An Embedding Framework for Consistent Polyhedral Surrogates

Jessie Finocchiaro
jefi8453@colorado.edu
CU Boulder

Rafael Frongillo
raf@colorado.edu
CU Boulder

Bo Waggoner
bwag@colorado.edu
CU Boulder

Abstract

We formalize and study the natural approach of designing convex surrogate loss functions via embeddings for problems such as classification, ranking, or structured prediction. 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 some way 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

[JF:

- Add figures of Bayes risks of 0-1, logistic, and hinge losses (Jessie: Have made them, not sure where would be the best place to insert them/if you think this would be helpful)
- Formalize reverse implication proof of Proposition 1 with projections into the affine hull and back
- Mention how our focus on polyhedral risks and general links extends DKR Section 3.1 (probably after Prop 1)
- Replace Proposition 2 with Lemma 7?
- Theorem 2: move from n-dimensional construction to n-1 dimensional construction (or add n-1 dim construction in appendix if it gets too hairy?)
- Links: cool to point out that when L is polyhedral, prop{L} has a finite range (this result) and so does its (multivalued map) inverse (trim result)
- Links: sketch vs full proof
- Commentary on how any loss embedding ell must be polyhedral ish

]

Convex surrogate losses are a central building block in machine learning for finite prediction problems such as classification and structured prediction tasks. A growing body of work seeks to design and analyze convex surrogates for given loss functions, and more broadly, understand the best empirical risk minimization bounds that can found for a surrogate, for which consistency is a necessary condition. For example, recent work has developed tools to bound the required number of dimensions of the surrogate's hypothesis space [15, 28]. Yet in some cases these bounds are far from tight, such

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

as for *abstain loss* (classification with an abstain option) [5, 28, 29, 38, 39]. 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 roughly maps approximately L-optimal reports to ℓ-optimal reports, then consistency follows [2]. Common examples of this general construction include hinge loss as a surrogate for 0-1 loss and the abstain surrogate mentioned above [29].

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.

Related works. The literature on convex surrogates focuses mainly on smooth surrogate losses [4–6, 8, 9, 24, 30, 35, 40]. Nevertheless, nonsmooth losses, such as the polyhedral losses we consider, have been proposed and studied for a variety of classification-like problems [21, 36, 37]. Moreover **?**] describes the impact of the hypothesis class has on consistency, and when consistency relative to the hypothesis class differs from Bayes consistency; the latter is what we describe in this paper when we say "consistency."

A notable addition to this literature is Ramaswamy et al. [29], 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 $n-1$ dimensions. Their surrogate is a natural example of an embedding (cf. § 6.1), and serves as inspiration for our work.

While property elicitation has by now an extensive literature [12, 14, 17, 19, 20, 25, 32, 33], these works are mostly concerned with point estimation problems. Literature directly connecting property elicitation to consistency is sparse. However, Agarwal and Agarwal [2] consider single-valued properties in finite outcome settings, whereas finite properties elicited by general convex losses are necessarily set-valued. **?**] additionally relates indirect property elicitation to consistency in finite outcome settings when one is given either a target loss or property in both discrete and continuous prediction settings, assuming surrogates attain their infimum in expectation over all distributions over the outcomes.

JF: cvx-flats paper; might need to throw on arxiv if we want this ref.

JF: flag

2 Setting

For discrete prediction problems like classification, due to hardness of directly optimizing a given discrete loss, many machine learning algorithms minimize 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 study *calibration* (Definition 4), which is equivalent to consistency in finite outcome settings [6, 28**?**] and depends solely on the conditional distribution over $\mathcal{Y}$. Consistency is a prerequisite to obtain empirical risk minimization bounds, motivating our firm requirement of constructing consistent surrogates.

2.1 Notation and Losses

Let $\mathcal{Y}$ be a finite 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 (requiring $\|p\|_1 = 1$). We write p_y for the probability of outcome $y \in \mathcal{Y}$ drawn from $p \in \Delta_{\mathcal{Y}}$. We first discuss the conditional setting, with just labels $\mathcal{Y}$ and no features $\mathcal{X}$, and show in § 2.3 how these notions relate to the usual $\mathcal{X} \times \mathcal{Y}$ setting considering features.

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 $\underline{\ell}$ with the infimum over $\mathcal{R}$.

JF: discrete vs target?

BTW: Cut generic loss

For example, 0-1 loss is a discrete loss with $\mathcal{R} = \mathcal{Y} = \{-1, 1\}$ given by $\ell_{0\text{-}1}(r)_y = \mathbb{1}\{r \neq y\}$, with Bayes risk $\underline{\ell_{0\text{-}1}}(p) = 1 - \max_{y \in \mathcal{Y}} p_y$. Two important surrogates for $\ell_{0\text{-}1}$ are hinge loss $L_{\text{hinge}}(u)_y = (1 - yu)_+$, where $(x)_+ = \max(x, 0)$, and logistic loss $L(u)_y = \log(1 + \exp(-yu))$ for $u \in \mathbb{R}$. See Figures 1, 2, and 3 for a visualization of the Bayes risks of 0-1, Hinge, and Logistic losses, respectively.

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 [31].

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. To motivate our focus on polyhedral losses, we echo Ramaswamy et al. [29, Section 1.2], who note that smooth surrogates often encode much more information than necessary, and in these cases non-smooth surrogates are the best candidates to achieve a low input dimension d.

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 $r = 1$ and $r = -1$ 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 $\Gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}} \setminus \{\emptyset\}$.

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 $\Gamma_r := \{p \in \Delta_{\mathcal{Y}} : r \in \Gamma(p)\}$.*

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. In general, by optimal, we mean minimizing an associated loss function in expectation over p, which we formalize shortly. Note that our definitions align such that discrete losses elicit finite properties; both are non-redundant in the correct senses. 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 L elicits a unique property, 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, which require that the excess loss of some report is bounded by (a constant times) the excess loss of the linked report.

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 $p \in \Delta_{\mathcal{Y}}$,*

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

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

It is well-known in finite-outcome settings that calibration 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. Note that in the finite-outcome setting, calibration is necessary and sufficient for consistency from a generalization of Tewari and Bartlett [34] given by Ramaswamy and Agarwal [28]. While our notion of embedding is sufficient for calibration (and therefore consistency), it is worth noting that it is not *necessary* for these conditions. For example, while logistic loss does not embed 0-1 loss, the surrogate and link for logistic loss are consistent with respect to 0-1 loss.

JF: is equivalent to?

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 (Embedding a loss)**.** *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*

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

It is not immediately 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 obvious. We address the question of when embeddings lead to calibrated links in § 5.

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

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

In particular, we see that $-1 \in \text{prop}[\ell_{0\text{-}1}](p) \iff p_1 \in [0, 1/2] \iff -1 \in \text{prop}[L_{\text{hinge}}](p)$, and $1 \in \text{prop}[\ell_{0\text{-}1}](p) \iff p_1 \in [1/2, 1] \iff 1 \in \text{prop}[L_{\text{hinge}}](p)$. With both conditions of Definition 5 satisfied, we conclude that L_{hinge} embeds $2\ell_{0\text{-}1}$. In this particular case, it is known (L_{hinge}, ψ) is calibrated for $\psi(u) = \text{sgn}(u)$; in § 5 we show that, perhaps surprisingly, all embeddings imply the existence of a link so that the surrogate is calibrated.

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 the question of when such surrogates are consistent to § 5.

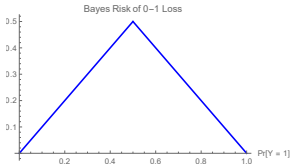


Figure 1: Bayes Risk of the 0-1 loss.

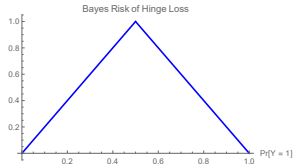


Figure 2: Bayes Risk of Hinge Loss.

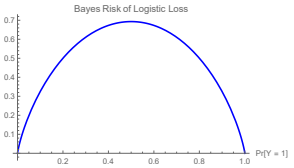


Figure 3: Bayes Risk of the logistic loss.

To begin, we observe that 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. [JF: Intuition: $\implies$: Take the injection given and take $\mathcal{U} = \varphi(\mathcal{R})$. Since we embed, the loss values match at these embedding points, and we know the surrogate has no "strictly better" report than the embedding, otherwise we would contradict the fact that L embeds ℓ, so the embedded report must be in the argmin of the surrogate restricted to $\mathcal{U}$, that is, $L|_{\mathcal{U}}$. At any $p \in \Delta_{\mathcal{Y}}$, suppose $r \in \arg\min_{r'} \langle \ell(r), p \rangle$. As $\ell(r) = L(\varphi(r))$, we then have $\langle p, \ell(r) \rangle = \langle p, L(\varphi(r)) \rangle$, and therefore we observe $\underline{L|_{\mathcal{U}}} = \underline{\ell}$ on $\Delta_{\mathcal{Y}}$. The property γ' elicited by $L|_{\mathcal{U}}$ is then nonempty for all $p \in \Delta_{\mathcal{Y}}$. Moreover, we have $\underline{L} = \underline{L|_{\mathcal{U}}}$ by construction of γ' and its nondegeneracy. $\impliedby$: We know $\underline{\ell}$ is polyhedral, and classic results (Aurenhammer) show that the projection of the facets of the epigraph of $\underline{\ell}$ form a power diagram on the simplex. Each cell of this power diagram corresponds to a level set of γ_r. A corollary of each level set being a cell in a power diagram is that each γ_r is full-dimensional. We know there is a report u_r such that $u_r \in \Gamma(\tilde{p})$ for some $\tilde{p} \in \mathring{\gamma}_r$. Moreover, the point $(p, -\langle p, L(u_r) \rangle)$ forms a supporting hyperplane of $-\underline{L}$ (and therefore $-\underline{\ell}$ as well) if and only if $u_r \in \Gamma(p)$ for $p \in \Gamma_{u_r}$. As $(p, -\langle p, L(u_r) \rangle)$ also forms a supporting hyperplane of $-\underline{\ell}$, we must then have $p \in \gamma_r$. As this is true for all $p \in \Gamma_{u_r}$, we must have $\Gamma_{u_r} = \gamma_r$. Losses match by considering the supporting hyperplanes to the facets of the epigraphs of the Bayes Risks are the same, and the loss values are uniquely determined by the final coordinate by this final coordinate of the supporting hyperplane.] 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 embedding condition (i), we have

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

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 $\Delta_{\mathcal{Y}}$ by taking the appropriate projection into $\mathbb{R}^{n-1}_+$, so that interiors are well-defined, and $\underline{\ell}$ may be differentiable on the interior of $\Delta_{\mathcal{Y}}$. (For intuition, one can consider the relative interiors in $\mathbb{R}^n_+$, but we need to consider interiors in order to apply previous results.) 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 5, whose cells are full-dimensional (in $n-1$ dimensions) 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 $v \in \mathbb{R}^d$ the hyperplane $v \mapsto \langle v, -L(u_r) \rangle$ supports the epigraph E_L of $-\underline{L}$ at the point $(p, -\langle p, L(v) \rangle)$ if and only if $v \in \Gamma(p)$. 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.

RF: $\impliedby$: First 2 sentences are good, but then it fogs up a bit. 3rd needs another careful pass, e.g., supporting hyperplane at what point? 4th cannot be right since you don't cover all reports r [JF: Took another pass]

BTW: RF: This is on the informal side, mostly because of the "working in the affine hull" bit. It can all be formalized though, probably should be for the journal version: just take an appropriate linear projection of $\mathbb{R}^n$ to $\mathbb{R}^{n-1}$, and work with the projected versions of everything, and then project back.

□

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 the Bayes risk, a function from distributions over $\mathcal{Y}$ to the reals, may be polyhedral even if the loss itself is not. For example, this can be seen by a loss that is polyhedral in the convex hull of $\phi(\mathcal{R})$, but smooth outside that region.

Previous work from Duchi et al. [9, Proposition 4] realized the significance of matching Bayes risks for calibration with respect to the 0-1 loss, but their result relies the Bayes risk of the surrogate being strictly concave. As polyhedral functions are never strictly concave, Proposition 2 extends their result to the embedding setting.

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

Proof. [JF: Intuition: $\Leftarrow$: There are a finite set of full-dimensional level sets, which we enumerate. We then use these enumerated reports to define a finite property γ such that $u_i \in \Gamma(p) \implies i \in \gamma(p)$, and take the discrete loss values equal to original loss values at the enumerated points.] 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 the epigraph of $\underline{L}$ onto $\Delta_\mathcal{Y}$ forms a power diagram by Theorem 5 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. □

Combining Proposition 2 with the observation that polyhedral losses have polyhedral Bayes risks (Lemma 7), 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, 13]). Note however that the number of dimensions d required could be as large as $|\mathcal{Y}|$, which is particularly undesirable in structured prediction problems with exponentially many outcomes. Recent work [11, 28] yield characterizations for bounding the convex calibration and embedding dimensions.

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 [31, 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. □

BTW: FUTURE: to get $n - 1$, just use the same trick as before and note that the Bayes risk doesn't change. Should be a couple lines.

Moreover, to construct an embedding of dimension $n - 1$ instead of n (see Finocchiaro et al. [11] for significance of dimension), we use the same trick as in the proof of Proposition 1. For $p \in \Delta_\mathcal{Y}$,

consider p_{-n} to be the first $n-1$ coordinates of p identically projected into $n-1$ dimensional space subject to the affine constraint $\sum_y p_y = 1$, so that $p_n = 1 - \sum_{y=1}^{n-1} p_y$. Take $L' : \mathbb{R}^{n-1} \to \mathbb{R}^{\mathcal{Y}}$ defined by $u \mapsto (-\ell')^*(u)\mathbb{1} - u$, where ℓ' is ℓ projected into $n-1$ dimensions as in Proposition 1, so that $\langle \ell, p \rangle = \ell'(p_{-n})$ for all $p \in \Delta_{\mathcal{Y}}$. The Bayes risks $\underline{L}(p)$ and $\underline{L'}(p_{-n})$ are equal for all $p \in \Delta_{\mathcal{Y}}$; thus, the result holds in $n-1$ dimensions as well.

JF: Leaving this change in the macro since we haven't made the change in proposition 1 yet.

We will see that embedding implies the existence of a consistent surrogate and link, so this result tells us that, for discrete losses, the embedding framework is sufficient to understand the existence of consistent surrogates for the given loss. Even if there is a smooth surrogate for a given discrete loss, there is also a polyhedral loss that is calibrated for the given loss. What's more is that there is a polyhedral *embedding* that is calibrated for the discrete loss.

4 Embedding properties

BTW: JESSIE: Commented out entire appendix section for the time being. All that was there that isn't here is some commentary about positive normal sets.

While Definition 5 gives the notion of one *loss* embedding another, we now generalize the notion of one *property* embedding another.

Definition 6. *A property $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathbb{R}^d$ embeds a property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ if there exists some injective embedding $\varphi : \mathcal{R} \to \mathbb{R}^d$ such that for all $p \in \Delta_{\mathcal{Y}}$ and $r \in \mathcal{R}$, we have $r \in \gamma(p) \iff \varphi(r) \in \Gamma(p)$. Moreover, we say a loss embeds a property when it embeds a loss eliciting the property.*

BTW: JESSIE: Should this section be before or after embedding losses, since the results in this section are used in the proofs in the embedding losses section? Narrative of having props then losses: here is this toolkit we have for understanding this notion of embedding, now here is how we apply it to the machine learning thing you actually care about. Narrative of losses then props: BOOM here's this cool thing, and btw this is how we show it. I think for a journal I'm inclined towards the latter, but that opinion is not very strong.

By condition (ii.) of Definition 5, we then have L embedding ℓ implies $\text{prop}[L]$ embeds $\text{prop}[\ell]$. However, Definitions 5 and 6 are not immediately equivalent because property embedding does not capture the requirement of matching losses on embedded points; i.e., that $L(\varphi(r)) = \ell(r)$ for all $r \in \mathcal{R}$. However, Lemma 1 shows an equivalence follows without loss of generality.

Lemma 1. *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a loss function. If the property $\Gamma := \text{prop}[L]$ embeds the finite property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$, then there is a discrete loss ℓ such that ℓ elicits γ and L embeds ℓ.*

Proof. Consider a loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, defined so that $\ell : r \mapsto L(\varphi(r))$, i.e. $\ell = L|_{\varphi(\mathcal{R})}$, where φ is the same embedding given by Γ embedding γ. Now, we claim that L embeds ℓ. Using Proposition 1, showing $\underline{L} = \underline{\ell}$ suffices to show that L embeds ℓ.

First, we show that for all $p \in \Delta_{\mathcal{Y}}$, we have $(1) \inf_{u \in \mathbb{R}^d} \langle p, L(u) \rangle = (2) \inf_{r \in \mathcal{R}} \langle p, L(\varphi(r)) \rangle$, and the equality of risks follows. $(1) \leq (2)$ follows from the fact that $\varphi(\mathcal{R}) \subset \mathbb{R}^d$ and definition of infimum. Consider that if we did not have $(1) \geq (2)$ for some $p \in \Delta_{\mathcal{Y}}$, then there would be no r such that $\varphi(r) \in \Gamma(p)$. Therefore the property it embeds, γ is degenerate at p, i.e. $\gamma(p) = \emptyset$, yielding a contradiction. This gives us $(1) = (2)$, so we have $\underline{L} = \underline{L|_{\varphi(\mathcal{R})}} = \underline{\ell(r)}$ by construction of ℓ.

□

JF: New to journal version

RF: All the important pieces are here, but take it slower. State ℓ as a definition (let ℓ be given by...) and show why L embeds it.[JF: Took another pass but changed technique, adding in this detail. Now very similar to the proof of Prop 1 forward direction.]

By Lemma 1, we can work directly with properties and set aside the losses which elicit them, which is often more convenient. Moreover, we can assume that finite properties are non-redundant without much loss of generality.

Definition 7 (Finite property, non-redundant)**.** *A property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ is* redundant *if for some $r, r' \in \mathcal{R}$ with $r \neq r'$, 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.*

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 $\langle p, L \rangle$ is flat, then there will be an uncountable set of reports which also minimize the expected loss. As results in property elicitation typically assume properties are non-redundant (e.g. [13, 15]), it is useful to consider a transformation which removes redundant level sets, captured by the *trim* operation.

Definition 8. *Given an elicitable property $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$, we define $\text{trim}(\Gamma) = \{\Gamma_u : u \in \mathcal{R} \text{ s.t. } \neg \exists u' \in \mathcal{R}, u' \neq u, \Gamma_u \subsetneq \Gamma_{u'}\}$ as the set of maximal level sets of Γ.*

Before we state Proposition 3, which is 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 5), and its proof is omitted. The second is closely related to the definition of 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 of 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 $p \in \Delta_{\mathcal{Y}}$. □

BTW: JESSIE: Again, figures/an example might help clarify this

When a property Γ embeds a finite property γ, we can show that the level sets of γ correspond exactly to their embedded level sets of Γ, and that these embedded level sets are exactly $\mathrm{trim}(\Gamma)$.

JF: This lemma new to journal version

Lemma 4. *Let Γ be an elicitable property. If Γ embeds a (non-redundant) finite property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ by the injection φ, then $\{\gamma_r : r \in \mathcal{R}\} = \{\Gamma_u : u \in \varphi(\mathcal{R})\} = \mathrm{trim}(\Gamma)$.*

Proof. Let L elicit Γ. Take ℓ to be the loss embedded by L from Lemma 1. Moreover, by Proposition 1, we have $\underline{L} = \underline{\ell}$. As γ is finite (and non-redundant by assumption), we know that each level set of γ must be full-dimensional. For all $r \in \mathcal{R}$, we know $\gamma_r = \Gamma_{\varphi(r)}$, so we must also have $\Gamma_{\varphi(r)}$ full-dimensional. Moreover, we have $\{\gamma_r : r \in \mathcal{R}\} = \{\Gamma_{\varphi(r)} : r \in \mathcal{R}\}$ as a corollary since this is true for each $r \in \mathcal{R}$. Since the cell $\Gamma_{\varphi(r)}$ is the projection of a facet of the epigraph of $\underline{L}$, which is concave, any other level set intersecting $\Gamma_{\varphi(r)}$ can be considered in one of two cases: First, if the level set Γ_u (for $u \notin \varphi(\mathcal{R})$) is a projection of a lower-dimensional face of $\underline{L}$, which is contained in a facet of the epigraph. Therefore, we must have $\Gamma_u \subset \Gamma_{\varphi(r)}$ for some $r \in \mathcal{R}$, and therefore $\Gamma_u \notin \mathrm{trim}(\Gamma)$. Second, $\Gamma_u = \Gamma_{\varphi(r)}$ for some $r \in \mathcal{R}$, then the level set is in both $\{\Gamma_{\varphi(r)} : r \in \mathcal{R}\}$ and $\mathrm{trim}(\Gamma)$. □

RF: Implicitly? Justify the claim. [JF: In definition of finite properties, we say that we assume we are talking about non-redundant.]

RF: The main action is here. (1) Epigraph of $-\underline{L}$. (2) There is really only one case: every level set is a face of the power diagram, which must be contained (weakly, so $\subseteq$) in a facet. (3) To show that level sets are faces, you'll want to appeal to some piece of a proof in the prev section.

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 only a slight 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. [JF: Consolidated; original below.] Let L elicit Γ.

$1 \Rightarrow 2$: If Γ embeds a finite γ, then Lemma 1 says there is be a discrete loss ℓ embedded by L such that ℓ elicits γ. By Lemma 4, we then observe that $\mathrm{trim}(\Gamma) = \{\Gamma_{\varphi(r)} : r \in \mathcal{R}\} = \{\gamma_r : r \in \mathcal{R}\}$.

RF: Again, a bit too brief. Consider adding some of this statement into the Lemma. [JF: Is this better?]

As $\mathcal{R}$ is finite, we conclude that $\mathrm{trim}(\Gamma)$ is also finite. Moreover, as γ is nondegenerate (recall that $\gamma : \Delta\mathcal{Y} \rightrightarrows \mathcal{R}$ is shorthand for $\gamma : \Delta\mathcal{Y} \to 2^{\mathcal{R}} \setminus \emptyset$), we have that $\cup\mathrm{trim}(\Gamma) = \cup_{r\in\mathcal{R}}\gamma_r = \Delta\mathcal{Y}$.

RF: These next two directions together seem like what we did before for 2 ⇒ 3, correct? (Not as complete though; see the step where we justify the conditions of Lemma 3.) I'm actually happy for them to be separated, but let's make sure the argument is just as complete. [JF: Added in the conditions of Lemma 3 to be a bit more thorough.]

$2 \Rightarrow 1$: As $\mathrm{trim}(\Gamma)$ is finite, we know $\mathrm{trim}(\Gamma) = \{\Gamma_{u_i}\}_{i=1}^k$ for some finite k. Take $\mathcal{U} = \{u_1, \ldots, u_k\}$, and consider L eliciting Γ and $\ell := L|_{\mathcal{U}}$. We have $\mathcal{U} \subset \mathbb{R}^d$ and $\Gamma(p) \cap \mathcal{U} \neq \emptyset$ for all p, since we have $p \in \Gamma_{u_i}$ for some $u_i \in \mathcal{U}$. Then we can apply Lemma 3 to observe that L embeds ℓ, and therefore Γ embeds $\mathrm{prop}[\ell]$, which is finite as ℓ is discrete.

$1 \Rightarrow 3$: Let $\varphi : \mathcal{R} \to \mathbb{R}^d$ by the embedding function, and take $\Theta = \{\Gamma_{\varphi(r)} : r \in \mathcal{R}\}$. The latter set is equal to $\{\gamma_r : r \in \mathcal{R}\}$ by Γ embedding γ. Therefore, we know each $\theta \in \Theta = \gamma_r$ for some $r \in \mathcal{R}$, which is full-dimensional in the simplex. Moreover, since γ is non-degenerate, we have $\cup\Theta = \cup_r\gamma_r = \Delta\mathcal{Y}$.

$3 \Rightarrow 1$: [JF: Intuition: Take $\gamma : p \mapsto \Gamma(p) \cap \varphi(\mathcal{R})$.] 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.

□

JF: Can we remove this?

Original. 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 5.) 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. □

4.1 Refining properties

BTW: JESSIE: This needs more narrative text. I also don't know if this should go in the Appendix, given it doesn't really connect to the rest of the paper, though I think the point it makes (surrogates work for nearly any indirectly elicited property) is worth having in the body.

Definition 9. *Let $\Gamma : \Delta\mathcal{Y} \rightrightarrows \mathcal{R}$ and $\Gamma' : \Delta\mathcal{Y} \rightrightarrows \mathcal{R}'$. Then Γ'* refines *Γ if for all $r' \in \mathcal{R}'$, we have $\Gamma'_{r'} \subseteq \Gamma_r$ for some $r \in \mathcal{R}$. That is, the cells of Γ' are all contained in the cells of Γ.*

Theorem 3. *Every polyhedral loss embeds a finite elicitable property. Moreover, a polyhedral loss L indirectly elicits a finite elicitable property γ if and only if γ is finite and L embeds a property which refines γ.*

Proof. [JF: Intuition: There are only a finite set of possible vertices of the loss, and claim that for each $p \in \Delta\mathcal{Y}$, one of these vertices is a minimizer of $\langle p, L(\cdot)\rangle$. Rinse and repeat for vertices of the expected loss on all possible (finite) supports. As there is a some vertex in the property value for every $p \in \Delta\mathcal{Y}$, we have $\mathrm{trim}(\Gamma)$ finite, which yields the results via Prop 3 somehow? Unclear on that gap.] [JF: Cut everything but last paragraph] The first statement is a trivial corollary of Theorem 1.

RF: I might be delusional, but this second part ended up being much slicker than I'd thought, by essentially chaining definitions and maps. Please check!

For the second part, let $\gamma' : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}'$ be the finite elicitable property embedded by L, with embedding $\varphi : \mathcal{R}' \to \mathbb{R}^d$, and let ψ be a link to a non-redundant elicitable property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$. Then letting $\psi' = (\psi \circ \varphi) : \mathcal{R}' \to \mathcal{R}$, we see that ψ' is a link from γ' to γ: for all $r' \in \mathcal{R}'$, we have $\gamma'_{r'} = \text{prop}[L]_{\varphi(r')} \subseteq \gamma_{\psi(\varphi(r'))} = \gamma_{\psi'(r')}$. In particular, γ' refines γ, and as γ' is finite, γ must be finite. □

[JF: I am going to suggest cutting this; I don't think it adds much to the narrative and forward refs results from calibrated links.]

Corollary 1. *If a polyhedral loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}$ indirectly elicits a property $\gamma' : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}'$, then for any loss ℓ' eliciting γ', there exists a link ψ such that (L, ψ) is calibrated with respect to ℓ'.*

JF: Added 7.24.20. Probably too long to be a Corollary proof, but would rather start too long than too short. Also the proof has a forward ref

Proof. By Theorem 3, we know that $\Gamma := \text{prop}[L]$ embeds a property $\gamma \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ which refines γ'. Therefore, these exists a calibrated link $\psi : \mathbb{R}^d \to \mathcal{R}$ from Γ to γ. [JF: Relies on results in the next section...] Moreover, as γ refines γ', take any link $\psi' : r \mapsto r'$, where $\gamma_r \subseteq \gamma'_{r'}$. (We know such a link exists by refinement.) Consider $\tilde{\psi} := \psi \circ \psi'$.

Now, we show that $(L, \tilde{\psi})$ is calibrated with respect to ℓ'. Consider that since Γ embeds γ, we have a link $\bar{\psi}$ such that $(L, \bar{\psi})$ is calibrated with respect to ℓ eliciting γ. It then suffices to show

$$\begin{aligned} \inf_u \langle p, L(u) \rangle &< \inf_{u : \psi(u) \notin \gamma(p)} \langle p, L(u) \rangle \\ &\leq \inf_{u : \psi'(u) \notin \gamma'(p)} \langle p, L(u) \rangle \ . \end{aligned}$$

This first statement is true as embedding implies calibration (Theorem 4), and the second is true if $\{u : \psi'(u) \notin \gamma'(p)\} \subseteq \{u : \psi(u) \notin \gamma(p)\}$; we show the contrapositive. Take $u \in \mathbb{R}^d$ so that $\psi(u) \in \gamma(p)$.

$$\begin{aligned} \psi(u) \in \gamma(p) &\iff \psi' \circ \psi(u) \in \psi' \circ \gamma(p) \\ &\iff \tilde{\psi}(u) \in \gamma'(p) \ , \end{aligned}$$

where the last inequality follows by construction of ψ'. □

5 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 C 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 choice $\psi(u) \in \mathcal{R}$ that is ℓ-optimal for both distributions. 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 6). 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 that these choices are well-defined uses a chain of arguments involving the Bayes risk, ultimately showing that if u lies in multiple such sets U, the corresponding report sets R_U all intersect at some $r =: \psi(u)$.

Intuitively, to ensure calibration, 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 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 10. *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.

BTW: JOURNAL: cool to point out that when L is polyhedral, $\mathrm{prop}[L]$ has a finite range (this result) and so does its (multivalued map) inverse (trim result)!

Theorem 4. *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. □

BTW: FUTURE: we should comment in the discussion section that we probably can show that *any* loss embedding ℓ must be polyhedral-ish, meaning polyhedral except for stuff that is never optimal. This theorem would then not need the "polyhedral" part. This is related to the "convex envelope conjecture", that if L embeds ℓ via φ, you can just take the loss L' such that L_y is the convex envelope of points $\{(\varphi(r), L(r)_y)\}_{r \in \mathcal{R}}$.

Note that the construction given above in Definition 10 is not necessarily computationally efficient as the number of labels n grows. In practice this potential inefficiency is not typically a concern, as the family of losses typically has some closed form expression in terms of n, and thus the construction can proceed at the symbolic level. We illustrate this formulaic approach in § 6.1.

6 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. The examples in § 6, with the exception of the abstain surrogate given by Ramaswamy et al. [29], all present a surrogate that use the link $\psi : u \mapsto \mathrm{sgn}(u)$. While this may sometimes yield a calibrated link, this is not always the case. In fact, we will see that most of these examples do not yield calibrated surrogates, although proving that there is *no* calibrated link for a given surrogate is quite difficult. Our results suggest that it is possible that some of the given surrogates are calibrated, but perhaps one must use a nontraditional link in order to calibrate the loss, such as the thickening construction given here.

BTW: FUTURE: This prop and theorem give an excellent reason to focus on embeddings, since other techniques do not necessarily give you separated links for free. Since we know we get them for free, we can just focus on the property, and study elicitation complexity; we know if we have a link at all it can be taken to be separated. [[Is this true?]]

6.1 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. [29] study the loss $\ell_\alpha : [n] \cup \{\perp\} \to \mathbb{R}_+^{\mathcal{Y}}$ defined by $\ell_\alpha(r)_y = 0$ if $r = y$, α if $r = \perp$, and 1 otherwise. Here, the report $\perp$ 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. [29] 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,

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

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 $\varphi(\perp) = 0 \in \mathbb{R}^d$. Moreover, the link function ψ can be recovered from Theorem 4 with norm $\|\cdot\|_\infty$ and $\epsilon = 1/2$; see Figure 4(L). Hence, our framework would have simplified the

BTW: RF: The top-k and hypercube/set examples (except abstain) have the link built in, e.g. as the sign function), and they just work around it. Our results may suggest that it's worth thinking "outside the box" (HAH, genious...) and looking for embeddings which are not the hypercube. [JF: Added a bit discussing this, starting at "the examples in ...".]

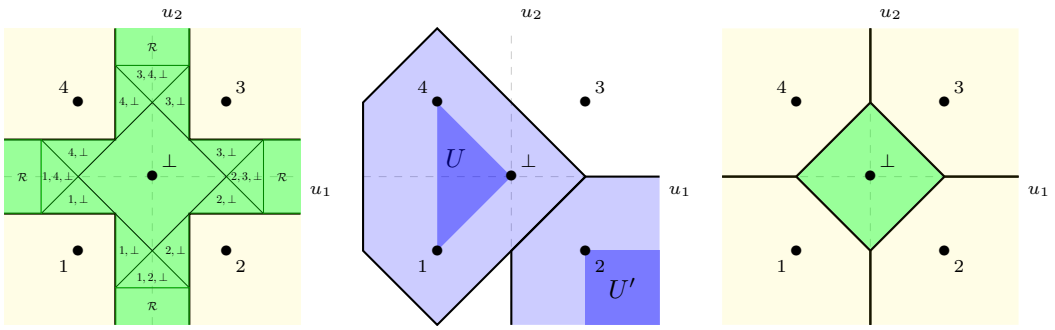


Figure 4: Constructing links for the abstain surrogate $L_{1/2}$ with $d = 2$. The embedding is shown in bold labeled by the corresponding reports. (L) The link envelope Ψ resulting from Theorem 4 using $\|\cdot\|_\infty$ and $\epsilon = 1/2$, and a possible link ψ which matches eq. (5) from [29]. (M) An illustration of the thickened sets from Definition 10 for two sets $U \in \mathcal{U}$, using $\|\cdot\|_1$ and $\epsilon = 1$. (R) The Ψ and ψ from Theorem 4 using $\|\cdot\|_1$ and $\epsilon = 1$.

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 4(R):

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

Theorem 4 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 $\perp$ to much less of the surrogate space $\mathbb{R}^d$. It would be interesting to compare the two links in practice.

6.2 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 [16, 18, 26]. 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 we may write $\ell^g(r)_y = g(\{i \in [k] : r_i \neq y_i\})$ for some set function $g : 2^{[k]} \to \mathbb{R}$. For example, Hamming loss is 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 [37] 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 the full version of the paper [10], 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

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

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 $\psi(u)_i = \operatorname{sgn}(u_i)$, with ties broken arbitrarily.

Let us consider the coefficients $g_\emptyset = 0$, $g_1 = g_2 = g_{1,2} = 1$, for which ℓ^g is merely 0-1 loss on $\mathcal{Y}$. For consistency, for any distribution $p \in \Delta_\mathcal{Y}$, we must have that whenever $u \in \arg\min_{u' \in \mathbb{R}^2} \langle p, L^g(u') \rangle$, the outcome $\psi(u)$ must be the most likely, i.e., in $\arg\max_{y \in \mathcal{Y}} p(y)$. Simplifying eq. (7), however, we have

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

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 can simply take a distribution which disagrees with the way ties are broken in ψ. For example, if we take $\operatorname{sgn}(0) = 1$, then under $p((1,1)) = p((1,-1)) = p((-1,1)) = 0.2$ and $p((-1,-1)) = 0.4$, we have $\{0\} = \arg\min_{u \in \mathbb{R}^2} \langle p, L^g(u) \rangle$, yet we also have $\psi(0) = (1,1) \notin \{(-1,-1)\} = \arg\min_{r \in \mathcal{R}} \langle p, \ell^g(r) \rangle$.

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 this 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 the full version of the paper [10], 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.

6.3 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 $\psi(u) = \{u_{[1]}, \ldots, u_{[k]}\}$, where $u_{[i]}$ is the i^{th} largest entry of u.

Lapin et al. [21, 22, 23] provide the following convex surrogate loss for this problem, which Yang and Koyejo [36] show to be inconsistent:[1]

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

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 14), we know from Theorem 1 that it embeds *some* discrete loss ℓ^k, and from Theorem 4 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 $L^2(u)_y = \left(1 - u_y + \frac{1}{2}(u_{[1]} + u_{[2]} - \min(1, u_y))\right)_+$. 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 $\ell^2(r)_y = 1 - r_y + \frac{1}{2}\langle r, \mathbb{1} - e_y \rangle$ otherwise. Observe that ℓ^2 is just $\ell^{\text{top-2}}$ with an extra term punishing weight on elements other than y, and a reward for a weight of 2 on y.

Moreover, we can visually inspect the corresponding properties (Fig. 5) 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.

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

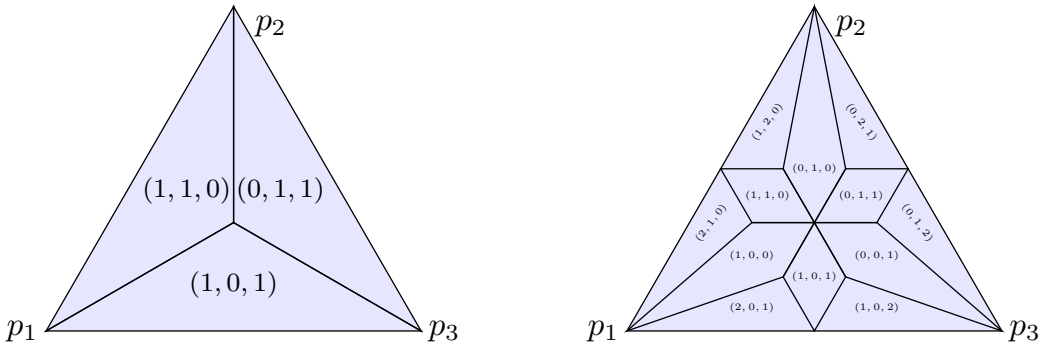


Figure 5: Minimizers of $\langle p, \ell^{\text{top-2}} \rangle$ and $\langle p, \ell^2 \rangle$, respectively, varying p over Δ_3.

BTW: New example using hierarchical surrogate

6.4 Embedding hierarchical classifications

Our final example comes from Ramaswamy et al. [27], who present a convex, calibrated surrogate for the hierarchical classification problem. Given a tree $H = ([n], E, W)$ over finite class labels, they consider the discrete loss

$$\ell^H(r, y) = \text{shortest path length in } H \text{ between } r \text{ and } y \tag{10}$$

In this setting, the outcomes $\mathcal{Y}$ are composed of all nodes in the tree and not just the leaves.

In [27, Theorem 1], Ramaswamy et al. show the property $\gamma^H := \text{prop}[\ell^H]$ is the deepest node i in the tree H such that the probability that node i or one of its descendants is the outcome, denoted $S_i(p)$, is greater than or equal to $1/2$. They additionally note that this property is agnostic to the weight W of the tree. As before, consider p_i to be the probability that node i is the ground truth label, and $S_i(p)$ to be the probability that node i or one of its descendants is the ground truth. For an example, see Figure 6, where predicting node 3 minimizes the expected tree loss over the distribution $\vec{p} = (0, 0.2, 0.1, 0.4, 0.3)$.

Letting h be the height of the tree H, Ramaswamy et al. present an embedding of ℓ^H that consists of learning h $\text{abstain}_{1/2}$ embeddings, where the outcomes at the j^{th} embedding are the nodes on level j of the tree. At each highest level, one proceeds down the path of the tree given by the prediction of the current abstain embedding. If one predicts $\perp$ at level j, then we predict the node predicted at the $(j-1)^{st}$ level.

Note that the given *Cascade Surrogate* of [27, Equation 2] simply requires the use of a surrogate that is calibrated with respect to $\text{abstain}_{1/2}$, so we can use the BEP surrogate of [29] to concatenate h $\lceil \log_2(n) \rceil$ dimensional embeddings for the tree loss, yielding an embedding for ℓ^H.

[JF: Some notes on dimension of this surrogate.] It is worth noting that the cascade surrogate is not the always most efficient in terms of dimension. For example, the cascading surrogate on the 3 node tree H in Figure 7 takes two 1-dimensional optimization problems for distributions p such that $p_1 < 1/2$. However, the property (Figure 8) elicited by tree distance for H in Figure 7 is embeddable by a real-valued surrogate thanks to the characterization of [11, § 3], so one can see that the dimension cascade surrogate does not yield a tight bound. This leaves the question of embedding dimension open for the hierarchical surrogate embedding the discrete tree loss ℓ^H.

7 Conclusions

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 embeddings 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 (Theorem 4). In fact, we conjecture that *any* loss embedding a discrete ℓ must be

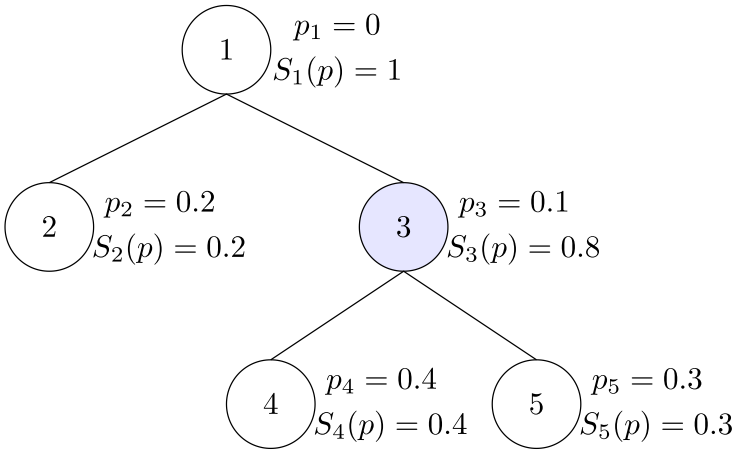


Figure 6: Tree H with distribution $\vec{p} = (0, .2, .1, .4, .3)$. $\gamma^H(\vec{p}) = 3$.

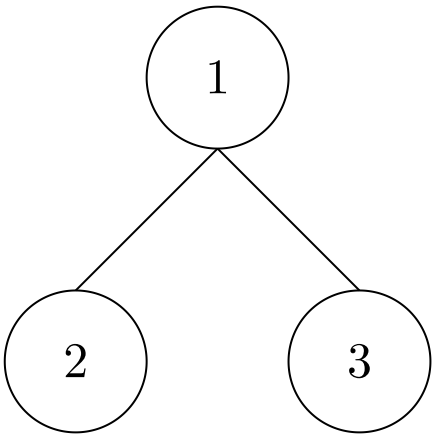


Figure 7: 3 node tree whose property we evaluate in next image.

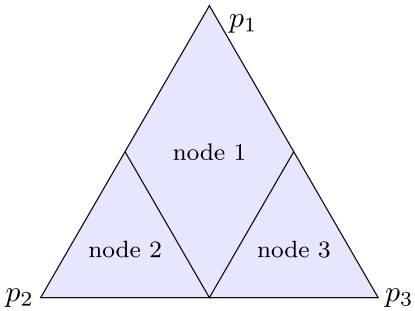


Figure 8: Property elicited by unweighted 3-node tree.

polyhedral on the convex hull of the embedded reports. (The convex hull of the embedded reports follows since any point not in the convex hull will never minimize the expected loss.) Since discrete losses have a finite set of reports, and in turn, minimizers, any surrogate embedding the discrete loss must also have a finite set of unique minimizers. This is in turn related to another conjecture about the "convex envelope" of embeddings: if L embeds ℓ by the embedding φ, the (polyhedral) surrogate L' such that L'_y is the convex envelope of $\{(\varphi(r), L(r)_y)\}_{r \in \mathcal{R}}$ also embeds ℓ. 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. In particular, our link construction recovers the link function proposed by Ramaswamy et al. [29] for abstain loss, as well as another simpler link based on the L_1 norm.

Acknowledgements

We thank Arpit Agarwal and Peter Bartlett for many early discussions, which led to several important insights. We thank Eric Balkanski for help with a lemma about submodular functions. This material is based upon work supported by the National Science Foundation under Grants No. 1657598 and No. DGE 1650115.

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

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

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

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

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

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

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

[16] 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.

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

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Multilabel reductions: what is my loss optimising? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 10600–10611. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9245-multilabel-reductions-what-is-my-loss-optimising.pdf.

[25] 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.

[26] 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.

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

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

[29] 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.

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

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

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

[33] 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.

[34] Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *The Journal of Machine Learning Research*, 8:1007–1025, 2007. URL http://dl.acm.org/citation.cfm?id=1390325.

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

[36] Forest Yang and Sanmi Koyejo. On the consistency of top-k surrogate losses. *CoRR*, abs/1901.11141, 2019. URL http://arxiv.org/abs/1901.11141.

[37] 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.

[38] 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.

[39] 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.

[40] Mingyuan Zhang, Harish G Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label f-measure. 2020.

A Power diagrams

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

Definition 11. *A* cell complex *in $\mathbb{R}^d$ is a set C of faces (of dimension $0, \dots, d$) which (i) union to $\mathbb{R}^d$, (ii) have pairwise disjoint relative interiors, and (iii) any nonempty intersection of faces F, F' in C is a face of F and F' and an element of C.*

Definition 12. *Given sites $s_1, \dots, s_k \in \mathbb{R}^d$ and weights $w_1, \dots, w_k \geq 0$, the corresponding* power diagram *is the cell complex given by*

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

Definition 13. *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 5 (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 5. *Let $f_1, \dots, f_m : \mathbb{R}^d \to \mathbb{R}$ be polyhedral convex functions. The power diagram induced by $\sum_{i=1}^m p_i f_i$ is the same for all $p \in \mathring{\Delta}_{\mathcal{Y}}$.*

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 [31], f, f' are polyhedral. We now show that f is differentiable whenever f' is differentiable:

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

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

B Polyhedral losses

Lemma 6. *Let $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ be a polyhedral loss, and let $\Gamma = \operatorname{prop}[L]$. 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 5, we have that the power diagram $D_{\mathcal{Y}}$ induced by the projection of $P(p)$ onto $\mathbb{R}^d$ is the same for any $p \in \mathring{\Delta}_{\mathcal{Y}}$. Let $\mathcal{F}_{\mathcal{Y}}$ be the set of faces of $D_{\mathcal{Y}}$, which by the above are the set of faces of $P(p)$ projected onto $\mathbb{R}^d$ for any $p \in \mathring{\Delta}_{\mathcal{Y}}$.

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

Now 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 $\mathcal{Y}' \subseteq \mathcal{Y}$. □

Lemma 7. *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 6, $\mathcal{U} = \Gamma(\Delta_{\mathcal{Y}})$ 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 $\underline{L} = \underline{L|_{U'}}$, which is polyhedral as U' is finite. □

BTW: RF: interesting: we didn't need Proposition 2 after all. This already gives the discrete loss L embeds!

C 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

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

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 8. *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.

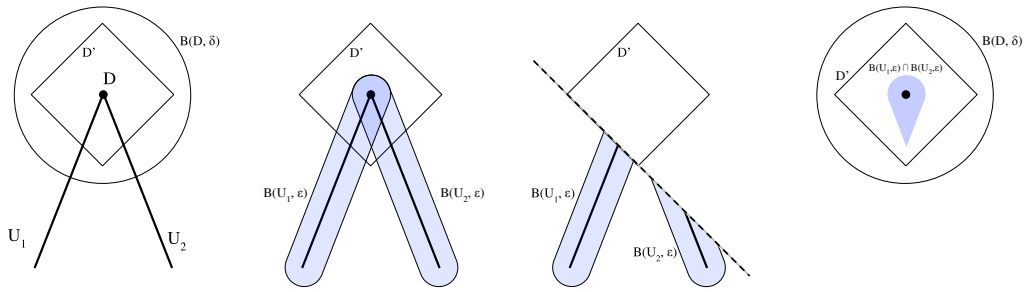
For each $U \in \mathcal{U}'$, for any p such that $U = \Gamma(p)$, we have in particular that u is optimal for p, so $p \in \Gamma_u$, so $p \in \gamma_r$, so r is optimal for p. This implies that $\phi(r)$, the embedding point, is optimal for p, so $\phi(r) \in U$. This holds for all $U \in \mathcal{U}'$,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

Lemma 9. *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

Figure 9: Illustration of a special case of the proof of Lemma 10 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 $B(D, \delta)$.



form the higher-dimensional polyhedron $\{(x, y, z) : x \in D, y \in S', z = x + y\}$, then project onto the z coordinates.)

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

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

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

Lemma 10. *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 9, 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). □

Lemma 11. *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 10, 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 2. *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 11 gives an ϵ. We take the minimum over these finitely many choices. □

Theorem 6. *For all small enough ϵ, the epsilon-thickened link ψ (Definition 10) 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 2. 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 2, because of our choice of ϵ, the family themselves has nonempty intersection. By Lemma 8, 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 12. *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 13. *If L is a polyhedral loss, then for each p, there exists a constant c such that, for all u,*

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

RF: What I changed: linear ftn → affine; name the cell U_f for f; also name set of functions $\mathcal{F}$, normals V_f, etc

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.

RF: NOTE: we need to define "exposes" or just phrase in terms of normals: v is normal to U at u^*

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 12,

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

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, $U_f \cap U = \cup_i \mathrm{ri}(F_i)$. For each i, let $V_i = \{v \in V : \exists u^* \in \mathrm{ri}(F_i),\ (u^*,v) \in R_f\}$. 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 $A \cap \mathrm{ri}(F_i)$ and thus must contain $A \cap \mathrm{ri}(F_i)$. We conclude that $(u',v) \in R_f$ for all other $u' \in \mathrm{ri}(F_i)$. Thus, we have $\{(u^*,v) \in R_f : u^* \in F_i\} = F_i \times V_i$. For closure, pick any $u^* \in \mathrm{ri}(F_i)$, 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 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. □

RF: The idea here is that if you can start in $\mathrm{ri}(F_i)$ and move in direction v and land immediately in U_f, then you can do that anywhere from $\mathrm{ri}(F_i)$; otherwise U_f intersects only part of $\mathrm{ri}(F_i)$ (at least when restricting to A), a contradiction.

RF: Thus is totally bogus actually, since the limiting point could be u^* itself. Need a different approach I think.

Theorem 7. *For small enough ϵ, the ϵ-thickened link ψ (Definition 10) 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 6. 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 13, 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. □

Proof of Theorem 4. The two claims are Theorems 6 and 7. □

D Top-k surrogate

[JF: Remove 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 14. *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. □