Consistent Polyhedral Surrogates for Top-k Classification and Variants

Jessie Finocchiaro [1] **Rafael Frongillo** [1] **Emma Goodwill** [1] **Anish Thilagar** [1]

Abstract

Top-k classification is a generalization of multiclass classification used widely in information retrieval, image classification, and other extreme classification settings. Several hinge-like (piecewise-linear) surrogates have been proposed for the problem, yet all are either non-convex or inconsistent. For the proposed hinge-like surrogates that are convex (i.e., polyhedral), we apply the recent embedding framework of Finocchiaro et al. (2019; 2022) to determine the prediction problem for which the surrogate is consistent. These problems can all be interpreted as variants of top-k classification, which may be better aligned with some applications. We leverage this analysis to derive constraints on the conditional label distributions under which these proposed surrogates become consistent for top-k. It has been further suggested that every convex hinge-like surrogate must be inconsistent for top-k. Yet, we use the same embedding framework to give the first consistent polyhedral surrogate for this problem.

1. Introduction

Top-k classification is commonly used in image recognition (Akata et al., 2013; Karpathy et al., 2014; Russakovsky et al., 2015) and action analysis (Furnari et al., 2018), search querying (Ailon and Mohri, 2008; Reddi et al., 2019), and recommender systems more broadly (Adomavicius and Zhang, 2016; Billsus et al., 1998; Deshpande and Karypis, 2004). For example, in information retrieval, a page of k results may be displayed out of $n \gg k$ total webpages available, with success indicated by a user clicking one of these k. This scenario can be captured by the *top-k loss*: given a set S of labels, $|S| = k$, and the true label y, assign loss 1 if $y \notin S$, and 0 otherwise. As top-k loss is discrete, it is typically computationally hard to optimize. Therefore, top-k learning algorithms typically employ a *surrogate loss*. Common desiderata for surrogate losses are that they be *convex*, and thus easier to optimize, and that they be *statistically consistent*, meaning they solve the original problem (here: top-k) when given enough data. Another consideration is whether the surrogate is *smooth (e.g. differentiable) or piecewise-linear ("hinge-like")*. This consideration is related to whether the surrogate will implicitly learn the underlying conditional label distribution, which generally is a harder learning problem than the original; for example, the entire label distribution contains more information than the set of k most likely labels. Typically, smooth surrogates, such as cross-entropy, implicitly learn the entire label distribution.[1] Conventional wisdom has been that piecewise-linear surrogates are more "efficient" in the sense that they learn only what is relevant for the original problem. Moreover, piecewise-linear and convex surrogates give rise to linear surrogate regret bounds, whereas most smooth surrogates do not (Frongillo and Waggoner, 2021).

Combining the above desiderata, we would like a surrogate which is both *polyhedral* (convex and piecewise-linear) and consistent for top-k classification. Unfortunately, while many piecewise-linear surrogates have been proposed for top-k, they are all either non-convex or inconsistent (Lapin et al., 2015; 2016; 2018; Reddi et al., 2019; Yang and Koyejo, 2020). Moreover, the results and writing of both Lapin et al. (2016, pg.6) and Yang and Koyejo (2020, pg.1) suggest that perhaps no such surrogate exists for top-k.

We resolve this open question by presenting the first consistent polyhedral surrogate for top-k classification (§ 4). Our proof uses embedding framework of Finocchiaro et al. (2019; 2022). We also use the embedding framework to analyze three previous polyhedral surrogates in the literature which are inconsistent for top-k (§ 3). For each we show (a) what discrete prediction problem the surrogate is actually solving, in all cases a natural variant of top-k, and (b) a constraint on the conditional label distributions such that the surrogate becomes consistent for top-k. Finally, we evaluate the performance of our surrogate compared to these

[1]University of Colorado Boulder Department of Computer Science, Boulder, CO, USA. Correspondence to: Jessie Finocchiaro <Jessica.Finocchiaro@colorado.edu>, Anish Thilagar <anish@colorado.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Concretely, consider any surrogate whose Bayes risk is strictly concave, which is the case for most smooth surrogates. For each surrogate prediction u, it can minimize expected loss for at most one conditional label distribution p; otherwise the Bayes risk would be flat on the line segment between two such distributions. Thus, one can infer p from the u returned by the model.

previous surrogates (§ 5).

2. Setting

We consider predictions in a discrete set $\mathcal{R}$ over a finite set of labels $\mathcal{Y} = \{1, \dots, n\}$, and conditional label distributions $\Delta_\mathcal{Y}$. In top-k classification, predictions take the form of size-k subsets of labels, $\mathcal{R} = \mathcal{R}_k := \{S \subseteq \mathcal{Y} \mid |S| = k\}$. Top-$k$ loss $\ell_k : \mathcal{R}_k \times \mathcal{Y} \to \mathbb{R}_+$ simply tests whether the actual label lies in the set,

$$\ell_k(S, y) = \mathbb{1}\{y \notin S\} \,, \tag{1}$$

where $\mathbb{1}\{E\}$ is 1 if event E is true, and 0 otherwise. In reasoning about top-k and variants, it is often useful to denote $u_{[i]}$ to be the i^{th} largest element of the vector $u \in \mathbb{R}^n$. Moreover, the set of possible top-k indices $T_k : \mathbb{R}^n \to 2^{\mathcal{R}_k}$ is given by $T_k : u \mapsto \arg\max_{S \in \mathcal{R}_k} \langle \mathbb{1}_S, u \rangle$. Observe $|T_k(u)| > 1$ if and only if $u_{[k]} = u_{[k+1]}$. Additionally, we denote the sum of these top k elements by $\sigma_k(u) = \max_{S \in \mathcal{R}_k} \langle \mathbb{1}_S, u \rangle$.

2.1. Consistency, Property Elicitation, and Calibration

Discrete losses such as ℓ_k are hard to optimize directly, so a consistent surrogate is sought instead with better optimization guarantees. In essence, a surrogate and link are *consistent* with respect to a discrete target loss if approaching the optimal surrogate loss implies approaching the optimal target loss when the link function is applied to the surrogate predictions. We will phrase consistency in terms of the equivalent notion of calibration (Bartlett and Wegkamp, 2008; Ramaswamy and Agarwal, 2016; Steinwart and Christmann, 2008; Tewari and Bartlett, 2007).

Before defining calibration, we first introduce properties, which encode the optimal predictions for a loss as a function of the conditional label distribution. Here $\mathcal{P} \subseteq \Delta_\mathcal{Y}$.

Definition 2.1. A *property* is a function $\Gamma : \mathcal{P} \to 2^\mathcal{R} \setminus \{\emptyset\}$, which we more succinctly denote $\Gamma : \mathcal{P} \rightrightarrows \mathcal{R}$. A loss $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ *elicits* a property $\Gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ if

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

A loss L is *minimizable* if $\mathbb{E}_{Y \sim p} L(\cdot, Y)$ attains its infimum for all $p \in \mathcal{P}$. Every minimizable loss L elicits a unique property, which we denote $\text{prop}[L]$.

As an example, the property elicited by top-k loss is $\gamma_k = \text{prop}[\ell_k]$, which is given by

$$\begin{aligned} \gamma_k(p) &= \arg\min_{S \in \mathcal{R}_k} \langle p, \ell_k(S, \cdot) \rangle \\ &= \arg\min_{S \in \mathcal{R}_k} \sum_{i \notin S} p_i \\ &= T_k|_{\Delta_\mathcal{Y}}(p) \,. \end{aligned} \tag{2}$$

Definition 2.2. 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 over $\mathcal{P} \subseteq \Delta_\mathcal{Y}$ if for all $p \in \mathcal{P}$,

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

We simply say L is calibrated with respect to ℓ if there exists a link ψ such that (L, ψ) is calibrated with respect to ℓ.

One can think of $\mathcal{P}$ as the set of possible conditional label distributions conditioned on some feature vector. We consider $\mathcal{P} = \Delta_\mathcal{Y}$ unless otherwise specified.

2.2. Embedding Framework for Polyhedral Surrogates

We rely heavily on the embedding framework of Finocchiaro et al. (2019; 2022), which gives tools to analyze and construct consistent polyhedral surrogates. An embedding maps the finite set of target predictions to a *representative set* of surrogate predictions.

Definition 2.3 (Representative set)**.** A set $\mathcal{S} \subseteq \mathcal{R}$ is *representative* for a property $\Gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ if, for all $p \in \mathcal{P}$, we have $\Gamma(p) \cap \mathcal{S} \neq \emptyset$. We say $\mathcal{S}$ is representative for a loss L if it is representative for the property $\text{prop}[L]$.

Definition 2.4 (Embedding)**.** A loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ *embeds* a discrete 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

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

In other words, a surrogate embeds a discrete target loss if the loss values match at the embedded points, and moreover, a target prediction is optimal exactly when its embedded prediction is optimal for the surrogate.

Embeddings are closely tied to polyhedral surrogates; in particular, every polyhedral surrogate embeds some discrete loss (Finocchiaro et al., 2022). We will primarily use the following results. Throughout, for a loss $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ and set $\mathcal{S} \subseteq \mathcal{R}$, we denote by $L|_\mathcal{S}$ the loss on $\mathcal{S} \times \mathcal{Y}$ given by $L|_\mathcal{S}(u, y) = L(u, y)$, i.e., the restriction of L to $\mathcal{S}$.

Theorem 2.5 (Finocchiaro et al. (2022))**.**

1. *Every polyhedral loss L has a finite representative set.*
2. *If $\mathcal{S}$ is a finite representative set for L, then L embeds the discrete loss $L|_\mathcal{S}$.*
3. *If L embeds ℓ, then there exists a link ψ such that (L, ψ) is calibrated with respect to ℓ.*

These correspond to Lemma 2, Proposition 1, and Theorem 2 in that work, respectively. The authors also provide a construction for the calibrated link ψ, as well as a construction for a calibrated polyhedral surrogate given any discrete loss; we discuss both of these additional tools in § 4.

3. Previous Polyhedral Surrogates

Lapin et al. (2015) proposes a nonconvex surrogate for top-k prediction, as well as convex upper bounds on this surrogate in (Lapin et al., 2016), denoted $L^{(2)}$ and $L^{(3)}$ here to parallel their notation. Yang and Koyejo (2020) show that $L^{(2)}$ and $L^{(3)}$ are inconsistent for ℓ_k classification, and introduce another inconsistent surrogate, which we denote $L^{(4)}$.

All three losses $L^{(2)}$, $L^{(3)}$, and $L^{(4)}$ are polyhedral; as such Theorem 2.5 implies that they all embed *some* discrete loss. It is not immediately clear, however, what exactly these discrete losses are for each surrogate. In this section, we derive a target loss that each surrogate embeds, which in each case is an interesting variant of the original top-k problem.

Deriving the loss embedded by an inconsistent surrogate also allows one to understand when it would be consistent for the intended target. In particular, by looking at the geometry of the property elicited by the surrogate, we can derive a constraint on the set of conditional label distributions under which it becomes consistent for top-k. One can view these results as a refinement of inconsistency results; for example, Yang and Koyejo (2020, Proposition 4.2) characterizes the set of distributions such that the surrogate report $u = \vec{0} \in \mathbb{R}^n$ is optimal, a subset of the set of distributions we eliminate.

In summary, then, we strive in this section to answer two questions about $L^{(2)}$, $L^{(3)}$, and $L^{(4)}$: (i) What discrete loss does the surrogate embed? (ii) On which conditional label distributions is the surrogate actually consistent for top-k?

To answer (i), we find a finite representative set and apply Theorem 2.5, which shows that restricting to that set gives an embedding. To find this set, we first observe that these surrogates are all invariant in the $\mathbb{1}$ direction, meaning $L(u, y) = L(u + \alpha\mathbb{1}, y)$ for all $\alpha \in \mathbb{R}$. Furthermore, we can fix the lowest $n - k - 1$ elements of U to be the same as $u_{[k+1]}$, as this can only improve the loss on any outcome. We can therefore restrict our attention to the set of reports

$$U = \{u \in \mathbb{R}^n_+ \mid u_{[k+1]} = 0 = u_{[n]}\}, \tag{4}$$

which is representative, although infinite. In some cases, we further restrict U to a region where the positive part operator $(\cdot)_+$ can be removed. In each case, we partition the resulting set into polytope regions over which the surrogate is affine; in other words, we find the pieces for which the loss is piecewise linear. By the theory of polyhedral functions, for each conditional label distribution, at least one vertex of one of these regions must be a minimizer of the expected loss. The union of all such vertices therefore yields a finite representative set. As a final step, in each case we reparameterize this set of vertices with a bijection to a more natural prediction set, which more transparently reveals a variant of the top-k problem. Applying such a bijection preserves the embedding by Definition 2.4.

To answer (ii), we observe that in all cases, inconsistency is driven by surrogate reports for which the set of top-k elements is ambiguous, thus forcing the link to break a tie. Specifically, for reports $u \in \mathbb{R}^n$ with $u_{[k]} = u_{[k+1]}$, we have multiple options for $T_k(u)$, yet ψ_k must select one. Let $U_{\text{ambig}} = \{u \in \mathbb{R}^n \mid u_{[k]} = u_{[k+1]}\}$ be the set of these ambiguous surrogate reports. Whenever a report $u \in U_{\text{ambig}}$ is optimal for a conditional label distribution p for which $T_k(p)$ is *not* ambiguous, i.e. $p_{[k]} > p_{[k+1]}$, we will have inconsistency. Therefore, $(L^{(i)}, \psi_k)$ is consistent with respect to ℓ_k on the set $\mathcal{P}^{(i)} := \{p \in \Delta_\mathcal{Y} \mid \text{prop}[L^{(i)}](p) \cap U_{\text{ambig}} = \emptyset\}$ of conditional label distributions for which there is no ambiguous optimal report.

3.1. Analysis of $L^{(2)}$

The surrogate $L^{(2)}$ proposed by Lapin et al. (2016) is given by

$$L^{(2)}(u, y) = \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u - e_y)_{[i]}\right)_+. \tag{5}$$

We will derive a discrete loss $\ell^{(2)}$ in eq. (6) that $L^{(2)}$ embeds, and then use it to characterize the set of distributions $\mathcal{P}^{(2)}$ on which $(L^{(2)}, \psi_k)$ is consistent with respect to ℓ_k. See § A for all omitted details.

By our strategy outlined above, we begin with the set U (eq. (4)), which is representative for $L^{(2)}$. We then construct the bounded region $U^{(2)}_+ \subset U$ in which the positive part operator in eq. (5) is not activated, and show $U^{(2)}_+$ is representative. We next partition $U^{(2)}_+$ into polytope regions over which $L^{(2)}$ is affine. When restricting to $U^{(2)}_+$, the only way $L^{(2)}(\cdot, y)$ fails to be affine is in the top-k elements of a prediction $(u - e_y)$ changing. Observe that, up to tie-breaking, the top k elements of $(u - e_y)$ are the same as the top k elements of u if and only if $u_y \geq 1 = 1 + u_{[k+1]}$. $L^{(2)}$ is therefore affine on regions where $\text{sign}(u_i - 1)$ is constant for all $i \in \{1, \ldots, k\}$. Further examining these affine regions reveals that their vertices are the points $u \in \mathbb{R}^n_+$ such that $u_i \in \{0, 1, c(u)\}$ for a particular value $c(u) > 1$ that depends on how many entries of u are nonzero and how many are strictly greater than 1.

Taking the union of these vertices, we arrive at a finite representative set for $L^{(2)}$. Theorem 2.5 now states that $L^{(2)}$ embeds $L^{(2)}$ restricted to this vertex set. To state this

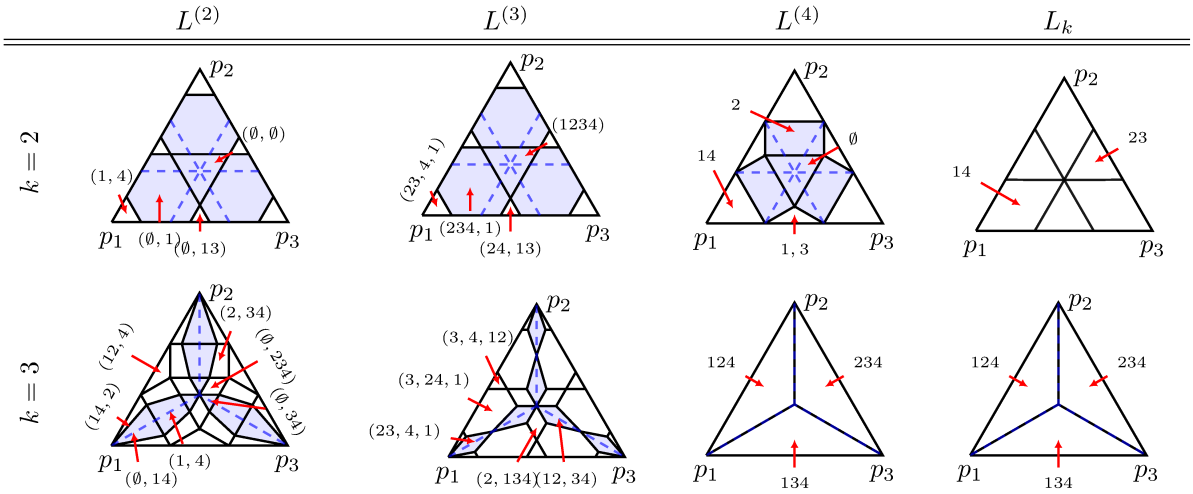

Table 1: Visualizations of the minimizers of the losses (embedded by) $L^{(2)}$, $L^{(3)}$, $L^{(4)}$, and L_k with $n = 4$ and $k \in \{2, 3\}$, fixing $p_4 = 1/4$. The dashed blue lines give sets of distributions p corresponding to the same report u such that $u \in \text{prop}[L_k](p)$. As we must link reports deterministically, we want each of the bold, black cells to be fully contained in a cell from the blue dashed cells. Blue regions cross the dashed blue lines and suggest where deciding how to construct a link ψ is ambiguous, as $|T_k(u)| > 1$. White regions are therefore where the surrogate and any top-k link are consistent, e.g., $\mathcal{P}^{(i)}$. On the right, L_k shows our proposed surrogate that is consistent for top-k classification, demonstrated by no blue regions. Each of the cells corresponds to a subset of distributions where exactly k reports are optimal.

discrete loss more intuitively, we simply reparameterize these vertices, letting M be the set of entries equal to 1, and H the set strictly greater than 1. Letting $\mathcal{R}^{(2)}$ be the set of valid pairs (H, M), namely disjoint and with $|H \cup M| \leq k$, we arrive at the following discrete loss $\ell^{(2)} : \mathcal{R}^{(2)} \times \mathcal{Y} \to \mathbb{R}$ embedded by $L^{(2)}$.

$$\ell^{(2)}((H, M), y) = \begin{cases} 0 & y \in H \\ \frac{|H|+|M|-1}{k-|H|} & y \in M \\ \frac{|H|+|M|-1}{k-|H|} + \frac{k+1}{k} & \text{otherwise} \end{cases} \quad (6)$$

One can regard H as the "high labels", with high likelihood of being the ground truth label, and M the "medium labels", with some likelihood. One therefore attains loss 0 if they were highly confident in the ground truth label, and accumulate a loss that grows in the size of H and M otherwise.

By our observations above, consistency with respect to top-k is achieved whenever the optimal report is some (H, M) with $|H \cup M| = k$. This condition can be written as follows, where $h^*(p) = \max\{i \in \{0, \ldots, k\} \mid p_{[i]} > \frac{1-\sigma_{i-1}(p)}{k-(i-1)}\}$.

Corollary 3.1. *Define*

$$\mathcal{P}^{(2)} := \left\{ p \in \Delta_\mathcal{Y} \mid p_{[k]} > \frac{(1 - \sigma_{h^*(p)}(p))}{(k+1)(k-h^*(p))} \right\}. \quad (7)$$

$L^{(2)}$ is consistent with respect to ℓ_k on $\mathcal{P}^{(2)}$.

3.2. Analysis of $L^{(3)}$

Lapin et al. (2016) give two convex upper bounds on the proposed top-k surrogate from (Lapin et al., 2015): $L^{(2)}$ studied in § 3.1, and $L^{(3)}$, defined as follows.

$$L^{(3)}(u, y) = \frac{1}{k} \sum_{i=1}^{k} \left[1 - u_y + (u - e_y)_{[i]}\right]_+ \quad (8)$$

While similar to $L^{(2)}$, the placement of the positive part operator changes the analysis of the surrogate significantly. See § B for all omitted details.

As above, it suffices to identify sources of non-affineness on U (eq. (4)) to construct a finite representative set for $L^{(3)}$. Non-affineness of $L^{(3)}$ is introduced by the positive part operator and the ordering of the top-k elements of a prediction $u \in U$. Unlike $L^{(2)}$, the positive part operator is applied to each term of the summand, so we cannot immediately ignore this operator by restricting to a bounded representative region. Instead, let us simultaneously fix (1) a set $S \in \mathcal{R}_k$ to be indices of the top-k elements of u, and (2) sets $\vec{V} = \{V_y \subseteq S \setminus \{y\} \mid y \in \mathcal{Y}\}$ corresponding to induces when the positive part operator is not activated for $L^{(3)}(u, y)$. For any such $S, \vec{V}$, therefore, we define the region $A^{S,\vec{V}}$ to be all points $u \in U$ with (1) $S \in T_k(u)$ and (2) for all $y \in \mathcal{Y}$, we have $u_i + 1 \geq u_y$ for all $i \in V_y$, and $u_i + 1 \leq u_y$ for all $i \notin V_y$. By the above reasoning, $L^{(3)}$ is affine on the set $A^{S,\vec{V}}$ for each choice of $S, \vec{V}$.

The union of the vertices of each $A^{S,\vec{V}}$ region is therefore

a finite representative set, and $L^{(3)}$ embeds $L^{(3)}$ restricted to these vertices. Upon inspection of the geometry of the $A^{S,\vec{V}}$ regions, we show that the vertices of each are in fact a subset of $\mathbb{Z}_k^n$. A more intuitive form for this discrete loss can therefore be expressed in terms of ordered partitions, where index i is in the j^{th} partition Q_j when $u_i = j$. Formally, we reparameterize the vertices as ordered partitions $Q \in \mathcal{R}^{(3)}$, where

$$\mathcal{R}^{(3)} = \{Q=(Q_0,\dots,Q_s) \mid s\leq k, Q_i\cap Q_j = \emptyset\,\forall i\neq j, \\ |Q_1,\cup\dots\cup Q_s| \leq k, Q_i\neq\emptyset\,\forall i\} \,.$$

We now have that $L^{(3)}$ embeds $\ell^{(3)} : \mathcal{R}^{(3)} \times \mathcal{Y} \to \mathbb{R}$, given by

$$\ell^{(3)}(Q,y) = \begin{cases} \frac{1}{k}\left(|Q_j| - 1 + \sum\limits_{i>j} |Q_i|(i-j+1)\right) & j>0 \\ \frac{1}{k}\sum\limits_{i=1}^{s} |Q_i|(i+1) & j=0 \end{cases}$$

where $y \in Q_j$. For intuition, $\ell^{(3)}$ allows for predictions with more granularity than $\ell^{(2)}$, where the higher index i of the partition Q_i is, the more confident one is in outcomes in Q_i. The punishment for error again grows in the number of indices one reports high confidence in, as well as the number of partitions.

In order to characterize the regions where $(L^{(3)}, \psi_k)$ is consistent with respect to ℓ_k, we can study where $\ell^{(3)}$ can be unambiguously linked to ℓ_k. In particular, one can do so for any $p \in \Delta_\mathcal{Y}$ such that $|Q_0| = n - k$ for $Q \in \text{prop}[\ell^{(3)}](p)$.

Corollary 3.2. *$L^{(3)}$ is consistent with respect to ℓ_k on $\mathcal{P}^{(3)} = \{p \in \Delta_\mathcal{Y} \mid p_{[k+1]} > \frac{1}{k+1} \wedge \frac{\sum_{i=k+1}^{n} p_{[i]}}{k-1} \geq p_{[k]}\}$.*

3.3. Analysis of $L^{(4)}$

Observing that $L^{(2)}$ and $L^{(3)}$ are inconsistent with respect to ℓ_k, Yang and Koyejo (2020) propose $L^{(4)}$ as in eq (9), changing the summation from elements of $(u - e_y)$ to elements of $u_{\setminus y} \in \mathbb{R}^{n-1}$: the elements of u excluding u_y. See § C for all omitted details.

$$L^{(4)}(u,y) = \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]}\right)_+ \tag{9}$$

Again following the strategy outlined above, we begin with the set U, which is representative for $L^{(4)}$. Here we also further restrict to the set of points $U_+^{(4)} \subseteq U$ yielding a nonnegative argument to the positive part operator, and show that $U_+^{(4)}$ is also representative for $L^{(4)}$. Within $U_+^{(4)}$, we observe that the only way $L^{(4)}(\cdot, y)$ fails to be affine is when the top k elements of $u_{\setminus y}$ change. Since all elements of U have at most k nonzero entries already, it therefore suffices to select a subset T of nonzero indices. For any $T \subseteq [n]$ with $|T| \leq k$, let us therefore define the set A^T to be all points $u \in \mathbb{R}^n$ such that $0 \leq u_i \leq 1 + \frac{1}{k}\sum_{j\in T, j\neq i} u_j$ for $i \in T$, and $u_i = 0$ for $i \notin T$. For any $p \in \Delta_\mathcal{Y}$, the function $u \mapsto \left\langle L^{(4)}(u,\cdot), p\right\rangle$ is affine on each region A^T, and moreover, they partition the representative set $U_+^{(4)}$.

Taking the union of vertices of each A^T set, we arrive at a finite representative set for $L^{(4)}$. Carefully examining the geometry of the A^T sets, one sees that these vertices are the points $u \in \mathbb{R}^n$ such that each element is either 0 or $\frac{k}{k+1-|T|}$. Therefore, the finite representative set for $L^{(4)}$ can be reparameterized as $\mathcal{R}^{(4)} = \{T \subseteq [n] \mid |T| \leq k\}$, and thus $L^{(4)}$ embeds $\ell^{(4)} : \mathcal{R}^{(4)} \times \mathcal{Y} \to \mathbb{R}$ given by

$$\ell^{(4)}(T,y) = \begin{cases} 0 & y \in T \\ \frac{k+1}{k+1-|T|} & y \notin T \end{cases} .$$

Intuitively, $\ell^{(4)}$ is a variant of top-k where one may report any set of labels of size $m \leq k$, and the stakes for being incorrect increase in m. Therefore, the loss incentivizes one to report smaller sets only when sufficiently confident.

Following this intuition, consistency therefore arises whenever the conditional label distribution does not lead to such high confidence that the optimal report is a set of size $m < k$. We characterize such distributions as follows.

Corollary 3.3. *$L^{(4)}$ is consistent with respect to ℓ_k on $\mathcal{P}^{(4)} := \{p \in \Delta_\mathcal{Y} \mid p_{[k]} > 1 - \sigma_k(p)\}$.*

4. A New Consistent Surrogate

Yang and Koyejo (2020) show that the polyhedral surrogates analyzed in § 3 are not consistent for top-k. They further suggest that perhaps *no* polyhedral surrogate can be consistent. On the other hand, the embedding framework of Finocchiaro et al. (2019; 2022) shows that every discrete loss has a consistent polyhedral surrogate. As their result is constructive, we apply it to the top-k loss ℓ_k, giving the first consistent polyhedral surrogate, L_k, for the problem (§ 4.1). The embedding framework relies on constructing a link from scratch, rather than using a pre-specified link function. As such, in principle their surrogate construction could yield a surrogate which is not consistent when paired with ψ_k, but only with a different link entirely. Interestingly, we further show that in particular (L_k, ψ_k) is consistent with respect to ℓ_k (§ 4.2).

4.1. Formulating L_k

To show that every discrete loss is embedded by a consistent polyhedral surrogate, Finocchiaro et al. give the following construction. Their construction echoes similar constructions in the literature (cf. Asif et al. (2015), Farnia and Tse (2016), Fathony et al. (2016), Duchi et al. (2018).) Recall

that the *Bayes risk* of a loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ is the function $\underline{\ell} : \Delta_\mathcal{Y} \to \mathbb{R}$, $\underline{\ell} : p \mapsto \min_{r \in \mathcal{R}} \langle p, \ell(r, \cdot)\rangle$.

Theorem 4.1 (Finocchiaro et al. (2022, Theorem 4))**.** *Any discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ is embedded by the consistent surrogate $L(u, y) = (-\underline{\ell})^*(u) - u_y$ where $(\cdot)^*$ denotes the convex conjugate.*

The Bayes risk of ℓ_k is

$$\underline{\ell_k}(p) = \inf_{S \in \mathcal{R}_k} \langle p, \ell_k(S, \cdot)\rangle = 1 - \sigma_k(p) \; .$$

By Theorem 4.1, the following loss function L_k therefore embeds ℓ_k, with consistency (for some link function) following from Theorem 2.5.

$$\begin{aligned} L_k(u, y) &= (-\underline{\ell_k})^*(u) - u_y \\ &= \sup_{p \in \Delta_\mathcal{Y}} \left(\langle p, u\rangle + \underline{\ell_k}(p)\right) - u_y \\ &= \sup_{p \in \Delta_\mathcal{Y}} \left(\langle p, u\rangle + 1 - \sigma_k(p)\right) - u_y \; . \end{aligned} \tag{10}$$

Choosing p to be uniform on the m largest indices of u (which we justify in § D.2), this expression simplifies to

$$= \max_{1 \le m \le n} \left\{ \tfrac{\sigma_m(u)}{m} + \left(1 - \tfrac{k}{m}\right)_+ \right\} - u_y \; . \tag{11}$$

Since $\frac{\sigma_m(u)}{m}$ is non-increasing in m, and $1 - \frac{k}{m} \le 0$ for $0 < m \le k$, the $m = 1$ case will dominate the $1 < m \le k$ cases. Therefore, we can further simplify the loss,

$$= \max\left\{ u_{[1]}, \max_{k < m \le n} \left\{ \tfrac{\sigma_m(u)}{m} + 1 - \tfrac{k}{m} \right\} \right\} - u_y \; .$$

In this form, it is clear to see that the surrogate is piecewise linear, as a maximum of affine functions (recall that σ_m can itself be written as a maximum).

4.2. The Argmax Link is Calibrated

From Theorem 2.5, there exists some link function $\psi : \mathbb{R}^n \to \mathcal{R}_k$ mapping the report space of L_k back to the that of ℓ_k, such that (L_k, ψ) is consistent with respect to ℓ_k. It remains to actually find this link ψ. In fact, we will show that one can take $\psi = \psi_k$, the canonical argmax link.

Recall that consistency is characterized by calibration (Definition 2.2), which says that linking to a non-ℓ_k-optimal report should be strictly L_k-suboptimal. To show that ψ_k is calibrated, we in turn use another equivalent condition, that ψ_k be ϵ-separated (Finocchiaro et al., 2022, Definition 8) with respect to ℓ_k and L_k. Recall that all minimizable losses elicit a property (Definition 2.1), which is just a map from distributions to all optimal reports under that loss.

Definition 4.2. Given a discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ and surrogate $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$, let $\Gamma = \text{prop}[L]$ and $\gamma = \text{prop}[\ell]$ be their respective properties. The link $\psi : \mathbb{R}^d \to \mathcal{R}$ is ϵ-separated with respect to (L, ℓ) if for all $p \in \Delta_\mathcal{Y}$, $u \in \Gamma(p)$, and $u' \in \mathbb{R}^d$ such that $\psi(u') \notin \gamma(p)$, we have $\|u - u'\|_\infty \ge \epsilon$.

Calibration and ϵ-separation are equivalent for polyhedral surrogates (Finocchiaro et al., 2022, Theorem 5).

To show ϵ-separation, we first must characterize the properties of ℓ_k and L_k. Eq. (2) gives us $\text{prop}[\ell_k] = \gamma_k$. Let $\Gamma_k = \text{prop}[L_k]$.

Recall that the report space of ℓ_k is $\mathcal{R}_k = \{S \subseteq \mathcal{Y} \mid |S| = k\}$. Let $\mathcal{T} = \{\mathbb{1}_S \mid S \in \mathcal{R}_k\}$ be the set of indicators for the elements of $\mathcal{R}_k$. Then, $\tau_k(u) = \arg\max_{t \in \mathcal{T}} \langle t, u\rangle$ is the set of possible indicators of the top k elements of u. Note that $|\tau_k(u)| > 1$ if and only if $u_{[k]} = u_{[k+1]}$.

Lemma 4.3. *Let* cone *denote the convex cone. Then,*

$$\Gamma_k(p) = \text{hull}(\tau_k(p)) - \text{cone}\{\mathbb{1}_i \mid p_i = 0\} + \bigcup_{\alpha \in \mathbb{R}} \{\alpha \mathbb{1}\} \; .$$

The proof, deferred to § D.1, relies on the connection between Γ_k and the subgradients of $-\underline{\ell_k}$. With this characterization of γ_k and Γ_k, we can prove that ψ_k is calibrated.

Theorem 4.4. *(L_k, ψ_k) is calibrated with respect to ℓ_k.*

Proof. First, we show ψ_k is ϵ-separated with respect to Γ_k and γ_k. Let $\epsilon = \frac{1}{2n}$. Fix any $p \in \Delta_\mathcal{Y}$, and choose any $u \in \Gamma_k(p)$. Choose α such that $u - \alpha\mathbb{1} \in \text{hull}(\tau_k(p)) - \text{cone}\{\mathbb{1}_i \mid p_i = 0\}$. We need to show for every u' with $\psi_k(u') \notin \gamma_k(p)$, $\|u - u'\|_\infty \ge \frac{1}{2n}$.

Case 1: $p_{[k]} > 0$. Since $u \in \Gamma_k(p)$, Lemma 4.3 implies every element of u is at most $1 + \alpha$, so we have $\sigma_{k-1}(u) \le (k-1)(1+\alpha)$. Let $S = \text{support}(\gamma_k(p))$, the set of indices i with $p_{[i]} \ge p_{[k]} > 0$. Lemma 4.3 also implies $\sum_{i \in S} u_i = k + \alpha|S|$. Since $u_{[k]}$ is the largest element of S that is not in the top $k-1$ elements of u, we have

$$\begin{aligned} u_{[k]} &\ge \frac{\left(\sum_{i \in S} u_i\right) - \sigma_{k-1}(u)}{|S| - (k-1)} \\ &= \frac{k + \alpha|S| - (k-1)(1+\alpha)}{|S| - (k-1)} \\ &= \frac{1}{|S| - (k-1)} + \alpha \\ &> \frac{1}{n} + \alpha \; . \end{aligned}$$

Now, pick any u' such that $\psi_k(u') \notin \gamma_k(p)$. Since $\psi_k(u')$ is some top-k index set of u', and by eq. (2) $\gamma_k(p)$ is every possible top-k index set of p, then there must be some index

$j \in \psi_k(u')$ such that $p_j < p_{[k]}$. Then by Lemma 4.3, $u_j \leq \alpha$.

We proceed by contradiction. Assume $\|u - u'\|_\infty < \frac{1}{2n}$. Therefore for every index i, we have $|u_i - u'_i| < \frac{1}{2n}$. Since $u_{[k]} > \alpha + \frac{1}{n}$, for every $i \in \psi_k(u)$, we must have $u'_i > u_i - \frac{1}{2n} \geq u_{[k]} - \frac{1}{2n} \geq \alpha + \frac{1}{2n}$. Since $u_j \leq \alpha$, we also must have $u'_j < \alpha + \frac{1}{2n}$. However, that means there are $|\psi_k(u)| = k$ elements of u' which are larger than u'_j, so $j \notin \psi_k(u')$, a contradiction. Therefore, $\|u - u'\|_\infty \geq \frac{1}{2n}$.

Case 2: $p_{[k]} = 0$. Let $S = \{i | p_{[i]} > 0\}$. Therefore, for all $i \in S$, $u_i = 1 + \alpha$. Since $p_{[k]} = 0$, S must be contained by element of $\gamma_k(p)$. Choose any u' such that $\psi_k(u') \notin \gamma_k(p)$. By eq (2) every element of $\gamma_k(p)$ contains S, so there must be some index $j \in S$ such that $u'_j \leq u'_{[k]}$.

We again proceed by contradiction, and assume $\|u - u'\|_\infty < \frac{1}{2n}$. Since $u_j = 1 + \alpha$, we must have $u'_j > 1 + \alpha - \frac{1}{2n}$. However, since $u'_j \leq u'_{[k]}$, there must be $k - (|S| - 1)$ elements of u' that are greater than u'_j but not in S. Formally, choose any set $T \subseteq ([n] \setminus S) \cup \psi_k(u')$ with $|T| = k - (|S| - 1)$. For every $i \in T$ we have $u'_i > u'_j$, so

$$\begin{aligned}
\sum_{i \in T} u_i &\geq \left(1 + \alpha - \frac{1}{2n}\right) |T| \\
&= \left(1 + \alpha - \frac{1}{2n}\right) (k - |S| + 1) \\
&= (1 + \alpha)(k - |S|) + \alpha + 1 - \frac{k - |S| + 1}{2n} \\
&> (k - |S|)(1 + \alpha) + \alpha \,.
\end{aligned}$$

However, by Lemma 4.3, the maximum sum of any $k - |S| + 1$ elements of $[n] \setminus S$ is $(k - |S|) + (k - |S| + 1)\alpha = (k - |S|)(1 + \alpha) + \alpha$, a contradiction. Thus, $\|u - u'\|_\infty \geq \frac{1}{2n}$.

Therefore, in either case, ψ_k is ϵ-separated with respect to (Γ_k, γ_k). Finally, by Finocchiaro et al. (2022, Theorem 5), (L_k, ψ_k) is calibrated with respect to ℓ_k. □

5. Numerical Comparison

We have seen that L_k is consistent for top-k classification, while $L^{(2)}, L^{(3)}$, and $L^{(4)}$ are not. In general, therefore, we expect these inconsistent losses to have worse top-k performance than L_k. We now quantify this gap for the case $n = 5$ and $k = 3$, by computing the expected difference in top-k loss obtained as a result of optimizing each of the four surrogates.

Recall from Definition 2.1 that we have $\text{prop}[L](p) = \arg\min_{u \in \mathbb{R}^n} \langle p, L(u, \cdot) \rangle$ as the minimizers of the expected loss of L under p. For each surrogate L we measure their expected *risk*: the top-k loss obtained by optimizing L

$$\text{Risk}(L) = \mathop{\mathbb{E}}_{p \sim D} \left[\langle p, \ell_k \left(\psi_k \left(\text{prop}[L](p) \right), \cdot \right) \rangle \right] \,,$$

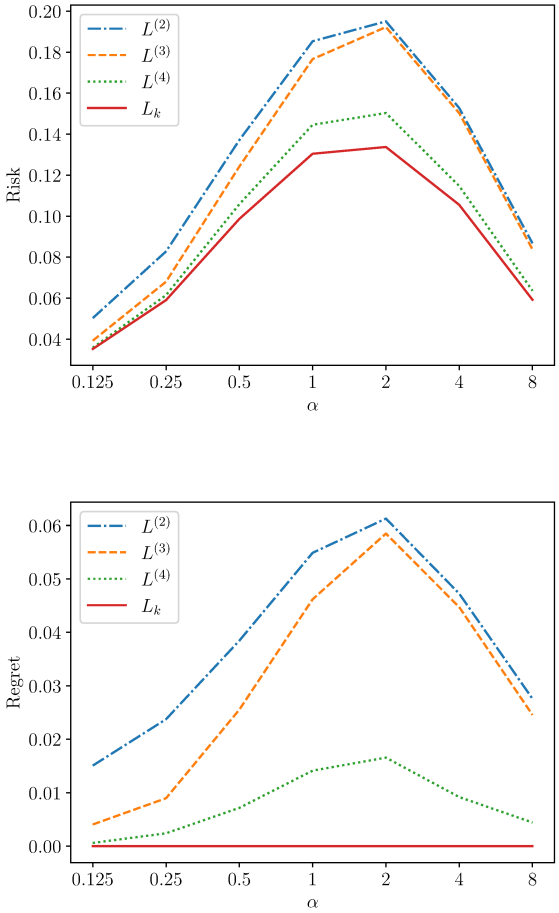

Figure 1: The top-k risk (top) and regret (bottom) from surrogate risk minimization of $L^{(2)}$, $L^{(3)}$, $L^{(4)}$, and L_k, for $n = 5$ and $k = 3$. For each choice of α, 1000 conditional label distributions were drawn from Dirichlet$(\alpha, \alpha, 1, 1, 1)$.

and *regret*: the risk minus the true optimal top-k loss.

$$\text{Regret}(L) = \text{Risk}(L) - \mathop{\mathbb{E}}_{p \sim D} \left[\min_{r \in \mathcal{R}_k} \langle p, \ell_k(r, \cdot) \rangle \right] \,.$$

Here p is a conditional label distribution, which we draw from $D = \text{Dirichlet}(\alpha, \alpha, 1, 1, 1)$, with α varied from 2^{-3} to 2^3. We take the ψ_k that breaks ties lexicographically. The results of these trials are shown in Figure 1.

When α is large, D concentrates on conditional label distributions with most of their weight on the first two labels, and for small α, it concentrates on those with weight on the last three. As $k = 3$, we expect all surrogates to perform well in these regimes, since it is relatively easy to select the most likely labels. For intermediate values, the distribution is closer to uniform, and the loss increases for all surrogates. However, the inconsistent surrogates incur the largest in-

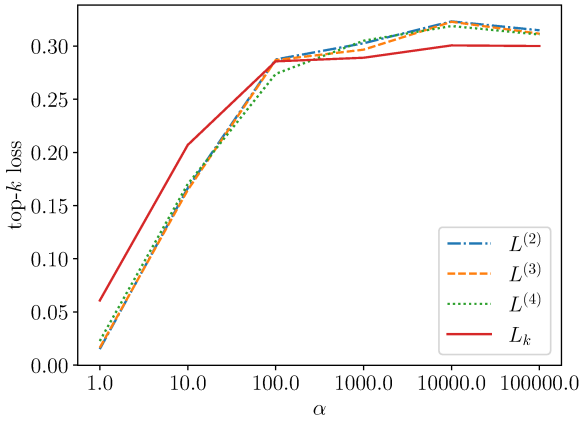

Figure 2: The empirical top-k test loss for each loss trained on a dataset with conditional label distributions sampled from Dirichlet(αp).

crease, and therefore largest regret, as they are more likely to link to a suboptimal set when $p_{[k]}$ is close to $p_{[k+1]}$.

As expected, L_k incurs no regret, since it is consistent. We also see that of the inconsistent surrogates, $L^{(2)}$ incurs the most regret, while $L^{(4)}$ incurs the least. This observation aligns with Table 1, which shows that $L^{(2)}$ has the largest inconsistent regions, while $L^{(4)}$ has the smallest.

Next, we verify this performance empirically. We fix $p = (.15, .15, .15, .2, .35)$, a point where $L^{(2)}, L^{(3)}$, and $L^{(4)}$ are inconsistent. For each value of α, we sample 10000 conditional label distributions $p_i \sim \text{Dirichlet}(\alpha p)$; we take the feature vector $x_i = p_i$ and draw the label $y_i \sim p_i$. For each dataset and each surrogate loss function, we train a linear model for 200 epochs using Adam with a learning rate of 0.01. Finally, for each α, we create a test set with 1000 samples in the same fashion. We then compute the top-k loss of the model trained for each surrogate loss, and plot the results in Figure 2.

For large α, the conditional labels are concentrated on a region where L_k is consistent but the other surrogate losses are not. In this regime, L_k clearly obtains a better top-k test loss. For smaller α, the conditional distributions are more evenly distributed on Δ_y, and in this regime L_k actually performs worse than the inconsistent surrogates. One explanation for this worse performance could be the shallowness of its gradients.

6. Discussion

In § 3, we apply the embedding framework of Finocchiaro et al. (2019; 2022) to analyze previously proposed, yet inconsistent, surrogates for top-k classification. The goal of this analysis is two-fold: first, to uncover the discrete losses for which these surrogates are consistent, and second, to characterize distributional conditions sufficent to render them consistent for top-k classification. We believe this general line of inquiry will be useful for other polyhedral surrogates in the literature known to be inconsistent for their desired target. In particular, while it is clearly useful to understand the circumstances in which these surrogates would be consistent, we also believe it would be useful to uncover the variants of the intended target which are embedded by these inconsistent surrogates.

To illustrate, consider the surrogate $L^{(4)}$, analyzed in § 3.3. We showed $L^{(4)}$ to be consistent for the target loss $\ell^{(4)}(T, y) = \frac{k+1}{k+1-|T|}\mathbb{1}\{y \notin T\}$, which allows one to predict any set of labels T with $|T| \leq k$. While $L^{(4)}$ is therefore consistent for top-k only when optimal sets T have size k, in practice, the extra flexibility to report smaller sets may be of use. That is, while common practice is to use $L^{(4)}$ with the argmax link ψ_k, which always yields a set of size k, it may be advantageous to use a link $\psi^{(4)}$ that makes $L^{(4)}$ consistent for $\ell^{(4)}$, which could link to sets strictly smaller than k. For example, suppose a search engine has $k = 10$ spaces to show on the first page, but given a specific query x, the model $h(x)$ links to $T = \psi^{(4)}(h(x))$ where $|T| = 7$. Given this information, the search engine may prefer to show only the results in T to reduce visual clutter, or perhaps serve advertisements in the remaining 3 slots. It is of course rare that a practical decision problem lines up exactly with the canonical discrete loss studied by machine learning researchers—exploring the variants of these canonical problems lurking behind inconsistent polyhedral surrogates may therefore be a useful line of research. We expect the general technique outlined in § 3 would apply readily to other such surrogates.

In § 4, we gave the first polyhedral surrogate that is consistent for top-k classification. This result contributes to an ongoing discussion in the literature about the relative benefits of smooth and polyhedral surrogates. While it has been suggested that no polyhedral surrogate could be consistent for top-k, our surrogate emphasizes the broader finding of Finocchiaro et al. (2022), that in fact *every* discrete target loss has a consistent polyhedral surrogate. Moreover, any smooth proper loss, with an appropriate link, suffices as a smooth surrogate (Williamson et al., 2016). The question is therefore not one of existence but of when and why smooth surrogates or polyhedral surrogates may be preferable. In particular, an important open direction is to study the relationship between smoothness, consistency, convergence rates, and excess risk tradeoffs for top-k classification, as well as other discrete prediction tasks.

Finally, while we give the first polyhedral surrogate that is consistent for top-k, it remains to compare it to other surrogates in practice beyond our limited experiments.

Acknowledgements

The authors would like to thank Enrique Nueve and the anonymous reviewers for their helpful suggestions. We also thank Forest Yang and Sanmi Koyejo for providing implementations of previously studied surrogates. This material is based upon work supported by the National Science Foundation under Grant No. IIS-2045347.

References

Gediminas Adomavicius and Jingjing Zhang. Classification, ranking, and top-k stability of recommendation algorithms. *INFORMS Journal on Computing*, 28(1):129–147, 2016.

Nir Ailon and Mehryar Mohri. An efficient reduction of ranking to classification. In *Proceedings of The 21st Annual Conference on Learning Theory (COLT 2008)*, Helsinki, Finland, 2008. URL http://www.cs.nyu.edu/~mohri/postscript/learning_ranking.pdf.

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Good practice in large-scale learning for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):507–520, 2013.

Kaiser Asif, Wei Xing, Sima Behpour, and Brian D Ziebart. Adversarial cost-sensitive classification. In *UAI*, pages 92–101, 2015.

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.

Daniel Billsus, Michael J Pazzani, et al. Learning collaborative information filters. In *ICML*, volume 98, pages 46–54, 1998.

Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177, 2004.

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

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

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

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. An embedding framework for the design and analysis of consistent polyhedral surrogates. *arXiv*, 2022.

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

Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

Branko Grünbaum, Victor Klee, Micha A Perles, and Geoffrey Colin Shephard. *Convex polytopes*, volume 16. Springer, 1967.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

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

Maksim Lapin, Matthias Hein, and Bernt Schiele. Loss functions for top-k error: Analysis and insights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1468–1477, 2016.

Maksim Lapin, Matthias Hein, and Bernt Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1533–1554, 2018.

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

Sashank J Reddi, Satyen Kale, Felix Yu, Daniel Holtmann-Rice, Jiecao Chen, and Sanjiv Kumar. Stochastic negative mining for learning with large output spaces. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1940–1949. PMLR, 2019.

Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 1997.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(5), 2007.

Robert Williamson, Elodie Vernet, Mark Reid, et al. Composite multiclass losses. 2016.

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

Figure 3: The general embedding procedure used to analyze $L^{(2)}$, $L^{(3)}$, and $L^{(4)}$.

A. Additional Derivations for $L^{(2)}$

Throughout this section, consider the surrogate loss

$$L^{(2)}(u, y) = \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u - e_y)_{[i]}\right)_+ . \tag{5}$$

We proceed as follows: find a bounded representative region for $L^{(2)}$, find the subsets of that region on which $u \mapsto L^{(2)}(u, y)$ is affine for all $y \in \mathcal{Y}$, enumerate the vertices of these regions as a finite representative set (since $L^{(2)}$ is polyhedral). By Theorem 2.5, $L^{(2)}$ embeds its restriction to these vertices. We can then study the property elicited by the embedded loss and compare it to the top-k property to understand which distributional assumptions are needed for top-k consistency. This procedure is in Figure 3.

In this section, we take $\bar{u}$ to be the average of the top k elements of u, $\bar{u} = \frac{\sigma_k(u)}{k}$. Moreover, we denote $\bar{u}_{-i} = \frac{1}{k-1}(\sigma_k(u) - u_{[i]})$ be the averages of the first k sorted elements of u and the average of the first $k-1$ sorted elements of u besides the i^{th} element, respectively. When $u \in \mathcal{U}_2$ as defined below, $\bar{u}_{-i}$ is the average of the top-k elements of u if $i \notin T_k(u)$ and the top $k-1$ of $u_{\setminus i}$ otherwise.

A.1. The Bounded Representative Region

We initially bound our report set with upper and lower bounds, and show the restricted set is representative. We first observe that $L^{(2)}$ is invariant in the $\mathbb{1}$ direction, which is necessary for our first restriction.

Lemma A.1 (Invariance in the $\mathbb{1}$ direction)**.** *$L^{(2)}(u, y) = L^{(2)}(u + \alpha\mathbb{1}, y)$ for all $\alpha \in \mathbb{R}$ and $y \in \mathcal{Y}$.*

Proof.

$$\begin{aligned}
L^{(2)}(u + \alpha\mathbb{1}, y) &= \left(1 - (u_y + \alpha) + \frac{1}{k}\sum_{i=1}^{k}(u + \alpha\mathbb{1} - e_y)_{[i]}\right)_+ \\
&= \left(1 - (u_y + \alpha) + \frac{1}{k}\alpha k + \frac{1}{k}\sum_{i=1}^{k}(u - e_y)_{[i]}\right)_+ \\
&= \left(1 - u_y - \alpha + \alpha + \frac{1}{k}\sum_{i=1}^{k}(u - e_y)_{[i]}\right)_+ \\
&= \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u - e_y)_{[i]}\right)_+ \\
&= L^{(2)}(u, y)
\end{aligned}$$

□

We now introduce our first restriction on reports and show it is representative.

Lemma A.2. $R_2^{\texttt{low}} := \{u \in \mathbb{R}^n_+ \mid \|u\|_0 \leq k\} = \{u \in \mathbb{R}^n \mid u_{[k+1]} = 0 = u_{[n]}\}$ *is a representative set for* $L^{(2)}$.

Proof. By Lemma A.1, we can fix $u_{[k+1]} = 0$ without loss of generality. We then have $u_{[i]} \leq 0$ for all $i \geq k+1$. Consider $u' = \max(u, \vec{0})$ such that $u'_{[i]} = u_{[i]}$ for all i such that $u_{[i]} \geq 0$, and $u'_{[i]} = 0$ otherwise. Observe that u and u' have the same ordering on their elements and $u' \in R_2^{\texttt{low}}$ by construction. We want to show that $L^{(2)}(u, y) \geq L^{(2)}(u', y)$ for all $y \in \mathcal{Y}$, and representativeness of $R_2^{\texttt{low}}$ follows.

If $u_y \leq u_{[k+1]} = 0$, then

$$\begin{aligned}
L^{(2)}(u, y) &= \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u - e_y)_{[i]}\right)_+ \\
&\geq \left(1 + \frac{1}{k}\sum_{i=1}^{k}(u' - e_y)_{[i]}\right)_+ && \text{top-}k+1 \text{ elements of } u \text{ and } u' \text{ are the same and } u_y \leq 0 \\
&= L^{(2)}(u', y)\ .
\end{aligned}$$

The inequality comes from the equality of the first $k+1$ sorted elements of u and u', combined with setting $u_y \leq u'_y = 0$ in this case.

Now, if $u_y \geq u_{[k+1]}$ and $u_y \geq 1$, observe that $\bar{u} = \bar{u}'$. We then have

$$\begin{aligned}
L^{(2)}(u, y) &= \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u - e_y)_{[i]}\right)_+ \\
&= \left(1 - u_y + \bar{u} - \frac{1}{k}\right)_+ && \text{substitution of summand by case} \\
&= \left(1 - u'_y + \bar{u}' - \frac{1}{k}\right)_+ && \bar{u} = \bar{u}' \\
&= L^{(2)}(u', y)\ .
\end{aligned}$$

Now suppose that $u_y \geq u_{[k+1]}$ and $u_y \in [0, 1)$; as the top-$k+1$ elements of u and u' are the same, observe $(u - e_y)_{[k]} = u_{[k+1]} = 0 = (u' - e_y)_{[k]}$.

$$\begin{aligned}
L^{(2)}(u, y) &= \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k-1}u_{[i]} + 0\right)_+ && \text{case} \\
&= \left(1 - u'_y + \frac{1}{k}\sum_{i=1}^{k-1}u'_{[i]}\right)_+ && u \text{ matches } u' \text{ on top } k+1 \text{ elements and } u_y = u'_y \\
&= \left(1 - u'_y + \frac{1}{k}\sum_{i=1}^{k}(u' - e_y)_{[i]}\right)_+ \\
&= L^{(2)}(u', y)\ .
\end{aligned}$$

Since $L^{(2)}(u, y) \geq L^{(2)}(u', y)$ for all $y \in \mathcal{Y}$, this also holds for the expected loss for all $p \in \Delta_\mathcal{Y}$. Thus, $R_2^{\texttt{low}}$ is representative. □

It follows from construction of $R_2^{\texttt{low}}$ that we can take $u_{[k+1]} = \ldots, = u_{[n]} = 0$ and still have a representative set.

Now, consider the set $R_2^{\texttt{high}} := \{u \in \mathbb{R}_+^n \mid u_i \leq \bar{u}_{-i} + 1 \;\; \forall i \in [n]\}$. While $R_2^{\texttt{low}}$ gives a lower bound on a representative region, $R_2^{\texttt{high}}$ gives an upper bound.

Lemma A.3. *The set $\mathcal{U}_2 := R_2^{\texttt{low}} \cap R_2^{\texttt{high}}$ is representative.*

Proof. Since we have already proven $R_2^{\texttt{low}}$ is representative in Lemma A.2, suppose $u \in R_2^{\texttt{low}}$. Any $u \notin R_2^{\texttt{high}}$ must have some element $i \in [n]$ such that $u_i > \bar{u}_{-i} + 1$. Consider u' as follows: for all $y \in [n]$ such that $u_y > \bar{u}_{-y} + 1$, reassign such a $u'_y := \bar{u}_{-y} + 1$. We proceed in two cases, showing below that $L^{(2)}(u', y) = L^{(2)}(u, y) = 0$ due to the positive part operator. In the second case, $L^{(2)}(u', j) \leq L^{(2)}(u, j)$ for all $j \neq y$ as $\bar{u} < \bar{u}'$. Moreover, $u' \in R_2^{\texttt{high}}$ by construction.

First, we consider when the outcome y is the modified element of u. We write $u_y = \bar{u}_{-y} + 1 + \epsilon$ for some $\epsilon > 0$ and $u'_y = \bar{u}_{-y} + 1$, with $u_j = u'_j$ for all $j \neq y$.

$$\begin{aligned}
L^{(2)}(u, y) &= \left(1 - (\bar{u}_{-y} + \epsilon + 1) + \frac{1}{k}\left(\sum_{j=1, j\neq y}^{k} u_j + \bar{u}_{-y} + \epsilon\right)\right)_+ \\
&= \left(1 - (\bar{u}_{-y} + \epsilon + 1) + \frac{1}{k}\left((k-1)\bar{u}_{-y} + \bar{u}_{-y} + \epsilon\right)\right)_+ \\
&= \left(-\bar{u}_{-y} - \epsilon + \bar{u}_{-y} + \frac{\epsilon}{k}\right)_+ \\
&= (-\frac{k-1}{k}\epsilon)_+ \\
&= 0
\end{aligned}$$

When $\epsilon = 0$, we recover u', in which case we observe the same result from $L^{(2)}(u', y) = (-\frac{k-1}{k}0)_+ = 0$. Thus, the losses are equal on the outcome y.

Now, let us consider $z \neq y$. Since $u \in R_2^{\texttt{low}}$, we have $\bar{u} \geq 0$ and $\bar{u}_{-i} \geq 0$ for any $i \in [n]$. Therefore, if $u_y > \bar{u}_{-y} + 1$, then we have $u_y > u_{[k+1]}$ as $u_y + 1 \geq 1 > 0 = u_{[k+1]}$. Now, for outcome $z \neq y$ (with $u_z \leq \bar{u}_{-z} + 1$, and therefore $u_z = u'_z$), we have

$$\begin{aligned}
L^{(2)}(u, z) &= \left(1 - u_z + \frac{1}{k}\sum_{i=1}^{k}(u - e_z)_{[i]}\right)_+ \\
&= \left(1 - u'_z + \frac{1}{k}\sum_{i=1}^{k}(u' - e_z)_{[i]} + \frac{\epsilon}{k}\right)_+ && u'_y \text{ is in the top } k \text{ elements of } (u' - e_z) \text{ and } u_z = u'_z \\
&\geq \left(1 - u'_z + \frac{1}{k}\sum_{i=1}^{k}(u' - e_z)_{[i]}\right)_+ && \text{Since } \epsilon > 0 \\
&= L^{(2)}(u', z)
\end{aligned}$$

If there is more than one index y such that $u_y > \bar{u}_{-y} + 1$, we can repeat this procedure in decsending order so the result holds.

Therefore, if $u \in \arg\min_r L^{(2)}(r, y)$, then so is some $u' \in \mathcal{U}_2$ for each $y \in \mathcal{Y}$, and we can say the same of the expected loss $\mathbb{E}_p L^{(2)}(u, \cdot)$ for all $p \in \Delta_\mathcal{Y}$. Thus, $\arg\min_u \mathbb{E}_p L^{(2)}(u, \cdot) \cap \mathcal{U}_2$ is nonempty for all $p \in \Delta_\mathcal{Y}$ and therefore $\mathcal{U}_2$ is representative. □

Re-writing the surrogate without the positive part operator. For any $u \in \mathcal{U}_2$, we can rewrite $L^{(2)}|_{\mathcal{U}_2}(u, y) = L^{(2)}(u, y) = 1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u - e_y)_{[i]}$, removing the positive part operator, as the term inside is always nonnegative for $u \in \mathcal{U}_2$. This allows us to re-write the loss as follows:

$$L^{(2)}(u, y) = 1 - u_y + \bar{u} - \frac{1}{k}\min(u_y, 1)\,. \tag{12}$$

Moreover, we can evaluate the expected loss

$$\begin{aligned}\mathbb{E}_p L^{(2)}(u,\cdot) &= \sum_y p_y \left(1 - u_y + \bar{u} - \frac{1}{k}\min(u_y, 1)\right) \\ &= \sum_y p_y(1+\bar{u}) - \sum_y p_y u_y - \sum_y p_y \frac{1}{k}\min(u_y,1) \\ &= 1 - \langle p, u\rangle + \bar{u} - \frac{1}{k}\langle p, \min(u, \mathbb{1})\rangle \ . \end{aligned} \tag{13}$$

A.2. Affine Regions and a Finite Representative Set

Since the loss $L^{(2)}$ is polyhedral, it has a finite set of minimizers (Finocchiaro et al., 2019, Lemma 2). Upon finding a finite representative set $\mathcal{R}^{(2)} \subseteq \mathcal{U}_2$, we can apply Theorem 2.5(2) and study the property elicited by $L^{(2)}|_{\mathcal{R}^{(2)}}$ via embeddings, and how it compares to the top-k property $\gamma_k := \text{prop}[\ell_k]$ under the argmax link.

As we showed $\mathcal{U}_2$ is representative in Lemma A.3, consider the following set

$$\mathcal{R}^{(2)} := \left\{ \frac{|M| + k - 1}{k - |H|}\mathbf{1}_H + \mathbf{1}_M : H, M \subset [n], H \cap M = \emptyset, |H| + |M| \leq k, |H| < k \right\} .$$

We will show that $\mathcal{R}^{(2)}$ enumerates the vertices of the regions where the function $u \mapsto \mathbb{E}_p L^{(2)}(u,\cdot)$ must be affine, regardless of $p \in \Delta_\mathcal{Y}$. Moreover, is the expected loss is polyhedral, it minimized on at least one face of these affine regions; since each face contains at least one vertex in $\mathcal{R}^{(2)}$, we will conclude $\mathcal{R}^{(2)}$ is representative.

Lemma A.4. *Fix a set $T \subseteq [n]$ such that $|T| = k$ and any $S \subseteq T$. Then $L^{(2)}(\cdot, y)$ is affine on the set $A^{T,S} := \{u \in \mathcal{U}_2 \mid T \in T_k(u) \wedge u_y \in [0,1] \forall y \in S \wedge u_y \in [1, 1 + \frac{1}{k-1}\sum_{j \in T_k(u)\setminus\{i\}} u_j] \forall y \in T \setminus S\}$ for all $y \in \mathcal{Y}$.*

Proof. First, observe the that $u \mapsto L^{(2)}(u,y)$ is affine in the first two terms of eq. (12) for all $u \in \mathcal{U}_2$, and nonlinearity is only introduced in the last two terms of eq. (13). Fix any set $T \subseteq [n]$ of size k. We denote by $A^T := \{u \in \mathcal{U}_2 \mid T_k(u) = T\}$ as the set of u whose top k elements are exactly the elements of T. Observe that $u \mapsto \bar{u}$ is affine in A^T for each T since $\bar{u}$ is the sum of the top k elements of u, regardless of their relative order.

Now, since $u \in \mathcal{U}_2 \supseteq R_2^{\texttt{low}}$, we have $u_{[k+1]} = \ldots = u_{[n]} = 0$, we impose k constraints constructing $R_2^{\texttt{low}}$ given by $0 \leq u_i$ for $i \in T$. Moreover, there are k constraints constructing $R_2^{\texttt{high}}$, given by $u_i \leq 1 + \bar{u}_{-i} = 1 + \frac{1}{k-1}\sum_{j \in T_k(u)\setminus\{i\}} u_j$ for all $i \in T$.

We now consider affineness of $u \mapsto \frac{1}{k}\min(u, \mathbb{1})$, where there is a "switch" of affine regions at $u_y = 1$ for each y. For a fixed set T and $u \in A^T$, consider any $S \subseteq T$. Construct the region $A^{T,S} = \{u \in A^T \mid u_y \in [0,1] \forall y \in S \wedge u_y \in [1, 1 + \frac{1}{k-1}\sum_{j \in T_k(u)\setminus\{i\}} u_j] \forall y \in T \setminus S\}$. Observe that $A^{T,S} \subseteq A^T$.

Since $A^{T,S} \subseteq A^T$, we know that $u \mapsto \bar{u}$ is affine on $A^{T,S}$, and construct $A^{T,S}$ so that $u \mapsto -\frac{1}{k}\min(u, \mathbb{1})$ is affine on $A^{T,S}$. Therefore, $u \mapsto L(u,y)$ is affine on $A^{T,S}$ for all $y \in \mathcal{Y}$, T of size k, and $S \subseteq T$ as it is the sum of affine functions. □

As vertices of the $A^{T,S}$ regions are formed by the intersections of these affine regions, we can now enumerate the vertices of the $A^{T,S}$ regions with $\mathcal{R}^{(2)}$.

Lemma A.5. *Let $\text{vert}(A^{T,S})$ be the vertices of the the region $A^{T,S}$, and let $\mathcal{V} := \cup_{T,S||T|=k,S\subseteq T}\text{vert}(A^{T,S})$. Then $\mathcal{V} \subseteq \mathcal{R}^{(2)}$.*

Proof. By a corollary of Lemma A.4, the function $u \mapsto \mathbb{E}_p L^{(2)}(u,\cdot)$ is affine on $A^{T,S}$ for any $T \subseteq [n]$ of size k and $S \subseteq T$ for all $p \in \Delta_\mathcal{Y}$. We now proceed to compute $\mathcal{V}$ by finding k equalities imposed on $A^{T,S}$ (Grünbaum et al., 1967). Vertices of each $A^{T,S}$ region are formed by the intersection of n hyperplanes technically, but with T fixed, the other $n-k$ come from the requirement $u_i = 0 \forall i \in [n] \setminus T$.

Fix T, S such that $|T| = k$ and $S \subseteq T$. We then have vertices at each of these 2^k possible equalities, given by the following constraints.

$$\forall i \in S,\ 0 \leq u_i \leq 1$$
$$\forall i \in T \setminus S,\ 1 \leq u_i \leq 1 + \frac{1}{k-1} \sum_{j \in T_k(u) \setminus \{i\}} u_j$$

Iterating over each of these 2^k inequalities, we see that vertices are generated at points $0, 1$ or some constant $c_{S,T} \geq 1$ for each choice of inequalities for T and S.

It suffices to show that for each T and S as above, there is a H and M satisfying the requirements of $\mathcal{R}^{(2)}$. In particular, we take $M = \{i \in S \mid u_i = 1\}$, and $H := \{i \in S \mid u_i > 1\}$. By construction, we have $H \cap M = \emptyset$, and $|H| + |M| \leq k$. Thus, every $v \in \mathcal{V}$ is contained in $\mathcal{R}^{(2)}$. □

Corollary A.6. *$\mathcal{R}^{(2)}$ is a finite representative set for $L^{(2)}$.*

A.3. The Loss Embedded by $L^{(2)}$

Corollary A.7. *$L^{(2)}$ embeds $L^{(2)}|_{\mathcal{R}^{(2)}}$.*

We can now evaluate the restricted function and obtain it in the form of a loss matrix.

$$L^{(2)}|_{\mathcal{R}^{(2)}}(r, y) = \begin{cases} 0 & r_y = \bar{r}_{-y} + 1 \\ \bar{r} - \frac{1}{k} & r_y = 1 \\ 1 + \bar{r} & r_y = 0 \end{cases}. \tag{14}$$

We can equivalently relabel the reports in $\mathcal{R}^{(2)}$ via a bijection Φ designating y as an element of H if $u_y > 1$, and y is an element of M if $u_y = 1$.

$$\hat{\ell}_2((H, M), y) = \begin{cases} 0 & y \in H \\ \frac{|H|+|M|-1}{k-|H|} & y \in M \\ \left(\frac{|H|+|M|-1}{k-|H|}\right) + \frac{k+1}{k} & \text{otherwise} \end{cases}$$

A.4. The Property Elicited by the Embedded Loss

The next natural question is to consider is whether or not $(L^{(2)}, \psi_k)$ is calibrated with respect to ℓ_k. In order to answer this, we necessarily need to understand something about $\Gamma := \text{prop}[L^{(2)}]$, which we will study through $\gamma := \text{prop}[\hat{\ell}_2]$.

In the previous subsection, we saw the construction of "high" (H), "meduim" (M), and "low" ($L := [n] \setminus (H \cup M)$) bins for the elements of $\mathcal{R}^{(2)}$ via the bijection Φ. However, because of the nature of $\mathcal{U}_2$, there is a dependence of multiple coordinates for an optimal report of $L^{(2)}$. That is, for a fixed probability distribution $p \in \Delta_\mathcal{Y}$, there may be coordinates $i \in [n]$ with "enough" weight for $i \in H \cup M$, but there is sometimes a benefit in expected loss for this surrogate by artificially bumping up from the "low" to "middle" bin when possible because doing so cranks up the constant on the "high" reports, yielding better expected loss. That is, sometimes an algorithm is confident enough in its "high" labels that it is optimal to take an additional expected loss on some "lower" labels.

Next, we characterize the distributions p such that $\mathbb{E}_p \hat{\ell}_2((H \cup \{i\}, M), \cdot) \leq \mathbb{E}_p \hat{\ell}_2((H, M \cup \{i\}), \cdot)$.

Lemma A.8. *Fix some $(H, M) \in \Phi(\mathcal{R}^{(2)})$ and consider any index $i \in [n] \setminus (H \cup M)$. Consider $u \in \Phi(\mathcal{R}^{(2)})$ such that $u = (H \cup \{i\}, M)$ and $u' = (H, M \cup \{i\})$. Then $\mathbb{E}_p \hat{\ell}_2(u, \cdot) \leq \mathbb{E}_p \hat{\ell}_2(u', \cdot)$ if and only if $p_i \geq (1 - \sigma_H(p))(\frac{1}{k-h})$.*

Proof. Let $h = |H|$ and $m = |M|$.

$$\begin{aligned}
\mathbb{E}_p\hat{\ell}_2(u,\cdot) &\leq \mathbb{E}_p\hat{\ell}_2(u',\cdot) \\
(1-(\sigma_H(p)+p_i))\left(\frac{k(h+1+m)}{k-h-1}\right) &\leq (1-\sigma_H(p))\left(\frac{k(h+1+m)}{k-h}\right) \\
\frac{(k(1-\sigma_H(p)))}{(k-h-1)(k-h)} &\leq p_i\frac{k}{k-h-1} \\
\frac{1-\sigma_H(p)}{k-h} &\leq p_i\,.
\end{aligned}$$

□

Observe that for $H = \emptyset$, this inequality becomes $p_i \geq \frac{1}{k}$. Now, we characterize the distributions p such that $\mathbb{E}_p\hat{\ell}_2((H, M\cup\{i\}),\cdot) \leq \mathbb{E}_p\hat{\ell}_2((H,M),\cdot)$.

A.4.1. $M \succeq_i L$

Lemma A.9. *Fix $(H,M) \in \Phi(\mathcal{R}^{(2)})$ and consider any index $i \in [n]\setminus(H\cup M)$. Consider $u \in \Phi(\mathcal{R}^{(2)})$ such that $(H, M\cup\{i\})$ and $u' = (H,M)$. Then $\mathbb{E}_p\hat{\ell}_2(u,\cdot) \leq \mathbb{E}_p\hat{\ell}_2(u',\cdot)$ if and only if $p_i \geq (\frac{h}{k} - \sigma_H(p))(\frac{k}{(k-h)(k+1)}) + \frac{1}{k+1}$.*

Proof. Let $h = |H|$ and $m = |M|$.

$$\begin{aligned}
\mathbb{E}_p\hat{\ell}_2(u,\cdot) &\leq \mathbb{E}_p\hat{\ell}_2(u',\cdot) \\
(1-\sigma_H(p))\tfrac{k(h+m+1)}{k-h} + \tfrac{(k+1)(1-\sigma_H(p)-\sigma_M(p)-p_i)}{k} &\leq (1-\sigma_H(p))\tfrac{k(h+m)}{k-h} + \tfrac{(k+1)(1-\sigma_H(p)-\sigma_M(p))}{k} \\
\frac{k(1-\sigma_H(p))}{k-h} &\leq \frac{k+1}{k}p_i \\
\frac{1-\sigma_H(p)}{(k+1)(k-h)} &\leq p_i\,.
\end{aligned}$$

□

Lemmas A.8 and A.9 now provide testable conditions to yield $\mathrm{prop}[\hat{\ell}_2]$ as $\mathcal{R}^{(2)}$ is finite. Now let us consider how one wants to assign indices to each of these three bins.

Consider first that we can calculate the set of indices that should be designated in H.

$$h^*(p) = \max\left\{ i \in \{0,\dots,k-1\} \mid p_{[i]} > \frac{1-\sum_{j=1}^{i-1} p_{[j]}}{k-(i-1)} \right\} \tag{15}$$

Now, let us consider $p_{h^*(p)} := \sigma_{h^*(p)}(p) = \sum_{j=1}^{h^*(p)} p_{[j]}$ to determine which elements of p should be designated in M.

$$m^*(p) = \max\{ j \in \{0,\dots,k\} \mid p_{[j]} > \frac{1-p_{h^*(p)}}{(k+1)(k-h^*(p))}\} \tag{16}$$

A.5. Characterizing Consistency of $L^{(2)}$ with Respect to ℓ_k

We have consistency via the canonical argmax link ψ_k when the optimal surrogate reports u have $u_{[k]} > 0 = u_{[k+1]}$, since its top-k set is unique. For intuition, consider that inconsistency means that any sequence of reports $\{u_i\}$ approaching the $L^{(2)}$ optimum and applying the link (e.g., $\{r_i\} = \{\psi(u_i)\}$ approaches the $L^{(2)}|_{\mathcal{R}^{(2)}}$ optimum; equivalently, approaching the $L^{(2)}$ optimum implies that $(H_i.M_i) = \{\Phi(r_i)\}$ approaches the $\hat{\ell}_2$ optimum.

Consider some distributions p, p' such that $T_k(p) \neq T_k(p')$ but $(H,M) \in \mathrm{prop}[\hat{\ell}_2](p) = \mathrm{prop}[\hat{\ell}_2](p')$ with $|H\cup M| < k$. As the link must be deterministic, given $u = \Phi^{-1}((H,M))$, the link must choose some ordering over the elements $S \subseteq [n]$ such that $u_i = 0$ for all $i \in S$. Even if this ordering aligns with $T_k(p)$, it will not align with $T_k(p')$ as they are not equal;

hence the ambiguity in $T_k(u)$ makes it impossible for consistency to hold at both p and p'. Thus, we will only have consistency guaranteed at distributions $p \in \Delta_\mathcal{Y}$ such that there is a value $u \in \text{prop}[L^{(2)}](p)$ with $T_k(u)$ unambiguous. The distributions where this condition holds are exactly the p for which $m^*(p) = k$.

Lemma A.10. *Let $L : \mathbb{R}^n \times \mathcal{Y} \to \mathbb{R}_+$ be a polyhedral loss which embeds $\hat{\ell} : \hat{\mathcal{R}} \times \mathcal{Y} \to \mathbb{R}_+$. Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ be a target loss. Let $\mathcal{P} \subseteq \Delta_\mathcal{Y}$. Let $\hat{\gamma} = \text{prop}[\hat{\ell}]$, $\gamma = \text{prop}[\ell]$. If for all $\hat{r} \in \hat{\mathcal{R}}$, there exists some $r \in \mathcal{R}$ such that $\{p \in \mathcal{P} : \hat{r} \in \hat{\gamma}(p)\} \subseteq \{p \in \mathcal{P} : r \in \gamma(p)\}$, then there exists a link function $\psi : \mathbb{R}^n \to \mathcal{R}$ such that (L, ψ) is calibrated with respect to ℓ on $\mathcal{P}$.*

Proof. The proof is essentially the same as that of Finocchiaro et al. (2022, Theorem 8), but restricted to $\mathcal{P} \subseteq \Delta_\mathcal{Y}$. Since L embeds $\hat{\ell}$, let $\hat{\psi} : \mathbb{R}^n \to \hat{\mathcal{R}}$ be a link function such that $(L, \hat{\psi})$ is calibrated with respect to $\hat{\ell}$ (Theorem 2.5). By Finocchiaro et al. (2022, Lemma 4), $\hat{\ell}$ indirectly elicits γ for some link function $\psi^\mathcal{R} : \hat{\mathcal{R}} \to \mathcal{R}$. Then for any $r \in \hat{\mathcal{R}}$ and $p \in \mathcal{P}$, $r \in \hat{\gamma}(p) \implies \psi^\mathcal{R}(r) \in \gamma(p)$.

Now, let $\psi = \psi^\mathcal{R} \circ \hat{\psi} : \mathcal{R}^n \to \mathcal{R}$. We will show (L, ψ) is calibrated with respect to ℓ on $\mathcal{P}$. By the construction of ψ, for any $p \in \mathcal{P}$ and $u \in \mathbb{R}^d$, if $\hat{\psi}(u) \in \hat{\gamma}(p)$, then $\psi(u) = \psi^\mathcal{R}(\hat{\psi}(u)) \in \gamma(p)$. Similarly, if $\psi(u) \notin \gamma(p)$, then $\hat{\psi}(u) \notin \hat{\gamma}(p)$. Therefore,

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

Since $(L, \hat{\psi})$ is calibrated with respect to $\hat{\ell}$, we obtain

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

so (L, ψ) is calibrated with resepct to ℓ on $\mathcal{P}$. □

Corollary 3.1. *Define*

$$\mathcal{P}^{(2)} := \left\{ p \in \Delta_\mathcal{Y} \mid p_{[k]} > \frac{(1 - \sigma_{h^*(p)}(p))}{(k+1)(k - h^*(p))} \right\} . \tag{7}$$

$L^{(2)}$ is consistent with respect to ℓ_k on $\mathcal{P}^{(2)}$.

B. Additional Derivations for $L^{(3)}$

Recall that we have

$$L^{(3)}(u, y) = \frac{1}{k} \sum_{i=1}^{k} \left[1 - u_y + (u - e_y)_{[i]} \right]_+ \tag{17}$$

We follow the same general procedure as § A; see Figure 3 for an outline.

B.1. Finding a Representative Region

As with $L^{(2)}$, we can show that $L^{(3)}$ is invariant in the ones direction.

Lemma B.1 (Invariance in the $\mathbb{1}$ direction)**.** *$L^{(3)}(u, y) = L^{(3)}(u + \alpha \mathbb{1}, y)$ for all $\alpha \in \mathbb{R}$.*

Proof.

$$\begin{aligned}
L^{(3)}(u + \alpha\mathbb{1}, y) &= \frac{1}{k} \sum_{i=1}^{k} \left[1 - (u_y + \alpha) + (u + \alpha\mathbb{1} - e_y)_{[i]} \right]_+ \\
&= \frac{1}{k} \sum_{i=1}^{k} \left[1 - u_y - \alpha + (u - e_y)_{[i]} + \alpha \right]_+ \\
&= \frac{1}{k} \sum_{i=1}^{k} \left[1 - u_y + (u - e_y)_{[i]} \right]_+ \\
&= L^{(3)}(u, y) .
\end{aligned}$$

□

As before, we can then set $u_{[k+1]} = 0$ without loss of generality, and show that we can restrict to the representative set $R_3^{\texttt{low}} := \{u \in \mathbb{R}^n_+ \mid \|u\|_0 \leq k\}$ (Lemma B.2). Throughout, let $T(u) \in T_k(u)$ be some choice of top-k elements of u so that $|T(u)| = k$. For a fixed choice $T(u)$, we additionally consider $V_y(u) := \{i \in T(u) \mid u_i > u_y - 1\} \setminus \{y\}$.

Lemma B.2 ($R_3^{\texttt{low}}$ **is representative for** $L^{(3)}$**).** *Consider $u \in \mathbb{R}^n_+$ such that $u_{[k+1]} = 0$, and u' such that $u'_{[i]} = u_{[i]}$ for $i \leq k+1$, and $u'_{[i]} = 0$ otherwise. For all $y \in \mathcal{Y}$, $L^{(3)}(u, y) \geq L^{(3)}(u', y)$.*

Proof. Observe that there is a choice of T such that $T(u) = T(u')$; we proceed with this choice, though any other choice of $T(u') \in T_k(u)$ results in the same loss values. Consider two cases: first, if $y \in T(u)$, and then if $y \notin T(u)$.

Case 1: $y \in T(u)$ follows trivially since the elements being summed over are equal (e.g., $u_i = u'_i \forall i \in T(u) = T(u')$), so the losses are equal.

Case 2: $y \notin T(u)$

$$\begin{aligned}
L^{(3)}(u, y) &= \frac{1}{k} \sum_{i \in T(u)} \left[1 - u_y + (u - e_y)_{[i]}\right]_+ \\
&= \frac{1}{k} \sum_{i \in T(u)} \left[1 - u_y + u_i\right]_+ \\
&= \frac{1}{k} \sum_{i \in T(u')} \left[1 - u_y + u'_i\right]_+ \\
&\geq \frac{1}{k} \sum_{i \in T(u')} \left[1 + u'_i\right]_+ \\
&= L^{(3)}(u', y) \, .
\end{aligned}$$

By Lemma B.1, u is invariant in the ones direction, so without loss of generality we can set $u_{[k+1]} = 0$. The cases above show we can set $u_{[j]} = 0$ for any $j > k+1$ without increasing the loss on any outcome. Together, these results imply that $R_3^{\texttt{low}} = \{u \in \mathbb{R}^n_+ \mid \|u\|_0 \leq k\}$ is representative. □

We continue towards a finite representative set, showing that each element of u should be no more than 1 greater than the next lowest element

Lemma B.3. *Consider $u \in R_3^{\texttt{low}}$ and $i \in \{1, \ldots, k\}$ such that $u_{[i]} = u_{[i+1]} + 1 + \varepsilon$ for some $\varepsilon > 0$. Take u' such that $u'_{[j]} = u_{[j]} - \varepsilon$ for all $j \in \{1, \ldots, i\}$. Then there exists a choice of T such that $T(u) = T(u')$ and $(1 - u_y + u_j)_+ \geq (1 - u'_y + u'_j)_+$ for all $j \in V_y(u)$, so $V_y(u') \subseteq V_y(u)$.*

Proof. First, observe $T_k(u) = T_k(u')$, as we are only shifting at most the top k elements of u, and they are being shifted in a way that preserves them as the top-k. Thus, by taking $T(u)$ to be a function of $T_k(u)$, a choice of T such that $T(u) = T(u')$ exists.

For any outcome $y \in \mathcal{Y}$ and index $j \in [n]$, there are four possible cases for the change in u_y and u_j: (1) neither is modified (e.g., $u_y = u'_y$ and $u_j = u'_j$); (2) just u_y is modified (e.g., $u_y = u'_y + \varepsilon$ and $u_j = u'_j$); (3) just u_j is modified (e.g., $u_y = u'_y$ and $u_j = u'_j + \varepsilon$); and (4) both are modified (e.g., $u_y = u'_y + \varepsilon$ and $u_j = u'_j + \varepsilon$). Cases 1 and 4 are immediate, $(1 - u_y + u_j)_+ = (1 - u'_y + u'_j)_+$ by substitution.

Case 2: $u_y = u'_y + \varepsilon$, $u_j = u'_j$. For this case to occur, $u_y \geq u_{[i]}$ and $u_{[i]} > u_j$. Therefore, $u_y > u_i + 1 \geq u_j$, violating the construction of $V_y(u)$.

Case 3: $u_y = u'_y$, $u_j = u'_j + \varepsilon$. By the case, we have $u_j > u'_j \geq u_{[i+1]} + 1 \geq u'_y = u_y$. As $u'_j < u_j$, we immediately have $(1 - u_y + u_j)_+ \geq (1 - u'_y + u'_j)_+ \geq 0$.

□

Let us denote the set $R_3^{\texttt{high}} := \{u \in \mathbb{R}^n \mid u_{[i+1]} \leq u_{[i]} \leq u_{[i+1]} + 1 \; \forall i \in (1, \ldots, k)\}$. We now give a bounded representative set for $L^{(3)}$.

Lemma B.4. *The set $\mathcal{U}_3 := R_3^{\texttt{high}} \cap R_3^{\texttt{low}}$ is representative for $L^{(3)}$.*

Proof. Fix any $u \in R_3^{\texttt{low}}$ such that for some $i \in \{1, \dots, k\}$ and $\varepsilon > 0$, $u_{[i]} = u_{[i+1]} + 1 + \varepsilon$. Take u' such that $u'_{[j]} = u_{[j]} - \varepsilon$ for all $j \in \{1, \dots, i\}$. We want to show $L^{(3)}(u, y) \geq L^{(3)}(u', y)$ for all $y \in \mathcal{Y}$.

By construction of $V_y(u)$, we can write

$$L^{(3)}(u, y) = \sum_{j \in V_y(u)} (1 - u_y + u_j) \,. \tag{18}$$

Moreover, we have the existence of a T such that $T(u) = T(u')$ and $V_y(u) \supseteq V_y(u')$ by Lemma B.3.

We can consider 3 cases for any $j \in V_y(u)$: (1) $u'_y = u_y$ and $u'_j = u_j$; (2) $u'_y + \varepsilon = u_y$ and $u'_j + \varepsilon = u_j$; and (3) $u'_y = u_y$ and $u'_j + \varepsilon = u_j$. For cases (1) and (2), we immediately have $(1 - u_y + u_j) = (1 - u'_y + u'_j)$, and for (3), we have $(1 - u_y + u_j) = (1 - u'_y + u'_j + \varepsilon) > (1 - u'_y + u'_j)$.

$$\begin{aligned}
L^{(3)}(u, y) &= \sum_{j \in V_y(u)} (1 - u_y + u_j) \\
&\geq \sum_{j \in V_y(u')} (1 - u_y + u_j) && \text{Since } V_y(u) \supseteq V_u(u') \\
&\geq \sum_{j \in V_y(u')} (1 - u'_y + u'_j) && \text{By substitution} \\
&= L^{(3)}(u', y) \,.
\end{aligned}$$

As this is true for all $y \in \mathcal{Y}$, we have $\mathbb{E}_p L^{(3)}(u, \cdot) \geq \mathbb{E}_p L^{(3)}(u', \cdot)$ for all $p \in \Delta_\mathcal{Y}$, yielding the result. □

B.2. Characterizing Affineness

Furthermore, we can show that $u \mapsto \mathbb{E}_p L^{(3)}(u, \cdot)$ is affine on the following regions for all $p \in \Delta_\mathcal{Y}$.

Lemma B.5. *Fix a set $T \subseteq n$ such that $|T| = k$ and the set $\vec{V} = \{V_y \mid V_y \subseteq T, y \in \mathcal{Y}\}$.*

$$A^{T,\vec{V}} = \{u \in \mathcal{U}_3 \mid T \in T_k(u) \wedge V_y = V_y(u), \forall y \in \mathcal{Y}\} \,.$$

Then $u \mapsto \mathbb{E}_p L^{(3)}(u, \cdot)$ is affine on each $A^{T,\vec{V}}$ for all $p \in \Delta_\mathcal{Y}$.

Proof. Nonaffineness in $u \mapsto L^{(3)}(u, y)$ for any $y \in \mathcal{Y}$ is imposed where there is a change in $T(\cdot)$ or in $V_y(\cdot)$ since we can write $L^{(3)}(u, y) = \sum_{i \in V_y(u)} (1 - u_y + u_i)$ as in eq. (18). As non-affineness is only introduced in the terms of the summand, we construct $A^{T,\vec{V}}$ so that $T(u) \in T_k(u)$ and $V_y(u)$ is constant on $A^{T,\vec{V}}$, and thus the terms of the summand are constant on $A^{T,\vec{V}}$. Therefore, $u \mapsto \mathbb{E}_p L^{(3)}(u, \cdot)$ is affine on $A^{T,\vec{V}}$ for all $p \in \Delta_\mathcal{Y}$. □

B.3. Constructing a Finite Representative Set

When constructing a finite representative set, it is sufficient to consider the vertices of these affine regions; thus, Lemma B.5 yields a finite representative set as follows.

Corollary B.6. *$\mathcal{R}^{(3)} := \mathcal{U}_3 \cap \mathbb{Z}_k^n$ is a finite representative for $L^{(3)}$.*

Thus, we can think of the loss $L^{(3)}|_{\mathcal{R}^{(3)}}$ as taking in as predictions an ordered partition of size at most k partitions. As with $L^{(2)}$, we can relabel the elements of $\mathcal{R}^{(3)}$ via some bijection Φ; in particular, we consider a bijection to ordered partitions as follows. Let $\mathcal{Q} = \{(Q_0, Q_1, \dots, Q_s) \mid s \leq k, Q_i \cap Q_j = \emptyset \forall i \neq j, |Q_s \cup \dots \cup Q_1| \leq k\}$. Let $\Phi : \mathcal{R}^{(3)} \to \mathcal{Q}$ be the bijection $u \mapsto (\{i \in [n] \mid u_i = 0\}, \{i \in [n] \mid u_i = 1\}, \dots, \{i \in [n] \mid u_i = s\})$. Then we can denote $\hat{\ell}_3$ such that $L^{(3)}(u, y) = \hat{\ell}_3(\Phi(u), y)$ for all $u \in \mathcal{R}^{(3)}$.

$$\hat{\ell}_3(Q, y) = \begin{cases} \frac{1}{k}\left(|Q_j| - 1 + \sum_{i>j} |Q_i|(i - j + 1)\right) & j > 0 \\ \frac{1}{k}\left(\sum_{i=1}^{s} |Q_i|(i+1)\right) & j = 0 \end{cases} , \tag{19}$$

where $y \in Q_j$.

B.4. Analyzing the Loss Embedded by $L^{(3)}$: Characterizing Consistency

Now that we have the finite representative set $\mathcal{R}^{(3)}$ for $L^{(3)}$, we can characterize the property elicited by $L^{(3)}$.

Lemma B.7. *Fix $u \in \mathcal{R}^{(3)}$ with $u_{[k]} = 1$, and consider $u' \in \mathcal{R}^{(3)}$ such that $u'_{[k]} = 0$ and $u_{[i]} = u'_{[i]}$ for all $i \in \{1, \dots, k-1\}$. Then $\mathbb{E}_p L^{(3)}(u, \cdot) \geq \mathbb{E}_p L^{(3)}(u', \cdot) \iff \frac{\sum_{i=k+1}^{n} p_{[i]}}{k-1} \geq p_{[k]}$.*

Proof. First, observe that we are not changing the relative order of elements of u and u', so there is a choice $T \in T_k$ such that $T(u) = T(u')$, and for each y, the loss is positive on the same set of indices.

$$\begin{aligned}
\mathbb{E}_p L^{(3)}(u, \cdot) &\geq \mathbb{E}_p L^{(3)}(u', \cdot) \\
\sum_{y \neq [k]} p_y \sum_{i \in V_y(u)} (1 - u_y + u_i) + p_{[k]} \sum_{i \in T(u) \setminus k} (1 - u_{[k]} + u_i) &\geq \sum_{y \neq [k]} p_y \sum_{i \in V_y(u)} (1 - u'_y + u'_i) + p_{[k]} \sum_{i \in T(u') \setminus k} (1 - u'_{[k]} + u'_i) \\
\sum_{y : p_y < p_{[k]}} p_y - p_{[k]}(k-1) &\geq 0 \\
\sum_{y : p_y < p_{[k]}} p_y &\geq p_{[k]}(k-1) \\
\frac{\sum_{y : p_y < p_{[k]}} p_y}{k-1} &\geq p_{[k]} \\
\frac{\sum_{i=k+1}^{n} p_{[i]}}{k-1} &\geq p_{[k]} \, .
\end{aligned}$$

The result follows. □

This result partially characterizes when it is better to keep the k^{th} element of u as 0: when it only imposes change in that one element. This is particularly important to characterize inconsistency for top-k; if $u_{[k]} = u_{[k+1]} = 0$, then $|T_k(u)| > 1$ for $u \in \text{prop}[L^{(3)}]$, so how to link u is ambiguous.

However, we also need to understand when it is beneficial to bump every higher element up by 1, which is given by the following result.

Lemma B.8. *Fix $u \in \mathcal{R}_3$ with $u_{[j]} = 0$, and consider $u' \in \mathcal{R}^{(3)}$ such that $u'_{[j]} = 1$ and $u_{[i]} + 1 = u'_{[i]}$ for all $i = 1, \dots, j$. Then $L^{(3)}(u; p) \geq L^{(3)}(u'; p) \iff p_{[j+1]} \geq \frac{1}{k+1}$.*

Proof.

$$\begin{aligned}
\mathbb{E}_p L^{(3)}(u, \cdot) &\geq \mathbb{E}_p L^{(3)}(u', \cdot) \\
\sum_{i=j+1}^{n} p_{[i]}(k+j) + \sum_{i=1}^{j}(j-1) &\geq \sum_{i=j+2}^{n} p_{[i]}(k+j+1) + \sum_{i=1}^{j+1}(j) \\
p_{[j+1]}(k+j) &\geq \sum_{i=j+2}^{n} p_{[i]} + p_{[j+1]}(j-1) + \sum_{i=1}^{j+1} p_{[i]} \\
p_{[j+1]}(k+j) &\geq 1 + p_{[j+1]}(j-1) \\
p_{[j+1]} &\geq \frac{1}{k+1} \, .
\end{aligned}$$

□

Lemmas B.7 and B.8 together characterize the the distributions $p \in \Delta_{\mathcal{Y}}$ where the report $u \in \text{prop}[L^{(3)}](p) \cap \mathcal{R}^{(3)}$ has $u_{[k]} > 0$. Thus, for $u \in \text{prop}[L^{(3)}](p)$ for such distributions p, $u_{[k]} > 0$ and therefore $|T_k(u)| = 1$. Applying Lemma A.10, we obtain the desired consistency result.

Corollary 3.2. *$L^{(3)}$ is consistent with respect to ℓ_k on $\mathcal{P}^{(3)} = \{p \in \Delta_{\mathcal{Y}} \mid p_{[k+1]} > \frac{1}{k+1} \wedge \frac{\sum_{i=k+1}^{n} p_{[i]}}{k-1} \geq p_{[k]}\}$.*

C. Additional Derivations for $L^{(4)}$

Recall that for a report $u \in \mathbb{R}^n$ and label $y \in \mathcal{Y}$,

$$L^{(4)}(u, y) = \left(1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]}\right)_+ .$$

We again follow the procedure in § A to find a representative region for $L^{(4)}$.

C.1. Constructing a Bounded, Representative Region for $L^{(4)}$

To establish a bounded, representative region for $L^{(4)}$, we must first show that $L^{(4)}$ is invariant in the $\mathbb{1}$ direction.

Lemma C.1 (Invariance in the $\mathbb{1}$ direction)**.** *$L^{(4)}(u, y) = L^{(4)}(u + \alpha\mathbb{1}, y)$ for all $\alpha \in \mathcal{R}$.*

Proof.

$$\begin{aligned}
L^{(4)}(u + \alpha\mathbb{1}, y) &= (1 - (u_y + \alpha) + \frac{1}{k}\sum_{i=1}^{k}((u + \alpha\mathbb{1})_{\setminus y})_{[i]})_+ \\
&= (1 - u_y - \alpha + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y} + \alpha\mathbb{1}_{\setminus y})_{[i]})_+ \\
&= (1 - u_y - \alpha + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]} + \frac{1}{k}k\alpha)_+ \\
&= (1 - u_y - \alpha + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]} + \alpha)_+ \\
&= (1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]})_+ \\
&= L^{(4)}(u, y)
\end{aligned}$$

□

Let the sets $R_4^{\texttt{low}}$ and $R_4^{\texttt{high}}$ be defined as follows:

- $R_4^{\texttt{low}} = \{u \in \mathbb{R}^n_+ \mid ||u||_0 \leq k\}$
- $R_4^{\texttt{high}} = \{u \in \mathbb{R}^n_+ \mid u_y \leq 1 + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]} \quad \forall y \in \mathcal{Y}\}$.

We will show in Theorem C.3 that the intersection $\mathcal{U}_4 := R_4^{\texttt{low}} \cap R_4^{\texttt{high}}$ is representative.

Lemma C.2. *$R_4^{\texttt{low}}$ is a representative set for $L^{(4)}$.*

Proof. Suppose that $u \in \mathbb{R}^n$ where $u_{[k+1]} = 0$. By Lemma C.1, $u_{[k+1]} = 0$ is without loss of generality. Let $u' = \max(u, \vec{0})$ be the element-wise max, which is in $R_4^{\texttt{low}}$ by construction. It suffices to show that for all $y \in \mathcal{Y}, L^{(4)}(u, y) \geq L^{(4)}(u', y)$.

By construction, there is a set $S \subseteq \mathcal{Y}, |S| = k$ such that $S \in T_k(u) \cap T_k(u')$. We proceed in two cases: if $y \in S$, and if $y \notin S$.

Case 1: $y \in S$:
In this case, we have $u_y = u'_y \geq 0$.

$$\begin{aligned}
L^{(4)}(u, y) &= (1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]})_+ \\
&= (1 - u'_y + \frac{1}{k}\sum_{i=1}^{k}(u'_{\setminus y})_{[i]})_+ \\
&= L^{(4)}(u', y) .
\end{aligned}$$

Case 2: $y \notin S$: In this case, we have $u_y \leq u'_y = 0$. Moreover, $\sum_{i=1}^k (u_{\setminus y})_{[i]} = \sum_{j \in S} u_j = \sum_{j \in S} u'_j = \sum_{i=1}^k (u'_{\setminus y})_{[i]}$, as $S \in T_k(u_{\setminus y}) \cap T_k(u'_{\setminus y})$.

$$\begin{aligned} L^{(4)}(u, y) &= (1 - u_y + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]})_+ \\ &\geq (1 - 0 + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]})_+ \\ &= (1 - u'_y + \frac{1}{k} \sum_{i=1}^k (u'_{\setminus y})_{[i]})_+ \\ &= L^{(4)}(u', y) \ . \end{aligned}$$

Therefore, for all y, we have $L^{(4)}(u, y) \geq L^{(4)}(u', y)$. Thus, $R_4^{\texttt{low}}$ is representative. □

Using $R_4^{\texttt{low}}$ as a starting point, we now proceed to show $\mathcal{U}_4 := R_4^{\texttt{low}} \cap R_4^{\texttt{high}}$ is a representative set for $L^{(4)}$.

Theorem C.3. *The set $\mathcal{U}_4 := R_4^{\texttt{low}} \cap R_4^{\texttt{high}}$ is a representative set for $L^{(4)}$.*

Proof. Since $R_4^{\texttt{low}}$ is representative by Lemma C.2, consider $u \in R_4^{\texttt{low}}$. Moreover, if $u \notin R_4^{\texttt{high}}$, construct $u' \in \mathbb{R}_+^n$ such that

$$u'_y = \begin{cases} 1 + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]} & u_y > 1 + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]} \\ u_y & u_y \leq 1 + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]} \end{cases} .$$

Observe that $u' \in R_4^{\texttt{low}} \cap R_4^{\texttt{high}}$ by construction and $\forall y \in \mathcal{Y}, u_y \geq u'_y$.
Since $u \notin R_4^{\texttt{high}}$, there is a $y \in \mathcal{Y}$ such that $u_y > 1 + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]}$; we can equivalently write

$$u_y = \left(1 + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]}\right) + \epsilon, \tag{20}$$

for some $\epsilon > 0$. We now proceed in two cases: considering the ground truth $y' = y$ and $y' \neq y$.

Case 1: Suppose y is the ground truth label:

$$\begin{aligned} L^{(4)}(u, y) &= (1 - u_y + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]})_+ \\ &= \left(1 - (1 + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]} + \epsilon) + \frac{1}{k} \sum_{i=1}^k (u_{\setminus y})_{[i]}\right)_+ \\ &= (-\epsilon)_+ && \text{where } \epsilon > 0 \implies -\epsilon < 0 \\ &= 0 \ . \end{aligned}$$

As u'_y is of the same form of eq. (20) with $\epsilon = 0$, we observe equality as $(\epsilon)_+ = (0)_+ = 0$. Therefore, $L^{(4)}(u, y) = 0 = L^{(4)}(u', y)$, and $L^{(4)}(u, y) \geq L^{(4)}(u', y)$ immediately. **Case 2:** Let $j \neq y$ be the ground truth label.

$$L^{(4)}(u, j) = (1 - u_j + \frac{1}{k} \sum_{i=1}^k (u_{\setminus j})_{[i]})_+ \ .$$

By the case, we have $u_j = u'_j$.

$$\begin{aligned} L^{(4)}(u, j) &= (1 - u_j + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus j})_{[i]})_+ \\ &= (1 - u'_j + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus j})_{[i]})_+ \\ &\geq (1 - u'_j + \frac{1}{k}\sum_{i=1}^{k}(u'_{\setminus j})_{[i]})_+ && \text{as } u \geq u' \text{ element-wise} \\ &= L^{(4)}(u', j)\ . \end{aligned}$$

Therefore, $L^{(4)}(u, j) \geq L^{(4)}(u', j)$. Thus, we conclude

$$L^{(4)}(u, y) \geq L^{(4)}(u', y) \quad \forall y \in \mathcal{Y},$$

and therefore $\mathcal{U}_4 := R_4^{\texttt{high}} \cap R_4^{\texttt{low}}$ is a bounded, infinite, representative set for $L^{(4)}$. □

C.2. Characterizing Affineness of $L^{(4)}$

For $u \in \mathcal{U}_4$, we know that $\forall y \in \mathcal{Y}$

$$u_y \leq 1 + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]} \qquad \text{as } u \in R_4^{\texttt{high}} \tag{21}$$

$$\implies 0 \leq 1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]}\ . \tag{22}$$

Therefore, for all ground truth labels $y \in \mathcal{Y}$ and $u \in \mathcal{U}_4$, we have

$$\begin{aligned} L^{(4)}|_{\mathcal{U}_4}(u, y) &= (1 - u_j + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]})_+ \\ &= 1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]} \end{aligned}$$

is an equivalent way to write the loss $L^{(4)}$ when restricting the domain to $\mathcal{U}_4$. When restricting to $u \in \mathcal{U}_4$, we may denote $L^{(4)}(u, y) = L^{(4)}|_{\mathcal{U}_4}(u, y)$ for brevity and drop the positive part operator.

Now consider a set $T \subseteq \mathcal{Y}$ such that $|T| \leq k$. Let us define the region

$$A_4^T = \left\{ u \in \mathcal{U}_4 \mid \begin{cases} 0 \leq u_y \leq 1 + \frac{1}{k}\sum_{i \in T, i \neq y} u_i & y \in T \\ u_y = 0 & y \notin T \end{cases} \right\}$$

We claim, for any $y \in \mathcal{Y}$, the function $u \mapsto L^{(4)}(u, y)$ is affine on A^T, and note that $A^T \subseteq \mathcal{U}_4$ for all T by construction.

Lemma C.4. *For all $y \in \mathcal{Y}$ and set $T \subseteq \mathcal{Y}$ such that $|T| \leq k$, the function $u \mapsto \mathbb{E}_p L^{(4)}(u, \cdot)$ defined on $\mathcal{U}_4$ is affine on A^T.*

Proof. Fix $y \in \mathcal{Y}$ and $T \subseteq \mathcal{Y}$ such that $|T| \leq k$. Note that for $u \in \mathcal{U}_4$,

$$L^{(4)}(u, y) = 1 - u_y + \frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]}$$

The first two terms of this loss are linear in u; therefore $\frac{1}{k}\sum_{i=1}^{k}(u_{\setminus y})_{[i]}$ is the only term with non-linearity. Moreover, this term results from the ordering of the top k elements of $u_{\setminus y}$. Given that $|T| \leq k$ and all elements of $u \notin T$ are 0, we have that $T \in T_k(u_{\setminus y})$ Therefore, $u \mapsto Li4(u, y)$ will be linear for $u \in A^T$. □

This result yields affine regions over which $u \mapsto L^{(4)}(u, y)$ is affine for each $y \in \mathcal{Y}$. The vertices of these affine regions yield a finite representative set for $L^{(4)}$.

C.3. Constructing a Finite Representative Set for $L^{(4)}$

Each set T has a finite set of vertices according to the two inequalities shown in the definition of A^T above. Since $|T| \leq k$, there are a finite number of possible sets T ($\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + ... + \binom{n}{k} = 2^n$ possible sets in particular). Therefore,

$$\bigcup_{T \subseteq \mathcal{Y},\, |T| \leq k} A^T = \mathcal{U}_4$$

has a finite number of vertices.

According to the boundaries of the halfspaces defining A^T, the vertices of A^T must be such points u such that $u_y = 0$ or $u_y = 1 + \frac{1}{k}\sum_{i \in T, i \neq y} u_i$ for each $y \in \mathcal{Y}$. Consider when $u_y = 1 + \frac{1}{k}\sum_{i \in T, i \neq y} u_i$, which we will refer to as the "bumped up" value of u_y.

Theorem C.5. *Fix $T \subseteq \mathcal{Y}$ such that $1 \leq |T| \leq k$. For a vertex u in the region A^T with $y \in \mathcal{Y}$ such that $u_y = 1 + \frac{1}{k}\sum_{i \in T, i \neq y} u_i$, then $\forall i \in \mathcal{Y}$ such that $u_i \neq 0 \implies u_i = u_y$.*

Proof. Let $u_y = 1 + \frac{1}{k}\sum_{i \in T, i \neq y} u_i$ and $j \in T \implies u_j = 1 + \frac{1}{k}\sum_{i \in T, i \neq j} u_j$. We will show that $u_y = u_j$, so all "bumped up" elements of u must be equal to one another:

$$\begin{aligned}
u_y &= 1 + \frac{1}{k}\sum_{i \in T, i \neq y} u_i \\
u_y &= 1 + \frac{1}{k}\sum_{i \in T, i \neq j} u_i + \frac{1}{k}u_j - \frac{1}{k}u_y \\
\frac{k+1}{k}u_y &= u_j + \frac{1}{k}u_j \\
\frac{k+1}{k}u_y &= \frac{k+1}{k}u_j \\
u_y &= u_j
\end{aligned}$$

Therefore, for any two arbitrary elements $y, j \in T, u_j = u_y$. □

Therefore, the closed for over vertices u of the region A^T are as follows:

$$\begin{aligned}
u_y &= 1 + \frac{1}{k}\sum_{i \in T, i \neq y} u_i \\
u_y &= 1 + \frac{1}{k}\sum_{i \in T} u_i - \frac{1}{k}u_y \\
u_y &= 1 + \frac{1}{k}|T|u_y - \frac{1}{k}u_y \\
u_y &= \frac{k}{k+1-|T|}\,.
\end{aligned}$$

Thus, all of the vertices of each A^T occur at $u \in \mathcal{U}_4$ such that $u_y = 0$ or $\frac{k}{k+1-|T|}$ for all $y \in \mathcal{Y}$ where $|T| \leq k$ is the number of non-zero elements of $u \in A^T$. consider the set of subsets $\mathcal{T} = \{T \subseteq \{1, \ldots, n\} \mid |T| \leq k\}$ and finite report set vertices of the A^T sets by $\mathcal{R}^{(4)} := \{\frac{k}{k+1-|T|}\mathbb{1}_T \mid T \in \mathcal{T}\}$.

C.4. Characterizing the Loss Embedded by $L^{(4)}$

We reparameterize the vertices of the A^T sets by their defining set T by the bijection $\Phi : T \mapsto \frac{k}{k+1-|T|}\mathbb{1}_T$. We define the reparameterization $\hat{\ell}_4$ such that $L^{(4)}(\Phi(T), y) = L^{(4)}|_{\mathcal{U}_4}(\Phi(T), y) = \hat{\ell}_4(T, y)$ for all $u \in \mathcal{T}$.

We know that $L^{(4)}$ embeds $L^{(4)}|_{\mathcal{U}_4}$, and therefore also embeds

$$\hat{\ell}_4(T,y) = \begin{cases} 1 - \frac{k}{k+1-|T|} + \frac{1}{k}(|T|-1)\frac{k}{k+1-|T|} & y \in T \\ 1 + \frac{1}{k}|T|\frac{k}{k+1-|T|} & y \notin T \end{cases} = \begin{cases} 0 & y \in T \\ \frac{k+1}{k+1-|T|} & y \notin T \end{cases} . \tag{23}$$

In a slight abuse of notation, for a set $T \subset [n]$ and $p \in \Delta_{\mathcal{Y}}$, let $\sigma_T(p) = \sum_{i \in T} p_i$. Therefore, the expected value of $\hat{\ell}_4$ is

$$\mathbb{E}_p \hat{\ell}_4(T,\cdot) = \sum_{y \in T} p_y(0) + \sum_{y \notin T} p_y(\frac{k+1}{k+1-|T|}) = (1-\sigma_T(p))(\frac{k+1}{k+1-|T|}) .$$

Now, suppose we have some set $T \in \mathcal{T}$ as defined above; we will analyze $\text{prop}[\hat{\ell}_4]$ to determine the necessary probability p_i some in $i \in \mathcal{Y}$ so that $i \in T \in \text{prop}[\hat{\ell}_4](p)$. In other words, if we have some set T of labels corresponding to scores in u of $\frac{k+1}{k+1-|T|}$, then we will bump the score of some label, $z \notin T$, up to $\frac{k+1}{k+1-|T|-1}$ (changing all of the non-zero scores in u to this value as well) if it surpasses a particular probability threshold. We will find this probability boundary below by seeing what probability $z \in \mathcal{Y}$ must achieve in order to meet or lower the expected loss:

$$\mathbb{E}_p \hat{\ell}_4(T,\cdot) \geq \mathbb{E}_p \hat{\ell}_4(T \cup \{z\},\cdot) .$$

By doing so, we are determining the probability of p_z such that, for a fixed $p \in \Delta_{\mathcal{Y}}$, we have $T \in \text{prop}[\hat{\ell}_4](p) \implies T \cup \{z\} \in \text{prop}[\hat{\ell}_4](p)$.
This boundary is given:

$$\begin{aligned} \mathbb{E}_p \hat{\ell}_4(T,\cdot) &= \mathbb{E}_p \hat{\ell}_4(T \cup \{z\},\cdot) \\ (1-\sigma_T(p))\frac{k+1}{k+1-|T|} &= (1-\sigma_T(p)-p_z)\frac{k+1}{k+1-|T|-1} \\ (k-|T|)(1-\sigma_T(p)) &= (k+1-|T|)(1-\sigma_T(p)-p_z) \\ 0 &= 1-\sigma_T(p)-p_z(k+1-|T|) \\ p_z &= \frac{1-\sigma_T(p)}{k+1-|T|} \end{aligned} \tag{24}$$

Therefore, to add the element z to the set T

$$p_z \geq \frac{1-\sigma_T(p)}{k+1-|T|}.$$

Iteratively adding elements such that the above boundary holds will be necessary and sufficient to form an optimal set $M^* \subset [n]$ of labels that minimizes $\mathbb{E}_p \hat{\ell}_4(M^*,\cdot)$, and equivalently, $\Phi(M^*)$ minimizes $L^{(4)}$.

Theorem C.6. *Consider $\gamma_4 := \text{prop}[\hat{\ell}_4]$. Fix $p \in \Delta_{\mathcal{Y}}$ and $T \in \mathcal{T}$ be such that T is the top-$|T|$ elements of p with $|T| \leq k-1$. Consider $z \in [n] \setminus T$ such that $p_z \geq \frac{1-\sigma_T(p)}{k+1-|T|}$ and $p_i \leq p_z$ for all $i \in [n] \setminus T$. Then z must be an element of M^* for some $M^* \in \gamma_4(p)$.*

Proof. For intuition, T is a set of labels at least as likely as label z. Observe that there is an $M \in \gamma_4(p)$ such that $T \subseteq M$ since T is composed of the top-$|T|$ elements of p, and replacing any $t \in T$ with $t' \notin T$ cannot decrease expected loss as the denominator stays the same and T is composed of the top-$|T|$ elements of p.

It is not necessarily the case that $M = M^*$ as we may have $|\gamma_4(p)| > 1$ and the top-k elements of the property value are ambiguous. If $|\gamma_4(p)| = 1$, however, then we must have $M = M^*$. Suppose $z \notin M$ (otherwise this proof is trivial), we

have two cases:
Case 1: $T \subsetneq M$, e.g., $\exists z' \in M$ such that $z' \notin T$.

$$\begin{aligned}\mathbb{E}_p \hat{\ell}_4(M, \cdot) &= (k+1)\frac{1 - p_{z'} - \sigma_{M\setminus\{z'\}}(p)}{k+1-|M|} \\ &\geq (k+1)\frac{1 - p_z - \sigma_{M\setminus\{z'\}}(p)}{k+1-|M|} \quad \text{from } p'_z \leq p_z \ .\end{aligned}$$

Therefore, $\exists M^*$ which is optimal such that $z \in M^* = M\setminus\{z'\} \cup \{z\}$.
Case 2: $T = M$. By the assumptions and choice of z,

$$p_z \geq \frac{1 - \sigma_T(p)}{k+1-|T|} \ .$$

Therefore, using the bound from eq. (24), we have

$$\mathbb{E}_p \hat{\ell}_4(T \cup \{z\}, \cdot) \leq \mathbb{E}_p \hat{\ell}_4(T, \cdot) = \mathbb{E}_p \hat{\ell}_4(M, \cdot) \ .$$

As M is optimal, $M^* = T \cup \{z\}$ is also optimal.

From both cases above, we can conclude $z \in M^*$ for some optimal set M^*. □

This will enable us to characterize γ_4 in Theorem C.8. However, we first need the following Lemma.

Lemma C.7. *For $a, c \in \mathbb{R}_+$ and $b, d \in \mathbb{R}_{++}$ with $b > d$,*

$$\frac{c}{d} < \frac{a}{b} \implies \frac{a+c}{b+d} < \frac{a}{b}.$$

Proof.

$$\begin{aligned}&\frac{c}{d} < \frac{a}{b} \\ \iff\ & cb < ad \\ \iff\ & ab + cb < ab + ad \\ \iff\ & \frac{a+c}{b+d} < \frac{a}{b}\end{aligned}$$

□

Now we obtain the following result to characterize γ_4.

Theorem C.8. *Fix $p \in \Delta_\mathcal{Y}$, and consider any $T \subset [n]$ which minimizes $\mathbb{E}_p \hat{\ell}_4(T, \cdot)$, i.e., $T \in \gamma_4(p)$. Then for all $z \in T$, we have $p_z \geq \frac{1-\sigma_T(p)}{k+1-|T|}$.*

Proof. We will show the contrapositive. For $T \in \gamma_4(p)$. Suppose there was a $z \in T$ such that $p_z < \frac{1-\sigma_T(p)}{k+1-T}$. We will contradict optimality of T by showing $\mathbb{E}_p \hat{\ell}_4(T \setminus \{z\}, \cdot) < \mathbb{E}_p \hat{\ell}_4(T, \cdot)$. Denote $M := T \setminus \{z\}$.

Note, that if we let $c = p_z \in \mathbb{R}_+$, $d = 1 \in \mathbb{R}_{++}$, $a = (1 - \sigma_M(p)) \in \mathbb{R}_+$, and $b = (k + 1 - |T|) \in \mathbb{R}_{++}$, then we have $p_z < \frac{1-\sigma_T(p)}{k+1-|T|} \iff \frac{c}{d} < \frac{a}{b}$. Thus, we can apply Lemma C.7 to observe

$$\begin{aligned}&\frac{a+c}{b+d} < \frac{a}{b} \\ &\frac{1 - \sigma_T(p) + p_z}{k - (|T|-1) + 1} < \frac{1 - \sigma_T(p)}{k+1-|T|} \\ \implies\ & (k+1)\frac{1 - \sigma_M(p)}{k - |M| + 1} < (k+1)\frac{1 - \sigma_T(p)}{k+1-|T|} \\ \implies\ & \mathbb{E}_p \hat{\ell}_4(M, \cdot) < \mathbb{E}_p \hat{\ell}_4(T, \cdot)\end{aligned}$$

Therefore, the expected loss on $M \subset [n]$ is strictly lower than on T; thus, $T \notin \gamma_4(p)$. Thus for any $T \in \gamma_4(p)$, we must have $p_z \geq \frac{1-\sigma_T(p)}{k+1-|T|}$ for all $z \in T$. □

By Theorem C.8, we can conclude that iteratively adding elements $z \in [n]$ (in increasing order of corresponding probability) such that

$$p_z \geq \frac{1-\sigma_T(p)}{k+1-|T|}$$

to a set $T \subseteq [n]$, that is initially the empty set, is necessary and sufficient to form the optimal set $M^* \subset [n]$ that minimizes $\mathbb{E}_p \hat{\ell}_4(M^*, \cdot)$. That is, $\text{prop}[L^{(4)}|_{\mathcal{U}_4}]$ can be computed by implementing a greedy algorithm.

C.5. A sketch of $\text{prop}[L^{(4)}]$

Let $T \subset [n]$ where the elements of T have been iteratively added in decreasing order of probability so long as $|T| \leq k$ and the probability of the added item meets the boundary condition defined above. Suppose $|T| \leq k-1$, then from our derivation of the probability needed to add an element z to T, we can rewrite the boundary condition as adding an element $u_{[j]}$ with $j \in \{1, 2, ..., k\}$, so long as

$$p_{[j]} \geq \frac{1-\sum_{i=1}^{j-1} p_{[i]}}{k+2-j}.$$

We can rewrite rewrite the above as

$$(k+1-j)p_{[j]} \geq 1-\sigma_j(p)\ . \tag{25}$$

Let $j_1 \in [n]$ be the largest j such that

$$(k+1-j)p_{[j]} > 1-\sigma_j(p)\ . \tag{26}$$

Let $j_2 \in [n]$ be the largest j such that

$$(k+1-j)p_{[j]} \geq 1-\sigma_j(p).$$

Lemma C.9. *For all $p \in \Delta_\mathcal{Y}$ and j_1 as in eq. (26), we have $j_1 \leq k$.*

Proof. Suppose for the sake of contradiction that $j_1 > k$, and therefore $j_1 \geq k+1$. Then, we have

$$(k+1-j_1)p_{[j_1]} \leq 0$$

By definition of j_1,

$$\begin{aligned}(k+1-j_1)p_{[j_1]} > 1-\sigma_{j_1}(p) \implies 0 &> 1-\sigma_{j_1}(p)\\ \sigma_{j_1}(p) &> 1\ .\end{aligned}$$

However, this contradicts that $\sigma_{j_1}(p) \leq \sum_{i=1}^n p_{[i]} = 1$, as $p \in \Delta_\mathcal{Y}$. Thus, we conclude that $j_1 \leq k$. □

Note that if for some $j \in [n]$,

$$(k+1-j)p_{[j]} = 1-\sigma_j(p),$$

then the expected loss will not change by "bumping up" the corresponding element in u. Therefore, we are indifferent to "bumping up" this element or not.

From the above definitions define two sets $H : \Delta_\mathcal{Y} \to 2^{[n]}$ and $I : \Delta_\mathcal{Y} \to 2^{[n]}$ as follows:

$$\begin{aligned}H(p) &= \left\{i \in [n] \mid p_i \geq p_{[j_1]}\right\}\\ I(p) &= \left\{i \in [n] \mid p_{[j_2]} \leq p_i < p_{[j_1]}\right\}\end{aligned}$$

Note, that $T = H(p) \cup I(p)$ is a minimizing set of indices for $\mathbb{E}_p \hat{\ell}_4(T, \cdot)$ when we "bump up" exactly those corresponding elements in $H(p)$. If p is understood from context, then we simply denote $H(p) = H$, etc.

Intuitively, H ("high") is the set of elements that bumping up (including in the report set T) will result in a lower expected loss. I ("indifferent") is the set of elements that bumping up will not affect expected loss, meaning we are indifferent to bumping them up.

From these definitions, we can see that the set of all $H \cup I^*$ where $I^* \in P(I)$ (the power set 2^I) such that $|I^*| \leq k - j_1$, will have an expected loss equal to the expected loss associated with the set $H \cup I$. And $H \cup I$ is the exact set constructed by iteratively adding elements according to the boundary condition defined above (and we established above that this is the strategy for forming an optimal report set when $|H \cup I| \leq k$). Therefore, the set of all $H \cup I^*$ where $I^* \in P(I)$ such that $|I^*| \leq k - j_1$, will be representative. In particular, there is an $I^* \in P(I)$ (e.g., $I^* = \emptyset$) such that $|H \cup I^*| \leq k$, so that $\Psi(T) \in \mathcal{R}^{(4)}$, where $T = H \cup I^*$.

We can conclude that the property elicited by $\hat{\ell}_4$ is given

$$\operatorname{prop}[L^{(4)}](p) = \left\{ \frac{k+1}{k+1-|T|} \mathbb{1}_T \mid T = H \cup I^*, I^* \in P(I), |I^*| \leq k - j_1 \right\} ,$$

where $P(I)$ is the power set of set I, and H and I are functions of p.

C.6. Characterizing Consistency of $L^{(4)}$

From this, we can conclude that $\hat{\ell}_4$ indirectly elicits top-k when $j_1 = k$ because in all other cases $\operatorname{prop}[\hat{\ell}_4]$ will return a set with cardinality greater than 1 which will require the breaking of ties. This breaking of ties is dependent on the link utilized, which in this case is the $\arg\max$; however, as established we would be breaking ties between sets that result in the same expected loss of $L^{(4)}$. This means that we would be breaking ties arbitrarily. The only case in which this does not occur is when we are not indifferent between bumping up any elements u_i, u_j where $i, j \in [n], i \neq j$. This occurs when $j_1 = k$, resulting in

$$\begin{aligned} (k+1-k)p_{[k]} &> 1 - \sigma_k \qquad \text{by definition of } j_1 \\ p_{[k]} &> 1 - \sigma_k \end{aligned}$$

Therefore by Lemma A.10, we know that $L^{(4)}$ is guaranteed consistency with top-k when $p_{[k]} > 1 - \sigma_k$.

Corollary 3.3. *$L^{(4)}$ is consistent with respect to ℓ_k on $\mathcal{P}^{(4)} := \{p \in \Delta_\mathcal{Y} \mid p_{[k]} > 1 - \sigma_k(p)\}$.*

D. Additional Derivations for L_k

D.1. Proof of Lemma 4.3

As L_k is a proper polyhedral function, we know that it attains its infimum (Rockafellar, 1997, Corollary 19.3.1), and thus Γ is well-defined on $\Delta_\mathcal{Y}$. Let $G(u) = (-\underline{\ell_k})^*(u)$ and $I_{\Delta_\mathcal{Y}}$ be the convex indicator function that is 0 on $\Delta_\mathcal{Y}$ and ∞ on $\mathbb{R}^n \setminus \Delta_\mathcal{Y}$. Then, $G^*(p) = -\underline{\ell_k}(p) = \sigma_k(p) + I_{\Delta_\mathcal{Y}}(p) - 1$.

Lemma D.1. $\Gamma(p) = \partial G^*(p)$.

Proof. As L_k is a proper convex function, Rockafellar (1997, Theorem 23.5) yields

$$\begin{aligned} u \in \partial G^*(p) &\iff G(u) + G^*(p) = \langle u, p \rangle && \text{Rockafellar (1997, Theorem 23.5)} \\ &\iff \langle p, L_k(u, \cdot) \rangle = -G^*(p) && \text{Finocchiaro et al. (2022, Theorem 4)} \\ &\iff \langle p, L_k(u, \cdot) \rangle = \underline{L_k}(p) && \text{Finocchiaro et al. (2022, Theorem 4)} \\ &\iff u \in \arg\min_{u'} \langle p, L_k(u', \cdot) \rangle = \Gamma(p) . \end{aligned}$$

□

Therefore, we just need to characterize the subgradients of G^*. As L_k is polyhedral, we know that is is the pointwise maximum of a finite number of affine (and therefore convex) functions. This enables us to use a result from Hiriart-Urruty and Lemaréchal (2012) to rewrite the subdifferential of G^* in order to characterize $\Gamma(p)$ for all $p \in \operatorname{relint}(\Delta_\mathcal{Y})$.

Theorem D.2 (Hiriart-Urruty and Lemaréchal (2012)[D.4.3.2])**.** *Let $f_1, \ldots f_m$ be convex functions from $\mathbb{R}^n \to \mathbb{R}$. Then,*

$$\partial \left(\max_i f_i(x) \right) = \operatorname{hull} \left\{ \cup_i \partial f_i(x) \mid i \in \arg\max_j f_j(x) \right\} .$$

Lemma D.3. *For all $p \in \Delta_\mathcal{Y}$, we have $\partial\sigma_k(p) = \text{hull}\{\tau_k(p)\}$.*

Proof. Let $f_t(p) = \langle t, p\rangle$ for each $t \in \mathcal{T}$. By affineness, $\partial f_t(p) = \{t\}$. Now, recalling the definition of σ_k, we can write

$$\begin{aligned}
\partial\sigma_k(p) &= \partial\left(\max_{t\in\mathcal{T}} \langle t, p\rangle\right) \\
&= \partial\left(\max_{t\in\mathcal{T}} f_t(p)\right) \\
&= \text{hull}\left\{\cup_t \left(\partial f_t(p)\right) \mid t \in \arg\max_{t'} f_{t'}(p)\right\} && \text{Theorem D.2} \\
&= \text{hull}\left\{\cup_t \{t\} \mid t \in \arg\max_{t'} f_{t'}(p)\right\} \\
&= \text{hull}\left\{\arg\max_t f_t(p)\right\} \\
&= \text{hull}\left\{\arg\max_t \langle t, p\rangle\right\} \\
&= \text{hull}\left\{\tau_k(p)\right\} \ .
\end{aligned}$$

□

Lemma D.4. *For all p on the relative boundary of $\Delta_\mathcal{Y}$, (that is, $\Delta_\mathcal{Y} \setminus \text{relint}(\Delta_\mathcal{Y})$),*

$$\partial I_{\Delta_\mathcal{Y}}(p) = \cup_{\alpha\in\mathcal{R}}\{\alpha\mathbb{1}\} - \text{cone}\{\mathbb{1}_i | p_i = 0\} \ .$$

Moreover, $\partial I_{\Delta_\mathcal{Y}}(p) = \vec{0}$ for all $p \in \text{relint}(\Delta_\mathcal{Y})$.

Proof. We can define the simplex as the set of points $p \in \mathbb{R}^n$ that satisfies the constraints $\langle p, \mathbb{1}\rangle = 1$, and $\langle p, \mathbb{1}_i\rangle \geq 0$ for all $1 \leq i \leq n$. Let I_0 be the convex indicator of the first constraint, such that $I_0(p) = 0$ when $\langle p, \mathbb{1}\rangle = 1$, and $I_0(p) = \infty$ otherwise. Similarly, let I_i be the convex indicators such that $I_i(p) = 0$ if $\langle p, \mathbb{1}_i\rangle \geq 0$, and $I_i(p) = \infty$ otherwise. A point $p \in \mathbb{R}^n$ will be in $\Delta_\mathcal{Y}$ precisely when all $n+1$ constraints are satisfied, which is exactly when all the indicators are 0. Therefore, we can rewrite

$$I_{\Delta_\mathcal{Y}}(p) = \sum_{i=0}^{n} I_i(p) \ .$$

For any $p \in \Delta_\mathcal{Y}$, we have $\partial I_0(p) = \{\alpha_0\mathbb{1} | \alpha_0 \in \mathbb{R}\}$. For $1 \leq i \leq n$, $\partial I_i(p) = \{\vec{0}\}$ if $p_i > 0$, and $\partial I_i(p) = \{-\alpha_i\mathbb{1}_i \mid \alpha_i > 0\}$ if $p_i = 0$.

The subgradient of a sum of convex functions is the Minkowski sum of their individual subgradients (Rockafellar, 1997, Theorem 23.8). Now, we observe,

$$\begin{aligned}
\partial I_{\Delta_\mathcal{Y}}(p) &= \partial\left(\sum_{i=0}^{n} I_i(p)\right) \\
&= \sum_{i=0}^{n} \partial I_i(p) \\
&= \partial I_0(p) + \sum_{i=1}^{n} \partial I_i(p) \\
&= \cup_{\alpha_0\in\mathbb{R}}\{\alpha_0\mathbb{1}\} + \sum_{i=1}^{n}\{-\alpha_i\mathbb{1}_i | p_i = 0, \alpha_i \geq 0\} \\
&= \cup_{\alpha\in\mathcal{R}}\{\alpha\mathbb{1}\} - \text{cone}\{\mathbb{1}_i | p_i = 0\} \ .
\end{aligned}$$

□

Lemma D.5. $\Gamma(p) = \text{hull}\{\tau_k(p)\} - \text{cone}\{\mathbb{1}_i | p_i = 0\} + \cup_{\alpha\in\mathcal{R}}\{\alpha\mathbb{1}\}$.

Proof.

$$\begin{aligned}
\Gamma(p) &= \partial G^*(p) && \text{Lemma D.1}\\
&= \partial\left(\sigma_k(p) + I_{\Delta_\mathcal{Y}}(p) - 1\right)\\
&= \partial\sigma_k(p) + \partial I_{\Delta_\mathcal{Y}}(p) - 0\\
&= \text{hull}\{\tau_k(p)\} - \text{cone}\{\mathbb{1}_i | p_i = 0\} + \cup_{\alpha\in\mathcal{R}}\{\alpha\mathbb{1}\}\,. && \text{Lemma D.3 and Lemma D.4}
\end{aligned}$$

□

D.2. Equivalence of Equations 10 and 11

Lemma D.6.

$$L_k(u, y) = \max_{1\le m\le n}\left\{\tfrac{\sigma_m(u)}{m} + \left(1 - \tfrac{k}{m}\right)_+\right\} - u_y\,.$$

Proof. By Equation 10,

$$L_k(u, y) = \sup_{p\in\Delta_\mathcal{Y}}\left(\langle p, u\rangle - \sigma_k(p)\right) + 1 - u_y\,.$$

Without loss of generality, we may assume u is sorted. Since $\sigma_k(p)$ is not order dependent, and $\langle p, u\rangle$ will be maximized when the elements of p have the same ordering as the elements of u, we can assume p is sorted as well. Let $\text{sort}(\Delta_\mathcal{Y})$ denote the subset of vectors $p \in \Delta_\mathcal{Y}$ that are sorted. The loss then simplifies to

$$L_k(u, y) = \sup_{p\in\text{sort}(\Delta_\mathcal{Y})}\left(\sum_{i=1}^{k} p_i(u_i - 1) + \sum_{i=k+1}^{n} p_i u_i\right) + 1 - u_y\,.$$

Let v be the vector such that for $i \le k$, $v_i = u_i - 1$, and for $i > k$, $v_i = u_i$. We can then reduce to

$$\begin{aligned}
&= \sup_{p\in\text{sort}(\Delta_\mathcal{Y})}\left(\sum_{i=1}^{n} p_i v_i\right) + 1 - u_y\\
&= \sup_{p\in\text{sort}(\Delta_\mathcal{Y})}\langle p, v\rangle + 1 - u_y\,.
\end{aligned}$$

We claim that, for any fixed v, there exists a $p \in \arg\sup_{p'\in\Delta_\mathcal{Y}}\langle p', v\rangle$ such that $p = \mathbb{1}_M/|M|$ for some set $M \subseteq [n]$.

We proceed by contradiction. Assume that there is no p that is exactly $\frac{1}{m}$ on m indices that achieves the supremum. Let $U = \arg\sup_{p\in\Delta_\mathcal{Y}}\langle p, v\rangle \subseteq \Delta_\mathcal{Y}$ be the set of (sorted) distributions that do achieve the supremum. Since $\Delta_\mathcal{Y}$ is compact and $\langle p, v\rangle$ is linear, U is nonempty. By assumption, for every $q \in U$, there must be some index m such that $q_1 = q_m > q_{m+1} > 0$. Choose any q with the maximal such m. Let $\mu = \frac{1}{m}\sum_{i=1}^{m} v_i$ be the average of the first m elements of v. Then, we have

$$\langle q, v\rangle = q_m\mu + \sum_{i=m+1}^{n} q_i v_i\,.$$

If $m\mu > \sum_{i>m} v_i q_i$, we can choose a sufficiently small $\epsilon > 0$ and set $q'_i = q_1 - \epsilon\frac{1-m\mu}{m}$ for $i \le m$ and $q'_i = (1+\epsilon)q_i$ for $i > m$ to get a new distribution $q' \in \Delta_\mathcal{Y}$. Using this q' instead of q increases $\langle q, v\rangle$, so $q \notin U$, a contradiction. If instead $m\mu < \sum_{i>m} v_i q_i$, we can instead choose a sufficiently large $\epsilon < 0$, and achieve the same result. If instead $m\mu = \sum_{i>m} v_i q_i$, we can choose ϵ such that $q'_m = q'_{m+1}$, so we did not choose the q with the maximal m, also a contradiction.

Therefore, there is some sorted p that is $\frac{1}{m}$ on exactly m indicies that achieves the supremum. We therefore need only consider this set of distributions. Plugging this into the original equation, we get

$$
\begin{aligned}
L_k(u, y) &= \sup_{p \in \Delta_{\mathcal{Y}}} \left(\sum_{i=1}^{k} p_i (u_i - 1) + \sum_{i=k+1}^{n} p_i u_i \right) + 1 - u_y \\
&= \max_{1 \leq m \leq n} \left(\sum_{i=1}^{k} \frac{1}{m} (u_i - 1) \right) + 1 - u_y \\
&= \max_{1 \leq m \leq n} \left\{ \tfrac{\sigma_m(u)}{m} + \left(1 - \tfrac{k}{m}\right)_+ \right\} - u_y \, .
\end{aligned}
$$

□