Proceedings of Machine Learning Research vol 125:1–28, 2020 33rd Annual Conference on Learning Theory

Embedding Dimension of Polyhedral Losses

Jessie Finocchiaro JESSICA.FINOCCHIARO@COLORADO.EDU
Rafael Frongillo RAF@COLORADO.EDU
Bo Waggoner BWAG@COLORADO.EDU
CU Boulder

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

A common technique in supervised learning with discrete losses, such as 0-1 loss, is to optimize a convex surrogate loss over $\mathbb{R}^d$, calibrated with respect to the original loss. In particular, recent work has investigated embedding the original predictions (e.g. labels) as points in $\mathbb{R}^d$, showing an equivalence to using polyhedral surrogates. In this work, we study the notion of the *embedding dimension* of a given discrete loss: the minimum dimension d such that an embedding exists. We characterize d-embeddability for all d, with a particularly tight characterization for $d = 1$ (embedding into the real line), and useful necessary conditions for $d > 1$ in the form of a quadratic feasibility program. We illustrate our results with novel lower bounds for abstain loss.

Keywords: Calibrated surrogates, convex surrogates, proper scoring rules

1. Introduction

In supervised machine learning, one typically measures performance of a model using some loss function ℓ(prediction, observation) assigning a punishment for error comparing one's prediction to the outcome observed in nature. In particular, we study *discrete* losses, in which the predictions lie in a finite set. These are popular for many categorical tasks such as classification, ranking, top-k, set inclusion, and other structured prediction tasks.

Optimizing such discrete losses over a dataset (i.e. empirical risk minimization), however, is typically computationally hard. One therefore generally resorts to optimizing a computationally nice *surrogate loss* L. The key requirement is *calibration*, namely that optimizing surrogate loss allows one to recover the discrete prediction that optimizes the original ℓ-loss. Calibrated surrogate losses yields desirable statistical guarantees such as consistency and excess risk bounds (Tewari and Bartlett, 2007).

The question therefore becomes: Given a discrete loss ℓ, how do we design calibrated surrogates L? In this paper we study *polyhedral* surrogates, meaning piecewise-linear convex functions over prediction space $\mathbb{R}^d$. Finocchiaro et al. (2019) show that polyhedral surrogates are in a strong sense equivalent to an *embedding*: a 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 the authors show that every discrete loss ℓ can be so embedded, in the worst case the dimension required is $d = n-1$ where n is the number of possible observations. A surrogate dimension d significantly below n, such as $O(\log n)$ for classification with an abstain option (Ramaswamy

et al., 2018), can lead to faster downstream optimization and computation, an effect that grows with n; this motivates us to understand when this dimension can be low.

In this work, we define and investigate the embedding dimension of discrete losses, and characterize the d-embeddable losses for each d. Beginning with $d = 1$, i.e. embedding into the real line, we offer a complete characterization via a variety of conceptual and testable/constructive conditions (§ 3). We also show that, 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 (§ 4). A particular contribution is to isolate and investigate the optimality condition, which we significantly reduce from a search over sets of polytopes to a quadratic feasibility program (Definition 17), 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 (Ramaswamy and Agarwal, 2016; Ramaswamy et al., 2018) (§ 5). In particular we apply both our 1-dimension characterization and higher-dimensional quadratic program to obtain previously unknown lower bounds. We conclude with discussion and open questions (§ 6).

1.1. Related work

Zhang (2004); Bartlett et al. (2006) study when a convex, calibrated surrogate exists for a given discrete loss, and in a finite-outcome setting, Tewari and Bartlett (2007) show that calibration (Definition 4) is necessary and sufficient for consistency.

Several works have investigated the problem of reducing the dimensionality of a surrogate loss. Frongillo and Kash (2015) propose *elicitation complexity*, which roughly captures the minimum dimension d of a surrogate loss L in a particular class (not necessarily convex or calibrated) for a given loss ℓ (not necessarily discrete). In the case of convex calibrated surrogates for discrete losses, Ramaswamy and Agarwal (2016) propose the notion of *convex calibration dimension*, which is the minimum dimension of a convex surrogate loss (not necessarily polyhedral) that is calibrated with respect to a given discrete loss. While Ramaswamy and Agarwal (2016) present a condition for obtaining lower bounds on convex calibration dimension for any discrete loss, called *feasible subspace dimension*, it yields vacuous bounds for important examples such as the abstain loss ℓ_α loss (eq. (1)), where we only know the convex calibration to be in the range $1 \leq d \leq \lceil \log_2(n) \rceil$ (Ramaswamy et al., 2018). One contribution of this work is to provide the first nontrivial lower bounds for this loss (see § 6).

In order to concisely discuss embedding dimension, we appeal to notation and terminology from the field of property elicitation (Savage, 1971; Osband and Reichelstein, 1985; Gneiting and Raftery, 2007; Lambert and Shoham, 2009; Lambert, 2018), relating it to the language of calibrated surrogates as needed.

2. Setting

The base object of study is a *discrete loss function* $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ where $\mathcal{R}$ and $\mathcal{Y}$ are finite sets. Here $\ell(r)_y$ is the loss of prediction $r \in \mathcal{R}$ (the report space) on $y \in \mathcal{Y}$ (the observation or label space). For simplicity we let $\mathcal{Y} = [n]$ throughout (where $[n] = \{1, \ldots, n\}$). The set of

probability distributions on $\mathcal{Y}$ is denoted $\Delta_{\mathcal{Y}} \subseteq \mathbb{R}^{\mathcal{Y}}_{\geq 0}$, represented as vectors of probabilities. We write p_y for the probability of outcome $y \in \mathcal{Y}$ drawn from $p \in \Delta_{\mathcal{Y}}$. As is assumed by Finocchiaro et al. (2019), we assume the given discrete loss is *non-redundant*, meaning every report r uniquely minimizes expected loss for some distribution $p \in \Delta_{\mathcal{Y}}$.

We similarly denote surrogate losses in dimension d by $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$, with predictions typically written $u \in \mathbb{R}^d$. The expected losses when $Y \sim p$ can be written $\langle p, \ell(r) \rangle$ and $\langle p, L(u) \rangle$. For example, 0-1 loss on two labels is a discrete loss with $\mathcal{R} = \mathcal{Y} = \{-1, 1\}$ given by $\ell_{0\text{-}1}(r)_y = \mathbf{1}\{r \neq y\}$. Two important surrogates for $\ell_{0\text{-}1}$ are hinge loss $L_{\text{hinge}}(u)_y = \max\{0\ ,\ 1 - yu\}$ and logistic loss $L(u)_y = \log(1 + \exp(-yu))$ for $u \in \mathbb{R}$.

Most of the surrogates L we consider will be *polyhedral*, meaning piecewise linear and convex. 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 (Rockafellar, 1997).

Definition 1 (Polyhedral loss) *A loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ is* polyhedral *if $L(u)_y$ is a polyhedral (piecewise linear convex) function of u for each $y \in \mathcal{Y}$.*

For example, hinge loss is polyhedral, whereas logistic loss is not. In order to study calibration of surrogates (Definition 4 below), it is useful to formalize the following functions that describe the reports that minimize expected ℓ-loss and L-loss under p.

Definition 2 ((Finite) property, level set, elicited) *A* property *is a function $\Gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}'}$ for some set $\mathcal{R}'$ where $\Gamma(p) \neq \emptyset$ for all p. It is* finite *if $|\mathcal{R}'| < \infty$. The* level set *of $r \in \mathcal{R}'$ is the set $\Gamma_r = \{p \in \Delta_{\mathcal{Y}} : r \in \Gamma(p)\}$. A loss function $L : \mathcal{R}' \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$* elicits *$\Gamma$ if $\Gamma(p) = \arg\min_{r \in \mathcal{R}'} \langle p, L(r) \rangle$.*

In this paper we write ℓ for discrete losses (i.e., when $|\mathcal{R}|$ is finite), and γ instead of Γ for the property elicited by ℓ, reserving Γ and L for properties and losses on report space $\mathcal{R} = \mathbb{R}^d$. For example, 0-1 loss elicits the mode, which is formalized as a property on report space $\mathcal{R} = \mathcal{Y}$ via $\gamma(p) = \arg\max_{y \in \mathcal{Y}} p_y$. Notice in this example $\gamma(p)$ is a singleton set when the mode of p is unique and e.g. is $\mathcal{Y}$ when p is uniform.

Polyhedral losses are motivated partly because they correspond to a natural surrogate construction technique: embedding the discrete predictions $\mathcal{R}$ as points in $\mathbb{R}^d$ and linearly interpolating between embedded reports. The key condition required on the embedding and surrogate loss L is that a discrete prediction $r \in \mathcal{R}$ should minimize expected ℓ-loss if and only if its embedding point minimizes expected L-loss.

Definition 3 (Embedding of a loss) *A loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$* embeds *a loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ in dimension d with* embedding function *$\varphi : \mathcal{R} \to \mathbb{R}^d$ if: (i) φ is injective; (ii) for all $r \in \mathcal{R}$, $\ell(r) = L(\varphi(r))$; and (iii) for all $r \in \mathcal{R}$, $\gamma_r = \Gamma_{\varphi(r)}$. We simply say L* embeds *ℓ if some such φ exists.*

Note that embedding does not immediately imply a key desirable property, consistency, which is equivalent to calibration in finite-outcome settings. Informally, calibration means that if one minimizes the expected surrogate loss arbitrarily well over $\mathbb{R}^d$, then applies the link function to obtain a discrete prediction r, then r exactly minimizes ℓ-loss.

Definition 4 (Calibrated surrogate) *A surrogate loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ and* link function *$\psi : \mathbb{R}^d \to \mathcal{R}$ are* calibrated *for a loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ if for all $p \in \Delta_{\mathcal{Y}}$,*

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

We simply say L can be calibrated *to ℓ if such a link function exists.*

As we alluded to above, calibration is not actually the trait we desire of a surrogate; rather, we seek consistency. Bartlett et al. (2006); Tewari and Bartlett (2007) show that, in a finite-outcome setting, the presence of calibration and consistency are equivalent conditions. By consistency, we mean the standard notion where the expected loss over $\mathcal{X} \times \mathcal{Y}$ for a sequence of hypotheses h_m converging to the optimal expected surrogate loss implies the expected discrete loss over the linked hypotheses $\psi \circ h_m$ converges to the optimal discrete loss. The equivalence of these conditions allows us to focus on calibration, so we can abstract away the input space $\mathcal{X}$ and simply focus on probability distributions over the outcomes $\mathcal{Y}$.

Theorem 5 (Finocchiaro et al. (2019)) *Given a discrete loss ℓ, (1) there exists a polyhedral loss L embedding ℓ; and (2) there exists a link function calibrating L to ℓ.*

In other words, polyhedral surrogates are equivalent to embeddings of discrete losses ℓ in $\mathbb{R}^d$, and they come with a guarantee of calibration. This raises the question studied in this paper: for a given discrete loss, in what dimension d does such a surrogate exist? This question is captured by the following quantity.

Definition 6 (Embedding dimension) *We say a discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ is* d-embeddable *if there exists a polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ that embeds it. The* embedding dimension *of ℓ is the smallest d such that ℓ is d-embeddable.*

A number of things are already known about embedding dimension. Many surrogates in the literature provide upper bounds; we highlight in particular the *abstain loss* (Ramaswamy et al., 2018) and seen in eq. (1), in which one wants to predict the most likely outcome *only if* confident in the outcome, and otherwise abstain.

$$\ell_\alpha(r)_y = \begin{cases} 0 & r = y \\ \alpha & r = \perp \\ 1 & r \notin \{y, \perp\} \end{cases} \tag{1}$$

Here $\mathcal{R} = \mathcal{Y} \cup \{\perp\}$. For $\alpha \leq 1/2$, Ramaswamy et al. (2018) 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 $\perp$ is embedded at the origin.

In general (e.g. Ramaswamy and Agarwal (2016, 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 (2016) 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. We discuss further in § 6.

3. One-dimensional embeddings

In this section, we completely characterize 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). 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.

3.1. General characterization via property elicitation

We begin with conditions on the property elicited by a discrete loss. The following condition of Lambert (2018, Theorem 3), that a finite property is *orderable*, states that any two level sets intersect in a hyperplane, or not at all.

Definition 7 (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. The proof involves several intermediate results which we state later in this section; see § 3.3 for more details.

Theorem 8 *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 9 (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 $\gamma_r \cap \gamma_{r'} \cap \operatorname{relint}(\Delta_{\mathcal{Y}}) \neq \emptyset$, where $\operatorname{relint}(\Delta_{\mathcal{Y}})$ 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 10, proven in Appendix A.

Proposition 10 *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.*

Combining Proposition 10 and Theorem 8, 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

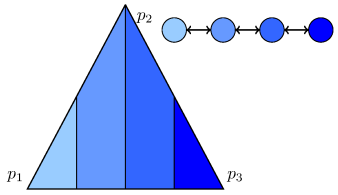

Figure 1: Level sets and intersection graph for a given property, $|\mathcal{Y}| = 3$.

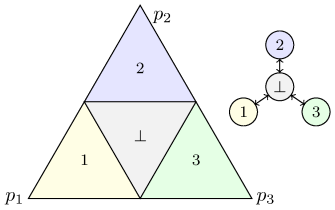

Figure 2: Level sets and intersection graph for the $\text{abstain}_{1/2}$ property, $\mathcal{Y} = \{1, 2, 3\}$.

to be a path. We give an example of a direct application in § 5. By visualizing the level sets of γ as a power diagram (generalization of Voronoi diagram) in the simplex like Lambert and Shoham (2009), we can also use Proposition 10 to perform a visual test for orderability, and thus 1-embeddability (Figures 1 and 2).

3.2. Constructing a surrogate

While Theorem 8 and Proposition 10 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 $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ 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

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

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

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 (§ 3.3). 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.

3.3. Including general convex surrogates

We summarize the above results in the following theorem, together with one additional result, whose proof is in Appendix A: if ℓ has any calibrated convex surrogate at all, it must have a polyhedral one. (See Conjecture 23 regarding higher dimensional generalizations.)

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 $L : \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$; (6) ℓ has some convex calibrated surrogate loss $L : \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$.*

For the proof, note that (1) $\iff$ (2) was shown in Proposition 10, while (4) $\iff$ (5) follows from Theorem 5, and (5) $\implies$ (6) is immediate from the definitions. We therefore prove (1) $\implies$ (3) $\implies$ (5) and (6) $\implies$ (1).

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 (§ 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 (§ 4.2, Theorem 15). 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 (§ 4.3, Theorem 18). 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 19).

4.1. Setup: subgradient sets at embedding points.

Recall that if $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ is embedded by $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$, then each $r \in \mathcal{R}$ is embedded at some point $\varphi(r) \in \mathbb{R}^d$. In particular, $\varphi(r)$ must minimize $\langle p, L(\cdot) \rangle$ if and only if r

minimizes $\langle p, \ell(\cdot)\rangle$. 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 $\mathbf{0}$ is in its subgradient set at u. Therefore, we consider collections of sets T_y^r, which intuitively aspire to be the subgradient sets 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 (Rockafellar, 1997).

Definition 13 ($\mathcal{T}$, $D(\mathcal{T})$) *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 $D(\mathcal{T}) = \{p \in \Delta_\mathcal{Y} : \mathbf{0} \in \oplus_p \mathcal{T}\}$.

Note that $T_y = T_y^r$ 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 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, we have $D(\mathcal{T}) = \Gamma_u$, the level set for u of the property Γ elicited by L.

Lemma 14 *Let $L : \mathbb{R}^d \to \mathbb{R}^\mathcal{Y}_{\geq 0}$ 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 $\mathbf{0} \in \partial f(u)$. We thus have $p \in \Gamma_u \iff u \in \arg\min_{u'} \langle p, L(u')\rangle \iff \mathbf{0} \in \partial\langle p, L(u)\rangle = \bigoplus_{y\in\mathcal{Y}} p_y \partial L(u)_y = \oplus_p \mathcal{T} \iff p \in D(\mathcal{T})$. 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 $r \in \mathcal{R}$.

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 $L : \mathbb{R}^d \to \mathbb{R}^\mathcal{Y}_{\geq 0}$ 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

1. Recall that a subgradient of e.g. the convex function $L(\cdot)_y : \mathbb{R}^d \to \mathbb{R}$ 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'.

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 14, 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 15 *Let $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}$ be a discrete loss with, for each $r \in \mathcal{R}$, $\gamma_r = \{p : r \in \arg\min_{r'} \langle p, \ell(r) \rangle\}$. Then ℓ is d-embeddable if and only if there exists a collection of polytopes $\mathcal{T}^r = \{T^r_y : y \in \mathcal{Y}\}$ 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 $\varphi : \mathcal{R} \to \mathbb{R}^d$ and loss functions $\{L_y : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_{\geq 0}\}_{y \in \mathcal{Y}}$ 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 14 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 14 implies that $\Gamma_{\varphi(r)} = D(\mathcal{T}^r) = \gamma_r$. ■

4.3. Characterizing optimality

We now focus entirely on the optimality condition of Theorem 15, 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 15, 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 1 (Halfspace condition) *A collection of polytopes $\mathcal{T}$ and a polytope $C \subseteq \Delta_\mathcal{Y}$ defined by $C = \{p \in \Delta_\mathcal{Y} : Bp \geq \mathbf{0}\}$* 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 2 (Vertex condition) *A collection of polytopes $\mathcal{T}$ and a polytope $C \subseteq \Delta_\mathcal{Y}$ defined by $C = \text{conv}(\{p^1, \ldots, p^l\})$* satisfy the vertex condition *if for all $j \in [l]$, $0 \in \oplus_{p^j} \mathcal{T}$.*

Theorem 16 *Let the polytopes $\mathcal{T} = \{T_y \subseteq \mathbb{R}^d : y \in \mathcal{Y}\}$ and C be given, with $C = \text{conv}(\{p^1, \ldots, p^l\}) = \{p : Bp \geq \mathbf{0}\}$ 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.*

The two conditions above give us a much better understanding of when a given set of polytopes $\mathcal{T}$ satisfies the optimality condition, and the proof of Theorem 16 is shown in Appendix C. 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 $\{x_y^j \in \mathbb{R}^d : j \in [l], y \in \mathcal{Y}\}$ which witness $0 \in \oplus_{p^j} \mathcal{T}$ for the vertex condition, where the vector x_y^j is the y^{th} column of X^j.

Definition 17 (Quadratic Feasibility Program)

Given: *$d \in \mathbb{N}$, a polytope $C = \{p \in \Delta_{\mathcal{Y}} : Bp \geq \mathbf{0}\} = \text{conv}(\{p^1, \ldots, p^l\}) \subseteq \Delta_{\mathcal{Y}}$, 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 $\{x_y^j\}$.*

Constraints:

$$VX^j \leq B \qquad \text{(pointwise, } \forall j \in [l]\text{)} \tag{2}$$

$$\sum_{y=1}^{n} p_y^j x_y^j = \mathbf{0} \qquad (\forall j \in [l]) \tag{3}$$

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 15. 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 18 *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 17) is a feasible.*

Proof By Theorem 16, 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 3). 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 2) 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 19 *Given a discrete loss ℓ eliciting γ, if there is a report $r \in \mathcal{R}$ such that the quadratic program (Definition 17) 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: do there exist a set of rank-d matrices that are pointwise dominated by B, sharing the left factor V, whose right factors X^j respectively satisfy a subspace constraint? We will see in § 5 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 (Finocchiaro et al., 2019, Theorem 2), found by taking 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 (because $\Delta_{\mathcal{Y}}$ is $n - 1$ dimensional to begin with).

5. Example: Abstain loss

5.1. Abstain, $d = 1$

One classification-like problem that is of particular interest is the abstain property, elicited by the loss ℓ_α given in Equation (1). The property $\gamma = \text{abstain}_\alpha$ for $\alpha \in (0, 1)$ can be verified:

$$\text{abstain}_\alpha(p) = \begin{cases} \arg\max_{y \in \mathcal{Y}} p_y & \max_y p_y \geq 1 - \alpha \\ \perp & \text{otherwise} \end{cases}. \tag{4}$$

Ramaswamy et al. (2018) 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 (2016) 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 20 *For $n \geq 3$ and $\alpha < 1$, the abstain loss ℓ_α is not 1-embeddable.*

Proof Consider the intersection graph of $\gamma := \text{abstain}_\alpha$: the node associated with $\gamma_\perp$ 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 2. ■

5.2. Abstain, $\alpha = 1/2$, $d = 2$

We now use our d-dimensional characterization and some observations about the $\text{abstain}_{1/2}$ property to improve lower bounds from those given by Ramaswamy and Agarwal (2016).

Proposition 21 *The quadratic feasibility program (Definition 17) 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 22 *The abstain loss with $\alpha = 1/2$ on $n \geq 5$ outcomes has embedding dimension at least 3.*

We have two proofs, both delegated to the appendix. The first proof is a direct geometric one utilizing the same observations. We show that packing too many normal vectors into a half-circle yields contradictions, implying that a 5-outcome embedding into 2 dimensions is impossible (and incidentally characterizing the structure of all possible 4-outcome embeddings). The second is to directly obtain from mathematical software that the program is infeasible. We use observations about $\ell_{1/2}$ to make a number of simplifications to the program first. This reduces the quadratic program to the problem: *Given a real-valued matrix M in which some entries are missing, but have bounds of the form $[a, b]$, do there exist legal values for these entries such that M is rank d?*

6. Discussion

Essentially the only other known lower-bound technique for dimensionality of calibrated surrogates is the *feasible subspace dimension* of Ramaswamy and Agarwal (2016). This crux of this technique is also an optimality condition on a surrogate loss, showing that if $\mathbf{0}$ 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 $\mathbf{0}$ 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.

Thus, although feasible subspace dimension applies to any convex surrogate, the relation to our techniques is an interesting future direction. The advantage of our approach for generic d is that our optimality conditions and quadratic feasibility program consider the structure of the entire level set γ_r, rather than just a single witness point. This allows us to prove lower bounds on the abstain loss for $\alpha \leq 1/2$, while feasible subspace dimension cannot. Proving the following conjecture would even more closely relate the two techniques (a disproof would also be extremely interesting):

Conjecture 23 *If a discrete loss ℓ has a d-dimensional calibrated convex surrogate, then it is d-embeddable (i.e. has a d-dimensional calibrated polyhedral surrogate). In other words, embedding dimension always equals convex calibration dimension.*

Our results show the conjecture is true when $d = 1$.

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, as stated in Conjecture 23. 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 (2015, Example 6).

Acknowledgments

Jessie Finocchiaro was funded by NSF Graduate Research Fellowship under grant No. DGE 1650115. This material is based upon work supported by the National Science Foundation under Grant No. 1657598. We also thank Peter Bartlett, Nishant Mehta, and Hari Ramaswamy for helpful comments.

References

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.

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.

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.

Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. In *Advances in neural information processing systems*, 2019.

Rafael Frongillo and Ian A. Kash. On Elicitation Complexity. In *Advances in Neural Information Processing Systems 29*, 2015.

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

Nicolas S. Lambert. Elicitation and evaluation of statistical forecasts. 2018. URL https://web.stanford.edu/~nlambert/papers/elicitability.pdf.

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.

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.

Harish G Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *The Journal of Machine Learning Research*, 17(1):397–441, 2016.

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.

R.T. Rockafellar. *Convex analysis*, volume 28 of *Princeton Mathematics Series*. Princeton University Press, 1997.

L.J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, pages 783–801, 1971.

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.

Christophe Weibel. Minkowski sums of polytopes. Technical report, EPFL, 2007.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

Günter M Ziegler. *Lectures on polytopes*, volume 152. Springer Science & Business Media, 2012.

Appendix A. 1-dimensional characterization omitted proofs

Proposition 24 *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*.*

Proof [Proof of Proposition 10] ($\implies$) The intersection graph is constructed by adding an edge for each halfspace, which connects only two nodes. If three level sets intersected on $\text{relint}(\Delta_{\mathcal{Y}})$, then the level boundary for any two would not be a halfspace (this follows because the level sets form a power diagram, e.g. (Lambert and Shoham, 2009)). 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. ■

We will make substantial use of the following general definition.

Definition 25 *A property $\Gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}}$ 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.

As described below Theorem 12, the proof of that statement, as well as Theorem 11, follow from the following two results.

Theorem 26 *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 (2018, Theorem 4), we have positively-oriented normals $v_i \in \mathbb{R}^{\mathcal{Y}}$ for all $i \in \{1, \ldots, k-1\}$ 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 $i \in \{2, \ldots, k-1\}$, 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 $\text{sgn}(\langle v_i, p\rangle)$ 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

$$\begin{aligned}
\langle v(i), p\rangle &= \frac{1}{1+c}\left(v(i)_y + \tfrac{1}{2}(|v(i+1)_y|/v(i+1)_{y'} + |v(i)_y|/v(i)_{y'})v(i)_{y'}\right) \\
&> \frac{1}{1+c}\left(v(i)_y + (|v(i)_y|/v(i)_{y'})v(i)_{y'}\right) = 0 \\
\langle v(i+1), p\rangle &= \frac{1}{1+c}\left(v(i+1)_y + \tfrac{1}{2}(|v(i+1)_y|/v(i+1)_{y'} + |v(i)_y|/v(i)_{y'})v(i)_{y'}\right) \\
&< \frac{1}{1+c}\left(v(i+1)_y + (|v(i+1)_y|/v(i+1)_{y'})v(i+1)_{y'}\right) = 0\ ,
\end{aligned}$$

thus violating the observation that $\operatorname{sgn}(\langle v(i), p\rangle)$ is monotone in i.

Conditions $\implies$ 1-embeddable: (correctness of construction)

First, observe that that $\lambda(i)$ satisfies $\lambda^+(i) \le \lambda(i) \le \lambda^-(i)$, 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 $i \ge 2$. 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 $\partial^-\ell(\varphi(r_i)) \le \partial^+\ell(\varphi(r_i))$ 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 $\partial^-\ell(\varphi(r_1)) = -K\mathbb{1} \le v(2) = \partial^+\ell(\varphi(r_1))$ and $\partial^-\ell(\varphi(r_k)) = \Lambda_{k-1}v(k) \le \Lambda_{k-1}K\mathbb{1} = \partial^+\ell(\varphi(r_k))$.

It remains to show that for all $i \in \{2, \ldots, k-1\}$ and all $y \in \mathcal{Y}$, we have $\Lambda_{i-1}v(i)_y \le \Lambda_i v(i+1)_y$, which by definition of Λ is equivalent to $v(i)_y \le \lambda(i)v(i+1)_y$. By our first condition, the possible pairs $(\operatorname{sgn}(v(i)_y), \operatorname{sgn}(v(i+1)_y))$ 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 $\lambda(i) \le \lambda^-(i) \le v(i)_y/v(i+1)_y$. Recalling that both $v(i)_y$ and $v(i+1)_y$ are negative, we conclude $v(i)_y \le \lambda(i)v(i+1)_y$. In the $(+,+)$ case, we have $\lambda(i) \ge \lambda^+(i) \ge v(i)_y/v(i+1)_y$, and again $v(i)_y \le \lambda(i)v(i+1)_y$.

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 $\langle p, \ell(r)\rangle = \langle p, L(\varphi(r))\rangle \le \langle p, L(\varphi(r'))\rangle = \langle p, \ell(r')\rangle$, 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 $i \in \{2, \dots, k-1\}$. We have

$$\begin{aligned}
0 \in \partial\langle p, L(\varphi(r_i))\rangle &\iff \partial^-\langle p, \ell(\varphi(r_i))\rangle \le 0 \le \partial^+\langle p, \ell(\varphi(r_i))\rangle \\
&\iff \langle p, \partial^-\ell(\varphi(r_i))\rangle \le 0 \le \langle p, \partial^+\ell(\varphi(r_i))\rangle \\
&\iff \langle p, \Lambda_{i-1}v(i)\rangle \le 0 \le \langle p, \Lambda_i v(i+1)\rangle \\
&\iff \langle p, v(i)\rangle \le 0 \le \langle p, v(i+1)\rangle \\
&\iff \langle p, \ell(r_i) - \ell(r_{i-1})\rangle \le 0 \le \langle p, \ell(r_{i+1}) - \ell(r_i)\rangle \\
&\iff \langle p, \ell(r_i)\rangle \le \langle p, \ell(r_{i-1})\rangle \text{ and } \langle p, \ell(r_i)\rangle \le \langle p, \ell(r_{i+1})\rangle \ .
\end{aligned}$$

For $i = 1$, similar reasoning gives that optimality is equivalent to the condition $\langle p, \ell(r_1)\rangle \le \langle p, \ell(r_2)\rangle$, and for $i = k$, $\langle p, \ell(r_k)\rangle \le \langle p, \ell(r_{k-1})\rangle$. (Note that the other conditions, $-K \le 0$ or $0 \le \Lambda_{k-1}K$, are true regardless of p.) In particular, if $p \in \gamma_{r_i}$, then we have $\langle p, \ell(r_i)\rangle \le \langle p, \ell(r_{i-1})\rangle$ for $i \ge 2$, and $\langle p, \ell(r_i)\rangle \le \langle p, \ell(r_{i+1})\rangle$ for $i \le k-1$, 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 25 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 $\mathcal{R} = \{r_1, \dots, r_k\}$ such that $r_i < r_{i+1}$ for all $i \in \{1, \dots, k-1\}$. We now have $\gamma_{r_i} \cap \gamma_{r_{i+1}} = \{p \in \Delta\mathcal{Y} : \langle a(r_{i+1}), p\rangle \le 0 \le \langle b(r_i), p\rangle\}$. 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 $b(r_i) \le a(r_{i+1})$ 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 $\Gamma[L]$ be the unique property elicited by the loss L.

Proposition 27 *If convex $L : \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}$ indirectly elicits a finite elicitable property γ, then γ is orderable.*

Proof Let $\gamma : \Delta\mathcal{Y} \to 2^{\mathcal{R}} \setminus \{\emptyset\}$. From Lemma 28 below, $\Gamma := \Gamma[L]$ is monotone. Let $\psi : \mathbb{R} \to \mathcal{R}$ be the calibrated link from Γ to γ. From Lemma 29, 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 $\overline{P}_r = \{p \in \Delta\mathcal{Y} : \langle a(r), p\rangle \le 0 \le \langle b(r), p\rangle\}$. (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 $p_r \in \operatorname{int}(\gamma)_r$ and picking $u_r \in \Gamma(p_r)$, we order $\mathcal{R} = \{r_1, \dots, r_k\}$ so that $u_{r_i} < u_{r_{i+1}}$ for all $i \in \{1, \dots, k-1\}$. (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 $i \in \{1, \dots, k-1\}$. By monotonicity of Γ, we must have $a(r_i) \le b(r_i) \le a(r_{i+1}) \le b(r_{i+1})$. 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 $i \in \{1, \dots, k-1\}$, γ is orderable. ■

The proof of Proposition 27 uses the following results.

Lemma 28 *For any convex $L : \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}_+$, the property $\Gamma[L]$ 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 $r \in \Gamma[L](p) \iff 0 \in \partial\langle p, L(r)\rangle \iff \langle a(r), p\rangle \leq 0 \leq \langle b(r), p\rangle$, showing the first condition. The second condition follows as the subgradients of L are monotone functions (see e.g. Rockafellar (1997, Theorem 24.1)). ■

Lemma 29 *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 33, γ_r has nonempty interior, so we must have some $p \in \operatorname{int}(\gamma)_r \setminus \overline{P}_r$. 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 32, the level sets of γ form a power diagram, and in particular a cell complex, so we have contradicted point (ii) of Definition 30: the relative interiors of the faces must not be disjoint. Hence, for all $r \in \mathcal{R}$ we have $\gamma_r = \overline{P}_r$. ■

Definition 30 *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 31 *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*

$$\operatorname{cell}(s_i) = \{x \in \mathbb{R}^d : \forall j \in \{1, \ldots, k\} \, \|x - s_i\|^2 - w_i \leq \|x - s_j\| - w_j\} \ . \tag{5}$$

Theorem 32 ((Aurenhammer, 1987)) *A cell complex is affinely equivalent to a convex polyhedron if and only if it is a power diagram.*

Lemma 33 *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 $n - 1$).*

A.1. Example construction of real-valued embedding

For concreteness, let us now construct an embedding via the loss given in Theorem 11. We start with the ordered discrete loss given in Table A.1

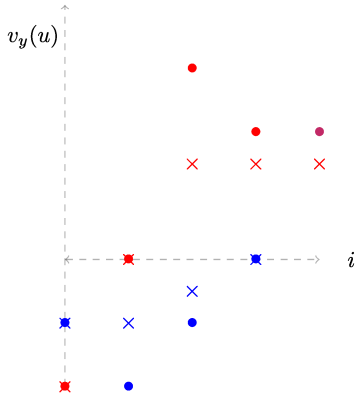

Figure 3: • represents the original v_i, where blue is used for $v(\cdot)_{y_1}$ and red for $v(\cdot)_{y_2}$. The × symbol of the same color is the Λ-corrected directional derivative to force monotonicity.

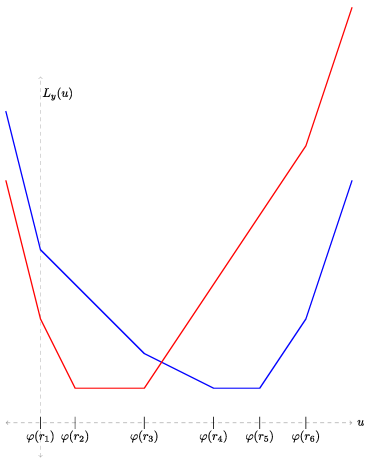

Figure 4: Our constructed embedding L for the discrete loss ℓ given in Table A.1.

	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 1: Ordered discrete loss that we embed.

Given this loss, we can calculate $v(i)_y$ for both losses, shown by the • in Figure 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 4.

Appendix B. Polytope notes

Here, we recall some standard definitions from the theory of Convex Polytopes.

Definition 34 *A* polyhedra *in $\mathbb{R}^d$ is defined by the intersection of a finite number of half-spaces. A* polytope *is a bounded polyhedra.*

Definition 35 (Valid inequality) *Let S be a set in $\mathbb{R}^d$. A* valid inequality *for S is an inequality that holds for all vectors in S. That is, the pair (a, β) is a valid inequality for S if and only if*

$$\langle a, x \rangle \leq \beta \quad \forall x \in S \ .$$

Definition 36 (Face) *For any valid inequality of a polytope, the subset of the polytope of vectors which are tight for the inequality is called a* face *of the polytope. That is, the set F is a face of the polytope T if and only if*

$$F = \{x \in T : \langle a, x\rangle = \beta\}$$

for some valid inequality (a, β) of T.

Definition 37 (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 38 (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 39 ((Weibel, 2007, 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 40 ((Weibel, 2007, Theorem 3.1.6)) *The supporting function of a Minkowski sum is the sum of the supporting functions of its summands.*

Weibel (2007) 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* $E = (e^y)_{y=1}^n$.

Definition 41 *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 $\mathcal{T} := \{T_1, \ldots, T_n\}$, 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 $\mathbf{0} \in \oplus_p \mathcal{T}$ by substituting $z = \mathbf{0}$. 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 40 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 $\mathbf{0} \in \oplus_p \mathcal{T}$ if and only if $Ep \geq 0$ by substitution.

We assume $p \in \Delta_{\mathcal{Y}}$, so we now describe the cell $D(\mathcal{T}) := \{p \in \Delta_{\mathcal{Y}} : Ep \geq \mathbf{0}\}$ as the set of distributions such that $\mathbf{0} \in \oplus_p \mathcal{T}$. We will see in Lemma 42 that this definition is equivalent to the definition of $D(\mathcal{T})$ in Definition 13.

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 $\mathbf{0} \in T(p)$ for all $p \in C$? We know that if we are given $\mathcal{T}$ and a complete set of normals V, we can describe $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \mathbf{0}\}$.

Lemma 42 *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 $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \mathbf{0}\}$.*

Then $\{p \in \Delta_{\mathcal{Y}} : \mathbf{0} \in \oplus_p \mathcal{T}\} = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \mathbf{0}\}$.

Proof First, let us fix a distribution $p \in \Delta_{\mathcal{Y}}$. By Theorem 40, 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 $\mathbf{0} \in \oplus_p \mathcal{T} \iff Ep \geq \mathbf{0}$. 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 $\mathbf{0}$ is in the Minkowski sum in terms of the minimal rank matrix describing the cell.

Proposition 43 *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 $D(\mathcal{T}) =$*

$\{p \in \Delta_{\mathcal{Y}} : Ep \geq \mathbf{0}\}$ *and take the minimal rank* $B \in \mathbb{R}^{k \times n}$ *such that we have the given cell* $C = \{p \in \Delta_{\mathcal{Y}} : Bp \geq \mathbf{0}\}$.

Then $\{p \in \Delta_{\mathcal{Y}} : \mathbf{0} \in \oplus_p \mathcal{T}\} = C$ *if and only if* $C = D(\mathcal{T})$.

Proof By Lemma 42, we have $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : \mathbf{0} \in \oplus_p \mathcal{T}\}$, and the result follows. ■

Definition 44 *We say a vector* v *is* redundant *with respect to matrix* Y *if we have* $\{z : Yz \geq \boldsymbol{b}\} = \{z : [Y; v]z \geq \boldsymbol{b}^*\}$, *where* $b^* = [b; c]$ *for some constant* $c \in \mathbb{R}$.

Proposition 45 *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* $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \mathbf{0}\}$ *and take the minimal matrix* B *such that a given cell* $C = \{p \in \Delta_{\mathcal{Y}} : Bp \geq \mathbf{0}\}$.

Then $\{p \in \Delta_{\mathcal{Y}} : \mathbf{0} \in \oplus_p \mathcal{T}\} = C$ *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 $C = \{p \in \Delta_{\mathcal{Y}} : \mathbf{0} \in \oplus_y p_y T_y\}$. By Proposition 43, we know that $C = D^{\mathcal{T}} := \{p \in \Delta_{\mathcal{Y}} : Ep \geq \mathbf{0}\}$. Then we have $\{p \in \Delta_{\mathcal{Y}} : Bp \geq \mathbf{0}\} = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \mathbf{0}\}$. 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 $D(\mathcal{T}) = \{p \in \Delta_{\mathcal{Y}} : Ep \geq \mathbf{0}\} = \{p \in \Delta_{\mathcal{Y}} : Bp \geq \mathbf{0}\} = C$.

Then $D(\mathcal{T}) = C$, and by Proposition 43, we have $C = \{p \in \Delta_{\mathcal{Y}} : \mathbf{0} \in \oplus_p \mathcal{T}\}$. ■

Appendix C. d-dimensional omitted proofs

C.1. Details for the proof

The following condition formalizes a necessary condition on $\mathcal{T}$ in terms of the halfspace representation of C; the subsequent one formalizes a sufficient condition using the vertex representation.

Proof [Proof of Theorem 16] ($\implies$) 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 46, there is a finite collection of vectors $w_1, \ldots, w_K \in \mathbb{R}^d$ and such that $\mathbf{0} \in \oplus_p \mathcal{T}$ 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 48 , 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 1 and 2 hold. Then by the vertex condition, $p^j \in D(\mathcal{T})$ for all $j \in [\ell]$. Because $D(\mathcal{T})$ is convex (Lemma 47), 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 46 *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 47 *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 (Weibel, 2007, 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 46, 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 (Weibel, 2007, 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 $x = \mathbf{0}$, we see $\mathbf{0} \in \oplus_p \mathcal{T} \iff E^* p \geq \mathbf{0} \iff p \in D(\mathcal{T})$ by Lemma 42, which defines a polytope by construction of E^*. ■

Lemma 48 *Let $C = \{p : Bp \geq \mathbf{0}\}$ where B has the minimum possible number of rows to capture C, and suppose $C = \{p : Ep \geq \mathbf{0}\}$. Then for each row in B there is some (unique) row in E that is equal to αB for some positive α.*

Proof Ziegler (2012, 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 $\mathbf{0}$) 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 $x \in C = \{x : x : Bx \geq 0\}$. ■

Appendix D. Simplifying the QFP for $\text{abstain}_{1/2}$, $d = 2$

In order to prove Proposition 21, 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 49 *For the abstain loss $\ell_{1/2}$, the level set of abstain satisfies $\gamma_\perp = \text{conv}\{(\delta_y + \delta_{y'})/2 : y, y' \in \mathcal{Y}, y < y'\} = \{p : Bp \geq \mathbf{0}\}$ where δ_y puts probability one on y and $B = \mathbb{1}\mathbb{1}^\intercal - 2I \in \mathbb{R}^{5\times 5}$, 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 $\sum_{y\in\mathcal{Y}} p_y^j (A^{-1}X^j)_y = \mathbf{0}$. If we let a_m be the m^{th} row of A^{-1}, then the m^{th} row of the vector equality is

$$\begin{aligned} 0 &= \sum_{y\in\mathcal{Y}} p_y^j \langle a_m, x_y^j\rangle \\ &= \langle a_m, \sum_{y\in\mathcal{Y}} p_y^j x_y^j\rangle \\ &= 0 \end{aligned}$$

so the program is feasible. ■

Corollary 50 *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 $(1, 0, \ldots, 0)$. ■

2. Note one can show $V = 0$ is not feasible unless B is a trivial property, i.e. essentially has no rows at all.

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 51 *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 $\frac{1}{2}x_i^{ij} + \frac{1}{2}x_j^{ij} = \mathbf{0}$. ■

Lemma 52 *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 51. ■

Lemma 53 *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 51 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. ■

Lemma 54 *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. By Lemma 53, 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 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 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 $\langle v_3, x\rangle = 1$;

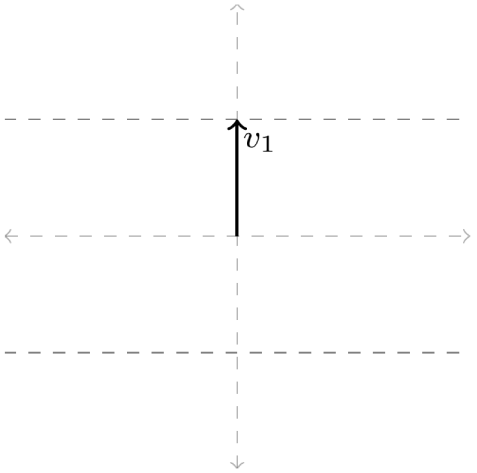

Figure 5: Example of v_1.

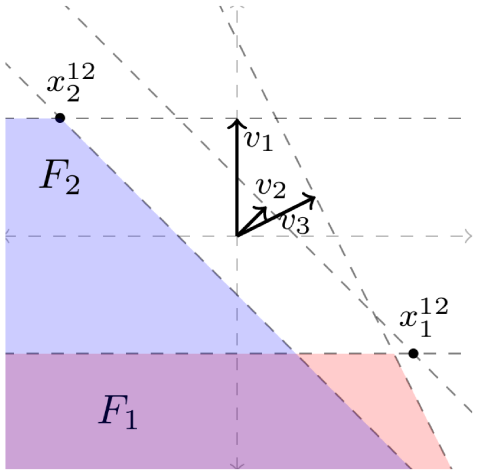

Figure 6: If $x_1^{12} \neq x_1^{13}$, we have a contradiction.

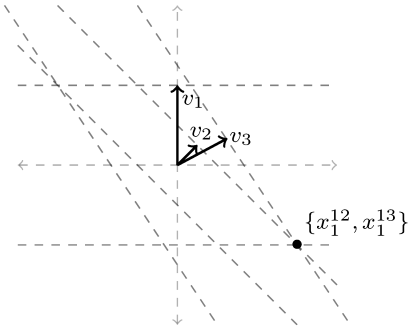

Figure 7: Reducing to the case where $x_1^{12} = x_1^{13}$.

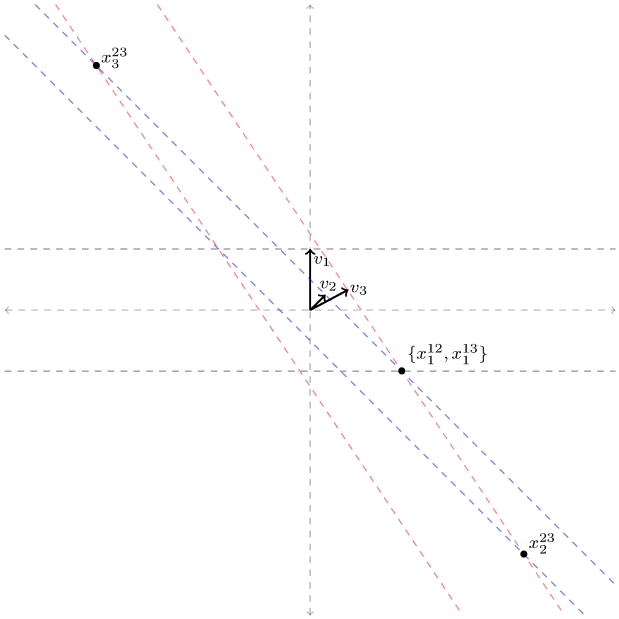

Figure 8: The intersection of $\langle v_2, x\rangle = -1$ and $\langle v_3, x\rangle = 1$ occurs outside the region $\langle v_1, x\rangle \in [-1, 1]$.

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 8. This proves there is no feasible solution with three normals lying strictly in the same halfspace through the origin. ■

Proposition 55 *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 54. So the clockwise angle between v_1 and v_3 is at least 180°. Since there are no duplicate angles (Lemma 52), this implies that the clockwise angle between v_4 and v_1, which includes v_5, is strictly less than 180°. This contradicts Lemma 54. ■

Proof [Matrix rank proof of Proposition 21] We can formulate the QFP given in Section 4 for $\text{abstain}_{1/2}$ as a matrix rank problem. First, as observed in Lemma 51, the vertex constraints are exactly equivalent to requiring that, in our notation, $x_i^{ij} = -x_j^{ij}$ for all $i < j$. Therefore, we can substitute in for all variables of the form x_j^{ij} and eliminate them from the problem. Second, we can observe that the program is only easier to satisfy if one drops the halfspace constraints on all variables of the form x_m^{ij}, $i < j$, $m \notin \{i, j\}$. This allows us to drop all such variables, and we are now left with only the variables $v_1, \ldots, v_k$ and $\{x_i^{ij} : i < j\}$, with the understanding that $x_j^{ij} = -x_i^{ij}$. We can therefore simplify the following constraints of the QFP:

$$\begin{aligned}
&\langle v_i, x_i^{ij}\rangle \leq B_{ii} = -1\\
&\langle v_i, x_j^{ij}\rangle \leq B_{ij} = 1\\
\implies &\langle v_i, x_i^{ij}\rangle = -1\\
&\langle v_i, x_i^{ji}\rangle \leq B_{ii} = -1\\
&\langle v_i, x_j^{ji}\rangle \leq B_{ij} = 1\\
\implies &\langle v_i, x_j^{ji}\rangle = 1\\
&\langle v_i, x_j^{jm}\rangle \leq B_{ij} = 1\\
&\langle v_i, x_m^{jm}\rangle \leq B_{im} = 1\\
\implies &\langle v_i, x_j^{jm}\rangle \in [-1, 1].
\end{aligned}$$

This gives us the following simplified feasibility problem.

$$\begin{aligned}
\langle v_i, x_i^{ij}\rangle &= -1 && (\forall i < j)\\
\langle v_i, x_j^{ji}\rangle &= 1 && (\forall i < j)\\
\langle v_i, x_j^{jm}\rangle &\in [-1, 1] && (\forall j < m, i \notin \{j, m\}).
\end{aligned}$$

If we consider the matrix V and construct Y whose columns are $\{x_i^{ij} : i < j\}$, this problem asks us to find such a V and Y whose product $M = VY$ is a matrix with certain fixed entries and others bounded.

In particular, with $n = 4$, we obtain the following matrix rank problem: does there exist

$$M_4 = \begin{bmatrix} -1 & -1 & -1 & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & -1 & -1 & \cdot \\ \cdot & 1 & \cdot & 1 & \cdot & -1 \\ \cdot & \cdot & 1 & \cdot & 1 & 1 \end{bmatrix}$$

where each unknown entry $\cdot$ is in $[-1, 1]$, of rank $d = 2$?

Since M_4 is a submatrix of the matrix obtained when $n = 5$, any rank-2 solution for the large matrix given below must also solve the above problem.

$$M_5 = \begin{bmatrix} -1 & -1 & -1 & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & -1 & -1 & \cdot & \cdot & -1 & \cdot & \cdot \\ \cdot & 1 & \cdot & 1 & \cdot & -1 & \cdot & \cdot & -1 & \cdot \\ \cdot & \cdot & 1 & \cdot & 1 & 1 & \cdot & \cdot & \cdot & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 \end{bmatrix}$$

We have verified with Mathematica that M_4 has rank 2 if and only if it is in the following form, up to a couple symmetries.

$$\begin{bmatrix} -1 & -1 & -1 & 1 & 1 & a \\ 1 & 1 & 1 & -1 & -1 & -a \\ b & 1 & -1 & 1 & -1 & -1 \\ -b & -1 & 1 & -1 & 1 & 1 \end{bmatrix}$$

with $a, b \in [-1, 1]$. Plugging in this solution set into the larger matrix yields

$$M_5' = \begin{bmatrix} -1 & -1 & -1 & 1 & 1 & a & -1 & \cdot & \cdot & \cdot \\ 1 & 1 & 1 & -1 & -1 & a & \cdot & -1 & \cdot & \cdot \\ b & 1 & -1 & 1 & -1 & -1 & \cdot & \cdot & -1 & \cdot \\ -b & -1 & 1 & -1 & 1 & 1 & \cdot & \cdot & \cdot & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 \end{bmatrix}$$

In particular, the submatrix of the last four columns alone cannot be completed.

■