An Embedding Framework for Consistent Polyhedral Surrogates

Anonymous Author(s)
Affiliation
Address
email

Abstract

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

1 Introduction

Convex surrogate losses are a central building block in machine learning for classification and classification-like problems. A growing body of work seeks to design and analyze convex surrogates for given loss functions, and more broadly, understand when such surrogates can and cannot be found. For example, recent work has developed tools to bound the required number of dimensions of the surrogate's hypothesis space [12, 24]. Yet in some cases these bounds are far from tight, such as for *abstain loss* (classification with an abstain option) [4, 24, 25, 33, 34]. Furthermore, the kinds of strategies available for constructing surrogates, and their relative power, are not well-understood.

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

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

Related works. The literature on convex surrogates focuses mainly on smooth surrogate losses [4, 5, 7, 8, 26, 30]. Nevertheless, nonsmooth losses, such as the polyhedral losses we consider, have been proposed and studied for a variety of classification-like problems [18, 31, 32]. A notable

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

addition to this literature is Ramaswamy et al. [25], who argue that nonsmooth losses may enable dimension reduction of the prediction space (range of the surrogate hypothesis) relative to smooth losses, illustrating this conjecture with a surrogate for *abstain loss* needing only $\log n$ dimensions for n labels, whereas the best known smooth loss needs $n-1$. Their surrogate is a natural example of an embedding (cf. Section 5.1), and serves as inspiration for our work.

While property elicitation has by now an extensive literature [9, 11, 14, 16, 17, 21, 28, 29], these works are mostly concerned with point estimation problems. Literature directly connecting property elicitation to consistency is sparse, with the main reference being Agarwal and Agarwal [2]; note however that they consider single-valued properties, whereas properties elicited by general convex losses are necessarily set-valued.

2 Setting

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

2.1 Notation and Losses

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

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

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

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

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

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

2.2 Property Elicitation

To make headway, we will appeal to concepts and results from the property elicitation literature, which elevates the *property*, or map from distributions to optimal reports, as a central object to study in its own right. In our case, this map will often be multivalued, meaning a single distribution could yield multiple optimal reports. (For example, when $p = (1/2, 1/2)$, both $y = 1$ and $y = -1$ optimize 0-1 loss.) To this end, we will use double arrow notation to mean a mapping to all nonempty subsets, so that $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ is shorthand for $\Gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}} \setminus \emptyset$. See the discussion following Definition 3 for conventions regarding $\mathcal{R}$, Γ, γ, L, ℓ, etc.

Definition 2 (Property, level set)**.** *A* property *is a function* $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$*. The* level set *of* Γ *for report* r *is the set* $\Gamma_r := \{p : r \in \Gamma(p)\}$.

Intuitively, $\Gamma(p)$ is the set of reports which should be optimal for a given distribution p, and Γ_r is the set of distributions for which the report r should be optimal. For example, the *mode* is the property $\mathrm{mode}(p) = \arg\max_{y\in\mathcal{Y}} p_y$, 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*

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

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

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

2.3 Links and Embeddings

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

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

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

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

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

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

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

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

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

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

In particular, we see that $-1 \in \mathrm{prop}[\ell_{0\text{-}1}](p) \iff p_1 \in [0, 1/2] \iff -1 \in \mathrm{prop}[L_{\mathrm{hinge}}](p)$, and $1 \in \mathrm{prop}[\ell_{0\text{-}1}](p) \iff p_1 \in [1/2, 1] \iff 1 \in \mathrm{prop}[L_{\mathrm{hinge}}](p)$. 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 $(L_{\mathrm{hinge}}, \psi)$ is calibrated for $\psi(u) = \mathrm{sgn}(u)$; we address in Section 4 the interesting question of whether embeddings lead to calibration in general.

3 Embeddings and Polyhedral Losses

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

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

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

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

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

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

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

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

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

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

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

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

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

We now turn to the reverse direction: which discrete losses are embedded by some polyhedral loss? Perhaps surprisingly, we show that *every* discrete loss is embeddable, using a construction via convex conjugate duality which has appeared several times in the literature (e.g. [1, 8, 10]). Note however that the number of dimensions d required could be as large as $|\mathcal{Y}|$.

Theorem 2. *Every discrete loss ℓ is embedded by a polyhedral loss.*

Proof. Let $n = |\mathcal{Y}|$, and let $C : \mathbb{R}^n \to \mathbb{R}$ be given by $(-\underline{\ell})^*$, the convex conjugate of $-\underline{\ell}$. From standard results in convex analysis, C is polyhedral as $-\underline{\ell}$ is, and C is finite on all of $\mathbb{R}^{\mathcal{Y}}$ as the domain of $-\underline{\ell}$ is bounded [27, Corollary 13.3.1]. Note that $-\underline{\ell}$ is a closed convex function, as the infimum of affine functions, and thus $(-\underline{\ell})^{**} = -\underline{\ell}$. Define $L : \mathbb{R}^n \to \mathbb{R}^{\mathcal{Y}}$ by $L(u) = C(u)\mathbb{1} - u$, where $\mathbb{1} \in \mathbb{R}^{\mathcal{Y}}$ is the all-ones vector. We first show that L embeds ℓ, and then establish that the range of L is in fact $\mathbb{R}^{\mathcal{Y}}_+$, as desired.

We compute Bayes risks and apply Proposition 1 to see that L embeds ℓ. For any $p \in \Delta_{\mathcal{Y}}$, we have

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

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

4 Consistency via Calibrated Links

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

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

Intuitively, to create separation, we just need to "thicken" this construction by mapping all approximately-optimal points u to optimal reports r. Let $\mathcal{U}$ contain all optimal report sets U of the form above. A key step in the following definition will be to narrow down a "link envelope" Ψ where $\Psi(u)$ denotes the legal or valid choices for $\psi(u)$.

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

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

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

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

5 Application to Specific Surrogates

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

5.1 Consistency of abstain surrogate and link construction

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

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

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

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

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

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

Theorem 3 immediately gives calibration of $(L_{1/2}, \psi_1)$ with respect to $\ell_{1/2}$. Aside from its simplicity, one possible advantage of ψ_1 is that it appears to yield the same constant in generalization bounds as ψ, yet assigns $\bot$ to much less of the surrogate space $\mathbb{R}^d$. It would be interesting to compare the two links in practice.

5.2 Inconsistency of top-k losses

In certain classification problems, for example in information retrieval, it is common to predict a set of possible labels. As one instance, for $k < n$ the top-k classification problem has reports $\mathcal{R} := \{r \subseteq [n] : |r| = k\}$, with label $y \in [n]$. The natural discrete loss $\ell_{\text{top-}k} : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ is given by

$$\ell_{\text{top-}k}(r, y) = \mathbb{1}\{y \notin r\} \ , \tag{7}$$

which simply gives a penalty if the label was not in the reported set.

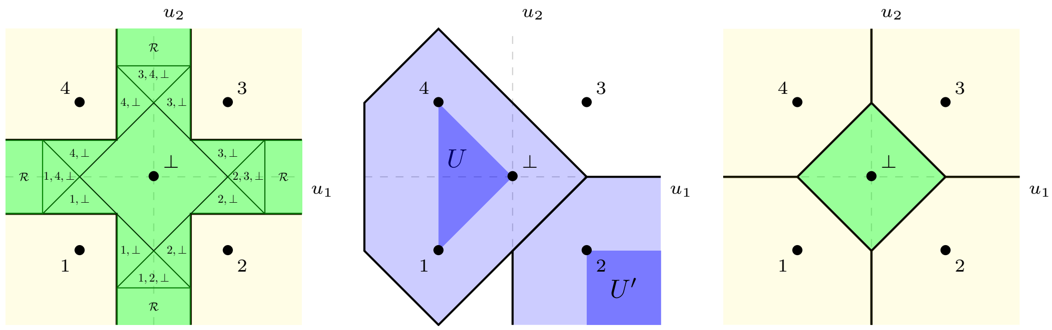

Figure 1: Constructing links for the abstain surrogate $L_{1/2}$ with $d = 2$. The embedding is shown in bold labeled by the corresponding reports. (L) The link envelope Ψ resulting from Theorem 3 using $\|\cdot\|_\infty$ and $\epsilon = 1/2$, and a possible link ψ which matches eq. (5) from [25]. (M) An illustration of the thickened sets from Definition 6 for two sets $U \in \mathcal{U}$, using $\|\cdot\|_1$ and $\epsilon = 1$. (R) The Ψ and ψ from Theorem 3 using $\|\cdot\|_1$ and $\epsilon = 1$.

Surrogates for this problem commonly take reports $u \in \mathbb{R}^n$, with the link $\psi(u) = \{u_{[1]}, \ldots, u_{[k]}\}$, where $x_{[u]}$ is the i^{th} largest component of u, with ties broken arbitrarily. Lapin et al. [18, 19, 20] provide the following convex surrogate loss for this problem, which Yang and Koyejo [31] show to be inconsistent:[1]

$$L'(u)_y := \left(\tfrac{1}{k} \textstyle\sum_{i=1}^{k} (u + \mathbb{1} - e_y)_{[i]} - u_y \right)_+ , \tag{8}$$

where e_y is 1 in component y and 0 elsewhere.

With our framework, we can say more. Specifically, while (L', ψ) is not consistent for $\ell_{\text{top-}k}$, since L' is polyhedral (Lemma 11), we know from Theorem 1 that it embeds *some* discrete loss ℓ', and from Theorem 3 there is a link ψ' such that (L', ψ') is calibrated (and consistent) for ℓ'. We therefore turn to deriving this discrete loss ℓ'.

In Lemma 12, we show that the set $\mathcal{U} = \{u \in \{0,1\}^n : \|u\|_0 \leq k\}$ is always represented among the optimizers of L', meaning for all $p \in \Delta_\mathcal{Y}$ we have $\text{prop}[L'](p) \cap \mathcal{U} \neq \emptyset$. From Lemma 3, then, L' must embed $L'|_\mathcal{U}$, which gives us $\ell' : \mathcal{R}' \to \mathbb{R}^\mathcal{Y}_+$ via the natural embedding from sets $\mathcal{R}' = \{r \subseteq [n] : |r| \leq k\}$ to their indicator vectors:

$$\ell'(r)_y = \mathbb{1}\{y \notin r\} + \tfrac{1}{k}|r \setminus \{y\}| = (1 + \tfrac{1}{k})\mathbb{1}\{y \notin r\} + \tfrac{1}{k}(|r| - 1) . \tag{9}$$

We now see that ℓ' is essentially $\ell_{\text{top-}k}$ (extended to sets smaller than k) plus an additional cardinality term which rewards smaller sets.

Knowledge of what loss L' actually embeds greatly simplifies the task of proving inconsistency with $\ell_{\text{top-}k}$. Specifically, we see that ℓ' allows sets of cardinality strictly less than k, so we can look for a distribution making one of these smaller sets optimal. Writing the expected loss, we have $\langle p, \ell'(r) \rangle = (1 + \frac{1}{k})(1 - p(r)) + \frac{1}{k}(|r| - 1) = 1 + \frac{1}{k}|r| - (1 + \frac{1}{k})p(r)$, where $p(r) = \sum_{i \in r} p_i$. So let us ask when it is better to drop an element i from r: we have $\langle p, \ell'(r) - \ell'(r \setminus \{i\}) \rangle = \frac{1}{k} - (1 + \frac{1}{k})p_i$, meaning we will drop elements from r until they all have weight at least $\frac{1}{k+1}$. In particular, for distributions close to uniform, $r = \emptyset$ is optimal, already giving us inconsistency. More generally, as $k < n$, the set $P = \{p \in \Delta_\mathcal{Y} : \max_{r \in \mathcal{R}'} p(r) < k/(k+1)\}$ is full-dimensional and guarantees that at least one of the top k labels has probability strictly less than $\frac{1}{k+1}$.

5.3 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 [13, 15, 22]. In the case of k binary predictions, these settings are typically formalized by taking the predictions and outcomes to be ± 1 vectors, so $\mathcal{R} = \mathcal{Y} = \{-1, 1\}^k$. One then defines a joint loss function, which is often merely a function of the set of mispredictions, meaning $\ell^g(r)_y =$

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

$g(\{i \in [k] : r_i \neq y_i\})$ for some function $g : 2^{[k]} \to \mathbb{R}$. For example, Hamming loss is simply given by $g(S) = |S|$. In an effort to provide a general convex surrogate for these settings when g is a submodular function, Yu and Blaschko [32] introduce the *Lovász hinge*, which leverages the well-known convex Lovász extension of submodular functions. While the authors provide theoretical justification and experiments, consistency of the Lovász hinge is left open, which we resolve.

Rather than formally define the Lovász hinge, we defer the complete analysis to the full version of the paper,[2] 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

$$L^g(u)_y = \max\Big\{(1 - u_1y_1)_+ g_1 + (1 - u_2y_2)_+(g_{1,2} - g_1), \\ (1 - u_2y_2)_+ g_2 + (1 - u_1y_1)_+(g_{1,2} - g_2)\Big\} \,, \quad (10)$$

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 $g_\emptyset = 0$, $g_1 = g_2 = g_{1,2} = 1$, for which ℓ^g is merely 0-1 loss on $\mathcal{Y}$. For consistency, then, for any distribution $p \in \Delta_{\mathcal{Y}}$, we must have that whenever $u \in \arg\min_{u' \in \mathbb{R}^2} p \cdot L^g(u')$, then $\psi(u)$ must be the most likely outcome, in $\arg\max_{y \in \mathcal{Y}} p(y)$. Simplifying eq. (10), however, we have

$$L^g(u)_y = \max\big\{(1 - u_1y_1)_+, (1 - u_2y_2)_+\big\} = \max\big\{1 - u_1y_1, 1 - u_2y_2, 0\big\} \,, \quad (11)$$

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 $\{0\} = \arg\min_{u \in \mathbb{R}^2} p \cdot L^g(u)$, yet $\psi(0) = (1,1) \notin \{(-1,-1)\} = \arg\min_{r \in \mathcal{R}} p \cdot \ell^g(r)$.

In fact, this example is typical: using our embedding framework, and characterizing when $0 \in \mathbb{R}^2$ is an embedded point, one can show that L^g is consistent if and only if $g_{1,2} = g_1 + g_2$. Moreover, in the linear case, which corresponds to g being *modular*, the Lovász hinge reduces to weighted Hamming loss, which is trivially consistent from the consistency of hinge loss for 0-1 loss. In the full version of the paper, 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 Conclusion and Future Directions

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

One open question of particular interest involves the dimension of the input to a surrogate; given a discrete loss, can we construct the surrogate that embeds it *of minimal dimension*? If we naïvely embed the reports into an n-dimensional space, the dimensionality of the problem scales with the number of possible labels n. As the dimension of the optimization problem is a function of this *embedding dimension* d, a promising direction is to leverage tools from elicitation complexity [12, 17] and convex calibration dimension [24] to understand when we can take $d \ll n$.

[2] Citation withheld for anonymity.

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] 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.

[5] 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`.

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

[7] 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.

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

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

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

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

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

[13] 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.

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

[15] 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.

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

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Bernardo Avila Pires, Csaba Szepesvari, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In *International Conference on Machine Learning*, pages 1391–1399, 2013.

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

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

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

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

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

[29] Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and Identification of Properties. In *Proceedings of The 27th Conference on Learning Theory*, pages 482–526, 2014.

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

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

[32] Jiaqian Yu and Matthew B Blaschko. The lovász hinge: A novel convex surrogate for submodular losses. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

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

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