Journal of Machine Learning Research 25 (2024) 1-60 Submitted 6/22; Revised 1/24; Published 2/24

An Embedding Framework for the Design and Analysis of Consistent Polyhedral Surrogates

Jessie Finocchiaro* JEFI8453@COLORADO.EDU
Department of Computer Science
Boston College
Chestnut Hill, MA, USA

Rafael M. Frongillo RAF@COLORADO.EDU
Department of Computer Science
University of Colorado Boulder
Boulder, CO, USA

Bo Waggoner BWAG@COLORADO.EDU
Department of Computer Science
University of Colorado Boulder
Boulder, CO, USA

Editor: Tong Zhang

Abstract

We formalize and study the natural approach of designing convex surrogate loss functions via embeddings, for discrete problems such as classification, ranking, or structured prediction. In this approach, one embeds each of the finitely many predictions (e.g. rankings) as a point in $\mathbb{R}^d$, assigns the original loss values to these points, and "convexifies" the loss in some way to obtain a surrogate. We establish a strong connection between this approach and polyhedral (piecewise-linear convex) surrogate losses: every discrete loss is embedded by some polyhedral loss, and every polyhedral loss embeds some discrete loss. Moreover, an embedding gives rise to a consistent link function as well as linear surrogate regret bounds. Our results are constructive, as we illustrate with several examples. In particular, our framework gives succinct proofs of consistency or inconsistency for existing polyhedral surrogates, and for inconsistent surrogates, it further reveals the discrete losses for which these surrogates are consistent. We go on to show additional structure of embeddings, such as the equivalence of embedding and matching Bayes risks, and the equivalence of various notions of non-redudancy. Using these results, we establish that indirect elicitation, a necessary condition for consistency, is also sufficient when working with polyhedral surrogates.

Keywords: Statistical consistency, surrogate loss functions, calibration, property elicitation

1. Introduction

In supervised learning, one tries to learn a hypothesis which fits labeled data as judged by a target loss function. Minimizing the target loss directly is typically computationally intractable for discrete prediction tasks like classification, ranking, and structured prediction. Instead, one typically minimizes a surrogate loss which is convex and therefore efficiently

∗. Work done while a student at the University of Colorado Boulder

minimized. After learning a surrogate hypothesis, a link function then translates back to the target problem. To ensure the surrogate and link properly correspond to the target problem, the surrogate must be *consistent*, meaning that minimizing the surrogate loss over enough data will also minimize the target loss.

While a growing body of work seeks to design and analyze consistent convex surrogates for particular target loss functions, to date much of this work has been ad-hoc. We lack general tools to systematically construct consistent convex surrogates, much less an understanding of the full class of consistent surrogates. For example, in multiclass and top-k classification, several proposed surrogates were proposed and adopted but later proved to be inconsistent (Yang and Koyejo, 2020; Crammer and Singer, 2001; Rifkin and Klautau, 2004). This state of affairs is even more dire for structured prediction, where in addition to convexity and consistency, one often requires a low *prediction dimension* (the dimension of the surrogate prediction space) as the label set can grow exponentially large. Clever constructions like the binary-encoded predictions (BEP) surrogate for multiclass classification with an abstain option (Ramaswamy et al., 2018), which achieves logarithmic prediction dimension, are the exception rather than the rule. In all of these settings, we lack a unifying framework that moves from a given target problem to a convex consistent surrogate and link function.

To address this shortcoming, we introduce a new framework motivated by a particularly natural approach for finding convex surrogates, wherein one "embeds" a discrete loss. Specifically, we say a convex surrogate L embeds a discrete target loss ℓ if there is an injective function, which we call an *embedding*, from the target reports (predictions) to $\mathbb{R}^d$ such that (i) the surrogate loss values match the target at the embedded reports, and (ii) a target report is ℓ-optimal if and only if its embedded report is L-optimal. (See § 2.4.) Common examples of this general construction include hinge loss as a surrogate for 0-1 loss and the BEP surrogate mentioned above (Ramaswamy et al., 2018).

Using this framework, we give several constructive results to design new consistent surrogates, as well as a suite of tools to analyze existing surrogates. As a first step, in § 3, we show that such an embedding scheme is intimately related to the class of *polyhedral* loss functions, i.e., those that are piecewise-linear and convex.

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

Crucially, we go on in § 4 to show that an embedding gives rise to a calibrated link function, and is therefore consistent with respect to the target loss.

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 in § 4.3 that any polyhedral surrogate achieves a linear surrogate regret bound, which allows one to translate generalization bounds from the surrogate to the target. Our results are constructive: given a discrete target loss, we show how to define a surrogate and construct a calibrated link, and given a polyhedral surrogate, we show how to find a discrete loss that it embeds.

We demonstrate the constructiveness of our framework in § 5 with several applications, many of which are subsequent to our work. In addition to constructing new surrogates, we

illustrate the power of our framework to analyze previously proposed polyhedral surrogates. For example, while we know that the inconsistent top-k polyhedral surrogates mentioned above are not consistent for top-k classification, our framework illuminates the problems for which they *are* consistent; it also yields restrictions on the underlying distribution which would render these surrogates top-k consistent (§ 5.5).

Underpinning our results are several observations, outlined in § 6, which formalize the idea that polyhedral losses "behave like" discrete losses. In particular, any polyhedral loss L has a finite *representative set* $\mathcal{S}$ of reports, such that for all distributions, some report in $\mathcal{S}$ is L-optimal. We show that L embeds the discrete loss $\ell = L|_{\mathcal{S}}$ given by restricting L to just the reports in $\mathcal{S}$. To go from a discrete loss to a polyhedral surrogate, we prove that the conditions of an embedding are equivalent to matching Bayes risks (Proposition 22), and use the fact that discrete losses and polyhedral losses both have polyhedral Bayes risks.

Finally, we also provide several observations beyond what is needed to prove our main results, which we view as conceptual contributions (§ 6, 7). Using tools from property elicitation, we show an equivalence between minimum representative sets (those of minimum cardinality) and "non-redundancy", wherein no report is dominated by another. We further show that, while a minimum representative set is not always unique, the loss values associated with it are unique, giving rise to a natural "trim" operation on losses. The paper concludes with an interesting observation (Theorem 32): while indirect property elicitation is generally a strictly weaker condition than consistency, the two are equivalent when restricting to the class of polyhedral surrogates.[1]

Taken together, we view our contributions as 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. We conclude with several directions for future work.

2. Setting

For discrete prediction problems like classification, the given (discrete) loss is often computationally intractable to optimize directly. Therefore, many machine learning algorithms instead minimize a surrogate loss function with better optimization qualities, such as convexity. To ensure that this surrogate loss successfully addresses the original target problem, one needs to establish statistical consistency, a minimal requirement that is a prerequisite for generalization bounds. Consistency roughly says that, in the limit as one obtains more and more data, the learned hypothesis approaches the best possible. Consistency also depends crucially on the choice of link function that maps surrogate reports (predictions) to target reports; see the discussion following Definition 8.

In this section, we introduce notation and concepts related to consistency that we use throughout. Consistency is often a difficult condition to work with directly, but in finite prediction settings, it is equivalent the simpler notion of *calibration* (Definition 6) which depends solely on the conditional distribution over the labels (Bartlett et al., 2006; Tewari

1. This result is also implicit in Ramaswamy and Agarwal (2016, Theorem 8); see § E.

and Bartlett, 2007; Ramaswamy and Agarwal, 2016). Even simpler than calibration is indirect elicitation, a weaker condition only requiring that optimal surrogate reports link to optimal target reports. Finally, we introduce a new precise notion of embedding, a special case of indirect elicitation, which forms the backbone of our approach.

2.1 Notation and Losses

Let $\mathcal{Y}$ be a finite label space, and throughout let $n = |\mathcal{Y}|$. Define $\mathbb{R}^{\mathcal{Y}}_+$ to be the nonnegative orthant in $\mathbb{R}^{\mathcal{Y}}$, i.e., $\mathbb{R}^{\mathcal{Y}}_+ = \{x \in \mathbb{R}^{\mathcal{Y}} : \forall y \in \mathcal{Y}\ x_y \geq 0\}$. Let $\Delta_{\mathcal{Y}} = \{p \in \mathbb{R}^{\mathcal{Y}}_+ : \|p\|_1 = 1\}$ be the set of probability distributions on $\mathcal{Y}$, represented as vectors. We will primarily focus on conditional distributions $p \in \Delta_{\mathcal{Y}}$ over labels, abstracting away the feature space $\mathcal{X}$; see § 2.3 for a discussion of the joint distribution over $\mathcal{X} \times \mathcal{Y}$.

A generic loss function, denoted $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, maps a report (prediction) r from a set $\mathcal{R}$ to the vector of loss values $L(r) = (L(r)_y)_{y \in \mathcal{Y}}$ for each possible outcome $y \in \mathcal{Y}$. We write the corresponding expected loss when $Y \sim p$ as $\langle p, L(r)\rangle$. The *Bayes risk* of a loss $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ is the function $\underline{L} : \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ given by $\underline{L}(p) := \inf_{r \in \mathcal{R}} \langle p, L(r)\rangle$. When restricting the domain of a loss L from $\mathcal{R}$ to $\mathcal{R}' \subseteq \mathcal{R}$, we write $L|_{\mathcal{R}'}$.

We assume that the target prediction problem is given in the form of a *target loss* $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ for some report set $\mathcal{R}$. A *discrete (target) loss* is such an ℓ where $\mathcal{R}$ is a finite set. Surrogate losses will take $\mathcal{R} = \mathbb{R}^d$ and be written $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, typically with reports written $u \in \mathbb{R}^d$.

For example, in binary classification, 0-1 loss is a discrete loss with $\mathcal{R} = \mathcal{Y} = \{-1, 1\}$ given by $\ell_{0\text{-}1}(r)_y = \mathbb{1}\{r \neq y\}$, with Bayes risk $\underline{\ell_{0\text{-}1}}(p) = 1 - \max_{y \in \mathcal{Y}} p_y$. Two widely-used surrogates for $\ell_{0\text{-}1}$ are hinge loss $L_{\text{hinge}}(u)_y = (1 - yu)_+$, where $(x)_+ = \max(x, 0)$, and logistic loss $L(u)_y = \log(1 + \exp(-yu))$ for $u \in \mathbb{R}$. See Figure 1 for a visualization of the Bayes risks of 0-1, hinge, and logistic losses, respectively.

Many of the surrogate losses we consider will be *polyhedral*, meaning piecewise linear and convex; we 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 (Rockafellar, 1997).

Definition 3 (Polyhedral loss) *A loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ is* polyhedral *if $L(u)_y$ is a polyhedral function of u for each $y \in \mathcal{Y}$.*

In the example above, hinge loss is polyhedral, whereas logistic loss is not.

2.2 Property Elicitation

We will frequently appeal to concepts and results from property elicitation (Savage, 1971; Osband and Reichelstein, 1985; Lambert et al., 2008; Gneiting, 2011; Steinwart et al., 2014; Frongillo and Kash, 2015a; Lambert, 2018). Here one studies *properties*, maps from (conditional) label distributions to reports, and asks when a property characterizes the reports that exactly minimize a loss. In our case, this map will at times be set-valued, meaning a single distribution could yield multiple optimal reports. For example, when $p = (1/2, 1/2)$, both $r = 1$ and $r = -1$ optimize 0-1 loss with $\langle p, L(1)\rangle = 1/2 = \langle p, L(-1)\rangle$. We will use

double arrow notation to denote a (non-empty) set-valued map, so that $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ is shorthand for $\Gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}} \setminus \{\emptyset\}$, where $2^{\mathcal{R}}$ denotes the power set of $\mathcal{R}$.

Definition 4 (Property, level set) *A* property *is a function $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$. The* level set *of Γ for report r is the set $\Gamma_r := \{p \in \Delta_{\mathcal{Y}} \mid r \in \Gamma(p)\}$. If $\mathcal{R}$ is finite, we call Γ a* finite property*.*

Intuitively, $\Gamma(p)$ is the set of reports which optimize expected loss under a given distribution p, and Γ_r is the set of distributions for which the report r optimizes the expected loss. For example, the *mode* is the property $\text{mode}(p) = \arg\max_{y \in \mathcal{Y}} p_y$, and captures the set of optimal reports for 0-1 loss: for each distribution over the labels, one should report the most likely label. In this case we say 0-1 loss (directly) *elicits* the mode, as we formalize below. literature (Savage, 1971; Osband and Reichelstein, 1985; Lambert et al., 2008), in which a report r is elicited from some forecaster by scoring her with a loss on the observed outcome y.

Definition 5 (Directly Elicits) *A loss $L : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$,* (directly) elicits *a property $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ if*

$$\forall p \in \Delta_{\mathcal{Y}}, \quad \Gamma(p) = \arg\min_{r \in \mathcal{R}} \langle p, L(r) \rangle \ . \tag{1}$$

If L elicits a property, that property is unique and we denote it $\text{prop}[L]$*.*

Since we have defined a property Γ to be nonempty, if the infimum of expected loss $\langle p, L(\cdot) \rangle$ is not attained for some $p \in \Delta_{\mathcal{Y}}$, then L does not elicit a property. We say that a loss L is *minimizable* if the infimum of $\langle p, L(\cdot) \rangle$ is attained for all $p \in \Delta_{\mathcal{Y}}$. For example, hinge loss is minimizable, whereas logistic loss is not (take $p = (0, 1)$ or $(1, 0)$).

We will typically denote general properties and losses with Γ and L, respectively. For surrogate losses and properties, recall that we typically consider the report set $\mathbb{R}^d$. For discrete target losses and properties, we will take $\mathcal{R}$ to be a finite set, and use lowercase notation γ and ℓ, respectively.

2.3 Calibration and Links

To assess whether a surrogate and link function align with the original loss, we turn to the common condition of *calibration*. Roughly, a surrogate and link are calibrated if the best possible expected loss achieved by linking to an incorrect report is strictly suboptimal.

Definition 6 *Let discrete loss $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$, proposed surrogate $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$, and function $\psi : \mathbb{R}^d \to \mathcal{R}$ be given. In this context, ψ is called a* link function*. We say (L, ψ) is* calibrated *with respect to ℓ if for all $p \in \Delta_{\mathcal{Y}}$,*

$$\inf_{u \in \mathbb{R}^d : \psi(u) \notin \text{prop}[\ell](p)} \langle p, L(u) \rangle > \inf_{u \in \mathbb{R}^d} \langle p, L(u) \rangle \ . \tag{2}$$

If (L, ψ) is calibrated with respect to ℓ, we call ψ a calibrated link*.*

It is well-known in finite-outcome settings that calibration is equivalent to *consistency*, as defined next (cf. Bartlett et al. (2006); Zhang (2004); Agarwal and Agarwal (2015)). Let feature space $\mathcal{X}$ and label space $\mathcal{Y}$ be given. Intuitively, a surrogate and link pair (L, ψ) is consistent with respect to ℓ if, for all data distributions $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, and all sequences of surrogate hypotheses $h_1, h_2, \ldots$ whose L-loss limits to the optimal surrogate loss L^* (in expectation over D), the ℓ-loss of the sequence the linked hypotheses limits to the optimal target loss ℓ^*.

Definition 7 *Given a loss $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ and link $\psi : \mathbb{R}^d \to \mathcal{R}$, the pair (L, ψ) is* consistent *with respect to a target loss $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$ if, for all $\mathcal{D} \in \Delta(X \times \mathcal{Y})$ and all sequences of measurable functions $\{h_m : \mathcal{X} \to \mathbb{R}\}$, we have*

$$\begin{aligned} &\mathbb{E}_{(X,Y)\sim\mathcal{D}} L(h_m(X), Y) \to \inf_h \mathbb{E}_{(X,Y)\sim\mathcal{D}} L(h(X), Y) \\ \implies \quad &\mathbb{E}_{(X,Y)\sim\mathcal{D}} \ell(\psi \circ h_m(X), Y) \to \inf_f \mathbb{E}_{(X,Y)\sim\mathcal{D}} \ell(\psi \circ h(X), Y) \ . \end{aligned}$$

When working with a restricted hypothesis class $\mathcal{H}$, as opposed to the set of all measurable functions in Definition 7, then the corresponding notion of consistency (called $\mathcal{H}$*-consistency*) is no longer equivalent to calibration (Long and Servedio, 2013; Zhang et al., 2020; Kuznetsov et al., 2014); see § 8 for further discussion.

Like the use of a surrogate and link pair in the calibration definition, one can also extend the earlier definition of property elicitation to *indirect (property) elicitation*, in which one applies a link to an elicited property to obtain a related property of interest.

Definition 8 *Let minimizable loss $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ and link $\psi : \mathbb{R}^d \to \mathcal{R}$ be given. The pair (L, ψ) indirectly elicits a property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ if for all $u \in \mathbb{R}^d$, we have $\Gamma_u \subseteq \gamma_{\psi(u)}$, where $\Gamma = \text{prop}[L]$. Moreover, we say L indirectly elicits γ if such a ψ exists, i.e., if for all $u \in \mathbb{R}^d$ there exists $r \in \mathcal{R}$ such that $\Gamma_u \subseteq \gamma_r$.*[2]

Indirect elicitation is a weaker condition than calibration (Steinwart and Christmann, 2008; Agarwal and Agarwal, 2015; Finocchiaro et al., 2021); we briefly sketch the argument. Suppose L is minimizable and (L, ψ) is calibrated with respect to ℓ, and set $\Gamma = \text{prop}[L]$ and $\gamma = \text{prop}[\ell]$. Let $u \in \mathbb{R}^d$ and $p \in \Gamma_u$. From eq. (2), if $\psi(u) \notin \gamma(p)$, then we would have $u \notin \Gamma(p)$, a contradiction to $p \in \Gamma_u$. Thus, $\psi(u) \in \gamma(p)$, so $p \in \gamma_{\psi(u)}$. In fact, indirect elicitation is strictly weaker; take hinge loss with the link $\psi(u) = -1$ for $u < 1$ and $\psi(u) = 1$ for $u \geq 1$. While this pair indirectly elicits the mode, we can show it is not calibrated with respect to 0-1 loss. Suppose $p = (0, 1) \in \Delta_{\mathcal{Y}}$ is the distribution putting all weight on $y = 1$, and consider any sequence $u_m \to 1$ with $u_m < 1$ for all m. Then the loss approaches the Bayes optimal, $L(1)_1 = 0$, but $\psi(u_n) = -1$ for all n, violating calibration. Agarwal and Agarwal (2015) were the first to formally connect property elicitation to calibration.

2.4 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 $\mathbb{R}^d$

2. The elicitation literature often refers to this latter condition as one property "refining" another (Frongillo and Kash, 2015b).

that agrees with ℓ at these points. This approach captures several surrogates proposed in the literature (e.g., Ramaswamy et al. (2015); Ramaswamy and Agarwal (2016); Lapin et al. (2015); Wang and Scott (2020); see § 5).

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 § 6; see also Wang and Scott (2020).) 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 (conditional) label distributions, at least one report $r \in \mathcal{S}$ minimizes the conditional expected loss.

Definition 9 (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* $\text{prop}[L]$.

Wang and Scott (2020) first study the notion of minimum representative sets under the name *embedding cardinality*.

We now define an embedding, which is a special case of indirect property elicitation. (This fact is non-trivial; see Lemma 50.) In addition to matching loss values, as described above, we require the original reports to be ℓ-optimal exactly when the corresponding embedded points are L-optimal. As we discuss following Proposition 12, this latter condition can be more simply stated: the representative set for the target must embed into a representative set for the surrogate.

Definition 10 (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*

$$r \in \text{prop}[\ell](p) \iff \varphi(r) \in \text{prop}[L](p) \ . \tag{3}$$

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

To illustrate the idea of embedding, let us closely examine hinge loss as a surrogate for 0-1 loss in binary classification. Recall that we have $\mathcal{R} = \mathcal{Y} = \{-1, +1\}$, with $L_{\text{hinge}}(u)_y = (1 - uy)_+$ and $\ell_{0\text{-}1}(r)_y := \mathbb{1}\{r \neq y\}$, typically with link function $\psi(u) = \text{sgn}(u)$, where $\text{sgn}(0) = 1$ without loss of generality. We will see that hinge loss embeds (2 times) 0-1 loss, via the identity embedding $\varphi(r) = r$. For condition (i), it is straightforward to check that $L_{\text{hinge}}(\varphi(r))_y = L_{\text{hinge}}(r)_y = 2\mathbb{1}\{r \neq y\} = 2\ell_{0\text{-}1}(r)_y$ 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 $p \in \Delta_\mathcal{Y}$:

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

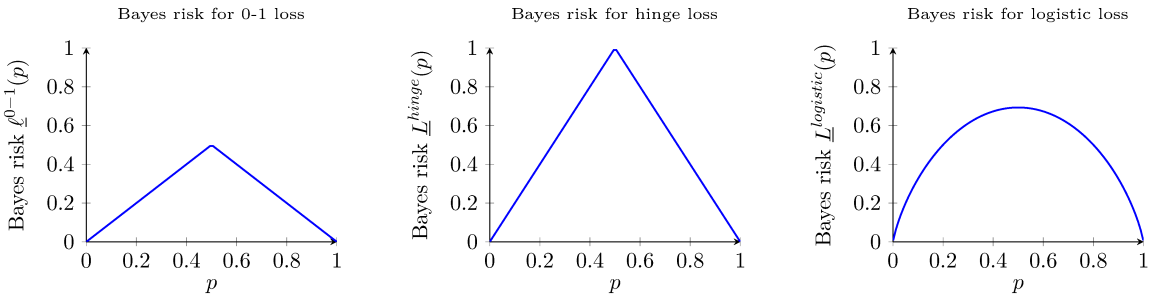


Figure 1: Bayes risks $\underline{L} : p \mapsto \inf_u \langle p, L(u) \rangle$ of 0-1, hinge, and logistic losses, respectively, plotted as a function of $p_1 = \Pr[Y = 1]$. Observe that the Bayes risks of 0-1 and hinge loss are both piecewise linear and concave, while the Bayes risk of logistic loss is not piecewise linear. Proposition 22 states that embedding is equivalent to matching Bayes risks, confirming that hinge loss (M) embeds twice 0-1 loss (L), but logistic loss (R) does not.

In particular, we see that $-1 \in \text{prop}[\ell_{0\text{-}1}](p) \iff p_1 \in [0, 1/2] \iff -1 \in \text{prop}[L_{\text{hinge}}](p)$, and $1 \in \text{prop}[\ell_{0\text{-}1}](p) \iff p_1 \in [1/2, 1] \iff 1 \in \text{prop}[L_{\text{hinge}}](p)$. With both conditions of Definition 10 satisfied, we can conclude that L_{hinge} embeds $2\ell_{0\text{-}1}$ with the representative set $\mathcal{S} = \{-1, 1\}$. By results in § 6.2, one could also show that L_{hinge} embeds $2\ell_{0\text{-}1}$ by the fact that their Bayes risks match (Figure 1).

In this particular example, it is known $(L_{\text{hinge}}, \text{sgn})$ is calibrated with respect to 0-1 loss (Bartlett et al., 2006, Example 4). Beyond this simple case, however, it is not clear whether an embedding will always yield a calibrated link. Indeed, while it is intuitively clear that embedded points should link back to their original reports, via $\psi(\varphi(r)) = r$, how to map the remaining values is far from obvious. Using the strong connection between embeddings and polyhedral surrogates in § 3, we give a construction to map the remaining values in § 4, showing that embeddings from polyhedral surrogates always yield calibration.

While our notion of embedding is sufficient for calibration (and therefore consistency), it is worth noting that an embedding is not *necessary* for these conditions. For example, while logistic loss does not embed 0-1 loss, logistic loss and the sign link are still consistent for 0-1 loss. When working with polyhedral surrogates, however, embeddings are necessary for calibration in a strong sense, as we discuss in § 7: if a polyhedral surrogate L has a calibrated link to some target ℓ, then L must embed some discrete target $\hat{\ell}$ which can then be linked to ℓ.

3. 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, culminating in Theorem 1. We defer the question of when such surrogates are consistent to § 4.

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 11 *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 11 leads to the following useful tool for finding embeddings: if a surrogate has a finite representative set, it embeds its restriction to the representative set.

Proposition 12 *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 $\Gamma = \text{prop}[L]$ and $\gamma = \text{prop}[L|_{\mathcal{S}}]$. 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 11, 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. ■

Proposition 12 reveals an equivalent definition of an embedding which can be more convenient. Given a representative set $\mathcal{S}$ for ℓ, an injection $\varphi : \mathcal{S} \to \mathbb{R}^d$ is an embedding if: (i) the loss values match as in Definition 10(i), and (ii) $\varphi(\mathcal{S})$ is representative for L. This new definition is clearly implied by Definition 10; for the converse, Proposition 12 states that L embeds $L|_{\varphi(\mathcal{S})}$, which by (i) is the same loss as ℓ up to relabeling via φ.

With Proposition 12 in hand, we now shift our focus to *polyhedral* (piecewise-linear and convex) surrogates. While polyhedral surrogates do not directly elicit finite properties, as their report sets are infinite, they do elicit properties with a finite range, meaning the set of possible optimal report sets is finite.

Lemma 13 *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral loss. Then L is minimizable and elicits a property $\Gamma := \text{prop}[L]$. Moreover, 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 [Sketch] See § A for the full proof. As L is bounded from below, L is minimizable from Rockafellar (1997, Corollary 19.3.1) . 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 (Lemma 40). Moreover, we have $\Gamma(p)$ is 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 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. ■

From Lemma 13, one can simply select a point from each of the finitely many optimal sets to obtain a finite representative set. Plugging this finite representative set into Proposition 12 then yields an embedding.

Theorem 14 *Every polyhedral loss L embeds a discrete loss.*

Proof Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral loss, and $\Gamma = \text{prop}[L]$. By Lemma 13, $\Gamma(\Delta_{\mathcal{Y}})$ is a 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 12 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 15 that *every* discrete loss is embeddable by a polyhedral surrogate. In the proof, we apply a result we will prove in § 6: a minimizable surrogate embeds a discrete loss if and only if their Bayes risks match (Proposition 22).

Theorem 15 *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 (Rockafellar, 1997, 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 22 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. ■

Combining Theorems 14 and 15, we have Theorem 1.

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

The proof of Theorem 15 uses a construction via convex conjugate duality similar to many constructions in the literature. For example, the min-max objective in the literature on adversarial prediction (Asif et al., 2015; Farnia and Tse, 2016; Fathony et al., 2016, 2018) is a special case of this construction when one unfolds the definition of the convex conjugate of $-\underline{\ell}$. Reid et al. (2012) construct a canonical link function for proper losses with differentiable Bayes risks; the link maps a report $p \in \Delta_\mathcal{Y}$ to the gradient of the Bayes risk at p, which uses the same duality as above. Duchi et al. (2018, Proposition 3) give essentially the same construction as ours, but only comment on the calibration of surrogates under such constructions for multiclass classification tasks given by strictly concave losses, which excludes polyhedral surrogates. Finally, a similar construction also appears in the design of prediction markets (Abernethy et al., 2013) and in connections between proper losses and mechanism design (Frongillo and Kash, 2014, 2021).

4. Consistency and Linear Regret Transfer via Separated Links

We have seen that every polyhedral loss embeds some discrete loss. The embedding itself tells us how to link the embedded points back to the discrete reports: link $\varphi(r)$ back 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, via a construction to generate calibrated links for *any* polyhedral surrogate. Recalling that calibration is equivalent to consistency for discrete targets, this result implies that an embedding always yields consistency.

The key idea behind our construction is the notion of *separation*, a condition which is equivalent to calibration for discrete prediction problems. Roughly, given a surrogate L and discrete target ℓ, a link is ϵ-separated if the distance between any L-optimal point in $\mathbb{R}^d$, and any point that links to an ℓ-suboptimal report, is at least ϵ. We also show how this characterization also leads to linear *regret transfer* or *surrogate regret* bounds.

4.1 Separation

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.

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 $u \in \mathbb{R}^d$, $p \in \Delta_\mathcal{Y}$ with $\psi(u) \notin \gamma(p)$, we have $d_\infty(u, \Gamma(p)) \geq \epsilon$, where $d_\infty(u, A) := \inf_{a \in A} \|u - a\|_\infty$.*[3] *Similarly,*

3. Frongillo and Waggoner (2021) define ϵ-separation with a strict inequality $d_\infty(u, \Gamma(p)) > \epsilon$; we adopt a weak inequality as it is more natural in applications. For example, taking hinge loss for binary classification, the sign link is 1-separated under the weak inequality, but only $(1 - \delta)$-separated for $\delta > 0$ under the strict inequality.

we say ψ is ϵ-separated with respect to L and ℓ if it is ϵ-separated with respect to $\text{prop}[L]$ *and* $\text{prop}[\ell]$.

Theorem 17 *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 $\epsilon > 0$.*

Intuitively, calibration of a polyhedral surrogate and separated link follows from two facts. First, Lemma 13 states that a polyhedral surrogate has only a finite set of "optimal report sets" $\Gamma(\Delta_{\mathcal{Y}}) := \{\Gamma(p) : p \in \Delta_{\mathcal{Y}}\}$. Second, for a given p, the expected surrogate loss for a suboptimal point scales with the distance from the optimal set $\Gamma(p)$ at some minimum linear rate $\alpha > 0$; this rate is related to *Hoffman constants*. Combining with the first fact gives a universal minimum constant α for all p. Now bringing in ϵ-separation, any surrogate report linking to a suboptimal target report has expected surrogate loss at least $\alpha \cdot \epsilon > 0$. On the other hand, if a link is not separated, then the same two facts imply that, for a sequence of surrogate reports that get arbitrarily close to an optimal report set while linking to a suboptimal target report, this sequence has conditional expected loss approaching optimal, violating calibration. See § B for the proof.

4.2 Consistency

To prove Theorem 2, that embedding implies calibration, we now show how to construct a calibrated link from an embedding. In light of Theorem 17, 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. Construction 1, discussed next, "thickens" a given embedding to produce a link. Theorem 18 states that, for a small enough choice of ϵ, that link is separated. See § D for the proof. We also discuss a more general construction in § 7, and an alternate approach using results from Ramaswamy and Agarwal (2016) in § E.

Theorem 18 *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 for $L, \ell, \epsilon, \|\cdot\|$ produces a nonempty set of links, all of which are ϵ-separated with respect to L and 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.*

Proof From Theorem 18, since L embeds ℓ, there exists $\epsilon > 0$ such that Construction 1 yields a link ψ which is ϵ-separated with respect to L and ℓ. By Theorem 17, since ψ is ϵ-separated for L and ℓ, the pair (L, ψ) is calibrated with respect to ℓ. ■

To set the stage for Construction 1, we sketch the two main steps in proving Theorem 18: (i) showing that one can define a link ψ on all possible optimal points of L; (ii) "thickening" ψ so that it is separated.

For (i), given the embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$, begin by linking each embedding point back to its original report, so that $\psi(\varphi(r)) = r$. Now we wish to determine $\psi(u)$ for

non-embedding points $u \in \mathbb{R}^d$ which optimize L for some distribution. Let $\Gamma = \text{prop}[L]$ and define $\mathcal{U} = \Gamma(\Delta_\mathcal{Y}) := \{\Gamma(p) \mid p \in \Delta_\mathcal{Y}\}$ to be the set of all possible optimal sets. For each $U \in \mathcal{U}$, we can define $R_U = \{r \in \mathcal{S} \mid \varphi(r) \in U\}$ to be the set of target reports which, by definition of embedding, are ℓ-optimal when U is L-optimal. We would like to restrict $\psi(U) \subseteq R_U$, so that optimal surrogate reports are mapped to optimal target reports. The challenge is that we could have a point $u \in U \cap U'$ for two optimal sets $U, U' \in \mathcal{U}$, and *a priori*, it could be that $R_U \cap R_{U'} = \emptyset$. The first step of the proof is to show that this scenario cannot arise.

$$\forall \mathcal{U}' \subseteq \mathcal{U},\ \cap_{U \in \mathcal{U}'} U \neq \emptyset \implies \cap_{U \in \mathcal{U}'} R_U \neq \emptyset\ . \tag{4}$$

Thus, for any $u \in \cup_{U \in \mathcal{U}} U$, we have a nonempty set of valid choices for $\psi(u)$. (Eq. (4) may appear similar to indirect elicitation; in fact the two conditions are equivalent, as we discuss in § 7.)

For (ii), we show that this link can be "thickened" by some positive ϵ, as described next. Let $U \in \mathcal{U}$. By the above, 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 R_U. For $\epsilon > 0$, this condition directly implies separation.

It is not clear that this linking is possible, however, because a point u may be in the intersection of several thickened sets U_ϵ, U'_ϵ, etc., corresponding to $\Gamma(p), \Gamma(p')$, etc. Therefore, we need to take each possible collection U, U', etc., show that their intersection (if nonempty) contains a legal choice for the link, and then thicken their intersection in an analogous way.

To do so, given $u \in U_\epsilon \cap U'_\epsilon \cap \ldots$, 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_ϵ, U'_ϵ, etc. The key claim is that, for small enough $\epsilon > 0$, $\Psi(u)$ is nonempty: at least one permitted value for $\psi(u)$ remains. This claim follows from a geometric result (Lemma 52) 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, eq. (4) implies that there exists a permitted choice of link for the intersection of the thickenings. It is crucial that, by Lemma 13, polyhedral surrogates only have finitely many sets of the form $U = \Gamma(p)$. Together, these observations yield a single sufficiently small 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 adjust Ψ to narrow down to only legal choices; we then pick from $\psi(u)$ from $\Psi(u)$ arbitrarily. Theorem 18 implies that, for all small enough ϵ, the resulting link ψ is well-defined at all points, and ϵ-separated.

Construction 1 (ϵ-thickened link) *Let $L : \mathbb{R}^d \to \mathbb{R}^\mathcal{Y}_+$, $\ell : \mathcal{R} \to \mathbb{R}^\mathcal{Y}_+$, $\epsilon > 0$, and a norm $\|\cdot\|$ be given, such that L is polyhedral and embeds ℓ via the embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$ for a representative set $\mathcal{S} \subseteq \mathcal{R}$. Define $\Gamma = \text{prop}[L]$ and $\mathcal{U} = \{\Gamma(p) \mid p \in \Delta_\mathcal{Y}\}$. For all $U \in \mathcal{U}$, define $R_U = \{r \in \mathcal{S} \mid \varphi(r) \in U\}$. The* ϵ-thickened link *ψ is constructed as follows. First, initialize the* link envelope *$\Psi : \mathbb{R}^d \to 2^\mathcal{S}$ by setting $\Psi(u) = \mathcal{S}$ for all u. Then for each $U \in \mathcal{U}$, for all points u such that $\inf_{u^* \in U} \|u^* - u\| < \epsilon$, update $\Psi(u) = \Psi(u) \cap R_U$. If we have $\Psi(u) \neq \emptyset$ for all $u \in \mathbb{R}^d$, then the construction* produces a link *$\psi \in \Psi$ pointwise, breaking ties arbitrarily.*

In § 7 we generalize this construction beyond embeddings, where we only require that L indirectly elicits prop$[\ell]$. There we will see that, perhaps surprisingly, this construction recovers every possible calibrated link.

Applying Construction 1 also enables one to verify the consistency of a given 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. See § 5.5 for an illustration of this test. On the other hand, if ψ^* cannot be produced from Construction 1, then it cannot be a calibrated link. This impossibility can be shown, for example, if there exists a point u where $\Psi(u)$ is empty for all $\epsilon > 0$.

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 § 5.2.

4.3 Surrogate regret bounds

Recall that the approach of surrogate risk minimization is to learn a hypothesis h that minimizes expected surrogate loss, then output hypothesis $\psi \circ h$, which hopefully minimizes expected target loss. We would like surrogates where a bound on surrogate loss of h immediately implies a bound on target loss of $\psi \circ h$. One can formalize this problem in terms of *regret*, as follows. Fix a data distribution $\mathcal{D}$. The *surrogate regret* R_L of h and the *target regret* R_ℓ of the implied hypothesis $\psi \circ h$, are given by

$$R_L(h; \mathcal{D}) = \mathbb{E}_{(X,Y)\sim\mathcal{D}} L(h(X))_Y - \inf_{h':\mathcal{X}\to\mathbb{R}^d} \mathbb{E}_{(X,Y)\sim\mathcal{D}} L(h'(X))_Y \ ,$$
$$R_\ell(\psi \circ h; \mathcal{D}) = \mathbb{E}_{(X,Y)\sim\mathcal{D}} \ell(\psi(h(X)))_Y - \inf_{g':\mathcal{X}\to\mathcal{R}} \mathbb{E}_{(X,Y)\sim\mathcal{D}} \ell(g'(X))_Y \ ,$$

where the infimum is taken over all measurable functions. The infimum represents the risk of the Bayes optimal hypothesis, so regret can be viewed as *excess risk* under the assumption that the learner's hypothesis class contains the Bayes optimal.

Consistency (Definition 7) means that if the surrogate regret of h converges to zero, then the target regret of $\psi \circ h$ does as well; in other words, $R_L(h; \mathcal{D}) \to 0$ implies $R_\ell(\psi \circ h; \mathcal{D}) \to 0$. Consistency is therefore a minimal requirement; in general, we are also interested in the *rate* at which the target regret diminishes, as a function of the number of data points m. A *regret transfer bound*, also called a *surrogate regret bound*, gives a guarantee on the relationship between the rates of convergence of R_L and R_ℓ. For example, a surrogate with a fast rate of convegence $R_L \to 0$ as $m \to \infty$ is not very useful if we nevertheless have a slow rate $R_\ell \to 0$.

We show that, for any polyhedral surrogate, the transfer is linear: if surrogate regret diminishes at a rate of $O(f(m))$, then the target rate is also $O(f(m))$. In particular, fast convergence in surrogate regret implies fast convergence in target regret.

Theorem 19 *Let (L, ψ) be consistent for a discrete loss ℓ, and L polyhedral. Then there exists $c > 0$ such that, for all measurable hypotheses h and data distributions $\mathcal{D}$, we have $R_\ell(\psi \circ h; \mathcal{D}) \leq c \cdot R_L(h; \mathcal{D})$.*

In the proof (§ C), we further show that the constant c can be decomposed in terms of three constants, which depend on L, ψ, and ℓ, respectively. Specifically, we may write $c = C_\ell H_L / \epsilon_\psi$, where H_L is the Hoffman constant for L, ϵ_ψ the separation of ψ, and C_ℓ the maximum loss gap of ℓ. This expression gives the intuition that larger link separation is generally better for performance. In some cases, this bound can be tightened, as we discuss in § C.2. See Frongillo and Waggoner (2021) for a quadratic lower bound on the rate transfer for sufficiently non-polyhedral surrogates.

5. 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. To warm up, we study the abstain surrogate given by Ramaswamy et al. (2018), which is an embedding, and show how to rederive their link function and surrogate regret bounds (§ 5.2). We then give three examples of subsequent works that use our framework, in the context of structured binary classification (§ 5.3), multiclass classification (§ 5.4), and top-k classification (§ 5.5). In each of these latter three examples, 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 link functions which appear challenging to derive otherwise.

5.1 Applying the embedding framework

When using our framework to study the consistency or inconsistency of an existing surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, often the first step is determining the loss it embeds. To do so, 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 12. From here one can further remove redundant reports until arriving at a tight embedding if desired, or re-label embedded reports to a more intuitive form; call this resulting embedded loss $\hat{\ell}$.

Once the embedded discrete loss $\hat{\ell}$ is known, the behavior of the surrogate L becomes more clear. In particular, we learn what problem L is actually solving, as captured by $\hat{\ell}$. If this problem $\hat{\ell}$ is not the desired target problem ℓ, we can still derive restrictions on conditional distributions (e.g., $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$) for which L is $\mathcal{P}$-calibrated for ℓ (i.e., the inequality holds for all $p \in \mathcal{P}$). Any level set of the embedded property $\hat{\gamma} = \text{prop}[\hat{\ell}]$ which spans multiple level sets of the target property $\gamma = \text{prop}[\ell]$ will lead to inconsistency for ℓ (Figure 2). To obtain consistency with respect to a desired target, therefore, it suffices to restrict to the union of level sets of $\hat{\gamma}$ which are each fully contained in some level set of γ.

With an embedding in hand, Construction 1 provides a calibrated link function from L to $\hat{\ell}$. This construction is especially beneficial in cases where the most intuitive link functions

4. In some cases, these regions do not have vertices, such as the top-k surrogates in § 5.5 which are invariant in the all-ones direction; here one can restrict to a subspace, or otherwise select among equivalent reports.

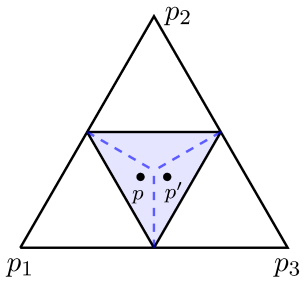


Figure 2: Using an embedding to show inconsistency. Let L be a surrogate embedding $\hat{\ell}$, and let ℓ be a desired target; here $L = L_{\text{BEP}}$, $\hat{\ell} = \ell_{\text{abs}}$ (§ 5.2) and ℓ is 0-1 loss for multiclass classification. Let $\Gamma = \text{prop}[L]$, $\hat{\gamma} = \text{prop}[\hat{\ell}]$, and $\gamma = \text{prop}[\ell]$ be the properties elicited by these losses; here $\gamma =$ mode, as 0-1 loss elicits the mode. The level sets of $\hat{\gamma}$ are given in solid black lines, and those of γ in dashed blue lines. To exhibit non-calibration, take distributions p, p' in the relative interior of the blue cell $\hat{\gamma}_{\hat{r}}$ (here $\hat{r} = \perp$ for ℓ_{abs}) but in different cells of γ. These distributions will satisfy $\hat{\gamma}(p) = \hat{\gamma}(p')$, and thus $\Gamma(p) \cap \mathcal{S} = \Gamma(p') \cap \mathcal{S}$ by definition of embedding, but $\gamma(p) \cap \gamma(p') = \emptyset$. Taking $u \in \Gamma(p) = \Gamma(p')$, it is impossible to define $\psi(u)$ to satisfy calibration, as $\psi(u)$ cannot be in $\gamma(p)$ and $\gamma(p')$ simultaneously. In particular, even though u is L-optimal for both p and p', $\psi(u)$ will be ℓ-suboptimal for at least one, violating calibration. We may impose restrictions on the conditional distributions to remove the blue cell, however, in order to satisfy calibration. For this example, the L_{BEP} is classification-calibrated if one restricts to the set of distributions where at least one label $y \in \mathcal{Y}$ has probability $p_y \geq \frac{1}{2}$.

are not calibrated, and no known calibrated link is known; see § 5.3 for a somewhat intricate example. Surrogate regret bounds then follow from Theorem 19, as we illustrate in § 5.2. In particular, our results imply the existence of linear regret transfer bounds bounds for several applications where no such bounds were known (§ 5.3, 5.4, 5.5).

Finally, our link construction can even be useful in cases where the search for consistent surrogates has been restricted to those accommodating a particular canonical link function ψ. For example, one typically uses the sign link for binary classification, and the argmax link (the k largest coordinates) for top-k classification (§ 5.5). As we show in Proposition 57, Construction 1 fully characterizes the set of possible calibrated link functions for a polyhedral embedding via the link envelope Ψ, so ψ is calibrated if and only if it is contained in Ψ for some $\epsilon > 0$. We demonstrate this approach for top-k classification in § 5.5. More generally, however, while such canonical 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.

5.2 Consistency of abstain surrogate and link construction

Several authors consider a variant of binary and multiclass classification, with the addition of an *abstain* option (Bartlett and Wegkamp, 2008; Ramaswamy et al., 2018; Madras et al., 2018; El-Yaniv and Wiener, 2010; Cortes et al., 2016). Ramaswamy et al. (2018) study the loss $\ell_{\text{abs}} : \mathcal{Y} \cup \{\perp\} \to \mathbb{R}^{\mathcal{Y}}_+$ defined by $\ell_{\text{abs}}(r)_y = 0$ if $r = y$, $1/2$ if $r = \perp$, and 1 otherwise. The report $\perp$ corresponds to "abstaining" to predict, in exchange for a constant loss regardless of outcome y. Ramaswamy et al. give the polyhedral *binary encoded predictions (BEP)* surrogate L_{BEP}, and the link ψ^∞ which they show is calibrated for ℓ_{abs}. Letting $d = \lceil \log_2 |\mathcal{Y}| \rceil$, their surrogate $L_{\text{BEP}} : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ is given by

$$L_{\text{BEP}}(u)_y = \max_{j \in [d]} (1 - \varphi(y)_j u_j)_+ \ , \tag{5}$$

where $\varphi : \mathcal{Y} \to \{-1, 1\}^d$ is an injection.[5] Observe that L_{BEP} is exactly hinge loss when $|\mathcal{Y}| = 2$ and thus $d = 1$. The authors show that the link ψ^∞ is calibrated, where

$$\psi^\infty(u) = \begin{cases} \perp & \min_{i \in [d]} |u_i| \leq 1/2 \\ \varphi^{-1}(\text{sgn}(u)) & \text{otherwise} \end{cases}, \tag{6}$$

and they go on to establish linear surrogate regret bounds for $(L_{\text{BEP}}, \psi^\infty)$.

Using our framework, one can show that L_{BEP} embeds (2 times) ℓ_{abs}, with the embedding given by φ above where we define $\varphi(\perp) = 0 \in \mathbb{R}^d$. (Following the general procedure outlined above, the regions where L_{BEP} is affine all have vertices in the set $\{-1, 1\}^d \cup \{0\}$, meaning it is representative, and L_{BEP} restricted to that set is precisely $2\ell_{\text{abs}} \circ \varphi^{-1}$.)

As an illustration, one can use the fact that L_{BEP} embeds ℓ_{abs} to verify that L_{BEP} is inconsistent for multiclass classification, i.e., with respect to 0-1 loss. In particular, since the abstain report $\perp$ is ℓ_{abs}-optimal whenever $\max_{y \in \mathcal{Y}} p_y \leq 1/2$, by the definition of embedding, the origin $0 \in \mathbb{R}^d$ is L_{BEP}-optimal for the same distributions. Recalling that 0-1 loss elicits the mode, one can now find two distributions with different modes but for which 0 is L_{BEP}-optimal, violating calibration (Figure 2).

Moreover, as we illustrate in Figure 3(L), the link ψ^∞ proposed by Ramaswamy et al. can be recovered from Construction 1 by choosing the norm $\|\cdot\|_\infty$ and $\epsilon = 1/2$ (or smaller). Hence, our framework could have simplified the process of finding ψ^∞, and the corresponding proof of consistency. It also could have simplified the derivation of surrogate regret bounds (§ 4.3); we show how to recover the tight bound of Ramaswamy et al. for the BEP surrogate in § C.2.

To illustrate these points further, consider the alternate link ψ^1 in Figure 3(R), given by

$$\psi^1(u) = \begin{cases} \perp & \|u\|_1 \leq 1 \\ \varphi^{-1}(\text{sgn}(u)) & \text{otherwise} \end{cases}. \tag{7}$$

This link is the result of Construction 1 for norm $\|\cdot\|_1$ and the choice $\epsilon = 1$, which proves calibration of (L_{BEP}, ψ^1) with respect to ℓ_{abs}. Aside from its simplicity, one possible advantage of ψ^1 is that it assigns $\perp$ to much less of the surrogate space $\mathbb{R}^d$.

5.3 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 (Hazan et al., 2010; Gao and Zhou, 2011; Osokin et al., 2017). 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 $\text{dis}(r, y) = \{i \in [k] \mid r_i \neq y_i\}$, meaning we may write $\ell^f(r)_y = f(\text{dis}(r, y))$ 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 (2018) introduce the *Lovász hinge* surrogate $L^f : \mathbb{R}^k \to \mathbb{R}_+^{\mathcal{Y}}$

5. To translate our notation to that of Ramaswamy et al. (2018), take $B = -\varphi$.

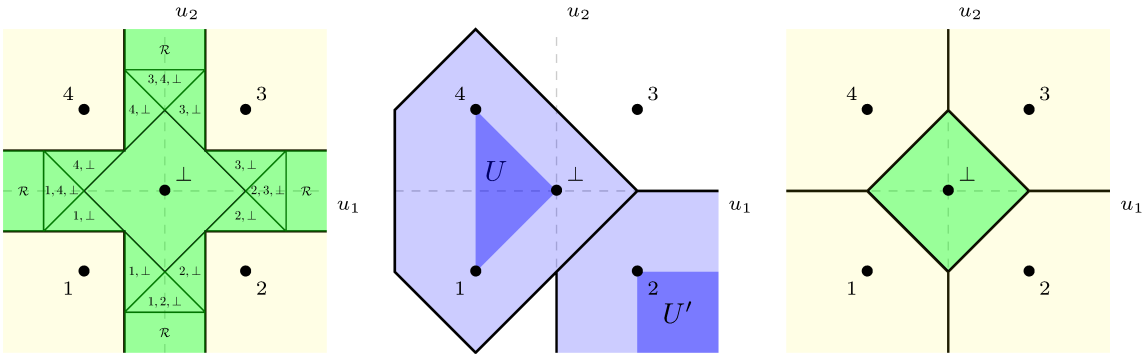


Figure 3: Designing links for L_{BEP} with $d = 2$ using Construction 1. The embedding is shown in bold labeled by the corresponding reports. (L) The link envelope Ψ resulting from Construction 1 using $\|\cdot\|_\infty$ and $\epsilon = 1/2$, and a possible link ψ which matches eq. (6) from Ramaswamy et al. (2018). (M) An illustration of the thickened sets for two sets $U, U' \in \mathcal{U}$, using $\|\cdot\|_1$ and $\epsilon = 1$. (R) The envelope Ψ and link ψ using $\|\cdot\|_1$ and $\epsilon = 1$.

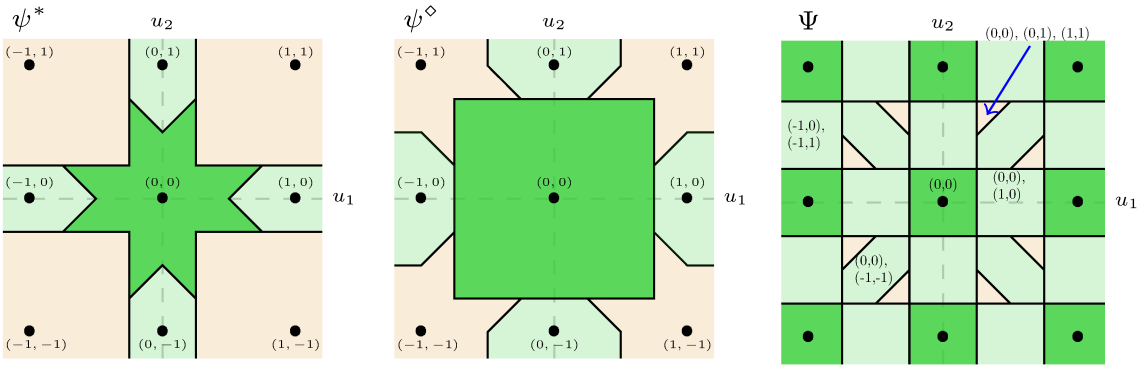


Figure 4: Links ψ^* and $\psi^\diamond$ such that (L^f, ψ^*) and $\psi^\diamond$ are calibrated with respect to ℓ^f_{abs} for all suitable f. Points in each region link to the embedding point contained in the region. Both are constructed via the link envelope Ψ from Construction 1, which yields possible choices for calibrated links.

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 is actually consistent for ℓ^f.

Finocchiaro et al. (2022b) 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 (in which case 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_{BEP}, for multiclass classification with an abtain option (§ 5.2), is the special case of L^f where $f(S) = \mathbb{1}\{S \neq \emptyset\}$.

To derive the discrete loss ℓ^f_{abs} that L^f embeds, the authors follow an approach similar to § 5.1 to show that the set $\mathcal{V} = \{-1, 0, 1\}^k$ is representative for L^f, for any choice of submodular and increasing f. From Proposition 12, 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

$$\ell^f_{\text{abs}}(v)_y = f(\text{dis}(v, y) \setminus \text{abs}(v)) + f(\text{dis}(v, y)) \ . \tag{8}$$

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

As Finocchiaro et al. observe, actually deteriving a calibrated link function in this case is nontrivial. Simple threshold links like for the BEP surrogate in § 5.2 are not always calibrated, thus casting doubt that a trial-and-error approach for finding the link would be successful. Instead, the authors leverage our thickened link construction (Construction 1) to derive two links ψ^* and $\psi^\diamond$, which have somewhat intricate geometric structure (Figure 4). Perhaps surprisingly, by deriving the link envelope Ψ which is contained in the envelopes for L^f for *all* submodular and increasing f, they prove that $\psi^*(u) \subseteq \Psi(u)$ and $\psi^\diamond(u) \subseteq \Psi(u)$ for all $u \in \mathbb{R}^d$. Thus, both (L^f, ψ^*) and $(L^f, \psi^\diamond)$ are simultaneously calibrated with respect to ℓ^f_{abs} for all such f.

5.4 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 and Watkins (1999) give an alternate surrogate for multiclass SVM prediction, defined as follows for predictions $u \in \mathbb{R}^n$,

$$L_{\text{WW}}(u)_y = \sum_{i \in \mathcal{Y} : i \neq y} (1 - (u_y - u_i))_+ \ . \tag{9}$$

This surrogate L_{WW} was later shown to be inconsistent with respect to 0-1 loss (Tewari and Bartlett, 2007; Liu, 2007).

Wang and Scott (2020) use our embedding framework to show that the Weston–Watkins hinge embeds an *ordered partition* loss ℓ_{OP}, as defined below. In turn, they recover the result of inconsistency with respect to 0-1 loss. The report space for ℓ_{OP} 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 $\ell_{\text{OP}} : \mathcal{T} \to \mathbb{R}^{\mathcal{Y}}_+$ is then defined

$$\ell_{\text{OP}}(T)_y = \sum_{i=1}^{s} (|T_i| \cdot \mathbb{1}\{y \notin T_{i-1}\}) - 1 \ .$$

6. To recover the partition of Wang and Scott (2020), take $S_i = T_i \setminus T_{i-1}$.

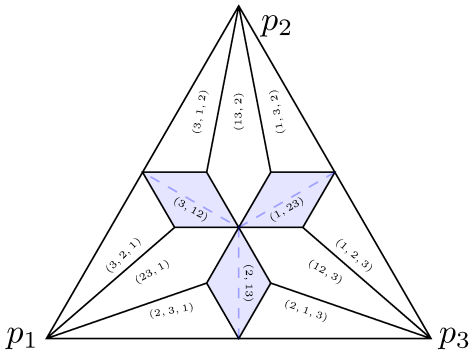


Figure 5: Level sets of prop$[\ell_{\text{OP}}]$ (solid lines) juxtaposed against the level sets of mode (dashed lines). For the same reason as in Figure 2, the level sets of prop$[\ell_{\text{OP}}]$ whose relative interiors span multiple cells of the mode, colored blue, cannot be properly linked to the mode. These offending cells correspond to reports whose highest partition has more than one element, where in the white cells, the "highest" element of the partition is well-defined.

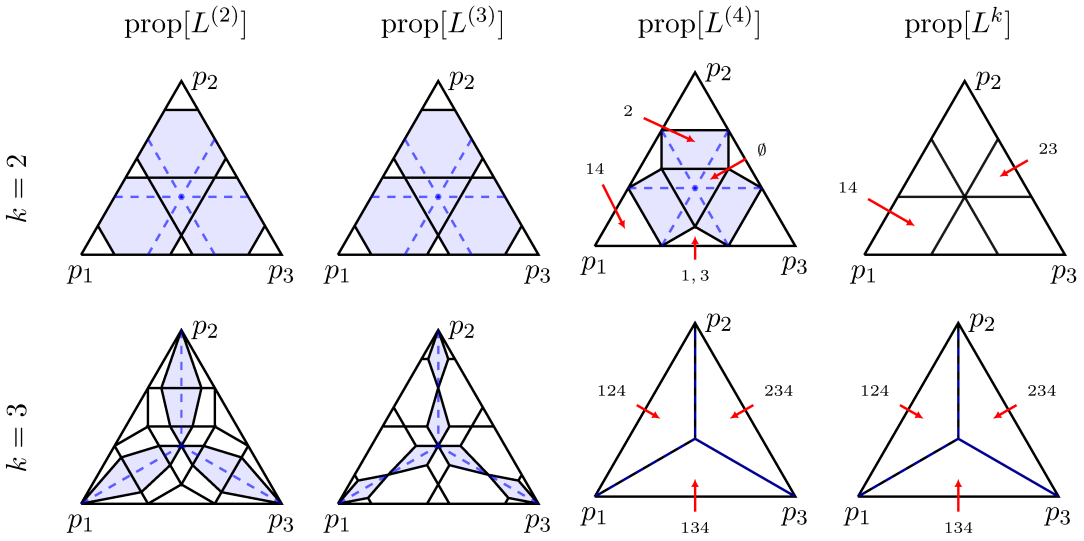


Table 1: Visualizations of the properties elicited by the losses (embedded by) $L^{(2)}$, $L^{(3)}$, $L^{(4)}$ studied by Yang and Koyejo (2020) and Finocchiaro et al. (2022a), and L^k in eq. (10). We take $n = 4$ and $k \in \{2, 3\}$, and visualize in 2 dimensions by fixing $p_4 = 1/4$. The blue-filled regions are cells of the surrogate property which cross the dashed blue lines of the target property, exhibiting inconsistency (see Figure 2). Intuitively, the inconsistency arises from ambiguity in the top k elements of the optimal surrogate report.

The loss ℓ_{OP} can be interpreted as a variation of 0-1 loss incorporating confidence: reports are a nested sequence of sets, and the penalty upon seeing label y is the cardinality of the first set containing y, plus the cardinality of all earlier sets.

Upon showing that L_{WW} embeds ℓ_{OP}, Wang and Scott then characterize prop$[\ell_{\text{OP}}]$. In the same manner as Figure 2, knowledge of the level sets of prop$[\ell_{\text{OP}}]$ clarifies which conditional distributions are the source of inconsistency for classification. Removing these distributions gives a set $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ such that L_{WW} and the canonical argmax link $\psi(u) : u \mapsto r \in \arg\max_y \langle e_y, u \rangle$ are calibrated with respect to 0-1 loss on $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ (i.e., such that eq. (2) holds for all $p \in \mathcal{P}$). See Figure 5 for an illustration.

5.5 Surrogates for top-k classification

In settings like object recognition and information retrieval, it is natural to predict a set S of labels. In top-k classification, one requires $|S| = k$, and given the true label y, the target loss is $\ell^{\text{top-}k}(S)_y = \mathbb{1}\{y \notin S\}$ (Lapin et al., 2015, 2016, 2018; Yang and Koyejo, 2020; Berrada et al., 2018; Rastegari et al., 2011; Reddi et al., 2019). 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 (2020) show that a number of previously proposed polyhedral losses, i.e., those which are convex and hinge-like, are inconsistent. They further conjecture that perhaps no surrogate could satisfy all three properties.

Finocchiaro et al. (2022a) 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 fewer than k labels. The conditional distributions for which these optimal reports differ from the optimal top-k reports are shown in Table 1 with $n = 4$ and $k \in \{2, 3\}$. (Recall $n = |\mathcal{Y}|$, the number of labels.)

For example, one of the surrogates is $L^{(4)}(u)_y = \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]}\right)_+$, where $u_{[i]}$ denotes the ith largest element of $u \in \mathbb{R}^n$. 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, using the same technique from Figure 2, the authors go on to derive constraints on the conditional distributions under which the proposed surrogates are actually consistent for top-k classification; these constraints are tighter than previous constraints (Yang and Koyejo, 2020).

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

$$L^k(u)_y = \max\left(u_{[1]}, \max_{m \in \{k+1,\ldots,n\}}\left[1 - \frac{k}{m} + \frac{1}{m}\sum_{i=1}^{m} u_{[i]}\right]\right) - u_y \ . \tag{10}$$

That is, they show that a hinge-like surrogate does exist which is both convex and consistent. 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 15 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 calibrated. 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 $\epsilon \leq \frac{1}{2n}$ (Finocchiaro et al., 2022a, Theorem 4.4).

Interestingly, by choosing $k = 1$, one obtains what appears to be a novel surrogate for 0-1 loss in multiclass settings,

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

In particular, this surrogate is consistent with respect to 0-1 loss using the canonical argmax link. By contrast, recall that the Weston-Watkins hinge is not consistent for any link (§ 5.4), despite being proposed as a surrogate for 0-1 loss.

6. Additional Structure of Embeddings

We have shown in § 3 a close 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 set (Lemma 20). This result implies that the embedding condition simplifies to matching Bayes risks (Proposition 22). We also use this result to understand deeper structure of embeddings, and the geometry of the underlying properties. In particular, we study a natural notion of a "trimmed" loss function (Definition 23), and connect this notion to tight embeddings, and to non-redundancy from property elicitation (Proposition 24).

6.1 Structure of polyhedral Bayes risks

While we have focused on polyhedral losses thus far, many of our results extend to losses with polyhedral Bayes risks, a strictly 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 14 constructs a finite representative set $\mathcal{S}$ for any polyhedral loss L, and thus $\underline{L} = \underline{L|_{\mathcal{S}}}$ by Lemma 11, which is polyhedral. Conversely, however, 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 6, which matches hinge loss on the interval $[-1, 1]$ but is strictly convex outside the interval $[-2, 2]$, still embeds twice 0-1 loss.

Much of our embedding framework relies on the existence of finite representative sets. Our main structural result is that a minimizable loss has a finite representative sets if and only if its Bayes risk is polyhedral. The proof looks at the facets (full-dimensional faces) of the Bayes risk, argues that each facet is generated by the loss at a particular report, and shows the (finite) set of these reports is representative. Along the way, we identify several other useful facts deriving from this same geometry; for example, a discrete loss tightly embedded by a loss are unique up to relabeling, any set-wise minimal representative set must be minimum in cardinality, and the level sets of the corresponding property are unique and full-dimensional. Together, these facts form Lemma 20, which we use throughout this section. See § F for omitted proofs.

Lemma 20 *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 $\Gamma = \text{prop}[L]$, 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 $\mathcal{V} \subseteq L(\mathcal{R}')$.*

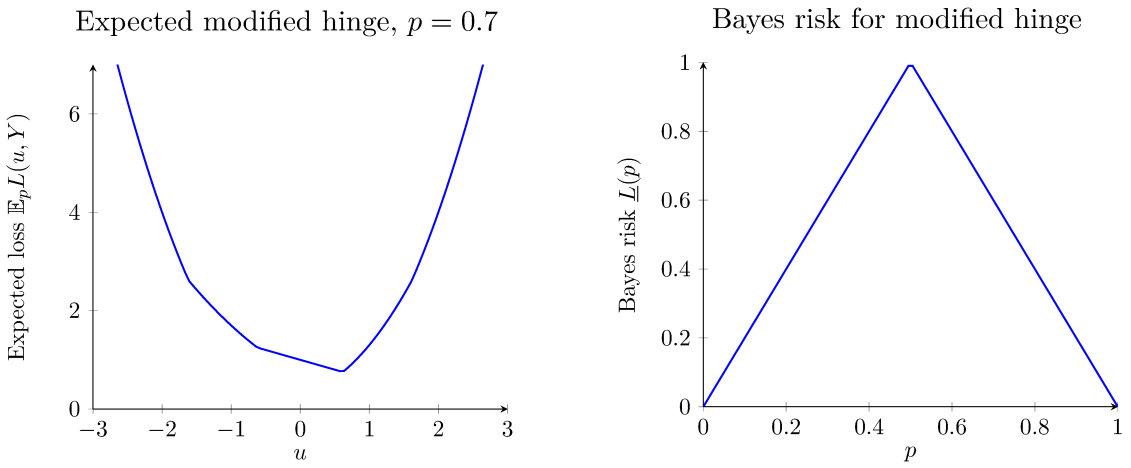


Figure 6: (L) Expected modified hinge loss for fixed distribution; (R) Bayes risk of modified hinge still matches the Bayes risk of hinge.

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 $\ell(\mathcal{R}') = \mathcal{V}$.*

Lemma 20 now allows us to observe the relationship between embeddings, finite representative sets and polyhedral Bayes risks.

Corollary 21 *The following are equivalent for any minimizable loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$.*

1. *$\underline{L}$ is polyhedral.*
2. *L has a finite representative set.*
3. *L embeds a discrete loss.*

Proof If $\underline{L}$ is polyhedral, then Lemma 20 implies that L has a finite representative set. For any surrogate L with a finite representative set $\mathcal{S}$, Proposition 12 implies that L embeds $L|_{\mathcal{S}}$, which is finite. Finally, if L embeds a discrete loss, then by definition L has a finite representative set $\mathcal{S}$, and Lemma 11 implies that $\underline{L} = \underline{L|_{\mathcal{S}}}$, which is polyhedral as the pointwise minimum of a finite set of affine functions. ■

From Corollary 21, 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 20 follow whenever L has a finite representative set.

6.2 Equivalent condition: matching Bayes risks

Lemma 20 leads to another appealing equivalent condition to an embedding: a surrogate embeds a discrete loss if and only if their Bayes risks match. The proof follows by mapping the two conditions of an embedding onto the geometric structure revealed by Lemma 20: (ii) if the properties have the same level sets, the Bayes risks have the same projections onto $\Delta_\mathcal{Y}$, and (i) if the loss values match, then the slopes of the Bayes risk must be identical as well.

Proposition 22 *Let discrete loss ℓ and minimizable loss L be given. Then L embeds ℓ if and only if $\underline{L} = \underline{\ell}$.*

Proof Define $\Gamma := \text{prop}[L]$ and $\gamma := \text{prop}[\ell]$. 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 11, 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

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

For the reverse implication, assume $\underline{L} = \underline{\ell}$, which are polyhedral functions as ℓ is discrete. From Lemma 20(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 12, 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

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

giving condition (ii). ■

We use this fact in the proof of Theorem 15 to show that every discrete loss is embedded by some polyhedral surrogate. See Figure 1 for an illustration.

6.3 Trimming a loss

Central to the structural results in Lemma 20 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 20(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 20(2).

Definition 23 (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 induce the same loss vectors, or have the same trim.

Proposition 24 *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* $\mathrm{trim}(L) = \mathrm{trim}(\ell)$. *Furthermore, L tightly embeds ℓ if and only if ℓ is injective and* $\mathrm{trim}(L) = \ell(\mathcal{R})$.

Proof As L has a finite representative set, it is minimizable. Proposition 22 gives L embeds ℓ if and only if $\underline{L} = \underline{\ell}$. If $\underline{L} = \underline{\ell}$, Lemma 20(2) gives $\mathrm{trim}(L) = \mathrm{trim}(\ell)$. For the converse, suppose $\mathrm{trim}(L) = \mathrm{trim}(\ell) =: \mathcal{V}$. Define the discrete loss $\ell_{\mathrm{trim}} : \mathcal{V} \to \mathcal{V}, v \mapsto v$. Then ℓ_{trim} is injective and $\ell_{\mathrm{trim}}(\mathcal{V}) = \mathcal{V}$, so from Lemma 20(8), both L and ℓ tightly embed ℓ_{trim}. We conclude $\underline{L} = \underline{\ell_{\mathrm{trim}}} = \underline{\ell}$ from Proposition 22. The second statement also follows directly from Lemma 20(8). ■

In a strong sonse, the trim operation reduces a loss to its core: the unique minimal set of loss vectors that drive its statistical behavior. One can therefore think of designing consistent convex surrogates as trying to "fill out" this minimal set with additional loss vectors so that one attains convexity while keeping trim the same.

6.4 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 (Frongillo and Kash, 2014; Lambert, 2018). Intuitively, a loss is non-redundant if no report is weakly dominated by another report.

Definition 25 (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 20, we can see that in fact these two notions are equivalent when L has a polyhedral Bayes risk.

Proposition 26 *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 $\Gamma = \text{prop}[L]$. 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 20(5) gives some minimum representative set $\mathcal{S} \subseteq \mathcal{R}'$. Suppose we had some $r \in \mathcal{R}' \setminus \mathcal{S}$. Now Lemma 20(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 27 *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 $\text{red}(\Gamma) := \{r \in \mathcal{R} \mid \exists r' \in \mathcal{R},\ \Gamma_r \subsetneq \Gamma_{r'}\}$ be the set of strictly redundant reports. Similarly, for minimizable L, let $\text{red}(L) := \text{red}(\text{prop}[L])$.

Proposition 28 *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 $\mathcal{R}' = \mathcal{R} \setminus \text{red}(L)$.*

Proof Let $\Gamma = \text{prop}[L]$. Let $\mathcal{S}$ be a minimum representative set for L, and let $s \in \mathcal{S}$. Suppose for a contradiction that $s \in \text{red}(\Gamma)$. Then we have some $r \in \mathcal{R}$ with $\Gamma_s \subsetneq \Gamma_r$. From Lemma 20(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 a minimum representative set. Thus $\mathcal{S} \subseteq \mathcal{R} \setminus \text{red}(\Gamma)$, which implies $\mathcal{R}' \subseteq \mathcal{R} \setminus \text{red}(\Gamma)$.

For the reverse inclusion, let $r \in \mathcal{R} \setminus \text{red}(\Gamma)$. Let $\mathcal{S}$ again be a minimum representative set for L. From Lemma 20(4,7), we have some $s \in \mathcal{S}$ such that $\Gamma_r \subseteq \Gamma_s$. By definition of $\text{red}(L)$, 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 20(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 29 *Let $L : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$ have a finite representative set. Then* $\text{trim}(L) = L(\mathcal{R} \setminus \text{red}(L))$.

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

6.5 A property elicitation perspective on trimmed losses

We conclude this section with a structural result similar to Lemma 20, but for properties. To do so, we must first generalize the definition of embeddeding to properties. 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 10

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, 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, but potentially with other reports "filling in the gaps" between the embedded reports of γ. When working with convex surrogates, 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 directly 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 the same full-dimensional level sets.

The proof relies heavily on Lemma 20. The statements about level sets use the following corollary of Proposition 24 for properties.

Corollary 30 *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 20(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 26, 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 31 *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 20 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 30 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 20(5), we have some minimum

representative set $\mathcal{U} \subseteq \mathcal{S}$ for L and Γ, implying statement 3. Moreover, Lemma 20(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 20(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 11 now states that L embeds $L|_{\mathcal{S}}$, so Γ embeds $\gamma := p \mapsto \Gamma(p) \cap \mathcal{S}$, giving Case 1. Finally, Corollary 30 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 13). Proposition 31 shows a complementary 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.

7. Polyhedral Indirect Elicitation Implies Consistency

As we have observed, consistency, and therefore calibration, implies indirect elicitation (§ 2.3). In general, indirect elicitation is simpler and weaker than calibration, since it only depends on the loss through the property it elicits, i.e., its exact minimizers. Surprisingly, for polyhedral surrogates, we show the converse: indirect elicitation implies calibration, and therefore consistency.

Theorem 32 *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a 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 with respect to ℓ.*

One technical detail is that the link function may have to change. That is, we will show that if (L, ψ) indirectly elicits $\mathrm{prop}[\ell]$, then there exists some potentially different link ψ' such that (L, ψ') is calibrated with respect to ℓ. To see why this change may be necessary, consider again the example from § 2.3: hinge loss with the link $\psi(u) = -1$ for $u < 1$ and $\psi(u) = 1$ for $u \geq 1$. Here indirect elicitation is achieved, since we have $\psi((-\infty, -1]) = \{-1\}$ and $\psi([1, \infty)) = \{1\}$, but the link is not ϵ-separated for any $\epsilon > 0$. In general, it is not clear whether one can always adjust the link in this case to achieve separation, and therefore calibration. Fortunately, for polyhedral surrogates, one can always "thicken" a given link to achieve separation.

We give two proofs of Theorem 32. The first is direct: we show, as foreshadowed in § 4, that our thickened link construction can be generalized for indirect elicitation. In fact, we will further prove that our general construction recovers every possible calibrated link function. The second proof highlights the central role that embeddings play when reasoning about polyhedral surrogates. Specifically, we will show that if a polyhedral surrogate indirectly elicits a finite property, the link function must "pass through" an embedding, giving calibration through Construction 1.

Finally, a third proof of Theorem 32 is implicit in Ramaswamy and Agarwal (2016, Theorem 8). The authors use an entirely different link construction involving the superprediction set of the surrogate loss. We discuss their result and how it relates to our work in § E.

7.1 Generalizing the thickened link construction

Given that Construction 1 uses the embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$ in a crucial role, it is not immediately clear how to generalize the construction beyond embeddings. Specifically, this crucial role is in the definition of R_U, the set of target reports which must be optimal whenever a given surrogate report set $U \subseteq \mathbb{R}^d$ is optimal. Using the embedding, we can simply define $R_U = \{r \in \mathcal{S} \mid \varphi(r) \in U\}$, since the definition of embedding means that those are exactly the target reports (among the representative set $\mathcal{S}$) which are optimal when U is.

Now suppose we merely know that L indirectly elicits some finite property $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$. In § D, we give Construction 2, which is the same as Construction 1 but with the following modification of R_U to $\hat{R}_U$. Let $\Gamma = \text{prop}[L]$ and $\mathcal{U} = \{\Gamma(p) \mid p \in \Delta_\mathcal{Y}\}$ as before. Then for all $U \in \mathcal{U}$, we define $\hat{R}_U := \{r \in \mathcal{R} \mid \Gamma_U \subseteq \gamma_r\}$, where $\Gamma_U := \{p \in \Delta_\mathcal{Y} \mid U = \Gamma(p)\}$ is the set of distributions for which U is the surrogate optimal set. In words, $\hat{R}_U$ is the set of reports r which may be linked to from points in U, in the sense that U being L-optimal implies r is ℓ-optimal. For the special case where L embeds ℓ, it is straightforward to verify that $R_U = \hat{R}_U \cap \mathcal{S}$. As a result, Construction 1 is the special case of Construction 2 where one is given an embedding and restricts to the representative set $\mathcal{S}$ (Lemma 49).

The main result of § D is that, if a polyhedral surrogate indirectly elicits a finite property, then for small enough $\epsilon > 0$, Construction 2 always produces a link (Proposition 56). Combined with the fact that, by design, the construction and our choice of $\hat{R}_U$ enforce separation, we have the following.

Proposition 33 *Let L be a polyhedral surrogate which indirectly elicits a finite property γ. Then there exists $\epsilon_0 > 0$ such that for all $0 < \epsilon \leq \epsilon_0$, Construction 2 for $L, \gamma, \epsilon, \|\cdot\|_\infty$ produces a separated link from* $\text{prop}[L]$ *to γ.*

Since separation is equivalent to calibration for polyhedral surrogates (Theorem 17), we now have Theorem 32: indirect elicitation implies calibration for polyhedral surrogates.

In fact, we can show something stronger: $\hat{R}_U$ enforces separation exactly, and therefore every possible calibrated link must arise from Construction 2.

Theorem 34 *A link ψ is calibrated for a given polyhedral surrogate L and discrete target ℓ if and only if there exists $\epsilon > 0$ such that ψ is produced by Construction 2 for $L, \text{prop}[\ell], \epsilon, \|\cdot\|$.*

7.2 Centrality of embeddings

To derive another proof of Theorem 32, we now show that, for polyhedral surrogates, indirect elicitation must always pass through an embedding. That is, if L indirectly elicits γ, then there is some loss ℓ which L embeds, such that ℓ indirectly elicits γ. This result holds more generally whenever L has a finite representative set, as in § 6.

Lemma 35 *Let $L : \mathbb{R}^d \to \mathbb{R}^\mathcal{Y}_+$ be polyhedral. Then L indirectly elicits a property γ if and only if L tightly embeds a discrete loss ℓ that indirectly elicits γ.*

Proof Let $\Gamma = \text{prop}[L]$. From Lemma 20(8), L tightly embeds a discrete loss. Furthermore, Lemma 20(4,7,8) implies that L indirectly elicits $\hat{\gamma} := \text{prop}[\ell]$ for any discrete loss ℓ that L tightly embeds. The link is any function $\hat{\psi} : u \mapsto r$ such that $\Gamma_u \subseteq \hat{\gamma}_r$ for all $u \in \mathbb{R}^d$.

We will prove the stronger statement that, for any property γ, and any loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ that L tightly embeds, L indirectly elicits γ if and only if ℓ indirectly elicits γ. If ℓ indirectly elicits γ via the link ψ, then L indirectly elicits γ by transitivity of subset inclusion, as $\Gamma_u \subseteq \hat{\gamma}_{\hat{\psi}(u)} \subseteq \gamma_{\psi \circ \hat{\psi}(u)}$ for all $u \in \mathbb{R}^d$. Conversely, suppose L indirectly elicits γ via the link ψ. As L tightly embeds ℓ, from Lemma 20(4,8), the level sets of $\hat{\gamma}$ are contained in the set $\{\Gamma_u \mid u \in \mathbb{R}^d\}$. Letting the map $\hat{\psi} : \mathcal{R} \to \mathbb{R}^d$ exhibit this containment, we have $\hat{\gamma}_r = \Gamma_{\hat{\psi}(r)} \subseteq \gamma_{\psi \circ \hat{\psi}(r)}$ for all $r \in \mathcal{R}$. ■

Proof [Alternate proof of Theorem 32] Let $\mathcal{R}$ be the range of γ, so that $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$, and let ℓ elicit γ. By Lemma 35, L tightly embeds a discrete loss $\hat{\ell} : \hat{\mathcal{R}} \to \mathbb{R}^{\mathcal{Y}}_+$ such that $\hat{\ell}$ indirectly elicits γ; let $\psi^{\mathcal{R}} : \hat{\mathcal{R}} \to \mathcal{R}$ be the corresponding link function. Let $\hat{\gamma} := \text{prop}[\hat{\ell}]$ be the property that $\hat{\ell}$ directly elicits. Then for all $r \in \hat{\mathcal{R}}$ and $p \in \Delta_{\mathcal{Y}}$ we have $r \in \hat{\gamma}(p) \implies \psi^{\mathcal{R}}(r) \in \gamma(p)$. Moreover, Construction 1 gives a link function $\hat{\psi} : \mathbb{R}^d \to \hat{\mathcal{R}}$ such that $(L, \hat{\psi})$ is calibrated with respect to $\hat{\ell}$.

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

$$\{u \in \mathbb{R}^d \mid \psi(u) \notin \gamma(p)\} \subseteq \{u \in \mathbb{R}^d \mid \hat{\psi}(u) \notin \hat{\gamma}(p)\} \ .$$

Combined with the fact that $(L, \hat{\psi})$ is calibrated with respect to $\hat{\ell}$, we have

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

showing calibration of ψ. ■

8. Conclusion

In this work, we introduce an embedding framework to design and analyze consistent, convex surrogates for discrete prediction tasks. Our results are constructive; as we outline in § 5, they can be fruitfully applied to a range of tasks, from designing new surrogates and link functions to understanding the consistency or inconsistency of existing surrogates. Beyond these tools, our results shed light on fundamental questions about the design of consistent surrogates.

Perhaps the most pressing open direction is simply to apply our framework to prediction problems of interest. We hope that the discussion in § 5.1, and the detailed examples in subsequent works, serve as useful guidelines for doing so. A particularly promising domain to apply our framework is structured prediction, where relatively few consistent surrogates are known. Indeed, our framework has already been applied to submodular structured problems (§ 5.3) and to max-margin losses (Nowak et al., 2022).

Beyond applying our framework, we see several interesting directions for theoretical research. Below we outline several such directions.

Prediction dimension It can be 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 convex elicitation complexity (Frongillo and Kash, 2021). Theorem 15 constructs a consistent surrogate for any discrete loss, with prediction dimension $d = n := |\mathcal{Y}|$. In some settings, such as structured prediction and information retrieval, a prediction dimension of $d = n$ can be prohibitively large. For example, in § 5.3 we discuss structured problems which decompose as k simple subproblems, like pixel classification for image segmentation. The Lovász hinge has prediction dimension k for this problem, whereas our construction would give one with $d = n = 2^k$, an impractical number even for relatively small images. While one could achieve $d = n - 1$ with a simple modification to our construction,[7] it is unclear when and how the prediction dimension could be further lowered.

Beyond studying convex elicitation complexity directly (Ramaswamy and Agarwal, 2016; Finocchiaro et al., 2021; Frongillo and Kash, 2021), one promising approach to this question is to first understand the minimum d for which a polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ embeds ℓ, called the *embedding dimension* of ℓ, and then relate this dimension to polyhedral, or general convex, elicitation complexity. One reason this approach may be fruitful is that embeddings have much more structure than general convex losses, such as the fact that calibrated links arise automatically (Theorem 18). Yet from § 7 and similar observations, it may well be that the lowest possible prediction dimension is achieved by an embedding.

In previous work, we introduce and present some bounds for embedding dimension based on optimality conditions (Finocchiaro et al., 2020). We show in particular that a target loss has embedding dimension 1 if and only if it hase convex elicitation complexity 1, underscoring the possibility that these quantities may be the same for all discrete losses. It is unclear if these bounds are tight or if they can be improved by leveraging information about adjacent level sets of an embedded property. Moreover, beyond the fact that the embedding dimension upper bounds convex elicitation complexity, it remains to understand the relationship between these two quantities in dimensions greater than 1.

Polyhedral vs. smooth surrogates The literature on convex surrogates focuses mainly on smooth surrogate losses (Crammer and Singer, 2001; Bartlett et al., 2006; Bartlett and Wegkamp, 2008; Duchi et al., 2018; Williamson et al., 2016; Reid and Williamson, 2010; Menon et al., 2019; Zhang et al., 2020; Bao et al., 2020). In practice, minimizing such surrogates often implicitly fits a model to the full conditional label distributions. On the other hand, Ramaswamy et al. (2018, Section 1.2) contend that optimizing nonsmooth losses may enable reduction of the prediction dimension while maintaining consistency relative to smooth losses, improving downstream efficiency of the learning algorithm.[8] While generalization rates may suffer for nonsmooth losses, polyhedral surrogates achieve linear regret transfer bounds (§ 4.3), so the target generalization rates may remain the same; see also Frongillo and Waggoner (2021). Even further, Lapin et al. (2016) suggest that optimizing a nonsmooth loss that directly captures the target problem of interest, rather than a smooth one that implicitly fits to the full conditional label distributions, can improve performance in limited data settings. We would like to verify this intuition, with specific cases or broad results comparing smooth and polyhedral losses.

7. One can always reduce to $d = n - 1$ in Theorem 15 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.
8. Polyhedral losses may be more challenging to optimize in some cases than smooth losses, so the prediction dimension may need to be much smaller, as in the BEP surrogate, until one sees a computational benefit.

Indirect elicitation as a condition for consistency It is well-known that consistency is equivalent to calibration (Definition 6) for discrete target problems. As calibration is a much easier condition to work with, and in particular only involves the conditional distributions, calibration is the main tool in the literature on consistency for discrete target problems. It is easy to verify that calibration in turn implies indirect elicitation, meaning that exact minimizers of the surrogate loss are linked to exact minimizers of the target. In § 7, we show that, when restricting to the class of polyhedral surrogates, indirect elicitation is actually *equivalent* to calibration, and therefore consistency. As indirect elicitation is an even simpler condition than calibration, an important line of future work is to identify other classes of surrogates for which this equivalence holds.

Extensions to $\mathcal{H}$-consistency Throughout, we rely heavily on the fact that calibration is equivalent to consistency when the hypothesis class $\mathcal{H}$ in question is the set of all measurable functions. Calibration is no longer equivalent to consistency, however, when the hypothesis class $\mathcal{H}$ is restricted. Consistency in this case is called *$\mathcal{H}$-consistency*. In classification, calibration can fail to be sufficent for $\mathcal{H}$-consistency even in the realizable setting (Long and Servedio, 2013; Kuznetsov et al., 2014; Awasthi et al., 2021a,b). Realizability and similar assumptions often rule out the conditional label distributions causing inconsistency issues in § 5. Under such assumptions, $\mathcal{H}$-consistency essentially reduces to an alignment of the surrogate hypothesis class and the target class $\mathcal{H}$, as mediated by the link function (Zhang and Agarwal, 2020). We expect that our link construction (§ 4.2) and characterization (§ 7.1) could help extend these results beyond classification when using polyhedral surrogates. Work has also begun to derive surrogate regret bounds for restricted classes $\mathcal{H}$, called *$\mathcal{H}$-consistency bounds* (Awasthi et al., 2022a,b). Here we expect that the linear surrogate regret bounds we derive (§ 4.3) could be applied.

Superprediction sets An interesting direction is to understand consistent surrogates by studying their superprediction sets, as has been done for proper losses (Williamson, 2014). The superprediction set of a loss is the set of loss vectors weakly dominated by the range of the loss: $\{v \in \mathbb{R}^{\mathcal{Y}} \mid \exists r\ L(r) \leq v\}$, where the inequality holds pointwise. One appealing aspect of the superprediction set is that it ignores the surrogate reports and focuses directly on the set of loss vectors, in a similar fashion to the trim operation in § 6.3. In particular, taking inspiration from Ramaswamy and Agarwal (2016), it may be that questions about the required prediction dimension (see above) could be more readily answered by trying to find low-dimensional structures in the superprediction set of the target loss.

Convex envelope Finally, recall that we motivated the idea of an embedding as a way to "convexify" a discrete loss. It is not clear, however, how embeddings relate to the convex envelope operation, which is perhaps the most direct way to perform this convexification given the map φ. For example, suppose $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ embeds $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ via the embedding $\varphi : \mathcal{R} \to \mathbb{R}^d$, and consider the (polyhedral) surrogate $L' : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ given by $L'_y = (L_y + \mathbb{1}_{\varphi(\mathcal{R})})^{**}$, where here $\mathbb{1}$ denotes the convex indicator and $(\cdot)^{**}$ the biconjugate. (One might also consider similar operations that keep L' finite-valued.) When is it the case that L' also embeds ℓ? Conversely, we would like to know when the construction in Theorem 15 can be viewed as a convex envelope.

Acknowledgements

We thank Arpit Agarwal and Peter Bartlett for many early discussions and insights, Stephen Becker for a reference to Hoffman constants, and Dhamma Kimpara for observations that led to § E. We thank Drona Khurana, Nishant Mehta, Enrique Nueve, and Anish Thilagar for other suggestions. This material is based upon work supported by the National Science Foundation under Grant Nos. CCF-1657598, IIS-2045347, and DGE-1650115.

References

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`.

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`.

Kaiser Asif, Wei Xing, Sima Behpour, and Brian D Ziebart. Adversarial cost-sensitive classification. In *Conference on Uncertainty in Artificial Intelligence*, pages 92–101, 2015.

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`.

Pranjal Awasthi, Natalie Frank, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Calibration and consistency of adversarial surrogate losses. *Advances in Neural Information Processing Systems*, 34:9804–9815, 2021a.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. A finer calibration analysis for adversarial robustness. *arXiv preprint arXiv:2105.01550*, 2021b.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pages 1117–1174. PMLR, 2022a.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-class h-consistency bounds. *Advances in neural information processing systems*, 35:782–795, 2022b.

Han Bao, Clayton Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. *The Conference on Learning Theory*, 2020.

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.

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`.

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`.

S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016.

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.

John Duchi, Khashayar Khosravi, and Feng Ruan. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018.

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`.

Farzan Farnia and David Tse. A minimax approach to supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Rizal Fathony, Anqi Liu, Kaiser Asif, and Brian Ziebart. Adversarial multiclass classification: A risk minimization perspective. *Advances in Neural Information Processing Systems*, 29, 2016.

Rizal Fathony, Kaiser Asif, Anqi Liu, Mohammad Ali Bashiri, Wei Xing, Sima Behpour, Xinhua Zhang, and Brian D Ziebart. Consistent robust adversarial prediction for general multiclass classification. *arXiv preprint arXiv:1812.07526*, 2018.

Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. Embedding dimension of polyhedral losses. *The Conference on Learning Theory*, 2020.

Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. Unifying lower bounds on prediction dimension of convex surrogates. *Advances in Neural Information Processing Systems*, 34, 2021.

Jessie Finocchiaro, Rafael Frongillo, Emma Goodwill, and Anish Thilagar. Consistent polyhedral surrogates for top-k classification and variants. *International Conference on Machine Learning*, 2022a.

Jessie Finocchiaro, Rafael Frongillo, and Enrique Nueve. The structured abstain problem and the lovász hinge. *Conference on Learning Theory*, 2022b.

Rafael Frongillo and Ian Kash. General truthfulness characterizations via convex analysis. In *Web and Internet Economics*, pages 354–370. Springer, 2014.

Rafael Frongillo and Ian Kash. Vector-Valued Property Elicitation. In *Proceedings of the 28th Conference on Learning Theory*, pages 1–18, 2015a.

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

Rafael Frongillo and Bo Waggoner. Surrogate regret bounds for polyhedral losses. *Advances in Neural Information Processing Systems*, 34, 2021.

Rafael M Frongillo and Ian A Kash. Elicitation complexity of statistical properties. *Biometrika*, 108(4):857–879, 2021.

Jean Gallier. Notes on convex sets, polytopes, polyhedra, combinatorial topology, voronoi diagrams and delaunay triangulations, 2008.

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.

T. Gneiting. Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

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.

Alan J Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4), 1952.

Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Multi-class deep boosting. *Advances in Neural Information Processing Systems*, 27, 2014.

Nicolas S. Lambert. Elicitation and evaluation of statistical forecasts. *Resubmitted to* Econometrica, 2018. URL `https://web.stanford.edu/~nlambert/papers/elicitability.pdf`.

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.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Top-k multiclass svm. In *Advances in Neural Information Processing Systems*, pages 325–333, 2015.

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.

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.

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.

Phil Long and Rocco Servedio. Consistency versus realizable h-consistency for multiclass classification. In *International Conference on Machine Learning*, pages 801–809. PMLR, 2013.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31, 2018.

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.

Alex Nowak, Alessandro Rudi, and Francis Bach. On the consistency of max-margin losses. In *International Conference on Artificial Intelligence and Statistics*, pages 4612–4633. PMLR, 2022.

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.

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.

Harish Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Convex calibrated surrogates for hierarchical classification. In *International Conference on Machine Learning*, pages 1852–1860, 2015.

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, and Shivani Agarwal. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1): 530–554, 2018.

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.

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.

Mark D Reid, Robert C Williamson, and Peng Sun. The convexity and design of composite multiclass losses. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 243–250, 2012.

M.D. Reid and R.C. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 9999:2387–2422, 2010.

Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.

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.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, September 2008. ISBN 978-0-387-77242-4. Google-Books-ID: HUnqnrpYt4IC.

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.

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`.

Yutong Wang and Clayton Scott. Weston-watkins hinge loss and ordered partitions. *Advances in neural information processing systems*, 2020.

Jason Weston and Chris Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the 7th European Symposium on Artificial Neural Networks*, 1999.

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`.

Robert C Williamson, Elodie Vernet, and Mark D Reid. Composite multiclass losses. *Journal of Machine Learning Research*, 17(223):1–52, 2016.

Forest Yang and Sanmi Koyejo. On the consistency of top-k surrogate losses. In *International Conference on Machine Learning*, pages 10727–10735. PMLR, 2020.

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.

Constantin Zalinescu. Sharp estimates for hoffman's constant for systems of linear inequalities and equalities. *SIAM Journal on Optimization*, 14(2):517–533, 2003.

Mingyuan Zhang and Shivani Agarwal. Bayes consistency vs. h-consistency: The interplay between surrogate loss functions and the scoring function class. *Advances in neural information processing systems*, 33:16927–16936, 2020.

Mingyuan Zhang, Harish G Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label f-measure. *International Conference on Machine Learning*, 2020.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

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

Appendix A. Power diagrams

We begin with several definitions from Aurenhammer Aurenhammer (1987).

Definition 36 *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 37 *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*

$$\text{cell}(s_i) = \{x \in \mathbb{R}^d : \forall j \in \{1, \dots, k\} \, \|x - s_i\|^2 - w_i \leq \|x - s_j\|^2 - w_j\} \ . \tag{11}$$

Definition 38 *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.*

Some of the convex polyhedra we study are the (negative) Bayes risks of loss functions, whose projections onto $\Delta_{\mathcal{Y}}$ form the level sets of the property they elicit. The following result from Aurenhammer (1987) therefore immediately applies to elicitable properties. We also make use of the same structure for the loss itself; 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 either case, we refer to the resulting power diagram as being *induced* by the convex function.

Theorem 39 (Aurenhammer (Aurenhammer, 1987)) *A cell complex is affinely equivalent to a convex polyhedron if and only if it is a power diagram.*

We extend Theorem 39 to a weighted sum of convex functions, showing that the induced power diagram is the same for any choice of strictly positive weights.

Lemma 40 *Let $f_1, \dots, f_m : \mathbb{R}^d \to \mathbb{R}$ be polyhedral convex functions. The power diagram induced by $\sum_{i=1}^m p_i f_i$ is the same for all $p \in \text{inter}(\Delta_{\mathcal{Y}})$.*

Proof For any polyhedral convex function g with epigraph P, the proof of Aurenhammer (1987, 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 $\text{inter}(F')$ 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 Rockafellar (1997), f, f' are polyhedral. We now show that f is differentiable whenever f' is differentiable:

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

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 13.

Lemma 41 *Let $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ be a polyhedral loss. Then L is minimizable and elicits a property $\Gamma := \text{prop}[L]$. Moreover, 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 0), and thus its infimum is finite. Therefore, we can apply Rockafellar (1997, Corollary 19.3.1) to conclude that its infimum is attained for all $p \in \Delta_{\mathcal{Y}}$ and is therefore minimizable. Thus, L 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 40, 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 $p \in \text{inter}(\Delta_{\mathcal{Y}})$. 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 $p \in \text{inter}(\Delta_{\mathcal{Y}})$.

We claim for all $p \in \text{inter}(\Delta_{\mathcal{Y}})$, 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 $p \notin \text{inter}(\Delta_{\mathcal{Y}})$, 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 $\mathcal{Y}' \subseteq \mathcal{Y}$. ■

Appendix B. 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 17, that calibration is equivalent to separation (Definition 16); and Theorem 18, existence of a separated link. Theorem 17 is proven in this section and Theorem 18 is proven in Appendix D.

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* $R_\ell(r, p) = \langle p, \ell(r)\rangle - \underline{\ell}(p)$. We will use these functions again when we prove surrogate regret bounds (§ C).

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 (2007, Lemma 6).

Lemma 42 *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* $\text{prop}[L]$ *and* $\text{prop}[\ell]$.

Proof Let $\Gamma := \text{prop}[L]$ and $\gamma := \text{prop}[\ell]$. 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 13). 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

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

contradicting the calibration of ψ. ■

For the other direction, we will make use of Hoffman constants for systems of linear inequalities. See Zalinescu (2003) for a modern treatment.

Theorem 43 (Hoffman constant Hoffman (1952)) *Given a matrix $A \in \mathbb{R}^{m \times n}$, there exists some smallest $H(A) \geq 0$, called the* Hoffman constant *(with respect to $\|\cdot\|_\infty$), such that for all $b \in \mathbb{R}^m$ and all $x \in \mathbb{R}^n$,*

$$d_\infty(x, S(A,b)) \leq H(A) \|(Ax - b)_+\|_\infty \ , \tag{12}$$

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

Lemma 44 *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral loss with $\Gamma = \text{prop}[L]$. Then for any fixed p, there exists some smallest constant $H_{L,p} \geq 0$ such that $d_\infty(u, \Gamma(p)) \leq H_{L,p} R_L(u,p)$ 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 $\langle p, L(u)\rangle = \max_{1 \leq j \leq m} a_j \cdot u + c_j$. 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

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

Similarly, we have $\max_i \, (Au - b)_i = \langle p, L(u)\rangle - \underline{L}(p) = R_L(u,p) \geq 0$. 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 43, we have

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

■

Given discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, define the constant $C_\ell = \max_{r, r' \in \mathcal{R}, y \in \mathcal{Y}} \ell(r)_y - \ell(r')_y$. We are now ready to prove Theorem 17.

Theorem 17 *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 $\gamma = \text{prop}[\ell]$ and $\Gamma = \text{prop}[L]$. From Lemma 42, 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 44, $\psi(u) \notin \gamma(p)$ implies

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

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

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 $R_\ell(\psi(u),p) \geq C'_{\ell,p}$. Rearranging, we have

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

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. ■

Appendix C. Surrogate Regret Bounds

C.1 Proof of Theorem 19

Lemma 45 *Suppose (L,ψ) indirectly elicits ℓ and let $\Gamma = \text{prop}[L]$. Then for any fixed $u, u^* \in \mathbb{R}^d$ and $r \in \mathcal{R}$, the functions $R_L(u,\cdot)$ and $R_\ell(r,\cdot)$ are linear in their second arguments on Γ_{u^*}.*

Proof Let $u^* \in \mathbb{R}^d$ and $p \in \Gamma_{u^*}$. By definition, for all $p \in \Gamma_{u^*}$, $\underline{L}(p) = \langle p, L(u^*)\rangle$. So for fixed u,

$$R_L(u,p) = \langle p, L(u)\rangle - \langle p, L(u^*)\rangle = \langle p, L(u) - L(u^*)\rangle,$$

a linear function of p on Γ_{u^*}. Next, by the definition of indirect elicitation, there exists r^* such that $\Gamma_{u^*} \subseteq \gamma_{r^*}$. By the same argument, for fixed r, $R_\ell(r,p) = \langle p, \ell(r) - \ell(r^*)\rangle$, a linear function of p on γ_{r^*} and thus on Γ_{u^*}. ■

Recall the definitions of C_ℓ and $H_{L,p}$ from § B.

Lemma 46 *Let $\ell : \mathcal{R} \to \mathbb{R}^\mathcal{Y}_+$ be a discrete target loss, $L : \mathbb{R}^d \to \mathbb{R}^\mathcal{Y}_+$ be a polyhedral surrogate loss, and $\psi : \mathbb{R}^d \to \mathcal{R}$ a link function. If (L,ψ) indirectly elicit ℓ and ψ is ϵ-separated, then for all u and p,*

$$R_\ell(\psi(u),p) \leq \frac{C_\ell H_{L,p}}{\epsilon} R_L(u,p).$$

Proof If $\psi(u) \in \gamma(p)$, then $R_\ell(u,p) = 0$ and we are done. Otherwise, applying the definition of ϵ-separated and Lemma 44,

$$\begin{aligned}\epsilon &< d_\infty(u, \Gamma(p)) \\ &\leq H_{L,p} R_L(u,p).\end{aligned}$$

So $R_\ell(\psi(u),p) \leq C_\ell \leq \frac{C_\ell H_{L,p}}{\epsilon} R_L(u,p)$. ■

We can now restate and prove Theorem 19.

Theorem 47 (Theorem 19) *Let $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ be discrete, $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ polyhedral, and $\psi : \mathbb{R}^d \to \mathcal{R}$. If (L, ψ) are consistent for ℓ, then there exists constants $\epsilon_\psi, H_L > 0$ such that*

$$(\forall h, \mathcal{D}) \quad R_\ell(\psi \circ h; \mathcal{D}) \leq \frac{C_\ell H_L}{\epsilon_\psi} R_L(h; \mathcal{D}) \ .$$

Proof Let $\Gamma = \text{prop}[L]$. From Lemma 20, there is a finite set $U \subset \mathbb{R}^d$ of predictions such that (a) for each $u \in U$, the level set Γ_u is a polytope (see e.g. Lemma 75), and (b) $\cup_{u \in U} \Gamma_u = \Delta_{\mathcal{Y}}$. For each $u \in U$ let $\mathcal{Q}_u \subset \Delta_{\mathcal{Y}}$ be the finite set of vertices of the polytope Γ_u, and define the finite set $\mathcal{Q} = \cup_{u \in U} \mathcal{Q}_u$. Let $H_L := \max_{q \in \mathcal{Q}} H_{L,q}$.

By Lemma 42, ψ is ϵ-separated for some $\epsilon > 0$; let $\epsilon_\psi = \epsilon$. By Lemma 46, for each $q \in \mathcal{Q}$,

$$R_\ell(\psi(u), q) \leq \frac{C_\ell H_{L,q}}{\epsilon_\psi} R_L(u,q) \leq \frac{C_\ell H_L}{\epsilon_\psi} R_L(u,q)$$

for all $u \in \mathbb{R}^d$. Now consider a general $p \in \Delta_{\mathcal{Y}}$, which is in some full-dimensional polytope level set Γ_u. Write $p = \sum_{q \in \mathcal{Q}_u} \beta(q) q$ for some convex combination $\beta \in \Delta_{\mathcal{Q}_u}$. By Lemma 45, R_L and R_ℓ are linear in the second argument on Γ_u, so for any $u' \in \mathbb{R}^d$,

$$\begin{aligned}R_\ell(\psi(u'), p) &= \sum_{q \in \mathcal{Q}_u} \beta(q) R_\ell(\psi(u'), q) \\ &\leq \sum_{q \in \mathcal{Q}_u} \beta(q) \frac{C_\ell H_L}{\epsilon_\psi} R_L(u', q) \\ &\leq \frac{C_\ell H_L}{\epsilon_\psi} \sum_{q \in \mathcal{Q}_u} \beta(q) R_L(u', q) \\ &= \frac{C_\ell H_L}{\epsilon_\psi} R_L(u', p).\end{aligned}$$

The result for $\mathcal{D}$ now holds by linearity of expectation over $\mathcal{D}$. ■

C.2 Tighter bounds

Our goal in proving Theorem 19 is to show a broad result that consistent polyhedral losses always yield linear regret bounds. As one may expect given the generality of the result,

however, the specific constant we derive may be loose in some cases. We now discuss some techniques to further tighten the constant.

Let us consider the tightest possible constant c^* for which $R_\ell(\psi \circ h; \mathcal{D}) \leq c^* R_L(h; \mathcal{D})$ for all h and $\mathcal{D}$. In general, for a fixed p, there is some smallest c_p^* such that $R_\ell(\psi(u), p) \leq c_p^* R_L(u, p)$ for all u. It therefore follows from our results that $c^* = \max_{p \in \mathcal{Q}} c_p^*$ for the finite set $\mathcal{Q}$ used in the proof, i.e., the vertices of the full-dimensional level sets of $\Gamma = \text{prop}[L]$.

Above, we bounded $c_p^* \leq \frac{C_\ell H_{L,p}}{\epsilon_\psi}$. The intuition is that some u at distance $\geq \epsilon_\psi$ from $\Gamma(p)$, the optimal set, may link to a "bad" report $r = \psi(u) \notin \gamma(p)$. The rate at which L grows is at least $H_{L,p}$, so the surrogate loss at u may be as small as $\frac{\epsilon_\psi}{H_{L,p}}$, while the target regret may be as high as $C_\ell = \max_{r',p'} R_\ell(r', p')$. The ratio of regrets is therefore bounded by $\frac{H_{L,p} C_\ell}{\epsilon_\psi}$.

The tightest possible bound, on the other hand, is $c_p^* = \sup_{u:\psi(u) \notin \gamma(p)} \frac{R_\ell(\psi(u),p)}{R_L(u,p)}$. This bound can be smaller if the values of numerator and denominator are correlated across u. For example, u may only be ϵ_ψ-close to the optimal set when it links to reports $\psi(u)$ with lower target regret; or L may have a smaller slope in the direction where the link's separation is larger than ϵ.

To illustrate with a concrete example, consider the *binary encoded predictions (BEP) surrogate* of Ramaswamy et al. (2018), which we discuss in § 5.2. The target loss here is the abstain loss, $\ell(r, y) = \frac{1}{2}$ if $r = \perp$, otherwise $\ell(r, y) = \mathbb{1}\{r \neq y\}$. Letting $d = \lceil \log_2 |\mathcal{Y}| \rceil$, the BEP surrogate $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ is given by $L(u)_y = \max_{j \in [d]} (1 - \varphi(y)_j u_j)_+$, where $\varphi : \mathcal{Y} \to \{-1, 1\}^d$ is an injection. The associated link is $\psi(u) = \perp$ if $\|u\|_\infty \leq \frac{1}{2}$, otherwise $\psi(u) = \arg\min_{y \in \mathcal{Y}} \|B(y) - u\|_\infty$.

One can show for $p = \delta_y$, the distribution with full support on some $y \in \mathcal{Y}$, that $L(u)_y = d_\infty(u, \Gamma(p))$ exactly, giving $H_{L,p} = 1$. It is almost immediate that $\epsilon_\psi = \frac{1}{2}$. Meanwhile, $R_\ell(r, p) \leq 1$, giving us an upper bound $c_p^* \leq \frac{(1)(1)}{1/2} = 2$. The exact constant as given by Ramaswamy et al., however, is $c^* = 1$. The looseness stems from the fact that for $p = \delta_y$, the closest reports u to the optimal set, i.e., at distance only $\epsilon_\psi = \frac{1}{2}$ away, do not link to reports maximizing target regret; they link to the abstain report $\perp$, which has regret only $\frac{1}{2}$. With this correction, and an observation that all u linking to reports $y' \neq y$ are at distance at least $\frac{3}{2}$ from $\Gamma(p)$, we restore the tight bound $c_p^* \leq 1$. A similar but slightly more involved calculation can be carried out for the other vertices $p \in \mathcal{Q}$, which turn out to be all vertices of the form $\frac{1}{2}\delta_y + \frac{1}{2}\delta_{y'}$.

Finally, while we use $\|\cdot\|_\infty$ to define the minimum-slope H_L and the separation ϵ_ψ, in principle one could use another norm. One reason for restricting to $\|\cdot\|_\infty$ is that it is more compatible with Hoffman constants. However, all definitions hold for other norms and so does the main upper bound, as existence of an H_L and ϵ_ψ in $\|\cdot\|_\infty$ imply existence of constants for other norms. These constants may change for different norms, and in particular, the optimal overall constant may arise from a norm other than $\|\cdot\|_\infty$.

Appendix D. Existence of a Separated Link

In this section, we prove Theorem 18 from § 4, as discussed at the beginning of Appendix B: embeddings give rise to separated links. The crux of the proof is showing that embeddings imply eq. (4), the intersection condition on optimal sets, and that this condition is sufficient

for the construction to produce a link. Calibration then follows by the fact every link produced by Construction 1 is separated, and therefore calibrated.

In fact, we will show that the approach outlined above also suffices for the more general case where the given polyhedral surrogate indirectly elicits a given finite property. This more general setting will allow us to prove the results from § 7, and in particular, that indirect elicitation implies calibration for polyhedral surrogates.

To relate back to embeddings, we split the first phase into two: (i) an embedding is a special case of indirect elicitation (Lemma 50), and (ii) indirect elicitation is equivalent to the intersection condition (Lemma 51). We will then reason instead about Construction 2, a generalization of Construction 1 for indirect elicitation.

D.1 A more general construction

As described in § 7, the task of generalizing Construction 1 beyond embeddings reduces to carefully generalizing the definition of R_U. Informally, R_U in Construction 1 is the set of target reports which must be ℓ-optimal whenever U is L-optimal. There we define R_U simply as $\{r \in \mathcal{S} \mid \varphi(r) \in U\}$ where $\mathcal{S}$ is the given representative set. For the more general case, we can define R_U as follows.

Definition 48 *For polyhedral loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, and finite propert $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$, we will define*

- $\Gamma = \text{prop}[L]$,
- $\mathcal{U} = \{\Gamma(p) \mid p \in \Delta_{\mathcal{Y}}\}$,
- $\Gamma_U := \{p \in \Delta_{\mathcal{Y}} \mid U = \Gamma(p)\}$ *for all* $U \in \mathcal{U}$,
- $R_U := \{r \in \mathcal{R} \mid \Gamma_U \subseteq \gamma_r\}$ *for all* $U \in \mathcal{U}$.

Construction 2 is essentially the same as Construction 1 but with the definition of R_U above.

Construction 2 (General ϵ-thickened link) *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$, $\epsilon > 0$, and a norm $\|\cdot\|$ be given, such that L is polyhedral and indirectly elicits γ. Let $\mathcal{U}$ and R_U be defined as in Definition 48. The* ϵ-thickened link *ψ is constructed as follows. First, initialize the* link envelope *$\Psi : \mathbb{R}^d \to 2^{\mathcal{R}}$ 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 $\Psi(u) = \Psi(u) \cap R_U$. If we have $\Psi(u) \neq \emptyset$ for all $u \in \mathbb{R}^d$, then the construction* produces a link *$\psi \in \Psi$ pointwise, breaking ties arbitrarily.*

It is straightforward to show that Construction 1 is a special case of Construction 2.

Lemma 49 *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral surrogate which embeds $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ via the representative set $\mathcal{S} \subseteq \mathcal{R}$ and embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$. Then for any $\epsilon > 0$ and norm $\|\cdot\|$, Construction 1 for $L, \ell, \epsilon, \|\cdot\|$ is equivalent to Construction 2 for $L, \text{prop}[\ell|_{\mathcal{S}}], \epsilon, \|\cdot\|$.*

Proof Let $\gamma = \text{prop}[\ell|_{\mathcal{S}}]$, which is also given by $\gamma : p \mapsto \text{prop}[\ell](p) \cap \mathcal{S}$. Let $U \in \mathcal{U}$. Then for $r \in \mathcal{S}$, we have $r \in R_U \iff \Gamma_U \subseteq \gamma_r \iff \Gamma_U \subseteq \Gamma_{\varphi(r)} \iff (\Gamma(p) = U \implies \varphi(r) \in$

$\Gamma(p)) \iff \varphi(r) \in U$. As we have $\mathcal{R} = \mathcal{S}$ in Construction 2 for $L, \text{prop}[\ell|_{\mathcal{S}}], \epsilon, \|\cdot\|$, we conclude $R_U = \{r \in \mathcal{S} \mid \varphi(r) \in U\}$, exactly as in Construction 1 for $L, \ell, \epsilon, \|\cdot\|$. The equivalence of the two constructions follows. ■

D.2 Indirect elicitation and optimal set intersection

We first show (i), that embedding is a special case of indirect elicitation.

Lemma 50 *If L embeds ℓ, then L indirectly elicits* $\text{prop}[\ell]$.

Proof Let $\Gamma = \text{prop}[L]$ and $\gamma = \text{prop}[\ell]$. From Proposition 31, we have $\text{trim}(\gamma) = \text{trim}(\Gamma) =: \Theta$. By definition of trim, for any $u \in \mathbb{R}^d$, we have some $\theta \in \Theta$ such that $\Gamma_u \subseteq \theta$. Since $\text{trim}(\gamma) = \Theta$, we have some $r \in \mathcal{R}$ such that $\theta = \gamma_r$, giving $\Gamma_u \subseteq \gamma_r$. ■

We next show (ii), the equivalence of indirect elicitation and the following intersection condition.

Lemma 51 *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be polyhedral and $\gamma : \Delta\mathcal{Y} \rightrightarrows \mathcal{R}$ be a finite property. Let $\mathcal{U}$ and R_U be defined as in Definition 48. Then L indirectly elicits γ if and only if the following condition holds*

$$\forall \mathcal{U}' \subseteq \mathcal{U},\ \cap_{U \in \mathcal{U}'} U \neq \emptyset \implies \cap_{U \in \mathcal{U}'} R_U \neq \emptyset\ . \tag{13}$$

Proof First assume L indirectly elicits γ. As $\mathcal{U}$ is the range of Γ, we have for all $u \in \mathbb{R}^d$ that $\Gamma_u = \cup\{\Gamma_U \mid U \in \mathcal{U}, u \in U\}$. Suppose $\cap_{U \in \mathcal{U}'} U \neq \emptyset$; let $u \in \cap_{U \in \mathcal{U}'} U$. As $u \in U$ for all $U \in \mathcal{U}'$, we have $\Gamma_U \subseteq \Gamma_u$ for all $U \in \mathcal{U}'$. By indirect elicitation, there exists some $r \in \mathcal{R}$ such that $\Gamma_u \subseteq \gamma_r$. Thus, for all $U \in \mathcal{U}'$, we have $\Gamma_U \subseteq \gamma_r$ and thus $r \in R_U$. We conclude $\cap_{U \in \mathcal{U}'} R_U \neq \emptyset$.

For the converse, let $u \in \mathbb{R}^d$. If $\Gamma_u = \emptyset$, then $\Gamma_u \subseteq \gamma_r$ for any $r \in \mathcal{R}$. Otherwise, $\Gamma_u \neq \emptyset$, and the set $\mathcal{U}'_u = \{U \in \mathcal{U} \mid u \in U\}$ is nonempty. Moreover, $\cap \mathcal{U}'_u \neq \emptyset$ as $u \in \cap \mathcal{U}'_u$. Eq. (13) now gives some $r \in \cap\{R_U \mid U \in \mathcal{U}'_u\}$. By definition of R_U, for all $U \in \mathcal{U}'_u$ we have $\Gamma_U \emptyset \gamma_r$. Thus $\Gamma_u = \cup\{\Gamma_U \mid U \in \mathcal{U}'_u\} \subseteq \gamma_r$, showing indirect elicitation. ■

D.3 Convex geometry for separation

Let some norm $\|\cdot\|$ on $\mathbb{R}^d$ be given. Given a set $T \subseteq \mathbb{R}^d$ and a point $u \in \mathbb{R}^d$, let $d(T, u) = \inf_{t \in T} \|t - u\|$. Given two sets $T, T' \subseteq \mathbb{R}^d$, let $d(T, T') = \inf_{t \in T, t' \in T'} \|t - t'\|$. Finally, for $T \subseteq \mathbb{R}^d$ and $\epsilon > 0$, let the "thickening" $B(T, \epsilon)$ be defined as

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

The goal of this subsection is to prove the first part of step (iii): for small enough $\epsilon > 0$, if any set of ϵ-thickened optimal sets intersect, then the optimal sets themselves must intersect. We will conclude that, for small enough ϵ, the link envelope Ψ is non-empty everywhere, meaning there will be legal choices left over for the link.

Lemma 52 *Let $\mathcal{U}$ be defined as in Definition 48. 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 $\cap_j B(U_j, \epsilon) = \emptyset$.*

The next few geometric results build to Lemma 52.

Lemma 53 *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 (Boyd and Vandenberghe, 2004, 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

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

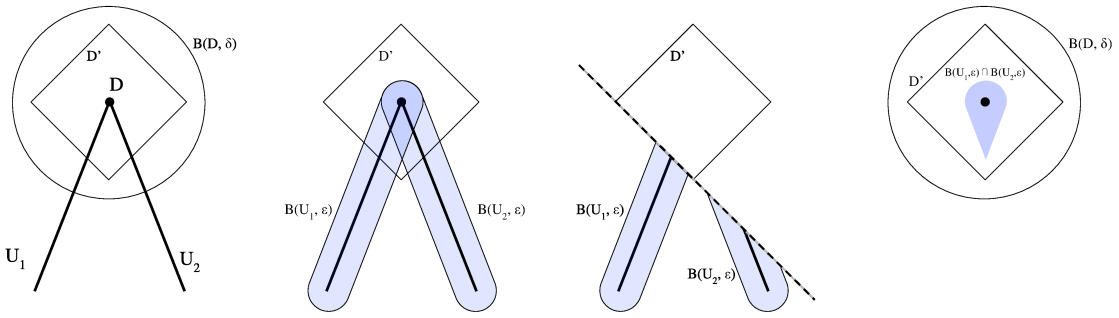
Now let D' be the interior of D'', i.e. if $D'' = \{x : Ax \leq b\}$, then we let $D' = \{x : Ax < b\}$. We retain $D' \subseteq B(D, \epsilon)$. Further, we retain $D \subseteq D'$, because D is contained in the interior of $D'' = D + S'$. (Proof: if $x \in D$, then for some γ, $x + B(\{\vec{0}\}, \gamma)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

Lemma 54 *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 $0 < \epsilon \leq \epsilon_0$, $\cap_j B(U_j, \epsilon) \subseteq B(\cap_j U_j, \delta)$.*

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 53, 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). ■

Figure 7: Illustration of a special case of the proof of Lemma 54 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 $B(D, \delta)$.



Lemma 55 *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 54, 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. ■

The proof of Lemma 52 now follows: for each $\mathcal{U}' \subseteq \mathcal{U}$, Lemma 55 gives an $\epsilon_0(\mathcal{U}') > 0$; we take the minimum of $\epsilon_0(\mathcal{U}')$ over the finitely many choices of $\mathcal{U}'$.

D.4 Separation of the general construction

We now prove the main results for link construction from § 4 and § 7. Specifically, we show that indirect elicitation implies that Construction 2 produces a link, and moreover, a link is produced if and only if it is separated. As we have established above that Construction 1 is a special case, the results specific to embeddings will follow.

Proposition 56 *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be polyhedral and $\gamma : \Delta\mathcal{Y} \rightrightarrows \mathcal{R}$ be finite. If L indirectly elicits γ, then there exists $\epsilon_0 > 0$ such that, for all $0 < \epsilon \leq \epsilon_0$, Construction 2 for L, γ, ϵ, $\|\cdot\|_\infty$, produces a link.*

Proof Fix a small enough ϵ_0 as promised by Lemma 52. Let $u \in \mathbb{R}^d$ and $\mathcal{U}'_u = \{U \in \mathcal{U} \mid d_\infty(u, U) < \epsilon\}$. From Construction 2, we have $\Psi(u) = \cap\{R_U \mid U \in \mathcal{U}'_u\}$. Since

$u \in \cap\{B(U,\epsilon) \mid U \in \mathcal{U}'_u\}$ by definition, Lemma 52 and our choice of ϵ give $\cap\mathcal{U}'_u \neq \emptyset$. By Lemma 51, we have $\Psi(u) = \cap\{R_U \mid U \in \mathcal{U}'_u\} \neq \emptyset$. ■

Perhaps surprisingly, one can also show that *every* calibrated link from a polyhedral surrogate to a discrete loss is produced by Construction 1. This result follows from the fact, stated now, that Construction 1 with $\|\cdot\|_\infty$ is exactly enforcing ϵ-separation. From Theorem 17, every calibrated link ψ is therefore output by the construction for some sufficiently small ϵ, in the sense that ψ is one of the valid choices within the link envelope.

Proposition 57 *Let polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, finite property $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$, and $\epsilon > 0$ be given. Then Construction 2 for $L, \gamma, \epsilon, \|\cdot\|_\infty$ produces a link ψ if and only if ψ is ϵ-separated.*

Proof Let $\mathcal{U}'_u = \{U \in \mathcal{U} \mid d_\infty(u, U) < \epsilon\}$. Then we have the following chain of equivalences.

$$\begin{aligned}
&\forall u \in \mathbb{R}^d, \;\; \psi(u) \in \Psi(u) \\
\iff &\forall u \in \mathbb{R}^d, U \in \mathcal{U}'_u, \;\; \psi(u) \in R_U \\
\iff &\forall u \in \mathbb{R}^d, U \in \mathcal{U}'_u, \;\; \Gamma_U \subseteq \gamma_{\psi(u)} \\
\iff &\forall u \in \mathbb{R}^d, p \in \Delta_\mathcal{Y} \text{ s.t. } \Gamma(p) \in \mathcal{U}'_u, \;\; p \in \gamma_{\psi(u)} \\
\iff &\forall u \in \mathbb{R}^d, p \in \Delta_\mathcal{Y} \text{ s.t. } d_\infty(u, \Gamma(p)) < \epsilon, \;\; \psi(u) \in \gamma(p) \\
\iff &\psi \text{ is } \epsilon\text{-separated} .
\end{aligned}$$

■

From the equivalence of calibration and separation for polyhedral surrogates (Theorem 17), we now have Theorem 34: the construction produces exactly the set of calibrated links. (The move from $\|\cdot\|_\infty$ to a general norm follows from norm equivalence in finite-dimensional vector spaces.) Combined with Proposition 56, we also have Proposition 33: if L indirectly elicits γ, the construction produces a calibrated link.

Returning to embeddings, recall that Construction 1 is a special case of Construction 2 by Lemma 49. Moreover, embeddings are a special case of indirect elicitation (Lemma 50). As Proposition 56 guarantees that Construction 2 produces a link, and Proposition 57 that every link produced is separated, we now have Theorem 18 which we restate. (Note that the converse need not hold for Construction 1, and indeed, that construction may miss some separated links which make use of reports outside $\mathcal{S}$.)

Theorem 18 *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 for $L, \ell, \epsilon, \|\cdot\|$ produces a nonempty set of links, all of which are ϵ-separated with respect to L and ℓ.*

Appendix E. Connection to Ramaswamy and Agarwal (2016)

Ramaswamy and Agarwal (2016) give an impressive array of consistency results for general prediction tasks. Among them, in their Theorem 8, is a general sufficient condition for consistency. We restate their result here in our notation.

Theorem 58 (Ramaswamy and Agarwal (2016, Theorem 8)) *Let $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ for $\mathcal{R}$ finite, and $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$. Let $\gamma = \text{prop}[\ell]$ and $\Gamma = \text{prop}[L]$. Suppose there exists a finite set $\mathcal{S} \subset \mathbb{R}^d$ such that $\bigcup_{u \in \mathcal{S}} \Gamma_u = \Delta_{\mathcal{Y}}$ and for each $u \in \mathcal{S}$ there exists $r \in \mathcal{R}$ such that $\Gamma_u \subseteq \gamma_r$. Then L is calibrated with respect to ℓ.*

In other words, if there exists a finite representative set $\mathcal{S} \subset \mathbb{R}^d$ for L such that $L|_{\mathcal{S}}$ indirectly elicits γ, then L is calibrated with respect to ℓ. Of course, if L indirectly elicits γ and has a finite representative set, then this condition holds. Thus, not only does this result prove Theorem 32 for polyhedral surrogates, it also shows that the result extends to the setting of § 6: surrogates with finite representative sets.

The crux of their result, as in ours, in the construction of a link function exhibiting calibration. Their construction is quite different from our Construction 1, in that it operates in the loss vector space $\mathbb{R}^{\mathcal{Y}}_+$ rather than the surrogate report space $\mathbb{R}^d$. Specifically, it operates on the *superprediction set* $L(\mathbb{R}^d) + \mathbb{R}^{\mathcal{Y}}_+$, where the $+$ here is Minkowski addition. An arbitrary element of the superprediction set is decomposed into a convex combination of elements of $L(\mathcal{S})$, plus an element of $\mathbb{R}^{\mathcal{Y}}_+$. This convex combination is then thresholded at $1/|\mathcal{S}|$, where the link function may output $r \in \mathcal{R}$ if the weight on any $u \in \mathcal{S}$ with $\Gamma_u \subseteq \gamma_r$ is at least $1/|\mathcal{S}|$. One can see that this construction does achieve separation, but it is less explicitly aligned with any norm in the surrogate report space. It would be interesting to further compare the two constructions.

Appendix F. Proving Lemma 20

In this section, we give a careful treatment of the results on convex polyhedra needed to prove Lemma 20.

F.1 General definitions for polyhedra

We begin with general definitions of polyhedra. See also Ziegler (2012) and Gallier (2008).

Definition 59 (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 some $(w, b) \in \mathbb{R}^d \times \mathbb{R}$.*

Definition 60 (Hyperplane) *A hyperplane is a set of the form $H_{(w,b)} := \{x \in \mathbb{R}^d \mid \langle x, w\rangle = b\}$ for some $(w, b) \in \mathbb{R}^d \times \mathbb{R}$.*

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 closed halfspaces corresponding to the hyperplane $\partial H^+ = H$.

Definition 61 (Polyhedron halfspace representation (Ziegler, 2012)) *A* polyhedron *P is an intersection of a finite set of closed halfspaces $\mathcal{H}$ presented in the form $P = \cap\mathcal{H}$. Here, we say $\mathcal{H}$ is a halfspace representation for P.*

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

Definition 62 (Supports) *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 63 (Face, facet) *Let $P \subseteq \mathbb{R}^d$ be a convex polyhedron. A* 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 $\dim(F) := \dim(\text{affhull}(F))$. A face F with $\dim(F) = \dim(\text{affhull}(P)) - 1$ is called a facet.

While one traditionally considers P to be a trivial face of itself, we exclude this case throughout.

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, as this irredundant set corresponds to the facets of a polyhedron in a natural way.

Definition 64 (Irredundant; adapted from Gallier (2008)) *Let $P = \cap\mathcal{H}$ for a finite set of closed 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 set of closed halfspaces $\mathcal{H}'$ such that $|\mathcal{H}'| < |\mathcal{H}|$, and redundant otherwise. Moreover, we call $\mathcal{H}$ irredundant for P if $\cap\mathcal{H}$ is an irredundant decomposition of P.*

Gallier (2008) shows that every d-dimensional 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 65 (Gallier (2008)) *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. As P has a finite halfspace representation, it must have a smallest halfspace representation $\mathcal{H}^*$. That is, $|\mathcal{H}^*| = \min\{|\mathcal{H}| : P = \cap\mathcal{H}, \mathcal{H} \text{ finite}\}$. As a smallest halfspace representation, $\mathcal{H}^*$ is irredundant by definition. Gallier (2008, Proposition 4.5(i)) then states that $\mathcal{H}^*$ is unique, giving (i). Additionally, (ii) is shown by (Gallier, 2008, Proposition 4.5(ii)).

It remains to show (iii). Let $\mathcal{H}$ be a finite set of closed halfspaces such that $P = \cap\mathcal{H}$. As noted in the last sentence of the proof of (Gallier, 2008, Proposition 4.5), the hyperplanes defining facets are unique: if F is a facet of P and H, H' are hyperplanes with $F = H \cap P = H' \cap P$, then it must be the case that $H = H'$. It therefore suffices to show that, for each facet F of P, there is an $H^+ \in \mathcal{H}$ such that $F = H \cap P$. Gallier (2008, Proposition 3.17) observes that, for all $x \in \partial P$, there exists some hyperplane $\hat{H}$ such that $\hat{H}$ supports P at x. Since $x \in \text{relint}(F)$ is in exactly one face of P, namely F, there must be a unique $H^+ \in \mathcal{H}$ such that $F = H \cap P$. ■

F.2 Specializing to $\mathbb{R}^{\mathcal{Y}} \times \mathbb{R}$

Within this appendix, we use some self-contained notation to work in the function graph space $\mathbb{R}^{\mathcal{Y}} \times \mathbb{R}$. 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 $d := |\mathcal{Y}| + 1$.

Given any $v \in \mathbb{R}^{\mathcal{Y}}_+$, define $H^+_v := H^+_{(v,-1)} = \{(x, c) \in \mathbb{R}^{\mathcal{Y}}_+ \times \mathbb{R} \mid \langle v, x\rangle \geq c\}$. 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 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 $\mathcal{V}$. Similarly, let $\mathcal{H}_{\mathcal{Y}} = \{H^+_y \mid y \in \mathcal{Y}\}$.

For any $S \subseteq \mathbb{R}^k$, let $\delta(\cdot \mid S) : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$ be the convex indicator function, given by $\delta(x \mid S) = 0$ if $x \in S$ and ∞ otherwise. Throughout, we will work with a concave function $g_{\mathcal{V}}$ generated by a set $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$ of the following form.

Definition 66 *Given a set $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$, define the function $g_{\mathcal{V}} : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}_+ \cup \{-\infty\}$ by*

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

We denote the hypograph of a function $g : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R} \cup \{-\infty\}$ by $\text{hypo}(g) = \{(x, c) \mid c \leq g(x)\} \subseteq \mathbb{R}^{\mathcal{Y}} \times \mathbb{R}$.

A first observation is that the region generated by the intersection of the H^+_y halfspaces restricts the hypograph $g_{\mathcal{V}}$ to be finite only on the nonnegative orthant for any $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{Y}}_+$.

Lemma 67 $\cap\mathcal{H}_{\mathcal{Y}} = \mathbb{R}^{\mathcal{Y}}_+ \times \mathbb{R}$.

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 $x \in \mathbb{R}^d_+$. ■

F.3 Hypographs of extended Bayes risks

We now apply this polyhedral perspective to the Bayes risk of a loss function, extended to "unnormalized distributions", i.e., all of $\mathbb{R}^{\mathcal{Y}}_+$. Given a minimizable loss function $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, define the 1-homogeneous extension of its Bayes risk as $\underline{L}_+ : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R} \cup \{\infty\}$, $x \mapsto \inf_{r \in \mathcal{R}} \langle x, L(r)\rangle - \delta(x \mid \mathbb{R}^{\mathcal{Y}}_+)$. In other words, letting $L(\mathcal{R}) := \{L(r) \mid r \in \mathcal{R}\} \subseteq \mathbb{R}^{\mathcal{Y}}_+$, we have $\underline{L}_+ = g_{L(\mathcal{R})}$. Observe that $L(\mathcal{R})$ and $\mathcal{H}_{L(\mathcal{R})}$ may be infinite sets.

Claim 1 *Suppose we are given a minimizable $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ with polyhedral extended risk $\underline{L}_+$. Then* $\text{hypo}(g_{L(\mathcal{R})}) = \cap(\mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{L(\mathcal{R})})$.

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

$$\begin{aligned}
(x, c) \in \text{hypo}(g_{L(\mathcal{R})}) &\iff g_{L(\mathcal{R})}(x) \geq c && \text{definition of hypograph} \\
&\iff \underline{L}_+(x) \geq c && g_{L(\mathcal{R})} = \underline{L}_+ \\
&\iff \langle v, x\rangle \geq c \ \forall v \in L(\mathcal{R}) && \text{definition of } \underline{L}_+ \\
&\iff (x, c) \in H_v^+ \ \forall v \in L(\mathcal{R}) && \text{definition of } H_v^+ \\
&\iff (x, c) \in \cap\mathcal{H}_{L(\mathcal{R})}
\end{aligned}$$

Combining the two equalities, we have $\text{hypo}(g_{L(\mathcal{R})}) = \cap(\mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{L(\mathcal{R})})$. ■

F.4 Finding the unique smallest subset of loss vectors

Lemma 68 *Consider a loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ with polyhedral extended risk $\underline{L}_+$. There is a unique irredundant set $\mathcal{H}^*$ of closed halfspaces such that* $\text{hypo}(g_{L(\mathcal{R})}) = \cap\mathcal{H}^*$. *Moreover, for each $H^+ \in \mathcal{H}^*$ and H such that $H = \partial H^+$, the face* $\text{hypo}(g_{L(\mathcal{R})}) \cap H$ *is a facet. Moreover, $\mathcal{H}_{\mathcal{Y}} \subseteq \mathcal{H}^*$.*

Proof As $g_{L(\mathcal{R})}$ is nonnegative on $\mathbb{R}^{\mathcal{Y}}_+$, the set $\text{hypo}(g_{L(\mathcal{R})})$ therefore contains $\{(x, c) \mid x \in \mathbb{R}^{\mathcal{Y}}_+, c \leq 0\}$, which is $(|\mathcal{Y}| + 1)$-dimensional. Therefore $\text{hypo}(g_{L(\mathcal{R})})$ is full-dimensional. Take $\mathcal{H}^*$ to be the unique irredundant and finite set of closed halfspaces such that $\text{hypo}(g_{L(\mathcal{R})}) = \cap\mathcal{H}^*$ from Theorem 65(i). Now, $\text{hypo}(g_{L(\mathcal{R})}) \cap H$ for any $H \in \mathcal{H}^*$ being a facet follows immediately from Theorem 65(ii) and (iii).

To show that $\mathcal{H}_{\mathcal{Y}} \subseteq \mathcal{H}^*$, it suffices from Theorem 65(ii) to show that each $F_y := H_y \cap \text{hypo}(g_{L(\mathcal{R})})$ for $y \in \mathcal{Y}$ is a facet. From Claim 1, since $H_y^+ \in \mathcal{H}_{\mathcal{Y}}$, we have $\text{hypo}(g_{L(\mathcal{R})}) \subseteq H_y^+$. We have $F_y = H_y \cap \text{hypo}(g_{L(\mathcal{R})}) = \{(x, c) \mid x \in \mathbb{R}^{\mathcal{Y}}_+, x_y = 0, c \leq g_{L(\mathcal{R})}(x)\} \supseteq \{(x, c) \mid x \in \mathbb{R}^{\mathcal{Y}}_+, x_y = 0, c \leq 0\}$ as $g_{L(\mathcal{R})}(x) \geq 0$. Thus, F_y is a nonempty face of $\text{hypo}(g_{L(\mathcal{R})})$, and contains a $|\mathcal{Y}|$-dimensional set, so must be a facet. ■

We now use the above result about the unique irredundant halfspace decomposition of $\text{hypo}(g_{\mathcal{V}})$ to observe a unique finite set of loss vectors generating these halfspaces.

Corollary 69 *Given a minimizable loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ with polyhedral extended risk, consider the unique irredundant set $\mathcal{H}^*$ given by Lemma 68. Then (i) there is a unique finite set $\mathcal{V}^* \subseteq \mathbb{R}^{\mathcal{Y}}_+$ such that $\mathcal{H}^* = \mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{\mathcal{V}^*}$. (ii) $F_v := H_v \cap \text{hypo}(g_{L(\mathcal{R})})$ is a facet of* $\text{hypo}(g_{L(\mathcal{R})})$ *for each $v \in \mathcal{V}^*$. (iii) $g_{L(\mathcal{R})}(x) = \min_{v \in \mathcal{V}^*} \langle v, x\rangle - \delta(x \mid \mathbb{R}^{\mathcal{Y}}_+) = g_{\mathcal{V}^*}(x)$.*

Proof (i) Since $\text{hypo}(g_{L(\mathcal{R})})$ is full-dimensional, the facets of $\text{hypo}(g_{L(\mathcal{R})})$ are uniquely determined by the hyperplanes H such that $H = \partial H^+$ and $H^+ \in \mathcal{H}^*$ by Lemma 68. Any facet must then be some intersection of an $H_y \cap \text{hypo}(g_{L(\mathcal{R})})$ or $H_v \cap \text{hypo}(g_{L(\mathcal{R})})$. Since $\mathcal{H}_{\mathcal{Y}} \subseteq \mathcal{H}^*$ by Lemma 68, take $\mathcal{H}_{\mathcal{V}^*} := \mathcal{H}^* \setminus \mathcal{H}_{\mathcal{Y}}$, and $\mathcal{V}^*$ the unique set generating $\mathcal{H}_{\mathcal{V}^*}$. (ii) Moreover, $H_v \in \mathcal{H}_{\mathcal{V}^*} \subseteq \mathcal{H}^*$ generates the facet $F_v = H_v \cap \text{hypo}(g_{L(\mathcal{R})})$ of $\text{hypo}(g_{L(\mathcal{R})})$ by Lemma 68. (iii) The result holds if $\text{hypo}(g_{L(\mathcal{R})}) = \text{hypo}(g_{\mathcal{V}^*})$. By construction, $\text{hypo}(g_{L(\mathcal{R})}) = \cap\mathcal{H}^* = \cap(\mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{\mathcal{V}^*}) = \{(x, c) \in \mathbb{R}^{\mathcal{Y}}_+ \times \mathbb{R} \mid \langle v^*, x\rangle \geq c$ for all

$v^* \in \mathcal{V}^*\} = \text{hypo}(g_{\mathcal{V}^*})$. The result follows from equality of the hypographs. ■

The above establishes our primary assumption for the rest of this section.

Assumption 1 *$L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ is a minimizable loss function such that $\underline{L}_+ = g_{L(\mathcal{R})}$ is polyhedral. $\mathcal{V}^* \subseteq L(\mathcal{R})$ is the unique finite set such that $g_{L(\mathcal{R})} = g_{\mathcal{V}^*}$ and $\text{hypo}(g_{L(\mathcal{R})}) = \text{hypo}(g_{\mathcal{V}^*}) = \cap(\mathcal{H}_{\mathcal{V}^*} \cup \mathcal{H}_{\mathcal{Y}})$, the last of which is irredundant.*

Proposition 70 *Given a minimizable loss $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ such that $\underline{L}_+$ is polyhedral, there exists a finite $\mathcal{V}^*$ satisfying Assumption 1.*

Proof Consider the unique $\mathcal{V}^* \subseteq L(\mathcal{R})$ that is given in Corollary 69 such that $g_{L(\mathcal{R})} = g_{\mathcal{V}^*}$. Finally, by construction in Lemma 68, we additionally have $\mathcal{H}^* = \mathcal{H}_{\mathcal{Y}} \cup \mathcal{H}_{\mathcal{V}^*}$, and $\cap\mathcal{H}^*$ is irredundant. ■

F.5 Extended Bayes risk to extended level sets: projecting from $\mathbb{R}^d_+$ to $\mathbb{R}^{\mathcal{Y}}_+$

When the loss L is clear, we denote the faces of $\text{hypo}(g_{L(\mathcal{R})})$ by $F_v := H_v \cap \text{hypo}(g_{L(\mathcal{R})})$. We now define the projection $\pi : \mathbb{R}^{\mathcal{Y}} \times \mathbb{R} \to \mathbb{R}^{\mathcal{Y}}, (x, c) \mapsto x$.

In this subsection, we will establish results about the dimensionality of extended level sets, and conditions under which subsets of these extended level sets cover the nonnegative orthant $\mathbb{R}^{\mathcal{Y}}_+$. As a first step, the extended level sets generated by $\mathcal{V}^*$ cover the nonnegative orthant.

Corollary 71 *Consider $L, \mathcal{V}^*$ satisfying Assumption 1. Then $\cup_{v \in \mathcal{V}^*} \pi(F_v) = \mathbb{R}^{\mathcal{Y}}_+$.*

Moreover, the projection π preserves dimension of faces.

Claim 2 *Consider $L, \mathcal{V}^*$ satisfying Assumption 1. Then $\dim(F_v) = \dim(\pi(F_v))$.*

Proof Recall from Definition 63 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 will show that their projections $\{\pi(z_i)\}_i$ are affinely independent, giving the result.

Let $a_1 + \ldots + a_{|\mathcal{Y}|+1} = 0$, such that $\sum_i a_i \pi(z_i) = 0$. As $z_i \in F_v \subseteq H_v$, we have some $\{x_i\}_i$ such that $z_i = (x_i, \langle v, x_i \rangle)$ for all i. By assumption, then, $0 = \sum_i a_i \pi(z_i) = \sum_i a_i x_i$. Thus, $\sum_i a_i z_i = \sum_i a_i (x_i, \langle v, x_i \rangle) = (\sum_i a_i x_i, \langle v, \sum_i a_i x_i \rangle) = (0, 0) = 0 \in \mathbb{R}^{|\mathcal{Y}|+1}$. By affine independence of the z_i, we conclude $a_i = 0$ for all i. Thus, the set $\{\pi(z_i)\}_i$ is affinely independent. ■

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 $v \in L(\mathcal{R})$.

Lemma 72 *Given $L, \mathcal{V}^*$ satisfying Assumption 1, fix $x \in \mathbb{R}^{\mathcal{Y}}_+$. For any $v \in L(\mathcal{R})$, the following are equivalent:*

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

(2) $\langle v, x\rangle = g_{\mathcal{V}^*}(x)$*;*
(3) $v \in \arg\min_{v' \in L(\mathcal{R})} \langle v', x\rangle$*; and*
(4) $x \in \pi(F_v)$ *.*

Proof

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

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 $(x, g_{\mathcal{V}^*}(x)) \in F_v$. ■

Claim 3 *Consider* $L, \mathcal{V}^*$ *satisfying Assumption 1. For all* $v \in \mathcal{V}^*$*,* $\pi(F_v)$ *is full dimensional in* $\mathbb{R}^{\mathcal{Y}}_+$*.*

Proof By Corollary 69, F_v is a facet of $\text{hypo}(g_{L(\mathcal{R})})$ in $\mathbb{R}^d_+$, meaning it is $(d-1)$-dimensional. Moreover, Claim 2 states that the dimension of F_v is preserved for each $v \in \mathcal{V}^*$. Thus, $d - 1 = |\mathcal{Y}| = \dim(F_v) = \dim(\pi(F_v))$. ■

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

Claim 4 *Consider* $L, \mathcal{V}^*$ *satisfying Assumption 1. 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

($\implies$) For contraposition, suppose $\mathcal{V}^* \not\subseteq \mathcal{V}'$. Then $\exists v \in \mathcal{V}^* \setminus \mathcal{V}'$. Observe that $\mathcal{V}^*$ is unique, $\mathcal{H}^*$ is irredundant by assumption, and $\pi(F_v)$ is full-dimensional in $\mathbb{R}^{\mathcal{Y}}_+$ by Claim 3. 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 71, 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 $\{F_v\}_{v \in \mathcal{V}'}$ for some $\mathcal{V}'$ will cover $\mathbb{R}^{\mathcal{Y}}_+$ if and only if $\mathcal{V}^* \subseteq \mathcal{V}'$. Given a finite set $\mathcal{V}^* \subset \mathbb{R}^{\mathcal{Y}}_+$, we denote $\Lambda_S := \{\pi(F_v) \mid v \in S\}$ as the set of projected facets generated by S.

Claim 5 *Consider $L, \mathcal{V}^*$ satisfying Assumption 1, 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 4. Explicitly, for all $v \in \mathcal{V}^*$ we also have $v \in \mathcal{V}'$, so, $\pi(F_v) \in \Lambda_{\mathcal{V}^*} \cap \Lambda_{\mathcal{V}'} = \Lambda_{\mathcal{V}^*}$.

$\impliedby$ If $\Lambda_{\mathcal{V}^*} \subseteq \Lambda_{\mathcal{V}'}$, then $\cup_{v^* \in \mathcal{V}^*} \pi(F_{v^*}) \subseteq \cup_{v \in \mathcal{V}'} \pi(F_v)$. By Corollary 71, we have $\cup_{v^* \in \mathcal{V}^*} \pi(F_{v^*}) = \mathbb{R}^{\mathcal{Y}}_+$, so $\mathbb{R}^{\mathcal{Y}}_+ \subseteq \cup_{v \in \mathcal{V}'} \pi(F_v)$. The other direction of subset inequality following from $\text{hypo}(g_{\mathcal{V}^*})$ being finite only on $\mathbb{R}^{\mathcal{Y}}_+$. ■

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

Corollary 73 *Consider $L, \mathcal{V}^*$ satisfying Assumption 1. For any $v \in L(\mathcal{R})$, there is a $v^* \in \mathcal{V}^*$ such that $\pi(F_v) \subseteq \pi(F_{v^*})$.*

Proof First, observe that $F_v \subseteq F_{v^*}$ as F_{v^*} is a facet by construction of $\mathcal{V}^*$ and Theorem 65 (ii). As each face of a polyhedron is contained in a facet, we have $F_v \subseteq F_{v^*}$. Moreover, $\pi(F_v) \subseteq \pi(F_{v^*})$ follows immediately as a corollary. ■

F.6 Translating to properties: projecting from $\mathbb{R}^{\mathcal{Y}}_+$ to $\Delta_{\mathcal{Y}}$

Definition 74 *For any polyhedral concave function $f_{\mathcal{V}}$ with domain on $\Delta_{\mathcal{Y}}$, we denote the function $f_{\mathcal{V}}(p) = \min_{v \in \mathcal{V}} \langle p, v \rangle - \delta(p \mid \Delta_{\mathcal{Y}})$, where $\mathcal{V} \subset \mathbb{R}^{\mathcal{Y}}_+$ is a finite set.*

We now make a few observations about the extensions of 1-homogeneous polyhedral functions.

Claim 6 *For any minimizable function L such that $\underline{L}$ is polyhedral, its extension $\underline{L}_+$ is also polyhedral.*

Proof We will think of $\underline{L}$ as defined $\underline{L} : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}_+ \cup \{-\infty\}$ with $\text{dom}(\underline{L}) = \Delta_{\mathcal{Y}}$. 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 $\underline{L}(p) = -\infty$. Moreover, since $\underline{L}$ is polyhedral, it is finitely generated (Rockafellar, 1997, Proposition 19.1.2) and can be written

$$\begin{aligned} \underline{L}(p) &= \min(\langle p, b_1 \rangle - \beta_1, \ldots, \langle p, b_k \rangle - \beta_k) - \delta(p \mid \Delta_{\mathcal{Y}}) \\ &= \min(\langle p, b_1 - \beta_1 \mathbb{1} \rangle, \ldots, \langle p, b_k - \beta_k \mathbb{1} \rangle) - \delta(p \mid \Delta_{\mathcal{Y}}) \\ &= \min_{(b,\beta) \in \mathcal{B}} \langle p, b - \mathbb{1}\beta \rangle - \delta(p \mid \Delta_{\mathcal{Y}}) \ , \end{aligned}$$

where $\mathcal{B} = \{(b_i, \beta_i)\}_{i=1}^k$ from the second line. We claim that $\underline{L}_+(x) = \min_{(b,\beta) \in \mathcal{B}} \langle x, b - \mathbb{1}\beta \rangle - \delta(x \mid \mathbb{R}^{\mathcal{Y}}_+)$, as this form is clearly 1-homogenous and agrees with $\underline{L}$ on $\Delta_{\mathcal{Y}}$. As $\mathcal{B}$ is a finite set, $\underline{L}_+$ is polyhedral. ■

Claim 7 *Consider $L, \mathcal{V}^*$ satisfying Assumption 1. 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 $p \in \Delta_{\mathcal{Y}}$.*

Now, we 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 8 *Consider $L, \mathcal{V}^*$ as in Assumption 1. For all $v \in \mathcal{V}^*$, consider the face F_v of* $\text{hypo}(g_{L(\mathcal{R})})$. *Then $\theta(v) = \pi(F_v) \cap \Delta_\mathcal{Y}$.*

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

$$\begin{aligned}
p \in \theta(v) &\iff \langle v, p\rangle = f_{\mathcal{V}^*}(p) && \text{Definition of } \theta \\
&\iff \langle v, p\rangle = \min_{v' \in \mathcal{V}^*} \langle v', p\rangle && f_{\mathcal{V}^*} = g_{\mathcal{V}^*} \text{ on } \Delta_\mathcal{Y} \text{ (Claim 7)} \\
&\iff v \in \arg\min_{v' \in \mathcal{V}^*} \langle v', p\rangle \\
&\iff p \in \pi(F_v)\ . && \text{Lemma 72}
\end{aligned}$$

■

F.7 Moving back from $\Delta_\mathcal{Y}$ to $\mathbb{R}^\mathcal{Y}_+$

Now that we have translated from $\mathbb{R}^d_+$ to $\mathbb{R}^\mathcal{Y}_+$ in § F.5 and from $\mathbb{R}^\mathcal{Y}_+$ to $\Delta_\mathcal{Y}$ in § F.6, we take some final steps to prove Lemma 20 by showing equivalences from $\Delta_\mathcal{Y}$ to $\mathbb{R}^\mathcal{Y}_+$.

Lemma 75 *Consider $L, \mathcal{V}^*$ satisfying Assumption 1. 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

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

The rest of the result follows from Claim 8. ■

Lemma 76 *Consider $L, \mathcal{V}^*$ satisfying Assumption 1. 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 $g(cx) = \min_{v \in \mathcal{V}^*} \langle v, cx\rangle = c \min_{v \in \mathcal{V}^*} \langle v, x\rangle = cg(x)$ for any $c > 0$ by linearity of the inner product. ■

Every minimizable loss L elicits a unique property $\Gamma := \text{prop}[L]$, 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 77 *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}^*$ satisfying Assumption 1. 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

$$\begin{aligned}
p \in \Gamma_r &= \{p' \in \Delta_{\mathcal{Y}} \mid r \in \arg\min_{r' \in \mathcal{R}} \langle L(r'), p' \rangle\} && \text{Definition of level set} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid v \in \arg\min_{v' \in L(\mathcal{R})} \langle v', p' \rangle\} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid \langle v, p' \rangle = \min_{v' \in L(\mathcal{R})} \langle v', p' \rangle\} && L \text{ minimizable} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid \langle v, p' \rangle = g_{\mathcal{V}^*}(p')\} && \text{Assumption 1} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid c\langle v, p' \rangle = c g_{\mathcal{V}^*}(p')\} \\
&= \{p' \in \Delta_{\mathcal{Y}} \mid \langle v, cp' \rangle = g_{\mathcal{V}^*}(cp')\} && \text{Lemma 76} \\
\implies cp &\in \{x \in \mathbb{R}^{\mathcal{Y}}_+ \mid \langle v, x \rangle = g_{\mathcal{V}^*}(x)\} = \bar{\Gamma}_r \ .
\end{aligned}$$

■

Lemma 78 *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}^*$ satisfying Assumption 1. For any $r \in \mathcal{R}$ with $v = L(r)$, $\bar{\Gamma}_r = \pi(F_v)$.*

Proof

$$\begin{aligned}
\bar{\Gamma}_r &= \{x \in \mathbb{R}^{\mathcal{Y}}_+ \mid \langle L(r), x \rangle = \underline{L}_+(x)\} && \text{Definition of } \bar{\Gamma}_r \\
&= \{x \in \mathbb{R}^{\mathcal{Y}}_+ \mid \langle L(r), x \rangle = g_{L(\mathcal{R})}(x)\} && \text{Assumption 1} \\
&= \{x \in \mathbb{R}^{\mathcal{Y}}_+ \mid \langle v, x \rangle = g_{L(\mathcal{R})}(x)\} && v = L(r) \\
&= \pi(F_v) && \text{Since } F_v = \{(x, g_{L(\mathcal{R})}(x)) \mid \langle v, x \rangle = g_{L(\mathcal{R})}(x)\}
\end{aligned}$$

■

Claim 9 *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}^*$ satisfying Assumption 1. 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 $g_{\mathcal{V}^*}$ is 1-homogeneous via Lemma 76. 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 77. Finally, we show the extended level set is exactly the projection $\pi(F_v)$ in Lemma 78. As a corollary, we chain the results to observe $\cup_{r \in \mathcal{R}'} \Gamma_r = \Delta_{\mathcal{Y}} \implies \cup_{r \in \mathcal{R}'} \bar{\Gamma}_r = \mathbb{R}^{\mathcal{Y}}_+ \implies \cup_{v \in L(\mathcal{R}')} \pi(F_v) = \mathbb{R}^{\mathcal{Y}}_+$, 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 75, 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. ■

F.8 Proving Lemma 20

We now proceed with a few final lemmas that ultimately yield the proof of Lemma 20.

Lemma 79 *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}^*$ satisfying Assumption 1. 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 9 and Claim 4 to yield the result. ■

Define $\Theta_S := \{\theta(v) \mid v \in S\}$; it follows that $\Theta_{\mathcal{V}^*}$ is exactly the set of level sets of the property elicited by L.

Lemma 80 *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}^*$ satisfying Assumption 1. A finite set $\mathcal{R}' \subseteq \mathcal{R}$ with $\mathcal{V}' = L(\mathcal{R}')$ is representative if and only if $\Theta_{\mathcal{V}^*} \subseteq \Theta_{\mathcal{V}'}$.*

Proof Chain Claim 9 and Claim 5 to yield the result. ■

With $L, \mathcal{V}^*$ satisfying Assumption 1, we additionally let $\mathcal{R}^*$ be a set such that $\mathcal{V}^* := L(\mathcal{R}^*)$. Such a set exists as $\mathcal{V}^* \subseteq L(\mathcal{R})$ in Assumption 1, though it is not necessarily unique.

Corollary 81 *Consider $L, \mathcal{V}^*$ satisfying Assumption 1, and $\mathcal{R}^*$ such that $\mathcal{V}^* = L(\mathcal{R}^*)$. Moreover, suppose L elicits Γ. $\Theta_{\mathcal{V}^*} = \{\Gamma_r \mid r \in \mathcal{R}^*\}$.*

Lemma 82 *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}^*$ satisfying Assumption 1. Moreover, let $\Gamma := \text{prop}[L]$. $\Theta_{\mathcal{V}^*} = \{\Gamma_r \mid r \in \mathcal{R}, \dim(\Gamma_r) = |\mathcal{Y}| - 1\}$.*

Proof From Claim 3, 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 75, we have $\theta(v) = \Gamma_r = \pi(F_v) \cap \Delta_{\mathcal{Y}}$ is full-dimensional relative to the simplex. The result follows. ■

Lemma 83 *Consider $L : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+, \mathcal{V}^*$ satisfying Assumption 1, and $\mathcal{R}^* \subseteq \mathcal{R}$ the set such that $\mathcal{V}^* = L(\mathcal{R}^*)$. Moreover, consider $\Gamma := \text{prop}[L]$. 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 73, there is a $v^* \in \mathcal{V}^* \subseteq L(\mathcal{R})$ 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 75. The result follows. ■

We are now ready to apply the above to prove Lemma 20. In particular, any loss L satisfying the assumptions of Lemma 20 has some $g_{L(\mathcal{R})} = \underline{L}_+$ as in this section.

Lemma 84 *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 $\Gamma = \text{prop}[L]$, 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 Since $\underline{L}$ is polyhedral, so is $\underline{L}_+$ by Claim 6. Therefore, we have satisfied the requirements of Proposition 70, and can conclude there is a finite set $\mathcal{V}^* \subseteq L(\mathcal{R})$ such that $\underline{L}_+ = g_{\mathcal{V}^*}$ satisfying Assumption 1. Hence, there is a finite set $\mathcal{R}^*$ such that $\mathcal{V}^* = L(\mathcal{R}^*)$. For all $x \in \mathbb{R}_+^{\mathcal{Y}}$, there exists $v \in \mathcal{V}^*$ such that $\langle v, x\rangle = g_{\mathcal{V}^*}(x)$, and thus some $r \in \mathcal{R}^*$ such that $\langle L(r), p\rangle = g_{\mathcal{V}^*}(x)$. Thus, for all $p \in \Delta_{\mathcal{Y}}$, there exists $r \in \mathcal{R}^*$ such that $\underline{L}(p) = \underline{L}_+(p) = g_{\mathcal{V}^*}(p) = \langle L(r), p\rangle$ and therefore $r \in \text{prop}[L](p)$. As this is true for all $p \in \Delta_{\mathcal{Y}}$, $\mathcal{R}^*$ is representative for L.

Now consider the itemized statements. Lemma 79 is exactly statement (1). This immediately implies statement (2). Moreover, Lemma 80 is exactly statement (3), and again statement (4) immediately follows. Statement (5) is a corollary of the existence of a finite representative set, which follows since $\mathcal{V}^* \subseteq L(\mathcal{R})$. Statement (6) is exactly Lemma 82. Statement (7) is exactly Lemma 83. Finally, Statement (8) follows as a corollary of statement (2) and Corollary 27. ■