An Embedding Framework for Consistent Polyhedral Surrogates

Jessie Finocchiaro
jessica.finocchiaro@colorado.edu
CU Boulder

Rafael Frongillo
raf@colorado.edu
CU Boulder

Bo Waggoner
bwag@colorado.edu
Microsoft Research, CU Boulder

Abstract

We formalize and study the natural approach of designing convex surrogate loss functions via embeddings for problems such as classification or ranking. In this approach, one embeds each of the finitely many predictions (e.g. classes) as a point in $\mathbb{R}^d$, assigns the original loss values to these points, and convexifies the loss in between to obtain a surrogate. We prove that this approach is equivalent, in a strong sense, to working with polyhedral (piecewise linear convex) losses. Moreover, given any polyhedral loss L, we give a construction of a link function through which L is a consistent surrogate for the loss it embeds. We go on to illustrate the power of this embedding framework with succinct proofs of consistency or inconsistency of various polyhedral surrogates in the literature.

1 Introduction

Convex surrogate losses are a central building block in machine learning for classification and classification-like problems. A growing body of work seeks to design and analyze convex surrogates for given loss functions, and more broadly, understand when such surrogates can and cannot be found. For example, recent work has developed tools to bound the required number of dimensions of the surrogate's hypothesis space [13, 24]. Yet in some cases these bounds are far from tight, such as for *abstain loss* (classification with an abstain option) [5, 24, 25, 33, 34]. Furthermore, the kinds of strategies available for constructing surrogates, and their relative power, are not well-understood.

We augment this literature by studying a particularly natural approach for finding convex surrogates, wherein one "embeds" a discrete loss. Specifically, we say a convex surrogate L embeds a discrete loss ℓ if there is an injective embedding from the discrete reports (predictions) to a vector space such that (i) the original loss values are recovered, and (ii) a report is ℓ-optimal if and only if the embedded report is L-optimal. If this embedding can be extended to a calibrated link function, which maps approximately L-optimal reports to ℓ-optimal reports, consistency follows [2]. Common examples which follow this general construction include hinge loss as a surrogate for 0-1 loss and the abstain surrogate mentioned above.

Using tools from property elicitation, we show a tight relationship between such embeddings and the class of polyhedral (piecewise-linear convex) loss functions. In particular, by focusing on Bayes risks, we show that every discrete loss is embedded by some polyhedral loss, and every polyhedral loss function embeds some discrete loss. Moreover, we show that any polyhedral loss gives rise to a calibrated link function to the loss it embeds, thus giving a very general framework to construct consistent convex surrogates for arbitrary losses.

Preprint. Under review.

Related works.   The literature on convex surrogates focuses mainly on smooth surrogate losses [5, 6, 8, 9, 26, 30]. Nevertheless, nonsmooth losses, such as the polyhedral losses we consider, have been proposed and studied for a variety of classification-like problems [19, 31, 32]. A notable addition to this literature is Ramaswamy et al. [25], who argue that nonsmooth losses may enable dimension reduction of the prediction space (range of the surrogate hypothesis) relative to smooth losses, illustrating this conjecture with a surrogate for *abstain loss* needing only $\log n$ dimensions for n labels, whereas the best known smooth loss needs $n - 1$. Their surrogate is a natural example of an embedding (cf. Section 5.1), and serves as inspiration for our work.

While property elicitation has by now an extensive literature [10, 12, 15, 17, 18, 22, 28, 29], these works are mostly concerned with point estimation problems. Literature directly connecting property elicitation to consistency is sparse, with the main reference being Agarwal and Agarwal [2]; note however that they consider single-valued properties, whereas properties elicited by general convex losses are necessarily set-valued.

2   Setting

For discrete prediction problems like classification, due to hardness of directly optimizing a given discrete loss, many machine learning algorithms can be thought of as minimizing a surrogate loss function with better optimization qualities, e.g., convexity. Of course, to show that this surrogate loss successfully addresses the original problem, one needs to establish consistency, which depends crucially on the choice of link function that maps surrogate reports (predictions) to original reports. After introducing notation, and terminology from property elicitation, we thus give a sufficient condition for consistency (Def. 4) which depends solely on the conditional distribution over $\mathcal{Y}$.

2.1   Notation and Losses

Let $\mathcal{Y}$ be a finite outcome (label) space, and throughout let $n = |\mathcal{Y}|$. The set of probability distributions on $\mathcal{Y}$ is denoted $\Delta_{\mathcal{Y}} \subseteq \mathbb{R}^{\mathcal{Y}}$, represented as vectors of probabilities. We write p_y for the probability of outcome $y \in \mathcal{Y}$ drawn from $p \in \Delta_{\mathcal{Y}}$.

We assume that a given discrete prediction problem, such as classification, is given in the form of a *discrete loss* $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$, which maps a report (prediction) r from a finite set $\mathcal{R}$ to the vector of loss values $\ell(r) = (\ell(r)_y)_{y \in \mathcal{Y}}$ for each possible outcome $y \in \mathcal{Y}$. We will assume throughout that the given discrete loss is *non-redundant*, meaning every report is uniquely optimal (minimizes expected loss) for some distribution $p \in \Delta_{\mathcal{Y}}$. Similarly, surrogate losses will be written $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$, typically with reports written $u \in \mathbb{R}^d$. We write the corresponding expected loss when $Y \sim p$ as $\langle p, \ell(r) \rangle$ and $\langle p, L(u) \rangle$. The *Bayes risk* of a loss $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ is the function $\underline{L} : \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ given by $\underline{L}(p) := \inf_{u \in \mathbb{R}^d} \langle p, L(u) \rangle$; naturally for discrete losses we write $\underline{\ell}$ (and the infimum is over $\mathcal{R}$).

For example, 0-1 loss is a discrete loss with $\mathcal{R} = \mathcal{Y} = \{-1, 1\}$ given by $\ell_{\text{0-1}}(r)_y = \mathbb{1}\{r \neq y\}$, with Bayes risk $\underline{\ell_{\text{0-1}}}(p) = 1 - \max_{y \in \mathcal{Y}} p_y$. Two important surrogates for $\ell_{\text{0-1}}$ 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 $u \in \mathbb{R}$.

Most of the surrogates L we consider will be *polyhedral*, meaning piecewise linear and convex; we therefore briefly recall the relevant definitions. In $\mathbb{R}^d$, a *polyhedral set* or *polyhedron* is the intersection of a finite number of closed halfspaces. A *polytope* is a bounded polyhedral set. A convex function $f : \mathbb{R}^d \to \mathbb{R}$ is *polyhedral* if its epigraph is polyhedral, or equivalently, if it can be written as a pointwise maximum of a finite set of affine functions [27].

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

For example, hinge loss is polyhedral, whereas logistic loss is not.

2.2   Property Elicitation

To make headway, we will appeal to concepts and results from the property elicitation literature, which elevates the *property*, or map from distributions to optimal reports, as a central object to study in its own right. In our case, this map will often be multivalued, meaning a single distribution could

yield multiple optimal reports. (For example, when $p = (1/2, 1/2)$, both $y = 1$ and $y = -1$ optimize 0-1 loss.) To this end, we will use double arrow notation to mean a mapping to all nonempty subsets, so that $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ is shorthand for $\Gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}} \setminus \emptyset$. See the discussion following Definition 3 for conventions regarding $\mathcal{R}, \Gamma, \gamma, L, \ell$, etc.

Definition 2 (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* $\Gamma_r := \{p : r \in \Gamma(p)\}$.

Intuitively, $\Gamma(p)$ is the set of reports which should be optimal for a given distribution p, and Γ_r is the set of distributions for which the report r should be optimal. 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 *elicits* the mode, as we formalize below.

Definition 3 (Elicits). *A loss* $L : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$, elicits *a property* $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ *if*

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

As Γ *is uniquely defined by* L, *we write* $\text{prop}[L]$ *to refer to the property elicited by a loss* L.

For finite properties (those with $|\mathcal{R}| < \infty$) and discrete losses, we will use lowercase notation γ and ℓ, respectively, with reports $r \in \mathcal{R}$; for surrogate properties and losses we use Γ and L, with reports $u \in \mathbb{R}^d$. For general properties and losses, we will also use Γ and L, as above.

2.3   Links and Embeddings

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

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

It is well-known that calibration implies *consistency*, in the following sense (cf. [2]). Given feature space $\mathcal{X}$, fix a distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$. Let L^* be the best possible expected L-loss achieved by any hypothesis $H : \mathcal{X} \to \mathbb{R}^d$, and ℓ^* the best expected ℓ-loss for any hypothesis $h : \mathcal{X} \to \mathcal{R}$, respectively. Then (L, ψ) is consistent if a sequence of surrogate hypotheses $H_1, H_2, \ldots$ whose L-loss limits to L^*, then the ℓ-loss of $\psi \circ H_1, \psi \circ H_2, \ldots$ limits to ℓ^*. As Definition 4 does not involve the feature space $\mathcal{X}$, we will drop it for the remainder of the paper.

Several consistent convex surrogates in the literature can be thought of as "embeddings", wherein one maps the discrete reports to a vector space, and finds a convex loss which agrees with the original loss. A key condition is that the original reports should be optimal exactly when the corresponding embedded points are optimal. We formalize this notion as follows.

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

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

Note that it is not clear if embeddings give rise to calibrated links; indeed, apart from mapping the embedded points back to their original reports via $\psi(\varphi(r)) = r$, how to map the remaining values is far from clear. We address the question of when embeddings lead to calibrated links in Section 4.

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

3

to check that $L_{\text{hinge}}(r)_y = 2\ell_{\text{0-1}}(r)_y$ for all $r, y \in \{-1, 1\}$. For condition (ii), let us compute the property each loss elicits, i.e., the set of optimal reports for each p:

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

In particular, we see that $-1 \in \text{prop}[\ell_{\text{0-1}}](p) \iff p_1 \in [0, 1/2] \iff -1 \in \text{prop}[L_{\text{hinge}}](p)$, and $1 \in \text{prop}[\ell_{\text{0-1}}](p) \iff p_1 \in [1/2, 1] \iff 1 \in \text{prop}[L_{\text{hinge}}](p)$. With both conditions of Definition 5 satisfied, we conclude that L_{hinge} embeds $2\ell_{\text{0-1}}$. In this particular case, it is known (L_{hinge}, ψ) is calibrated for $\psi(u) = \text{sgn}(u)$; we address in Section 4 the interesting question of whether embeddings lead to calibration in general.

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. We defer to the following section the question of when such surrogates are consistent.

To begin, we observe that, somewhat surprisingly, our embedding condition in Definition 5 is equivalent to merely matching Bayes risks. This useful fact will drive many of our results.

Proposition 1. *A loss L embeds discrete loss ℓ if and only if $\underline{L} = \underline{\ell}$.*

Proof. Throughout we have $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$, $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$, and define $\Gamma = \text{prop}[L]$ and $\gamma = \text{prop}[\ell]$. Suppose L embeds ℓ via the embedding φ. Letting $\mathcal{U} := \varphi(\mathcal{R})$, define $\gamma' : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{U}$ by $\gamma' : p \mapsto \Gamma(p) \cap \mathcal{U}$. To see that $\gamma'(p) \neq \emptyset$ for all $p \in \Delta_{\mathcal{Y}}$, note that by the definition of γ as the property elicited by ℓ we have some $r \in \gamma(p)$, and by the embedding condition (3), $\varphi(r) \in \Gamma(p)$. By Lemma 3, we see that $L|_{\mathcal{U}}$ (the loss L with reports restricted to $\mathcal{U}$) elicits γ' and $\underline{L} = \underline{L|_{\mathcal{U}}}$. As $L(\varphi(\cdot)) = \ell(\cdot)$ by the embedding, we have

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

for all $p \in \Delta_{\mathcal{Y}}$. Combining with the above, we now have $\underline{L} = \underline{\ell}$.

For the reverse implication, assume that $\underline{L} = \underline{\ell}$. In what follows, we implicitly work in the affine hull of $\Delta_{\mathcal{Y}}$, so that interiors are well-defined, and $\underline{\ell}$ may be differentiable on the (relative) interior of $\Delta_{\mathcal{Y}}$. Since ℓ is discrete, $-\underline{\ell}$ is polyhedral as the pointwise maximum of a finite set of linear functions. The projection of its epigraph E_{ℓ} onto $\Delta_{\mathcal{Y}}$ forms a power diagram by Theorem 4, whose cells are full-dimensional and correspond to the level sets γ_r of $\gamma = \text{prop}[\ell]$.

For each $r \in \mathcal{R}$, let p_r be a distribution in the interior of γ_r, and let $u_r \in \Gamma(p)$. Observe that, by definition of the Bayes risk and Γ, for all $u \in \mathbb{R}^d$ the hyperplane $v \mapsto \langle v, -L(u_r) \rangle$ supports the epigraph E_L of $-\underline{L}$ at the point $(p, -\langle p, L(u) \rangle)$ if and only if $u \in \Gamma(p)$. Thus, the hyperplane $v \mapsto \langle v, -L(u_r) \rangle$ supports $E_L = E_{\ell}$ at the point $(p_r, -\langle p_r, L(u_r) \rangle)$, and thus does so at the entire facet $\{(p, -\langle p, L(u_r) \rangle) : p \in \gamma_r\}$; by the above, $u_r \in \Gamma(p)$ for all such distributions as well. We conclude that $u_r \in \Gamma(p) \iff p \in \gamma_r \iff r \in \gamma(p)$, satisfying condition (3) for $\varphi : r \mapsto u_r$. To see that the loss values match, we merely note that the supporting hyperplanes to the facets of E_L and E_{ℓ} are the same, and the loss values are uniquely determined by the supporting hyperplane. (In particular, if h supports the facet corresponding to γ_r, we have $\ell(r)_y = L(u_r)_y = h(\delta_y)$, where δ_y is the point distribution on outcome y.) $\square$

From this more succinct embedding condition, we can in turn simplify the condition that a loss embeds *some* discrete loss: it does if and only if its Bayes risk is polyhedral. (We say a concave function is polyhedral if its negation is a polyhedral convex function.) Note that Bayes risk, a function from distributions over $\mathcal{Y}$ to the reals, may be polyhedral even if the loss itself is not.

Proposition 2. *A loss L embeds a discrete loss if and only if $\underline{L}$ is polyhedral.*

4

Proof. If L embeds ℓ, Proposition 1 gives us $\underline{L} = \underline{\ell}$, and its proof already argued that $\underline{\ell}$ is polyhedral. For the converse, let $\underline{L}$ be polyhedral; we again examine the proof of Proposition 1. The projection of $\underline{L}$ onto $\Delta_{\mathcal{Y}}$ forms a power diagram by Theorem 4 with finitely many cells $C_1, \ldots, C_k$, which we can index by $\mathcal{R} := \{1, \ldots, k\}$. Defining the property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ by $\gamma_r = C_r$ for $r \in \mathcal{R}$, we see that the same construction gives us points $u_r \in \mathbb{R}^d$ such that $u_r \in \Gamma(p) \iff r \in \gamma(p)$. Defining $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$ by $\ell(r) = L(u_r)$, the same proof shows that L embeds ℓ. $\qquad\square$

Combining Proposition 2 with the observation that polyhedral losses have polyhedral Bayes risks (Lemma 5), we obtain the first direction of our equivalence between polyhedral losses and embedding.

Theorem 1. *Every polyhedral loss L embeds a discrete loss.*

We now turn to the reverse direction: which discrete losses are embedded by some polyhedral loss? Perhaps surprisingly, we show that *every* discrete loss is embeddable, using a construction via convex conjugate duality which has appeared several times in the literature (e.g. [1, 9, 11]). Note however that the number of dimensions d required could be as large as $|\mathcal{Y}|$.

Theorem 2. *Every discrete loss ℓ 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 [27, 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. 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 1 to see that L embeds ℓ. 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. $\qquad\square$

4 Consistency via Calibrated Links

We have now seen the tight relationship between polyhedral losses and embeddings; in particular, every polyhedral loss embeds some discrete loss. The embedding itself tells us how to link the embedded points back to the discrete reports (map $\varphi(r)$ to r), but it is not clear when this link can be extended to the remaining reports, and whether such a link can lead to consistency. In this section, we give a construction to generate calibrated links for *any* polyhedral loss.

Appendix D contains the full proof; this section provides a sketch along with the main construction and result. The first step is to give a link ψ such that exactly minimizing expected surrogate loss L, followed by applying ψ, always exactly minimizes expected original loss ℓ. The existence of such a link is somewhat subtle, because in general some point u that is far from any embedding point can minimize expected loss for two very different distributions p, p', making it unclear whether there exists a link choice $\psi(u)$ that is simultaneously optimal for both. We show that as we vary p over $\Delta_{\mathcal{Y}}$, there are only finitely many sets of the form $U = \arg\min_{u \in \mathbb{R}^d} \langle p, L(u) \rangle$ (Lemma 4). Associating each U with $R_U \subseteq \mathcal{R}$, the set of reports whose embedding points are in U, we enforce that all points in U link to some report in R_U. (As a special case, embedding points must link to their corresponding reports.) Proving this is well-defined uses a chain of arguments involving the Bayes risk, ultimately showing that if u lies in multiple U, the corresponding report sets R_U all intersect at some $r =: \psi(u)$.

Intuitively, to create separation, we just need to "thicken" this construction by mapping all approximately-optimal points u to optimal reports r. Let $\mathcal{U}$ contain all optimal report sets U

5

of the form above. A key step in the following definition will be to narrow down a "link envelope" Ψ where $\Psi(u)$ denotes the legal or valid choices for $\psi(u)$.

Definition 6. *Given a polyhedral L that embeds some ℓ, an $\epsilon > 0$, and a norm $\|\cdot\|$, the ϵ-thickened link ψ is constructed as follows. First, initialize $\Psi : \mathbb{R}^d \rightrightarrows \mathcal{R}$ by setting $\Psi(u) = \mathcal{R}$ for all u. Then for each $U \in \mathcal{U}$, for all points u such that $\inf_{u^* \in U} \|u^* - u\| < \epsilon$, update $\Psi(u) = \Psi(u) \cap R_U$. Finally, define $\psi(u) \in \Psi(u)$, breaking ties arbitrarily. If $\Psi(u)$ became empty, then leave $\psi(u)$ undefined.*

Theorem 3. *Let L be polyhedral, and ℓ the discrete loss it embeds from Theorem 1. Then for small enough $\epsilon > 0$, the ϵ-thickened link ψ is well-defined and, furthermore, is a calibrated link from L to ℓ.*

Sketch. Well-defined: For the initial construction above, we argued that if some collection such as U, U', U'' overlap at a u, then their report sets $R_U, R_{U'}, R_{U''}$ also overlap, so there is a valid choice $r = \psi(u)$. Now, we thicken all sets $U \in \mathcal{U}$ by a small enough ϵ; it can be shown that if the *thickened* sets overlap at u, then U, U', U'' themselves overlap, so again $R_U, R_{U'}, R_{U''}$ overlap and there is a valid chioce $r = \psi(u)$.

Calibrated: By construction of the thickened link, if u maps to an incorrect report, i.e. $\psi(u) \notin \gamma(p)$, then u must have at least distance ϵ to the optimal set U. We then show that the minimal gradient of the expected loss along any direction away from U is lower-bounded, giving a constant excess expected loss at u. $\qquad\square$

5 Application to Specific Surrogates

Our results give a framework to construct consistent surrogates and link functions for any discrete loss, but they also provide a way to verify the consistency or inconsistency of given surrogates. Below, we illustrate the power of this framework with specific examples from the literature, as well as new examples. In some cases we simplify existing proofs, while in others we give new results, such as a new calibrated link for abstain loss, and the inconsistency of the recently proposed Lovász hinge.

5.1 Consistency of abstain surrogate and link construction

In classification settings with a large number of labels, several authors consider a variant of classification, with the addition of a "reject" or *abstain* option. For example, Ramaswamy et al. [25] study the loss $\ell_\alpha : [n] \cup \{\bot\} \to \mathbb{R}_+^{\mathcal{Y}}$ defined by $\ell_\alpha(r)_y = 0$ if $r = y$, α if $r = \bot$, and 1 otherwise. Here, the report $\bot$ corresponds to "abstaining" if no label is sufficiently likely, specifically, if no $y \in \mathcal{Y}$ has $p_y \geq 1 - \alpha$. Ramaswamy et al. [25] provide a polyhedral surrogate for ℓ_α, which we present here for $\alpha = 1/2$. Letting $d = \lceil \log_2(n) \rceil$ their surrogate is $L_{1/2} : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ given by

$$L_{1/2}(u)_y = \left(\max_{j \in [d]} B(y)_j u_j + 1 \right)_+ , \tag{4}$$

where $B : [n] \to \{-1, 1\}^d$ is a arbitrary injection; let us assume $n = 2^d$ so that we have a bijection. Consistency is proven for the following link function,

$$\psi(u) = \begin{cases} \bot & \min_{i \in [d]} |u_i| \leq 1/2 \\ B^{-1}(\mathrm{sgn}(-u)) & \text{otherwise} \end{cases} . \tag{5}$$

In light of our framework, we can see that $L_{1/2}$ is an excellent example of an embedding, where $\varphi(y) = B(y)$ and $\varphi(\bot) = 0 \in \mathbb{R}^d$. Moreover, the link function ψ can be recovered from Theorem 3 with norm $\|\cdot\|_\infty$ and $\epsilon = 1/2$; see Figure 1(L). Hence, our framework would have simplified the process of finding such a link, and the corresponding proof of consistency. To illustrate this point further, we give an alternate link ψ_1 corresponding to $\|\cdot\|_1$ and $\epsilon = 1$, shown in Figure 1(R):

$$\psi_1(u) = \begin{cases} \bot & \|u\|_1 \leq 1 \\ B^{-1}(\mathrm{sgn}(-u)) & \text{otherwise} \end{cases} . \tag{6}$$

Theorem 3 immediately gives calibration of $(L_{1/2}, \psi_1)$ with respect to $\ell_{1/2}$. Aside from its simplicity, one possible advantage of ψ_1 is that it appears to yield the same constant in generalization bounds as ψ, yet assigns $\bot$ to much less of the surrogate space $\mathbb{R}^d$. It would be interesting to compare the two links in practice.
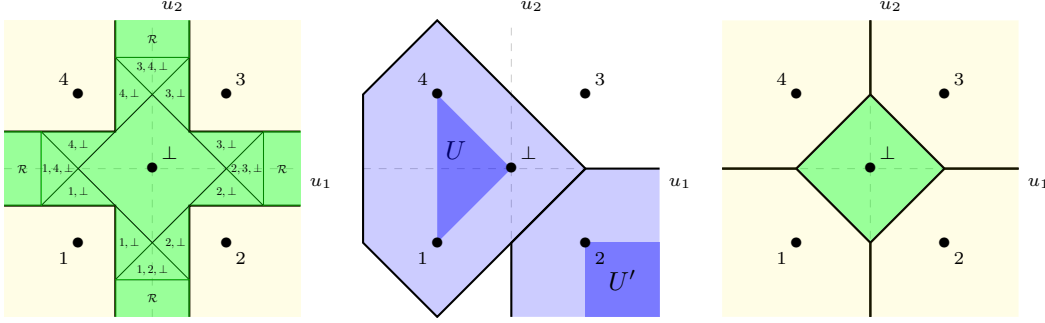
Figure 1: Constructing links for the abstain surrogate $L_{1/2}$ with $d = 2$. The embedding is shown in bold labeled by the corresponding reports. (L) The link envelope Ψ resulting from Theorem 3 using $\|\cdot\|_\infty$ and $\epsilon = 1/2$, and a possible link ψ which matches eq. (5) from [25]. (M) An illustration of the thickened sets from Definition 6 for two sets $U \in \mathcal{U}$, using $\|\cdot\|_1$ and $\epsilon = 1$. (R) The Ψ and ψ from Theorem 3 using $\|\cdot\|_1$ and $\epsilon = 1$.

5.2 Inconsistency of Lovász hinge

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 [14, 16, 23]. In the case of k binary predictions, these settings are typically formalized by taking the predictions and outcomes to be ± 1 vectors, so $\mathcal{R} = \mathcal{Y} = \{-1, 1\}^k$. One then defines a joint loss function, which is often merely a function of the set of mispredictions, meaning $\ell^g(r)_y = g(\{i \in [k] : r_i \neq y_i\})$ for some function $g : 2^{[k]} \to \mathbb{R}$. For example, Hamming loss is simply given by $g(S) = |S|$. In an effort to provide a general convex surrogate for these settings when g is a submodular function, Yu and Blaschko [32] introduce the *Lovász hinge*, which leverages the well-known convex Lovász extension of submodular functions. While the authors provide theoretical justification and experiments, consistency of the Lovász hinge is left open, which we resolve.

Rather than formally define the Lovász hinge, we defer the complete analysis to Appendix E, and focus here on the $k = 2$ case. For brevity, we write $g_\emptyset := g(\emptyset)$, $g_{1,2} := g(\{1, 2\})$, etc. Assuming g is normalized and increasing (meaning $g_{1,2} \geq \{g_1, g_2\} \geq g_\emptyset = 0$), the Lovász hinge $L : \mathbb{R}^k \to \mathbb{R}_+^{\mathcal{Y}}$ is given by

$$L^g(u)_y = \max\Big\{(1 - u_1 y_1)_+ g_1 + (1 - u_2 y_2)_+(g_{1,2} - g_1),$$
$$(1 - u_2 y_2)_+ g_2 + (1 - u_1 y_1)_+(g_{1,2} - g_2)\Big\}, \quad (7)$$

where $(x)_+ = \max\{x, 0\}$. We will explore the range of values of g for which L^g is consistent, where the link function $\psi : \mathbb{R}^2 \to \{-1, 1\}^2$ is fixed as $\psi(u)_i = \text{sgn}(u_i)$, with ties broken arbitrarily.

Let us consider $g_\emptyset = 0$, $g_1 = g_2 = g_{1,2} = 1$, for which ℓ^g is merely 0-1 loss on $\mathcal{Y}$. For consistency, then, for any distribution $p \in \Delta_\mathcal{Y}$, we must have that whenever $u \in \arg\min_{u' \in \mathbb{R}^2} p \cdot L^g(u')$, then $\psi(u)$ must be the most likely outcome, in $\arg\max_{y \in \mathcal{Y}} p(y)$. Simplifying eq. (13), however, we have

$$L^g(u)_y = \max\{(1 - u_1 y_1)_+, (1 - u_2 y_2)_+\} = \max\{1 - u_1 y_1, 1 - u_2 y_2, 0\}, \quad (8)$$

which is exactly the abstain surrogate (4) for $d = 2$. We immediately conclude that L^g cannot be consistent with ℓ^g, as the origin will be the unique optimal report for L^g under distributions with $p_y < 0.5$ for all y, and one simply take a distribution which disagrees with the way ties are broken in ψ. For example, if we take $\text{sgn}(0) = 1$, then under $p((1, 1)) = p((1, -1)) = p((-1, 1)) = 0.2$ and $p((-1, -1)) = 0.4$, we have $\{0\} = \arg\min_{u \in \mathbb{R}^2} p \cdot L^g(u)$, yet $\psi(0) = (1, 1) \notin \{(-1, -1)\} = \arg\min_{r \in \mathcal{R}} p \cdot \ell^g(r)$.

In fact, this example is typical: using our embedding framework, and characterizing when $0 \in \mathbb{R}^2$ is an embedded point, one can show that L^g is consistent if and only if $g_{1,2} = g_1 + g_2$. Moreover, in the linear case, which corresponds to g being *modular*, the Lovász hinge reduces to weighted Hamming loss, which is trivially consistent from the consistency of hinge loss for 0-1 loss. In Appendix E, we generalize this observation for all k: L^g is consistent if and only if g is modular. In other words, even for $k > 2$, the only consistent Lovász hinge is weighted Hamming loss. These results cast doubt on the effectiveness of the Lovász hinge in practice.
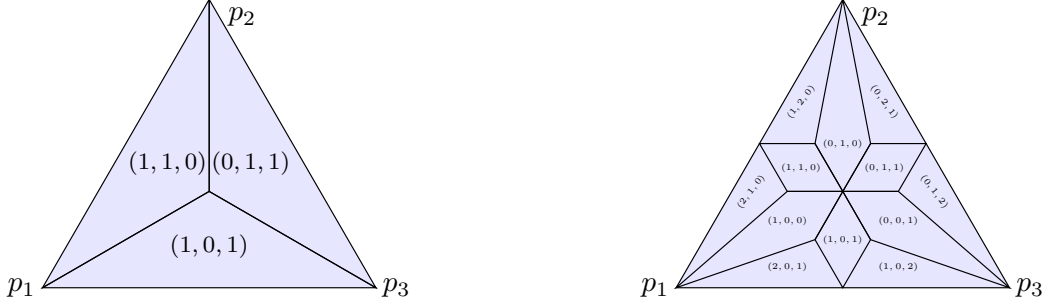
7

Figure 2: Minimizers of $\langle p, \ell^{\text{top-2}} \rangle$ and $\langle p, \ell^2 \rangle$, respectively, varying p over Δ_3.

5.3 Inconsistency of top-k losses

In certain classification problems when ground truth may be ambiguous, such as object identification, it is common to predict a set of possible labels. As one instance, the top-k classification problem is to predict the set of k most likely labels; formally, we have $\mathcal{R} := \{r \in \{0,1\}^n : \|r\|_0 = k\}$, $1 < k < n$, $\mathcal{Y} = [n]$, and discrete loss $\ell^{\text{top-}k}(r)_y = 1 - r_y$. Surrogates for this problem commonly take reports $u \in \mathbb{R}^n$, with the link $\psi(u) = \{u_{[1]}, \dots, u_{[k]}\}$, where $u_{[i]}$ is the ith largest entry of u.

Lapin et al. [19, 20, 21] provide the following convex surrogate loss for this problem, which Yang and Koyejo [31] show to be inconsistent:[1]

$$L^k(u)_y := \left(1 - u_y + \tfrac{1}{k} \sum_{i=1}^{k} (u - e_y)_{[i]}\right)_+ , \tag{9}$$

where e_y is 1 in component y and 0 elsewhere. With our framework, we can say more. Specifically, while (L^k, ψ) is not consistent for $\ell^{\text{top-}k}$, since L^k is polyhedral (Lemma 17), we know from Theorem 1 that it embeds *some* discrete loss ℓ^k, and from Theorem 3 there is a link ψ' such that (L^k, ψ') is calibrated (and consistent) for ℓ^k. We therefore turn to deriving this discrete loss ℓ^k.

For concreteness, consider the case with $k = 2$ over $n = 3$ outcomes. We can re-write $L^2(u)_y = \left(1 - u_y + \tfrac{1}{2}(u_{[1]} + u_{[2]} - \min(1, u_y))\right)_+$. By inspection, we can derive the properties elicited by $\ell^{\text{top-2}}$ and L^2, respectively, which reveals that the set $\mathcal{R}'$ consisting of all permutations of $(1, 0, 0)$, $(1, 1, 0)$, and $(2, 1, 0)$, are always represented among the minimizers of L^2. Thus, L^2 embeds the loss $\ell^2(r)_y = 0$ if $r_y = 2$ or $\ell^2(r)_y = 1 - r_y + \tfrac{1}{2}\langle r, \mathbb{1} - e_y \rangle$ otherwise. Observe that this is like $\ell^{\text{top-2}}$, with a punishment for reporting weight on elements of r other than the outcome and a reward for showing a "higher confidence" in the correct outcome (i.e. $r_y = 2$). Moreover, we can visually inspect the corresponding properties (Fig. 2) to immediately see why L^2 is inconsistent: for distributions where the two least likely labels are roughly equally (un)likely, the minimizer will put all weight on the most likely label, and thus fail to distinguish the other two. More generally, L^2 cannot be consistent because the property it embeds does not "refine" (subdivide) the top-k property, so not just ψ, but *no* link function, could make L^2 consistent.

6 Conclusion and Future Directions

This paper formalizes an intuitive way to design convex surrogate losses for classification-like problems—by embedding the reports into $\mathbb{R}^d$. We establish a close relationship between embedding and polyhedral surrogates, showing both that every polyhedral loss embeds a discrete loss (Theorem 1) and that every discrete loss is embedded by some polyhedral loss (Theorem 2). We then construct a calibrated link function from any polyhedral loss to the discrete loss it embeds, giving consistency for all such losses. We conclude with examples of how the embedding framework presented can be applied to understand existing surrogates in the literature, including those for the abstain loss, top-k loss, and Lovász hinge.

---

[1]Yang and Koyejo also introduce a consistent surrogate, but it is non-convex.

One open question of particular interest involves the dimension of the input to a surrogate; given a discrete loss, can we construct the surrogate that embeds it *of minimal dimension*? If we naïvely embed the reports into an n-dimensional space, the dimensionality of the problem scales with the number of possible labels n. As the dimension of the optimization problem is a function of this *embedding dimension d*, a promising direction is to leverage tools from elicitation complexity [13, 18] and convex calibration dimension [24] to understand when we can take $d \ll n$.

Acknowledgements

References

[1] 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.

[2] 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.

[3] 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.

[4] Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373, 2013.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] Tobias Fissler, Johanna F Ziegel, and others. Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016.

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

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

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

[14] 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.

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

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] Harish G Ramaswamy, Ambuj Tewari, Shivani Agarwal, et al. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1):530–554, 2018.

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

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

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

[29] 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.

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

[31] Forest Yang and Sanmi Koyejo. On the consistency of top-k surrogate losses, 01 2018.

[32] 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.

[33] Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(Jan):111–130, 2010.

[34] Chong Zhang, Wenbo Wang, and Xingye Qiao. On reject and refine options in multicategory classification. *Journal of the American Statistical Association*, 113(522):730–745, 2018. doi: 10.1080/01621459.2017.1282372. URL https://doi.org/10.1080/01621459.2017.1282372.

A Power diagrams

First, we present several definitions from Aurenhammer [3].

Definition 7. *A* cell complex *in $\mathbb{R}^d$ is a set C of faces (of dimension $0, \ldots, d$) which (i) union to $\mathbb{R}^d$, (ii) have pairwise disjoint relative interiors, and (iii) any nonempty intersection of faces F, F' in C is a face of F and F' and an element of C.*

Definition 8. *Given sites $s_1, \ldots, s_k \in \mathbb{R}^d$ and weights $w_1, \ldots, w_k \geq 0$, the corresponding* power diagram *is the cell complex given by*

$$\text{cell}(s_i) = \{x \in \mathbb{R}^d : \forall j \in \{1, \ldots, k\} \, \|x - s_i\|^2 - w_i \leq \|x - s_j\| - w_j\} . \tag{10}$$

Definition 9. *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.*

Theorem 4 (Aurenhammer [3]). *A cell complex is affinely equivalent to a convex polyhedron if and only if it is a power diagram.*

In particular, one can consider the epigraph of a polyhedral convex function on $\mathbb{R}^d$ and the projection down to $\mathbb{R}^d$; in this case we call the resulting power diagram *induced* by the convex function. We extend Aurenhammer's result to a weighted sum of convex functions, showing that the induced power diagram is the same for any choice of strictly positive weights.

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

Proof. For any convex function g with epigraph P, the proof of Aurenhammer [3, 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 $\mathring{F}'$ 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 [27], f, f' are polyhedral. We now show that f is differentiable whenever f' is differentiable:

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

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

B Embedding properties

It is often convenient to work directly with properties and set aside the losses which elicit them. To this end, we say a property to embeds another if eq. (3) holds. We begin with the notion of redundancy.

Definition 10 (Finite property, non-redundant). *A property* $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ *is* redundant *if for some* $r, r' \in \mathcal{R}$ *we have* $\Gamma_r \subseteq \Gamma_{r'}$*, and* non-redundant *otherwise.* Γ *is* finite *if it is non-redundant and* $\mathcal{R}$ *is a finite set. We typically write finite properties in lower case, as* γ.

When working with convex losses which are not strictly convex, one quickly encounters redundant properties: if $\langle p, L(\cdot) \rangle$ is minimized by a point where $p \cdot L$ is flat, then there will be an uncountable set of reports which also minimize the loss. As results in property elicitation typically assume non-redundant properties (e.g. [11, 13]), it is useful to consider a transformation which removes redundant level sets. We capture this transformation as the trim operator presented below.

Definition 11. *Given an elicitable property* $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$*, we define* $\mathrm{trim}(\Gamma) = \{\Gamma_u : u \in \mathcal{R} \text{ s.t. } \neg\exists u' \in \mathcal{R}, u' \neq u, \Gamma_u \subsetneq \Gamma_{u'}\}$ *as the set of maximal level sets of* Γ.

Take note that the unlabeled property $\mathrm{trim}(\Gamma)$ is non-redundant, meaning that for any $\theta \in \mathrm{trim}(\Gamma)$, there is no level set $\theta' \in \mathrm{trim}(\Gamma)$ such that $\theta \subset \theta'$.

Before we state the Proposition needed to prove many of the statements in Section 3, we will need to general lemmas about properties and their losses. The first follows from standard results relating finite properties to power diagrams (see Theorem 4 in the appendix), and its proof is omitted. The second is closely related to the trim operator: it states that if some subset of the reports are always represented among the minimizers of a loss, then one may remove all other reports and elicit the same property (with those other reports removed).

Lemma 2. *Let* γ *be a finite (non-redundant) property elicited by a loss* L*. Then the negative Bayes risk* G *of* L *is polyhedral, and the level sets of* γ *are the projections of the facets of the epigraph of* G *onto* $\Delta_{\mathcal{Y}}$*, and thus form a power diagram. In particular, the level sets* γ *are full-dimensional in* $\Delta_{\mathcal{Y}}$ *(i.e., of dimension* $n - 1$*).*

Lemma 3. *Let* L *elicit* $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}_1$*, and let* $\mathcal{R}_2 \subseteq \mathcal{R}_1$ *such that* $\Gamma(p) \cap \mathcal{R}_2 \neq \emptyset$ *for all* $p \in \Delta_{\mathcal{Y}}$*. Then* $L|_{\mathcal{R}_2}$ *(L restricted to* $\mathcal{R}_2$*) elicits* $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}_2$ *defined by* $\gamma(p) = \Gamma(p) \cap \mathcal{R}_2$*. Moreover, the Bayes risks of* L *and* $L|_{\mathcal{R}_2}$ *are the same.*

Proof. Let $p \in \Delta_{\mathcal{Y}}$ be fixed throughout. First let $r \in \gamma(p) = \Gamma(p) \cap \mathcal{R}_2$. Then $r \in \Gamma(p) = \arg\min_{u \in \mathcal{R}_1} \langle p, L(u) \rangle$, so as $r \in \mathcal{R}_2$ we have in particular $r \in \arg\min_{u \in \mathcal{R}_2} \langle p, L(u) \rangle$. For the other direction, suppose $r \in \arg\min_{u \in \mathcal{R}_2} \langle p, L(u) \rangle$. By our assumption, we must have some $r^* \in \Gamma(p) \cap \mathcal{R}_2$. On the one hand, $r^* \in \Gamma(p) = \arg\min_{u \in \mathcal{R}_1} \langle p, L(u) \rangle$. On the other, as $r^* \in \mathcal{R}_2$, we certainly have $r^* \in \arg\min_{u \in \mathcal{R}_2} \langle p, L(u) \rangle$. But now we must have $\langle p, L(r) \rangle = \langle p, L(r^*) \rangle$, and thus $r \in \arg\min_{u \in \mathcal{R}_1} \langle p, L(u) \rangle = \Gamma(p)$ as well. We now see $r \in \Gamma(p) \cap \mathcal{R}_2$. Finally, the equality of the Bayes risks $\min_{u \in \mathcal{R}_1} \langle p, L(u) \rangle = \min_{u \in \mathcal{R}_2} \langle p, L(u) \rangle$ follows immediately by the above, as $\emptyset \neq \Gamma(p) \cap \mathcal{R}_2 \subseteq \Gamma(p)$ for all $p \in \Delta_{\mathcal{Y}}$. $\square$

We now state a useful result for proving the existence of an embedding loss, which shows remarkable structure of embeddable properties, and the properties that embed them. First, we conclude that any embeddable property must be elicitable. We also conclude that 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, just with other reports filling in the gaps between the embedded reports of γ. (When working with convex losses, these extra reports are typically the convex hull of the embedded reports.) In this sense, we can regard embedding as a minor departure from direct elicitation: if a loss L elicits Γ which embeds γ, we can think of L as essentially eliciting γ itself. Finally, we have an important converse: if Γ has finitely many full-dimensional level sets, or if $\mathrm{trim}(\Gamma)$ is finite, then Γ must embed some finite elicitable property with those same level sets.

Proposition 3. *Let* $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathbb{R}^d$ *be an elicitable property. The following are equivalent:*

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

2. $\mathrm{trim}(\Gamma)$ *is a finite set, and* $\cup \mathrm{trim}(\Gamma) = \Delta_{\mathcal{Y}}$.

3. *There is a finite set of full-dimensional level sets* Θ *of* Γ*, and* $\cup \Theta = \Delta_{\mathcal{Y}}$.

Moreover, when any of the above hold, $\{\gamma_r : r \in \mathcal{R}\} = \mathrm{trim}(\Gamma) = \Theta$*, and* γ *is elicitable.*

Proof. Let L elicit Γ.

$1 \Rightarrow 2$: By the embedding condition, taking $\mathcal{R}_1 = \mathbb{R}^d$ and $\mathcal{R}_2 = \varphi(\mathcal{R})$ satisfies the conditions of Lemma 3: for all $p \in \Delta_{\mathcal{Y}}$, as $\gamma(p) \neq \emptyset$ by definition, we have some $r \in \gamma(p)$ and thus some $\varphi(r) \in \Gamma(p)$. Let $G(p) := -\min_{u \in \mathbb{R}^d} \langle p, L(u) \rangle$ be the negative Bayes risk of L, which is convex, and $G_{\mathcal{R}}$ that of $L|_{\varphi(\mathcal{R})}$. By the Lemma, we also have $G = G_{\mathcal{R}}$. As γ is finite, G is polyhedral. Moreover, the projection of the epigraph of G onto $\Delta_{\mathcal{Y}}$ forms a power diagram, with the facets projecting onto the level sets of γ, the cells of the power diagram. (See Theorem 4.) As L elicits Γ, for all $u \in \mathbb{R}^d$, the hyperplane $p \mapsto \langle p, L(u) \rangle$ is a supporting hyperplane of the epigraph of G at $(p, G(p))$ if and only if $u \in \Gamma(p)$. This supporting hyperplane exposes some face F of the epigraph of G, which must be contained in some facet F'. Thus, the projection of F, which is Γ_u, must be contained in the projection of F', which is a level set of γ. We conclude that $\Gamma_u \subseteq \gamma_r$ for some $r \in \mathcal{R}$. Hence, $\mathrm{trim}(\Gamma) = \{\gamma_r : r \in \mathcal{R}\}$, which is finite, and unions to $\Delta_{\mathcal{Y}}$.

$2 \Rightarrow 3$: let $\mathcal{R} = \{u_1, \ldots, u_k\} \subseteq \mathbb{R}^d$ be a set of distinct reports such that $\mathrm{trim}(\Gamma) = \{\Gamma_{u_1}, \ldots, \Gamma_{u_k}\}$. Now as $\cup \, \mathrm{trim}(\Gamma) = \Delta_{\mathcal{Y}}$, for any $p \in \Delta_{\mathcal{Y}}$, we have $p \in \Gamma_{u_i}$ for some $u_i \in \mathcal{R}$, and thus $\Gamma(p) \cap \mathcal{R} \neq \emptyset$. We now satisfy the conditions of Lemma 3 with $\mathcal{R}_1 = \mathbb{R}^d$ and $\mathcal{R}_2 = \mathcal{R}$. The property $\gamma : p \mapsto \Gamma(p) \cap \mathcal{R}$ is non-redundant by the definition of trim, finite, and elicitable. Now from Lemma 2, the level sets $\Theta = \{\gamma_r : r \in \mathcal{R}\}$ are full-dimensional, and union to $\Delta_{\mathcal{Y}}$. Statement 3 then follows from the fact that $\gamma_r = \Gamma_r$ for all $r \in \mathcal{R}$.

$3 \Rightarrow 1$: let $\Theta = \{\theta_1, \ldots, \theta_k\}$. For all $i \in \{1, \ldots, k\}$ let $u_i \in \mathbb{R}^d$ such that $\Gamma_{u_i} = \theta_i$. Now define $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \{1, \ldots, k\}$ by $\gamma(p) = \{i : p \in \theta_i\}$, which is non-degenerate as $\cup \Theta = \Delta_{\mathcal{Y}}$. By construction, we have $\gamma_i = \theta_i = \Gamma_{u_i}$ for all i, so letting $\varphi(i) = u_i$ we satisfy the definition of embedding, namely statement 1. $\qquad \square$

C Polyhedral losses

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

Proof. For all p, let $P(p)$ be the epigraph of the convex function $u \mapsto \langle p, L(u) \rangle$. From Lemma 1, we have that the power diagram $D_{\mathcal{Y}}$ induced by the projection of $P(p)$ onto $\mathbb{R}^d$ is the same for any $p \in \mathring{\Delta}_{\mathcal{Y}}$. Let $\mathcal{F}_{\mathcal{Y}}$ be the set of faces of $D_{\mathcal{Y}}$, which by the above are the set of faces of $P(p)$ projected onto $\mathbb{R}^d$ for any $p \in \mathring{\Delta}_{\mathcal{Y}}$.

We claim for all $p \in \mathring{\Delta}_{\mathcal{Y}}$, that $\Gamma(p) \in \mathcal{F}_{\mathcal{Y}}$. To see this, let $u \in \Gamma(p)$, and $u' = (u, \langle p, L(u) \rangle) \in P(p)$. The optimality of u is equivalent to u' being contained in the face F of $P(p)$ exposed by the normal $(0, \ldots, 0, -1) \in \mathbb{R}^{d+1}$. 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 consider $\mathcal{Y}' \subset \mathcal{Y}, \mathcal{Y}' \neq \emptyset$. 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 $\mathcal{Y}' \subseteq \mathcal{Y}$. $\qquad \square$

Lemma 5. *If L is polyhedral, $\underline{L}$ is polyhedral.*

Proof. Let $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ be a polyhedral loss, and $\Gamma = \mathrm{prop}[L]$. By Lemma 4, $\mathcal{U} = \Gamma(\Delta_{\mathcal{Y}})$ is finite. For each $U \in \mathcal{U}$, select $u_U \in U$, and let $U' = \{u_U : U \in \mathcal{U}\}$. Then for all $p \in \Delta_{\mathcal{Y}}$ we have $\Gamma(p) \cap U' \neq \emptyset$, so Lemma 3 gives us $\underline{L} = L|_{U'}$, which is polyhedral as U' is finite. $\qquad \square$

D Thickened link and calibration

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

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

14

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

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

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

Lemma 6. *Let* $\mathcal{U}' \subseteq \mathcal{U}$*. If* $\cap_{U \in \mathcal{U}'} U \neq \emptyset$ *then* $\cap_{U \in \mathcal{U}'} R_U \neq \emptyset$*.*

Proof. Let $u \in \cap_{U \in \mathcal{U}'} U$. Then we claim there is some r such that $\Gamma_u \subseteq \gamma_r$. This follows from Proposition 3, which shows that $\mathrm{trim}(\Gamma) = \{\gamma_r : r \in \mathcal{R}\}$. Each Γ_u is either in $\mathrm{trim}(\Gamma)$ or is contained in some set in $\mathrm{trim}(\Gamma)$, by definition, proving the claim.

For each $U \in \mathcal{U}'$, for any p such that $U = \Gamma(p)$, we have in particular that u is optimal for p, so $p \in \Gamma_u$, so $p \in \gamma_r$, so r is optimal for p. This implies that $\phi(r)$, the embedding point, is optimal for p, so $\phi(r) \in U$. This holds for all $U \in \mathcal{U}'$, so $r \in \cap_{U \in \mathcal{U}'} R_U$, so it is nonempty. $\square$

Lemma 7. *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 [7, 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. $\square$

Lemma 8. *Let* $\{U_j : j \in \mathcal{J}\}$ *be a finite collection of closed, convex sets with* $\cap_{j \in \mathcal{J}} U_j \neq \emptyset$*. Then there exists* $\epsilon > 0$ *such that* $\cap_j B(U_j, \epsilon) \subseteq B(\cap_j U_j, \delta)$*.*

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 7, 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 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 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 be the minimum over these finitely many ϵ_F (one per halfspace). $\square$

Figure 3: Illustration of a special case of the proof of Lemma 8 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 $B(D, \delta)$.
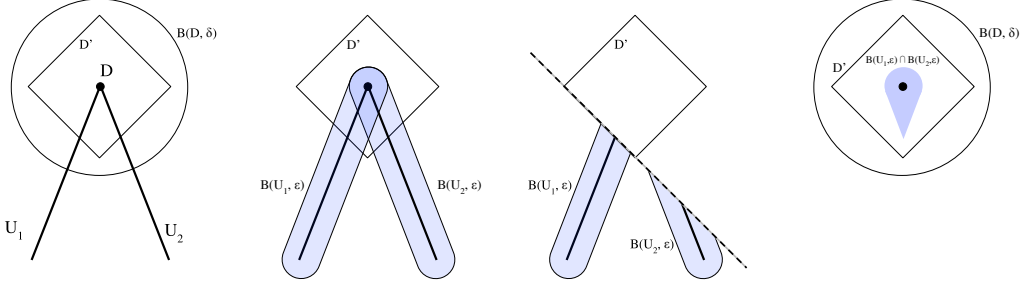
Lemma 9. *Let $\{U_j : j \in \mathcal{J}\}$ be a finite collection of nonempty closed, convex sets with $\cap_{j \in \mathcal{J}} U_j = \emptyset$. Then for all $\delta > 0$, there exists $\epsilon > 0$ such that $\cap_{j \in \mathcal{J}} B(U_j, \epsilon) = \emptyset$.*

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 ϵ. By Lemma 8, for any $\epsilon > 0$, there exists $\delta > 0$ such that $\cap_{j' \neq j} B(U_{j'}, \delta) \subseteq B(U', \epsilon/3)$. Then we have $B(U_j, \epsilon/3) \cap B(U', \epsilon/3) = \emptyset$.

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 small enough δ we have $\cap_{j' \neq j} B(U_{j'}, \delta) = \emptyset$, which proves this case. □

Corollary 1. *There exists a small enough $\epsilon > 0$ such that, for any subset $\{U_j : j \in \mathcal{J}\}$ of $\mathcal{U}$, if $\cap_j U_j = \emptyset$, then $\cap_j B(U_j, \epsilon) = \emptyset$.*

Proof. For each subset, Lemma 9 gives an ϵ. We take the minimum over these finitely many choices. □

Theorem 5. *For all small enough ϵ, the epsilon-thickened link ψ (Definition 6) is a well-defined link function from $\mathcal{R}'$ to $\mathcal{R}$, i.e. $\psi(u) \neq \perp$ for all u.*

Proof. Fix a small enough ϵ as promised by Corollary 1. Consider any $u \in \mathcal{R}'$. If u is not in $B(U, \epsilon)$ for any $U \in \mathcal{U}$, then we have $\Psi(u) = \mathcal{R}$, so it is nonempty. Otherwise, let $\{U_j : j \in \mathcal{J}\}$ be the family whose thickenings intersect at u. By Corollary 1, because of our choice of ϵ, the family themselves has nonempty intersection. By Lemma 6, their corresponding report sets $\{R_j : j \in \mathcal{J}\}$ also intersect at some r, so $\Psi(u)$ is nonempty. □

In the rest of the section, for shorthand, we write $L(u; p) := \langle p, L(u) \rangle$ and similarly $\ell(r; p)$.

Lemma 10. *Let U be a convex, closed set and $u \notin U$. Then $\inf_{u^* \in U} \|u - u^*\|$ is achieved by some unique $u^* \in U$. Furthermore, u^* is the unique member of U such that $u = u^* + \alpha v$ for some $\alpha > 0$ and unit vector v that exposes u^*.*

Proof. Unique achievement of the infimum is well-known. (Achievement follows e.g. because the set $U \cap \{u' : \|u - u'\| \leq d(U, u) + 1\}$ is closed and compact, so the continuous function $u' \mapsto \|u - u'\|$ achieves its infimum. Uniqueness follows because for two different points u', u'' at the same distance from u, the point $0.5u' + 0.5u''$ is strictly closer and also lies in the convex set U.) Now suppose $u = u' + \alpha' v'$ where v' is a unit vector exposing u'. Then U is contained in the halfspace $\{u'' : \langle u'', v'' \rangle \leq \langle u', v' \rangle\}$. But every point in this halfspace is distance at least α' from u, as $\|u - u''\| \geq \langle v, u - u'' \rangle \geq \langle v, u - u' \rangle = \alpha'$. So u' uniquely achieves this minimum distance. □

Lemma 11. *If L is a polyhedral loss, then for each p, there exists a constant c such that, for all u,*

$$L(u; p) - \inf_{u^* \in \mathcal{R}'} L(u^*; p) \geq c \cdot d(\Gamma(p), u).$$

Proof. Fix p and let $U = \Gamma(p)$. If $u \in U$, then both sides are zero. So it remains to find a c such that the inequality holds for all $u \notin U$.

$L(\cdot; p)$ is a convex polyhedral function, so it is the pointwise maximum over finitely many affine functions. Recall that $\underline{L}(p) = \min_u L(u; p)$, the Bayes risk. Construct the convex polyhedral function $\hat{L}(\cdot; p)$ by dropping from the maximum those affine functions that are never equal to L for any $u^* \in U$. We have $\hat{L}(u^*; p) = \underline{L}(p)$ for all $u^* \in U$ and $\hat{L}(u; p) \leq L(u; p)$ for all $u \notin U$. Now $\hat{L}$ is also a maximum over finitely many affine functions $\mathcal{F}$. Each such function $f \in \mathcal{F}$ is equal to $\hat{L}$ above a closed, convex cell $U_f \subseteq \mathbb{R}^d$ in the power diagram formed by projecting $\hat{L}(\cdot; p)$. If f has nonzero gradient, then $U_f \cap U$ is a face of U. We will prove that there exists $c_f > 0$ such that, for all $u \in U_f$,

$$\hat{L}(u; p) \geq \underline{L}(p) + c_f \cdot d(U, u).$$

Taking c to be the minimum of c_f over the finitely many $f \in \mathcal{F}$ with nonzero gradient (which covers all points $u \notin U$) will complete the proof.

Consider the set of unit vectors $V = \{v \in \mathbb{R}^d : \|v\| = 1\}$ and the boundary of U, denoted ∂U. For any $u^* \in \partial U$, $v \in V$ such that v exposes u^*, let $G_{u^*, v} = \{u^* + \beta v : \beta \geq 0\}$ be the ray leaving U from u^* in direction v. For each $f \in \mathcal{F}$, we define the set $R_f \subseteq \partial U \times V$ to be the points (u^*, v) such that there exists $\epsilon > 0$ with $G_{u^*, v} \cap U_f = G_{u^*, v} \cap B(u^*, \epsilon)$; that is, such that the ray $G_{u^*, v}$ starts its journey in U_f. Futhermore, define $U^*_{f, v} = \{u^* \in \partial U : (u^*, v) \in R_f\}$ and $V_f = \{v \in V : \exists u^* \in U_f \cap U, (u^*, v) \in R_f\}$. (That is, $U^*_{f, v}$ is the set of points from which the ray in direction v begins in U_f, and V_f is the set of all normal directions in which some ray begins in U_f.) Finally, define $G_f = \cup_{(u^*, v) \in R_f} G_{u^*, v}$ as the union of all such rays beginning in U_f. Note that $\cup_{f \in \mathcal{F}} G_f \supseteq \mathbb{R}^d \setminus U$; this follows as every point not in U is on a normal ray out of U, which must begin in some cell U_f.

We will prove the following steps:

1. For all $f \in \mathcal{F}$, $v \in V_f$, there exists a constant $c_{f,v} > 0$ such that $L(u; p) \geq \underline{L}(p) + c_{f,v} \cdot d(U, u)$ for all $u \in G_{u^*, v}$ and all $u^* \in U^*_{f, v}$.

2. For all $f \in \mathcal{F}$, the set V_f is compact, and the map $v \mapsto c_{f,v}$ is continuous on V_f.

3. Hence, there is an infimum $c_f > 0$ such that $f(u) \geq \underline{L}(p) + c_f \cdot d(U, u)$ for all $u \in G_f$.

4. Let $c = \min\{c_f : f \in \mathcal{F}, \nabla f \neq 0\}$; then $L(u; p) \geq \underline{L}(p) + c \cdot d(U, u)$ for all $u \notin U$.

(1) Let ∇f denote the gradient of the affine function f. Note that because u^* is on the boundary of U, we have $f(u^*) = \hat{L}(u^*; p) = \underline{L}(p)$. So we can write, using Lemma 10,

$$\begin{aligned}
f(u) &= f(u^*) + (\nabla f) \cdot (u - u^*) \\
f(u) &= f(u^*) + (\nabla f) \cdot (d(u^*, u)v) \\
&= \underline{L}(p) + c_{f,v} \cdot d(U, u)
\end{aligned}$$

where $c_{f,v} = (\nabla f) \cdot v$. We must have $c_{f,v} > 0$ because the set U minimizes $L(\cdot; p)$, so $f(u) > f(u^*) = \underline{L}(p)$. The result now follows as $L(u; p) \geq \hat{L}(u; p) \geq f(u)$.

(2) The intersection $U_f \cap U$ is a face of U, and thus decomposes as the union of relative interiors of subfaces, $U_f \cap U = \cup_i \text{ri}(F_i)$. For each i, let $V_i = \{v \in V : \exists u^* \in \text{ri}(F_i), (u^*, v) \in R_f\}$. For any $v \in V_i$, we may consider the power diagram restricted to A, the affine hull of $\{u + \alpha v : u \in U, \alpha \in \mathbb{R}\}$. As there is some $u^* \in U$ such that $(u^*, v) \in R_f$, in particular, $U_f \cap A$ intersects $A \cap \text{ri}(F_i)$ and thus must contain $A \cap \text{ri}(F_i)$. We conclude that $(u', v) \in R_f$ for all other $u' \in \text{ri}(F_i)$. Thus, we have $\{(u^*, v) \in R_f : u^* \in F_i\} = F_i \times V_i$. For closure, pick any $u^* \in \text{ri}(F_i)$, and consider a sequence $\{v_j\}_j$ with $(u^*, v_j) \in R_f$, and corresponding witnesses $\{\epsilon_j\}_j$. Then we have $u^* + \epsilon_j v_j \in U_f$ for

17

all j, and as $u^* \in U_f$ and U_f is closed and convex, the limiting point must be contained in U_f as well. We have now shown V_f to be the union of finitely many closed convex sets, and thus closed. Boundedness follows as V is bounded. Finally, $c_{f,v}$ is linear in v, and thus continuous.

Steps (3) and (4) are immediate and complete the proof. $\qquad\square$

Theorem 6. *For small enough ϵ, the ϵ-thickened link ψ (Definition 6) satisfies that, for all p, there exists $\delta > 0$ such that, for all $u \in \mathcal{R}'$,*

$$L(u; p) - \inf_{u^* \in \mathcal{R}'} L(u^*; p) \geq \delta \left[\ell(\psi(u); p) - \min_{r^* \in \mathcal{R}} \ell(r^*; p) \right].$$

Proof. We take the ϵ thickened link, which is well-defined by Theorem 5. Fix p and let $U = \Gamma(p)$. The left-hand side is nonnegative, so it suffices to prove the result for all u such that the right side is strictly positive, i.e. for all u such that $\psi(u) \notin \gamma(p)$. By definition of the ϵ-thickened link, we must have $d(U, u) \geq \epsilon$. By Lemma 11, we have $L(u; p) - \inf_{u^*} L(u^*; p) \geq C$ where $C = c\epsilon$ for some $c > 0$. This holds for all u. Meanwhile,

$$\ell(\psi(u); p) - \min_{r^*} \ell(r^*; p) \leq \max_{r \in \mathcal{R}} \ell(r; p) - \min_{r^* \in \mathcal{R}} \ell(r^*; p) =: D,$$

for some constant D. This also holds for all u. Set $\delta = \frac{C}{D}$ to complete the proof. $\qquad\square$

Proof of Theorem 3. The two claims are Theorems 5 and 6. $\qquad\square$

E   Lovász Hinge

The Lovász hinge, introduced by Yu and Blaschko [32], is a (convex) polyhedral surrogate for discrete losses described in terms of a submodular function, based on the well-known Lovász extension. We will study this surrogate using our framework, first identifying the loss it embeds, and then leveraging this loss to find a proof of inconsistency. As defining the Lovász hinge takes care, we begin with definitions.

E.1   Notation and Definitions

Let $N = \{1, \ldots, k\}$ be the index set for our binary predictions, with outcomes $\mathcal{Y} = 2^N$ corresponding to the set of labels which are assigned $+1$. To map to the usual labels $\{-1, 1\}$, for any $S \subseteq N$, we let $\mathbb{1}_S \in \{0, 1\}^k$ with $(\mathbb{1}_S)_i = 1 \iff i \in S$ be the 0-1 indicator for S, and we let $\chi_S \in \{-1, 1\}^k$ with $\chi_S = 2\mathbb{1}_S - \mathbb{1}$ be the ± 1 indicator. For clarity of exposition, we will depart from our usual notation for loss functions, writing a discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ and surrogate $L : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$, and writing expected loss $L(u; p)$. The link will be the sign function $\psi = \mathrm{sgn}$, with ties broken arbitrarily.

A set function $f : 2^N \to \mathbb{R}$ is *submodular* if for all $S, T \subseteq N$ we have $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$. A function is *supermodular* if the inequality is reversed, and *modular* if it holds with equality, for all $S, T \subseteq N$. The function f is *increasing* if we have $f(S \cup T) \geq f(S)$, again for all $S, T \subseteq N$.

We are interested in convex surrogates for the following discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{R} = \mathcal{Y} = 2^N$,

$$\ell^f(A, S) = f(A \triangle S), \tag{11}$$

where $\triangle$ is the symmetric difference operator, defined by $S \triangle T = (S \backslash T) \cup (T \backslash S)$. Note: throughout we assume $\triangle$ has operator precedence over $\backslash, \cap,$ and $\cup$. In words, ℓ^f measures the joint error of our k predictions by computing the set of mispredictions (elements in A but not S and vice versa) and calling the set function f.

A natural approach to deriving convex surrogates in the case of submodular functions f is the Lovász extension, which is known to be convex when f if (and only if) submodular. Given any set-valued function $f : 2^N \to \mathbb{R}$, its *Lovász extension* $F : \mathbb{R}^k \to \mathbb{R}$ is given by

$$F(u) = \mathbb{E}[f(\{i : u_i \geq \Theta\})], \tag{12}$$

where Θ is a uniformly distributed random variable on $[0, 1]$. There are several equivalent formulations for the Lovász extension; see Bach [4, Definition 3.1].

Given set function f with Lovász extension F, Yu and Blaschko [32] define the *Lovász hinge* as the loss $L^f : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ given by

$$L^f(u, S) = \begin{cases} F\big((\mathbb{1} - u \odot \chi_S)_+\big) & \text{if } f \text{ is increasing} \\ \big(F(\mathbb{1} - u \odot \chi_S)\big)_+ & \text{otherwise} \end{cases}, \tag{13}$$

where $v(u \odot y)_i = u_i y_i$ is the Hadamard (element-wise) product and $((u)_+)_i = \max(u_i, 0)$. In what follows, we focus on the increasing case, which is the most natural: when you make an additional error, your loss cannot decrease.

E.2 What does L embed?

From well-known facts about the Lovász extension (see Lemma 13 below), L^f is certainly polyhedral, and thus by our framework we know it must embed a discrete loss $\hat{\ell}^f$, which may or may not be the same as ℓ^f. As with the top-k example, we begin our analysis by calculating $\hat{\ell}^f$.

Let $\Gamma^f = \operatorname{prop}[L^f]$ and $\mathcal{U} = \{-1, 0, 1\}^k$. Note that, for disjoint sets $A, B \subseteq N$, we have $\chi_A + \mathbb{1}_B \in \mathcal{U}$, which at coordinate i evaluates to 1 for $i \in A$, 0 for $i \in B$, and -1 otherwise. Moreover, every point in $\mathcal{U}$ can be uniquely described in this way. Finally, observe $\chi_A + \mathbb{1}_B = \chi_A \odot \mathbb{1}_{N \setminus B}$.

We will show that for every distribution p, an element of $\mathcal{U}$ is always represented in the minimizers of L^f, i.e., $\Gamma^f(p)$. First, we show that we may restrict to the filled hypercube $[-1, 1]^k$ without loss of generality.

Lemma 12. *Let $f : 2^N \to \mathbb{R}_+$ be increasing and normalized. Then for all $p \in \Delta_{\mathcal{Y}}$, $\Gamma(p) \cap [-1, 1]^k \neq \emptyset$.*

Proof. Let $u \in \Gamma^f(p)$ such that $|u_i| > 1$ for some $i \in [k]$, and furthermore suppose $|u_i|$ is the smallest value among all such coordinates, i.e., $|u_i| = \min\{|u_j| : |u_j| > 1\}$. We show that $u' \in \Gamma^f(p)$ where $u'_j = u_j$ for $j \neq i$ and $u'_i = \operatorname{sgn}(u_i)$ so that $|u'_i| = 1$; the result then follows by iterating this argument until there are no entries with $|u_i| > 1$. In fact, we will show the stronger statement that for all $S \in \mathcal{Y}$, $L^f(u', S) \leq L^f(u, S)$. Let $w = \mathbb{1} - u \odot \chi_S$ and $w' = \mathbb{1} - u' \odot \chi_S$, and note that $L^f(u, S) = F(w_+)$ and $L^f(u', S) = F(w'_+)$.

First, consider the case that $(\chi_S)_i u_i > 0$; that is, if $u_i > 0$ and $i \in S$, or $u_i < 0$ and $i \notin S$. Here $1 - (\chi_S)_i u_i = 1 - |u_i| < 0$, so $w_i < 0$. For u', we similarly have $w'_i = 1 - |u'_i| = 0$. As u and u' differ only at index i, the same holds for w and w'; we thus have $w_+ = w'_+$, so the loss remains unchanged.

In the other case, $(\chi_S)_i u_i < 0$, we have $w_i = 1 + |u_i| > 2$ and $w'_i = 1 + |u'_i| = 2$, and again the other entries are identical. In particular, $w'_i \leq w_i$. Moreover, we claim that there is no other value in between, i.e., there is no index $j \neq i$ such that $w'_i < w_j < w_i$. This follows from our assumption on i: if we had such a j, then we must have $2 < 1 - (\chi_S)_j u_j < 1 + |u_i|$, which can only occur when $\operatorname{sgn}(u_j) \neq \chi_S$, and thus $-(\chi_S)_j u_j = |u_j|$; we conclude $1 < |u_j| < |u_i|$, a contradiction. Thus, for all $j \neq i$, we have either $w_j \leq w'_i \leq w_i$ or $w'_i \leq w_i \leq w_j$. In light of $w'_j = w_j$, this is equivalent to either (a) $w_j \leq w_i$ and $w'_j \leq w'_i$, or (b) $w_i \leq w_j$ and $w'_i \leq w'_j$. Thus, there is a permutation π which orders the elements of both w and w' simultaneously: for all $j, j' \in [k]$, $j < j'$, we have both $w_{\pi_j} \geq w_{\pi_{j'}}$ and $w'_{\pi_j} \geq w'_{\pi_{j'}}$. By another common representation of the Lovász extension [4, Equation 3.1], we thus have $F(w_+) - F(w'_+) = ((w_+)_i - (w'_+)_i)(f(T \cup \{i\}) - f(T)) = (|u_i| - 1)(f(T \cup \{i\}) - f(T)) > 0$, where $T = \{\pi_1, \ldots, \pi_j\}$ such that $\pi_{j+1} = i$, and we have used the fact that f is increasing and $|u_i| > 1$. $\square$

From Lemma 12, we may now simplify the loss. When $u \in [-1, 1]^k$, we simply have $L^f(u, S) = F(\mathbb{1} - u \odot \chi_S)$, as the coordinates of the input to F are nonnegative. We now further restrict to $\mathcal{U}$.

Lemma 13. *Let $\Gamma = \operatorname{prop}[L^f]$. Then for all $p \in \Delta_{\mathcal{Y}}$, $\Gamma(p) \cap \mathcal{U} \neq \emptyset$.*

Proof. We will construct polytopes $P_\pi^A \subseteq [-1, 1]^k$ for every set $A \subseteq N$ and permutation $\pi \in N^N$, satisfying three conditions: (i) these polytopes cover the hypercube, meaning $\cup_{A, \pi} P_\pi^A = [-1, 1]^k$,

(ii) P_π^A is the convex hull of points in $\mathcal{U}$, and (iii) for all $S \subseteq N$, $L(\cdot, S)$ is linear on P_π^A. The result will then follow, as $L(\cdot; p)$ will be also linear on each P_π^A, and thus minimized at a vertex.

To begin, let us recall the polyhedra on which the Lovász extension is linear; for any permutation π, define $P_\pi = \{u \in \mathbb{R}^k : u_{\pi_1} \geq \cdots \geq u_{\pi_k}\}$. It is clear from the definition that F is linear on P_π; see also Equation 3.1, and "Linear interpolation on simplices" (pg. 167) in Bach [4]. We will use these polyhedra to identify the regions where $L(\cdot, S)$ is linear simultaneously for all outcomes $S \in \mathcal{Y}$. For any $A \subseteq N$ and permutation π, let

$$P_\pi^A = \{u \in [-1, 1]^k : u \odot \chi_A \in P_\pi \cap \mathbb{R}_+^k\} . \tag{14}$$

That is, P_π^A contains all points u such that $u \odot \chi_A$ is nonnegative (meaning $\mathrm{sgn}(u)$ matches χ_A, breaking ties at 0 favorably) and such that the coordinates of $u \odot \chi_A$ are in increasing order according to π.

Condition (i) follows immediately: for any $u \in [-1, 1]^k$, let $A = \{i : u_i \geq 0\}$, and π be any permutation ordering the elements of $u \odot \chi_A$. For condition (ii), note that for any $u \in P_\pi^A$, as $u \odot \chi_A \in P_\pi$ we may write

$$u \odot \chi_A = \sum_{i=1}^{k-1} \left[\left(u_{\pi_i}(\chi_A)_{\pi_i} - u_{\pi_{i+1}}(\chi_A)_{\pi_{i+1}} \right) \mathbb{1}_{\{\pi_1, \ldots, \pi_i\}} \right] + u_{\pi_k}(\chi_A)_{\pi_k} \mathbb{1} , \tag{15}$$

which is a convex combination (again, see Bach [4, pg. 167]). We simply apply the involution $\odot \chi_A$ again, to obtain

$$u = \sum_{i=1}^{k-1} \left(u_{\pi_i}(\chi_A)_{\pi_i} - u_{\pi_{i+1}}(\chi_A)_{\pi_{i+1}} \right) \mathbb{1}_{\{\pi_1, \ldots, \pi_i\}} \odot \chi_A + u_{\pi_k}(\chi_A)_{\pi_k} \chi_A , \tag{16}$$

and condition (ii) follows as $\mathbb{1}_B \cdot \chi_A \in \mathcal{U}$ for all sets $B \subseteq N$.

Finally, for condition (iii), fix a subset $A \subseteq N$ and permutation π. For each outcome $S \subseteq N$, we will construct an alternate permutation π^S such that for all $u \in P_\pi^A$ we have $\mathbb{1} - u \odot \chi_S \in P_{\pi^S}$. As F is linear on $P_{\pi'}$ for all permutations π', we will conclude that for any fixed subset S the loss $L(u, S) = F(\mathbb{1} - u \odot \chi_S)$ will be linear in u on P_π^A.

To construct π^S, we "shatter" the permutation π into two pieces, depending on whether an index is in $A \triangle S$ or not. In particular, note that if $i \in A \triangle S$, then for all $u \in P_\pi^A$ we have $u_i(\chi_A)_i \geq 0$ and $(\chi_A)_i = -(\chi_S)_i$, so $(\mathbb{1} - u \odot \chi_S)_i = 1 - u_i(\chi_S)_i = 1 + u_i(\chi_A)_i \geq 1$. Similarly, when $i \notin A \triangle S$, then $(\chi_A)_i = (\chi_S)_i$, so $(\mathbb{1} - u \odot \chi_S)_i = 1 - u_i(\chi_S)_i = 1 - u_i(\chi_A)_i \leq 1$. As π orders the elements $u_i(\chi_S)_i$ in decreasing order, we see that the following permutation π^S will order the elements of $\mathbb{1} - u \odot \chi_S$ in decreasing order: sort the elements in $A \triangle S$ according to π, followed by the remaining elements according to the reverse of π. As the definition of π^S is independent of u, we see that $u \in P_\pi^A \implies \mathbb{1} - u \odot \chi_S \in P_{\pi^S}$, as desired. $\quad\square$

We can now see that L^f embeds the loss given by L^f restricted to $\mathcal{U}$, as $\mathcal{U}$ is a finite set. In fact, we can write this loss entirely in terms of f itself.

Lemma 14. *Let $\hat{\ell}^f : \hat{\mathcal{R}} \times \mathcal{Y} \to \mathbb{R}_+$, where $\hat{\mathcal{R}} = \{(A, B) \in 2^N \times 2^N : A \cap B = \emptyset\}$, be given by*

$$\hat{\ell}^f((A, B), S) = L^f(\chi_A + \mathbb{1}_B, S) = f(A \triangle S \setminus B) + f(A \triangle S \cup B) . \tag{17}$$

Then the Lovász hinge L^f embeds $\hat{\ell}^f$.

Proof. As observed above, the set $\mathcal{U}$ is in bijection with $\hat{\mathcal{R}}$, using the transformation $u = \chi_A + \mathbb{1}_B$. As also observed above, we may write $u = \chi_A \odot \mathbb{1}_{N \setminus B}$ as well. Combining Lemma 13 with 3, we see that L^f embeds $L^f|_{\mathcal{U}}$. It thus only remains to verify the form (17). We have $L^f(u, S) = F(x)$ where $x = \mathbb{1} - u \odot \chi_S = \mathbb{1} - \mathbb{1}_{N \setminus B} \odot \chi_A \odot \chi_S = \mathbb{1} + \mathbb{1}_{N \setminus B} \odot \chi_{A \triangle S} = \mathbb{1}_B + 2\mathbb{1}_{A \triangle S \setminus B}$. As $(A \triangle S \setminus B) \cup B = A \triangle S \cup B$, the result follows from [4, Prop 3.1(h)]. $\quad\square$

From the form (17), we see that $\hat{\ell}^f$ matches 2ℓ when $B = \emptyset$, just as hinge loss embeds twice 0-1 loss. When B is nonempty, it acts as an "abstain set", guaranteeing some loss in the second term, but removing errors in the first term.

E.3 Inconsistency

In light of the previous results, we can see that to show inconsistency we may focus on reports (A, B) with $B \neq \emptyset$. Intuitively, if such a report is ever optimal, then L^f has a "blind spot" with respect to the indices in B, and we can leverage this to "fool" L^f. In particular, we will focus on the uniform distribution $\bar{p}$ on $\mathcal{Y}$, and perturb it slightly depending on B to find an optimal point $u \in \mathcal{U}$ which maps to a suboptimal report. In fact, we will show that one can always find such a point violating calibration, unless f is modular.

Given our focus on the uniform distribution, the following definition will be useful: for any set function f, let $\bar{f} := \mathbb{E}_{S \sim \bar{p}}[f(S)] = 2^{-k} \sum_{S \subseteq N} f(S)$. The next two lemmas relate $\bar{f}$ and $f(N)$ to expected loss and modularity.

Lemma 15. *For all* $(A, B) \in \hat{\mathcal{R}}$, $\hat{\ell}^f((A, B); \bar{p}) \geq f(N)$. *For all* $A \subseteq N$, $\hat{\ell}^f((A, \emptyset); \bar{p}) = 2\bar{f}$.

Proof. Letting $\overline{B} := N \setminus B$ for short

$$\hat{\ell}^f((A, B); \bar{p}) = 2^{-k} \sum_{S \subseteq N} f(A \triangle S \setminus B) + f(A \triangle S \cup B)$$

$$= 2^{-|\overline{B}|} \sum_{T \subseteq \overline{B}} f(T) + f(T \cup B)$$

$$= \frac{1}{2} 2^{-|\overline{B}|} \sum_{T \subseteq \overline{B}} f(T) + f(\overline{B} \setminus T) + f(T \cup B) + f((\overline{B} \setminus T) \cup B)$$

$$\geq \frac{1}{2} \left(f(\overline{B}) + f(\emptyset) + f(N) + f(B) \right)$$

$$\geq \frac{1}{2} \left(f(N) + f(N) \right) = f(N) \,,$$

where we use submodularity in both inequalities. The second statement follows from the second equality above after setting $B = \emptyset$, as then $\overline{B} = N$ and thus T ranges over all of 2^N. $\square$

Lemma 16. *Let* f *be submodular and normalized. Then* $\bar{f} \geq f(N)/2$, *and* f *is modular if and only if* $\bar{f} = f(N)/2$.

Proof. The inequality follows from Lemma 15 with $B = \emptyset$. Next, note that if f modular we trivially have $\bar{f} = f(N)/2$. If f is submodular but not modular, we must have some $S \subseteq N$ and $i \in S$ such that $f(S) - f(S \setminus \{i\}) < f(\{i\})$. By submodularity, we conclude that $f(N) - f(N \setminus \{i\}) < f(\{i\})$ as well; rearranging, $f(\{i\}) + f(N \setminus \{i\}) > f(N) = f(N) + f(\emptyset)$. Again examining the proof of Lemma 15, we see that the first inequality must be strict, as we have one such $T \subseteq N$, namely $T = \{i\}$, for which the inequality in submodularity is strict. $\square$

Theorem 7. *Let* f *be submodular, normalized, and increasing. Then* (L^f, sgn) *is consistent if and only if* f *is modular.*

Proof. If f is modular, then F is linear, and $L^f(\cdot; p)$ is linear on $[-1, 1]^k$. We conclude that $L^f(\cdot; p)$ is minimized at a vertex of the hypercube, meaning L^f embeds $2\ell^f$. (Equivalently, there is always an optimal report $(A, \emptyset) \in \hat{\mathcal{R}}$ for $\hat{\ell}^f$.) Calibration and consistency then follow.

Now suppose f is submodular but not modular. As f is increasing, we will assume without loss of generality that $f(\{i\}) > 0$ for all $i \in N$, which is equivalent to $f(S) > 0$ for all $S \neq \emptyset$; otherwise, $f(T) = f(T \setminus \{i\})$ for all $T \subseteq N$, so discard i from N and continue. In particular, we have $\{\emptyset\} = \arg\min_{S \subseteq N} f(S)$.

Define $\epsilon = \bar{f}/(2\bar{f} - f(N))$, which is strictly positive by Lemma 16 and submodularity of f. Let $p = (1 - \epsilon)\bar{p} + \epsilon\delta_\emptyset$, where again $\bar{p}$ is the uniform distribution, and $\delta_\emptyset$ is the point distribution on $\emptyset$.

From Lemma 15, for all $A \subseteq N$ we have

$$
\begin{aligned}
\hat{\ell}^f((A, \emptyset); p) &= (1 - \epsilon)2\bar{f} + \epsilon\, \hat{\ell}^f((A, \emptyset), \emptyset) \\
&= (1 - \epsilon)2\bar{f} + \epsilon\, 2f(A) \\
&\geq (1 - \epsilon)2\bar{f} > f(N) = \hat{\ell}^f((\emptyset, N); p) \,.
\end{aligned}
$$

As we have some report with strictly lower loss than all reports of the form $(A, \emptyset)$, we conclude that we must have some $(A, B) \in \mathrm{prop}[\hat{\ell}^f](p)$ with $B \neq \emptyset$. We can also see that $\mathrm{prop}[\ell^f](p) = \{\emptyset\}$ by the second equality and the fact that $\{\emptyset\} = \arg\min_{S \subseteq N} f(S)$.

Revisiting L^f, from Lemma 14 and the map between $\mathcal{U}$ and $\hat{\mathcal{R}}$, we have some $u \in \Gamma^f(p)$ which we can write $u = \chi_A + \mathbb{1}_B$. Let $T \subseteq N$ such that $\chi_T = \mathrm{sgn}(u)$ after breaking ties, and note that $A \subseteq T \subseteq A \cup B$. If $T \neq \emptyset$, we are done, as by the above $\emptyset$ optimizes ℓ^f, so we have violated calibration and therefore consistency.

Otherwise, $T = \emptyset$, so $A = \emptyset$ as well. In this case, we will modify p to put weight on $B \neq \emptyset$ instead of $\emptyset$, and will find that u is still optimal for L^f, again violating calibration. To show optimality, let $c = L^f(u; p) = \underline{L}^f(p)$, and note that by symmetry of L^f, for any $S \subseteq N$ we have $c = \underline{L}^f(p^S)$ as well, where $p^S = (1 - \epsilon)\bar{p} + \epsilon\delta_S$. In particular, this will hold for $p^{\bar{B}}$. By Lemma 14, we have $L^f(u, B) = f(\emptyset \triangle B \setminus B) + f(\emptyset \triangle B \cup B) = f(\overline{B}) + f(N) = f(\emptyset \triangle \emptyset \setminus B) + f(\emptyset \triangle \emptyset \cup B) = L^f(u, \emptyset)$. Thus, $L^f(u, p^B) = (1 - \epsilon)L^f(u, \bar{p}) + \epsilon L^f(u, B) = (1 - \epsilon)L^f(u, \bar{p}) + \epsilon L^f(u, \emptyset) = L^f(u, p) = c$, so u is still optimal. As $\chi_B \neq \chi_T = \mathrm{sgn}(u)$, we are done. □

F   Top-k surrogate

Throughout this section, consider the surrogate and discrete loss $L^k(u)_y = \left(\frac{1}{k}\sum_{i=1}^{k}(u + \mathbb{1} - e_y)_{[i]} - u_y\right)_+$ given in Equation 9.

Lemma 17. L^k is a polyhedral loss.

Proof. Observe L^k can be written as the pointwise max of $\binom{n}{k} + 1$ terms, where the $\binom{n}{k}$ terms are selecting the k elements of $u + \mathbb{1} - e_y$, and the max comes from selecting the u_i elements with highest weight. □