{p}$ be the right inverse of Γ, so that $\Gamma(\hat{p}(r)) = r$ for all $r \in \mathcal{R}$, we may define

$$V(r, y) = \nabla_{\hat{p}(r)}\Gamma \cdot (\delta_y - \hat{p}(r)) \ , \tag{4.2}$$

a form taken from Frongillo and Kash [38, Proposition 18].

Now for any p with $\Gamma(p) = r$, as the level set is a hyperplane intersected with $\mathcal{P}$, we must have $\Gamma(\alpha p + (1 - \alpha)\hat{p}(r)) = r$, and we conclude $\nabla_{\hat{p}(r)}\Gamma \cdot (p - \hat{p}(r)) = 0$. (Simply take the derivative with respect to α.) Thus, as $\nabla\Gamma \neq 0$, the vector $\nabla_{\hat{p}(r)}\Gamma - \nabla_{\hat{p}(r)}\Gamma \cdot \hat{p}(r)\mathbb{1}$ defines the same hyperplane as $\{p : \Gamma(p) = r\}$, and thus V identifies Γ. That V is also bounded and oriented follows easily from our assumptions. As V has a bounded derivative everywhere by assumption, it satisfies Condition 1, and convex elicitability then follows from Theorem 14. □

4.4 Sketch of Proposition 11 and Intuition

We now give a sketch of the construction of the weight function λ in Proposition 11. See § 4.7.2 for the full proof. For the purposes of this section, let us simplify our three conditions as follows:

Condition 1'. Every $f \in \mathcal{F}$ is continuously differentiable.

Condition 2'. Each $f \in \mathcal{F}$ has a single zero, and moves from negative to positive.

Condition 3'. When $f > 0 > g$, the ratio g/f is increasing.

Two function case. To begin, let us consider two functions satisfying Conditions 1', 2', and 3', such that $f > 0 > g$ on the interval $\mathcal{I}$. We wish to find some $\lambda : \mathcal{I} \to \mathbb{R}_+$ making both λf and λg strictly increasing. By Condition 3', we know g/f is increasing on $\mathring{\mathcal{I}}$. Let us choose λ as follows,

$$\lambda(r) := (-f(r)g(r))^{-1/2} \ . \tag{4.3}$$

As $-(fg)(r) > 0$ for all $r \in \mathcal{I}$, we have $\lambda(r) > 0$ as well. Moreover, one easily checks that $\lambda f = \sqrt{-f/g}$ and $\lambda g = \sqrt{-g/f}$, which are both increasing as monotonic transformations of g/f.

General case. More generally, we wish to find a λ such that for all $x \in \mathring{\mathcal{R}}$, $\frac{d}{dx}(\lambda f)(x) > 0$. When $f > 0$, this constraint is equivalent to $\frac{d}{dx}\log(\lambda f)(x) > 0$, which is in turn equivalent to $-\frac{d}{dx}\log\lambda(x) < \frac{d}{dx}\log f(x)$. Similarly, if $f(x) < 0$, then we need $-\frac{d}{dx}\log\lambda(x) > \frac{d}{dx}\log(-f(x))$. Finally, the case $f(x) = 0$ follows easily from Condition 2', as $\frac{d}{dx}f(x) > 0$ and $\lambda > 0$. Combining these constraints, we see that for all $f > 0$ and all $g < 0$, we must have

$$\tfrac{d}{dx}\log(-g(x)) < -\tfrac{d}{dx}\log\lambda(x) < \tfrac{d}{dx}\log f(x) \ . \tag{4.4}$$

In order for these constraints to be feasible, we must have $\frac{d}{dx}\log(-g(x)) < \frac{d}{dx}\log f(x)$ for all $f < 0 < g$, which is seen to be equivalent to Condition 3' after some manipulation.

Perhaps the most natural way to satisfy constraint (4.4) is to simply take the midpoint between the maximum lower bound $\underline{m} : \mathcal{R} \to \mathbb{R}$ and minimum upper bound $\overline{m} : \mathcal{R} \to \mathbb{R}$ defined as follows:

$$\overline{m}(x) := -\sup_{g\in\mathcal{F}:g(x)<0} \tfrac{d}{dx}\log(-g(x)) \ , \qquad \underline{m}(x) := -\inf_{f\in\mathcal{F}:f(x)>0} \tfrac{d}{dx}\log(f(x)) \ .$$

This yields the following construction (where $r_0 \in \mathring{\mathcal{R}}$ is arbitrary),

$$h(x) = \frac{1}{2}\left(\overline{m}(x) + \underline{m}(x)\right) \ , \quad \lambda(x) = \exp\left(\int_{r_0}^{x} h(z)dz\right) \ , \tag{4.5}$$

where one notes $h(x) = \frac{d}{dx}\log\lambda(x)$. Provided our three conditions hold, we now have a positive weight function λ satisfying the constraint (4.4), and we conclude that λf is increasing for all $f \in \mathcal{F}$.

Let us observe that our general construction in eq. (4.5) really is a generalization of the two-function case in eq. (4.3). That is, we are primarily concerned with the "most decreasing" and "least increasing" functions, which allows us to focus on two functions instead of the entire set $\mathcal{F}$. When we only have two functions $f > 0 > g$, eq. (4.5) reduces to $h(x) = -\frac{1}{2}\left(\frac{d}{dx}\log(-g(x)) + \frac{d}{dx}\log f(x)\right)$, whence $\lambda(x) = \exp\left(\frac{1}{2}\log(-g(x)f(x))\right) = 1/\sqrt{-g(x)f(x)}$.

Hurdles and technicalities. As stated, the above construction has two issues, which we now briefly identify and describe how our proof circumvents. First, in general our functions f will pass through 0, possibly making h and therefore λ unbounded. Recall that we only needed to satisfy eq. (4.4), and thus rather than taking the midpoint of the lower and upper bounds as in eq. (4.5), which will diverge whenever one of the bounds diverges, we can always choose h in a slightly more clever manner to be closer to the smallest magnitude bound. See the Appendix for one such construction.

The second problem is that our actual Condition 1 allows for nondifferentiability, which arises in settings of particular interest, like Proposition 10. Fortunately, in the finite-outcome setting, it is essentially without loss of generality to consider continuous $f \in \mathcal{F}$ (see Theorem 13). We can therefore address the finite nondifferentiabilities using continuity arguments, allowing us to focus on the set $\mathcal{I}_c \subseteq \mathcal{I}$ where every $f \in \mathcal{F}$ is continuously differentiable.

4.5 Examples

To illustrate the constructive nature of Theorem 14, we now give two examples. The first is the Beta family scoring rule found in Buja et al. [16, §11] and Gneiting and Raftery [46, §3], which we use to illustrate the construction itself. The second is a simple elicitable property for which the obvious identification function does not give a convex loss; we show how to convexify it.

1. Beta families. Consider the Beta family of loss functions discussed by Buja et al. [16], which elicit the mean over outcomes $\mathcal{Y} = \{0, 1\}$, with $\mathcal{R} = [0, 1]$. After some manipulation, one can write the loss and identification function as follows, for any $\alpha, \beta > -1$,

$$L(r,y) = \int_0^r z^{\alpha-1}(1-z)^{\beta-1}(z-y)dz \qquad\qquad V(r,y) = r^{\alpha-1}(1-r)^{\beta-1}(r-y)\ .$$

While some choices of the parameters yield convex losses, such as $\alpha = \beta = 0$ (log loss) and $\alpha = \beta = 1$ (squared loss), not all do, e.g. $\alpha = 1/5$, $\beta = -1/2$.

We choose $\lambda(r) = r^{1/2-\alpha}(1-r)^{1/2-\beta}$, giving the identification function $V'(r,y) = r^{1/2}(1-r)^{1/2}(r-y)$, which is itself in the Beta family with $\alpha = \beta = 1/2$. Intergrating V' yields the following convex loss,

$$L'(r,y) = \int_0^r z^{1/2}(1-z)^{1/2}(z-y)dz = \arcsin(\sqrt{|y-r|}) - \sqrt{r(1-r)}\ , \tag{4.6}$$

also discovered by Buja et al., which serves as a intermediary between log and squared loss.

2. A quadratic property. Let $\mathcal{Y} = \{1,2,3\}$, and $\Gamma(p) = \frac{1-\sqrt{1-4p_1p_2+2p_2^2}}{2p_2}$, where $\Gamma(p) = p_1$ when $p_2 = 0$ for continuity (from L'Hôpital's rule). Here, p_y denotes the probability outcome y is observed. Some of the level sets of Γ can be seen in Figure 4.2. A very natural choice of identification function for Γ is $V(r,1) = r-1$, $V(r,2) = \frac{1}{2} + r - r^2$, $V(r,3) = r$, as one readily verifies. Yet we see in Figure 4.1(b) that $V(\cdot,2)$ is not strictly increasing, so the loss given by integrating V will not be convex.

The set $\mathcal{F} = \{V(\cdot,y)\}_{y\in\mathcal{Y}}$ satisfies Conditions 1–3, however, and thus we may use our construction to obtain a positive function λ for which $L(r,y) = \int_{r_0}^r \lambda(x)V(x,y)dx$ elicits Γ and is convex in r. Unfortunately, for this particular example, the construction given in the proof of Proposition 11 produces a somewhat unwieldy function λ. Fortunately, while that constructed λ is guaranteed to make λf monotone for every function f in $\mathcal{F}$, it is generally not unique, and in many cases a simpler choice of λ can be found. In particular, our proof shows that *any* function h satisfying the criteria of [26, Claim 1] will lead to suitable choice of λ; among these criteria are that $h(r) = -\frac{d}{dr}\log\lambda(r)$ must lie between certain lower and upper bounds (solid orange and blue in Figure 4.3) for all r. We illustrate this example construction in Figure 4.3; for the case of our quadratic property, the choice $-\frac{d}{dr}\log\lambda(r) = h(r) = 4r-1$ (shown as dashed blue) suffices, yielding a simple $\lambda(r) = \exp(2r^2 - r)$. This choice of λ gives

$$\begin{aligned}\lambda(r)V(r,1) &= \exp(2r^2 - r)(r-1) \\ \lambda(r)V(r,2) &= \exp(2r^2 - r)((1/2) + r - r^2) \\ \lambda(r)V(r,3) &= r\exp(2r^2 - r)\ ,\end{aligned}$$

which we can integrate to obtain a convex loss.

4.6 Chapter conclusion

We have shown that all real-valued properties over finite outcomes, which are identified by a mostly-smooth continuous identification function, are convex elicitable. Beyond natural relevance to machine learning, and statistical estimation more broadly, these results bring insights into the area of information elicitation. For example, a generalization of a common prediction market framework, the Scoring Rule Market, is well-defined for any loss function [41, 57]. Yet it is not clear whether practical markets exist for any elicitable property. Among the practical considerations are axioms such as Tractable Trading (TT), which states that participants can compute their optimal trade/action under a budget [3], and Bounded Trader Budget (BTB), which states that traders with arbitrarily small budgets can still fruitfully participate in the market [41]. Our results imply that essentially every continuous real-valued elicitable property over finite outcomes has a market mechanism which satisfies these axioms. There are likely also implications for wagering mechanisms [58] and forecasting competitions [90], among other settings in information elicitation.

4.6.1 Future work

Infinite outcomes. A challenging but important extension would be to allow infinite $\mathcal{Y}$, for example, $\mathcal{Y} = [0,1] \subseteq \mathbb{R}$. As discussed following Theorem 14, many pieces of our argument extend immediately, such as the argument establishing Condition 3. We believe the key hurdle to such an extension will be in Proposition 11, as several quantities become harder to control. As one example, function h used to obtain λ might be constructed as the midpoint of some upper and lower bounds, which may not be attained in the infinite case. Extending to infinite outcomes requires the relaxation of our continuity assumption, as many properties of interest have discontinuous identification functions in the infinite-outcome space, like the median.

Vector-valued properties. Finally, we would like to extend our construction to vector-valued properties $\Gamma : \mathcal{P} \to \mathbb{R}^k$. In light of our results, this question is only interesting for properties which are not vectors of elicitable properties: if the k components of Γ are themselves elicitable, we may construct a convex loss for each, and the sum will be a convex loss eliciting Γ. Unfortunately, we lack a characterization of elicitable vector-valued properties, so the question of whether all elicitable vector-valued properties are convex elicitable seems even further from reach.

4.7 Chapter appendix

4.7.1 Omitted proofs

Theorem 14. *For $\mathcal{P} = \Delta(\mathcal{Y})$, let $\Gamma : \mathcal{P} \to \mathcal{R}$ be a continuous, nowhere-locally-constant property which is identified by a bounded and oriented $V : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$. If $\mathcal{F} = \{V(\cdot, y)\}_{y \in \mathcal{Y}}$ satisfies Condition 1, then Γ is convex elicitable.*

Proof. We have assumed all $f \in \mathcal{F} = \{V(\cdot, y)\}_{y \in \mathcal{Y}}$ are bounded, oriented, and satisfy Condition 1, and thus to apply Proposition 11, we need only establish Conditions 2 and 3. A fact we use throughout is that $V(r, y) = \mathbb{E}_{Y \sim \delta_y} V(r, Y)$, where δ_y is the point distribution on $y \in \mathcal{Y}$.

To establish Condition 2, we first observe that boundedness of each $f \in \mathcal{F}$ follows by assumption. Second, we show that each f has at most one zero on $\mathring{\mathcal{R}}$. As V identifies Γ, note that $V(r, y) = 0 \iff \Gamma(\delta_y) = r$ when $r \in \mathring{\mathcal{R}}$. As Γ is single-valued, there can be at most one such $r \in \mathring{\mathcal{R}}$. Third, we must show that if f has a zero on $\mathring{\mathcal{R}}$, it changes sign from negative to positive at that zero, and if not, f never changes sign on $\mathring{\mathcal{R}}$. The first case follows from the fact that $\Gamma(\delta_y) = r$ and that V is oriented. For the second case, $V(\cdot, y)$ has no zero on $\mathring{\mathcal{R}}$, and thus by continuity of V, cannot change sign on $\mathring{\mathcal{R}}$. Fourth, to see that $\mathcal{F}$ has at least one nonzero function for all $r \in \mathring{\mathcal{R}}$, note that if $V(r, y) = 0$ for all $y \in \mathcal{Y}$, then $\mathbb{E}_{Y \sim p} V(r, Y) = 0$ for all $p \in \mathcal{P}$. Thus, as V identifies Γ

and $r \in \mathring{\mathcal{R}}$, we would have $\Gamma(p) = r$ for all p, contradicting nowhere-locally-constancy.

For Condition 3, consider $V(\cdot, y_0), V(\cdot, y_1) \in \mathcal{F}$ and open interval $\mathcal{I}' = (a, b)$ such that $V(r, y_0) > 0 > V(r, y_1)$ for all $r \in \mathcal{I}'$. We define $p_\alpha = (1-\alpha)\delta_{y_0} + \alpha\delta_{y_1}$ and $\gamma(\alpha) = \Gamma(p_\alpha)$ for $\alpha \in [0, 1]$. Since Γ is continuous and nowhere-locally-constant, Steinwart et al. [83, Cor. 9] implies that Γ is quasi-monotone, which in turn implies that γ is nondecreasing on $[0, 1]$.

We first show $\mathcal{I}' \subseteq \gamma([0, 1]) = [\gamma(0), \gamma(1)]$. By definition of $\mathcal{I}'$, we know $r \in \mathcal{I}' \implies V(r, y_1) < 0 < V(r, y_0)$ and the orientation of V then implies $\Gamma(\delta_{y_1}) > r > \Gamma(\delta_{y_0})$. Thus, $\Gamma(\delta_{y_1}) = \gamma(1) \geq b > a \geq \Gamma(\delta_{y_0}) = \gamma(0)$, with the strict inequality since $\mathcal{I}'$ is nonempty. We then see that $r \in (a, b) \implies r \in [\Gamma(\delta_{y_0}), \Gamma(\delta_{y_1})] = \gamma([0, 1])$, and therefore $\mathcal{I}' \subseteq \gamma([0, 1])$

Next, we show that γ is not only nondecreasing but strictly increasing on $A = \gamma^{-1}(\mathcal{I}')$. Note that A is itself an open interval as γ is continuous. Let $\alpha, \alpha' \in A$, and suppose for a contradiction that $\gamma(\alpha) = \gamma(\alpha') = r \in \mathcal{I}' \subseteq \mathring{\mathcal{R}}$. Then $\Gamma(p_\alpha) = \Gamma(p_{\alpha'}) = r$, and as V identifies Γ, we have $\mathbb{E}_{Y\sim p_\alpha} V(r, Y) = \mathbb{E}_{Y\sim p_{\alpha'}} V(r, Y) = 0$. Thus, $\mathbb{E}_{Y\sim p_0} V(r, Y) = (\alpha' \mathbb{E}_{Y\sim p_\alpha} V(r, Y) - \alpha \mathbb{E}_{Y\sim p_{\alpha'}} V(r, Y))/(\alpha' - \alpha) = 0$, and similarly for p_1. By identifiability again, we must now have $\Gamma(p_0) = \Gamma(p_1) = r$, contradicting $\Gamma(p_0) < \Gamma(p_1)$ as observed above.

Since V identifies Γ, we have for $\alpha \in A$,

$$\begin{aligned} 0 = \mathbb{E}_{Y\sim p_\alpha} V(\gamma(\alpha), Y) &= (1-\alpha)\mathbb{E}_{Y\sim\delta_{y_0}}[V(\gamma(\alpha), Y)] + \alpha\mathbb{E}_{Y\sim\delta_{y_1}}[V(\gamma(\alpha), Y)] \\ &= (1-\alpha)V(\gamma(\alpha), y_0) + \alpha V(\gamma(\alpha), y_1) \ , \end{aligned}$$

from which we conclude the function $F(\alpha) = V(\gamma(\alpha), y_1)/V(\gamma(\alpha), y_0) = (\alpha - 1)/\alpha = 1 - 1/\alpha$, which is strictly increasing in α. Observe that as γ is strictly increasing on A, its inverse is strictly increasing on $\mathcal{I}'$. Thus $V(r, p_1)/V(r, p_0) = F(\gamma^{-1}(r)) = 1 - 1/\gamma^{-1}(r)$ is strictly increasing on $\mathcal{I}'$, as desired.

As we have now established that $\mathcal{F}$ satisfies Conditions 1-3, Proposition 11 yields a weight function $\lambda : \mathcal{R} \to \mathbb{R}_+$ such that for all $y \in \mathcal{Y}$, the map $r \mapsto \lambda(r)V(r, y)$ is strictly increasing on $\mathring{\mathcal{R}}$. Thus, fixing $r_0 \in \mathring{\mathcal{R}}$, the loss $L(r, y) = \int_{r_0}^{r} \lambda(r')V(r', y)dr'$ is convex in r for each $y \in \mathcal{Y}$, as noted by Rockafellar [77, Theorem 24.2]. Moreover, as $\lambda > 0$, L elicits Γ by Lambert [55, Theorem 6]. □

4.7.2 Proving Proposition 11

Let $\mathcal{I} \subseteq \mathbb{R}$ be an interval, and $\mathcal{F}$ a finite set of functions $f : \mathcal{I} \to \mathbb{R}$. As a reminder, we designate $\mathbb{R}_+ := \{r : r \in \mathbb{R}, r > 0\}$.

In Theorem 14, the family $\mathcal{F}$ will be the set of functions $\{V(\cdot, y)\}_{y\in\mathcal{Y}}$ identifying Γ. For ease of exposition, given any $\mathcal{F}$, we define $\mathcal{F}_+(x) = \{f \in \mathcal{F} : f(x) > 0\}$ to be the subset of functions strictly positive at x and $\mathcal{F}_-(x) = \{f \in \mathcal{F} : f(x) < 0\}$ strictly negative. For any function $f : \mathcal{I} \to \mathbb{R}$, let $f^+(x) = \lim_{u\to x^+} f(u)$ and $f^-(x) = \lim_{u\to x^-} f(u)$ denote the right and left limits of f at x, respectively.

For convenience, we recall our three conditions, and the statement to be proved.

Condition 1. *Every $f : \mathcal{I} \to \mathbb{R} \in \mathcal{F}$ is continuous on $\mathring{\mathcal{I}}$, and continuously differentiable on $\mathring{\mathcal{I}}$ except on a finite set $S_f \subsetneq \mathring{\mathcal{I}}$. When f is differentiable, $\frac{d}{dx}f(x)$ is finite. Additionally, if $x \in \mathring{\mathcal{I}}$ and $f(x) = 0$, then for all z in some open neighborhood U of x, $\frac{d}{dz}f(z) \geq 0$ whenever f is differentiable.*

Condition 2. *Every $f \in \mathcal{F}$ is bounded and has at most one zero $x_f \in \mathring{\mathcal{I}}$ so that if x_f exists, $f(x) < 0$ for $x < x_f$ and $f(x) > 0$ for $x > x_f$. If f does not have a zero on $\mathring{\mathcal{I}}$, then either $f(x) < 0$ or $f(x) > 0$ for all $x \in \mathring{\mathcal{I}}$. For all $x \in \mathring{\mathcal{I}}$, at least one function $f \in \mathcal{F}$ is nonzero at x.*

Condition 3. *For all $f, g \in \mathcal{F}$ and all open subintervals $\mathcal{I}' \subseteq \mathring{\mathcal{I}}$ such that $f > 0 > g$ on $\mathcal{I}'$, the function $\frac{g}{f}$ is strictly increasing on $\mathcal{I}'$.*

Proposition 11. *If $\mathcal{F}$ satisfies Condition 1, 2, and 3, then there exists a function $\lambda : \mathcal{I} \to \mathbb{R}_+$ so that λf is increasing over $\mathring{\mathcal{I}}$ for every $f \in \mathcal{F}$.*

Define $S_{\mathcal{F}} = \bigcup_{f\in\mathcal{F}} S_f$ to be the set of all "problem points" in $\mathring{\mathcal{I}}$, where one or more functions fail to be differentiable. Note that $S_{\mathcal{F}}$ is finite, as $\mathcal{F}$ is finite, and S_f is finite for every $f \in \mathcal{F}$ by Condition 1. Let $\hat{S} = S_{\mathcal{F}} \bigcup \partial\mathcal{I}$, where $\partial\mathcal{I}$ denotes the (possibly empty) boundary of interval $\mathcal{I}$. Additionally, define $\mathcal{I}_c := \mathcal{I} \setminus \hat{S} \subseteq \mathring{\mathcal{I}}$, which is an open set as the union of open sets.

We define the functions $\overline{m} : \mathcal{I}_c \to \mathbb{R}$ and $\underline{m} : \mathcal{I}_c \to \mathbb{R}$

$$\overline{m}(x) := -\sup_{g \in \mathcal{F}_-(x)} \tfrac{d}{dx} \log(-g(x)) \qquad \underline{m}(x) := -\inf_{f \in \mathcal{F}_+(x)} \tfrac{d}{dx} \log(f(x))$$
$$\overline{m}'(x) := \max\left(\overline{m}(x) - 1, \frac{\underline{m}(x) + \overline{m}(x)}{2}\right) \qquad \underline{m}'(x) := \min\left(\underline{m}(x) + 1, \frac{\underline{m}(x) + \overline{m}(x)}{2}\right) \;,$$

and finally define $h : \mathcal{I} \to \mathbb{R}$ and $\lambda : \mathcal{I} \to \mathbb{R}_+$ as below.

$$h(x) := \min\left(\max\left(\underline{m}'(x), 0\right), \overline{m}'(x)\right) \tag{4.7}$$
$$\lambda(x) := \exp\left(\int_0^x h(u)du\right) \;, \tag{4.8}$$

where we let $h(x) = 0$ for $x \in \hat{S}$.

Lemma 29. *Let $\varphi : (x, y) \to \mathbb{R}$ be continuously differentiable and nondecreasing. Then $\frac{d}{dz}\varphi(z) \geq 0$ for all $z \in (x, y)$. Moreover, φ is strictly increasing if and only if $Z = \{z \in (x, y) : \frac{d}{dz}\varphi(z) = 0\}$ is totally disconnected.*

Proof. The first part of the statement follows from invoking the Mean Value Theorem.

The converse of the statement is more nontrivial. Assume the differentiable function φ is strictly increasing. As we know increasing $\implies$ non-decreasing $\implies$ nonnegative derivative, it remains to be seen that the derivative is not 0 on any subinterval of (x, y). Suppose there was some subinterval $(a, b) \subseteq (x, y)$ so that ϕ had zero derivative on (a, b). By Mean Value Theorem, we could then see φ is locally constant on (a, b); a contradiction. Therefore, the set of points where φ has 0 derivative must be totally disconnected.

Now assume $\frac{d}{dz}\varphi(z)$ is strictly positive except for a totally disconnected set Z where ϕ has 0 derivative, and recall $\frac{d}{dz}\varphi$ is continuous on (x, y). As the set Z is totally disconnected, the derivative of φ is not zero on any interval $(a, b) \subseteq (x, y)$. As Z is totally disconnected, we then see that for some $z \in (a, b)$, $\frac{d}{dz}\varphi(z) = \varphi(b) - \varphi(a)/(b - a) > 0$ by Mean Value Theorem, and therefore φ is strictly increasing as we can see $\varphi(a) < \varphi(b)$. Therefore, φ is strictly increasing on (x, y). □

We will show that the product λf increasing on $\mathring{\mathcal{I}}$ for every function $f \in \mathcal{F}$. We begin with three claims and then turn to the proof, interspersing technical lemmas which we prove afterwards.

Lemma 30. *If Condition 3 for $\mathcal{F}$ is satisfied on the interval $(a, b) \subseteq \mathcal{I}_c$ then for all $x \in (a, b)$ and all $f \in \mathcal{F}_+(x)$, $g \in \mathcal{F}_-(x)$ we have $\frac{d}{dx}\log(-g(x)) \leq \frac{d}{dx}\log f(x)$, where $Z = \{x \in (a, b) : \frac{d}{dx}\log(-g(x)) = \frac{d}{dx}\log f(x)\}$ is totally disconnected.*

Proof. Since f and g have well-defined, bounded derivative at $x \in \mathcal{I}_c$ (and $f(x) \neq 0$), the function $\frac{g}{f}$ has a well-defined derivative at x, and strictly increasing on (a, b) by Condition 3, as $(a, b) \subseteq \mathcal{I}_c$. By Lemma 29, as $\frac{g}{f}$ is strictly increasing on (a, b), we must have $\frac{d}{dx}\left(\frac{g(x)}{f(x)}\right) \geq 0 \ \forall x \in (a, b)$ where $Z = \{x : \frac{d}{dx}(\frac{g(x)}{f(x)}) = 0\}$ is totally disconnected.

We can see that

$$\begin{aligned} 0 < \tfrac{d}{dx}\left(\frac{g(x)}{f(x)}\right) &= \frac{\frac{d}{dx}g(x)f(x) - \frac{d}{dx}f(x)g(x)}{(f(x))^2} \\ \iff 0 < \tfrac{d}{dx}g(x)f(x) &- \tfrac{d}{dx}f(x)g(x) \\ \iff \frac{\frac{d}{dx}g(x)}{g(x)} &< \frac{\frac{d}{dx}f(x)}{f(x)} \\ \iff \tfrac{d}{dx}\log(-g(x)) &< \tfrac{d}{dx}\log(f(x)) \ . \end{aligned}$$

Similarly, we conclude $\frac{d}{dx}\log(-g(x)) \leq \frac{d}{dx}\log(f(x))$ for all $f \in \mathcal{F}_+(x)$, $g \in \mathcal{F}_-(x)$, and Lemma 29 shows that Z is totally disconnected. □

Lemma 31. *Let $\mathcal{F}$ satisfy Conditions 1 and 2, and let $\delta > 0$ be given. Let $\mathcal{I}'$ refer to any compact subinterval of $\mathring{\mathcal{I}}$. For all $f \in \mathcal{F}$ and ϵ sufficiently small, there exists an $M > 0$ such that for all $x \in \mathcal{I}_c \cap \mathcal{I}'$, the following two conditions hold: (i) if $|f(x)| \geq \epsilon$, $|\frac{d}{dx}\log|f(x)|| < C$, (ii) if $|f(x)| < \epsilon$, $\frac{d}{dx}f(x) \geq 0$.*

Proof of Lemma 31. Let $f \in \mathcal{F}$ be given. If $f(x) \neq 0$ for all $x \in \mathcal{I}'$, then by continuity of f from Condition 1, we have some $\epsilon_1 > 0$ such that $|f(x)| > \epsilon_1$ for all $x \in \mathcal{I}'$. We know that $|\frac{d}{dx}f(x)| \leq C$ by Condition 1 whenever $x \in \mathcal{I}_c$, implying $|\frac{d}{dx}\log|f(x)|| = |\frac{d}{dx}f(x)/f(x)| \leq C/\epsilon_1 =: M$. Thus, our condition (i) holds for any $\epsilon < \epsilon_1$, and condition (ii) never occurs.

On the other hand, if f has some zero $x_f \in \mathring{\mathcal{I}}$, then by Condition 1, we have some open neighborhood U of x_f such that $\frac{d}{dz}f(z) \geq 0$ for all $z \in U \cap \mathcal{I}_c$. On the closed set $W = \mathcal{I}' \setminus U \subset \mathring{\mathcal{I}}$, we have $|f| > 0$ by Condition 2, and by continuity of f, we moreover have ϵ_2 such that $|f| \geq \epsilon_2$ on W. Note that for any $\epsilon < \epsilon_2$, we have $\{x \in \mathcal{I}' : |f(x)| < \epsilon\} \subseteq U$. Now for any $x \in \mathcal{I}_c$ we have: (i) $|f(x)| \geq \epsilon \implies |\frac{d}{dx} \log|f(x)|| = |\frac{d}{dx} f(x)/f(x)| \leq C/\epsilon =: M$ by the above argument, and (ii) $|f(x)| < \epsilon \implies x \in U \implies \frac{d}{dx} f(x) \geq 0$. □

Lemma 31 allows us to conclude that either $\overline{m}$ or $\underline{m}$ is bounded by $M := C/\epsilon$ on $\mathcal{I}'$.

Claim 16. *The function $h : \mathcal{I} \to \mathbb{R}$ defined by eq. (4.7) satisfies*

(1) h is continuous on $\mathcal{I}_c$.

(2) $\underline{m}(x) \leq h(x) \leq \overline{m}(x)$ for all $x \in \mathcal{I}_c$.

(3) $\underline{m}(x) < h(x) < \overline{m}(x)$ for all but a totally disconnected set of $x \in \mathcal{I}_c$.

(4) h is bounded on any compact subinterval $\mathcal{I}'$ of $\mathring{\mathcal{I}}$.

Proof. Statement 1: As h is defined as the max and min of finitely many continuous functions on $\mathcal{I}_c$, it is continuous on $\mathcal{I}_c$.

Statement 2: We give the function $h(x) = \max(\min(\underline{m}'(x), 0), \overline{m}'(x))$, consider $\overline{m}'(x) \leq \overline{m}(x)$ and $\underline{m}'(x) \geq \underline{m}(x)$.

If $\overline{m}'(x) > 0$, $0 \leq h(x) \leq \overline{m}'(x) \leq \overline{m}(x)$, and we can then say that $h(x) \leq \overline{m}(x)$. Similarly, if $\overline{m}'(x) < 0$, then $0 \geq \overline{m}(x) \geq \overline{m}'(x) = h(x)$ by definition of h. If $\overline{m}'(x) = 0$, then lastly we see $\overline{m}(x) \geq \overline{m}'(x) = h(x) = 0$.

Therefore, $\overline{m}(x) \geq \overline{m}'(x) \geq h(x)$. It remains to be seen that $\underline{m}(x) \leq \underline{m}'(x) \leq h(x)$. If $\underline{m}'(x) > 0$, then we can see $\underline{m}'(x) = h(x) \geq \underline{m}(x)$ since $\underline{m}'(x) \leq \overline{m}'(x)$ for all $x \in \mathcal{I}$. If $\underline{m}'(x) < 0$, we can see that $0 \geq h(x) \geq \underline{m}'(x) \geq \underline{m}(x)$. Thus, we observe that $h(x) \geq \underline{m}'(x)$.

Statement 3: By Lemma 30, for all $g, f \in \mathcal{F}$ so that $g < 0 < f$ on the interval $\mathcal{I}' \subseteq \mathcal{I}_c$, $Z = \{x \in \mathcal{I}' : \frac{d}{dx} \log(-g(x)) = \frac{d}{dx} \log(f(x))\}$ is totally disconnected. By definition of $\overline{m}$ and $\underline{m}$, we conclude $\underline{m}(x) < \overline{m}(x)$ on all but a totally disconnected set. By inspection, $h(x) = \overline{m}(x)$ if and only if $h(x) = \underline{m}(x)$, as this is the only way that $\overline{m}(x) = \overline{m}'(x)$ and $\underline{m}(x) = \underline{m}'(x)$, which are used to construct h. Thus, $\underline{m}(x) < h(x) < \overline{m}(x)$ for all $x \in Z$, which is totally disconnected.

Statement 4: Let $\mathcal{I}'$ be a compact subinterval of $\mathring{\mathcal{I}}$. By Lemma 31, for every $f \in \mathcal{F}$ we have ϵ_f and M_f such that for every $x \in \mathcal{I}' \cap \mathcal{I}_c$, if (i) $|f(x)| \geq \epsilon_f$, $|\frac{d}{dx}\log|f(x)|| < M_f$, and (ii) if $|f(x)| < \epsilon_f$, $\frac{d}{dx}f(x) \geq 0$. Note that similar to the proof of Lemma 31, for ϵ' sufficiently small, we will have for all $x \in \mathcal{I}'$ that there is at least one $f \in \mathcal{F}$ such that $|f(x)| > \epsilon'$. (This uses the uniqueness of roots by Condition 2 and continuity of each element of $\mathcal{F}$ by Condition 1.) Finally, let $\epsilon = \min(\epsilon', \min_f \epsilon_f)$ and $M = \max_f M_f$.

Now suppose $|f(x)| < \epsilon$. If $f(x) > 0$, then $\frac{d}{dx}\log f(x) = \frac{d}{dx}f(x)/f(x) \geq 0$ as $\frac{d}{dx}f(x) \geq 0$ by (i). Similarly, if $f(x) < 0$, then $\frac{d}{dx}\log -f(x) = \frac{d}{dx}f(x)/f(x) \leq 0$. We conclude that for any $x \in \mathcal{I}_c$, either $|\overline{m}(x)| < M$ or $\underline{m}(x) \leq 0$, by definition of $\underline{m}$. Similarly, for any $x \in \mathcal{I}_c$, either $|\overline{m}(x)| < M$ or $\overline{m}(x) \geq 0$. Moreover, as $\epsilon \leq \epsilon'$, for all $x \in \mathcal{I}' \cap \mathcal{I}_c$ at least one function $f \in \mathcal{F}$ has $|f(x)| \geq \epsilon$, and we must have $|\overline{m}(x)| < M$ or $|\underline{m}(x)| < M$.

Let us then consider three cases:

- $|\overline{m}(x)| < M$ and $|\underline{m}(x)| < M$. By definition of h, it is clear that $|h(x)| < M + 1$.
- $|\overline{m}(x)| < M$ and $\underline{m}(x) \leq 0$. Here $|\overline{m}'(x)| < M + 1$ and either $\underline{m}'(x) \leq 0$ or $0 \leq \underline{m}'(x) \leq \underline{m}(x) \leq \overline{m}(x)$ (by Lemma 30, we have $\overline{m} \geq \underline{m}$), and in both cases $0 \leq h(x) \leq M + 1$.
- $\overline{m}(x) \geq 0$ and $|\underline{m}(x)| < M$. Symmetrically, we have $-M - 1 \leq h(x) \leq 0$.

We conclude that $|h(x)| < M + 1$ on $\mathcal{I}'$. As $\mathcal{I}'$ is arbitrary, we then observe $|h(x)| < M + 1$ on any compact subinterval of $\mathcal{I}$.

□

Claim 17. *The function h constructed in eq. (4.7) is Lebesgue integrable and λ constructed in eq. (4.8) is continuously differentiable for all $x \in \mathcal{I}_c$. Moreover, λf is continuous on $\mathring{\mathcal{I}}$ for all $f \in \mathcal{F}$.*

Proof. Observe that the function h is bounded on every compact subinterval $\mathcal{I}'$ of $\mathring{\mathcal{I}}$ by Statement (4) of Claim 16, and is continuous on $\mathcal{I}_c$ by Statement (1). As $\mathcal{I}_c = \mathcal{I} \setminus \hat{S}$ has full Lebesgue measure over $\mathring{\mathcal{I}}$, h is continuous almost everywhere on $\mathring{\mathcal{I}}$. Moreover, for every compact subinterval $\mathcal{I}' \subseteq \mathring{\mathcal{I}}$, the Riemann integral $\int_{\mathcal{I}'} h(z)dz$ exists and is bounded. Thus, the improper Riemann integral

$\lim_{r\to a^-}\int_{r_0}^{r} h(z)dz$ exists for $a \in \partial\mathcal{I}$ and is equal to the Lebesgue integral $\int_{r_0}^{a} h(z)dz$, as shown by Apostol [7, Theorem 10.33]; note however that we may have $\int_{r_0}^{a} h(z)dz = \infty$.

As Pouso [69] shows the Fundamental Theorem of Calculus can be applied to Lebesgue integrals and h is Lebesgue integrable on $\mathring{\mathcal{I}}$, $\varphi(r) = \int_{r_0}^{r} h(z)dz$ is differentiable where h is continuous, shown by Abbott [1, Theorem 7.5.1]. By statement (1) of Claim 16, we also know h is continuous on $\mathcal{I}_c$, so we know φ is differentiable on $\mathcal{I}_c$, and additionally observe $\frac{d}{dx}\varphi(x) = h(x)$. Thus we can say λ is differentiable for all $x \in \mathcal{I}_c$, and $\frac{d}{dx}\lambda(x) = h(x)\lambda(x)$. Therefore, at every $x \in \mathcal{I}_c$, λf is the product of two differentiable functions, and is therefore differentiable on $\mathcal{I}_c$.

Moreover, as the set of discontinuities of h is finite as $\hat{S}$ is finite, so we know φ is then continuous on $\mathring{\mathcal{I}}$. λ is then continuous on $\mathring{\mathcal{I}}$ as it is the exponent of a continuous function. As f is assumed continuous in Condition 1, the product λf is continuous on $\mathcal{I}'$.

(We note that *any* function h satisfying the statements in Claim 16 would be Lebesgue intregrable on $\mathring{\mathcal{I}}$, but the proof in Claim 16 is only provided for h as constructed in Equation 4.7.) □

Claim 18. *For every $x \in \mathcal{I}_c$ and every $f \in \mathcal{F}$, the derivative of λf at x is nonnegative, and 0 at a totally disconnected set of values.*

Proof. We observe $\frac{d}{dx}\lambda(x) = \lambda(x)h(x)$ for any $x \in \mathcal{I}_c$ given the construction of λ in Equation (4.8). By substitution, we now have

$$\tfrac{d}{dx}(\lambda f)(x) = \tfrac{d}{dx}\lambda(x)f(x) + \lambda(x)\tfrac{d}{dx}f(x) = \lambda(x)h(x)f(x) + \lambda(x)\tfrac{d}{dx}f(x)\ . \tag{4.9}$$

As $\lambda(x) > 0$, it therefore suffices to show $h(x)f(x) + \frac{d}{dx}f(x) \geq 0$ for any $f \in \mathcal{F}$.

Recall from Condition 1 that if $f(x) = 0$, then there is some open neighborhood U around x where $\frac{d}{dz}f(z) \geq 0$ for all $z \in U$. By Claim 16, Statement 2, and the definitions of $\overline{m}$ and $\underline{m}$, for all $x \in \mathcal{I}_c$, $g \in \mathcal{F}_-(x)$, and $f \in \mathcal{F}_+(x)$, we have

$$-\tfrac{d}{dx}\log(-g(x)) \geq \underline{m}(x) \geq h(x) \geq \overline{m}(x) \geq -\tfrac{d}{dx}\log(f(x))\ , \tag{4.10}$$

where the inequalities involving h are strict at all but a totally disconnected set by Statement 3.

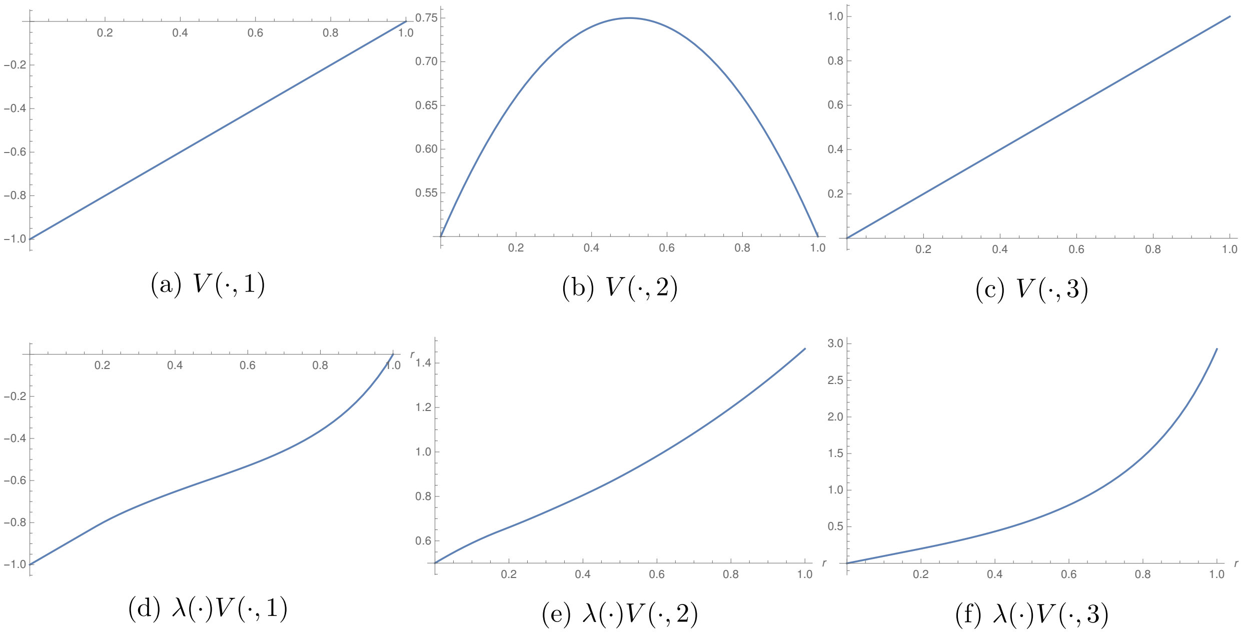
Writing $\frac{d}{dx}\log(f(x))$ as $\frac{\frac{d}{dx}f(x)}{f(x)}$, we can see that for any $f \in \mathcal{F}_+(x)$, $h(x)f(x) \geq -\frac{d}{dx}f(x)$ by some mild maniupulation of Statement (2) of Claim 16. Similarly, for $g \in \mathcal{F}_-(x)$, as g is negative, we can see that $h(x)g(x) \geq -\frac{d}{dx}g(x)$. Therefore, for all $f \in \mathcal{F}$ and $x \in \mathcal{I}_c$, we observe that $h(x)f(x) \geq -\frac{d}{dx}f(x)$, with equality only at a totally disconnected set of points by Statement (3) of Claim 16, as for all $f \in \mathcal{F}$, $h(x)f(x) > -\frac{d}{dx}f(x) \iff \overline{m}(x) > h(x) > \underline{m}(x)$ at all but a totally disconnected set. We stated earlier that it sufficed to show $h(x)f(x) > -\frac{d}{dx}f(x)$ for all $f \in \mathcal{F}$ in order to conclude $\frac{d}{dx}(\lambda f)(x) \geq 0$ for all x in $\mathcal{I}_c$. □

We now have the groundwork we need to prove the Proposition.

Proposition 11. *If $\mathcal{F}$ satisfies Condition 1, 2, and 3, then there exists a function $\lambda : \mathcal{I} \to \mathbb{R}_+$ so that λf is increasing over $\mathring{\mathcal{I}}$ for every $f \in \mathcal{F}$.*

Proof. Claim 17 shows that λf is continuous on $\mathring{\mathcal{I}}$ and continuously differentiable on $\mathcal{I}_c$. Claim 18 shows $\frac{d}{dx}(\lambda f)(x) \geq 0$ with $\frac{d}{dx}(\lambda f)(x) = 0$ at a totally disconnected set of $x \in \mathcal{I}_c$. As $\hat{S}$ is finite, note that $\mathcal{I}_c$ is then the disjoint union of open intervals $\mathcal{I}_1, \ldots, \mathcal{I}_k$. By Lemma 29, we have that λf is strictly increasing on each $\mathcal{I}_i$. As λf is continuous on $\mathring{\mathcal{I}}$, we further conclude that λf is strictly increasing on the closure of each $\mathcal{I}_i$ except on the boundary of $\mathcal{I}$; that is, λf is strictly increasing on $\mathrm{cl}(\mathcal{I}_i) \setminus \partial\mathcal{I}$. Finally, writing $\mathring{\mathcal{I}} = \cup_{i=1}^k \mathrm{cl}(\mathcal{I}_i) \setminus \partial\mathcal{I}$, we see that λf is strictly increasing on all of $\mathring{\mathcal{I}}$. □

This concludes the proof of Proposition 11.



(a) $V(\cdot, 1)$

(b) $V(\cdot, 2)$

(c) $V(\cdot, 3)$

(d) $\lambda(\cdot)V(\cdot, 1)$

(e) $\lambda(\cdot)V(\cdot, 2)$

(f) $\lambda(\cdot)V(\cdot, 3)$

Figure 4.1: The functions $V(\cdot, y)$ are not always increasing for all $y \in \mathcal{Y}$, but our function λ "monotonizes" them, as shown in (d)–(f).

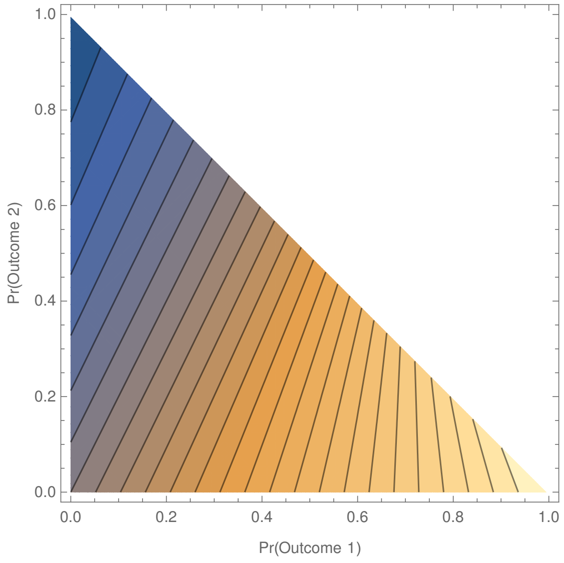


Figure 4.2: Level sets of Γ

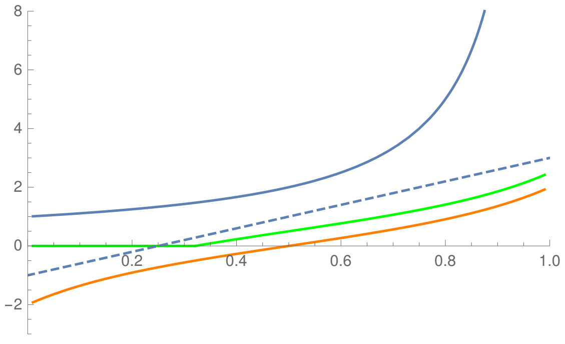


Figure 4.3: $\overline{m}(\cdot)$ in solid blue, $\underline{m}(\cdot)$ in orange, valid $h(\cdot)$ in green and dashed blue.

Chapter 5

Embedding Dimension of Polyhedral Surrogates

5.1 Introduction

In Chapter 1, we mentioned and discussed three desiderata of surrogate losses: convexity, consistency, and efficiency. This chapter focuses on the setting of Quadrant 1 of Table 2.1, and takes convexity and consistency as prerequisites, now asking how efficient a loss can be for a given task. We ask, given ℓ, how do we design consistent and convex surrogates $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$? In particular, this chapter examines *polyhedral embeddings*, which are piecewise-linear and convex embeddings over the prediction space $\mathbb{R}^d$. Chapter 3 shows that polyhedral surrogates are in a strong sense related to an *embedding* mapping of each discrete prediction r to some point $\varphi(r) \in \mathbb{R}^d$, which optimizes L-loss if and only if r optimizes ℓ-loss. While every discrete loss ℓ can be embedded, in the worst case the prediction dimension required is $d = n - 1$ where n is the number of possible labels; this may be exponentially high in some settings, such as structured prediction. This chapter studies the prediction dimension of polyhedral embeddings such as structured prediction. A prediction dimension d significantly below n, such as $O(\log n)$ for classification with an abstain option [72], can lead to faster downstream optimization and computation, an effect that grows with n; thus, we seek to understand for which target losses this dimension can be low.

This chapter defines and investigates the embedding dimension of discrete losses, and characterizes the d-embeddable losses for each d. Beginning with $d = 1$, i.e. embedding into the real line via $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$, we offer a complete characterization via a variety of conceptual and testable/constructive conditions (§ 5.3). Perhaps surprisingly, for $d = 1$, if *any* convex calibrated surrogate exists, then in particular a polyhedral one does. In higher dimensions, we observe a general characterization for d-embeddability in terms of certain optimality and monotonicity conditions (§ 5.4). In particular, we isolate and investigate the optimality condition, which we significantly reduce from a search over sets of polytopes to a quadratic feasibility program (Definition 37), yielding a new technique to prove lower bounds on the embedding dimension. Finally, we apply our characterizations to show new lower bounds on the embedding dimension for abstain loss, whose convex calibration dimension has been well-studied [71, 72] (§ 5.5). Both the 1-dimension characterization and higher-dimensional quadratic program obtain previously unknown lower bounds on embedding dimension and convex calibration dimension (cf. [71]).

The work in this chapter comes heavily from Finocchiaro et al. [29], published at COLT 2020.

5.2 Setting

We now formalize the notion of embedding dimension studied in this chapter. In order to concisely discuss embedding dimension, we appeal to notation and terminology from the field of property elicitation [46, 55, 56, 66, 79], relating it to the language of calibrated surrogates as needed; for intuition this is a translation from Quadrant 1 to Quadrant 3 in Table 2.1. Recall that in the Quadrant 1 setting of Table 2.1, Tewari and Bartlett [84] show that calibration (Definition 8) is necessary and sufficient for consistency. For simplicity, we assume the given discrete loss is *non-redundant*, meaning every report r uniquely minimizes expected loss for some distribution $p \in \Delta_\mathcal{Y}$.

In Chapter 3, we studied the notion of embedding. For a quick recap, a surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ embeds a discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ if there is an injection $\varphi : \mathcal{S} \to \mathbb{R}^d$ such that $\varphi(r)$ is L-optimal if and only if r is ℓ optimal for every $r \in \mathcal{S}$, where $\mathcal{S} \subseteq \mathcal{R}$ such that $\mathrm{prop}_{\Delta_\mathcal{Y}}[\ell](p) \cap \mathcal{S} \neq \emptyset$ for all $p \in \Delta_\mathcal{Y}$.

Definition 33 (Embedding dimension)**.** *We say a discrete loss $\ell : \mathcal{R} \to \mathbb{R}^\mathcal{Y}_+$ is d-embeddable if there exists a polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^\mathcal{Y}_+$ that embeds it. The embedding dimension of ℓ is the* *smallest d such that ℓ is d-embeddable.*

Several works have investigated the problem of reducing the prediction dimension of a surrogate loss while maintaining correctness desiderata, such as convex and consistency.

A number of upper bounds on embedding dimension are already known. Many surrogates in the literature provide upper bounds; we highlight in particular the *abstain loss* [72] in eq. (5.3), in which one wants to predict the most likely outcome *only if* confident in the outcome, and otherwise abstain. In general (e.g. Ramaswamy and Agarwal [71, Corollary 13]), a known convex-conjugate construction generically embeds any discrete loss on $\mathcal{Y} = [n]$ into $d = n - 1$ dimensions, giving a flat upper bound of $n - 1$ on embedding dimension.

Lower bounds exist but are rare. A lower bound on the dimensionality of *any* calibrated convex surrogate L implies in particular a lower bound on polyhedral surrogates. Ramaswamy and Agarwal [71] give such a lower bound via the technique of *feasible subspace dimension*, which is able to e.g. prove that embedding 0-1 loss on n labels requires dimension $n - 1$. However, this technique gives only the trivial $d \geq 1$ for the abstain family of losses above when $\alpha \leq 1/2$ because of their geometric structure. For an overview of different prediction dimension metrics, see § 2.4.

5.3 One-dimensional embeddings

We first give a complete characterization when a discrete loss ℓ can be embedded into the real line, i.e., when ℓ is 1-embeddable. Our first characterization is expressed in terms of the property γ that ℓ elicits, stating that ℓ is 1-embeddable if and only if γ is *orderable*, meaning the adjacency graph of its level sets is a path. For example, this characterization will immediately imply that embedding the abstain losses on $n \geq 3$ outcomes requires $d \geq 2$ dimensions (§ 5.5.2). While determining these adjacencies can be straightforward when ℓ has known symmetries, we also give a more constructive algorithm for testing 1-embeddability and constructing a 1-dimensional polyhedral surrogate. Finally, we show that the existence of any 1-dimensional convex calibrated implies 1-embeddability, showing that embeddings are without loss of generality in dimension 1.

After presenting and discussing this sequence of results, we observe that they can be collected as a set of six conditions on ℓ (Theorem 12) that are all pairwise equivalent, and in particular, are equivalent to 1-embeddability.

5.3.1 General characterization via property elicitation

We begin with conditions on the property elicited by a discrete loss. The following condition of Lambert [55, Theorem 3], that a finite property is *orderable*, states that any two level sets intersect in a hyperplane, or not at all.

Definition 34 (Orderable). *A finite property $\gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}}$ is orderable if there is an enumeration of $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$ such that for all $i \leq |\mathcal{R}| - 1$, we have $\gamma_{r_i} \cap \gamma_{r_{i+1}}$ is a hyperplane intersected with $\Delta_{\mathcal{Y}}$.*

In fact, we show that orderability characterizes 1-embeddability.

Theorem 10. *A discrete loss ℓ is 1-embeddable if and only if the property it elicits is orderable.*

We now give an equivalent condition to orderability which may be more intuitive: the adjacency graph of the level sets of γ, formed by connecting reports if their level sets intersect, must be a path. This graph can be easily established for discrete losses with known symmetries or other facts, such as abstain, the mode, or ranking losses.

Definition 35 (Intersection graph). *Given a discrete loss ℓ and associated finite property $\gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}}$ elicited by ℓ, the intersection graph has vertices $\mathcal{R}$ with an edge (r, r') if $\gamma_r \cap \gamma_{r'} \cap \mathrm{relint}(\Delta_{\mathcal{Y}}) \neq \emptyset$, where $\mathrm{relint}(\Delta_{\mathcal{Y}})$ is the relative interior of $\Delta_{\mathcal{Y}}$.*

If one can visualize level sets of a property, constructing the intersection graph yields an intuitive way to conceptualize orderability by Proposition 12.

Proposition 12. *A finite property γ is orderable iff its intersection graph is a path, i.e. a connected graph where two nodes have degree 1 and every other node has degree 2.*

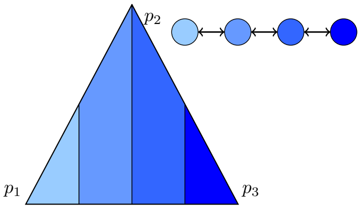


Figure 5.1: Level sets and intersection graph for a given property, $|\mathcal{Y}| = 3$.

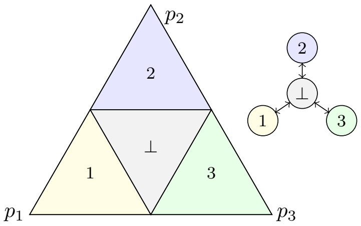


Figure 5.2: Level sets and intersection graph for the abstain$_{1/2}$ property, $\mathcal{Y} = \{1, 2, 3\}$.

Proof. ($\implies$) The intersection graph is constructed by adding an edge for each halfspace, which connects only two nodes. If three level sets intersected on relint$(\Delta_{\mathcal{Y}})$, then the level boundary for any two would not be a halfspace (this follows because the level sets form a power diagram, e.g. [56]). This yields a path for the intersection graph.

($\impliedby$) If the intersection graph forms a path, then we can enumerate the vertices from source to sink as $r_1, \ldots, r_{|\mathcal{R}|}$. The level sets are full-dimensional (in the simplex) convex polytopes whose intersections only occur in the relative boundary, as they form cells of a power diagram. Since γ_{r_1} intersects only with γ_{r_2} on the relative interior of the simplex, and both sets are convex, this intersection must be a hyperplane intersecting the simplex. (Otherwise, one of the sets would not be convex, or γ_{r_1} would intersect with some other level set on the relative interior. We can now "delete" γ_{r_1}, more formally, consider the convex polytope $\gamma_{r_2} \cup \ldots \cup \gamma_{r_{|\mathcal{R}|}}$. The same argument now applies to γ_{r_2}, giving that it is intersects with γ_{r_3} along a hyperplane intersected with the simplex; and so on. □

Combining Proposition 12 and Theorem 10, we see that in order for ℓ to be embedded onto the real line, it is necessary and sufficient for the intersection graph of the property γ to be a path. We give an example of a direct application in § 5.5.1. By visualizing the level sets of γ as a power diagram (generalization of Voronoi diagram) in the simplex like Lambert and Shoham [56], we can also use Proposition 12 to perform a visual test for orderability, and thus 1-embeddability (Figures 5.1 and 5.2).

5.3.2 Constructing a surrogate

While Theorem 10 and Proposition 12 are quite useful for discrete losses with known symmetries, they do not immediately provide an algorithm to test 1-embeddability of an arbitrary discrete loss ℓ, nor to construct the convex loss L which embeds it. We now turn to an algorithmic test, which actually builds a real-valued polyhedral calibrated surrogate in the event that ℓ is 1-embeddable that "stitches" linear functions together using weights Λ to insure continuity.

Theorem 11. *Let $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ be a discrete loss. Then ℓ is 1-embeddable if and only if there is an ordering $\mathcal{R} = \{r_1, \ldots, r_k\}$ of the reports such that the following two conditions hold, where $v(i)_y := \ell(r_i, y) - \ell(r_{i-1}, y)$:*

(1) For all $y \in \mathcal{Y}$, the sequence $\operatorname{sgn}(v(i)_y)$ is monotone in $i \in \{1, \ldots, k-1\}$,

(2) For all $i \in \{2, \ldots, k-1\}$

$$\lambda^-(i) = \min\left\{\frac{v(i)_y}{v(i+1)_y} : y \in \mathcal{Y}, v(i)_y, v(i+1)_y < 0\right\}$$
$$\lambda^+(i) = \max\left\{\frac{v(i)_y}{v(i+1)_y} : y \in \mathcal{Y}, v(i)_y, v(i+1)_y > 0\right\} ,$$

we have $\lambda^-(i) \geq \lambda^+(i)$. (We adopt the convention $\max(\emptyset) = -\infty$, $\min(\emptyset) = +\infty$.)

Moreover, when these conditions hold, the loss $L : \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}$ embeds ℓ with $\varphi : \mathcal{R} \to \mathbb{R}$, where

$$\varphi(r_i) = \sum_{j=1}^{i-1} 1/\Lambda_j \ , \quad \text{(where } \varphi(r_1) = 0\text{)}$$

$$L(u, y) = \begin{cases} \ell(r_1, y) - uK & u \leq \varphi(r_1) = 0 \\ \ell(r_i, y) + \Lambda_i \cdot (u - \varphi(r_i)) \cdot (\ell(r_{i+1}, y) - \ell(r_i, y)) & u \in [\varphi(r_i), \varphi(r_{i+1})] \\ \ell(r_k, y) + \Lambda_{k-1} \cdot (u - \varphi(r_k)) \cdot K & u \geq \varphi(r_k) \end{cases} ,$$

where $\lambda(i) = \min(\lambda^+(i), \max(\lambda^-(i), 1))$, $\Lambda_i := \prod_{j=2}^{i} \lambda(j)$, $\Lambda_1 = 1$, and $K = \max_{i \in \{2,\ldots,k\}, y \in \mathcal{Y}} |v(i)_y|$.

As intuition for the proof, note that the conditions of the theorem ensure the existence of a positive multiplier $\lambda(i)$ making $v(i) \leq \lambda(i)v(i+1)$ hold coordinate-wise; our choice of $\lambda(i)$ is but one option. The construction of L sets the left and right derivatives at an embedding point $\varphi(r_i)$ to be positive multiples of $v(i)$ and $v(i+1)$, respectively, using this inequality to maintain monotonicity, and hence convexity of L. The vectors $v(i), v(i+1)$ are chosen precisely to give the correct optimality conditions, so that for a given distribution, r_i is optimal for ℓ if and only if $\varphi(r_i)$ is optimal for L. The reverse direction, showing that these conditions are necessary for 1-embeddability, is much more involved. We can easily construct a link function in the case of $d = 1$, by taking the midpoints between the embedding points as cutoffs: $\psi(u) = \arg\min_{r \in \mathcal{R}} |u - \varphi(r)|$, breaking ties arbitrarily. Moreover, this ψ is an ϵ-thickened link via Construction 1.

We summarize the above results in the following theorem, together with one additional result: if ℓ has any calibrated convex surrogate at all, it must have a polyhedral one.

Theorem 12. *Let ℓ be a discrete loss eliciting a finite property γ. The following are equivalent: (1) γ is orderable; (2) the intersection graph of γ is a path; (3) the two conditions of Theorem 11 are satisfied; (4) ℓ is* 1*-embeddable; (5) ℓ has some polyhedral calibrated surrogate loss $L : \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}_{+}$; (6) ℓ has some convex calibrated surrogate loss $L : \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}_{+}$.*

For the proof, note that (1) $\iff$ (2) was shown in Proposition 12, while (4) $\iff$ (5) follows from Theorem 1, and (5) $\implies$ (6) is immediate from the definitions. [29, Theorem 1] therefore proves (1) $\implies$ (3) $\implies$ (5) and (6) $\implies$ (1).

5.4 Higher dimensions

Our characterization of 1-embeddable losses reveals a large class of properties are not 1-embeddable. In this section, we develop a characterization of d-embeddable discrete losses for $d \geq 2$. We begin with some basic facts and definitions about polytopes and their Minkowski sums, which naturally arise when considering the subgradients of a polyhedral surrogate loss (§ 5.4.1). From these definitions, we can state a somewhat immediate characterization of d-embeddable losses in terms of polytopes that satisfy certain optimality and monotonicity conditions (Theorem 13). We then explore the optimality condition further, and through facts about Minkowski sums, slowly remove mentions of polytopes from the condition until we arrive at a quadratic feasibility program to test whether such polytopes exist (Theorem 15). From our main characterization, dropping the monotonicity condition, this program gives a novel necessary condition for d-embeddability, yielding new lower bounds for embedding dimension (Corollary 14).

5.4.1 Setup: subgradient sets at embedding points.

Recall that if $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ with representative set $\mathcal{S}$ is embedded by $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, then each $r \in \mathcal{S}$ is embedded at some point $\varphi(r) \in \mathbb{R}^d$. In particular, $\varphi(r)$ must minimize $\mathbb{E}_p L(\cdot, Y)$ if and only if r minimizes $\mathbb{E}_p \ell(\cdot, Y)$. The key to our approach is to study as first-class objects the sets of all subgradients[1] of L at these embedding points. The question of whether a calibrated polyhedral surrogate exists in d dimensions essentially reduces to conditions on these sets alone. In particular, we use the fact that a convex function is minimized at u if and only if $\vec{0}$ is in its subdifferential (subgradient set) at u. Therefore, we consider collections of sets T^r_y, which intuitively aspire to be the subdifferentials of a calibrated polyhedral surrogate $L(\cdot, y)$ at $\varphi(r)$, denoted $\partial L(\varphi(r), y)$. Throughout, we often take r as implicit and suppress it from our notation for ease of exposition. Note that if $L(\cdot, y)$ is a polyhedral function on $\mathbb{R}^d$, then all of its subgradient sets are (bounded) closed polytopes [77].

Definition 36 ($\mathcal{T}$, $D(\mathcal{T})$)**.** *We write $\mathcal{T} = \{T_y \subseteq \mathbb{R}^d : y \in \mathcal{Y}\}$ to denote a collection of closed polytopes, with implicit parameter d. Given a distribution $p \in \Delta_{\mathcal{Y}}$, we write the p-weighted Minkowski sum of $\mathcal{T}$ as*

$$\oplus_p \mathcal{T} := \bigoplus_{y \in \mathcal{Y}} p_y T_y = \left\{ \sum_{y \in \mathcal{Y}} p_y x_y \;\middle|\; x_y \in T_y \; \forall y \in \mathcal{Y} \right\} ,$$

or in other words, the Minkowski sum of the scaled sets $\{p_y T_y : y \in \mathcal{Y}\}$. Finally, we associate with $\mathcal{T}$ a set of distributions $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : \vec{0} \in \oplus_p \mathcal{T}\}$.

Note that $T_y = T^r_y$ for some $r \in \mathcal{R}$; here, we are agnostic to the choice of r, so we omit its notation for clarity. The importance of the p-weighted Minkowski sum and of $D(\mathcal{T})$ are that they

[1] Recall that a subgradient of e.g. the convex function $L(\cdot, y) : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ at a point u is a vector $v \in \mathbb{R}^d$ such that $L(u', y) \geq L(u, y) + \langle v, u' - u \rangle$ for all u'.

capture the distributions p for which a point u minimizes expected loss, whenever $\mathcal{T}$ corresponds to the subgradient sets of some polyhedral L at u. In other words, under these conditions, if $T_y = T_y^{\varphi(r)}$, we have $D(\mathcal{T}) = \Gamma_{\varphi(r)}$, the level set for $\varphi(r)$ of the property Γ elicited by L embedding ℓ.

Lemma 32. *Let $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ be a polyhedral loss eliciting a property Γ. If for all $y \in \mathcal{Y}$ we have $T_y = \partial L(u, y)$ at some point $u \in \mathbb{R}^d$, then $D(\mathcal{T}) = \Gamma_u$.*

Proof. Recall that a convex function f is minimized at $u = \varphi(r)$ if and only if $\vec{0} \in \partial f(u)$. We thus have $p \in \Gamma_u \iff u \in \arg\min_{u'} \mathbb{E}_p L(u', Y) \iff \vec{0} \in \partial \mathbb{E}_p L(u, Y) = \bigoplus_{y \in \mathcal{Y}} p_y \partial L(u, y) = \oplus_p \mathcal{T} \iff p \in D(\mathcal{T})$. Here we used the basic fact that if f_1, f_2 are convex with subgradient sets T_1, T_2 at u, then $\alpha f_1 + \beta f_2$ has subgradient set $\alpha T_1 \oplus \beta T_2$, the Minkowski sum of the scaled sets. □

This fact will be vital for characterizing when ℓ is correctly embedded by some L whose subgradient sets are $\mathcal{T}^r$ for each $r \in \mathcal{R}$.

5.4.2 General characterization

We now give a general characterization of when a discrete loss ℓ can be embedded into d dimensions, i.e. when a consistent polyhedral surrogate $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ exists. Two conditions are required: *optimality* and *monotonicity*. Optimality enforces that the surrogate is minimized precisely when and where it should be. It says that for each discrete prediction r and set of distributions γ_r for which it is ℓ-optimal, there exists a collection of polytopes $\mathcal{T}^r$ such that, were they the subgradients of some polyhedral surrogate L at some point $\varphi(r)$, then $\varphi(r)$ would be L-optimal at the same set of distributions γ_r; more succinctly in light of Lemma 32, we require $D(\mathcal{T}^r) = \gamma_r$. Monotonicity says that these individual polytopes can indeed be glued together to form the subgradients of some convex loss function L.

Theorem 13. *Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ be a discrete loss with, for each $r \in \mathcal{R}$, $\gamma_r = \{p \in \Delta_{\mathcal{Y}} \mid r \in \arg\min_{r'} \mathbb{E}_p \ell(r, Y)\}$. Then ℓ is d-embeddable if and only if there exists a collection of polytopes $\mathcal{T}^r = \{T_y^r : y \in \mathcal{Y}\}$ for each $r \in \mathcal{R}$ such that both of the following hold:*

(1) (Optimality) For each r, we have $D(\mathcal{T}^r) = \gamma_r$.

(2) (Monotonicity) There exists an injective embedding function $\varphi : \mathcal{S} \to \mathbb{R}^d$ and loss functions $\{L(\cdot, y) : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}\}_{y \in \mathcal{Y}}$ such that for all $r \in \mathcal{R}$ and $y \in \mathcal{Y}$, we have $T^r_y = \partial L(\varphi(r), y)$ and for all $r \in \mathcal{R}$, we have $L(\varphi(r), y) = \ell(r, y)$.

Proof. ($\implies$) L embeds ℓ implies $\gamma_r = \Gamma_\varphi(r) = D(\mathcal{T}^r)$ by Lemma 32 for all $r \in \mathcal{R}$, thus we have optimality. Monotonicity follows directly from the embedding definition.

($\impliedby$) The first two embedding conditions hold by the assumption of φ in the monotonicity condition. The third condition is $\gamma_r = \Gamma_{\varphi(r)}$ for all r. From optimality, we have $\gamma_r = D(\mathcal{T}^r)$. Taking $\mathcal{T}^r = \{T^r_y : y \in \mathcal{Y}\}$, Lemma 32 implies that $\Gamma_{\varphi(r)} = D(\mathcal{T}^r) = \gamma_r$. □

5.4.3 Characterizing optimality

We now focus entirely on the optimality condition of Theorem 13, for two purposes. First, we aim to greatly narrow the search space for constructing low-dimensional surrogate loss functions for a given discrete loss. The tools we construct in this section aid in this task by constraining or constructing feasible subgradient sets $\mathcal{T}$ given a level set γ_r. Second, we wish to prove impossibilities, i.e., lower bounds on the embedding dimension of a given discrete loss (an apparently hard problem). For such lower bounds, it suffices to drop monotonicity from Theorem 13, leaving us with an independent optimality condition for each $r \in \mathcal{R}$, and show that for any one $r \in \mathcal{R}$, we could not have d-dimensional polytopes $\mathcal{T}$ satisfying $D(\mathcal{T}^r) = \gamma_r$.

At first glance, the optimality condition seems difficult to operationalize, as it involves the existence of polytopes, and even if said polytopes are given, it is unclear how to test whether $D(\mathcal{T}^r) = \gamma_r$. To begin, consider the latter problem, of understanding the set $D(\mathcal{T})$ in terms of descriptions of $\mathcal{T}$, and in particular, of writing conditions on $\mathcal{T}$ such that $D(\mathcal{T})$ is equal to a given polytope $C \subseteq \Delta_{\mathcal{Y}}$. We know that, by writing C in its halfspace and vertex representations, respectively, we can give two such conditions.

Condition 4 (Halfspace condition)**.** *A collection of polytopes $\mathcal{T}$ and a polytope $C \subseteq \Delta_\mathcal{Y}$ defined by $C = \{p \in \Delta_\mathcal{Y} : Bp \geq \vec{0}\}$ satisfy the halfspace condition if there exist $v_1, \ldots, v_k \in \mathbb{R}^d$ such that, for all $i \in [k]$ and $y \in \mathcal{Y}$, for all $x \in T_y$, we have $\langle v_i, x\rangle \leq B_{iy}$.*

Condition 5 (Vertex condition)**.** *A collection of polytopes $\mathcal{T}$ and a polytope $C \subseteq \Delta_\mathcal{Y}$ defined by $C = \mathrm{conv}(\{p^1, \ldots, p^l\})$ satisfy the vertex condition if for all $j \in [l]$, $0 \in \oplus_{p^j}\mathcal{T}$.*

We now show that, for any convex polytope C, satifying vertex and halfspace conditions are necessary and sufficient for C being equal to $D(\mathcal{T}) = \{p \in \Delta_\mathcal{Y} \mid \vec{0} \in \oplus_p\mathcal{T}\}$.

Theorem 14. *Let the polytopes $\mathcal{T} = \{T_y \subseteq \mathbb{R}^d : y \in \mathcal{Y}\}$ and C be given, with $C = \mathrm{conv}(\{p^1, \ldots, p^l\}) = \{p : Bp \geq \vec{0}\}$ for $B \in \mathbb{R}^{k\times n}$. We have $D(\mathcal{T}) = C$ if and only if both the halfspace and vertex conditions hold.*

Proof. ($\implies$) Suppose $D(\mathcal{T}) = C$. First, we note that the vertex condition is immediate: For all $j \in [\ell]$, $p^j \in C$ which gives $p^j \in D(\mathcal{T})$. To show the halfspace condition is satisfied, we first construct a matrix E such that $Ep \geq 0 \iff Bp \geq 0$, then use this construction to pick out the necessary vectors $v_1, \ldots, v_k$.

By Lemma 33, there is a finite collection of vectors $w_1, \ldots, w_K \in \mathbb{R}^d$ and such that $\vec{0} \in \oplus_p\mathcal{T}$ if and only if, for all w_i, $\sum_y p_y \max_{x\in T_y}\langle w_i, x\rangle \geq 0$. Hence, each vector w_i generates a row of a matrix $E \in \mathbb{R}^{K\times n}$ with $E_{iy} = \max_{x\in T_y}\langle w_i, x\rangle$, and we have $p \in D(\mathcal{T}) \iff Ep \geq 0$. By assumption of $D(\mathcal{T}) = C$, then, we have $Ep \geq 0 \iff Bp \geq 0$. By Lemma 35, because B has the minimum possible number of rows, each row of B appears (scaled by some positive constant) as a different row of E. Taking the collection of w_i corresponding to these rows and rescaling them by that positive constant, we get a collection of k vectors that we can rename $v_1, \ldots, v_k \in \mathbb{R}^d$, with $\max_{x\in T_y}\langle v_i, x\rangle = B_{iy}$, hence the halfspace condition is satisfied.

($\impliedby$) Suppose Conditions 4 and 5 hold. Then by the vertex condition, $p^j \in D(\mathcal{T})$ for all $j \in [\ell]$. Because $D(\mathcal{T})$ is convex (Lemma 34), this implies $C \subseteq D(\mathcal{T})$. To show $D(\mathcal{T}) \subseteq C$, let $p \in D(\mathcal{T})$; by definition, $0 \in \oplus_p\mathcal{T}$. Then in particular for each vector $v_1, \ldots, v_k$ guaranteed by the halfspace condition, we have

$$\begin{aligned}0 &\leq \max_{x \in \oplus_p \mathcal{T}} \langle v, x \rangle \\ &= \sum_{y \in \mathcal{Y}} p_y \max_{x \in T_y} \langle v_i, x \rangle \\ &\leq \sum_{y \in \mathcal{Y}} p_y B_{iy}.\end{aligned}$$

This proves $Bp \geq 0$, so $p \in C$. □

Lemma 33. *Given polytopes $\mathcal{T}$, there exists a finite set of normal vectors $w_1, \ldots, w_K \in \mathbb{R}^d$ such that, for all $p \in \Delta_\mathcal{Y}$, $\oplus_p \mathcal{T} = \{x : \langle w_i, x \rangle \leq \sum_{y \in \mathcal{Y}} p_y \max_{x \in T_y} \langle w_i, x \rangle\}$.*

Proof. For each p, $\oplus_p \mathcal{T}$ is a polytope. For each of the finitely many supports $(2^n - 1)$, we know $\oplus_p \mathcal{T}$ is a polytope, and every polytope can be defined by a finite, complete set of vectors for that polytope. As a two polytopes with the same support are combinatorially equivalent, they can be defined by the same facet enumeration, and any set of normals that is complete for $\oplus_p \mathcal{T}$ is also complete for a $\oplus_{p'} \mathcal{T}$ if $\text{supp}\,(p) = \text{supp}\,(p')$. We can simply concatenate these finite set of normals for the finite polytope supports, with some normals possibly becoming redundant. This yields finitely many normals defining the weighted Minkowski sum $\oplus_p \mathcal{T}$ for all $p \in \Delta_\mathcal{Y}$. □

Lemma 34. *For any $\mathcal{T}$, $D(\mathcal{T})$ is a polytope (in particular, is convex).*

Proof. Recall by definition, the notation $\oplus_p \mathcal{T} = \{\sum_y p_y x_y : x_y \in T_y (\forall y)\}$. Each T_y is a polytope, so $p_y T_y$ is a polytope. The Minkowski sum of polytopes is a polytope, so $\oplus_p \mathcal{T}$ is a polytope [86, Section 1.2]. Since $\oplus_p \mathcal{T}$ is a polytope for all $p \in \Delta_\mathcal{Y}$, we know there is a halfspace representation of normals V so that for all $y \in \mathcal{Y}$, we have $x \in p_y T_y \iff \langle V, x \rangle \leq p_y e^y$ for some matrix V and the support vector e^y, where $e^y_i = H_{T_y}(V_i)$. By Lemma 33, we know that there is a set of normals V^* that is complete for $T(p)$ for all $p \in \Delta_\mathcal{Y}$. We construct E^* as the support matrix for this complete set of normals. The support of the Minkowski sum for a given normal is the sum of the normals [86, Theorem 3.1.6], and so we we can take $x \in \oplus_p \mathcal{T} \iff \langle V^*, x \rangle \leq E^* p$. Substituting $x = \vec{0}$, we see

$\vec{0} \in \oplus_p \mathcal{T} \iff E^* p \geq \vec{0} \iff p \in D(\mathcal{T})$ by Lemma 44, which defines a polytope by construction of E^*. □

Lemma 35. *Let $C = \{p : Bp \geq \vec{0}\}$ where B has the minimum possible number of rows to capture C, and suppose $C = \{p : Ep \geq \vec{0}\}$. Then for each row in B there is some (unique) row in E that is equal to αB for some positive α.*

Proof. Ziegler [97, Exercise 2.15] alludes to this fact. Suppose there was a row j of B that did not appear (possibly scaled, because of the inequality on $\vec{0}$) in E. Then there is some $x \in \{x : Ex \geq 0\}$ so that $\langle B_i, x\rangle \geq 0$ for all $i \neq j$ and $\langle B_j, x\rangle < 0$ since B has the minimum number of rows required to capture C. This contradicts $x \in C = \{x : x : Bx \geq 0\}$. □

The two conditions in Theorem 14 give us a much better understanding of when a given set of polytopes $\mathcal{T}$ satisfies the optimality condition. We are still left with the problem, however, of understanding when such a set $\mathcal{T}$ exists. Intuitively, the biggest hurdle that remains is the quantification over sets of polytopes, a massive search space. Surprisingly, one can reduce this search to a quadratic feasibility program, which we now give. The key insight involves the halfspace condition, and observing that given a certain "complete" set of normal vectors, one can exactly describe the support function of $\oplus_p \mathcal{T}$ in terms of the support functions of each T_y and each normal vector v. From here, we use the fact that this description is linear in p, and can therefore relate it directly to the given matrix B.

Our program will consist of variables for the normal vectors $\{v_i \in \mathbb{R}^d : i \in [k]\}$ for the (relaxed) halfspace condition, as described above, and variables for vertices $\{x^j_y \in \mathbb{R}^d : j \in [l], y \in \mathcal{Y}\}$ which witness $0 \in \oplus_{p^j} \mathcal{T}$ for the vertex condition, where the vector x^j_y is the y^{th} column of X^j.

Definition 37 (Quadratic Feasibility Program)**.**

Given: *$d \in \mathbb{N}$, a polytope $C = \{p \in \Delta_{\mathcal{Y}} : Bp \geq \vec{0}\} = \mathrm{conv}(\{p^1, \ldots, p^l\}) \subseteq \Delta_{\mathcal{Y}}$, where $B \in \mathbb{R}^{k \times n}$ has a minimum number of rows.*

Variables: *$V \in \mathbb{R}^{k \times d}$ with rows $\{v_i\}$; $X^1, \ldots, X^l \in \mathbb{R}^{d \times n}$, where X^j has columns $\{x^j_y\}$.*

Constraints:

$$VX^j \leq B \qquad \textit{(pointwise, } \forall j \in [l]\textit{)} \tag{5.1}$$

$$\sum_{y=1}^{n} p_y^j x_y^j = \vec{0} \qquad \textit{(}\forall j \in [l]\textit{)} \tag{5.2}$$

Our main result of this section is that our quadratic program is feasible if and only if there exist some set of d-dimensional polytopes satisyfing the optimality condition in Theorem 13. As an immediate corollary, if some input $C = \gamma_r$ and d yields an infeasible program, then the embedding dimension of the loss ℓ is at least $d + 1$.

Theorem 15. *Given a convex polytope $C \subseteq \Delta_{\mathcal{Y}}$, there exist polytopes $\mathcal{T}$ in $\mathbb{R}^d$ such that $D(\mathcal{T}) = C$ if and only if the above quadratic program (Definition 37) is a feasible.*

Proof. By Theorem 14, it suffices to show that $\mathcal{T}$ satisfying the halfspace and vertex conditions exist if and only if the program is feasible.

($\implies$) By the vertex condition, for each $j \in [l]$, there exist witnesses $\{x_y^j \in T_y : y \in \mathcal{Y}\}$ satisfying the second constraint of the quadratic program (Inequality 5.2). By the halfspace condition, there exist normals $v_1, \ldots, v_k$ such that, for all i, for all $x \in T_y$, $\langle v_i, x \rangle \leq B_{iy}$; in particular, this applies to the above witnesses $x_y^j \in T_y$. Collecting $v_1, \ldots, v_k$ as the columns of V, this shows that the first constraint (Inequality 5.1) is satisfied.

($\impliedby$) We construct $T_y = \text{conv}(\{x_y^1, \ldots, x_y^l\})$. The second constraint of the quadratic program immediately implies the vertex condition. Taking $v_1, \ldots, v_k$ as the columns of V, the first constraint implies that for each x_y^j, we have $\langle v_i, x_y^j \rangle \leq B_{iy}$ for all i, j, y. Any point $x \in T_y$ is a convex combination of $x_y^1, \ldots, x_y^l$, so it satisfies $\langle v_i, x \rangle \leq B_{iy}$. This implies the halfspace condition. □

Corollary 14. *Given a discrete loss ℓ eliciting γ, if there is a report $r \in \mathcal{R}$ such that the quadratic program (Definition 37) is infeasible for input $C = \gamma_r$ and d, then the embedding dimension of ℓ is at least $d + 1$.*

The feasibility program can be viewed as a low-rank matrix problem, namely: does there exist a set of rank-d matrices that are pointwise dominated by B, sharing the left factor V, whose right

	y_1	y_2
r_1	5	3
r_2	4	1
r_3	2	1
r_4	1	4
r_5	1	6
r_6	3	8

Table 5.1: Ordered discrete loss matrix of ℓ which we embed.

factors X^j respectively satisfy a subspace constraint? We will see in § 5.5.3 that for the important example of abstain loss, the constraints simplify into a more pure low-rank matrix problem. In particular, for $d = n - 1$, a solution always exists via the construction in Theorem 4, which takes the convex conjugate of the negative Bayes risk of ℓ for each outcome and subtracting the report u, after which one can project down to $n - 1$ dimensions since $\dim(\text{affhull}(\Delta_{\mathcal{Y}})) = n - 1$.

5.5 Examples

5.5.1 Example construction of real-valued embedding

For concreteness, we now construct an embedding via the loss given in Theorem 11. We start with the ordered discrete loss given in Table 5.1

Given this loss, we can calculate $v(i)_y$ as in Theorem 11 for both losses, shown by the $\bullet$ in Figure 5.3.

Note here that we observe $K = 4$, $\Lambda = (1, 1/2, 1/2, 3/4, 3/4)$, and embedding points $\varphi(\mathcal{R}) = (0, 1, 3, 5, 19/3, 23/3)$. The polyhedral loss is then shown in Figure 5.4.

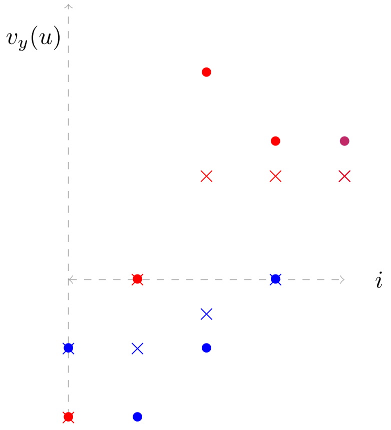


Figure 5.3: • represents the original v_i, where blue is used for $v(\cdot)_{y_1}$ and red for $v(\cdot)_{y_2}$. The × symbol of the same color is the Λ-corrected directional derivative to force monotonicity.

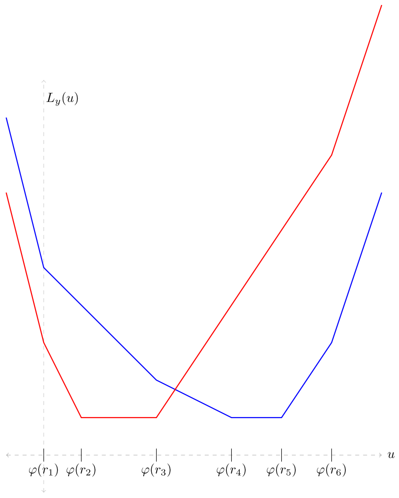


Figure 5.4: Our constructed embedding L for the discrete loss ℓ given in Table 5.1.

5.5.2 Abstain, $d = 1$

Consider the following abstain loss $\ell^f_{\text{abs}}\alpha : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$, where $\mathcal{R} = \mathcal{Y} \cup \{\bot\}$. For intuition, consider $\bot$ an "abstain" report.

$$\ell^f_{\text{abs}}\alpha(r, y) = \begin{cases} 0 & r = y \\ \alpha & r = \bot \\ 1 & r \notin \{y, \bot\} \end{cases} \tag{5.3}$$

For $\alpha \leq 1/2$, Ramaswamy et al. [72] give an elegant embedding of this loss on n outcomes into $d = \lceil \log_2(n) \rceil$ dimensions, where each $y \in \mathcal{Y}$ is embedded at a corner of the Boolean hypercube $\{-1, 1\}^d$ while $\bot$ is embedded at the origin.

One classification-like problem that is of particular interest is the abstain property, elicited by the loss $\ell^f_{\text{abs}}\alpha$ given in Equation (5.3). The property $\gamma = \text{abstain}_\alpha$ for $\alpha \in (0, 1)$ can be verified:

$$\text{abstain}_\alpha(p) = \begin{cases} \arg\max_{y \in \mathcal{Y}} p_y & \max_y p_y \geq 1 - \alpha \\ \bot & \text{otherwise} \end{cases} . \tag{5.4}$$

Ramaswamy et al. [72] study the abstain property in depth, presenting a $\lceil \log_2(n) \rceil$ dimensional embedding of the abstain property. However, it is unclear if this bound is tight, as the previously studied lower bounds of Ramaswamy and Agarwal [71] do not work well for this property, failing to give any lower bound tighter than the trivial dimension 1.

With our 1-dimensional characterization, we already observe a tighter lower bound.

Proposition 13. *For $n \geq 3$ and $\alpha < 1$, the abstain loss ℓ_α is not 1-embeddable.*

Proof. Consider the intersection graph of $\gamma := \text{abstain}_\alpha$, which is a spoke-like graph. In particular, the node associated with $\gamma_\bot$ has n edges, and since we assume $n \geq 3$, it cannot be a path. In fact, the intersection graph for this property is a star graph. For an example with $n = 3$ and $\alpha = 1/2$, see Figure 5.2. □

5.5.3 Abstain, $\alpha = 1/2$, $d = 2$

We now use our d-dimensional characterization and some observations about the abstain$_{1/2}$ property to improve lower bounds from those given by Ramaswamy and Agarwal [71]. We defer details to § 5.7.2.

Proposition 14. *The quadratic feasibility program (Definition 37) with input $C = \gamma_\perp = \{p \in \Delta_\mathcal{Y} : \max_y p_y \leq 1/2\}$, $n = 5$, and dimension parameter $d = 2$, is infeasible.*

Corollary 15. *The abstain loss with $\alpha = 1/2$ on $n \geq 5$ outcomes has embedding dimension at least 3.*

5.6 Chapter conclusion

Essentially the only other known lower-bound technique for dimensionality of calibrated surrogates is the *feasible subspace dimension* of Ramaswamy and Agarwal [71]. This crux of this technique is also an optimality condition on a surrogate loss, showing that if $\vec{0}$ is in the p-weighted Minkowski sum of the subgradient sets of L, then there is some local affine set of dimension $n-d-1$ such that $\vec{0}$ is also in the p'-weighted Minkowski sum for all p' in the set, and thus the set must be contained in γ_r. Therefore, for example, if the intersection of several level sets is a single vertex v (as in e.g. 0-1 loss for the uniform distribution), then the only such set can be of dimension 0, which gives a $d \geq n-1$ lower bound. Intuitively, the feasible subspace dimension bound uses *local* characteristics of the property around an examined distribution to give lower bounds, while the QFP in Definition 37 uses more global information.

Future Work. There are a few threads of future work: the first is to utilize monotonicity to see if we can construct even tighter lower bounds on embedding dimension. Second, we hope to understand when, if ever, embedding dimension is not equal to convex calibration dimension. Moreover, the restriction that we are calibrated over the entire simplex may be tighter than necessary in some contexts, and would be useful to understand the tradeoff between calibration and dimension of a surrogate loss. For one example where we can reduce surrogate dimension with a low-entropy assumption on the simplex, see Agarwal and Agarwal [6, Example 6].

5.7 Chapter appendix

5.7.1 1-dimensional characterization omitted proofs

We will make substantial use of the following general definition.

Definition 38. *A property $\Gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ is monotone if there are maps $a : \mathcal{R} \to \mathbb{R}^\mathcal{Y}$, $b : \mathcal{R} \to \mathbb{R}^\mathcal{Y}$ and a total ordering $<$ of $\mathcal{R}$ such that the following two conditions hold.*

(1) For all $r \in \mathcal{R}$, we have $\Gamma_r = \{p \in \Delta_\mathcal{Y} : \langle a(r), p\rangle \leq 0 \leq \langle b(r), p\rangle\}$.

(2) For all $r < r'$, we have $a(r) \leq b(r) \leq a(r') \leq b(r')$ (component-wise).

We have a property being orderable if and only if it is monotone since the maps a and b must define hyperplanes in the simplex in order for the ordering to be complete.

Theorems 11 and 12 follow from the following two results.

Theorem 16. *Let ℓ be a discrete loss eliciting a finite property γ. The following are equivalent: (1) γ is orderable; (2) the two conditions of Theorem 11 are satisfied; (3) ℓ is 1-embeddable; (4) γ is monotone. Moreover, when the conditions of Theorem 11 are satisfied, the loss L constructed does indeed embed ℓ.*

Proof. We will prove the chain of implications in order.

Orderable $\implies$ Conditions:

Let $\gamma : \Delta_\mathcal{Y} \to 2^\mathcal{R} \setminus \{\emptyset\}$ be finite and orderable. From Lambert [55, Theorem 4], we have positively-oriented normals $v_i \in \mathbb{R}^\mathcal{Y}$ for all $i \in \{1, \dots, k-1\}$ such that $\gamma_{r_i} \cap \gamma_{r_{i+1}} = \{p \in \Delta_\mathcal{Y} : \langle v_i, p\rangle = 0\}$, and moreover, for all $i \in \{2, \dots, k-1\}$, we have $\gamma_{r_i} = \{p \in \Delta_\mathcal{Y} : \langle v_{i-1}, p\rangle \leq 0 \leq \langle v_i, p\rangle\}$, while $\gamma_{r_1} = \{p \in \Delta_\mathcal{Y} : \langle v_1, p\rangle \leq 0\}$ and $\gamma_{r_k} = \{p \in \Delta_\mathcal{Y} : \langle v_{k-1}, p\rangle \leq 0\}$. From the positive orientation of the v_i, we have for all $p \in \Delta_\mathcal{Y}$ that $\mathrm{sgn}(\langle v_i, p\rangle)$ is monotone in i. In particular, it must be that for all y, $\text{sgn}((v_i)_y)$ is monotone in i, taking the distribution with all weight on outcome y, thus establishing the first condition.

For the second condition, suppose we had $\lambda^-(i) < \lambda^+(i)$. Then we would have $y, y' \in \mathcal{Y}$ such that $v(i)_y < 0$, $v(i+1)_y < 0$, $v(i)_{y'} > 0$, $v(i+1)_{y'} > 0$, and $0 < \frac{v(i)_y}{v(i+1)_y} < \frac{v(i)_{y'}}{v(i+1)_{y'}}$, which would in turn imply $|v(i)_y|/v(i)_{y'} < |v(i{+}1)_y|/v(i{+}1)_{y'}$. Letting $c = \frac{1}{2}\left(|v(i+1)_y|/v(i+1)_{y'} + |v(i)_y|/v(i)_{y'}\right)$ and taking p to be the distribution with weight $1/(1+c)$ on y and $c/(1+c)$ on y', we see that

$$\begin{aligned}
\langle v(i), p\rangle &= \frac{1}{1+c}\Big(v(i)_y + \tfrac{1}{2}(|v(i+1)_y|/v(i+1)_{y'} + |v(i)_y|/v(i)_{y'})v(i)_{y'}\Big) \\
&> \frac{1}{1+c}\left(v(i)_y + (|v(i)_y|/v(i)_{y'})v(i)_{y'}\right) = 0 \\
\langle v(i+1), p\rangle &= \frac{1}{1+c}\Big(v(i+1)_y + \tfrac{1}{2}(|v(i+1)_y|/v(i+1)_{y'} + |v(i)_y|/v(i)_{y'})v(i)_{y'}\Big) \\
&< \frac{1}{1+c}\left(v(i+1)_y + (|v(i+1)_y|/v(i+1)_{y'})v(i+1)_{y'}\right) = 0 \ ,
\end{aligned}$$

thus violating the observation that $\text{sgn}(\langle v(i), p\rangle)$ is monotone in i.

Conditions $\implies$ 1-embeddable: (correctness of construction)

First, observe that that $\lambda(i)$ satisfies $\lambda^+(i) \leq \lambda(i) \leq \lambda^-(i)$, and by the second condition, $\lambda(i) > 0$ even when either bound is infinite. Thus, $\Lambda_i > 0$ for all i, and so $\varphi(r_1) < \ldots < \varphi(r_k)$. By definition of L, we have $L(\varphi(r_1)) = \ell(r_1)$, and $L(\varphi(r_{i+1})) = \ell(r_i) + \Lambda_i \cdot (\varphi(r_{i+1}) - \varphi(r_i)) \cdot (\ell(r_{i+1}) - \ell(r_i))$ for all $i \geq 2$. Since $\varphi(r_{i+1}) - \varphi(r_i) = 1/\Lambda_i$ by our construction, we have $L(\varphi(r_{i+1})) = \ell(r_{i+1})$, so that $\ell(r) = L(\varphi(r))$ for all $r \in \mathcal{R}$. It remains therefore to show convexity of L and the optimality conditions.

For convexity, note that L is piecewise linear with the only possible nondifferentiable points being the embedding points $\varphi(r_1), \ldots, \varphi(r_k)$. Let us denote the left and right derivative operators for real-valued functions by ∂^- and ∂^+, respectively, and write $\partial^-\ell(u) = (\partial^-\ell(u)_y)_{y\in\mathcal{Y}} \in \mathbb{R}^{\mathcal{Y}}$, and similarly for $\partial^+\ell(u)$. To show convexity, then, we need only show $\partial^-\ell(\varphi(r_i)) \leq \partial^+\ell(\varphi(r_i))$ for all $i \in \{1, \ldots, k\}$, where the inequality holds coordinate-wise. By construction, we have $\partial^-\ell(\varphi(r_1)) = -K\mathbb{1}$ and $\partial^+\ell(\varphi(r_k)) = \Lambda_{k-1}K\mathbb{1}$, and for $i \in \{1, \ldots, k-1\}$ we have $\partial^+\ell(\varphi(r_i)) = \partial^-\ell(\varphi(r_{i+1})) = \Lambda_i v(i+1)$. By definition of K, we have $\partial^-\ell(\varphi(r_1)) = -K\mathbb{1} \leq v(2) = \partial^+\ell(\varphi(r_1))$ and $\partial^-\ell(\varphi(r_k)) = \Lambda_{k-1}v(k) \leq \Lambda_{k-1}K\mathbb{1} = \partial^+\ell(\varphi(r_k))$.

It remains to show that for all $i \in \{2, \ldots, k-1\}$ and all $y \in \mathcal{Y}$, we have $\Lambda_{i-1} v(i)_y \leq \Lambda_i v(i+1)_y$, which by definition of Λ is equivalent to $v(i)_y \leq \lambda(i) v(i+1)_y$. By our first condition, the possible pairs $(\mathrm{sgn}(v(i)_y), \mathrm{sgn}(v(i+1)_y))$ are $(-,-), (-,0), (-,+), (0,0), (0,+), (+,+)$, and given that $\lambda(i) > 0$, all are trivial except $(-,-)$ and $(+,+)$. In the $(-,-)$ case, we have by definition of $\lambda^-(i)$ that $\lambda(i) \leq \lambda^-(i) \leq v(i)_y / v(i+1)_y$. Recalling that both $v(i)_y$ and $v(i+1)_y$ are negative, we conclude $v(i)_y \leq \lambda(i) v(i+1)_y$. In the $(+,+)$ case, we have $\lambda(i) \geq \lambda^+(i) \geq v(i)_y / v(i+1)_y$, and again $v(i)_y \leq \lambda(i) v(i+1)_y$.

For optimality, consider any $r \in \mathcal{R}$ and any $p \in \Gamma_{\varphi(r)}$. By the matching of loss values, for every $r' \in \mathcal{R}$ we have $\langle p, \ell(r) \rangle = \langle p, L(\varphi(r)) \rangle \leq \langle p, L(\varphi(r')) \rangle = \langle p, \ell(r') \rangle$, which implies $p \in \gamma_r$. For the other direction, consider a distribution $p \in \Delta(\mathcal{Y})$, and the subgradient of $\langle p, L(\varphi(r_i)) \rangle$ for some $i \in \{2, \ldots, k-1\}$. We have

$$\begin{aligned}
0 \in \partial \langle p, L(\varphi(r_i)) \rangle &\iff \partial^- \langle p, \ell(\varphi(r_i)) \rangle \leq 0 \leq \partial^+ \langle p, \ell(\varphi(r_i)) \rangle \\
&\iff \langle p, \partial^- \ell(\varphi(r_i)) \rangle \leq 0 \leq \langle p, \partial^+ \ell(\varphi(r_i)) \rangle \\
&\iff \langle p, \Lambda_{i-1} v(i) \rangle \leq 0 \leq \langle p, \Lambda_i v(i+1) \rangle \\
&\iff \langle p, v(i) \rangle \leq 0 \leq \langle p, v(i+1) \rangle \\
&\iff \langle p, \ell(r_i) - \ell(r_{i-1}) \rangle \leq 0 \leq \langle p, \ell(r_{i+1}) - \ell(r_i) \rangle \\
&\iff \langle p, \ell(r_i) \rangle \leq \langle p, \ell(r_{i-1}) \rangle \text{ and } \langle p, \ell(r_i) \rangle \leq \langle p, \ell(r_{i+1}) \rangle \ .
\end{aligned}$$

For $i = 1$, similar reasoning gives that optimality is equivalent to the condition $\langle p, \ell(r_1) \rangle \leq \langle p, \ell(r_2) \rangle$, and for $i = k$, $\langle p, \ell(r_k) \rangle \leq \langle p, \ell(r_{k-1}) \rangle$. (Note that the other conditions, $-K \leq 0$ or $0 \leq \Lambda_{k-1} K$, are true regardless of p.) In particular, if $p \in \gamma_{r_i}$, then we have $\langle p, \ell(r_i) \rangle \leq \langle p, \ell(r_{i-1}) \rangle$ for $i \geq 2$, and $\langle p, \ell(r_i) \rangle \leq \langle p, \ell(r_{i+1}) \rangle$ for $i \leq k-1$, so for all i we have $0 \in \partial \langle p, L(\varphi(r_i)) \rangle$ and thus $p \in \Gamma_{\varphi(r_i)}$.

Embedding $\implies$ Monotone:

We trivially satsify the conditions of Definition 38 by taking $a(r_i) = \partial^- L(\varphi(r))$ and $b(r_i) = \partial^+ L(\varphi(r))$.

Monotone $\implies$ Orderable:

Let $\gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}} \setminus \{\emptyset\}$ be finite and monotone. Then we can use the total ordering of $\mathcal{R}$ to write $\mathcal{R} = \{r_1, \ldots, r_k\}$ such that $r_i < r_{i+1}$ for all $i \in \{1, \ldots, k-1\}$. We now have $\gamma_{r_i} \cap \gamma_{r_{i+1}} = \{p \in \Delta_{\mathcal{Y}} : \langle a(r_{i+1}), p\rangle \leq 0 \leq \langle b(r_i), p\rangle\}$. If this intersection is empty, then there must be some p with $\langle b(r_i), p\rangle < 0$ and $\langle a(r_{i+1}), p\rangle > 0$; by monotonicity, no earlier or later reports can be in $\gamma(p)$, so we see that $\gamma(p) = \emptyset$, a contradiction. Thus the intersection is nonempty, and as we also know $b(r_i) \leq a(r_{i+1})$ we conclude $b(r_i) = a(r_{i+1})$, and the intersection is the hyperplane defined by $b(r_i) = a(r_{i+1})$. □

Throughout the rest of this section, we let $\text{prop}_{\mathcal{P}}[L]$ be the unique property elicited by the loss L.

Lemma 36. *For any convex $L : \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}_+$, the property $\text{prop}_{\mathcal{P}}[L]$ is monotone.*

Proof. If L is convex and elicits Γ, let a, b be defined by $a(r)_y = \partial_- L(r)_y$ and $b(r) = \partial_+ L(r)_y$, that is, the left and right derivatives of $L(\cdot)_y$ at r, respectively. Then $\partial L(r)_y = [a(r)_y, b(r)_y]$. We now have $r \in \text{prop}_{\mathcal{P}}[L](p) \iff 0 \in \partial\langle p, L(r)\rangle \iff \langle a(r), p\rangle \leq 0 \leq \langle b(r), p\rangle$, showing the first condition. The second condition follows as the subgradients of L are monotone functions (see e.g. Rockafellar [77, Theorem 24.1]). □

Definition 39. *A cell complex in $\mathbb{R}^d$ is a set C of faces (of dimension $0, \ldots, d$) which (i) union to $\mathbb{R}^d$, (ii) have pairwise disjoint relative interiors, and (iii) any nonempty intersection of faces F, F' in C is a face of F and F' and an element of C.*

Definition 40. *Given sites $s_1, \ldots, s_k \in \mathbb{R}^d$ and weights $w_1, \ldots, w_k \geq 0$, the corresponding power diagram is the cell complex given by*

$$\text{cell}(s_i) = \{x \in \mathbb{R}^d : \forall j \in \{1, \ldots, k\} \; \|x - s_i\|^2 - w_i \leq \|x - s_j\| - w_j\} \; . \tag{5.5}$$

Theorem 17 ([8])**.** *A cell complex is affinely equivalent to a convex polyhedron if and only if it is a power diagram.*

Lemma 37. *Let γ be a finite (non-redundant) property elicited by a loss L. Then the negative Bayes risk G of L is polyhedral, and the level sets of γ are the projections of the facets of the epigraph of*

G onto $\Delta_\mathcal{Y}$, and thus form a power diagram. In particular, the level sets γ are full-dimensional in $\Delta_\mathcal{Y}$ (i.e., of dimension $n-1$).

Lemma 38. *Let $\gamma : \Delta_\mathcal{Y} \to 2^\mathcal{R} \setminus \{\emptyset\}$ be a finite elicitable property, and suppose there is a calibrated link ψ from an elicitable Γ to γ. For each $r \in \mathcal{R}$, define $P_r = \bigcup_{u \in \psi^{-1}(r)} \Gamma_u \subseteq \Delta_\mathcal{Y}$, and let $\overline{P}_r$ denote the closure of the convex hull of P_r. Then $\gamma_r = \overline{P}_r$ for all $r \in \mathcal{R}$.*

Proof. As $P_r \subseteq \gamma_r$ by the definition of calibration, and γ_r is closed and convex, we must have $\overline{P}_r \subseteq \gamma_r$. Furthermore, again by calibration of ψ, we must have $\bigcup_{r \in \mathcal{R}} P_r = \bigcup_{u \in \mathbb{R}} \Gamma_u = \Delta_\mathcal{Y}$, and thus $\bigcup_{r \in \mathcal{R}} \overline{P}_r = \Delta_\mathcal{Y}$ as well. Suppose for a contradiction that $\gamma_r \neq \overline{P}_r$ for some $r \in \mathcal{R}$. From Lemma 37, γ_r has nonempty interior, so we must have some $p \in \mathring{\gamma}_r \setminus \overline{P}_r$. But as $\bigcup_{r' \in \mathcal{R}} \overline{P}_{r'} = \Delta_\mathcal{Y}$, we then have some $r' \neq r$ with $p \in \overline{P}_{r'} \subseteq \gamma_{r'}$. By Theorem 17, the level sets of γ form a power diagram, and in particular a cell complex, so we have contradicted point (ii) of Definition 39: the relative interiors of the faces must not be disjoint. Hence, for all $r \in \mathcal{R}$ we have $\gamma_r = \overline{P}_r$. □

The preceding statements yield the following proposition.

Proposition 15. *If convex $L : \mathbb{R} \to \mathbb{R}^\mathcal{Y}$ indirectly elicits a finite elicitable property γ, then γ is orderable.*

Proof. Let $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$. From Lemma 36, $\Gamma := \mathrm{prop}_\mathcal{P}[L]$ is monotone. Let $\psi : \mathbb{R} \to \mathcal{R}$ be the calibrated link from Γ to γ. From Lemma 38, we have $\overline{P}_r = \gamma_r$ for all $r \in \mathcal{R}$, where $\overline{P}_r$ is the closure of the convex hull of $\bigcup_{u \in \psi^{-1}(r)} \Gamma_u$.

As Γ is monotone, we must have $a, b : \mathcal{R} \to \mathbb{R}^\mathcal{Y}$ such that $\overline{P}_r = \{p \in \Delta_\mathcal{Y} : \langle a(r), p \rangle \leq 0 \leq \langle b(r), p \rangle\}$. (Take $a(r)_y = \inf_{u \in \psi^{-1}(r)} a(u)_y$ and $b(r)_y = \sup_{u \in \psi^{-1}(r)} b(u)_y$.) Now taking $p_r \in \mathring{\gamma}_r$ and picking $u_r \in \Gamma(p_r)$, we order $\mathcal{R} = \{r_1, \ldots, r_k\}$ so that $u_{r_i} < u_{r_{i+1}}$ for all $i \in \{1, \ldots, k-1\}$. (The u_{r_i} must all be distinct, as we chose p_r so that $\gamma(p_r) = \{r\}$, so $\psi(u_{r_i}) = r_i$ for all i.)

Let $i \in \{1, \ldots, k-1\}$. By monotonicity of Γ, we must have $a(r_i) \leq b(r_i) \leq a(r_{i+1}) \leq b(r_{i+1})$. As $\bigcup_{r \in \mathcal{R}} \overline{P}_r = \bigcup_{r \in \mathcal{R}} \gamma_r = \Delta_\mathcal{Y}$, we must therefore have $b(r_i) = a(r_{i+1})$. Finally, we conclude $\gamma_{r_i} \cap \gamma_{r_{i+1}} = \{p \in \Delta_\mathcal{Y} : \langle b(r_i), p \rangle = 0\}$. As these statements hold for all $i \in \{1, \ldots, k-1\}$, γ is orderable. □

5.7.2 Omitted example

Geometric intuition for QFP on abstain$_{1/2}$**,** $d = 2$ In order to prove Proposition 14, we take some simplifying steps to the quadratic feasibility program for this specific problem. The strategy is to consider the level set $\gamma_\perp$, the set of distributions with modal mass at most $1/2$. We show that the quadratic feasibility program with this input cannot be satisfied with dimension 2 for $n = 5$.

Lemma 39. *For the abstain loss $\ell_{1/2}$, the level set of abstain satisfies $\gamma_\perp = \text{conv}\{(\delta_y + \delta_{y'})/2 : y, y' \in \mathcal{Y}, y < y'\} = \{p : Bp \geq \vec{0}\}$ where δ_y puts probability one on y and $B = \mathbb{1}\mathbb{1}^\intercal - 2I \in \mathbb{R}^{5\times5}$, i.e. has entries -1 on the diagonal and 1 everywhere else.*

Proof. Recall that $\gamma_\perp$ is the set of distributions p with $\max_y p_y = 1/2$. First, note that each distribution of the form $(1/2, 1/2, 0, 0, 0)$ and so on is in $\gamma_\perp$. Meanwhile, every such p can be written as a convex combination of these corners. Second, note that if $p \in \gamma_\perp$, then $p_y \leq 1/2$ for all $y \in \mathcal{Y}$. These constraints can be rewritten as $\langle p, b\rangle \geq 0$ where $b_y = -1$ and $b_{y'} = 1$ for all $y' \neq y$, literally requiring $p_y \leq \sum_{y' \neq y} p_{y'}$. □

Observation 1. *For any invertible $A \in \mathbb{R}^{d\times d}$, if $V, \{X^j : j \in [\ell]\}$ is a feasible solution to the quadratic feasibility program, then so is $(VA), \{A^{-1}X^j : j \in [\ell]\}$.*

Proof. The halfspace constraints are $(VA)(A^{-1}X^j) \leq B \iff VX^j \leq B$. The j^{th} vertex constraint is a vector equality $\sum_{y\in\mathcal{Y}} p^j_y (A^{-1}X^j)_y = \vec{0}$. If we let a_m be the m^{th} row of A^{-1}, then the m^{th} row of the vector equality is

$$\begin{aligned} 0 &= \sum_{y\in\mathcal{Y}} p^j_y \langle a_m, x^j_y\rangle \\ &= \langle a_m, \sum_{y\in\mathcal{Y}} p^j_y x^j_y\rangle \\ &= 0 \end{aligned}$$

so the program is feasible. □

Corollary 16. *If there is a feasible solution to the quadratic feasibility program, then there is a feasible solution where v_1 is the first standard basis vector and $\|v_i\| \leq \|v_1\| = 1$ for all i.*

Proof. In particular, we can take a series of matrices A in Observation 1 that permute the rows of V, scale[2] all rows by $\frac{1}{\|v_1\|}$, and linearly map v_1 to $(1, 0, \ldots, 0)$. □

Notation for the quadratic program. Recall that in the quadratic program, each vertex p in the convex-hull representation corresponds to a matrix variable X. Here, the vertices are indexed by a pair of distributions, so for each $i < j$, we refer to that vertex of $\gamma_\perp$ by $p^{ij} = (\delta_i + \delta_j)/2$, with corresponding variable X^{ij}. The yth column of this matrix is denoted $x_y^{ij} \in \mathbb{R}^d$.

Lemma 40. *In any feasible solution to the QFP for $\gamma_\perp$ and $\ell_{1/2}$, we have $x_i^{ij} = -x_j^{ij}$ for all $i < j$ in $\mathcal{Y}$.*

Proof. Directly from the vertex constraints: $p^{ij} = \frac{1}{2}\delta_i + \frac{1}{2}\delta_j$, so the ij constraint reduces to $\frac{1}{2}x_i^{ij} + \frac{1}{2}x_j^{ij} = \vec{0}$. □

Lemma 41. *There is no feasible solution to the QFP for $\gamma_\perp$ and $\ell_{1/2}$ where $v_i = cv_j$ for $c > 0$ and any $i \neq j$.*

Proof. There is an open halfspace through the origin strictly containing both the feasible regions $F_i = \{x : \langle v_i, x\rangle \leq -1, \langle v_j, x\rangle \leq 1 \; \forall j \neq i\}$ and F_j, so there is no set of witnesses such that $x_i^{ij} \in F_i$ and $x_j^{ij} \in F_j$, as this would contradict Lemma 40. □

Lemma 42. *For $d = 2$ and the level set $\gamma_\perp$ for $\ell_{1/2}$, any pair of linearly independent v_i, v_j rule out all except for a unique feasible value for x_i^{ij} and also for x_j^{ij}.*

Proof. From the halfspace constraints, we must have $\langle v_i, x_i^{ij}\rangle \leq -1$ and $\langle v_i, x_j^{ij}\rangle \leq 1$, which combines with Lemma 40 to give $\langle v_i, x_i^{ij}\rangle = 1$. This immediately also gives $\langle v_j, x_i^{ij}\rangle = -1$. This system of two inequalities in two dimensions has exactly one solution if v_i, v_j are linearly independent. □

[2] Note one can show $V = 0$ is not feasible unless B is a trivial property, i.e. essentially has no rows at all.

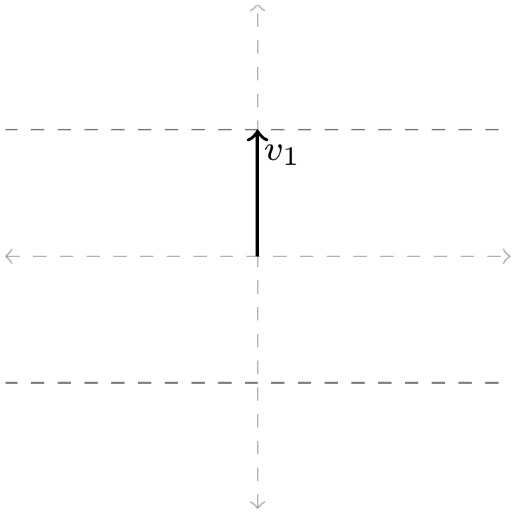


Figure 5.5: Example of v_1.

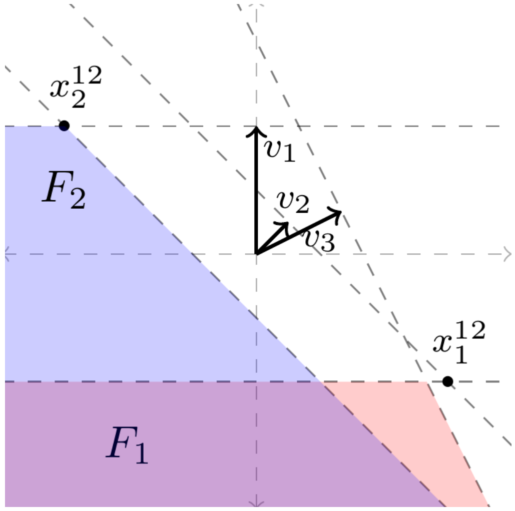


Figure 5.6: If $x_1^{12} \neq x_1^{13}$, we have a contradiction.

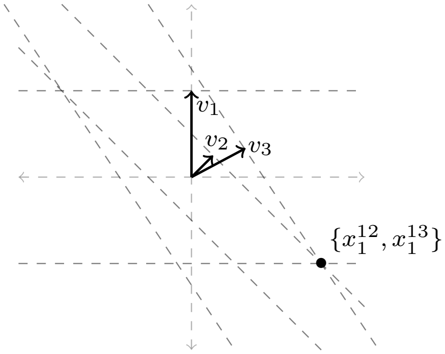


Figure 5.7: Reducing to the case where $x_1^{12} = x_1^{13}$.

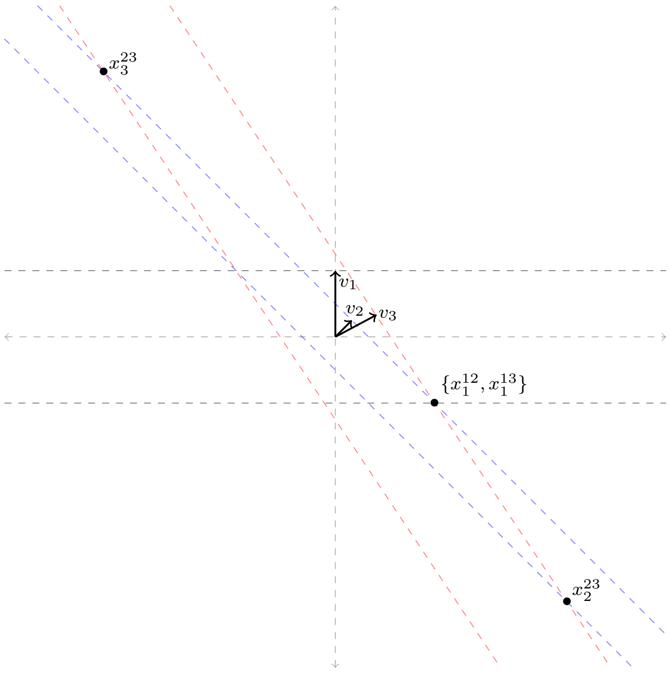


Figure 5.8: The intersection of $\langle v_2, x \rangle = -1$ and $\langle v_3, x \rangle = 1$ occurs outside the region $\langle v_1, x \rangle \in [-1, 1]$.

Lemma 43. *There is no feasible solution to the QFP for $\gamma_\perp$ and $\ell_{1/2}$ where three vectors v_i, v_j, v_m lie strictly within a halfspace through the origin (i.e. all within $180°$ of each other).*

Proof. Let three of the vectors be given, lying strictly inside a halfspace, and label them clockwise as v_1, v_2, v_3. WLOG suppose v_1 points vertically "up", as in Figure 5.5. By Lemma 42, the possible locations of the following points are all uniquely determined: x_i^{ij}, x_j^{ij} for $(i,j) \in \{(1,2),(1,3),(2,3)\}$. Both points x_1^{12} and x_1^{13} lie on the line $\langle v_1, x\rangle = -1$, i.e. a horizontal line below the origin. We have constraints $\langle v_2, x_1^{12}\rangle = 1$ and $\langle v_2, x_1^{13}\rangle \leq 1$. This implies x_1^{13} is left of x_1^{12} on the horizontal line $\langle v_1, x\rangle = 1$. But the symmetric constraints $\langle v_3, x_1^{13}\rangle = 1$ and $\langle v_3, x_1^{12}\rangle \leq 1$ imply symmetrically that x_1^{12} is left of x_1^{13} on the line. This implies we must have $x_1^{12} = x_1^{13}$. An example of this contradiction is shown in Figure 5.6.

If we consider the four lines $\langle v_2, x\rangle = 1, \langle v_2, x\rangle = -1, \langle v_3, x\rangle = 1, \langle v_3, x\rangle = -1$, we therefore have three points of intersection with the line $\langle v_1, x\rangle = 1$ and three with the line $\langle v_1, x\rangle = -1$, as shown in Figure 5.7. WLOG, these points from top left to top right are: x_2^{12} (which equals x_3^{13}), the intersection with $\langle v_2, x\rangle = 1$, and the intersection with $\langle v_3, x\rangle = 1$; and therefore from bottom left to bottom right are: the intersection with $\langle v_3, x\rangle = -1$, the intersection with $\langle v_2, x\rangle = -1$, and the intersection with x_1^{12} (which equals x_1^{13}).

This implies that the lines $\langle v_2, x\rangle = -1$ and $\langle v_3, x\rangle = 1$, in particular, do not intersect anywhere within the bounds of $\langle v_1, x\rangle \in [-1, 1]$. Therefore, either their intersection point x_2^{23} or its negative x_3^{23} violates the feasibility constraint $\langle v_1, x\rangle \leq 1$, as in Figure 5.8. This proves there is no feasible solution with three normals lying strictly in the same halfspace through the origin. □

Proposition 16. *The abstain loss $\ell_{1/2}$ with $n = 5$ is not 2-embeddable.*

Proof. Let any 5 vectors be given, numbered clockwise. v_1, v_2, v_3 cannot lie in a cone of strictly less than 180°, as this would contradict Lemma 43. So the clockwise angle between v_1 and v_3 is at least 180°. Since there are no duplicate angles (Lemma 41), this implies that the clockwise angle between v_4 and v_1, which includes v_5, is strictly less than 180°. This contradicts Lemma 43. □

5.8 Polytope notes

Here, we recall some additional standard definitions from the theory of convex polytopes.

Definition 41 (Supporting function)**.** *Let S be a nonempty bounded set in $\mathbb{R}^d$. We call the supporting function of S the function $H_S : \mathbb{R}^d \to \mathbb{R}$ by*

$$H_S(a) := \sup_{x \in S} \langle a, x \rangle \ .$$

Definition 42 (Minkowski sum)**.** *Let $S_1, S_2, \ldots, S_n$ be sets of vectors. We can define their Minkowski sum as the set of vectors which can be written as the sum of a vector in each set. Namely,*

$$S_1 \oplus \ldots \oplus S_n = \{x_1 + \ldots + x_n : x_i \in S_i \ \forall i\}$$

Theorem 18 ([86, Theorem 3.1.2])**.** *Let $T_1, \ldots, T_n$ be polytopes in $\mathbb{R}^d$ and let F be a face of the Minkowski sum $T := T_1 \oplus \ldots \oplus T_n$. Then there are faces $F_1, \ldots, F_n$ of $T_1, \ldots, T_n$ respectively such that $F = F_1 \oplus \ldots \oplus F_n$. Moreover, this decomposition is unique.*

Theorem 19 ([86, Theorem 3.1.6])**.** *The supporting function of a Minkowski sum is the sum of the supporting functions of its summands.*

Weibel [86] notes that:

> It is easy to see that the normal fan (undefined here, but consequently normal cones) of $p_i T_i$ does not change as long as p_i is positive. Since the normal fan of a Minkowski sum can be deduced from that of its summands, we can deduce from this that the combinatorial properties of $\oplus_p T_y$ stay the same as long as all p_i are positive.

Suppose we are given a polytope $T_y \in \mathbb{R}^d$ and set of vectors $V \in \mathbb{R}^{k \times d}$. Call $e^y \in \mathbb{R}^k$ the vector such that $e^y_i = \max_{x \in T_y} \langle v_i, x \rangle$. For a finite set $\mathcal{T} = \{T_1, , \ldots, T_n\}$, let us denote the *support matrix* $E = (e^y)^n_{y=1}$.

Definition 43. *We say a set of normals V is complete with respect to a polytope T_y if $T_y = \{x \in \mathbb{R}^d : Vx \leq e^y\}$.*

Moreover, we say V is complete with respect to the set of polytopes $\mathcal{T}$ if and only if V is complete with respect to each $T_y \in \mathcal{T}$.

We will suppose we start with a finite set of n polytopes $\mathcal{T} := \{T_1, \dots, T_n\}$, and we will call $T := T_1 \oplus \ldots \oplus T_n \in \mathbb{R}^d$ their Minkowski sum. We know that every polytope has both a halfspace and vertex representation ($\mathcal{H}$-representation and $\mathcal{V}$-representation, respectively.) By existence of the $\mathcal{H}$-representation, we know there must be a matrix $V \in \mathbb{R}^{k \times d}$ and vector $e \in \mathbb{R}^k$ such that $T = \{x \in \mathbb{R}^d : Vx \leq e\}$. In fact, with a complete set of normals V, we know that e can be the support vector of each of the normals. However, finding V is not always easy, so we assume that we are given V for now.

Now, for a given polytope $\oplus_p \mathcal{T}$, we want to ask when a given $z \in \mathbb{R}^d$ is in the polytope $\oplus_p \mathcal{T}$. We will later generalize to finding the set of $p \in \Delta_{\mathcal{Y}}$ for which $\vec{0} \in \oplus_p \mathcal{T}$ by substituting $z = \vec{0}$. Throughout, assume we have V which is complete for $\mathcal{T}$ and consider E defined by the support of each normal in V for all $T_y \in \mathcal{T}$. We denote $e^y = E_{;y}$ as the y^{th} column of E, or equivalently, the support vector for T_y given V.

Since we define $T_y = \{x : Vx \leq e^y\}$, we can multiply the right side of the inequality by the constant $p_y \geq 0$ to yield $p_y T_y = \{x : Vx \leq p_y e^y\}$. Taking the Minkowski sum of polytopes described by the same set of normals, we can take

$$\begin{aligned} \oplus_p \mathcal{T} &= \{x : Vx \leq p_1 E_{;1}\} \oplus \ldots \oplus \{x : Vx \leq p_n E_{;n}\} \\ &= \{x : Vx \leq p_1 E_{;1} + \ldots + p_n E_{;n}\} \\ &= \{x : Vx \leq Ep\} \ . \end{aligned}$$

The first to second line follows from Theorem 19 and preservation of inequalities under addition. Now, we have $z \in T(p) \iff \langle v_i, z \rangle \leq (Ep)_i$ for all $v_i \in V$.

Observe that this construction yields $\vec{0} \in \oplus_p \mathcal{T}$ if and only if $Ep \geq 0$ by substitution.

We assume $p \in \Delta_{\mathcal{Y}}$, so we now describe the cell $D(\mathcal{T}) := \{p \in \Delta_{\mathcal{Y}} : Ep \geq \vec{0}\}$ as the set of distributions such that $\vec{0} \in \oplus_p \mathcal{T}$. We will see in Lemma 44 that this definition is equivalent to the definition of $D(\mathcal{T})$ in Definition 36.

Given the complete set of normals V and constructing the support matrix for V and $\mathcal{T}$, E, we observe that E is unique up to rescaling. However, as discussed earlier, there are always multiple complete sets of normals for $\mathcal{T}$, and so in that sense, E is not unique.

We want to know the following: starting from $\mathcal{T}$, can we derive the cell $C \subseteq \Delta_{\mathcal{Y}}$ where $\vec{0} \in T(p)$ for all $p \in C$? We know that if we are given $\mathcal{T}$ and a complete set of normals V, we can describe $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \vec{0}\}$.

Lemma 44. *Suppose we are given polytopes $\mathcal{T} = \{T_1, \ldots, T_n\}$ and a set of normals V that is complete for $\mathcal{T}$. Take $E = (e_i^y)$ where $e_i^y = \max_{x \in T_y} \langle v_i, x \rangle$, and $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \vec{0}\}$.*

Then $\{p \in \Delta_{\mathcal{Y}} : \vec{0} \in \oplus_p \mathcal{T}\} = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \vec{0}\}$.

Proof. First, let us fix a distribution $p \in \Delta_{\mathcal{Y}}$. By Theorem 19, we have the support of the (weighted) Minkowski sum is the (weighted) sum of the support of each polytope, which we can re-write the weighted support as the product Ep.

Each halfspace is bounded by the support function of the weighted polytope by construction of E, so the support of the weighted polytope defined by an inequality on v_i can be described as $\langle v_i, z \rangle \leq \langle E_i, p \rangle$. Taking this for all v_i, we then have $\oplus_p \mathcal{T} = \{x \in \mathbb{R}^d : Vx \leq Ep\}$.

Therefore, for fixed p, we have $\vec{0} \in \oplus_p \mathcal{T} \iff Ep \geq \vec{0}$. As $p \in \Delta_{\mathcal{Y}}$ was arbitrary, we observe the stated set equality. □

The following result allows us to consider the sets of distributions for which $\vec{0}$ is in the Minkowski sum in terms of the minimal rank matrix describing the cell.

Proposition 17. *Suppose we are given polytopes $\mathcal{T} = \{T_1, \ldots, T_n\}$ and a set of normals V that is complete for $\mathcal{T}$. Take $E = (e_{iy})$ where $e_{iy} = \max_{x \in T_y} \langle v_i, x \rangle$, and take $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \vec{0}\}$ and take the minimal rank $B \in \mathbb{R}^{k \times n}$ such that we have the given cell $C = \{p \in \Delta_{\mathcal{Y}} : Bp \geq \vec{0}\}$.*

Then $\{p \in \Delta_{\mathcal{Y}} : \vec{0} \in \oplus_p \mathcal{T}\} = C$ if and only if $C = D(\mathcal{T})$.

Proof. By Lemma 44, we have $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : \vec{0} \in \oplus_p \mathcal{T}\}$, and the result follows. □

Definition 44. *We say a vector v is redundant with respect to matrix Y if we have $\{z : Yz \geq \vec{b}\} = \{z : [Y; v]z \geq \vec{b^*}\}$, where $b^* = [b; c]$ for some constant $c \in \mathbb{R}$.*

Proposition 18. *Suppose we have polytopes $\mathcal{T} = \{T_1, \ldots, T_n\}$ and a set of normals V that is complete for $\mathcal{T}$. Take $E = (e_i^y)$ where $e_i^y = \max_{x \in T_y} \langle v_i, x \rangle$, and take $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \vec{0}\}$ and take the minimal matrix B such that a given cell $C = \{p \in \Delta_{\mathcal{Y}} : Bp \geq \vec{0}\}$.*

Then $\{p \in \Delta_{\mathcal{Y}} : \vec{0} \in \oplus_p \mathcal{T}\} = C$ if and only the rows of B appear in E (possibly scaled) and every other row of E is redundant with respect to B.

Proof. ($\implies$) First, assume $C = \{p \in \Delta_{\mathcal{Y}} : \vec{0} \in \oplus_y p_y T_y\}$. By Proposition 17, we know that $C = D^{\mathcal{T}} := \{p \in \Delta_{\mathcal{Y}} : Ep \geq \vec{0}\}$. Then we have $\{p \in \Delta_{\mathcal{Y}} : Bp \geq \vec{0}\} = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \vec{0}\}$. As B is minimal, we must have that every row of B appears (possibly scaled) in E. Otherwise, we would contradict equality of the polytopes C and $D(\mathcal{T})$. Moreover, all rows in E not in B are redundant with respect to B by equality of the polytopes.

($\impliedby$) Suppose that all rows of B appear in E, and every other row of E is redundant with respect to B. Then we have $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \vec{0}\} = \{p \in \Delta_{\mathcal{Y}} : Bp \geq \vec{0}\} = C$.

Then $D(\mathcal{T}) = C$, and by Proposition 17, we have $C = \{p \in \Delta_{\mathcal{Y}} : \vec{0} \in \oplus_p \mathcal{T}\}$. □

Chapter 6

Lower bounding convex consistency dimension

6.1 Introduction

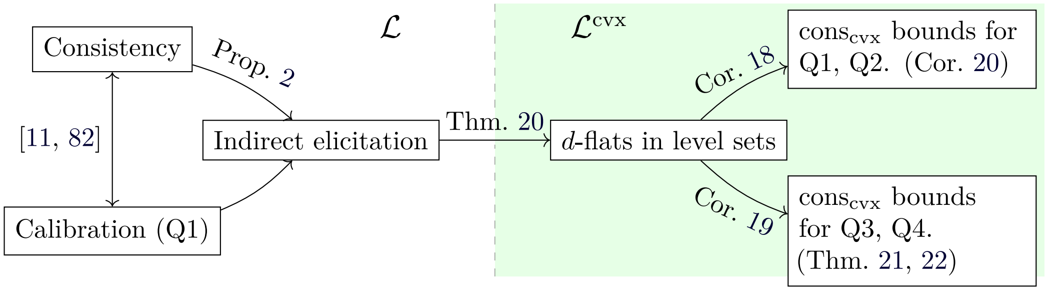
In § 2.3, we saw that indirect property elicitation is necessary for consistency. Moreover, in § 2.4, we saw a few metrics that measure prediction dimension, one of which is the embedding dimension introduced in Chapter 5. This chapter introduces a tool based on indirect property elicitation called d-flats that presents lower bounds on convex elicitation complexity, and in turn, on convex consistency dimension (Definition 11). These results can be applied in any of the settings introduced in Table 2.1, and we demonstrate this through applications to target problems such as abstain loss, variance, and conditional value at risk.

This chapter is based heavily on the work of Finocchiaro et al. [30], published at NeurIPS 2021.

6.2 Convex consistency dimension and elicitation complexity

Various works have studied the minimum prediction dimension d needed in order to construct a consistent surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$, typically through proxies such as calibration [6, 71, 82] and property elicitation [35, 38, 40]. Motivated by the importance of convex surrogates in machine learning, Ramaswamy and Agarwal [71] introduce the following definition.

Definition 11 (Convex consistency dimension)**.** *Given target loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ or property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$, its convex consistency dimension $\mathrm{cons}_{\mathrm{cvx}}(\cdot)$ is the minimum dimension d such that $\exists L \in \mathcal{L}_d^{\mathrm{cvx}}$ and link ψ such that (L, ψ) is consistent with respect to ℓ or γ.*



$\mathcal{L}$
$\mathcal{L}^{\text{cvx}}$
Consistency
Prop. 2
[11, 82]
Indirect elicitation
Thm. 20
d-flats in level sets
Calibration (Q1)
Cor. 18
Cor. 19
cons_{cvx} bounds for Q1, Q2. (Cor. 20)
cons_{cvx} bounds for Q3, Q4. (Thm. 21, 22)

Figure 6.1: Flow and implications of this chapter. Compared to calibration, we suggest indirect elicitation as a simpler but almost-as-powerful necessary condition for consistency. In particular, we obtain a testable condition (Theorem 20), based on d-flats, for the existence of a d-dimensional consistent convex surrogate. This condition recovers and strengthens existing calibration-based results for Q1, while simultaneously applying to other quadrants. We illustrate the breadth and power of d-flats by resolving two open questions for Q3 and Q4 in § 6.5.

In the case of a target property γ, Lambert et al. [57] similarly introduce the notion of *elicitation complexity*. Later generalized by Frongillo and Kash [40], elicitation complexity is the lowest prediction dimension of an elicitable property, from some class of properties, from which one can compute γ. We give here the definition for convex-elicitable properties.

Definition 12 (Convex elicitation complexity)**.** *Given a target property γ, the convex elicitation complexity* $\text{elic}_{\text{cvx}}(\gamma)$ *is the minimum dimension d such that there is a $L \in \mathcal{L}_d^{\text{cvx}}$ indirectly eliciting γ.*

As consistency implies indirect elicitation (Proposition 2), we have the following.

Corollary 17. *Given a property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ or loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ eliciting γ, we have* $\text{elic}_{\text{cvx}}(\gamma) \leq \text{cons}_{\text{cvx}}(\gamma) = \text{cons}_{\text{cvx}}(\ell)$.

The *embedding dimension* of Chapter 5 is a lower bound on both convex elicitation complexity of discrete properties and convex consistency dimension of discrete losses and finite statistics.

6.3 Lower bounding convex consistency dimension via d-flats

We now turn to the question of bounding the convex consistency dimension for a given task. From Proposition 2, given a target property γ or loss ℓ with $\gamma = \text{prop}_{\mathcal{P}}[\ell]$, this task reduces to lower bounding the convex consistency dimension of γ. Theorem 20, crystallized from the proofs of Ramaswamy and Agarwal [71, Theorem 16] and Agarwal and Agarwal [6, Theorem 9], considers a particular distribution p and surrogate prediction $u \in \mathbb{R}^d$ which is optimal for p. Theorem 20 will show that if d is small, then the level set $\{p \in \mathcal{P} : u \in \arg\min_{u'} \mathbb{E}_p L(u', Y)\}$ must be large; in fact, it must roughly contain a high-dimensional *flat* (of codimension d). By definition of indirect elicitation, there is some level set γ_r (where u is linked to r) containing this flat as well. We can then leverage the contrapositive of this result: if γ has a level set intricate enough not to contain any high-dimensional flats, then γ cannot have a low-dimensional consistent convex surrogate.

Definition 45 (d-flat)**.** *For $d \in \mathbb{N}$, a d-flat, or simply flat, is a nonempty set $F = \ker_{\mathcal{P}} W := \{q \in \mathcal{P} : \mathbb{E}_q W = \vec{0}\}$ for some measurable $W : \mathcal{Y} \to \mathbb{R}^d$.*

The following lemma yields consistency bounds when combined with Proposition 2. A similar result is found in Agarwal and Agarwal [6, Theorem 9], which bounds the dimension of level sets of a single-valued $\mathrm{prop}_{\mathcal{P}}[L]$. Theorem 20 instead bounds the dimension of flats contained in the level sets, an additional power which we leverage in our examples.

Theorem 20. *Let $\Gamma : \mathcal{P} \rightrightarrows \mathbb{R}^d$ be (directly) elicited by $L \in \mathcal{L}_d^{\mathrm{cvx}}$ for some $d \in \mathbb{N}$. Let $\mathcal{Y}$ be either a finite set, or $\mathcal{Y} = \mathbb{R}$, in which case we assume each $p \in \mathcal{P}$ admits a Lebesgue density supported on the same set for all $p \in \mathcal{P}$.*[1] *For all $u \in \mathrm{range}\,\Gamma$ and $p \in \Gamma_u$, there is some d-flat F such that $p \in F \subseteq \Gamma_u$.*

Proof. As L is convex and elicits Γ, we have $u \in \Gamma(p) \iff \vec{0} \in \partial \mathbb{E}_p L(u, Y)$. We proceed in two cases, depending on $|\mathcal{Y}|$.

Finite $\mathcal{Y}$: If $\mathcal{Y}$ is finite, this is additionally equivalent to $\vec{0} \in \oplus_y p_y \partial L(u, y)$, where $\oplus$ denotes the Minkowski sum [50, Theorem 4.1.1].[2] Expanding, we have $\oplus_y p_y \partial L(u, y) = \{\sum_{y \in \mathcal{Y}} p_y x_y \mid x_y \in \partial L(u, y)\ \forall y \in \mathcal{Y}\}$, and thus $Wp = \sum_y p_y x_y = \vec{0}$ where $W = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$; cf. [71, $\mathbf{A}^m$ in

[1] This assumption is largely for technical convenience, to ensure that $\mathcal{V}_{u,p}$ does not depend on p. Any such assumption would suffice, and we suspect even that condition can be relaxed.

[2] ∂ represents the subdifferential $\partial f(x) = \{z : f(x') - f(x) \geq \langle z, x' - x \rangle\ \forall x'\}$.

Theorem 16]. Let $V_{u,p} : \mathcal{Y} \to \mathbb{R}^d, y \mapsto W_y$ be the function encoding the columns of W. Observe that $\mathbb{E}_p V_{u,p} = \vec{0}$.

$\mathcal{Y} = \mathbb{R}$: Any $L \in \mathcal{L}_d^{\text{cvx}}$ satisfies the assumptions of [52], so we may interchange subdifferentiation and expectation. Specifically, letting $\mathcal{V}_{u,p} = \{V : \mathcal{Y} \to \mathbb{R}^d \mid V \text{ measurable}, V(y) \in \partial L(u, y) \text{ } p\text{-a.s.}\}$, we have $\partial \mathbb{E}_p L(u, Y) = \{\int V(y)dp(y) \mid V \in \mathcal{V}_{u,p}\}$. As $\vec{0} \in \partial \mathbb{E}_p L(u, Y)$, in particular, there is some $V_{u,p} \in \mathcal{V}_{u,p}$ such that $\mathbb{E}_p V_{u,p} = 0$. For any $q \in \mathcal{P}$, as by assumption q is supported on the same set as p, we have $V_{u,p}(y) \in \partial L(u, y)$ q-a.s., so that $V_{u,p} \in \mathcal{V}_{u,q}$. Thus, $\mathbb{E}_q V_{u,p} = 0$ implies $0 \in \partial \mathbb{E}_q L(u, Y)$ by the above.

In both cases, we take the flat $F := \ker_{\mathcal{P}} V_{u,p}$, and have $p \in F$ by construction. To see $F \subseteq \Gamma_u$, from the chain of equivalences above, we have for any $q \in \mathcal{P}$ that $q \in \ker_{\mathcal{P}} V_{u,p} \implies \vec{0} \in \partial \mathbb{E}_q L(u, Y) \implies u \in \Gamma(q) \implies q \in \Gamma_u$. □

Theorem 20 now allows us to derive bounds on convex consistency dimension by considering distributions and property values that are either single-valued (Corollary 18) or on the relative interior of the simplex with finite $\mathcal{Y}$ (Corollary 19). In order to apply Theorem 20 to various properties, we need the following lemmas about separating hyperplanes.

First, a hyperplane weakly separates two sets if its two closed halfspaces respectively contain the two sets.

Lemma 45 ([30, Lemma 2])**.** *If $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ is an elicitable property, then for any pair of predictions $r, r' \in \mathcal{R}$ where $\gamma_r \neq \gamma_{r'}$, there is a hyperplane $H = \{x \in \mathbb{R}^{\mathcal{Y}} : v \cdot x = 0\}$, for some $v \in \mathbb{R}^{\mathcal{Y}}$, that weakly separates γ_r and $\gamma_{r'}$ and has $\gamma_r \cap H = \gamma_{r'} \cap H = \gamma_r \cap \gamma_{r'}$.*

Proof. Let ℓ elicit γ. Let $v = \ell(r, \cdot) - \ell(r', \cdot)$, interpreted as a nonzero vector in $\mathbb{R}^{\mathcal{Y}}$. Let $H = \{q : v \cdot q = 0\}$. If $v \cdot q < 0$, then r' cannot be optimal, so $q \notin \gamma_{r'}$. So $\gamma_{r'} \subseteq \{q : v \cdot q \geq 0\}$. Symmetrically, $\gamma_r \subseteq \{q : v \cdot q \leq 0\}$. This is weak separation, and it immediately implies that $\gamma_r \cap \gamma_{r'} \subseteq H$. Finally, if and only if $v \cdot q = 0$, i.e. $q \in H$, by definition the expected losses of both reports are the same. So $q \in \gamma_r \cap H \iff q \in \gamma_{r'} \cap H$. This gives $\gamma_r \cap H = \gamma_{r'} \cap H = \gamma_r \cap \gamma_{r'} \cap H = \gamma_r \cap \gamma_{r'}$. □

Lemma 46. *Suppose we are given an elicitable property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$, where $\mathcal{Y}$ is finite, and distribution $p \in \mathrm{relint}(\mathcal{P})$ such that $p \in \gamma_r \cap \gamma_{r'}$ for $r, r' \in \mathcal{R}$. Then for any flat F containing p, $F \subseteq \gamma_r \iff F \subseteq \gamma_{r'}$.*

Proof. If $\gamma_r = \gamma_{r'}$, we are done. Otherwise, Lemma 45 gives a hyperplane $H = \{x \in \mathbb{R}^{\mathcal{Y}} : v \cdot x = 0\}$ and a guarantee that $\gamma_r \subseteq \{q \in \Delta_{\mathcal{Y}} : v \cdot q \leq 0\}$, while $\gamma_{r'} \subseteq \{q \in \Delta_{\mathcal{Y}} : v \cdot q \geq 0\}$, and finally $\gamma_r \cap \gamma_{r'} \subseteq H$.

Suppose $F \subseteq \gamma_r$; we wish to show $F \subseteq \gamma_{r'}$. Let $q \in F$. By Lemma 47(i), we have $p \in \mathrm{relint} F$, so there exists $\epsilon > 0$ so that $q' = p - \epsilon(q - p) \in F$.

Now, suppose for contradiction that $q \notin \gamma_{r'}$. Then $v \cdot q < 0$: containment in γ_r gives $v \cdot q \leq 0$, and if $v \cdot q = 0$ then $q \in \gamma_r \cap H \implies q \in \gamma_{r'}$, a contradiction. But, noting that $p \in H$, we have $v \cdot q' = -\epsilon(v \cdot q) > 0$, so q' is not in γ_r. This contradicts the assumption $F \subseteq \gamma_r$. Therefore, we must have $q \in \gamma_{r'}$, so we have shown $F \subseteq \gamma_{r'}$. Because r and r' were completely symmetric, this completes the proof. □

Now we can understand the application of Theorem 20.

Corollary 18. *Let target property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ and $d \in \mathbb{N}$ be given. Let $\mathcal{Y}$ be either a finite set, or $\mathcal{Y} = \mathbb{R}$, in which case we assume each $p \in \mathcal{P}$ admits a Lebesgue density supported on the same set for all $p \in \mathcal{P}$. Let $p \in \mathcal{P}$ with $|\gamma(p)| = 1$, and take $\gamma(p) = \{r\}$. If there is no d-flat F with $p \in F \subseteq \gamma_r$, then $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$.*

Proof. Let (L, ψ) indirectly elicit γ, where $L \in \mathcal{L}_d^{\mathrm{cvx}}$, and let $\Gamma = \mathrm{prop}_{\mathcal{P}}[L]$. As Γ is non-empty, there is some $u \in \Gamma(p)$. Since γ is single-valued at p, we have $r = \psi(u)$; by Theorem 20, we know there is a d-flat $F = \ker_{\mathcal{P}} V_{u,p}$ so that $p \in F \subseteq \Gamma_u$. By definition of indirect elicitation, we additionally have $\Gamma_u \subseteq \gamma_r$. Thus, we have $p \in F \subseteq \gamma_r$. If no flat F satisfies the above conditions, then no $L \in \mathcal{L}_d^{\mathrm{cvx}}$ indirectly elicits γ, so $\mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$, and recall $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma)$ by Corollary 17. □

Corollary 19. *Let an elicitable target property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ be given, where $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ is defined over a finite set of outcomes $\mathcal{Y}$, and let $d \in \mathbb{N}$. Let $p \in \mathrm{relint}\mathcal{P}$. If there is no d-flat F with $p \in F \subseteq \gamma_r$,* *then* $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$.

Proof. Let (L, ψ) indirectly elicit γ and the convex function L and elicit Γ. As Γ is non-empty, there is some $u \in \Gamma(p)$, and suppose $r' = \psi(u)$. Take $F \subseteq \Gamma_u$ to be the flat that exists by Theorem 20. If $r = r'$, then $p \in F \subseteq \Gamma_u \subseteq \gamma_r$ by indirect elicitation. Otherwise, by Lemma 46, for elicitable properties with $p \in \gamma_r \cap \gamma_{r'}$, we observe $p \in F \subseteq \gamma_r \iff p \in F \subseteq \gamma_{r'}$.

As above, if no flat F satisfies the above conditions, then no $L \in \mathcal{L}_d^{\mathrm{cvx}}$ indirectly elicits γ, so $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$, recalling Corollary 17 for the first inequality. □

6.3.1 Illustrating the condition in all four quadrants

We now illustrate how to apply Theorem 20 to construct lower bounds on convex consistency dimension for targets across all four quadrants of Table 2.1. Throughout the examples, we will have $|\mathcal{Y}| = 3$ so that the probability simplex can be visualized in two dimensions (Figure 6.3). For each, we take $d = 1$, and thus ask whether any 1-flat (a line in the figures) passes through the point p while staying within the corresponding level set.

Q1: Classification with an abstain option. The abstain target loss is a well-studied variation of 0-1 loss that allows for an "abstain" report that gives a lesser punishment 1/2 for abstaining, $r = \bot$ [19, 20, 64, 71, 72]. Formally, the target loss is $\ell^{1/2}(r, y) := \mathbf{1}\{r \not\in \{y, \bot\}\} + (1/2)\mathbf{1}\{r = \bot\}$. Since we are given a discrete target loss, this problem fits nicely into Quadrant 1.

To apply Theorem 20, we first consider the abstain property γ elicited by $\ell^{1/2}$, where one predicts the most likely outcome y if $Pr[Y = y] \geq 1/2$ and otherwise "abstains" by predicting $\bot$. For the depicted distribution $p \in \mathrm{relint}\gamma_\bot$, we cannot fit a 1-flat (line) fully contained in $\gamma_\bot$ that passes through p. By Corollary 19, we can conclude $\mathrm{cons}_{\mathrm{cvx}}(\gamma^{1/2}) \geq 2$ when $|\mathcal{Y}| = 3$, meaning there is no consistent convex surrogate in 1 dimension. This lower bound matches the upper bound from the convex surrogate of Ramaswamy and Agarwal [71].

Q2: Variation of hierarchical classification. Ramaswamy et al. [70] study hierarchical classification tasks, in which labels are arranged in a tree and one wishes to predict the deepest node in a tree that is "likely enough" [17, 91]. Consider the variation of this task where one can

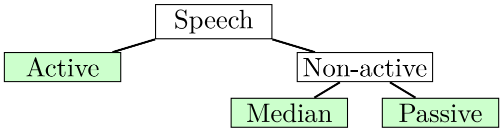


Figure 6.2: Hierarchical prediction example with labeling tone of speech. We take $\mathcal{Y} = \mathcal{R}$ to be the leaves of this tree, shown in blue.

only predict leaves of this tree. For example, Figure 6.2 depicts a speech classification task where speech is either active or non-active, and non-active is further subdivided into median and passive. It is natural to predict active if that label is more likely than both non-active labels combined, and otherwise to predict the most likely of median and passive:

$$\gamma(p) = \begin{cases} \text{active} & p_{\text{active}} \geq 1/2 \\ \text{median} & p_{\text{active}} \leq 1/2 \wedge p_{\text{median}} \geq p_{\text{passive}} \\ \text{passive} & p_{\text{active}} \leq 1/2 \wedge p_{\text{passive}} \geq p_{\text{median}} \end{cases} .$$

This "T-shaped" property, depicted in Figure 6.3 (Q2), falls under Quadrant 2, as it is not elicited by any target loss.[3] Like abstain, we cannot fit a 1-flat (line) entirely contained in the level set γ_{passive} through the depicted p, so Corollary 19 gives $\text{cons}_{\text{cvx}}(\gamma) = 2$.

Q3: Least-squares regression Squared loss is commonly used in machine learning and statistics for continuous estimation, making it the canonical choice for Quadrant 3. Squared loss is a 1-dimensional convex loss which elicits the mean $\Gamma(p) = \mathbb{E}_p[Y]$. Theorem 20 therefore states that we can fit a 1-flat through any distribution p while staying within the corresponding level set. In fact, the level sets of the mean are all exactly 1-flats, as demonstrated in Figure 6.3 (Q3).

Q4: Variance Consider the task of estimating the variance $\text{Var}(p) = \mathbb{E}_p[Y^2] - \mathbb{E}_p[Y]^2$. The variance is not (directly) elicitable as its level sets are not convex [57, 67], meaning this task falls under Quadrant 4. Interestingly, the fact that the variance is not elicitable does not yield a lower bound on elicitation complexity of 2, as it does not rule out the variance being a link of a real-valued convex-elicitable property; cf. Frongillo and Kash [40, Remark 1]. In § 6.5.1, we show $\text{elic}_{\text{cvx}}(\text{Var}) = 2$, meaning the lowest dimension of a convex loss to estimate conditional variance is 2. This lower bound will follow from Theorem 21 in § 6.5 using the fact that variance is the Bayes

[3] The cells of finite elicitable properties form power diagrams, a generalization of Voronoi diagrams, which disallow this "T-shaped" configuration [37, 56].

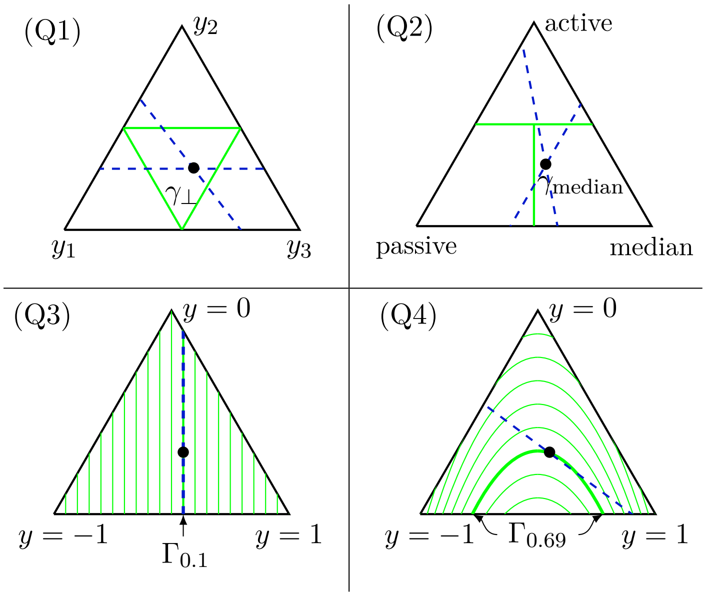


Figure 6.3: Example properties for each quadrant. Throughout, we take ● to be the distribution $p = (0.3, 0.3, 0.4)$ according to the left, top, and right outcomes respectively. (Q1,Q2) We cannot fit a 1-flat (line) through p without leaving the level sets $\gamma_\perp$ and γ_{median}, respectively; Theorem 20 implies that there is no 1-dimensional consistent convex surrogate for either problem. (Q3) Squared error is a 1-dimensional convex loss, and indeed it elicits the mean of Y, whose level sets are all 1-flats. (Q4) The level sets of the variance are curved and cannot fit a 1-flat; from Theorem 20 there is no 1-dimensional convex surrogate consistent for the variance.

risk of squared loss. While perhaps intuitively obvious, even this simple result is novel.

6.3.2 Relation to feasible subspace dimension

In Quadrant 1, Ramaswamy and Agarwal [71] give a lower bound on convex consistency dimension roughly by the co-dimension of the *subspace of feasible directions* $\mathcal{S}_\mathcal{C}(p)$ of a convex set $\mathcal{C}$ at a given distribution p such that $p \in \mathcal{C}$, which is loosely the "most full" subspace of $\mathcal{C}$ containing a neighborhood around p.

$$\mathcal{S}_\mathcal{C}(p) = \{v \in \mathbb{R}^n \mid \exists \epsilon_0 > 0 \text{ such that } p + \epsilon v \in \mathcal{C} \forall \epsilon \in (-\epsilon_0, \epsilon_0)\}$$

Theorem 20 subsumes the bounds given by Ramaswamy and Agarwal [71] by showing that, if there is a d-flat through p fully contained in a level set γ_r (so we can apply Theorem 20) then the subspace of feasible directions at the same $p \in \mathcal{C} := \gamma_r$ has co-dimension at most d, discussed in detail in § 6.4.1.

Proposition 19. *Suppose we are given a discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ eliciting the property $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$. Fix $p \in \text{relint}\Delta_\mathcal{Y}$ and take $r \in \mathcal{R}$ such that $p \in \gamma_r$. If $\text{cons}_{\text{cvx}}(\ell) = d$, then there exists a d-flat $F \subseteq \gamma_r$ through p. Moreover, F is a subspace of feasible directions over* *the set γ_r intersected with the simplex. Therefore,* $\operatorname{codim}(\mathcal{S}_{\gamma_r}(p)) \leq d$*, and in turn, this implies* $\operatorname{ccdim}(\ell) \geq d \geq \operatorname{codim}(\mathcal{S}_{\gamma_r}(p))$.

In other words, any d-flat through p is a subspace of feasible directions of co-dimension at most d, so Theorem 20 provides a weakly tighter lower bound on convex consistency dimension than Ramaswamy and Agarwal [71, Theorem 16]. In fact, the d-flats bound can be strictly tighter; in § 6.4 we show that the abstain example from Figure 6.3 (Q1) yields a d-flats lower bound of 2 and a feasible subspace dimension lower bound of 1. This gap stems from the fact that feasible subspace dimension uses only local information of the property to construct lower bounds, while d-flats in Theorem 20 allow us to additionally use global information. See Figure 6.4 in § 6.4 for an illustration.

6.4 Applications in Q1: Previous Lower Bounds and Comparisons

The main known technique for lower bounds on surrogate dimensions is given by Ramaswamy and Agarwal [71] for the Quadrant 1 (target loss and discrete predictions). The proof heavily builds around the "limits of sequences" in the definition of calibration. By restricting slightly to the broad class of minimizable losses $\mathcal{L}^{\text{cvx}}$, we show their bound follows relatively directly from Corollary 19. (We conjecture that the minimizability restriction to $\mathcal{L}^{\text{cvx}}$ can be lifted; see § 6.6.) Ramaswamy and Agarwal [71] construct what they call the subspace of feasible dimensions and give bounds in terms of its dimension.

Definition 46 (Subspace of feasible directions)**.** *The subspace of feasible directions $\mathcal{S}_\mathcal{C}(p)$ of a convex set $\mathcal{C} \subseteq \mathbb{R}^n$ at $p \in \mathcal{C}$ is $\mathcal{S}_\mathcal{C}(p) = \{v \in \mathbb{R}^n : \exists \epsilon_0 > 0$ such that $p + \epsilon v \in \mathcal{C}\ \forall \epsilon \in (-\epsilon_0, \epsilon_0)\}$.*

Ramaswamy and Agarwal [71] gives a lower bound on the dimensionality of all consistent convex surrogates, i.e. $\operatorname{cons}_{\text{cvx}}(\ell) \geq \|p\|_0 - \dim(\mathcal{S}_{\gamma_r}(p)) - 1$ for all p and $r \in \gamma(p)$, particularly in the setting where one is given a discrete prediction problem and target loss over finite outcomes. It turns out that the subspace of feasible directions is essentially a special case of a flat described by Theorem 20. So, by making a slight restriction to the class of minimizable convex surrogates

$\mathcal{L}^{\text{cvx}}$, we can derive this lower bound from our general technique in a way that we find shorter and simpler.

Corollary 20 ([71] Theorem 18)**.** *Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ be a discrete loss eliciting $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ with $\mathcal{Y}$ finite. Then for all $p \in \Delta_\mathcal{Y}$ and $r \in \gamma(p)$,*

$$\text{cons}_{\text{cvx}}(\gamma) \geq \|p\|_0 - \dim(\mathcal{S}_{\gamma_r}(p)) - 1 \ . \tag{6.1}$$

Sketch. If $\text{cons}_{\text{cvx}}(\gamma) \leq d$, then there is a $L \in \mathcal{L}^{\text{cvx}}_d$ so that L is consistent with respect to γ, and in turn, indirectly elicits γ. Theorem 20 says that there is some d-flat $F = \ker_\mathcal{P} V$ such that $p \in F \subseteq \gamma_r$. In particular, if $p \in \text{relint}\Delta_\mathcal{Y}$, we can see $\dim(F) = \dim(\mathcal{S}_{\gamma_r}(p))$. Since $\text{affhull}(\Delta_\mathcal{Y})$ has dimension $|\mathcal{Y}| - 1 = \|p\|_0 - 1$, by rank-nullity and $\text{rank}\,(V) \leq d$ (more precisely, the corresponding linear map $q \mapsto \mathbb{E}_q V$) we have $d \geq \|p\|_0 - 1 - \dim(\mathcal{S}_{\gamma_r}(p))$.

When $p \notin \text{relint}\Delta_\mathcal{Y}$, we can project down to the subsimplex on the support of p, again of dimension $\|p\|_0 - 1$, and modify L and ℓ accordingly. Now p is in the relative interior of this subsimplex, so the above gives $\text{cons}_{\text{cvx}}(\gamma) \geq \|p\|_0 - 1 - \dim(\mathcal{S}_{\gamma_r}(p))$, where now $\mathcal{S}$ is relative to $\mathbb{R}^{\text{supp}\,(p)}$. Finally, the feasible subspace dimension in the projected space is the same as in the original space because of p's location on a face of $\Delta_\mathcal{Y}$. □

There are some cases where the bound provided by Corollaries 18 and 19 is strictly tighter than the bound provided by feasible subspace dimension in Corollary 20. For an example of how Corollary 18 applies to a discrete property for which there is no target loss – a non-elicitable property, i.e. Quadrant 2, which is not considered by Ramaswamy et al. [72] – we refer the reader to Figure 6.3.

Example: Abstain Recall the abstain target loss $\ell^{abs}(r, y) := \mathbf{1}\{r \notin \{y, \bot\}\} + (1/2)\mathbf{1}\{r = \bot\}$, we can consider the *abstain property* it elicits, where one predicts the most likely outcome y if $Pr[Y = y|x] \geq 1/2$ and "abstain" by predicting $\bot$ otherwise. Ramaswamy and Agarwal [71] present a convex surrogate for the abstain loss that takes as input a prediction whose dimension is logarithmic in the number of outcomes, yielding new upper bounds on $\text{cons}_{\text{cvx}}(\ell^{abs})$ which are an exponential improvement over previous results, e.g., [21].

To lower bound the dimension of convex surrogates, we can consider two different distributions; in the first, our bound yields a strict gap over the feasible subspace dimension bound, and in the second, the bounds are equal. First, we choose $p = \bullet$ to be the uniform distribution (see Figure 6.4). In this case, the bound by feasible subspace dimension yields $\text{cons}_{\text{cvx}}(\ell^{abs}) \geq 3 - 2 - 1 = 0$, as the feasible subspace dimension is 2 since we are on the relative interior of the level set and simplex, as shown in Figure 6.4 (L).

However, consider any 1-flat containing $\bullet$. When intersected with the simplex, one can see that any line (a 1-flat, since $\bullet \in \text{relint}\Delta_{\mathcal{Y}}$) in the simplex through $\bullet$ also leaves the cell $\gamma_{\perp}$, which contains p. See Figure 6.4 (R) for intuition; a 1-flat through $p \in \text{relint}\Delta_{\mathcal{Y}}$ would be a line in such a figure. Therefore, we have no 1-flat containing p staying in $\gamma_{\perp}$, so we obtain a better lower bound, $\text{cons}_{\text{cvx}}(\ell^{abs}) \geq 2$. Combining this with the upper bounds given by [72], we observe the bound $\text{cons}_{\text{cvx}}(\ell^{abs}) = 2$ is tight in this case with $|\mathcal{Y}| = 3$.

Our bounds sometimes match those of [71]; consider the distribution $\star = (1/4, 1/4, 1/2)$, shown in Figure 6.4. The feasible subspace dimension of both $\gamma_{\perp}$ and γ_3 at $\star$ is 1, since one only moves toward the distributions $(0, 1/2, 1/2)$ and $(1/2, 0, 1/2)$ without leaving the level sets, and the three points are collinear in $\text{affhull}(\Delta_{\mathcal{Y}})$, suggesting $\mathcal{S}_{\gamma_{\perp}}(q) = 1$. This yields $\text{cons}_{\text{cvx}}(\ell^{abs}) \geq 3 - 1 - 1 = 1$. The same line segment defines a flat contained in both $\gamma_{\perp}$ and γ_3, so we have $\text{cons}_{\text{cvx}}(\ell^{abs}) \geq 1$ by Corollary 19, matching the feasible subspace dimension bound.

Bounds using d-flats appear to work well at distributions where previous bounds via feasible subspace dimension would have been vacuous. In essence, flats allow us a "global" view of the property we are eliciting, while the feasible subspace method only permits a "local" look at the property, so we find our method works better for distributions in $\text{relint}\Delta_{\mathcal{Y}}$.

6.4.1 Reconstructing Ramaswamy and Agarwal [71, Thm. 16]

Lemma 47. *Let the d-flat $F \subseteq \mathcal{P}$ (defined over finite $\mathcal{Y}$) contain some $p \in \text{relint}(\mathcal{P})$. Then*

(i) $p \in \text{relint}(F)$*;*

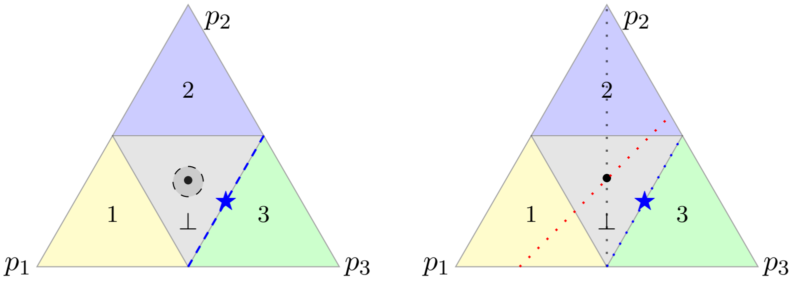


Figure 6.4: (Left) Feasible subspace dimension $\mathcal{S}_{\gamma_\perp}(\bullet) = 2$ and $\mathcal{S}_{\gamma_\perp}(\star) = 1$, giving the bound $\text{cons}_{\text{cvx}}(\ell^{abs}) \geq 3 - 1 - 1 = 1$. (Right) No 1-flat through $\bullet$ (a line since $\bullet \in \text{relint}\Delta_{\mathcal{Y}}$) stays fully contained in $\gamma_\perp$, so $\text{cons}_{\text{cvx}}(\ell^{abs}) \geq 2$.

(ii) $\dim(\mathcal{S}_F(p)) \geq \dim(\text{affhull}(\mathcal{P})) - d$.

Proof. As F is a d-flat, we have some $W : \mathcal{Y} \to \mathbb{R}^d$ such that $F = \ker_{\mathcal{P}} W$. Throughout, given a point (typically a distribution) p and convex set P, we define $P_p := P - \{p\}$. Define $T_W : \text{span}\,(\mathcal{P}_p) \to \mathbb{R}^d, v \mapsto \mathbb{E}_v W$.

(i) Since $p \in \text{relint}(\mathcal{P})$, for all $q \in \mathcal{P}$, there is some small enough $\epsilon > 0$ so that for all $\alpha \in (-\epsilon, \epsilon)$, the point $q_\alpha := p - \alpha(q - p)$ is still in $\mathcal{P}$. In particular, for $q \in F$, we claim $q_\alpha \in F$. As $p, q \in F$, we have $\mathbb{E}_p W = \mathbb{E}_q W = \vec{0}$. By linearity of expectation, we then have $\mathbb{E}_{q_\alpha} W = \vec{0}$. This implies $q_\alpha \in F$, and therefore $p \in \text{relint}F$.

(ii) We first show $\text{span}\,(F_p) = \mathcal{S}_F(p)$. First, take $v \in \mathcal{S}_F(p)$, and take ϵ_0 as in the definition. For $\epsilon = \epsilon_0/2$, we then have $p + \epsilon v \in F \implies \epsilon v \in F_p$, and therefore, $v \in \text{span}\,(F_p)$. Now take $v \in \text{span}\,(F_p)$. Since $p \in \text{relint}F$ (i), we have $\vec{0} \in \text{relint}F_p$. Therefore there is an $\epsilon_0 > 0$ so that $\epsilon v \in F_p$ for all $\epsilon \in (-\epsilon_0, \epsilon_0)$ by convexity of F. Therefore, $v \in \mathcal{S}_F(p)$, and we observe $\mathcal{S}_F(p) = \text{span}\,(F_p)$.

We now show $\mathcal{S}_F(p) = \ker(T_W)$. Observe that $\mathcal{S}_F(p) \subseteq \ker(T_W)$ follows trivially from the definitions of the two functions. Now let $v \in \ker(T_W)$, and $v' \in F_p$. This means $\mathbb{E}_v W = \vec{0}$, so it suffices to show $v = cv' \in F_p$, thus showing $v \in \mathcal{S}_F(p)$. Since $p \in \text{relint}\mathcal{P}$, we must have $\vec{0} \in \text{relint}F_p$, so we know there is some small enough $\epsilon > 0$ so that $-\alpha v' \in F_p$ for $\alpha \in (-\epsilon, \epsilon)$. Take $c = -\alpha$, and we conclude $v \in \mathcal{S}_F(p)$. Therefore, $\ker(T_W) = \mathcal{S}_F(p)$.

We finally want to show $\dim(\text{affhull}(\mathcal{P})) = \dim(\text{span}\,(\mathcal{P}_p))$. Consider that any $q \in \text{span}\,(\mathcal{P}_p)$ can be written as a scalar multiple of an element of $\mathcal{P}_p$, which can be written as a convex combination of elements of the minimal basis $\mathcal{P}_p$. In particular, since $\vec{0} \in \mathcal{P}_p$, it can be written as an affine combination of elements of the basis, so $\dim(\text{affhull}(\mathcal{P})) \geq \dim(\text{span}\,(\mathcal{P}_p))$. We also have $\text{affhull}(\mathcal{P}) - \{p\} \subseteq \text{span}\,(\mathcal{P}_p)$, so $\dim(\text{affhull}(\mathcal{P})) = \dim(\text{affhull}(\mathcal{P}) - \{p\}) \leq \text{span}\,(\mathcal{P}_p)$. Therefore, $\dim(\text{affhull}(\mathcal{P})) = \dim(\text{span}\,(\mathcal{P}_p))$.

As $\mathcal{Y}$ is a finite set, $\text{span}\,(\mathcal{P}_p)$ is a finite-dimensional vector space. The rank-nullity theorem states $\dim(\text{im}(T_W)) + \dim(\ker(T_W)) = \dim(\text{span}\,(\mathcal{P}_p)) = \dim(\text{affhull}(\mathcal{P}))$. As $\dim(\text{im}(T_W)) \leq d$, and we have shown above that $\mathcal{S}_F(p) = \text{span}\,(F_p) = \ker(T_W)$, the conclusion follows. □

Corollary 20 ([71] Theorem 18)**.** *Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ be a discrete loss eliciting $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ with $\mathcal{Y}$ finite. Then for all $p \in \Delta_\mathcal{Y}$ and $r \in \gamma(p)$,*

$$\text{cons}_{\text{cvx}}(\gamma) \geq \|p\|_0 - \dim(\mathcal{S}_{\gamma_r}(p)) - 1 \ . \tag{6.1}$$

Proof. Let $L \in \mathcal{L}_d^{\text{cvx}}$ be a calibrated surrogate for ℓ, and let $\Gamma := \text{prop}_{\Delta_\mathcal{Y}}[L]$. Consider $\mathcal{Y}' := \{y \in \mathcal{Y} : p_y > 0\}$ and $p' = (p_y)_{y \in \mathcal{Y}'} \in \Delta_{\mathcal{Y}'}$. Take $L' := L|_{\mathcal{Y}'}$ and $\ell' := \ell|_{\mathcal{Y}'}$. Define $h : \mathbb{R}^{\mathcal{Y}'} \to \mathbb{R}_\mathcal{Y}$ such that $h(q') = q$ such that $q_y = q'_y$ for $y \in \mathcal{Y}'$ and $q_y = 0$ otherwise. Take $\Gamma' = \Gamma \circ h$, $\gamma' = \gamma \circ h$.

We wish to first show L' indirectly elicits γ'. Since L indirectly elicits γ, we have a link ψ such that for all $u \in \mathbb{R}^d$, $\Gamma_u \subseteq \gamma_{\psi(u)}$. As $\Gamma'(q) = \Gamma(h(q))$ and $\gamma'(q) = \gamma(h(q))$, we have $q \in \Gamma'_u \iff h(q) \in \Gamma_u \implies h(q) \in \gamma_{\psi(u)} \iff (q_y)_{y \in \mathcal{Y}'} \in \gamma'_{\psi(u)}$, and therefore, L' indirectly elicits γ' via the link $\psi \circ \text{proj}\,(\mathcal{Y}')$, where $\text{proj}\,(\mathcal{Y}') : q \mapsto (q_y)_{y \in \mathcal{Y}'}$.

We aim to show $\dim(\mathcal{S}_{\gamma_r}(p)) \geq \dim(\mathcal{S}_{\gamma'_r}(p'))$. We do this by showing that $h(\mathcal{S}_{\gamma'_r}(p')) \subseteq \mathcal{S}_{\gamma_r}(p)$, and the result holds as h is linear and injective. Suppose $v \in h(\mathcal{S}_{\gamma'_r}(p'))$, then there exists a v' so that $v = h(v')$ and an $\epsilon_0 > 0$ such that $\epsilon v' + p' \in \gamma'_r$ for all $\epsilon \in (-\epsilon_0, \epsilon_0)$. Since h is linear and recall $h(\gamma'_r) \subseteq \gamma_r$, this implies $\epsilon v + p \in \gamma_r$ for all $\epsilon \in (-\epsilon_0, \epsilon_0)$. Therefore $v \in \mathcal{S}_{\gamma_r}(p)$, and the result follows.

As L' indirectly elicits γ', by Corollary 19, we know there exists a d-flat F with $p' \in F \subseteq \gamma'_r$. Taking $\mathcal{P} = \Delta_{\mathcal{Y}'}$, we know $p' \in \text{relint}\Delta_{\mathcal{Y}'}$ by construction, so we can apply Lemma 47(ii), which gives $\dim(\mathcal{S}_F(p')) \geq \dim(\text{affhull}(\Delta_{\mathcal{Y}'})) - d = \|p\|_0 - 1 - d$.[4] Additionally, $\mathcal{S}_F(p') \subseteq \mathcal{S}_{\gamma_r'}(p')$ by subset inclusion of the sets themselves. Chaining these results, we obtain

$$\dim(\mathcal{S}_{\gamma_r}(p)) \geq \dim(\mathcal{S}_{\gamma_r'}(p')) \geq \dim(\mathcal{S}_F(p')) \geq \|p\|_0 - 1 - d\ .$$

□

6.5 Applications in Q3, Q4: Variance, Risk Measures, Mode, and Modal Interval

We now turn to two main applications of Theorem 20: new lower bounds on the convex consistency dimension of risk measures (§ 6.5.2) and the mode and modal interval (§ 6.5.3). In both cases, we build on previous results due to Frongillo and Kash [39, 40] and Dearborn and Frongillo [22] which showed lower bounds with respect to *identifiable* properties; a property is d-identifiable if its level sets are all d-flats, as in Figure 6.3 (Q3). In contrast, properties elicited by convex losses are generally not identifiable, particularly when the loss is non-smooth. For example, the properties elicited by hinge loss and the abstain surrogate are not identifiable, as their level sets are not flats; see Figure 6.3 (Q1). It therefore might appear that entirely new ideas are needed. Indeed, both papers above pose developing similar bounds with respect to convex-elicitable properties as a major open question.

Using our d-flats framework, we resolve both open questions with new lower bounds in both settings. Our framework clarifies the relationship between d-identifiable properties and properties elicited by d-dimensional convex losses: the level sets of the former are d-flats by definition, while the level sets of the latter are *unions* of d-flats by Theorem 20. A careful examination of the arguments of Frongillo and Kash [39, 40] and Dearborn and Frongillo [22] reveals that they largely rely on the containment of d-flats in level sets, rather than the full structure of identifiable properties. As such, although quite subtle in the case of risk measures, the general structure of these previous proofs go through for convex-elicitable properties: since no d-flat could be contained in a particular level set,

[4] To reason about $\dim(\text{affhull}(\Delta_{\mathcal{Y}'})) = \|p\|_0 - 1$, observe that the uniform distribution on $\Delta_{\mathcal{Y}'}$ has full support and therefore requires $\|p\|_0 - 1$ elements in its basis.

no union of d-flats could be either. Our lower bounds therefore match both of these papers, though we conjecture that our convex consistency bounds could be tightened in some cases.

6.5.1 Lower-bounding the convex consistency dimension of the variance

Corollary 21. *Let $\mathcal{P}$ be a set of continuous Lebesgue densities on $\mathcal{Y} = \mathbb{R}$ with all $p \in \mathcal{P}$ having the same support. If there exist $p, q, q' \in \mathcal{P}$ with $\mathbb{E}_p Y = \mathbb{E}_q Y \neq \mathbb{E}_{q'} Y$ and $\mathrm{Var}(p) \neq \mathrm{Var}(q)$, then $\mathrm{cons}_{\mathrm{cvx}}(\mathrm{Var}) = \mathrm{elic}_{\mathrm{cvx}}(\mathrm{Var}) = 2$.*

Proof. For the upper bound, we may elicit the first two moments via the convex loss $L(r, y) = (r_1 - y)^2 + (r_2 - y^2)^2$, and recover the variance via $\psi(r) = r_2 - r_1^2$, giving $\mathrm{elic}_{\mathrm{cvx}}(\mathrm{Var}) \leq 2$. Now for the lower bound. Without loss of generality, $\mathbb{E}_q Y < \mathbb{E}_{q'} Y$. Let $r = \frac{1}{2}\mathbb{E}_q Y + \frac{1}{2}\mathbb{E}_{q'} Y$, and define $V : \mathcal{Y} \to \mathbb{R}, y \mapsto y - r$. Then $\ker_{\mathcal{P}} V = \{p' \in \mathcal{P} \mid \mathbb{E}_{p'} Y = r\} = \Gamma_r$ where $\Gamma : p' \mapsto \mathbb{E}_{p'} Y$ is the mean. As $\mathbb{E}_q Y < r < \mathbb{E}_{q'} Y$, we conclude $\mathbb{E}_q V < 0 < \mathbb{E}_{q'} V$. We have now satisfied Condition 6 for $d = 1$. To apply Theorem 21, it remains to show that Var is non-constant on Γ_r. By our assumptions and the definition of Var, we have $\mathbb{E}_p Y^2 \neq \mathbb{E}_q Y^2$. Letting $p_1 = \frac{1}{2}q + \frac{1}{2}q'$, $p_2 = \frac{1}{2}p + \frac{1}{2}q'$, we have $\mathbb{E}_{p_i} Y = r$ for $i \in \{1, 2\}$, but $\mathbb{E}_{p_1} Y^2 = \frac{1}{2}\mathbb{E}_q Y^2 + \frac{1}{2}\mathbb{E}_{q'} Y^2 \neq \frac{1}{2}\mathbb{E}_p Y^2 + \frac{1}{2}\mathbb{E}_{q'} Y^2 = \mathbb{E}_{p_2} Y^2$. As p_1, p_2 have the same mean but different second moments, we conclude $\mathrm{Var}(p_1) \neq \mathrm{Var}(p_2)$. □

6.5.2 Risk measures (Q4)

The problem of estimating a risk or uncertainty measure of Y is of central importance in financial regulation [4, 18, 34] and robust engineering design [13, 76, 80]. Risk measures include the upper confidence bound $\mathbb{E}[Y] + \lambda\sqrt{\mathrm{Var}[Y]}$, or the conditional value at risk (CVaR) defined below in eq. (6.2), in either conditional or unconditional contexts. Uncertainty measures include the variance, entropy, or norm of the distribution of Y. Risk and uncertainty measures are typically not elicitable, so this problem falls under Quadrant 4. Frongillo and Kash [39, 40] give prediction dimension lower bounds for a broad class of risk and uncertainty measures, namely Bayes risks. As stated above, these bounds are with respect to identifiable properties, and bounds for convex surrogates are left as a major open question.

We resolve this open question using our d-flats framework, giving a matching result for convex-elicitable properties (Theorem 21). First we recall the definition of the Bayes risk.

Definition 47. *Given loss function $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ for some report set $\mathcal{R}$, the Bayes risk of L is defined as $\underline{L}(p) := \inf_{r \in \mathcal{R}} \mathbb{E}_p L(r, Y)$.*

Condition 6. *For some $r \in \operatorname{range} \Gamma$, the level set $\Gamma_r = \ker_{\mathcal{P}} V$ is a d-flat presented by some $V : \mathcal{Y} \to \mathbb{R}^d$ such that $0 \in \operatorname{int} \{\mathbb{E}_p V : p \in \mathcal{P}\}$.*

Theorem 21. *Let $\mathcal{P}$ be a convex set of Lebesgue densities supported on the same set for all $p \in \mathcal{P}$. Let $\Gamma : \mathcal{P} \to \mathbb{R}^d$ satisfy Condition 6 for some $r \in \mathbb{R}^d$. Let $L \in \mathcal{L}^{\text{cvx}}$ elicit Γ such that $\underline{L}$ is non-constant on Γ_r. Then $\operatorname{cons}_{\text{cvx}}(\underline{L}) \geq \operatorname{elic}_{\text{cvx}}(\underline{L}) \geq d + 1$.*

To illustrate the theorem, we briefly apply it to one of the most prominent financial risk measures, the conditional value at risk (CVaR). Several other applications from Frongillo and Kash [39, 40], such as other risk measures, entropy, and norms, follow similarly. The authors observe that CVaR can be expressed as a Bayes risk; for $0 < \alpha < 1$, we may define

$$\mathrm{CVaR}_\alpha(p) = \inf_{r \in \mathbb{R}} \mathbb{E}_p \left\{ \tfrac{1}{\alpha}(r - Y)\mathbb{1}_{r \geq Y} - r \right\} \ , \tag{6.2}$$

which is the Bayes risk of the transformed pinball loss $L_\alpha(r, y) = \frac{1}{\alpha}(r - y)\mathbb{1}_{r \geq y} - r$. In turn, L_α elicits the α-quantile, the quantity $q_\alpha(p)$ such that $\mathbb{P}_p[Y \geq q_\alpha(p)] = \alpha$. Following Frongillo and Kash [40], we will restrict to the set $\mathcal{P}_q$ of probability measures over $\mathbb{R}$ with connected support and whose CDFs are strictly increasing on their support, so that q_α is single-valued. Under mild assumptions, we find that there is no consistent real-valued convex surrogate for CVaR_α.

Corollary 22. *Let $\mathcal{P}$ be a convex set of continuous Lebesgue densities on $\mathcal{Y} = \mathbb{R}$ with all $p \in \mathcal{P}$ having support on the same interval. If we have $p_1, p_2, p_3, p_2' \in \mathcal{P}$ with $q_\alpha(p_1) < q_\alpha(p_2) < q_\alpha(p_3)$ and $\mathrm{CVaR}_\alpha(p_2) \neq \mathrm{CVaR}_\alpha(p_2')$, then $\operatorname{cons}_{\text{cvx}}(\mathrm{CVaR}_\alpha) \geq \operatorname{elic}_{\text{cvx}}(\mathrm{CVaR}_\alpha) \geq 2$.*

As first shown by Fissler et al. [35], the pair $(\mathrm{CVaR}_\alpha, q_\alpha)$ is jointly identifiable and elicitable, but not by any convex loss [33, Prop. 4.2.31]. We conjecture the stronger statement $\operatorname{elic}_{\text{cvx}}(\mathrm{CVaR}_\alpha) \geq 3$, which if true would constitute an interesting gap between elicitation complexity for identifiable and convex-elicitable properties.

6.5.3 Mode and modal interval (Q4, Q3)

For finite $|\mathcal{Y}|$, the mode $\mathrm{mode}(p) = \arg\max_{y\in\mathcal{Y}} p(y)$ is elicited by 0-1 loss. By contrast, for $\mathcal{Y} = \mathbb{R}$, the mode is not elicitable [49], landing it in Quadrant 4. Defining the mode is subtle for general distributions; here let us assume p has a smooth and bounded Lesbegue density f_p, and define the mode the same way, $\mathrm{mode}(p) = \arg\max_{y\in\mathcal{Y}} f_p(y)$. Dearborn and Frongillo [22] recently showed a strong impossibility result, that the mode has countably infinite elicitation complexity with respect to identifiable properties. In other words, it is as hard to elicit the mode as the full distribution p itself. Complexity with respect to convex-elicitable properties is left as an important open question.

We resolve this question, with a matching infinite lower bound for convex-elicitable properties. In light of our d-flats framework, the result is nearly immediate, as the proof in Dearborn and Frongillo [22] already showed that the level sets of the mode cannot contain any d-flats.

Theorem 22 ([30, Theorem 3])**.** *The mode has* $\mathrm{cons}_{\mathrm{cvx}}(\mathrm{mode}) = \mathrm{elic}_{\mathrm{cvx}}(\mathrm{mode}) = \infty$ *(countably infinite) with respect to* $\mathcal{P}$*, the class of probability measures on* $\mathcal{Y} = \mathbb{R}$ *with a smooth and bounded density and such that* mode *is single-valued.*

Proof. The proof of Dearborn and Frongillo [22, Theorem 1] gives a distribution $p \in \mathcal{P}$ with $\mathrm{mode}(p) = 0 =: u$. It then introduces an arbitrary identification function $V : \hat{\mathcal{R}} \times \mathcal{Y} \to \mathbb{R}^k$, $k \in \mathbb{N}$, and value $r \in \hat{\mathcal{R}}$ such that $p \in \ker_{\mathcal{P}} V(r, \cdot)$. Letting $F = \ker_{\mathcal{P}} V(r, \cdot)$, we therefore have an arbitrary k-flat containing p. The proof then proceeds to construct some $p' \in F$ with $\mathrm{mode}(p') \neq u$. Corollary 19 now gives $\mathrm{cons}_{\mathrm{cvx}}(\mathrm{mode}) \geq \mathrm{elic}_{\mathrm{cvx}}(\mathrm{mode}) \geq k + 1$. As k was arbitrary, the result follows. □

A closely related property for any $\beta > 0$ is the (midpoint of the) modal interval of width 2β, given by $\gamma_\beta(p) = \arg\max_{x\in\mathbb{R}} p([x - \beta, x + \beta])$. Interestingly, unlike the mode for $\mathcal{Y} = \mathbb{R}$, the modal interval is elicitable, by the target loss $\ell_\beta(r, y) = \mathbb{1}\{|r - y| > \beta\}$. The problem of estimating the modal interval therefore could be thought of as falling under Quadrant 3.

As observed in Dearborn and Frongillo [22, Corollary 1], the properties mode and γ_β coincide with the family of distributions needed in their Theorem 1, meaning the conclusion of Theorem 22 transfers to the modal interval as well.

Corollary 23. *For any $\beta > 0$, the modal interval $\gamma_\beta : \mathcal{P}_\beta \to \mathbb{R}$ has $\mathrm{cons}_{\mathrm{cvx}}(\gamma_\beta) = \mathrm{elic}_{\mathrm{cvx}}(\gamma_\beta) = \infty$ (countably infinite) with respect to $\mathcal{P}_\beta$, the class probability measures on $\mathcal{Y} = \mathbb{R}$ with a smooth and bounded density, and such that* mode *and γ_β are single-valued.*

Thus, while γ_β is elicitable, it does not have any consistent finite-dimensional convex surrogate. While this statement may seem counter-intuitive, recall that the mode for finite $|\mathcal{Y}|$ has $\mathrm{cons}_{\mathrm{cvx}}(\mathrm{mode}) = |\mathcal{Y}| - 1$. Taking the limit as $|\mathcal{Y}| \to \infty$, one may therefore expect an infinite convex consistency dimension for both the mode and modal interval.

6.6 Chapter conclusion

This chapter introduces d-flats, a tool to generate lower bounds on the convex consistency dimension of general prediction tasks. This tool is simultaneously broader, stronger, and easier to understand than previous results. Its breadth is demonstrated by applying to multiple problem types simultaneously (§ 6.3), while its strength is demonstrated by proving new bounds on convex consistency dimension (§ 6.5), and ease is apparent when observing that indirect elicitation is a strictly weaker notion than calibration – the most common proxy for consistency. We then apply our framework to yield new bounds on convex consistency dimension for entropy, risk measures, the mode, and modal intervals.

Several important questions remain open. Particularly for the discrete settings, we would like to know whether one can lift the restriction that surrogates always achieve a minimum; we conjecture positively. The observation that our bounds are as tight as calibration-based bounds, yet we use the weaker condition of indirect elicitation, motivates the study of how much weaker indirect elicitation is than calibration. More broadly, we would like to characterize $\mathrm{cons}_{\mathrm{cvx}}$ and $\mathrm{elic}_{\mathrm{cvx}}$ and develop a general framework for constructing surrogates achieving the best possible prediction dimension.

6.7 Chapter appendix

6.7.1 Proof of Theorem 21

Throughout this section, we will assume $\mathcal{P}$ is convex. See Frongillo and Kash [40, §E.5] for a discussion of how to relax this assumption.

6.7.2 General setting of elicitation complexity

We briefly introduce the general notion of elicitation complexity, of which Definition 12 is a special case, as some statements are more naturally made in this general setting.

Definition 48. *Γ' refines Γ if for all $r' \in \operatorname{range}\Gamma'$ there exists $r \in \operatorname{range}\Gamma$ with $\Gamma'_{r'} \subseteq \Gamma_r$.*

Equivalently, Γ' refines Γ if there is a link function $\psi : \operatorname{range}\Gamma' \to \operatorname{range}\Gamma$ such that $\Gamma'_{r'} \subseteq \Gamma_{\psi(r')}$ for all $r' \in \operatorname{range}\Gamma'$.

Definition 49. *For $k \in \mathbb{N} \cup \{\infty\}$, let $\mathcal{E}_k(\mathcal{P})$ denote the class of all elicitable properties $\Gamma : \mathcal{P} \to \mathbb{R}^k$, and $\mathcal{E}(\mathcal{P}) := \bigcup_{k\in\mathbb{N}\cup\{\infty\}} \mathcal{E}_k(\mathcal{P})$. When $\mathcal{P}$ is implicit we simply write $\mathcal{E}$.*

Definition 50. *Let $\mathcal{C}$ be a class of properties. The elicitation complexity of a property Γ with respect to $\mathcal{C}$, denoted $\mathrm{elic}_{\mathcal{C}}(\Gamma)$, is the minimum value of $k \in \mathbb{N} \cup \{\infty\}$ such that there exists $\hat{\Gamma} \in \mathcal{C} \cap \mathcal{E}_k(\mathcal{P})$ that refines Γ.*

6.7.3 Supporting statements

Proposition 20 (Osband [67])**.** *Let Γ be elicitable. Then Γ_r is convex for all $r \in \operatorname{range}\Gamma$.*

Lemma 48 (Set-valued extension of Frongillo and Kash [40, Lemma 4])**.** *If Γ' refines Γ then $\mathrm{elic}_{\mathcal{C}}(\Gamma') \geq \mathrm{elic}_{\mathcal{C}}(\Gamma)$.*

Proof. As Γ' refines Γ, we have some $\psi : \operatorname{range}\Gamma' \to \operatorname{range}\Gamma$ such that for all $r' \in \operatorname{range}\Gamma'$ we have $\Gamma'_{r'} \subseteq \Gamma_{\psi(r')}$. Suppose we have $\hat{\Gamma} \in \mathcal{C}$ and $\varphi : \operatorname{range}\hat{\Gamma} \to \operatorname{range}\Gamma'$ such that for all $u \in \operatorname{range}\hat{\Gamma}$ we have $\hat{\Gamma}_u \subseteq \Gamma'_{\varphi(u)}$. Then for all $u \in \operatorname{range}\hat{\Gamma}$ we have $\hat{\Gamma}_u \subseteq \Gamma'_{\varphi(u)} \subseteq \Gamma_{(\psi\circ\varphi)(u)}$. In particular, if $\operatorname{elic}_{\mathcal{C}}(\Gamma') = m$, then we have such a $\hat{\Gamma} : \mathcal{P} \rightrightarrows \mathbb{R}^m$, and hence $\operatorname{elic}_{\mathcal{C}}(\Gamma) \leq m$. □

Lemma 49 (Frongillo and Kash [40, Lemma 8]). *Suppose $L \in \mathcal{L}$ elicits $\Gamma : \mathcal{P} \to \mathcal{R}$ and has Bayes risk $\underline{L}$. Then for any $p, p' \in \mathcal{P}$ with $\Gamma(p) \neq \Gamma(p')$, we have $\underline{L}(\lambda p + (1-\lambda)p') > \lambda\underline{L}(p) + (1-\lambda)\underline{L}(p')$ for all $\lambda \in (0,1)$.*

Lemma 50 (Adapted from Frongillo and Kash [40, Theorem 4]). *If L elicits a single-valued Γ, and $\hat{\Gamma}$ refines $\underline{L}$, then $\hat{\Gamma}$ refines Γ.*

Proof. Suppose for a contradiction that $\hat{\Gamma}$ does not refine Γ. Then we have some $u \in \operatorname{range}\hat{\Gamma}$ such that for all $r \in \operatorname{range}\Gamma$ we have $\hat{\Gamma}_u \not\subseteq \Gamma_r$. In particular, recalling that Γ is single-valued, we must have $p, p' \in \hat{\Gamma}_u$ such that $\Gamma(p) \neq \Gamma(p')$. Moreover, as $\hat{\Gamma}$ refines $\underline{L}$, we also have $\underline{L}(p) = \underline{L}(p')$. From Lemma 49 and $\lambda = 1/2$ we have $\underline{L}(q) > \frac{1}{2}\underline{L}(p) + \frac{1}{2}\underline{L}(p') = \underline{L}(p)$, where $q = \frac{1}{2}p + \frac{1}{2}p'$. As the level set $\hat{\Gamma}_u$ is convex by Proposition 20, we also have $q \in \hat{\Gamma}_u$, and hence $\underline{L}(q) = \underline{L}(p)$, a contradiction. □

Lemma 51 (Minor modifications from Frongillo and Kash [40]). *Let $\mathcal{V}$ be a real vector space. Let $f : \mathcal{V} \to \mathbb{R}^k$ be linear and $C \subseteq \mathcal{V}$ convex with $\operatorname{span} C = \mathcal{V}$, and let $m \in \mathbb{N}$. Suppose that $0 \in \operatorname{int} f(C)$, and for all $v \in S := C \cap \ker f$, there exists a linear $\hat{f}_v : \mathcal{V} \to \mathbb{R}^m$ with $v \in C \cap \ker \hat{f}_v \subseteq S$. Then $m \geq k$. If $m = k$, we additionally have $0 \in \operatorname{int} \hat{f}_v(C)$ for some $v \in S$.*

Proof. The condition $0 \in \operatorname{int} f(C)$ is equivalent to the existence of some $v_1, \ldots v_{k+1} \in C$ such that $0 \in \operatorname{int}\operatorname{conv}\{f(v_i) : i \in \{1, \ldots, k+1\}\}$. Let $\alpha_1, \ldots, \alpha_{k+1} > 0$, $\sum_{i=1}^{k+1} \alpha_i = 1$, such that $\sum_{i=1}^{k+1} \alpha_i f(v_i) = 0$. As these are barycentric coordinates, this choice of α_i is unique, a fact which will be important later. We will take $v = \sum_{i=1}^{k+1} \alpha_i v_i$, an element of C by convexity, and thus an element of S as $f(v) = 0$.

Let $\hat{f}_v : \mathcal{V} \to \mathbb{R}^m$ be linear with $v \in \hat{S} := C \cap \ker \hat{f}_v \subseteq S$. Let $\beta_1, \ldots, \beta_{k+1} \in \mathbb{R}$, $\sum_{i=1}^{k+1} \beta_i = 0$, such that $\sum_{i=1}^{k+1} \beta_i \hat{f}_v(v_i) = 0$. We will show that the β_i must be identically zero, i.e. that $\{\hat{f}_v(v_i) :$

$i \in \{1, \ldots, k+1\}\}$ are affinely independent. By construction, $v' := \sum_{i=1}^{k+1} \beta_i v_i \in \ker \hat{f}_v$, and as $v \in \ker \hat{f}_v$, for all $\lambda > 0$ we have $v_\lambda := v + \lambda v' = \sum_{i=1}^{k+1}(\alpha_i + \lambda\beta_i)v_i \in \ker \hat{f}_v$. Taking λ sufficiently small, we have $\gamma_i := \alpha_i + \lambda\beta_i > 0$ for all i, and $\sum_{i=1}^{k+1} \gamma_i = \sum_{i=1}^{k+1} \alpha_i + \lambda \sum_{i=1}^{k+1} \beta_i = 1$. By convexity of C, we have $v_\lambda \in C$. Now $v_\lambda \in C \cap \ker \hat{f}_v \subseteq S = C \cap \ker f$, and in particular $v_\lambda \in \ker f$. Thus, $f(v_\lambda) = \sum_{i=1}^{k+1} \gamma_i f(v_i) = 0$. By the uniqueness of barycentric coordinates, for all $i \in \{1, \ldots, k+1\}$, we must have $\gamma_i = \alpha_i$ and thus $\beta_i = 0$, as desired.

As $\hat{f}_v(C)$ contains $k+1$ affinely independent points, we have $m \geq \dim \mathrm{im} \hat{f}_v \geq k$. When $m = k$, by affine independence, the set $\mathrm{conv}\{\hat{f}_v(v_i) : i \in \{1, \ldots, k+1\}\}$ has dimension k in $\mathbb{R}^k$. As $0 = \hat{f}_v(v) = \sum_{i=1}^{k+1} \alpha_i \hat{f}_v(v_i)$, and $\alpha_i > 0$ for all i, we conclude $0 \in \mathrm{int}\,\mathrm{conv}\{\hat{f}_v(v_i) : i \in \{1, \ldots, k+1\}\} \subseteq \mathrm{int}\, \hat{f}_v(C)$. $\square$

Lemma 52 (Frongillo and Kash [40, Lemma 14])**.** *Let $\mathcal{V}$ be a real vector space. Let $f : \mathcal{V} \to \mathbb{R}^k$ be linear, $C \subseteq \mathcal{V}$ convex with* $\mathrm{span}\, C = \mathcal{V}$*, and let $S = C \cap \ker f$. If* $0 \in \mathrm{int}\, f(C)$ *then* $\mathrm{span}\, S = \ker f$*.*

6.7.4 Proving the lower bound for Bayes risks

Let $\mathcal{C}_d^*$ be the class of properties Γ which are elicited by a convex loss $L \in \mathcal{L}_d^{\mathrm{cvx}}$ for some $d \in \mathbb{N}$, and let $\mathcal{C}^* := \bigcup_{d \in \mathbb{N}} \mathcal{C}_d^*$. Then for all properties γ, if $\mathrm{elic}_{\mathcal{C}^*}(\gamma) < \infty$, we have $\mathrm{elic}_{\mathcal{C}^*}(\gamma) = \mathrm{elic}_{\mathrm{cvx}}(\gamma)$, a fact we use tacitly in the proof.

Theorem 21. *Let $\mathcal{P}$ be a convex set of Lebesgue densities supported on the same set for all $p \in \mathcal{P}$. Let $\Gamma : \mathcal{P} \to \mathbb{R}^d$ satisfy Condition 6 for some $r \in \mathbb{R}^d$. Let $L \in \mathcal{L}^{\mathrm{cvx}}$ elicit Γ such that $\underline{L}$ is non-constant on Γ_r. Then* $\mathrm{cons}_{\mathrm{cvx}}(\underline{L}) \geq \mathrm{elic}_{\mathrm{cvx}}(\underline{L}) \geq d+1$.

Proof. Let $V : \mathcal{Y} \to \mathbb{R}^d$ and r be given by the statement of the theorem and from Condition 6. Let $m = \mathrm{elic}_{\mathcal{C}^*}(\underline{L})$, so that we have $\hat{\Gamma} \in \mathcal{C}_m^*$ which refines $\underline{L}$. By Lemma 50 we have $\hat{\Gamma}$ refines Γ.

We now establish the conditions of Lemma 51 for $C = \mathcal{P}$. Let $f : \mathrm{span}\, \mathcal{P} \to \mathbb{R}^d$, $p \mapsto \mathbb{E}_p V$. From Condition 6, we have $0 \in \mathrm{int}\, f(\mathcal{P})$ and $\ker f \cap \mathcal{P} = \ker_{\mathcal{P}} V = \Gamma_r$. Now let $p \in \Gamma_r$ be arbitrary, and take any $u \in \hat{\Gamma}(p)$. As Γ is single-valued, $r \in \mathrm{range}\, \Gamma$ is the unique value with $p \in \Gamma_r$. As $\hat{\Gamma}$ refines Γ, there exists $r' \in \mathrm{range}\, \Gamma$ with $\hat{\Gamma}_u \subseteq \Gamma_{r'}$, and since $p \in \hat{\Gamma}_u$, we conclude $r' = r$ from the above. From Theorem 20, we have some $\hat{V}_{u,p}$ with $p \in \ker_{\mathcal{P}} \hat{V}_{u,p} \subseteq \hat{\Gamma}_u \subseteq \Gamma_r = \ker_{\mathcal{P}} V$. Letting $\hat{f}_p : \operatorname{span} \mathcal{P} \to \mathbb{R}^d$, $p \mapsto \mathbb{E}_p \hat{V}_{u,p}$, we have now satisfied the conditions of Lemma 51. We conclude $m \geq d$, and moreover, if $m = d$, then there exists some $q \in \Gamma_r$ such that $0 \in \operatorname{int} \hat{f}_q(\mathcal{P})$.

Now suppose $m = d$ for a contradiction. Let $\hat{S} := \ker f_q \cap \mathcal{P}$. Applying Lemma 52 to the functions f and $\hat{f}_q$ we have $\operatorname{span} \ker f = \operatorname{span} \Gamma_r$ and $\operatorname{span} \ker \hat{f}_q = \operatorname{span} \hat{S}$. As $\hat{S} \subseteq \Gamma_r$, we have $\ker \hat{f}_q = \operatorname{span} \hat{S} \subseteq \operatorname{span} \Gamma_r = \ker f$. By the first isomorphism theorem, we also have $\operatorname{codim} \ker \hat{f}_q = \operatorname{codim} \ker f = d$, as the images of these linear maps span all of $\mathbb{R}^d$. By the third isomorphism theorem we conclude $\Gamma_r = \hat{S}$. Moreover, as $\hat{S} \subseteq \hat{\Gamma}_u \subseteq \Gamma_r$, we have $\hat{S} = \hat{\Gamma}_u = \Gamma_r$.

We now see that $\underline{L}$ is constant on Γ_r since there is some link function $\psi : \mathbb{R}^m \to \mathbb{R}$ such that $\Gamma_r = \hat{\Gamma}_u \subseteq \underline{L}_{\psi(u)}$, meaning $\underline{L}(p) = \psi(u)$ for all $p \in \Gamma_r$. This statement contradicts the assumption that $\underline{L}$ is non-constant on Γ_r. □

Bibliography

[1] Stephen Abbott. Understanding analysis. Springer, 2001.

[2] Jacob Abernethy, Yiling Chen, and Jennifer Wortman Vaughan. Efficient market making via convex optimization, and a connection to online learning. ACM Transactions on Economics and Computation, 1(2):12, 2013. URL http://dl.acm.org/citation.cfm?id=2465777.

[3] Jacob D. Abernethy and Rafael M. Frongillo. A collaborative mechanism for crowdsourcing prediction problems. In Advances in Neural Information Processing Systems 24, pages 2600–2608, 2011.

[4] Carlo Acerbi and Balazs Szekely. Back-testing expected shortfall. Risk, November 2014.

[5] Adewole S. Adamson and Avery Smith. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatology, 154(11):1247–1248, 11 2018. ISSN 2168-6068. doi: 10.1001/jamadermatol.2018.2348. URL https://doi.org/10.1001/jamadermatol.2018.2348.

[6] Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In JMLR Workshop and Conference Proceedings, volume 40, pages 1–19, 2015. URL http://www.jmlr.org/proceedings/papers/v40/Agarwal15.pdf.

[7] Tom M Apostol. Mathematical analysis. 1974.

[8] Franz Aurenhammer. Power diagrams: properties, algorithms and applications. SIAM Journal on Computing, 16(1):78–96, 1987. URL http://epubs.siam.org/doi/pdf/10.1137/0216006.

[9] Han Bao, Clayton Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. The Conference on Learning Theory (COLT), 2020.

[10] Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. Journal of Machine Learning Research, 9(Aug):1823–1840, 2008.

[11] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. Journal of the American Statistical Association, 101(473):138–156, 2006. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214505000000907.

[12] Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Smooth loss functions for deep top-k classification. CoRR, abs/1802.07595, 2018. URL http://arxiv.org/abs/1802.07595.

[13] Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization–a comprehensive survey. Computer methods in applied mechanics and engineering, 196(33):3190–3218, 2007. URL http://www.sciencedirect.com/science/article/pii/S0045782507001259.

[14] S.P. Boyd and L. Vandenberghe. Convex optimization. Cambridge University Press, 2004.

[15] G.W. Brier. Verification of forecasts expressed in terms of probability. Monthly weather review, 78(1):1–3, 1950. ISSN 1520-0493.

[16] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Working draft, November, 3:13, 2005.

[17] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In Ninth European Conference on Speech Communication and Technology, 2005.

[18] Sean D. Campbell. A review of backtesting and backtesting procedures. Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, 2005. URL http://www.federalreserve.gov/pubs/FEDS/2005/200521/200521pap.pdf.

[19] Chi-Keung Chow. An optimum character recognition system using decision functions. IRE Transactions on Electronic Computers, (4):247–254, 1957.

[20] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In International Conference on Algorithmic Learning Theory, pages 67–82. Springer, 2016.

[21] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of machine learning research, 2(Dec):265–292, 2001.

[22] Krisztina Dearborn and Rafael Frongillo. On the indirect elicitability of the mode and modal interval. Annals of the Institute of Statistical Mathematics, 72(5):1095–1108, 2020.

[23] John Duchi, Khashayar Khosravi, Feng Ruan, et al. Multiclass classification, information, divergence and surrogate risk. The Annals of Statistics, 46(6B):3246–3275, 2018.

[24] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

[25] Jianqing Fan and Qiwei Yao. Efficient estimation of conditional variance functions in stochastic regression. Biometrika, 85(3):645–660, 1998.

[26] Jessica Finocchiaro and Rafael Frongillo. Convex elicitation of continuous properties. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 10425–10434. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8241-convex-elicitation-of-continuous-properties.pdf.

[27] Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. In Advances in neural information processing systems, 2019.

[28] Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates, 2019. URL https://arxiv.org/abs/1907.07330.

[29] Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. Embedding dimension of polyhedral losses. The Conference on Learning Theory, 2020.

[30] Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. Unifying prediction bounds for consistent convex surrogates. arXiv, 2021. URL `https://arxiv.org/pdf/2102.08218.pdf`.

[31] Jessie Finocchiaro, Rafael Frongillo, Emma Goodwill, and Anish Thilagar. Consistent polyhedral surrogates for top-k classification and variants. Preprint, 2022.

[32] Jessie Finocchiaro, Rafael Frongillo, and Enrique Nueve. The structured abstain problem and the lovász hinge. Preprint, 2022.

[33] Tobias Fissler. On higher order elicitability and some limit theorems on the Poisson and Wiener space. PhD thesis, 2017.

[34] Tobias Fissler, Johanna Fasciati-Ziegel, and Tilmann Gneiting. Expected Shortfall is jointly elicitable with Value at Risk-Implications for backtesting. Risk Magazine, 2016.

[35] Tobias Fissler, Johanna F Ziegel, and others. Higher order elicitability and Osband's principle. The Annals of Statistics, 44(4):1680–1707, 2016.

[36] Gerald B Folland. Real analysis: modern techniques and their applications, volume 40. John Wiley & Sons, 1999.

[37] Rafael Frongillo and Ian Kash. General truthfulness characterizations via convex analysis. In Web and Internet Economics, pages 354–370. Springer, 2014.

[38] Rafael Frongillo and Ian Kash. Vector-Valued Property Elicitation. In Proceedings of the 28th Conference on Learning Theory, pages 1–18, 2015.

[39] Rafael Frongillo and Ian A. Kash. On Elicitation Complexity. In Advances in Neural Information Processing Systems 29, 2015.

[40] Rafael Frongillo and Ian A Kash. Elicitation Complexity of Statistical Properties. Biometrika, 11 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa093. URL `https://doi.org/10.1093/biomet/asaa093`.

[41] Rafael Frongillo and Bo Waggoner. An Axiomatic Study of Scoring Rule Markets. Preprint, 2017.

[42] Rafael Frongillo and Bo Waggoner. Surrogate regret bounds for polyhedral losses. Advances in Neural Information Processing Systems, 34, 2021.

[43] Jean Gallier. Notes on convex sets, polytopes, polyhedra, combinatorial topology, voronoi diagrams and delaunay triangulations, 2008.

[44] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In Proceedings of the 24th annual conference on learning theory, pages 341–358, 2011.

[45] T. Gneiting. Making and Evaluating Point Forecasts. Journal of the American Statistical Association, 106(494):746–762, 2011.

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

[47] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[48] Tamir Hazan, Joseph Keshet, and David A McAllester. Direct loss minimization for structured prediction. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2010.

[49] C. Heinrich. The mode functional is not elicitable. *Biometrika*, page ast048, 2013. URL `http://biomet.oxfordjournals.org/content/early/2013/11/27/biomet.ast048.short`.

[50] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Science & Business Media, 2012.

[51] Alan J Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4), 1952.

[52] Aleksandr Davidovich Ioffe and Vladimir Mikhailovich Tikhomirov. On minimization of integral functionals. *Functional Analysis and Its Applications*, 3(3):218–227, 1969.

[53] Shalmali Joshi, Sonali Parbhoo, and Finale Doshi-Velez. Pre-emptive learning-to-defer for sequential medical decision-making under uncertainty, 2021. URL `https://arxiv.org/abs/2109.06312`.

[54] Daniel A Klain. The minkowski problem for polytopes. *Advances in Mathematics*, 185(2): 270–288, 2004.

[55] Nicolas S. Lambert. Elicitation and evaluation of statistical forecasts. 2018. URL `https://web.stanford.edu/~nlambert/papers/elicitability.pdf`.

[56] Nicolas S. Lambert and Yoav Shoham. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 109–118, 2009.

[57] Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

[58] Nicolas S Lambert, John Langford, Jennifer Wortman Vaughan, Yiling Chen, Daniel M Reeves, Yoav Shoham, and David M Pennock. An axiomatic characterization of wagering mechanisms. *Journal of Economic Theory*, 156:389–416, 2015.

[59] Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass svm. In *Advances in Neural Information Processing Systems*, pages 325–333, 2015.

[60] Maksim Lapin, Matthias Hein, and Bernt Schiele. Loss functions for top-k error: Analysis and insights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1468–1477, 2016.

[61] Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1533–1554, 2018.

[62] Yi Lin. A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82, 2004.

[63] Yufeng Liu. Fisher consistency of multicategory support vector machines. In Marina Meila and Xiaotong Shen, editors, Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, volume 2 of Proceedings of Machine Learning Research, pages 291–298, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR. URL https://proceedings.mlr.press/v2/liu07b.html.

[64] David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer, 2018.

[65] Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Multilabel reductions: what is my loss optimising? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 10600–10611. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9245-multilabel-reductions-what-is-my-loss-optimising.pdf.

[66] Kent Osband and Stefan Reichelstein. Information-eliciting compensation schemes. Journal of Public Economics, 27(1):107–115, June 1985. ISSN 0047-2727. doi: 10.1016/0047-2727(85)90031-3. URL http://www.sciencedirect.com/science/article/pii/0047272785900313.

[67] Kent Harold Osband. Providing Incentives for Better Cost Forecasting. University of California, Berkeley, 1985.

[68] Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In Advances in Neural Information Processing Systems, pages 302–313, 2017.

[69] Rodrigo López Pouso. A simple proof of the fundamental theorem of calculus for the lebesgue integral. arXiv preprint arXiv:1203.1462, 2012.

[70] Harish Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Convex calibrated surrogates for hierarchical classification. In International Conference on Machine Learning, pages 1852–1860, 2015.

[71] Harish G Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. The Journal of Machine Learning Research, 17(1):397–441, 2016.

[72] Harish G Ramaswamy, Ambuj Tewari, Shivani Agarwal, et al. Consistent algorithms for multiclass classification with an abstain option. Electronic Journal of Statistics, 12(1):530–554, 2018.

[73] Mohammad Rastegari, Chen Fang, and Lorenzo Torresani. Scalable object-class retrieval with approximate and top-k ranking. In 2011 International Conference on Computer Vision, pages 2659–2666, 2011. doi: 10.1109/ICCV.2011.6126556.

[74] Sashank J Reddi, Satyen Kale, Felix Yu, Daniel Holtmann-Rice, Jiecao Chen, and Sanjiv Kumar. Stochastic negative mining for learning with large output spaces. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 1940–1949. PMLR, 2019.

[75] M.D. Reid and R.C. Williamson. Composite binary losses. The Journal of Machine Learning Research, 9999:2387–2422, 2010.

[76] R. Tyrrell Rockafellar and Johannes O. Royset. On buffered failure probability in design and optimization of structures. Reliability Engineering & System Safety, 95(5):499–510, 2010. URL http://www.sciencedirect.com/science/article/pii/S0951832010000177.

[77] R.T. Rockafellar. Convex analysis, volume 28 of Princeton Mathematics Series. Princeton University Press, 1997.

[78] David Ruppert, M. P. Wand, Ulla Holst, and Ola Hösjer. Local polynomial variance-function estimation. Technometrics, 39(3):262–273, 1997. doi: 10.1080/00401706.1997.10485117. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1997.10485117.

[79] L.J. Savage. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, pages 783–801, 1971.

[80] Muhammad Shahbaz, Zhong-Hua Han, W. P. Song, and M. Nadeem Aizud. Surrogate-based robust design optimization of airfoil using inexpensive Monte Carlo method. In Applied Sciences and Technology (IBCAST), 2016 13th International Bhurban Conference on, pages 497–504. IEEE, 2016. URL http://ieeexplore.ieee.org/abstract/document/7429924/.

[81] Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

[82] Ingo Steinwart and Andreas Christmann. Support Vector Machines. Springer Science & Business Media, September 2008. ISBN 978-0-387-77242-4. Google-Books-ID: HUnqnrpYt4IC.

[83] Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and Identification of Properties. In Proceedings of The 27th Conference on Learning Theory, pages 482–526, 2014.

[84] Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. The Journal of Machine Learning Research, 8:1007–1025, 2007. URL http://dl.acm.org/citation.cfm?id=1390325.

[85] Yutong Wang and Clayton Scott. Weston-watkins hinge loss and ordered partitions. Advances in neural information processing systems, 2020.

[86] Christophe Weibel. Minkowski sums of polytopes. Technical report, EPFL, 2007.

[87] Jason Weston, Chris Watkins, et al. Support vector machines for multi-class pattern recognition.

[88] Robert C. Williamson. The geometry of losses. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, Proceedings of The 27th Conference on Learning Theory, volume 35 of Proceedings of Machine Learning Research, pages 1078–1108, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/williamson14.html.

[89] Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite multiclass losses. Journal of Machine Learning Research, 17(223):1–52, 2016.

[90] Jens Witkowski, Rupert Freeman, Jennifer Wortman Vaughan, David M Pennock, and Andreas Krause. Incentive-compatible forecasting competitions. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), 2018.

[91] Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, and Liming Chen. Hierarchical classification of emotional speech. IEEE Transactions on Multimedia, 37, 2007.

[92] Forest Yang and Sanmi Koyejo. On the consistency of top-k surrogate losses. CoRR, abs/1901.11141, 2019. URL `http://arxiv.org/abs/1901.11141`.

[93] Jiaqian Yu and Matthew B Blaschko. The lovász hinge: A novel convex surrogate for submodular losses. IEEE transactions on pattern analysis and machine intelligence, 2018.

[94] Constantin Zalinescu. Sharp estimates for hoffman's constant for systems of linear inequalities and equalities. SIAM Journal on Optimization, 14(2):517–533, 2003.

[95] Mingyuan Zhang, Harish G Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label f-measure. 2020.

[96] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research, 5(Oct):1225–1251, 2004.

[97] Günter M Ziegler. Lectures on polytopes, volume 152. Springer Science & Business Media, 2012.