Proceedings of Machine Learning Research vol 178:1–23, 2022 35th Annual Conference on Learning Theory

The Structured Abstain Problem and the Lovász Hinge

Jessie Finocchiaro JEFI8453@COLORADO.EDU
Rafael Frongillo RAF@COLORADO.EDU
Enrique Nueve ENNU6440@COLORADO.EDU
University of Colorado Boulder

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

The Lovász hinge is a convex surrogate recently proposed for structured binary classification, in which k binary predictions are made simultaneously and the error is judged by a submodular set function. Despite its wide usage in image segmentation and related problems, its consistency has remained open. We resolve this open question, showing that the Lovász hinge is inconsistent for its desired target unless the set function is modular. Leveraging a recent embedding framework, we instead derive the target loss for which the Lovász hinge is consistent. This target, which we call the structured abstain problem, allows one to abstain on any subset of the k predictions. We derive two link functions, each of which are consistent for all submodular set functions simultaneously.

1. Introduction

Structured prediction addresses a wide variety of machine learning tasks in which the error of several related predictions is best measured jointly, according to some underlying structure of the problem, rather than independently (Gao and Zhou, 2011; Hazan et al., 2010; Osokin et al., 2017; Tsochantaridis et al., 2005). This structure could be spatial (e.g., images and video), sequential (e.g., text), combinatorial (e.g., subgraphs), or a combination of the above. As traditional target losses such as 0-1 loss measure error independently, more complex target losses are often introduced to capture the joint structure of these problems.

As with most classification-like settings, optimizing a given discrete target loss is typically intractable. We therefore seek surrogate losses which are both convex, and thus efficient to optimize, and statistically consistent, meaning they actually solve the desired problem. Another important factor in structured prediction is that the number of possible labels and/or target predictions is often exponentially large. For example, in the structured binary classification problem, one makes k simultaneous binary predictions, yielding 2^k possible labels. In these settings, it is crucial to find a surrogate whose prediction space is low-dimensional relative to the relevant parameters.

In general, however, we lack surrogates satisfying all three desiderata: convex, consistent, and low-dimensional (McAllester, 2007; Nowozin, 2014). One promising low-dimensional surrogate for structured binary classification, the Lovász hinge, achieves convexity via the well-known Lovász extension for submodular set functions (Yu and Blaschko, 2018). Despite the fact that this surrogate and its generalizations (Berman et al., 2018) have been widely used, e.g. in image segmentation and processing (Athar et al., 2020; Chen et al., 2020; Neven et al., 2019), its consistency has thus far not been established.

Using the embedding framework of Finocchiaro et al. (2022), we show the inconsistency of Lovász hinge for structured binary classification (§ 4). Our proof relies on first determining what the Lovász hinge is actually consistent for: the *structured abstain problem*, a variation of structured binary prediction in which one may abstain on a subset of the predictions (§ 3). For reasons similar to classification with an abstain option (Bartlett and Wegkamp, 2008; Ramaswamy et al., 2018), this problem may be of interest to the structured prediction community. Finally, while the embedding framework shows that a calibrated link must exist, in our case actually deriving such a link is nontrivial. In § 5 we derive two complementary link functions, both of which are calibrated simultaneously for all submodular set functions parameterizing the problem.

2. Background

2.1. Notation

See Tables 1 and 2 in § A for full tables of notation. Throughout, we consider predictions over k binary events, yielding $n = 2^k$ total outcomes, with each label $y \in \mathcal{Y} = \{-1, 1\}^k$. Predictions are generically denoted $r \in \mathcal{R}$; we often take $\mathcal{R} = \mathcal{Y}$, or consider predictions $v \in \mathcal{V} := \{-1, 0, 1\}^k$ or $u \in \mathbb{R}^k$. Loss functions measure these predictions against the observed label $y \in \mathcal{Y}$. In general, we denote a discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ and surrogate $L : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}_+$. We also occasionally restrict a loss L to a domain $\mathcal{S} \subseteq \mathcal{R}$ and define $L|_{\mathcal{S}} : (u, y) \mapsto L(u, y)$ for all $u \in \mathcal{S}$.

Let $[k] := \{1, \dots, k\}$. When translating from vector functions to set functions, it is often useful to use the shorthand $\{u \le c\} := \{i \in [k] \mid u_i \le c\}$ for $u \in \mathbb{R}^k$, $c \in \mathbb{R}$, and similarly for other set comprehensions. Additionally, for any $S \subseteq [k]$, we let $\mathbb{1}_S \in \{0, 1\}^k$ with $(\mathbb{1}_S)_i = 1 \iff i \in S$ be the 0-1 indicator for S. Let $\mathcal{S}_k$ denote the set of permutations of $[k]$. For any permutation $\pi \in \mathcal{S}_k$, and any $i \in \{0, 1, \dots, k\}$, define $\mathbb{1}_{\pi,i} = \mathbb{1}_{\{\pi_1, \dots, \pi_i\}}$, where $\mathbb{1}_{\pi,0} = 0 \in \mathbb{R}^k$.

For $u, u' \in \mathbb{R}^k$, the Hadamard (element-wise) product $u \odot u' \in \mathbb{R}^k$ given by $(u \odot u')_i = u_i u'_i$ plays a prominent role. We extend $\odot$ to sets in the natural way; e.g., for $U \subseteq \mathbb{R}^k$ and $u' \in \mathbb{R}^k$, we define $U \odot u' = \{u \odot u' \mid u \in U\}$.

We often decompose elements of $u \in \mathbb{R}^k$ by their sign and absolute value. To this end, we define $\operatorname{sign} : \mathbb{R}^k \to \mathcal{V}$ to be the (element-wise) sign of u, and use the function $\operatorname{sign}^* : \mathbb{R}^k \to \mathcal{Y}$ to denote an arbitrary function that agrees with sign when $|u_i| \neq 0$ and break ties arbitrarily at 0. We let $|u| \in \mathbb{R}^k_+$ be the element-wise absolute value $|u|_i = |u_i|$, and frequently use the fact that $|u| = u \odot \operatorname{sign}^*(u) = u \odot \operatorname{sign}(u)$. We define $\overline{u} = \operatorname{sign}(u) \odot \min(|u|, \mathbb{1})$ to "clip" u to $[-1, 1]^k$. Finally, we denote $((u)_+)_i = \max(u_i, 0)$.

2.2. Submodular functions and the Lovász extension

A set function $f : 2^{[k]} \to \mathbb{R}$ is *submodular* if for all $S, T \subseteq [k]$ we have $f(S) + f(T) \ge f(S \cup T) + f(S \cap T)$. If this inequality is strict whenever S and T are incomparable, meaning $S \not\subseteq T$ and $T \not\subseteq S$, then we say f is *strictly submodular*. A function is *modular* if the submodular inequality holds with equality for all $S, T \subseteq [k]$. The function f is *increasing* if we have $f(S \cup T) \ge f(S)$ for all disjoint $S, T \subseteq [k]$, and *strictly increasing* if the inequality

is strict whenever $T \neq \varnothing$. Finally, we say f is *normalized* if $f(\varnothing) = 0$. Let $\mathcal{F}_k$ be the class of set functions $f : 2^{[k]} \to \mathbb{R}$ which are submodular, increasing, and normalized.

The structured binary classification problem is given by the following discrete loss $\ell^f : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$, with $\mathcal{R} = \mathcal{Y}$,

$$\ell^f(r, y) = f(\{r \odot y < 0\}) = f(\{i \in [k] \mid r_i \neq y_i\}) \ . \tag{1}$$

In words, ℓ^f measures the joint error of the k predictions by applying f to the set of mispredictions, i.e., indices corresponding to incorrect predictions. For the majority of the paper, we will consider $f \in \mathcal{F}_k$. In particular, we will make the natural assumption that f is increasing: making an additional error cannot decrease error. The assumption that f be normalized is without loss of generality.

A classic object related to submodular functions is the *Lovász extension* to $\mathbb{R}^k$ (Lovász, 1983), which is known to be convex when (and only when) f is submodular (Bach, 2013, Proposition 3.6). For any permutation $\pi \in \mathcal{S}_k$, define $P_\pi = \{x \in \mathbb{R}^k_+ \mid x_{\pi_1} \geq \cdots \geq x_{\pi_k}\}$, the set of nonnegative vectors ordered by π. The Lovász extension of a normalized set function $f : 2^{[k]} \to \mathbb{R}$ can be formulated in several equivalent ways (Bach, 2013, Definition 3.1).

$$F(x) = \max_{\pi \in \mathcal{S}_k} \sum_{i=1}^{k} x_{\pi_i} (f(\{\pi_1, \ldots, \pi_i\}) - f(\{\pi_1, \ldots, \pi_{i-1}\})) \ . \tag{2}$$

Given any $x \in \mathbb{R}^k_+$, the argmax in eq. (2) is the set $\{\pi \in \mathcal{S}_k \mid x \in P_\pi\}$, i.e., the set of all permutations that order the elements of x. For any $\pi \in \mathcal{S}_k$ such that $x \in P_\pi$, we may therefore write

$$F(x) = \sum_{i=1}^{k} x_{\pi_i} (f(\{\pi_1, \ldots, \pi_i\}) - f(\{\pi_1, \ldots, \pi_{i-1}\})) \ . \tag{3}$$

For any $f \in \mathcal{F}_k$, let F be the Lovász extension of f. Yu and Blaschko (2018) define the *Lovász hinge* as the loss $L^f : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}_+$ given as follows.

$$L^f(u, y) = F\big((\mathbb{1} - u \odot y)_+\big) \ . \tag{4}$$

The Lovász hinge is proposed as a surrogate for the structured binary classification problem in eq. (1), using the link sign^* to map surrogate predictions $u \in \mathbb{R}^k$ back to the discrete report space $\mathcal{R} = \mathcal{Y}$. From eq. (2), the Lovász extension is polyhedral (piecewise-linear and convex) as a maximum of a finite number of affine functions. Hence L^f is a polyhedral loss function.

Immediately from the definition, the fact that $\odot$ is symmetric, and $x \mapsto x \odot y$ is an involution for any $y \in \mathcal{Y}$, we have the following.

Lemma 1 *For all $u \in \mathbb{R}^k$ and $y, y' \in \mathcal{Y}$, $L^f(u, y) = L^f(u \odot y', y \odot y')$.*

2.3. Specific submodular functions

To illustrate the above definitions, we provide several examples. For the first, consider the case where f is modular. Modular set functions can be parameterized by any $w \in \mathbb{R}^k_+$,

so that $f_w(S) = \sum_{i \in S} w_i$. In this case ℓ^f reduces to *weighted Hamming loss*, and L^f to weighted hinge, the consistency of which is known (Gao and Zhou, 2011, Theorem 15).

$$\begin{aligned} L^{f_w}(u, y) &= \max_{\pi \in \mathcal{S}_k} \sum_{i=1}^{k} ((1 - u \odot y)_+)_{\pi_i} (f(\{\pi_1, \ldots, \pi_i\}) - f(\{\pi_1, \ldots, \pi_{i-1}\})) \\ &= \sum_{i=1}^{k} (1 - u_i y_i)_+ (w_i) \ . \end{aligned} \tag{5}$$

For another example, consider $f_{0\text{-}1}$ given by $f_{0\text{-}1}(\varnothing) = 0$ and $f_{0\text{-}1}(S) = 1$ for $S \neq \varnothing$. Here the Lovász hinge reduces to

$$\begin{aligned} L^{f_{0\text{-}1}}(u, y) &= \max_{\pi \in \mathcal{S}_k} \sum_{i=1}^{k} ((1 - u \odot y)_+)_{\pi_i} (f(\{\pi_1, \ldots, \pi_i\}) - f(\{\pi_1, \ldots, \pi_{i-1}\})) \\ &= \max_{i \in [k]} (1 - u_i y_i)_+ \ . \end{aligned} \tag{6}$$

In fact, $L^{f_{0\text{-}1}}$ is equivalent to the BEP surrogate by Ramaswamy et al. (2018) for the problem of multiclass classification with an abstain option. The target loss for this problem is $\ell_{1/2} : [n] \cup \{\perp\} \times [n] \to \mathbb{R}_+$ defined by $\ell_{1/2}(r, y) = 0$ if $r = y$, $1/2$ 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 $p_y \geq 1/2$. The BEP surrogate is given by

$$L_{\frac{1}{2}}(u, \hat{y}) = \left(\max_{j \in [k]} \ B(\hat{y})_j u_j + 1 \right)_+ \tag{7}$$

where $B : [n] \to \{-1, 1\}^k$ is an arbitrary injection. Substituting $y = -B(\hat{y})$ in eq. (7), and moving the $(\cdot)_+$ inside, we recover eq. (6).

Lastly, consider the function $f_\beta(S) = 1 - \beta^{|S|}$ where $\beta \in (0, 1)$ is a discount factor, as proposed by Yu and Blaschko (2018) with the parameter $-\log \beta$. The Lovász hinge for f_β has the following form,

$$\begin{aligned} L^{f_\beta}(u, y) &= \max_{\pi \in \mathcal{S}_k} \sum_{i=1}^{k} ((1 - u \odot y)_+)_{\pi_i} (f_\beta(\{\pi_1, \ldots, \pi_i\}) - f_\beta(\{\pi_1, \ldots, \pi_{i-1}\})) \\ &= (\beta^{-1} - 1) \max_{\pi \in \mathcal{S}_k} \sum_{i=1}^{k} ((1 - u \odot y)_+)_{\pi_i} \ \beta^i \ . \end{aligned} \tag{8}$$

As motivation for f_β, consider structured problems such as part-of-speech tagging and image segmentation, where additional errors on a single instance (sentence or image) may not be as dire as additional instances with errors. The "diminishing marginal return" behavior of f_β will therefore guide an algorithm to improve predictions on instances for which it is slightly wrong, and to de-prioritize intances for which it is extremely wrong; in other words, it encourages the model to cut its losses.

2.4. Property elicitation and calibration

When considering polyhedral (piecewise-linear and convex) losses, like the Lovász hinge (4), Finocchiaro et al. (2022) show that indirect property elicitation is equivalent to statistical consistency. Property elicitation is therefore an important tool to study consistent polyhedral surrogates for a given discrete loss.

Definition 2 *A property* $\Gamma : \Delta_{\mathcal{Y}} \to 2^{\mathcal{R}} \setminus \{\varnothing\}$ *is a function mapping distributions over labels to reports. A loss* $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ elicits *a property* Γ *if, for all* $p \in \Delta_{\mathcal{Y}}$,

$$\Gamma(p) = \arg\min_{r \in \mathcal{R}} \mathbb{E}_{Y \sim p} L(r, Y) \ .$$

Moreover, if $\mathbb{E}_{Y \sim p} L(\cdot, Y)$ *attains its infimum for all* $p \in \Delta_{\mathcal{Y}}$*, we say* L *is* minimizable*, and elicits some unique property, denoted* $\text{prop}[L]$.

Statistical consistency is a prerequisite for deriving excess risk bounds in empirical risk minimization problems. Roughly, we say a surrogate L and link (mapping surrogate reports $u \in \mathbb{R}^d$ to target reports) pair are consistent with respect to a target loss ℓ, if all possible data distributions, any sequence of hypotheses approaching the L-optimal expected loss will also approach the ℓ-optimal expected loss when the link is applied to each element of the sequence. See (Finocchiaro et al., 2021) for a more thorough treatment.

In order to connect property elicitation to statistical consistency, we work through the notion of calibration, which is equivalent to consistency in our setting (Bartlett et al., 2006; Ramaswamy and Agarwal, 2016; Zhang, 2004). One desirable characteristic of calibration over consistency is the ability to abstract features $x \in \mathcal{X}$ so that we can simply study the expected loss over labels through the distribution $p \in \Delta_{\mathcal{Y}}$. We often denote $L(u; p) := \mathbb{E}_{Y \sim p} L(u, Y)$, and $\ell(r; p) := \mathbb{E}_{Y \sim p} \ell(r, Y)$.

The definitions of consistency and calibration rely crucially on the existence of a link function $\psi : \mathbb{R}^d \to \mathcal{R}$ mapping surrogate reports to the target prediction space. For example, the sign link is a prominent link function for standard classification problems. Importantly, calibration and consistency are defined by a surrogate and link pair. Even if a seemingly natural link function is not calibrated for a target task alongside the surrogate, there may be another link that is calibrated for the task.

Definition 3 *Let* $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ *with* $|\mathcal{R}| < \infty$*. A surrogate* $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ *and link* $\psi : \mathbb{R}^d \to \mathcal{R}$ *pair* (L, ψ) *is* calibrated *with respect to* ℓ *if for all* $p \in \Delta_{\mathcal{Y}}$,

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

2.5. The embedding framework

We will lean heavily on the embedding framework of Finocchiaro et al. (2019, 2022). Given a discrete target loss, and a surrogate loss over $\mathbb{R}^k$, an embedding maps target reports into $\mathbb{R}^k$ so that the surrogate behaves the same as the target on the embedded points. The authors show that every polyhedral surrogate embeds some discrete loss, and show that an embedding implies consistency. To define embeddings, we first need a notion of representative sets, which allows one to ignore some target reports that are in some sense redundant.

Definition 4 *We say $\mathcal{S} \subseteq \mathcal{R}$ is* representative *with respect to the loss L if we have $\arg\min_u L(u;p) \cap \mathcal{S} \neq \varnothing$ for all $p \in \Delta_\mathcal{Y}$.*

Definition 5 (Embedding) *The loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$* embeds *a loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ if there exists a representative set $\mathcal{S}$ for ℓ and an injective embedding $\varphi : \mathcal{S} \to \mathbb{R}^d$ such that (i) for all $r \in \mathcal{S}$ and $y \in \mathcal{Y}$ we have $L(\varphi(r), y) = \ell(r, y)$, and (ii) for all $p \in \Delta_\mathcal{Y}, r \in \mathcal{S}$ we have*

$$r \in \text{prop}[\ell](p) \iff \varphi(r) \in \text{prop}[L](p) \ . \tag{9}$$

Embeddings are intimately tied to polyhedral losses as they have finite representative sets. Every discrete loss is embedded by some polyhedral surrogate (Finocchiaro et al., 2022, Thm. 4). A central tool for the present work, however, is the converse: every polyhedral loss embeds some discrete target loss, namely, its restriction to a finite representative set.

Theorem 6 ((Finocchiaro et al., 2022, Thm. 3, Prop. 1)) *A loss L with a finite representative set $\mathcal{S}$ embeds $L|_\mathcal{S}$. Moreover, every polyhedral L has a finite representative set.*

A central contribution of the embedding framework is to simplify proofs of consistency. In particular, if a surrogate $L : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}_+$ embeds a discrete target $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$, then there exists a calibrated link function $\psi : \mathbb{R}^k \to \mathcal{R}$ such that (L, ψ) is consistent with respect to ℓ. The proof is constructive, via the notion of *separated* link functions, a fact we will make use of in § 5; specifically, see Theorem 17.

3. Lovász hinge embeds the structured abstain problem

As the Lovász hinge is a polyhedral surrogate, Theorem 6 states that it embeds some discrete loss, which may or may not be the same as the intended target ℓ^f. As we saw in § 2.3, one special case, $L^{f_{0\text{-}1}}$, reduces to the BEP surrogate for multiclass classification with an abstain option, which implies that L^f cannot embed ℓ^f in general. In particular, whatever L^f embeds, it must allow the algorithm to abstain in some sense. We formalize this intuition by showing L^f embeds the discrete loss ℓ^f_{abs}, a variant of structured binary classification which allows abstention on any subset of the k labels. See § B for all omitted proofs.

3.1. The filled hypercube is representative

As a first step, we show that the filled hypercube $R := [-1, 1]^k$ is representative for L^f, and use this fact to later find a *finite* representative set for L^f and apply Theorem 6. In fact, we show the following stronger statement: surrogate reports outside the filled hypercube $[-1, 1]^k$ are dominated on each outcome.

Lemma 7 *For any $u \in \mathbb{R}^k$, we have $L^f(\overline{u}, y) \leq L^f(u, y)$ for all $y \in \mathcal{Y}$.*

Using this result, we may now simplify the Lovász hinge. When $u \in [-1, 1]^k$, we simply have

$$L^f|_R(u, y) = F(\mathbb{1} - u \odot y) \ , \tag{10}$$

as $\mathbb{1} - u \odot y$ is nonnegative.

3.2. Affine decomposition of L^f

We now give an affine decomposition of L^f on $[-1,1]^k$, which we use throughout. Recall that for any $\pi \in \mathcal{S}_k$ we define $P_\pi = \{x \in \mathbb{R}^k_+ \mid x_{\pi_1} \geq \cdots \geq x_{\pi_k}\}$. Letting $V_\pi = \{\mathbb{1}_{\pi,i} \mid i \in \{0,\ldots,k\}\} \subset \mathcal{V}$, we have $P_\pi = \operatorname{cone} V_\pi$, the conic hull of V_π, meaning every $x \in P_\pi$ can be written as a conic combination of elements of V_π. For all $i \in \{0,\ldots,k\}$, define the coefficients $\alpha_i : \mathbb{R}^k_+ \to \mathbb{R}$ as follows. For any $x \in \mathbb{R}^k_+$, define $\alpha_0(x) = 1 - x_{[1]} \in \mathbb{R}$, $\alpha_k(x) = x_{[k]} \geq 0$, and $\alpha_i(x) = x_{[i]} - x_{[i+1]} \geq 0$ for $i \in \{1,\ldots,k-1\}$. Then

$$x = \sum_{i=1}^{k} \alpha_i(x)\, \mathbb{1}_{\pi,i} = \sum_{i=0}^{k} \alpha_i(x)\, \mathbb{1}_{\pi,i} \ , \tag{11}$$

where we recall that $\mathbb{1}_{\pi,0} = \mathbf{0} \in \mathbb{R}^k$. We have $\alpha_i(x) \geq 0$ for all $i \in \{1,\ldots,k\}$, so the first equality gives the conic combination. In the case $x_{[1]} \leq 1$, we have $\alpha_i(x) \geq 0$ for all $i \in \{0,\ldots,k\}$. Since $\sum_{i=0}^k \alpha_i(x) = 1$, in that case the latter equality in eq. (11) is a convex combination. This yields $P_\pi \cap [0,1]^k = \operatorname{conv} V_\pi$.

It is clear from eq. (3) that F is affine on P_π for each $\pi \in \mathcal{S}_k$. We now identify the regions within $[-1,1]^k$ where $L^f(\cdot, y)$ is affine simultaneously for all outcomes $y \in \mathcal{Y}$, using these polyhedra and symmetry in y.

Motivated by the above, for any $y \in \mathcal{Y}$ and $\pi \in \mathcal{S}_k$, define

$$V_{\pi,y} = V_\pi \odot y = \{\mathbb{1}_{\pi,i} \odot y \mid i \in \{0,\ldots,k\}\} \subset \mathcal{V} \ , \tag{12}$$

$$P_{\pi,y} = \operatorname{conv}(V_{\pi,y}) = \operatorname{conv}(V_\pi) \odot y \subset [-1,1]^k \ . \tag{13}$$

Since $V_{\pi,y}$ is a set of affinely independent vectors, each $P_{\pi,y}$ is a simplex. Observe that for the case $y = \mathbb{1}$, we have $P_{\pi,\mathbb{1}} = P_\pi \cap [0,1]^k$. Indeed, the other $P_{\pi,y}$ sets are simply reflections of $P_{\pi,\mathbb{1}}$, as we may write $P_{\pi,y} = P_{\pi,\mathbb{1}} \odot y$. We now show that these regions union to the filled hypercube $[-1,1]^k$, and $L^f(\cdot, y)$ is affine on $P_{\pi,y}$ for each $y \in \mathcal{Y}$.

Lemma 8 *The sets $P_{\pi,y}$ satisfy the following.*

(i) $\cup_{y\in\mathcal{Y},\pi\in\mathcal{S}_k} P_{\pi,y} = [-1,1]^k$.

(ii) For all $f \in \mathcal{F}_k$, $y, y' \in \mathcal{Y}$, and $\pi \in \mathcal{S}_k$, the function $L^f(\cdot, y')$ is affine on $P_{\pi,y}$.

3.3. Embedding the structured abstain problem

Leveraging the affine decomposition given above, we will now show that the finite set $\mathcal{V} = \{-1,0,1\}^k$ must be representative for L^f. By Theorem 6, it will then follow that L^f embeds $\ell^f_{\text{abs}} := L^f|_{\mathcal{V}}$. As we describe below, we call ℓ^f_{abs} the *structured abstain problem* because the predictions $v \in \mathcal{V}$ allow one to "abstain" on an index i by setting $v_i = 0$.

Lemma 9 *Given a polyhedral loss function $L : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}_+$, let $\mathcal{C}$ be a collection of polyhedral subsets of $\mathbb{R}^k$ such that for all $y \in \mathcal{Y}$, $L(\cdot, y)$ is affine on each $C_i \in \mathcal{C}$, and denote* $\text{faces}(C_i)$ *as the set of faces of C_i. Let $R = \cup\mathcal{C}$ be the union of these polyhedral subsets. Then for all $p \in \Delta_\mathcal{Y}$,* $\text{prop}[L](p) \cap R = \cup\mathcal{F}$ *for some $\mathcal{F} \subseteq \cup_i \text{faces}(C_i)$.*

Proposition 10 *The set $\mathcal{V} = \{-1,0,1\}^k$ is representative for L^f.*

Proof Let $\mathcal{C} = \{P_{\pi,y} | \forall\ \pi \in \mathcal{S}_k, y \in \mathcal{Y}\}$ and $R = \cup\mathcal{C} = \cup_{\pi\in\mathcal{S}_k, y\in\mathcal{Y}} P_{\pi,y} = [-1,1]^k$ by Lemma 8(i). Since every $P_{\pi,y}$ is affine w.r.t L^f according to Lemma 8(ii), we have by Lemma 9 $\forall p \in \Delta_\mathcal{Y}$, $\text{prop}[L](p) \cap R = \cup\mathcal{F}$ where $\mathcal{F} \subseteq \cup_{\pi,y}\text{faces}(P_{\pi,y})$. Yet, by the construction of $P_{\pi,y}$, every face contains some number of vertices from $\mathcal{V}$. Therefore, $\forall\ p \in \Delta_\mathcal{Y}$, $\text{prop}[L](p) \cap \mathcal{V} \neq \varnothing$ which by definition means that $\mathcal{V}$ is representative for L^f. ■

Theorem 11 *The Lovász hinge L^f embeds $\ell^f_{abs} : \mathcal{V} \times \mathcal{Y} \to \mathbb{R}_+$ given by*

$$\ell^f_{abs}(v, y) = f(\{v \odot y < 0\}) + f(\{v \odot y \leq 0\})\ . \tag{14}$$

Proof From Proposition 10 and Theorem 6, L^f embeds $L^f|_\mathcal{V}$. It therefore remains only to establish the set-theoretic form of $L^f|_\mathcal{V}$ as the loss ℓ^f_{abs} in eq. (14).

Let $v \in \mathcal{V}, y \in \mathcal{Y}$ be given. We may write

$$\mathbb{1} - v \odot y = 0 \cdot \mathbb{1}_{\{v\odot y>0\}} + 1 \cdot \mathbb{1}_{\{v\odot y=0\}} + 2 \cdot \mathbb{1}_{\{v\odot y<0\}}\ .$$

Now combining eq. (10) and Bach (2013, Prop 3.1(h)), we may therefore write

$$\begin{aligned} L^f(v, y) &= F(\mathbb{1} - v \odot y) \\ &= (2-1)f(\{v \odot y < 0\}) + (1-0)f(\{v \odot y < 0\} \cup \{v \odot y = 0\}) + 0f([k]) \\ &= f(\{v \odot y < 0\}) + f(\{v \odot y \leq 0\})\ , \end{aligned}$$

as was to be shown. ■

We can interpret ℓ^f_{abs} as a structured abstain problem, where the algorithm is allowed to abstain on a given prediction by giving a zero instead of ± 1. Specifically, we can say the algorithm abstains on the set of indices $A_v = \{v = 0\}$.

To make this interpretation more clear, let $r = \text{sign}^*(v)$, which is forced to choose a label ± 1 for each zero prediction. The corresponding set of mispredictions for fixed $y \in \mathcal{Y}$ would be $M^y = \{r \odot y < 0\}$. We can rewrite eq. (14) in terms of these sets as $\ell^f_{\text{abs}}(v, y) = f(M^y \setminus A_v) + f(M^y \cup A_v)$. Contrasting with $\ell^f_{\text{abs}}(r, y) = 2f(\{r \odot y < 0\}) = f(M^y) + f(M^y)$, the abstain option allows one to reduce loss in the first term at the expense of a sure loss in the second term. Intuitively, when there is large uncertainty about the labels of a set of indices $A \subseteq [k]$, by submodularity the algorithm would prefer to abstain on A than take a chance on predicting.

When relating to submodularity, we will often find it useful to rewrite the misprediction set M^y above in terms of two sets of labels: $S_v = \{\text{sign}^*(v) > 0\}$ and $S_y = \{y > 0\}$. Then $M^y = S_v \triangle S_y$, and thus

$$\ell^f_{\text{abs}}(v, y) = f(S_v \triangle S_y \setminus A_v) + f(S_v \triangle S_y \cup A_v)\ , \tag{15}$$

where $\triangle$ is the symmetric difference operator $S \triangle T := (S \setminus T) \cup (T \setminus S)$. To avoid additional parentheses, throughout we assume $\triangle$ has operator precedence over $\setminus$, $\cap$, and $\cup$.

For $r \in \mathcal{Y}$, we have $\ell^f_{\text{abs}}(r, y) = 2\ell^f(r, y)$, meaning ℓ^f_{abs} matches (twice) ℓ^f on $\mathcal{Y}$. Were the "abstain" reports $v \in \mathcal{V} \setminus \mathcal{Y}$ dominated, then we would indeed have consistency. Following the above intuition, however, we can show that whenever f is submodular but not modular, there are situations where abstaining is uniquely optimal (relative to $\mathcal{V}$), leading to inconsistency.

4. Inconsistency for structured binary classification

Leveraging the embedded loss ℓ^f_{abs}, we now show that L^f is inconsistent for its intended target ℓ^f, except when f is modular. As the modular case is already well understood, under the name *weighted Hamming loss* (§ 2.3), this result essentially says that L^f is inconsistent for all nontrivial cases.

As L^f embeds ℓ^f_{abs}, to show inconsistency we may focus on reports $v \in \mathcal{V} \setminus \mathcal{Y}$, i.e., those that abstain on at least one index. Intuitively, if such a report is ever optimal, then L^f with the link sign* has a "blind spot" with respect to the indices in $A_v := \{v = 0\}$. We can leverage this blind spot to "fool" L^f, by making it link to an incorrect report. In particular, we will focus on the uniform distribution $\bar{p}$ on $\mathcal{Y}$, and perturb it slightly to find an optimal L^f point $v \in \mathcal{V}$ which maps to a ℓ^f suboptimal report sign$^*(v)$. In fact, we will show that one can always find such a point violating consistency, unless f is modular.

Given our focus on the uniform distribution, the following definition will be useful: for any set function f, let $\bar{f} := 2^{-k} \sum_{S \subseteq [k]} f(S) \in \mathbb{R}$. The next two lemmas relate $\bar{f}$ and $f([k])$ to expected loss and modularity. The proofs follow from summing the submodularity inequality over all possible subsets, and observing that at least one of them is strict when f is non-modular.

Lemma 12 *For all $v \in \mathcal{V}$, $\ell^f_{abs}(v; \bar{p}) \geq f([k])$. For all $r \in \mathcal{Y}$, $\ell^f_{abs}(r; \bar{p}) = 2\bar{f}$.*

Lemma 13 *Let f be submodular and normalized. Then $\bar{f} \geq f([k])/2$, and $\bar{f} = f([k])/2$ if and only if f is modular.*

Typical proofs of inconsistency identify a particular pair of distributions $p, p' \in \Delta(\mathcal{Y})$ for which the same surrogate report u is optimal, yet two distinct target reports are uniquely optimal for each, r for p and r' for p'. As u cannot link to both r and r', one concludes that the surrogate cannot be consistent. We follow this same general approach, but face one additional hurdle: we wish to show inconsistency of L^f for *all* non-modular f simultaneously. In particular, the distributions p, p' may need to depend on the choice f, so at first glance it may seem that such an argument would be quite complex. We achieve a relatively straightforward analysis by defining p, p' based on only a single parameter of f; the optimal surrogate report itself may be entirely governed by f, but will lead to inconsistency regardless.

The proof relies on a similar symmetry observation as Lemma 1, that $L^f(u \odot y', y \odot y') = L^f(u, y)$; in particular, prop$[L^f]$ has the same symmetry. For $p \in \Delta(\mathcal{Y})$ and $r \in \mathcal{Y}$, define $p \odot r \in \Delta(\mathcal{Y})$ by $(p \odot r)_y = p_{y \odot r}$.

Lemma 14 *For all $p \in \Delta(\mathcal{Y})$ and $r \in \mathcal{Y}$,* $\text{prop}[L^f](p \odot r) = \text{prop}[L^f](p) \odot r$.

Theorem 15 *Let f be submodular, normalized, and increasing. Then (L^f, sign) is consistent if and only if f is modular.*

Proof When f is modular, we may write $f = f_w$ for some $w \in \mathbb{R}^k_+$. Here L^{f_w} is weighted hinge loss (eq. (5)), which is known to be consistent for ℓ^{f_w}, which is weighted Hamming loss (Gao and Zhou, 2011, Theorem 15). (Briefly, for all $p \in \Delta_{\mathcal{Y}}$ the loss $L^{f_w}(\cdot; p)$ is linear on $[-1, 1]^k$, so it is minimized at a vertex $r \in \mathcal{Y}$. Hence $\mathcal{Y}$ is representative, so Theorem 6 gives that L^{f_w} embeds $L^{f_w}|_{\mathcal{Y}} = 2\ell^{f_w}$. Consistency follows from Theorem 17.)

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

Define $\epsilon = \bar{f}/(2\bar{f} - f([k]))$. We have $\epsilon > 0$ by Lemma 13 and submodularity of f. For any $y \in \mathcal{Y}$, let $p^y = (1-\epsilon)\bar{p} + \epsilon\delta_y$, where again $\bar{p}$ is the uniform distribution, and δ_y is the point distribution on y.

First, for all $r \in \mathcal{Y}$ with $r \neq y$, we have $\{r \odot y < 0\} \neq \varnothing = \{y \odot y < 0\}$. Since $\{\varnothing\} = \arg\min_{S \subseteq [k]} f(S)$, we have

$$\begin{aligned} \ell^f(r; p^y) &= (1-\epsilon)2\bar{f} + \epsilon\, 2f(\{r \odot y < 0\}) \\ &> (1-\epsilon)2\bar{f} + \epsilon\, 2f(\{y \odot y < 0\}) \\ &= \ell^f(y; p^y) \ , \end{aligned}$$

giving $\mathrm{prop}[\ell^f](p^y) = \{y\}$. On the other hand, from Lemma 12 and the fact that ℓ^f_{abs} agrees with ℓ^f, we have for all $r \in \mathcal{Y}$,

$$\ell^f_{\mathrm{abs}}(r; p^y) \geq \ell^f_{\mathrm{abs}}(y; p^y) = (1-\epsilon)2\bar{f} > f([k]) = \ell^f_{\mathrm{abs}}(0; p^y) \ .$$

We conclude there exists some optimal report $v \in \mathrm{prop}[\ell^f_{\mathrm{abs}}](p^y) \setminus \mathcal{Y}$. By Theorem 11, $v \in \mathrm{prop}[L^f](p^y)$ as well.

As $v \notin \mathcal{Y}$, in particular, $\{v = 0\} \neq \varnothing$. Now define $y' \in \mathcal{Y}$ to disagree with y on $\{v = 0\}$; formally, $y'_i = v_i$ if $v_i \neq 0$ and $y'_i = -y_i$ if $v_i = 0$. Although $y' \neq y$ (as $\{v = 0\} \neq \varnothing$), we have by construction that $v \odot (y \odot y') = v$. Furthermore, $p^y \odot (y \odot y') = p^{y'}$. By Theorem 11 and Lemma 14 then, $v \in \mathrm{prop}[L^f](p^{y'})$. By the above, however, we also have $\{y'\} = \mathrm{prop}[\ell^f](p^{y'})$. As $\mathrm{sign}^*(v)$ cannot be both y and y', at least one of p^y and $p^{y'}$ exhibits the inconsistency of L^f for ℓ^f. Specifically, calibration is violated (Definition 3) as v achieves the optimal L^f-loss for both p^y and $p^{y'}$, but for at least one, links to a report not in $\mathrm{prop}[\ell^f]$. ■

5. Constructing a calibrated link for $\ell^f_{\mathbf{abs}}$

As L^f embeds ℓ^f_{abs} from Theorem 11, Theorem 17 below further implies L^f is consistent with respect to ℓ^f_{abs} for some link function. Yet, the design of such a link function is not immediately clear. Indeed, natural choices turn out to be inconsistent in general, such as the threshold link ψ_c for $c > 0$ used by the BEP surrogate (§ 2.3), which given by $(\psi_c(u))_i = 0$ whenever $|u_i| < c$ and $(\psi_c(u))_i = \mathrm{sign}(u_i)$ otherwise (Figure 1). We instead follow the construction of an ϵ-separated link from Finocchiaro et al. (2022), resulting in two consistent link functions. Interestingly, while these links do not depend on f, they are calibrated with respect to ℓ^f_{abs} for all $f \in \mathcal{F}_k$ simultaneously. See § B for omitted proofs.

5.1. Approach via separated link functions

For any polyhedral loss L which embeds a target discrete loss ℓ, Finocchiaro et al. (2022) give a construction of a link function ψ such that (L, ψ) is calibrated with respect to ℓ. Their construction is based on ϵ-separation, as follows.

Definition 16 ((Finocchiaro et al., 2022, Construction 1)) *Let a polyhedral loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ that embeds some discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ be given, along with $\epsilon > 0$, and a norm $\|\cdot\|$. The* ϵ-thickened link envelope *$\Psi : \mathbb{R}^d \rightrightarrows \mathcal{R}$ is constructed as follows. Define $\mathcal{U} = \{\text{prop}[L](p) : p \in \Delta_\mathcal{Y}\}$ and, for each $U \in \mathcal{U}$, let $R_U = \{r \in \mathcal{R} : \varphi(r) \in U\}$, the reports whose embedding points are in U. Initialize by setting $\Psi(u) = \mathcal{R}$ for all $u \in \mathbb{R}^d$. Then for each $U \in \mathcal{U}$, and all points u such that $\inf_{u^* \in U} \|u^* - u\| < \epsilon$, update $\Psi(u) = \Psi(u) \cap R_U$.*

We say a link envelope Ψ is nonempty pointwise if $\Psi(u) \neq \varnothing$ for all $u \in \mathbb{R}^d$. Similarly, a link function ψ is pointwise contained in Ψ if $\psi(u) \in \Psi(u)$ for all $u \in \mathbb{R}^d$.

Theorem 17 ((Finocchiaro et al., 2022, Theorems 5, 6)) *Let $L : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}_+$ embed a discrete target $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$, and let Ψ be defined as in Definition 16. Then Ψ is nonempty pointwise for all sufficiently small ϵ. Furthermore, for any link function ψ pointwise contained in Ψ, the pair (L, ψ) is consistent with respect to ℓ.*

Essentially, this construction "thickens" each potentially optimal set and ensures surrogate report that is close to these regions must be linked to a representative report contained in that set. One can consider Ψ the resulting "link envelope", from which a calibrated link may be arbitrarily chosen pointwise.

To apply this construction to the Lovász hinge L^f, let Ψ^f be the envelope Ψ from Definition 16 applied to L^f. We immediately encounter a complication: as the link envelope Ψ^f depends on the choice of f, it is entirely possible that no single link function is contained in the envelopes Ψ^f for all $f \in \mathcal{F}_k$, i.e., is simultaneously calibrated for L^f for *all* such f. If no simultaneous link existed, the construction and analysis would have to be tailored carefully to each $f \in \mathcal{F}_k$. Interestingly, we show that such a simultaneous link does exist.

To find a link which is calibrated for all f, we identify certain structure which is common to Lovász hinges L^f. We encode this structure in a common link envelope $\hat{\Psi}$, and then show in Proposition 19 that, for all $f \in \mathcal{F}_k$ and $u \in \mathbb{R}^k$, we have $\hat{\Psi}(u) \subseteq \Psi^f(u)$. We then show that $\hat{\Psi}$ is nonempty for sufficiently small ϵ, meaning it contains a link option pointwise. This link is therefore contained in all the link envelopes Ψ^f for all f, and hence is calibrated with respect to ℓ^f_{abs} for all $f \in \mathcal{F}_k$ simultaneously.

5.2. The common link envelope $\hat{\Psi}$

We now present our link envelope $\hat{\Psi}$, used to construct calibrated links (Figure 1, left).

Definition 18 *Let $\mathcal{V}^{face} := \cup_{\pi \in \mathcal{S}_k, y \in \mathcal{Y}} 2^{V_{\pi,y}}$ be the subsets of $\mathcal{V}$ whose convex hulls are faces of some $P_{\pi,y}$ polytope. Define $\hat{\Psi} : \mathbb{R}^k \to 2^\mathcal{V}$ by $\hat{\Psi}(u) = \cap\{V \in \mathcal{V}^{face} \mid d_\infty(\text{conv}\, V, \overline{u}) < \epsilon\}$.*

Now we show that $\hat{\Psi} \subseteq \Psi^f$ pointwise. The proof uses the fact that both $\hat{\Psi}(u)$ and $\Psi^f(u)$ are constructed by the intersections of sets, and shows that the sets generating $\hat{\Psi}(u)$ are subsets of those generating $\Psi^f(u)$ for all $f \in \mathcal{F}_k$. In particular, every possible optimal set in the range of $\text{prop}[L^f]$ is a union of faces generated by convex hulls of elements of $\mathcal{V}^{\text{face}}$.

Proposition 19 *For all $f \in \mathcal{F}_k$ and $u \in \mathbb{R}^k$, we have $\hat{\Psi}(u) \subseteq \Psi^f(u)$.*

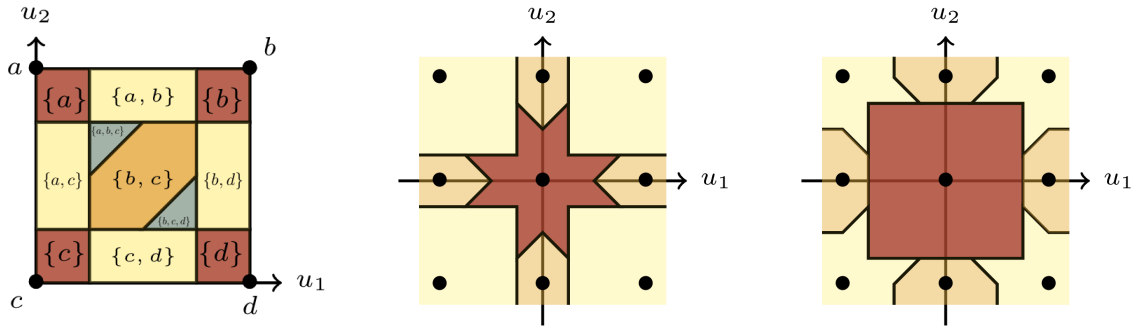

Figure 1: The link envelope $\hat{\Psi}$ (left) and link functions ψ^*_ϵ (middle) and $\psi^\diamond_\epsilon$ (right) for $k = 2$ and $\epsilon = \frac{1}{4}$. The envelope $\hat{\Psi}$ is pictured for $u \in \mathbb{R}^2_+$, with each region labeled by the value of $\hat{\Psi}$; a link is calibrated if it always links to one of the nodes in the region. The values for the link functions ψ^*_ϵ and $\psi^\diamond_\epsilon$ are given by the unique point $v \in \mathcal{V}$ that each depicted region contains. In particular, both links satisfy the constraints from $\hat{\Psi}$ (left) and thus are calibrated.

We now characterize the link envelope $\hat{\Psi}(u)$ in terms of the coordinates of u. In particular, $\hat{\Psi}(u)$ consists of the embedding points $v \in \mathcal{V}$ that make up the intersection of the faces from $\mathcal{V}^{\text{face}}$ that are ϵ close to u. We can express these points in terms of the ordered elements of $|u|$. In particular, such a point $v \in \mathcal{V}$ appears in the intersection exactly when the corresponding elements of $|u|$ are 2ϵ-far from each other, since otherwise we can find a face not containing v which is ϵ-close to u (Proposition 20). Therefore, Ψ is always nonempty when ϵ is small enough to guarantee a gap of at least 2ϵ in the ordered elements of $|u|$ (Lemma 21).

Proposition 20 *Let $u \in \mathbb{R}^k$, and let $\pi \in \mathcal{S}_k$ order the elements of $|u|$ (descending). For the purposes of the following, define $|u_{\pi_0}| = 1 + \epsilon$ and $|u_{\pi_{k+1}}| = -\epsilon$. Then we have*

$$\hat{\Psi}(u) = \{\mathbb{1}_{\pi,i} \odot \operatorname{sign}^*(u) \mid i \in \{0, 1, \ldots, k\},\ |u_{\pi_i}| \geq |u_{\pi_{i+1}}| + 2\epsilon\} \tag{16}$$

Lemma 21 *$\hat{\Psi}$ is nonempty pointwise if and only if $\epsilon \in (0, \frac{1}{2k}]$.*

5.3. Two calibrated link functions from $\hat{\Psi}$

We now proceed to construct two ϵ-separated links, ψ^*_ϵ, which abstains as little as possible, and $\psi^\diamond_\epsilon$, which abstains as much as possible. For sufficiently small ϵ, both links are pointwise contained in $\hat{\Psi}$, giving calibration from Theorem 17.

Definition 22 *Let $\epsilon > 0$ be fixed. Let $u \in \mathbb{R}^k$, and let $\pi \in \mathcal{S}_k$ order the elements of $|u|$. Given any $u \in \mathbb{R}^k$, let $i^* \in \{0, \ldots, k\}$ be the largest index i such that $|u_{\pi_i}| - |u_{\pi_{i+1}}| \geq 2\epsilon$ where we define $|u_{\pi_0}| = 1 + \epsilon$ and $|u_{\pi_{k+1}}| = -\epsilon$. Then define*

$$\psi^*_\epsilon(u) = \mathbb{1}_{\pi,i^*} \odot \operatorname{sign}^*(\overline{u})\ . \tag{17}$$

Similarly, let $i^\diamond \in \{0, \ldots, k\}$ be the smallest index i such that $|u_{\pi_i}| - |u_{\pi_{i+1}}| \geq 2\epsilon$ and define

$$\psi^\diamond_\epsilon(u) = \mathbb{1}_{\pi,i^\diamond} \odot \operatorname{sign}^*(\overline{u})\ . \tag{18}$$

Theorem 23 *Let $\epsilon \in (0, 1/2k]$, and fix any $f \in \mathcal{F}_k$. Then (L^f, ψ_ϵ^*) and $(L^f, \psi_\epsilon^\diamond)$ are well-defined and calibrated with respect to ℓ^f_{abs}.*

Proof Lemma 21 shows that the indices i^* and $i^\diamond$ in Definition 22 always exist when $\epsilon \in (0, \frac{1}{2k}]$, which shows that ψ_ϵ^* and $\psi_\diamond^*$ are well-defined. By construction, we have $\psi_\epsilon^*(u) \in \hat{\Psi}(u)$ and $\psi_\epsilon^\diamond(u) \in \hat{\Psi}(u)$ for all $u \in \mathbb{R}^k$. As Proposition 19 states that $\hat{\Psi} \subseteq \Psi^f$ pointwise, we then have $\psi_\epsilon^*, \psi_\epsilon^\diamond \in \Psi^f$ pointwise. Finally, Theorem 17 states that any link function contained in Ψ^f pointwise is calibrated. ■

The two proposed link functions, ψ_ϵ^* and $\psi_\epsilon^\diamond$, differ by how often one abstains vs the other. The first, ψ_ϵ^*, has a smaller abstain region which decreases in volume as ϵ decreases. Meanwhile, $\psi_\epsilon^\diamond$ has a larger abstain region which increases in volume as ϵ decreases. Based on one's preferred risk, either ψ_ϵ^* if risk seeking otherwise $\psi_\epsilon^\diamond$ if risk adverse could be used. The difference between how often ψ_ϵ^* and $\psi_\epsilon^\diamond$ abstain is demonstrated for $k = 2$ in Figure 1.

6. Discussion and conclusion

Despite the popularity of the Lovász hinge, we show in this work that it is inconsistent for structured binary prediction, its desired target. Instead, we show that it is consistent for the *structured abstain problem*, a variation of structured binary prediction in which one may abstain on a subset of predictions.

Our results crucially leverage the embedding framework of Finocchiaro et al. (2022). In particular, we rely heavily on the embedding framework to find a calibrated link function, as it allows us to prove calibration simultaneously for all submodular set functions parameterizing the problem.

Beyond investigating the utility of abstain options in practice, in analogy to the classification literature (Bartlett and Wegkamp, 2008; Ramaswamy et al., 2018), we see two important theoretical directions. First, for certain submodular functions $f \in \mathcal{F}_k$, the problem ℓ^f_{abs} may contain redundant reports; indeed, we know this must be the case for $f_{\text{0-1}}$, since every report $v \in \mathcal{V} \setminus \mathcal{Y}$ is dominated by 0. We would like to characterize the redundant reports for a given function f and modify the link function to avoid linking to them.

Second, our work sheds light on broader questions about when consistent convex surrogates $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ can be designed with low *prediction dimension* d. Recent works have developed tools to bound the prediction dimension required (Finocchiaro et al., 2021, 2020; Ramaswamy and Agarwal, 2016), yet general bounds, especially constructive upper bounds, remain elusive. In particular, structured prediction problems such as binary structured prediction often have exponentially large label sets $\mathcal{Y}$, and one seeks a consistent convex surrogate with prediction dimension logarithmic in $|\mathcal{Y}|$. Yet the BEP surrogate (7) has been perhaps the only such surrogate in the literature, with $d = \lceil \log |\mathcal{Y}| \rceil$. Our analysis adds an entire family of surrogates to this list, for any submodular function ($d = k$ and $|\mathcal{Y}| = 2^k$); we hope these additional positive examples could shed further light on the conditions required for a target loss to have a consistent low-dimensional convex surrogate.

Acknowledgements. We thank Anish Thilagar and Bo Waggoner for comments and suggestions, and Eric Balkanski for insights about submodular functions. We gratefully acknowledge support from the National Science Foundation under Grant No. IIS-2045347.

References

Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *European Conference on Computer Vision*, pages 158–177. Springer, 2020.

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

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(Aug):1823–1840, 2008.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018.

Arne Brondsted. *An introduction to convex polytopes*, volume 90. Springer Science & Business Media, 2012.

Stanley S Chang and Chi-Kwong Li. Certain isometries on rn. *Linear algebra and its applications*, 165:251–265, 1992.

Yiwei Chen, Jingtao Xu, Jiaqian Yu, Qiang Wang, ByungIn Yoo, and Jae-Joon Han. Afod: Adaptive focused discriminative segmentation tracker. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 666–682, Cham, 2020. Springer International Publishing. ISBN 978-3-030-68238-5.

Jessica Finocchiaro, Rafael Frongillo, and Bo Waggoner. Unifying lower bounds on prediction dimension of convex surrogates. In *Proceedings of Advances In Neural Information Processing Systems (NeurIPS)*, 2021.

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

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

Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for the design and analysis of consistent polyhedral surrogates. *arXiv*, 2022.

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.

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.

László Lovász. Submodular functions and convexity. In *Mathematical programming the state of the art*, pages 235–257. Springer, 1983.

David McAllester. Generalization bounds and consistency. *Predicting structured data*, pages 247–261, 2007.

Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8837–8845, 2019.

Sebastian Nowozin. Optimal decisions from probabilistic models: the intersection-over-union case. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 548–555, 2014.

Anton Osokin, Francis Bach, and Simon Lacoste-Julien. On structured prediction theory with calibrated convex surrogate losses. In *Advances in Neural Information Processing Systems*, pages 302–313, 2017.

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

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

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, Yasemin Altun, and Yoram Singer. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(9), 2005.

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.

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

Notation	Explanation
k	Number of binary events
$[k] := \{1, \dots, k\}$	Index set
$y \in \mathcal{Y} = \{-1, 1\}^k$	Label space
$v \in \mathcal{V} = \{-1, 0, 1\}^k$	(Abstain) prediction space
$r \in \mathcal{R}$	General prediction space
$R = [-1, 1]^k$	The filled ± 1 hypercube
$u \in \mathbb{R}^k$	Surrogate prediction space
$\{u \le c\} = \{i \in [k] \mid u_i \le c\}$	Set of indices of u less than c
$(u \odot u')_i = u_i u'_i$	Hadamard (element-wise) product
$U \odot u' = \{u \odot u' \mid u \in U\}$	Hadamard product on a set $U \subseteq \mathbb{R}^k$
$\text{sign} : \mathbb{R}^k \to \mathcal{V}$	Sign function including 0
$\text{sign}^* : \mathbb{R}^k \to \mathcal{Y}$	Sign function breaking ties arbitrarily at 0
$\lvert u \rvert \in \mathbb{R}^k_+$ s.t. $\lvert u \rvert_i = \lvert u_i \rvert$	Observe $\lvert u \rvert = u \odot \text{sign}^*(u) = u \odot \text{sign}(u)$
$\overline{u} = \text{sign}(u) \odot \min(\lvert u \rvert, \mathbb{1})$	"Clipping" of u to R
$\mathbb{1}_S \in \{0, 1\}^k$ s.t. $(\mathbb{1}_S)_i = 1 \iff i \in S$	$0-1$ Indicator on set $S \subseteq [k]$
$\pi \in \mathcal{S}_k$	Permutations of $[k]$
$f \in \mathcal{F}_k$	Set of normalized, increasing, and submodular set functions $f : 2^k \to \mathbb{R}_+$.
$\ell^f(r, y) = f(\{r \odot y < 0\})$	Structured binary classification eq. (1)
$F(x)$	Lovaśz extension for $x \in \mathbb{R}^k_+$ in eq. (2)
$L^f(u, y) = F((\mathbb{1} - u \odot y)_+)$	Lovász hinge eq. (4)
$\ell^f_{\text{abs}}(v, y) = f(\{v \odot y < 0\}) + f(\{v \odot y \le 0\})$	Structured abstain problem eq. (14)

Table 1: Table of general notation

Notation	Explanation
$\mathbb{1}_{\pi,i} = \mathbb{1}_{\{\pi_1, \dots, \pi_i\}}$ with $\mathbb{1}_{\pi,0} = \mathbf{0}$	Indicator of first i elements of π
$V_\pi = \{\mathbb{1}_{\pi,i} \mid i \in \{0, \dots, k\}\}$	Elements of $\mathcal{V}$ ordered by π
$V_{\pi,y} = V_\pi \odot y$	Signed elements of $\mathcal{V}$ ordered by π.
$P_\pi = \{x \in \mathbb{R}^k_+ \mid x_{\pi_1} \ge \dots \ge x_{\pi_k}\}$	elements of $\mathbb{R}^k_+$ ordered by π
$P_{\pi,y} = \text{conv}\, V_\pi \odot y$	Elements of P_π signed by y
$\mathcal{V}^{\text{face}} = \cup_{\pi \in \mathcal{S}_k, y \in \mathcal{Y}} 2^{V_{\pi,y}}$	Subsets of $\mathcal{V}$ whose convex hulls are faces of some $P_{\pi,y}$ polytope.
$\hat{\Psi}(u) = \cap\{V \in \mathcal{V}^{\text{face}} \mid d_\infty(\text{conv}\, V, u) < \epsilon\}$	Proposed general link envelope.
$\mathcal{U}^f = \text{prop}[L^f](\Delta_\mathcal{Y})$	Range of property elicited by Lovász hinge
$\Psi^f(u) = \cap\{U \in \mathcal{U}^f \mid d_\infty(U, u) < \epsilon\} \cap \mathcal{V}$	Link envelope for given $f \in \mathcal{F}_k$.

Table 2: Table of notation used for proofs

Appendix A. Notation tables

See Tables 1 and 2.

Appendix B. Omitted Proofs

B.1. Omitted Proofs from § 3

Lemma 7 *For any $u \in \mathbb{R}^k$, we have $L^f(\overline{u}, y) \leq L^f(u, y)$ for all $y \in \mathcal{Y}$.*

Proof Fix $y \in \mathcal{Y}$. Let $w = \mathbb{1} - u \odot y$ and $\overline{w} = \mathbb{1} - \overline{u} \odot y$, so that $L^f(u, y) = F(w_+)$ and $L^f(\overline{u}, y) = F(\overline{w}_+)$. We will first show that $\overline{w}_+ = \min(w_+, 2)$, where the minimum is element-wise.

For $i \in [k]$ such that $|u_i| \leq 1$, we have $\overline{u}_i = u_i$. Thus $(w_+)_i = (1 - u_i y_i)_+ = (1 - \overline{u}_i y_i)_+ = (\overline{w}_+)_i$. Furthermore, we have $0 \leq (w_+)_i = (\overline{w}_+)_i \leq 2$. Now suppose $|u_i| > 1$. If $y_i u_i > 0$, i.e., $\operatorname{sign}(u_i) = y_i$, then $1 - y_i u_i = 1 - |u_i| < 0$, so $(w_+)_i = 0$. For $\overline{u}$, we similarly have $(\overline{w}_+)_i = (1 - |\overline{u}_i|)_+ = 0$. In the other case, $y_i u_i < 0$, so $(w_+)_i = 1 + |u_i| > 2$ and $(\overline{w}_+)_i = 1 + |\overline{u}_i| = 2$. Therefore, we have $\overline{w}_+ = \min(w_+, 2)$.

Now, let $\pi \in \mathcal{S}_k$ be a permutation that orders the elements of w_+. Observe that π orders the elements of $\overline{w}_+$ as well, since the vectors are identical except for values above 2, which are all mapped to 2. By eq. (3), we thus have

$$\begin{aligned}
F(w_+) - F(\overline{w}_+) &= \sum_{i=1}^{k} (w_+)_{\pi_i} (f(\{\pi_1, \ldots, \pi_i\}) - f(\{\pi_1, \ldots, \pi_{i-1}\})) \\
&\qquad - \sum_{i=1}^{k} (\overline{w}_+)_{\pi_i} (f(\{\pi_1, \ldots, \pi_i\}) - f(\{\pi_1, \ldots, \pi_{i-1}\})) \\
&= \sum_{i=1}^{k} (w_+ - \overline{w}_+)_{\pi_i} (f(\{\pi_1, \ldots, \pi_i\}) - f(\{\pi_1, \ldots, \pi_{i-1}\})) \\
&\geq 0 \,,
\end{aligned}$$

where we have used the fact that f is increasing and $\overline{w}_+ \leq w_+$ element-wise. As y was arbitrary, this holds for all $y \in \mathcal{Y}$. ∎

Lemma 8 *The sets $P_{\pi,y}$ satisfy the following.*

(i) $\cup_{y \in \mathcal{Y}, \pi \in \mathcal{S}_k} P_{\pi,y} = [-1, 1]^k$.

(ii) For all $f \in \mathcal{F}_k$, $y, y' \in \mathcal{Y}$, and $\pi \in \mathcal{S}_k$, the function $L^f(\cdot, y')$ is affine on $P_{\pi,y}$.

Proof For (i), take any $u \in [-1, 1]^k$. Letting $y = \operatorname{sign}^*(u)$, we have $u \odot y = |u| \in \mathbb{R}^k_+$. Taking π to be any permutation ordering the elements of $u \odot y$, we have $u \odot y \in P_\pi \cap \mathbb{R}^k_+$. Notice, since $u \odot y \in P_\pi \cap \mathbb{R}^k_+$ and $u \in [-1, 1]^k$, we additionally have $u \odot y = |u| \in P_\pi \cap [0, 1]^k$. Since $\mathbb{1}_{\pi,i}$ for $i \in \{0, \ldots, k\}$ form V_π and P_π is the convex hull of points in V_π, showing there is an α such that $u = \sum_i \alpha_i \mathbb{1}_{\pi,i}$ suffices to conclude $u \in P_{\pi,y}$. We can write $u \odot y$ as the convex combination $u \odot y = \sum_{i=0}^{k} \alpha_i (u \odot y) \mathbb{1}_{\pi,i}$, as in eq. (11). Thus $u = u \odot y \odot y = \sum_{i=0}^{k} \alpha_i (u \odot y) \mathbb{1}_{\pi,i} \odot y$, so $u \in P_{\pi,y}$. Therefore, every $u \in [-1, 1]^k$ is in some $P_{\pi,y}$, we have $\cup_{y \in \mathcal{Y}, \pi \in \mathcal{S}_k} P_{\pi,y} \supseteq [-1, 1]^k$. Moreover, every $P_{\pi,y} \subseteq [-1, 1]^k$ by construction, and equality follows.

For (ii), first observe for all $\pi \in \mathcal{S}_k$, the function F is affine on P_π, immediately from eq. (3). To show $L^f(\cdot, y') = F((\mathbb{1} - u \odot y')_+)$ is affine on $P_{\pi,y}$ for all $y, y' \in \mathcal{Y}, \pi \in \mathcal{S}_k$, it

therefore suffices to show there exists some π' such that $\{\mathbb{1} - u \odot y' \mid u \in P_{\pi,y}\} \subseteq P_{\pi'}$. As $L^f(u, y') = F(\mathbb{1} - u \odot y')$ when $u \in [-1, 1]^k$, the result will follow.

We construct π', unraveling the permutation π into two permutations, depending on the sign of $y \odot y'$. Recall from the discussion following eq. (2) that π orders the elements of $u \odot y = |u|$ in decreasing order. Observe that $u \odot y' = u \odot (y \odot y) \odot y' = (u \odot y) \odot (y \odot y') = |u| \odot (y \odot y')$. Thus, π orders the elements of $u \odot y'$ in decreasing order among indices i with $y_i y'_i > 0$, and increasing order on the others. Therefore π orders the elements of $\mathbb{1} - u \odot y'$ in increasing order among indices i with $y_i y'_i > 0$, and decreasing order on the others. Taking π' to be the order given by sorting the elements in $\{y \odot y' < 0\}$ according to π, followed by the remaining elements according to the reverse of π, we have shown $\mathbb{1} - u \odot y' \in P_{\pi'}$. ■

We now introduce a lemma used in the proof of Lemma 9.

Lemma 24 *Let $L : \mathbb{R}^k \to \mathbb{R}_+$ be a polyhedral function that is affine on the polyhedron C. For any $x \in \text{relint}(C)$ and any $z \in C$, we have $\partial L(x) \subseteq \partial L(z)$.*

Proof Fix $x \in \text{relint}(C)$. Since L is affine on C, then there exists some $w' \in \mathbb{R}^k, b \in \mathbb{R}$ such that $L(z) = \langle w', z\rangle + b$ for all $z \in C$. Thus, we have $L(z) - L(x) = (\langle w', z\rangle + b) - (\langle w', x\rangle + b) = \langle w', z - x\rangle$ for all $z \in C$.

We claim that for all $w \in \partial L(x)$, and all $z \in C$, we have $\langle w, z - x\rangle = \langle w', z - x\rangle$. To prove this claim, observe that

$$\langle w', z - x\rangle = L(z) - L(x) \geq \langle w, z - x\rangle \text{ for all } z \in C \ , \tag{19}$$

by the subgradient inequality and affineness of L on C. Assume for a contradiction that $\langle w', z - x\rangle > \langle w, z - x\rangle$ for some $z \in C$. Since $x \in \text{relint}(C)$, there is an $\epsilon < 0$ such that $z' := x + \epsilon(z - x) \in C$. Therefore, we have

$$\langle w', z' - x\rangle = \langle w', \epsilon(z - x)\rangle = \epsilon\langle w', z - x\rangle < \epsilon\langle w, z - x\rangle = \langle w, \epsilon(z - x)\rangle = \langle w, z' - x\rangle \ ,$$

where we use the fact that $\epsilon < 0$ to flip the inequality. We have now contradicted eq. (19) for the point z'.

Since we now have $L(z) - L(x) = \langle w', z - x\rangle = \langle w, z - x\rangle$ for all $z \in C$, consider $w \in \partial L(x)$. Then we have, for all $v \in \mathbb{R}^k$,

$$\begin{aligned} L(v) - L(z) &= (L(v) - L(x)) + (L(x) - L(z)) \\ &\geq \langle w, v - x\rangle + \langle w, x - z\rangle \\ &= \langle w, v - z\rangle \ , \end{aligned}$$

where the inequality follows from the subgradient inequality and the claim. Thus $w \in \partial L(z)$, which completes the proof. ■

A corollary of Lemma 24 is that subdifferentials are constant on $\text{relint}(C)$ for any face C such that L is affine as the subset inclusion holds in both directions.

Lemma 9 *Given a polyhedral loss function $L : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}_+$, let $\mathcal{C}$ be a collection of polyhedral subsets of $\mathbb{R}^k$ such that for all $y \in \mathcal{Y}$, $L(\cdot, y)$ is affine on each $C_i \in \mathcal{C}$, and denote $\text{faces}(C_i)$ as the set of faces of C_i. Let $R = \cup\mathcal{C}$ be the union of these polyhedral subsets. Then for all $p \in \Delta_\mathcal{Y}$, $\text{prop}[L](p) \cap R = \cup\mathcal{F}$ for some $\mathcal{F} \subseteq \cup_i \text{faces}(C_i)$.*

Proof Fix $p \in \Delta_{\mathcal{Y}}$. For any $u \in R \cap \text{prop}[L](p)$, there is some $\mathcal{C}' \subseteq \mathcal{C}$ such that $u \in C_j$ for all $C_j \in \mathcal{C}'$. For now, let us simply consider any $C_j \in \mathcal{C}'$. Observe that $u \in \text{relint}(F_j)$ for exactly one face F_j of C_j.

By convexity of L, we have $u \in \text{prop}[L](p) \iff 0 \in \partial L(u;p)$. Moreover, as $u \in \text{relint}(F_j)$, we have $\partial L(u;p) \subseteq \partial L(z;p)$ for all $z \in F_j$ by Lemma 24. Thus, $0 \in \partial L(u;p)$ implies $0 \in \partial L(z;p)$ for all $z \in F_j$. Moreover, $0 \in \partial L(z;p)$ for all $z \in F_j$ if and only if $z \in \text{prop}[L](p)$ for all $z \in F_j$, and thus we have $F_j \subseteq \text{prop}[L](p)$.

As the value u and the index j were arbitrary, this holds for all such faces in $G(u) := \cup\{F_j \subseteq C_j \in \mathcal{C}' \mid u \in \text{relint}(F_j)\}$. Now, take $\mathcal{F} = \{G(u) \mid u \in R \cap \text{prop}[L](p)\}$; hence $\text{prop}[L](p) \cap R = \cup\mathcal{F}$. Moreover, $\mathcal{F} \subseteq \cup_i \text{faces}(C_i)$. ■

B.2. Omitted Proofs for § 4

Lemma 12 *For all $v \in \mathcal{V}$, $\ell^f_{abs}(v;\bar{p}) \geq f([k])$. For all $r \in \mathcal{Y}$, $\ell^f_{abs}(r;\bar{p}) = 2\bar{f}$.*

Proof Let $A_v = \{v = 0\}$ and $B_v = [k] \setminus A_v$. Recall that $\bar{p}$ is the uniform distribution on 2^k outcomes. Then we have

$$\begin{aligned}
\ell^f_{\text{abs}}(v;\bar{p}) &= 2^{-k} \sum_{S \subseteq [k]} f(S_v \triangle S \setminus A_v) + f(S_v \triangle S \cup A_v) \\
&= 2^{-|B_v|} \sum_{T \subseteq B_v} f(T) + f(T \cup A_v) \\
&= \frac{1}{2}\, 2^{-|B_v|} \sum_{T \subseteq B_v} f(T) + f(B_v \setminus T) + f(T \cup A_v) + f((B_v \setminus T) \cup A_v) \\
&\geq \frac{1}{2} \left(f(B_v) + f(\varnothing) + f([k]) + f(A_v)\right) \\
&\geq \frac{1}{2} \left(f([k]) + f([k])\right) = f([k]) \ ,
\end{aligned}$$

where we use submodularity in both inequalities. The second statement follows from the second equality above after setting $A_v = \varnothing$, as then $B_v = [k]$ and thus T ranges over all of $2^{[k]}$. ■

Lemma 13 *Let f be submodular and normalized. Then $\bar{f} \geq f([k])/2$, and $\bar{f} = f([k])/2$ if and only if f is modular.*

Proof The inequality follows from Lemma 12 with $r \in \mathcal{Y}$. Next, note that if f is modular we trivially have $\bar{f} = f([k])/2$. If f is submodular but not modular, we must have some $S \subseteq [k]$ and $i \in S$ such that $f(S) - f(S \setminus \{i\}) < f(\{i\})$. By submodularity, we conclude that $f([k]) - f([k] \setminus \{i\}) < f(\{i\})$ as well; rearranging, $f(\{i\}) + f([k] \setminus \{i\}) > f([k]) = f([k]) + f(\varnothing)$. Again examining the proof of Lemma 12, we see that the first inequality must be strict, as we have one such $T \subseteq [k]$, namely $T = \{i\}$, for which the inequality in submodularity is strict. ■

Lemma 14 *For all $p \in \Delta(\mathcal{Y})$ and $r \in \mathcal{Y}$, $\mathrm{prop}[L^f](p \odot r) = \mathrm{prop}[L^f](p) \odot r$.*

Proof We define $p \odot r \in \Delta(\mathcal{Y})$ by $(p \odot r)_y = p_{y \odot r}$.

$$\begin{aligned}
\mathrm{prop}[L^f](p \odot r) &= \operatorname*{arg\,min}_{u \in \mathbb{R}^k} \sum_{y \in \mathcal{Y}} (p \odot r)_y L^f(u, y) \\
&= \operatorname*{arg\,min}_{u \in \mathbb{R}^k} \sum_{y \in \mathcal{Y}} p_{y \odot r} L^f(u, y) && \text{Definition of } p \odot r \\
&= \operatorname*{arg\,min}_{u \in \mathbb{R}^k} \sum_{y \in \mathcal{Y}} p_{y \odot r} L^f(u \odot r, y \odot r) && \text{Lemma 1} \\
&= \operatorname*{arg\,min}_{u \in \mathbb{R}^k} \sum_{y' \in \mathcal{Y}} p_{y'} L^f(u \odot r, y') && \text{Substituting } y = y' \odot r \\
&= \left(\operatorname*{arg\,min}_{u' \in \mathbb{R}^k} \sum_{y' \in \mathcal{Y}} p_{y'} L^f(u', y') \right) \odot r \\
&= \mathrm{prop}[L^f](p) \odot r
\end{aligned}$$

■

B.3. Omitted proofs from § 5

Since $\overline{u} \in R$, "clipping" u' to $\overline{u'}$ can only reduce element-wise distance, and therefore $d_\infty(\overline{u}, \cdot)$ is still small, which allows us to restrict our attention to R.

Lemma 25 *Let $f \in \mathcal{F}_k$. For all $U \in \mathrm{prop}[L^f](\Delta_\mathcal{Y})$, $u \in \mathbb{R}^k$, and $0 < \epsilon < 2$, if $d_\infty(U, u) < \epsilon$ then $d_\infty(U \cap [-1, 1]^k, \overline{u}) < \epsilon$.*

Proof Since U is closed, we have some closest point $u' \in U$ to u, meaning $d_\infty(u', u) = d_\infty(U, u) < \epsilon$. As $\overline{u}' \in U$ by a corollary of Lemma 7, it suffices to show $d_\infty(\overline{u}, \overline{u}') < \epsilon$.

For each $i \in [k]$, we consider three cases. It suffices to show distance does not increase on each element by the choice of the $d_\infty(\cdot, \cdot)$ distance.

The cases are as follows: (i) $u_i = \overline{u}_i$ and $u' = \overline{u}'_i$, (ii) $u_i \neq \overline{u}_i$ and $u'_i \neq \overline{u}'_i$, and (iii) $u_i = \overline{u}_i$ and $u'_i \neq \overline{u}'_i$ (WLOG). Case (i) is trivial as $|u_i - u'_i| = |\overline{u}_i - \overline{u}'_i| < \epsilon$. In case (ii), we must have $\mathrm{sign}(u)_i = \mathrm{sign}(u')_i$ as $d_\infty(u, u') < \epsilon \implies |u_i - u'_i| < \epsilon$. If both u_i and u'_i are outside $[-1, 1]^k$, this inequality is only true (for $\epsilon < 2$) if the sign matches. Therefore $|\overline{u}_i - \overline{u}'_i| = |\mathrm{sign}(u)_i - \mathrm{sign}(u')_i| = 0 < \epsilon$. In case (iii), we have $\epsilon > |u_i - u'_i| > |u_i - 1| = |u_i - \overline{u}'_i|$. As absolute difference in each element does not increase, the $d_\infty(\cdot, \cdot)$ distance does not increase. ■

We now proceed to statements about the link envelope construction $\hat{\Psi}$.

Proposition 19 *For all $f \in \mathcal{F}_k$ and $u \in \mathbb{R}^k$, we have $\hat{\Psi}(u) \subseteq \Psi^f(u)$.*

Proof Let us define

$$\begin{aligned}
\mathcal{A}(u) &:= \{V \in \mathcal{V}^{\text{face}} \mid d_\infty(\mathrm{conv}\, V, \overline{u}) < \epsilon\}\ , \\
\mathcal{B}(u) &:= \{U \cap \mathcal{V} \mid U \in \mathcal{U}^f,\, d_\infty(U, u) < \epsilon\}\ ,
\end{aligned}$$

so that $\hat{\Psi}(u) = \cap\mathcal{A}(u)$ and $\Psi^f(u) = \cap\mathcal{B}(u)$. We wish to show $\cap\mathcal{A}(u) \subseteq \cap\mathcal{B}(u)$. It thus suffices to show the following claim: for all $B \in \mathcal{B}(u)$ we have some $A \in \mathcal{A}(u)$ with $A \subseteq B$. Since then $v \in \cap\mathcal{A}(u)$ implies $v \in A$ for all $A \in \mathcal{A}(u)$, which by the claim implies $v \in B$ for all $B \in \mathcal{B}(u)$ and thus $v \in \cap\mathcal{B}(u)$.

Let $B \in \mathcal{B}(u)$, so we may write $B = U \cap \mathcal{V}$ for $U \in \mathcal{U}^f$ with $d_\infty(U, u) < \epsilon$. By Lemma 25 we have $d_\infty(U \cap R, \overline{u}) = d_\infty(U, u) < \epsilon$. From Lemma 8, the set $R = [-1, 1]^k = \cup_{\pi \in \mathcal{S}_k, y \in \mathcal{Y}} P_{\pi,y}$ is the union of polyhedral subsets of $\mathbb{R}^k$, and $L(\cdot, y)$ is affine on each $P_{\pi,y}$. By Lemma 9, we then have $U \cap R = \cup\mathcal{F}$ for some $\mathcal{F} \subseteq \cup_{\pi,y}\text{faces}(P_{\pi,y})$. As each such face can be written as $\operatorname{conv} V$ for some $V \in \mathcal{V}^{\text{face}}$, we have some $\mathcal{V}' \subseteq \mathcal{V}^{\text{face}}$ such that $U \cap R = \cup\mathcal{F} = \cup_{V \in \mathcal{V}'} \operatorname{conv} V$. Now $\min_{V \in \mathcal{V}'} d_\infty(\operatorname{conv} V, \overline{u}) = d_\infty(U \cap R, \overline{u}) < \epsilon$, so we have some $V \in \mathcal{V}'$ such that $d_\infty(\operatorname{conv} V, \overline{u}) < \epsilon$. Thus $V \in \mathcal{A}(u)$ by definition. As $\operatorname{conv} V \subseteq U \cap R$, we have $V = (\operatorname{conv} V) \cap \mathcal{V} \subseteq (U \cap R) \cap \mathcal{V} = U \cap \mathcal{V} = B$, which proves the claim. ■

Lemma 26 *Fix $u \in [-1, 1]^k$, and consider π, y such that $u \in P_{\pi,y}$. Then $V^u_{\pi,y} := \{\mathbb{1}_{\pi,i} \odot y \mid i \in \{0, \dots, k\}, \alpha_i(|u|) \neq 0\}$ is the smallest (in cardinality) set of vertices such that $V^u_{\pi,y} \subseteq V_{\pi,y}$ and $u \in \operatorname{conv}(V^u_{\pi,y})$.*

Proof First, observe that $V^u_{\pi,y} \subseteq V_{\pi,y}$ by construction, as the first set is constructed the same as the second, with one additional constraint. Moreover, we have $u = \sum_{i=1}^k \alpha_i(|u|)u_i = \sum_{i:\alpha_i(|u|)\neq 0} \alpha_i(|u|)u_i \in \operatorname{conv} V^u_{\pi,y}$.

Now recall $P_{\pi,y}$ is a simplex (see "Linear interpolation on simplices" Bach (2013, pg. 167)) thus, by properties of simplex, each $u \in P_{\pi,y}$ has a unique convex combination expressed by the vertices of $V_{\pi,y}$ which are affinely independent (Brondsted, 2012, pg. 14, Thm 2.3). Therefore, every vertex i with a non-zero weighting $\alpha_i(|u|) \neq 0$ is necessary in order to express u as a convex combination due to the affine independence of the vertices. Thus, $V^u_{\pi,y} := \{\mathbb{1}_{\pi,i} \odot y \mid i \in \{0, \dots, k\}, \alpha_i(|u|) \neq 0\}$, and as $|V^u_{\pi,y}| < \infty$, has to be the smallest (in cardinality) set of vertices such such that $V^u_{\pi,y} \subseteq V_{\pi,y}$ and $u \in \operatorname{conv}(V^u_{\pi,y})$. ■

Moreover, $\hat{\Psi}$ is symmetric around signed permutations.

Lemma 27 *For all $u \in \mathbb{R}^k$, $y \in \mathcal{Y}$, and $\pi \in \mathcal{S}_k$, we have $\hat{\Psi}(\pi(u \odot y)) = \pi(\hat{\Psi}(u) \odot y)$, where we define $(\pi x)_i = x_{\pi_i}$ and we extend this operation to sets.*

Proof The proof that the permutation part ($\hat{\Psi}(\pi u) = \pi\hat{\Psi}(u)$) is straightforward from the definition. For sign changes, observe $\overline{u \odot y} = \operatorname{sign}(u \odot y)\dot{\min}(|u \odot y|, 1) = \operatorname{sign}(u) \odot y \odot \min(|u|, 1) = \overline{u} \odot y$. The operation $u \mapsto u \odot y$ is an isometry for the infinity norm as a special case of signed permutations, here the identity permutation (Chang and Li, 1992, Theorem 2.3). For all closed $U \subseteq \mathbb{R}^k$, we therefore have $d_\infty(\overline{u \odot y}, U \odot y) = d_\infty(\overline{u} \odot y, U \odot y) = d_\infty(\overline{u}, U)$. Therefore,

$$
\begin{aligned}
\hat{\Psi}(u \odot y) &= \cap\{V \in \mathcal{V}^{\text{face}} \mid d_\infty(\operatorname{conv} V, \overline{u \odot y}) < \epsilon\} \\
&= \cap\{V \in \mathcal{V}^{\text{face}} \mid d_\infty(\operatorname{conv} V \odot y, \overline{u}) < \epsilon\} && \overline{u \odot y} = \overline{u} \odot y, \text{ and } \overline{u} \odot y \odot y = \overline{u} \\
&&& \text{with } d_\infty \text{ preserved under } \odot. \\
&= \cap\{V \in \mathcal{V}^{\text{face}} \mid d_\infty(\operatorname{conv} V, \overline{u}) < \epsilon\} \odot y \\
&= \hat{\Psi}(u) \odot y \ .
\end{aligned}
$$

■

Proposition 20 *Let $u \in \mathbb{R}^k$, and let $\pi \in \mathcal{S}_k$ order the elements of $|u|$ (descending). For the purposes of the following, define $|u_{\pi_0}| = 1 + \epsilon$ and $|u_{\pi_{k+1}}| = -\epsilon$. Then we have*

$$\hat{\Psi}(u) = \{\mathbb{1}_{\pi,i} \odot \operatorname{sign}^*(u) \mid i \in \{0, 1, \dots, k\},\ |u_{\pi_i}| \geq |u_{\pi_{i+1}}| + 2\epsilon\} \tag{16}$$

Proof We will show the statement for $u \in \mathbb{R}^k_+$ with $u_1 \geq \cdots \geq u_k$, i.e., where $u \in P_{\pi^*}$ where π^* is the identity permutation. Lemma 27 then gives the result, as we now argue. For any $u \in \mathbb{R}^k$, let $\pi \in \mathcal{S}_k$ order the elements of $|u|$, and let $y = \operatorname{sign}^*(u)$. Then $\pi(u \odot y) = \pi|u| \in P_{\pi^*}$. Once we show eq. (16) is true on the unsigned, ordered case, eq. (16) gives $\hat{\Psi}(\pi|u|) = \{\mathbb{1}_{\pi^*,i} \mid i \in \{0, 1, \dots, k\},\ |u_{\pi_i}| \geq |u_{\pi_{i+1}}| + 2\epsilon\}$. Thus $\hat{\Psi}(u) = \hat{\Psi}(\pi(u \odot y)) = \pi(\hat{\Psi}(u) \odot y) = \{\pi(\mathbb{1}_{\pi^*,i} \odot y) \mid i \in \{0, 1, \dots, k\},\ |u_{\pi_i}| \geq |u_{\pi_{i+1}}| + 2\epsilon\} = \{\mathbb{1}_{\pi,i} \odot \operatorname{sign}^*(u) \mid i \in \{0, 1, \dots, k\},\ |u_{\pi_i}| \geq |u_{\pi_{i+1}}| + 2\epsilon\}$.

To begin, we show that for any $i \in \{0, 1, \dots, k\}$ where $u_i < u_{i+1} + 2\epsilon$, $\mathbb{1}_{\pi^*,i} \notin \hat{\Psi}(u)$ by the contrapositive. First, suppose that there exists an $i \in \{0, 1, \dots, k\}$ such that $u_i < u_{i+1} + 2\epsilon$. Since u is ordered, we know that $0 \leq u_i - u_{i+1} < 2\epsilon$.

Let $z = \frac{u_i + u_{i+1}}{2}$ and define $\hat{u}$ such that $\hat{u}_i = z$ and $\hat{u}_{i+1} = z$ while every other index of $\hat{u}$ is equal to u. Observe $u_i - z < \epsilon$ and $z - u_{i+1} < \epsilon$, and thus $d_\infty(u, \hat{u}) < \epsilon$ as $d_\infty(\cdot, \cdot)$ is measured component-wise. By Lemma 26 and construction of α in the first paragraph of § 3.2, we have $\alpha_i(\hat{u}) = \hat{u}_i - \hat{u}_{i+1} = 0$, we have $\hat{u} \in \operatorname{conv}(V_i)$, where $V_i := V^u_{\pi^*,y} \setminus \{\mathbb{1}_{\pi^*,i}\}$. Since $\hat{u} \in \operatorname{conv}(V_i)$ and $d_\infty(\hat{u}, u) < \epsilon$, we have $V_i \supseteq \hat{\Psi}(u)$, and therefore, for any $i \in \{0, 1, \dots, k\}$ such that $u_i < u_{i+1} + 2\epsilon$, $\mathbb{1}_{\pi^*,i} \notin \hat{\Psi}(u)$.

Now, for the converse, fix any $u \in P_{\pi^*}$ with $i \in \{0, 1, \dots, k\}$ such that $u_i \geq u_{i+1} + 2\epsilon$. For any $u' \in \mathbb{R}^k$ such that $d_\infty(u, u') < \epsilon$, we claim that $\alpha_i(u') \neq 0$, and therefore $\mathbb{1}_{\pi^*,i} \in \hat{\Psi}(u)$.

Assume there exists a $u' \in \mathbb{R}^k$ such that $d_\infty(u, u') < \epsilon$ for some $i \in \{0, 1, \dots, k\}$. Given that $d_\infty(u, u') < \epsilon$, $u'_j \in (u_j - \epsilon, u_j + \epsilon)\, \forall j \in \{0, \dots, k\}$: namely, for $j = i$ and $i + 1$. However, since $u_i - u_{i+1} \geq 2\epsilon$, $(u_i - \epsilon, u_i + \epsilon) \cap (u_{i+1} - \epsilon, u_{i+1} + \epsilon) = \varnothing$. Therefore, $\alpha_i(u') = u'_i - u'_{i+1} > 0$. By Lemma 26, we then have $\mathbb{1}_{\pi^*,i} \in V^u_{\pi^*,y}$, which is the smallest set V such that $d_\infty(\operatorname{conv} V, u) < \epsilon$, and is therefore in the intersection of all such sets; this intersection yields $\hat{\Psi}(u)$. Thus, we have $\hat{\Psi}(u) = \{\mathbb{1}_{\pi,i} \odot \operatorname{sign}^*(u) \mid i \in \{0, 1, \dots, k\},\ u_i \geq u_{i+1} + 2\epsilon\}$. ■

Lemma 21 *$\hat{\Psi}$ is nonempty pointwise if and only if $\epsilon \in (0, \frac{1}{2k}]$.*

Proof By Lemma 27, it suffices to show the statement for $u \in \mathbb{R}^k_+$. We will show the contrapositive in both directions: there exists $u \in \mathbb{R}^k_+$ such that $\hat{\Psi}(u) = \varnothing$ if and only if $\epsilon > \frac{1}{2k}$.

For any $u \in \mathbb{R}^k_+$, define $u_{k+1} = -\epsilon$ and $u_0 = 1 + \epsilon$ as in Proposition 20. From the characterization in Proposition 20 (eq. (16)), we have $\hat{\Psi}(u) = \varnothing$ if and only if

$$u_i - u_{i+1} < 2\epsilon \text{ for all } i \in \{0, 1, \dots, k\}\ . \tag{20}$$

We may also write

$$1 + \epsilon = u_0 = u_{k+1} + \sum_{i=0}^{k} (u_i - u_{i+1}) = \sum_{i=0}^{k} (u_i - u_{i+1}) - \epsilon\ . \tag{21}$$

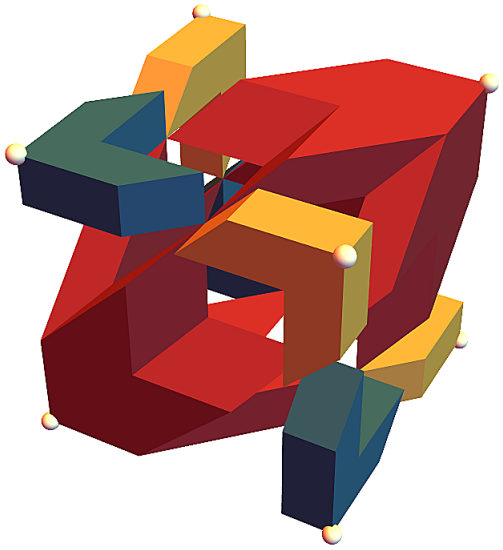

Figure 2: $\hat{\Psi}(u)$ for $u \in \mathbb{R}^3_+$ and $\epsilon = \frac{1}{6}$. Each colored region connected to a particular node corresponds to a $v \in \{0,1\}^3 \subseteq \mathcal{V}$ and at a point u, a calibrated link must link to one of the v in the region.

If there exists $u \in \mathbb{R}^k_+$ with $\hat{\Psi}(u) = \varnothing$, then eq. (20) and (21) together imply $1 + 2\epsilon = \sum_{i=0}^{k}(u_i - u_{i+1}) < (k+1)(2\epsilon)$, giving $\epsilon > \frac{1}{2k}$. For the converse, if $\epsilon > \frac{1}{2k}$, take $u \in \mathbb{R}^k_+$ given by $u_i = \frac{2i-1}{2k}$. Then $u_0 - u_1 = 1 + \epsilon - (1 - \frac{1}{2k}) < 2\epsilon$ and $u_k - u_{k+1} = \frac{1}{2k} + \epsilon < 2\epsilon$, and for $i \in \{1, \ldots, k-1\}$, we have $u_{i+1} - u_i = \frac{1}{k} < 2\epsilon$, giving eq. (20). ■