Surrogate Regret Bounds for Polyhedral Losses

Rafael Frongillo
University of Colorado Boulder
raf@colorado.edu

Bo Waggoner
University of Colorado Boulder
bwag@colorado.edu

Abstract

Surrogate risk minimization is an ubiquitous paradigm in supervised machine learning, wherein a target problem is solved by minimizing a surrogate loss on a dataset. Surrogate regret bounds, also called excess risk bounds, are a common tool to prove generalization rates for surrogate risk minimization. While surrogate regret bounds have been developed for certain classes of loss functions, such as proper losses, general results are relatively sparse. We provide two general results. The first gives a linear surrogate regret bound for any polyhedral (piecewise-linear and convex) surrogate, meaning that surrogate generalization rates translate directly to target rates. The second shows that for sufficiently non-polyhedral surrogates, the regret bound is a square root, meaning fast surrogate generalization rates translate to slow rates for the target. Together, these results suggest polyhedral surrogates are optimal in many cases.

1 Introduction

In supervised learning, our goal is to learn a hypothesis that solves some target problem given labeled data. The problem is often specified by a target loss ℓ that is discrete in the sense that the prediction lies in a finite set, as with classification, ranking, and structured prediction. As minimizing the target ℓ over a data set is typically NP-hard, we instead turn to the general paradigm of *surrogate risk minimization*. Here a surrogate loss function L is chosen and minimized instead of the target ℓ. A link function ψ then maps the surrogate predictions to target predictions.

Naturally, a central question in surrogate risk minimization is how to choose the surrogate L. A minimal requirement is statistical *consistency*, which roughly says that approaching the best possible surrogate loss will eventually lead to the best possible target loss. A more precise goal is to understand the rate at which this target minimization occurs, called the *generalization rate*: how quickly the target loss approaches the best possible, in terms of the number of data points. In other words, if we define the *regret* of a hypothesis to be its performance relative to the Bayes optimal hypothesis, we wish to know how quickly the ℓ-regret approaches zero.

Surrogate regret bounds are a common tool to prove generalization rates for surrogate risk minimization, by bounding the target regret as a function of the surrogate regret. This bound allows generalization rates for surrogate losses—of which there are many, often leveraging properties of L like strong convexity and smoothness—to imply rates for the target problem. Roughly speaking, a surrogate regret bound takes the form

$$\forall \text{ surrogate hypotheses } h, \quad \mathrm{Regret}_\ell(\psi \circ h) \leq \zeta(\,\mathrm{Regret}_L(h)\,)\,, \tag{1}$$

where the *regret transfer function* $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$ controls how surrogate regret translates to target regret. For example, if we have a sequence of surrogate hypotheses h_n achieving a fast rate $\mathrm{Regret}_L(h_n) = O(1/n)$, and we have a regret transfer of $\zeta(\epsilon) = \sqrt{\epsilon}$, then we can only guarantee a slower target rate of $\mathrm{Regret}_\ell(\psi \circ h_n) = O(1/\sqrt{n})$. The ideal regret transfer is *linear*, i.e. $\zeta(\epsilon) = C \cdot \epsilon$ for some $C > 0$, which would have directly translated the fast-rate guarantee to the target.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Much is known about surrogate regret bounds for particular problems, especially for binary classification and related problems like bipartite ranking; see § 1.1 for a discussion of the literature. General results which span multiple target problems, however, are sparse. Furthermore, we still lack guidance as to which surrogate will lead to the best overall target generalization rate, as this rate depends on both the surrogate generalization rate and the regret transfer function. For example, consider the choice between a *polyhedral* (piecewise-linear and convex) surrogate, like hinge loss, or a smooth surrogate like exponential or squared loss. The surrogate rate for polyhedral losses will likely be a slower $O(1/\sqrt{n})$, whereas a faster $O(1/n)$ rate is typical for smooth surrogates—but what is the tradeoff when it comes to the regret transfer function ζ, which ultimately decides the target generalization rate? To answer these types of questions, we need more general results.

Our first result is a general surrogate regret bound for polyhedral surrogates. Perhaps surprisingly, we show that *every* consistent surrogate in this class achieves a linear regret transfer, meaning their surrogate generalization rates translate directly to target rates, up to a constant factor.

Theorem 1 (Linear bound for polyhedral)**.** *Let L be any polyhedral surrogate and ψ any link function such that (L, ψ) is consistent for the target loss ℓ. Then L satisfies Equation 1 with a linear ζ.*

The recent embedding framework of Finocchiaro et al. [10] shows how to construct consistent polyhedral surrogates for any discrete target loss; Theorem 1 additionally endows these surrogates with linear regret bounds. The same framework shows that every polyhedral loss is a consistent surrogate for some discrete target. Theorem 1 therefore immediately applies to the Weston–Watkins surrogate [31], Lovász hinge [10, 33], and other polyhedral surrogates in the literature currently lacking regret bounds.[1]

We also give a partial converse: If the surrogate is sufficiently "non-polyhedral", i.e. smooth and locally strongly convex, then ζ cannot shrink faster than a square root. For these losses, therefore, even fast surrogate generalization rates would likely translate to slow rates for the target.

Theorem 2 (Lower bound for "non-polyhedral")**.** *Let L be a locally strongly convex surrogate with locally Lipschitz gradient. If L satisfies Equation 1 for a target loss ℓ, then there exists $c > 0$ such that, for all small enough $\epsilon \geq 0$, we have $\zeta(\epsilon) \geq c \cdot \sqrt{\epsilon}$.*

While it may seem intuitive that $\zeta = \sqrt{\cdot}$ for strongly convex losses, it is well-known that losses like exponential loss for binary classification also have $\zeta = \sqrt{\cdot}$, even though they are neither strongly convex nor strongly smooth [4]. Theorem 2 offers an explanation for this phenomenon, and together with Theorem 1 clarifies an intriguing trend in the literature: ζ is typically either linear or square-root.

Taken together, our results shed light on the dichotomy expressed above: slower polyhedral surrogate rates of $O(1/\sqrt{n})$ will translate to the target unchanged, while most smooth surrogates, even if achieving a fast surrogate generalization rate of $O(1/n)$, will likely yield the same target rate of $O(1/\sqrt{n})$. Our results therefore suggest an inherent tradeoff fast surrogate rates and optimal regret transfer, a phenomenon also observed by Mahdavi et al. [17] for binary classification. In particular, our results suggest that polyhedral losses may achieve an overall generalization rate to the target problem with the optimal dependence on the number of data points.

The proof of our upper bound, Theorem 1, is nonconstructive. After proving the main results (Sections 3 and 4), we make it constructive in Section 5. Specifically, the regret transfer function for a polyhedral loss can be bounded by $\zeta(t) \leq \alpha \cdot t$ where the constant is of the form $\alpha = \alpha_\ell \cdot \alpha_L \cdot \alpha_\psi$. Each component of α depends only on its respective object ℓ, L, and ψ. A corollary of the derivation is that a polyhedral loss L and link ψ are consistent for ℓ if and only if ψ is ϵ-*separated*, a condition introduced in Finocchiaro et al. [10].

1.1 Literature on surrogate regret bounds

Surrogate regret bounds were perhaps developed first by Zhang [37], as a method to prove statistical consistency for various surrogate losses for binary classification. These results were then generalized by Bartlett et al. [4] for margin losses, and Reid and Williamson [26] for composite proper losses.

[1] Many of these polyhedral surrogates mentioned are actually known to be *inconsistent* for their intended target problems. Nonetheless, the embedding framework of Finocchiaro et al. [10] shows that any polyhedral surrogate is consistent for *some* target problem, namely the one that it embeds. It is to this embedded target loss function, rather than the intended target, that our linear regret bounds apply.

These latter works and others give a fairly complete picture for binary classification: a characterization of calibrated surrogates, and precise closed forms for the optimal ζ for each. By extension, close variations on binary classification are also well-understood, such as cost-sensitive binary classification [28], binary classification with an abstain/reject option [3, 8, 34], bipartite ranking [1, 13, 18], and ordinal regression via reduction to binary classification [22]. For multiclass classification, we have nice consistency characterizations with respect to 0-1 loss [30], with some surrogate regret bounds [9, 23].

Beyond binary classification and 0-1 loss, our knowledge drops off precipitously, with few general results. There are some works which provide partial understand for specific target problems. For example, in the problem of multiclass classification with an abstain option, Ramaswamy et al. [25] give three consistent polyhedral surrogates (two of which are commonly used as inconsistent surrogates for 0-1 loss), along with linear surrogate regret bounds. (See also Ni et al. [19, Theorems 7 & 8] and Charoenphakdee et al. [6, Theorem 8] for other surrogate regret bounds for the same problem.) Ramaswamy et al. [24] show how to extend these surrogates to nested, hierarchical versions, again providing linear regret bounds. Motivated by these polyhedral surrogates, together with examples arising in top-k prediction [16, 32] and other structured prediction problems [33], Finocchiaro et al. [10] introduce an embedding framework for polyhedral surrogates. This framework shows that all polyhedral surrogates are consistent for some discrete target problem, and vice versa, but do not provide surrogate regret bounds. Our results fill this gap, giving linear regret bounds for all polyhedral losses, and thus all discrete target problems.

Perhaps closest in spirit to our work are several recent works on general structured prediction problems. Ciliberto et al. [7, Theorem 2] and Osokin et al. [21, Theorem 7] give surrogate regret bounds for a quadratic surrogate, for a general class of structed prediction problems. Blondel [5, Proposition 5] gives surrogate regret bounds for a tighter class of quadratic surrogates based on projection oracles. Nowak-Vila et al. [20, Theorem 4.4] give even more general bounds, using a broader class of surrogates based on estimating a vector-valued expectation. All of these works apply to a broad range of discrete target problems. The main difference to our work is that their upper bounds are all for strongly convex surrogates,[2] whereas ours are for polyhedral surrogates. Osokin et al. [21, Theorem 8] and Nowak-Vila et al. [20, Theorem 4.5] also provide lower bounds, in the same spirit as our Theorem 2 but only for the classes of surrogates they study. In particular, our results recover their bounds up to a constant factor.

Finally, in the binary classification setting, some works have used surrogate regret bounds to study the relationship between generalization *rates* for the surrogate and target, as discussed above. Zhang et al. [36] give detailed results about how surrogate regret bounds actually convert the surrogate rate to the target rate, by carefully analyzing the behavior of the regret tranfer function near 0. They argue that hinge loss achieves strictly better rates than smooth surrogates like logistic and exponential loss; notably, though, this claim implicitly assumes that generalization rate is the same for all surrogates under consideration. By contrast, Mahdavi et al. [17], as we do, consider that the surrogate generalization rate may *depend* on the choice of surrogate. They study a particular family of smoothed hinge losses, and show a direct tradeoff: smoother losses yield faster surrogate generalization rates, but "slower" regret transfers. Without placing further assumptions on the data, therefore, Mahdavi et al. [17] observe the same general phenomenon as we do: while smoothness may help surrogate generalization rates, it can hurt the regret transfer just as much. In particular, without further assumptions, polyhedral surrogates appear to achieve optimal overall generalization rates.

2 Preliminaries

Target problems. We consider supervised machine learning problems where data points have features in $\mathcal{X}$ and labels in $\mathcal{Y}$, a finite set. The *target problem* is to learn a hypothesis mapping $\mathcal{X}$ to a *report* (or *prediction*) *space* $\mathcal{R}$, possibly different from $\mathcal{Y}$. For example, binary classification has $\mathcal{Y} = \mathcal{R} = \{-1, +1\}$, whereas for ranking problems $\mathcal{R}$ may be the set of permutations of $\mathcal{Y}$. In this paper, we assume $\mathcal{R}$ is a finite set. The target problem is specified by the *target loss* $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$, which maps a report (prediction) $r \in \mathcal{R}$ to the vector $\ell(r)$ of nonnegative loss values for each label,

[2] These works express surrogate regret bounds with ζ^{-1} appearing on the left-hand side of eq. (1), and thus the terms "upper bound" and "lower bound" are reversed from their terminology.

i.e. $\ell(r) = (\ell(r)_y)_{y\in\mathcal{Y}}$. Given a hypothesis $g : \mathcal{X} \to \mathcal{R}$ and a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the loss is $\ell(g(x))_y$. We make a minimal non-redundancy assumption on ℓ, formalized in Section 2.1. We may refer to ℓ as a *discrete* target loss to emphasize our assumption that $\mathcal{R}$ and $\mathcal{Y}$ are finite.

Surrogate problems. We recall the surrogate risk minimization framework. First, one constructs a *surrogate loss function* $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$. Then, one minimizes it to learn a hypothesis of the form $h : \mathcal{X} \to \mathbb{R}^d$. Finally, one applies a *link function* $\psi : \mathbb{R}^d \to \mathcal{R}$, which maps h to the implied target hypothesis $\psi \circ h$.[3] In other words, if a surrogate prediction $h(x) \in \mathbb{R}^d$ implies a prediction $\psi(h(x)) \in \mathcal{R}$ for the target problem. For example, in binary classification, the target problem is given by 0-1 loss, $\ell_{\text{zo}}(r)_y = \mathbb{1}\{r \neq y\}$, with surrogate losses typically defined over $\mathbb{R}$. Common surrogate losses for 0-1 loss include hinge loss $L_{\text{hinge}}(u)_y = \max(1 - uy, 0)$, logistic loss $L_{\log}(u)_y = \ln(1 + \exp(-yu))$, and exponential loss $L_{\exp}(u)_y = \exp(-yu)$. These surrogates are typically associated with the link $\psi : u \mapsto \text{sign}(u)$, which maps negative predictions to -1 and nonnegative predictions to $+1$, breaking ties arbitrarily at 0.

Surrogate regret bounds. The suitability and efficiency of a chosen surrogate is often quantified with a *surrogate regret bound*. Formally, the *surrogate regret* of h with respect to data distribution $\mathcal{D}$ is given by $R_L(h; \mathcal{D}) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\, L(h(X))_Y - \inf_{h':\mathcal{X}\to\mathbb{R}^d} \mathbb{E}_{(X,Y)\sim\mathcal{D}}\, L(h'(X))_Y$, where the minimum is taken over all measurable functions. (One may equivalently assume the Bayes optimal hypothesis is in the function class.) Given h, the *target regret* of the implied hypothesis $\psi \circ h$ is $R_\ell(\psi \circ h; \mathcal{D}) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\, \ell(\psi(h(X)))_Y - \inf_{h':\mathcal{X}\to\mathcal{R}} \mathbb{E}_{(X,Y)\sim\mathcal{D}}\, \ell(h'(X))_Y$, where we recall that ψ is the link function.

A surrogate regret bound answers the following question: when we minimize the surrogate regret at a rate $f(n)$, given n data points, how fast does the target regret diminish? Formally, we say that $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$ is a *regret transfer function* if ζ is continuous at 0 and satisfies $\zeta(0) = 0$. Given a rerget transfer function ζ, we say that (L, ψ) guarantees a *regret transfer of* ζ for ℓ if

$$\forall \mathcal{D}, \forall h : \mathcal{X} \to \mathbb{R}^d, \quad R_\ell(\psi \circ h; \mathcal{D}) \leq \zeta(\, R_L(h; \mathcal{D})\,)\,. \tag{2}$$

We note that a classic minimal requirement for (L, ψ) is that they be *consistent* for ℓ, which is the case if there exists any regret transfer ζ guaranteed by (L, ψ) for ℓ [29]. In other words, consistency implies that $R_L(h; \mathcal{D}) \to 0 \implies R_\ell(\psi \circ h; \mathcal{D}) \to 0$, but says nothing about the relative rates.

Polyhedral surrogates. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called *polyhedral* if it can be written as a pointwise maximum of a finite set of affine functions [27, § 19]. The epigraph of a polyhedral function is a polyhedral set, i.e. the intersection of a finite set of closed halfspaces; thus polyhedral functions are always convex. We say a surrogate loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ is *polyhedral* if, for each fixed $y \in \mathcal{Y}$, the loss as a function of prediction, i.e. $r \mapsto L(r)_y$, is polyhedral.

2.1 Property elicitation and the conditional approach

Our technical approach is based on property elicitation, which studies the structure of ℓ and L as they relate to $\Delta_{\mathcal{Y}}$. This involves a *conditional* perspective. Given a data distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, for $(X, Y) \sim \mathcal{D}$, we consider the format of possible target predictions $g(x) = r \in \mathcal{R}$; surrogate predictions $h(x) = u \in \mathbb{R}^d$; and conditional distributions $\Pr[\cdot \mid X] = p \in \Delta_{\mathcal{Y}}$. By dropping the role of the hypotheses and the feature space $\mathcal{X}$, we can focus on the relationships of the functions $\ell(r)_y$ and $L(u)_y$ and the conditional distribution p.

In particular, in our notation the expected L-loss of prediction u on distribution p is simply the inner product $\langle p, L(u)\rangle$. The *Bayes risk* of L is $\underline{L}(p) = \inf_{u\in\mathbb{R}^d} \langle p, L(u)\rangle$, and similarly the Bayes risk of ℓ is $\underline{\ell}(p) = \inf_{r\in\mathcal{R}} \langle p, \ell(r)\rangle$. We abuse notation to define the *conditional surrogate regret* $R_L(u, p) = \langle p, L(u)\rangle - \underline{L}(p)$, and similarly the *conditional target regret* $R_\ell(r, p) = \langle p, \ell(r)\rangle - \underline{\ell}(p)$. Given a regret transfer function ζ, we say (L, ψ) guarantees a *conditional regret transfer* of ζ for ℓ if, for all $p \in \Delta_{\mathcal{Y}}$ and all $u \in \mathbb{R}^d$, $R_\ell(\psi(u), p) \leq \zeta(R_L(u, p))$.

The following is a direct result of Jensen's inequality, using that $R_L(h; \mathcal{D}) = \mathbb{E}_X\, R_L(h(X), p_X)$ where p_X is the conditional distribution on Y given X.

[3] The term "link function" can carry a more specific meaning in the statistics literature, for example when discussing generalized linear models. Our usage here is more general: any map from surrogate reports to target reports is a valid link function, though perhaps not a calibrated one.

Observation 1. *If (L, ψ) guarantee a conditional regret transfer of ζ for ℓ, and ζ is concave, then (L, ψ) guarantee a regret transfer of ζ for ℓ.*

Property elicitation. To begin, we define a *property*, which is essentially a statistic or summary of a distribution, such as the mode. We then define what it means for a loss function to *elicit* a property. We write $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows R$ as shorthand for $\Gamma : \Delta_{\mathcal{Y}} \to 2^R \setminus \{\emptyset\}$.

Definition 1 (Property, level set)**.** A *property* is a function $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows R$, for some set R. The *level set* of Γ for report r is the set $\Gamma_r = \{p \in \Delta_{\mathcal{Y}} : r \in \Gamma(p)\}$.

The following definition applies both to target losses and surrogate losses.

Definition 2 (Elicits, report space)**.** A loss function $f : R \to \mathbb{R}^{\mathcal{Y}}_+$, where the domain R is called the *report space*, *elicits* a property $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows R$ if

$$\forall p \in \Delta_{\mathcal{Y}}, \quad \Gamma(p) = \arg\min_{r \in R} \langle p, f(r) \rangle =: \mathrm{prop}[f](p) \ . \tag{3}$$

As each loss f elicits a unique property, we refer to it as $\mathrm{prop}[f]$. In particular, we will consistently use the notation $\Gamma = \mathrm{prop}[L]$ where L is a surrogate loss with report space $\mathbb{R}^d$, and we will use $\gamma = \mathrm{prop}[\ell]$ where ℓ is a target loss with finite report space $\mathcal{R}$. For example, $\mathrm{prop}[\ell_{\mathrm{zo}}]$ is simply the mode of p, given by $\mathrm{mode}(p) = \{-1, 1\}$ if $p = 1/2$, and $\mathrm{mode}(p) = \{\mathrm{sign}(2p - 1)\}$ otherwise.

We will assume throughout that the target loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ is *non-redundant*, meaning that all target reports $r \in \mathcal{R}$ are possibly uniquely optimal. Formally, letting $\gamma = \mathrm{prop}[\ell]$, we suppose that for all $r \in \mathcal{R}$, there exists $p \in \Delta_{\mathcal{Y}}$ with $\gamma(p) = \{r\}$.[4]

Calibration and consistency. It is known that consistency implies the weaker condition of *calibration*, which implies the still-weaker, but very useful, condition of *indirect elicitation*. To define these conditions, consider L, ψ, ℓ and $\gamma = \mathrm{prop}[\ell]$. Recall that $\gamma(p)$ is the set of target reports that are optimal for p. The typical definition of indirect elicitation, which we reformulate below, is as follows: a surrogate-link pair (L, ψ) *indirectly elicits* γ if $u \in \mathrm{prop}[L](p) \implies \psi(u) \in \gamma(p)$ for all $p \in \Delta_{\mathcal{Y}}$. Note that if $\psi(u) \notin \gamma(p)$, then u is a "bad" (suboptimal) surrogate prediction for p. Let $B_{L,\psi,\ell}(p) = \{R_L(u, p) \ : \ \psi(u) \notin \gamma(p)\}$, the set of expected surrogate losses resulting from bad predictions. Thus, we may equivalently state indirect elicitation as the requirement that suboptimal surrogate reports be bad, which clarifies the relationship to calibration: calibration further requires a nonzero gap between the expected loss of a "good" report and that of any bad one.

Definition 3 (Indirectly Elicits, Calibrated)**.** Let ℓ be a target loss and $\gamma = \mathrm{prop}[\ell]$. A surrogate-link pair (L, ψ) *indirectly elicits* γ if $0 \notin B_{L,\psi,\ell}(p)$ for all $p \in \Delta_{\mathcal{Y}}$. The pair (L, ψ) is *calibrated* for ℓ if $0 < \inf B_{L,\psi,\ell}(p)$ for all $p \in \Delta_{\mathcal{Y}}$.

Fact 1 ([4, 29])**.** *Let $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ be a target loss with finite $\mathcal{R}$ and $\mathcal{Y}$. If a surrogate and link (L, ψ) are consistent for ℓ, then (L, ψ) are calibrated for L. If (L, ψ) are calibrated for ℓ, then they indirectly elicit ℓ.*

For example, it is well-known that hinge loss L_{hinge} is calibrated and consistent for 0-1 loss ℓ_{zo} using the link $\psi : u \mapsto \mathrm{sign}(u)$. Recall that ℓ_{zo} elicits the mode. One can verify that indeed $(L_{\mathrm{hinge}}, \psi)$ indirectly elicit the mode; for example, $B_{L_{\mathrm{hinge}},\psi,\ell_{\mathrm{zo}}}(0) = B_{L_{\mathrm{hinge}},\psi,\ell_{\mathrm{zo}}}(1) = [1, \infty)$ and $B_{L_{\mathrm{hinge}},\psi,\ell_{\mathrm{zo}}}(1/2) = \emptyset$, neither of which contain 0.

3 Upper Bound: Polyhedral Surrogates

In this section, we describe the proof of Theorem 1, that polyhedral surrogates guarantee linear regret transfers. At a high level, the proof (whose details appear in Appendix A) has two parts:

1. Fix $p \in \Delta_{\mathcal{Y}}$. Prove a regret bound for this fixed p, namely a constant α_p such that $R_\ell(\psi(u), p) \leq \alpha_p R_L(u, p)$ for all $u \in \mathbb{R}^d$.
2. Apply this regret bound to each p in a carefully chosen finite subset of $\Delta_{\mathcal{Y}}$. Argue that the maximum α_p from this finite set suffices for the overal regret bound.

[4] Technically, this definition differs from the one given in literature [10, 11], but the two are equivalent for finite elicitable properties.

The first part will actually hold for any calibrated surrogate. The second part relies on a crucial fact about polyhedral losses L: their properties $\Gamma = \mathrm{prop}[L]$ cover the whole simplex with only a finite number of polyhedral level sets [10]. Thus, although the prediction space $\mathbb{R}^d$ is infinite, one can restrict to just a finite set of prediction points and always have a global minimizer of expected surrogate loss.

3.1 Linear transfer for fixed p

First, we establish a linear transfer function for any fixed conditional distribution p. This statement holds for any calibrated surrogate, not necessarily polyhedral.

Lemma 1. *Let $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$ be a discrete target loss and suppose the surrogate $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ and link $\psi : \mathbb{R}^d \to \mathcal{R}$ are calibrated for ℓ. Then for any $p \in \Delta_{\mathcal{Y}}$, there exists $\alpha_p \geq 0$ such that, for all $u \in \mathbb{R}^d$,*

$$R_\ell(\psi(u), p) \leq \alpha_p R_L(u, p).$$

While useful, this result is somewhat nonconstructive and potentially loose. In Section 5, we unpack the constant α_p for polyhedral losses in particular.

3.2 Linear overall transfer function

Given Lemma 1, we might hope to obtain that, for any calibrated surrogate, there exists $\alpha = \sup_p \alpha_p$ such that $R_\ell(\psi(u), p') \leq \alpha R_L(u, p')$ for all p'. However, we know this is false in general: not all surrogates yield linear regret transfers! Indeed, for many surrogates, the supremum will be $+\infty$.

However, polyhedral surrogates have a special structure that allows us to show α is finite. We first present some tools that hold for general surrogates, then the key implication from L being polyhedral.

Lemma 2 ([12])**.** *If (L, ψ) indirectly elicits γ, then $\Gamma = \mathrm{prop}[L]$ refines γ in the sense that, for all $u \in \mathbb{R}^d$, there exists $r \in \mathcal{R}$ such that $\Gamma_u \subseteq \gamma_r$.*

The next lemma states that surrogate regret is linear in p on any fixed level set Γ_{u^*} of $\Gamma = \mathrm{prop}[L]$. By combining this fact with Lemma 2, we obtain that target regret is linear on these level sets as well.

Lemma 3. *Suppose (L, ψ) indirectly elicits $\gamma = \mathrm{prop}[\ell]$ and let $\Gamma = \mathrm{prop}[L]$. Then for any fixed $u, u^* \in \mathbb{R}^d$ and $r \in \mathcal{R}$, the functions $R_L(u, \cdot)$ and $R_\ell(r, \cdot)$ are linear on Γ_{u^*} in their second arguments.*

A key fact we use about polyhedral losses is that they have a finite set of optimal sets.

Lemma 4. *If $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ is polyhedral, then $\Gamma = \mathrm{prop}[L]$ has a finite set of level sets that union to $\Delta_{\mathcal{Y}}$. Moreover, these level sets are polytopes.*

We are now ready to restate the main upper bound and sketch a proof.

Theorem 3 (Theorem 1 restated)**.** *Suppose the surrogate loss $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ and link $\psi : \mathbb{R}^d \to \mathcal{R}$ are consistent for the target loss $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$. If L is polyhedral, then (L, ψ) guarantee a linear regret transfer for ℓ, i.e. there exists $\alpha \geq 0$ such that, for all $\mathcal{D}$ and all measurable $h : \mathcal{X} \to \mathbb{R}^d$,*

$$R_\ell(\psi \circ h; \mathcal{D}) \leq \alpha R_L(h; \mathcal{D}).$$

The key point of the proof is that, by Lemma 4, the polyhedral loss L has a finite set $U \subset \mathbb{R}^d$ of predictions such that (a) for each $u \in U$, the level set Γ_u is a polytope, and (b) $\cup_{u \in U} \Gamma_u = \Delta_{\mathcal{Y}}$. Let $\mathcal{Q}$ be the finite set of all vertices of all these level sets. For any vertex $q \in \mathcal{Q}$, we have a linear regret transfer for this fixed q with constant α_q from Lemma 1. Lemma 3 allows us to write the regret of a general p as a convex combination of the regrets of its containing level set's vertices. So we obtain a bound of $\alpha = \max_{q \in \mathcal{Q}} \alpha_q$.

4 Lower bound

In this section, we show that if a surrogate loss is sufficiently "non-polyhedral", then the best regret transfer it can achieve is $\zeta(\epsilon)$ on the order of $\sqrt{\epsilon}$. (Proofs are deferred to Appendix B.) Specifically, we consider a relaxation of L being both strongly smooth and strongly convex, as we now formalize.

Let $\Gamma = \mathrm{prop}[L]$ and $\gamma = \mathrm{prop}[\ell]$. We will assume L is strongly smooth everywhere and strongly convex around some *boundary report for* L, ℓ, which we define to be a $u_0 \in \mathbb{R}^d$ such that, for some $r, r' \in \mathcal{R}$, there exists a distribution p_0 such that $p_0 \in \Gamma_{u_0} \cap \gamma_r \cap \gamma_{r'}$. For the conditional distribution p_0, the surrogate report u_0 minimizes expected L-loss, and it may link to either r or r', as either of these minimize expected ℓ-loss. Recall that an *open neighborhood* of u_0 is an open set in $\mathbb{R}^d$ containing u_0.

Assumption 1. $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ is a surrogate and $\psi : \mathbb{R}^d \to \mathcal{R}$ a link, consistent with the discrete target $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, satisfying the following: (i) For all $y \in \mathcal{Y}$, the function $L(\cdot)_y$ is differentiable with a locally Lipschitz gradient.[5] (ii) For some $\alpha > 0$ and some boundary report $u_0 \in \mathbb{R}^d$ for L, ℓ, the expected loss function $u \mapsto \langle p, L(u) \rangle$ is α-strongly convex on an open neighborhood of u_0.

This assumption is quite weak. In particular, because ℓ has a finite report set $\mathcal{R}$, boundary reports essentially always exist whenever (L, ψ) indirectly elicits ℓ. In particular, as the level sets of γ are closed, convex, and union to the simplex [10, 15], there must be some pair of reports $r, r' \in \mathcal{R}$ with $\gamma_r \cap \gamma_{r'} \neq \emptyset$. We can then take $p_0 \in \gamma_r \cap \gamma_{r'}$, and any $u_0 \in \Gamma(p_0)$. One caveat is that we may have $\Gamma(p_0) = \emptyset$. While some surrogates may fail to have a minimizer for some distributions p, this typically happens when p is on the boundary of the simplex, and not generally on the boundary between level sets. For example, consider exponential loss $L_{\exp}(u)_y = \exp(-uy)$, where $\mathrm{prop}[L_{\exp}](0) = \mathrm{prop}[L_{\exp}](1) = \emptyset$, but we still have $\mathrm{prop}[L_{\exp}](1/2) = \{0\}$.

The following result is our main lower bound.

Theorem 4 (Stronger version of Theorem 2.)**.** *Suppose the surrogate loss L and link ψ satisfy a regret transfer of ζ for a target loss ℓ. If L, ψ, and ℓ satisfy Assumption 1, then there exists $c > 0$ such that, for some $\epsilon^* > 0$, for all $0 \leq \epsilon < \epsilon^*$, $\zeta(\epsilon) \geq c\sqrt{\epsilon}$.*

The key idea of the proof is to fix a boundary report u_0 and consider a sequence of distributions $p_\lambda \to p_0$ as $\lambda \to 0$. Target regret $R_\ell(\psi(u_0), p_\lambda)$ shrinks linearly in λ, but surrogate regret $R_L(u_0, p_\lambda)$ shrinks quadratically. The latter holds roughly because expected surrogate loss is strongly convex in some neighborhood of u_0 and is strongly smooth elsewhere.[6]

Theorem 4 captures a wide range of surrogate losses beyond those that are strongly convex and strongly smooth. For example, exponential loss $L_{\exp}(u)_y = \exp(-uy)$ is not strongly convex, but does satisfy Assumption 1; for part (ii), the boundary report is $u_0 = 0$ as we saw above, and $L_{\exp}$ is $1/e$-strongly convex on $(-1, 1) \ni u_0$.

As another example, consider Huber loss for binary classification. Here $\mathcal{Y} = \{\pm 1\}$ and the prediction space is $\mathbb{R}$, and Huber loss can be defined by letting the hinge loss be $t = \max\{0, 1 - ry\}$ and setting $L(r)_y = t^2$ if $t \leq 2$, else $4(t-1)$. Theorem 4 does capture Huber, but needs the strength of Assumption 1(ii) as Huber is strongly convex on $[-1, 1]$ but linear outside that interval.

5 Unpacking the constant for polyhedral losses

Our main upper bound for polyhedral losses, Theorem 1, is mostly nonconstructive: it says polyhedral surrogates have some α such that $R_\ell(\psi \circ h; \mathcal{D}) \leq \alpha \cdot R_L(h; \mathcal{D})$. In this section, we make the result constructive by unpacking the implications of consistency for polyhedral surrogates. Proofs appear in Appendix C.

5.1 Contribution from the target, surrogate, and link

We will give a bound on the constant of the form $\alpha \leq \alpha_\ell \cdot \alpha_L \cdot \alpha_\psi$, i.e. one component each relating to the structure of the target loss, surrogate loss, and link. First, we will derive α_L using the theory of *Hoffman constants* from linear optimization [14]. Then, we will show that *any* link for a consistent polyhedral loss must satisfy a condition called ϵ-separation, introduced in [10]. We will set $\alpha_\psi = \frac{1}{\epsilon}$. With α_ℓ simply an upper bound on the target loss, we will finally prove $\alpha \leq \alpha_\ell \cdot \alpha_L \cdot \alpha_\psi$.

[5] A map $g : A \to B$ is locally Lipschitz if for every $a \in A$ there is an open neighborhood U of a such that $g|_A$ is Lipschitz continuous.

[6] A tempting alternative is to fix a distribution p_0 and consider a sequence of predictions. We know from Lemma 1 that this approach will fail, however: for any consistent surrogate and any fixed p_0, there is a linear regret transfer if we restrict to p_0.

5.1.1 Hoffman constants

The intuition we explore here is the linear growth rate of a given polyhedral function as we move away from its optimal set. In particular, consider the expected loss $u \mapsto \langle p, L(u)\rangle$. Expected loss should grow linearly with distance from the optimal set, which happens to be $\Gamma(p)$. In particular, the worst growth rate is governed roughly by the smallest slope of the function in any direction. This is captured by the *Hoffman constant* of an associated system of linear inequalities. In Appendix C, we use Hoffman constants to obtain the following result, where we define $H_{L,p}$.

Lemma 5 ([14])**.** *Let $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral loss with $\Gamma = \mathrm{prop}[L]$. Then for any fixed p, there exists some smallest constant $H_{L,p} \geq 0$ such that $d_\infty(u, \Gamma(p)) \leq H_{L,p} R_L(u, p)$ for all $u \in \mathbb{R}^d$.*

To foreshadow the usefulness of $H_{L,p}$, recall the proof strategy of the upper bound: use Lemma 1 to obtain a constant α_p for each p; then carefully choose a finite set of p's and take the maximum constant. In the same way, we will be able to set the "loss" constant $\alpha_L = \max_p H_{L,p}$. First, however, we will need to investigate separated links.

5.1.2 Separated link functions

A link is ϵ-separated if, for all pairs of surrogate reports u, u^* such that u^* is optimal for the surrogate and u links to a suboptimal target report, we have $\|u^* - u\|_\infty > \epsilon$. The definition was introduced by Finocchiaro et al. [10].

Definition 4 (Separated Link)**.** Let properties $\Gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathbb{R}^d$ and $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ be given. We say a link $\psi : \mathbb{R}^d \to \mathcal{R}$ is *ϵ-separated with respect to* Γ *and* γ if for all $u \in \mathbb{R}^d$ with $\psi(u) \notin \gamma(p)$, we have $d_\infty(u, \Gamma(p)) > \epsilon$, where $d_\infty(u, A) \doteq \inf_{a \in A} \|u - a\|_\infty$. Similarly, we say ψ is ϵ-separated with respect to L and ℓ if it is ϵ-separated with respect to $\mathrm{prop}[L]$ and $\mathrm{prop}[\ell]$.

Recall from the previous subsection that Hoffman constants allowed us to show that polyhedral expected losses grow linearly as we move away from the optimal set. Now, ϵ-separated link functions allow us to guarantee that we move at least a constant distance from the optimal set before we start linking to an "incorrect" report $r' \in \mathcal{R}$. But when does a polyhedral loss have an associated ϵ-separated link?

Finocchiaro et al. [10] showed that ϵ-separated links for polyhedral surrogates are calibrated. We show that the converse is in fact true: every calibrated link is ϵ-separated for some $\epsilon > 0$. The proof follows a similar argument to that of Tewari and Bartlett [30, Lemma 6].

Lemma 6. *Let polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, and link $\psi : \mathbb{R}^d \to \mathcal{R}$ be given such that (L, ψ) is calibrated with respect to ℓ. Then there exists $\epsilon > 0$ such that ψ is ϵ-separated with respect to $\Gamma \doteq \mathrm{prop}[L]$ and $\gamma \doteq \mathrm{prop}[\ell]$.*

5.1.3 Combining the loss and link

We can now follow the general proof strategy of the main upper bound, but constructively. Given Lemma 6, we can make the following definition.

Definition 5 (ϵ_ψ)**.** If (L, ψ) are calibrated for ℓ, then let $\epsilon_\psi \doteq \sup\{\epsilon \ : \ \psi \text{ is } \epsilon\text{-separated}\}$.

We will also need an upper bound $C_\ell = \max_{r,p} R_\ell(r, p)$ on the regret of ℓ. Since R_ℓ is convex in p, in particular this maximum is achieved at a vertex of the probability simplex.

Definition 6 (C_ℓ)**.** Given discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, define $C_\ell = \max_{r,r' \in \mathcal{R}, y \in \mathcal{Y}} \ell(r)_y - \ell(r')_y$.

Recall that Lemma 4 gave that, if L is polyhedral, then $\Gamma = \mathrm{prop}[L]$ has a finite set of full-dimensional level sets, each a polytope, that union to the simplex.

Definition 7 (H_L)**.** Given a polyhedral surrogate loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, let $\mathcal{Q}$ be the set of all vertices of the full-dimensional level sets of $\Gamma = \mathrm{prop}[L]$, and define $H_L \doteq \max_{p \in \mathcal{Q}} H_{L,p}$.

Theorem 5 (Constructive linear transfer)**.** *Let $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ be a discrete target loss, $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral surrogate loss, and $\psi : \mathbb{R}^d \to \mathcal{R}$ a link function. If (L, ψ) are consistent for ℓ, then*

$$(\forall h, \mathcal{D}) \quad R_\ell(\psi \circ h; \mathcal{D}) \leq \frac{C_\ell H_L}{\epsilon_\psi} R_L(h; \mathcal{D}) \ .$$

The proof closely mirrors the proof of the nonconstructive upper bound, Theorem 1, but using the constructive constants $C_\ell, H_L, \epsilon_\psi$ derived above. To summarize, we have shown the following.

Theorem 6. *Let $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$ be a discrete target loss, $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ be a polyhedral surrogate loss, and $\psi : \mathbb{R}^d \to \mathcal{R}$ a link function. The following are equivalent:*

1. *(L, ψ) is consistent for ℓ.*
2. *ψ is ϵ-separated with respect to L and ℓ for some $\epsilon > 0$.*
3. *(L, ψ) guarantees a linear regret transfer for ℓ.*

Furthermore, if any of the above hold, then in particular (L, ψ) guarantees the regret transfer $\zeta(t) = \left(\frac{C_\ell H_L}{\epsilon_\psi}\right) t$.

5.2 Further tightening

While our goal is not to focus on exact constants for specific problems, we remark on a few ways Theorem 5 may be loose and how one could compute the tightest possible constant α^* for which $R_\ell(\psi \circ h; \mathcal{D}) \leq \alpha^* R_L(h; \mathcal{D})$ for all h and $\mathcal{D}$. In general, for a fixed p, there is some smallest α_p^* such that $R_\ell(\psi(u), p) \leq \alpha_p^* R_L(u, p)$ for all u. Then, it follows from our results that $\alpha^* = \max_{p \in \mathcal{Q}} \alpha_p^*$ for the finite set $\mathcal{Q}$ used in the proof, i.e. the vertices of the full-dimensional level sets of $\Gamma = \mathrm{prop}[L]$.

Above, we bounded $\alpha_p^* \leq \frac{C_\ell H_{L,p}}{\epsilon_\psi}$. The intuition is that some u at distance $\geq \epsilon_\psi$ from $\Gamma(p)$, the optimal set, may link to a "bad" report $r = \psi(u) \notin \gamma(p)$. The rate at which L grows is at least $H_{L,p}$, so the surrogate loss at u may be as small as $\frac{\epsilon_\psi}{H_{L,p}}$, while the target regret may be as high as $C_\ell = \max_{r',p'} R_\ell(r', p')$. The ratio of regrets is therefore bounded by $\frac{H_{L,p} C_\ell}{\epsilon_\psi}$.

The tightest possible bound, on the other hand, is $\alpha^* = \sup_{u:\psi(u) \notin \gamma(p)} \frac{R_\ell(\psi(u),p)}{R_L(u,p)}$. This bound can be smaller if the values of numerator and denominator are correlated across u. For example, u may only be ϵ_ψ-close to the optimal set when it links to reports $\psi(u)$ with lower target regret; or L may have a smaller slope in the direction where the link's separation is larger than ϵ.

To illustrate with a concrete example, consider the *binary encoded predictions (BEP) surrogate* of [25] for the abstain target loss, $\ell(r, y) = \frac{1}{2}$ if $r = \perp$, otherwise $\ell(r, y) = \mathbb{1}[r \neq y]$. The surrogate involves an injective map $B : \mathcal{Y} \to \{-1, 1\}^d$ for $d = \lceil \log_2 |\mathcal{Y}| \rceil$. It is $L(u)_y = \max_{j=1\ldots d}(1 - u_d B(y)_d)_+$, where $(\cdot)_+$ indicates taking the maximum of the argument and zero. The associated link is $\psi(u) = \perp$ if $\min_{j=1\ldots d} |u_j| \leq \frac{1}{2}$, otherwise $\psi(u) = \arg\min_{y \in \mathcal{Y}} \|B(y) - u\|_\infty$.

One can show that for $p = \delta_y$, i.e. the distribution with full support on some $y \in \mathcal{Y}$, $L(u)_y = d_\infty(u, \Gamma(p))$ exactly, giving $H_{L,p} = 1$. It is almost immediate that $\epsilon_\psi = \frac{1}{2}$. Meanwhile, $R_\ell(r, p) \leq 1$, giving us an upper bound $\alpha_p^* \leq \frac{(1)(1)}{1/2} = 2$. In fact, this is slightly loose; the exact constant, given in [25] is 1. The looseness stems from the fact that for $p = \delta_y$, the closest reports u to the optimal set, i.e. at distance only $\epsilon_\psi = \frac{1}{2}$ away, do not link to reports maximizing target regret; they link to the abstain report $\perp$, which has regret only $\frac{1}{2}$. With this correction, and an observation that all u linking to reports $y' \neq y$ are at distance at least $\frac{3}{2}$ from $\Gamma(p)$, we restore the tight bound $\alpha_p^* \leq 1$. A similar but slightly more involved calculation can be carried out for the other vertices $p \in \mathcal{Q}$, which turn out to be all vertices of the form $\frac{1}{2}\delta_y + \frac{1}{2}\delta_{y'}$.

Finally, while we use $\|\cdot\|_\infty$ to define the minimum-slope H_L and the separation ϵ_ψ, in principle one could use another norm. One reason for restricting to $\|\cdot\|_\infty$ is that it is more compatible with Hoffman constants. However, all definitions hold for other norms and so does the main upper bound, as existence of an H_L and ϵ_ψ in $\|\cdot\|_\infty$ imply existence of constants for other norms. These constants may change for different norms, and in particular, the optimal overall constant may arise from a norm other than $\|\cdot\|_\infty$.

6 Discussion

We have shown two broad results about regret tranfer functions for surrogate losses. In particular, polyhedral surrogates always achieve a linear transfer, whereas "non-polyhedral" surrogates are generally square root. Section 5 outlines several directions to further refine the bound for polyhedral surrogates. Beyond these directions, an interesting question is to what extent our results hold when the label set or report set are infinite. For example, when the labels are the real line, pinball loss is polyhedral yet elicits a quantile, a very different case than the one we study. In particular, Lemma 4 will not give a finite set of optimal sets in this case. Nonetheless, we suspect similar results could go through under some regularity assumptions. Finally, in line with our results, we would like to better understand the tradeoff between smoothness of the surrogate and the overall generalization rate. One important direction along these lines is to extend the analysis of Mahdavi et al. [17] beyond binary classification.

Broader Impacts

As a theoretical work, the likely impacts of this paper take the form of downstream research and applications. Instead of direct applications, we anticipate this work leading to more investigation of surrogate losses to improve discrete prediction tasks. It may inform practitioners' choices of which surrogate losses they use for supervised learning tasks. Of course, such machine learning tasks can be solved for ethical or unethical purposes. We do not know of particular risk of negative impacts of this work beyond risks of supervised machine learning in general.

Acknowledgments and Disclosure of Funding

We thank Stephen Becker, Jessie Finocchiaro, and Nishant Mehta for insights, discussions, and references. This material is based upon work supported by the National Science Foundation under Grant No. IIS-2045347.

References

[1] Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *The Journal of Machine Learning Research*, 15(1):1653–1674, 2014.

[2] Franz Aurenhammer. A criterion for the affine equivalence of cell complexes in R d and convex polyhedra in R d+1. *Discrete & Computational Geometry*, 2(1):49–64, December 1987. ISSN 0179-5376, 1432-0444. doi: 10.1007/BF02187870. URL http://link.springer.com/article/10.1007/BF02187870.

[3] 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.

[4] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214505000000907.

[5] Mathieu Blondel. Structured prediction with projection oracles. *Advances in neural information processing systems*, 32:12145–12156, 2019.

[6] Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. Classification with rejection based on cost-sensitive classification. In *International Conference on Machine Learning*, pages 1507–1517. PMLR, 2021.

[7] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. *Advances in neural information processing systems*, 29:4412–4420, 2016.

[8] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pages 67–82. Springer, 2016.

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

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

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

[12] Rafael M Frongillo and Ian A Kash. Elicitation complexity of statistical properties. *Biometrika*, 2020.

[13] Wei Gao and Zhi-Hua Zhou. On the consistency of auc pairwise optimization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[14] Alan J Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4), 1952.

[15] Nicolas S. Lambert and Yoav Shoham. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 109–118, 2009.

[16] 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.

[17] Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Binary excess risk for smooth convex surrogates. *arXiv preprint arXiv:1402.1792*, 2014.

[18] Aditya Krishna Menon and Robert C Williamson. Bipartite ranking: a risk-theoretic perspective. *The Journal of Machine Learning Research*, 17(1):6766–6867, 2016.

[19] Chenri Ni, Nontawat Charoenphakdee, Junya Honda, and Masashi Sugiyama. On the calibration of multiclass classification with rejection. *arXiv preprint arXiv:1901.10655*, 2019.

[20] Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. A general theory for structured prediction with smooth convex surrogates. *arXiv preprint arXiv:1902.01958*, 2019.

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

[22] Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *Journal of Machine Learning Research*, 18:1–35, 2017.

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

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

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

[26] Mark D Reid and Robert C Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 897–904, 2009.

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

[28] Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6: 958–992, 2012.

[29] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, September 2008. ISBN 978-0-387-77242-4. Google-Books-ID: HUnqnrpYt4IC.

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

[31] Yutong Wang and Clayton Scott. Weston-watkins hinge loss and ordered partitions. *Advances in neural information processing systems*, 2020.

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

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

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

[35] Constantin Zalinescu. Sharp estimates for hoffman's constant for systems of linear inequalities and equalities. *SIAM Journal on Optimization*, 14(2):517–533, 2003.

[36] Jingwei Zhang, Tongliang Liu, and Dacheng Tao. On the rates of convergence from surrogate risk minimizers to the bayes optimal classifier. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[37] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

A Omitted Proofs: Upper Bound

This section contains omitted proofs from Section 3.

Lemma (Lemma 1)**.** *Let $\ell : \mathcal{R} \to \mathbb{R}_+^{\mathcal{Y}}$ be a discrete target loss and suppose the surrogate $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ and link $\psi : \mathbb{R}^d \to \mathcal{R}$ are calibrated for ℓ. Then for any $p \in \Delta_{\mathcal{Y}}$, there exists $\alpha_p \geq 0$ such that, for all $u \in \mathbb{R}^d$,*

$$R_\ell(\psi(u), p) \leq \alpha_p R_L(u, p).$$

Proof. Fix $p \in \Delta_{\mathcal{Y}}$. Let $C_p = \max_{r \in \mathcal{R}} R_\ell(r, p)$. The maximum exists because ℓ is discrete, i.e. $\mathcal{R}$ is finite. Meanwhile, recall that, when defining calibration, we let $B_{L,\psi,\ell}(p) = \{R_L(u, p) \,:\, \psi(u) \notin \gamma(p)\}$. Let $B_p = \inf B_{L,\psi,\ell}(p)$. By definition of calibration, we have $B_p > 0$.

To combine these bounds, let $\alpha_p = \frac{C_p}{B_p}$. Let $u \in \mathbb{R}^d$. There are two cases. If $\psi(u) \in \gamma(p)$, then $R_\ell(\psi(u), p) = 0 \leq R_L(u, p)$ immediately. If $\psi(u) \notin \gamma(p)$, then

$$\begin{aligned} R_\ell(\psi(u), p) &\leq C_p \\ &= \alpha_p \cdot B_p \\ &\leq \alpha_p R_L(u, p). \end{aligned}$$

□

Lemma (Lemma 2)**.** *If (L, ψ) indirectly elicits ℓ, then $\Gamma = \mathrm{prop}[L]$ refines $\gamma = \mathrm{prop}[\ell]$ in the sense that, for all $u \in \mathbb{R}^d$, there exists $r \in \mathcal{R}$ such that $\Gamma_u \subseteq \gamma_r$.*

Proof. For any u, let $r = \psi(u)$. By indirect elicitation, $u \in \Gamma(p) \implies r \in \gamma(p)$. So $\Gamma_u = \{p \,:\, u \in \Gamma(p)\} \subseteq \{p \,:\, r \in \gamma(p)\} = \gamma_r$. □

Lemma (Lemma 3)**.** *Suppose (L, ψ) indirectly elicits ℓ and let $\Gamma = \mathrm{prop}[L]$. Then for any fixed $u, u^* \in \mathbb{R}^d$ and $r \in \mathcal{R}$, the functions $R_L(u, \cdot)$ and $R_\ell(r, \cdot)$ are linear in their second arguments on Γ_{u^*}.*

Proof. Let $u^* \in \mathbb{R}^d$ and $p \in \Gamma_{u^*}$. By definition, for all $p \in \Gamma_{u^*}$, $\underline{L}(p) = \langle p, L(u^*) \rangle$. So for fixed u,

$$R_L(u, p) = \langle p, L(u) \rangle - \langle p, L(u^*) \rangle = \langle p, L(u) - L(u^*) \rangle,$$

a linear function of p on Γ_{u^*}. Next, by Lemma 2, there exists r^* such that $\Gamma_{u^*} \subseteq \gamma_{r^*}$. By the same argument, for fixed r, $R_\ell(r, p) = \langle p, \ell(r) - \ell(r^*) \rangle$, a linear function of p on γ_{r^*} and thus on Γ_{u^*}. □

Lemma (Lemma 4)**.** *If $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ is polyhedral, then $\Gamma = \mathrm{prop}[L]$ has a finite set of level sets that union to $\Delta_{\mathcal{Y}}$. Moreover, these level sets are polytopes.*

Proof. This statement can be deduced from the embedding framework of [10]. In particular, Lemma 5 of [10] states that if L is polyhedral, then its Bayes risk $\underline{L}$ is concave polyhedral, i.e. is the pointwise minimum of a finite set of affine functions. It follows that there exists a finite set $U \subset \mathbb{R}^d$ such that

$$\underline{L}(p) = \min_{u \in \mathbb{R}^d} \langle p, L(u) \rangle = \min_{u \in U} \langle p, L(u) \rangle \ . \tag{4}$$

We claim the level sets of U witness the claim. First, it is known (e.g. from theory of power diagrams, [2]) that if $\underline{L}$ is a polyhedral function represented as (4) and $u \in U$, then $\Gamma_u = \{p \in \Delta_{\mathcal{Y}} : \langle p, L(u) \rangle = \underline{L}(p)\}$ is a polytope. Finally, suppose for contradiction that there exists $p \in \Delta_{\mathcal{Y}}$, $p \notin \cup_{u \in U} \Gamma_u$. Then there must be some $u' \notin U$ with $p \in \Gamma_{u'}$, implying that $\langle p, L(u') \rangle > \max_{u \in U} \langle p, L(u) \rangle$, contradicting (4). □

Theorem (Theorem 3)**.** *Suppose the surrogate loss $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ and link $\psi : \mathbb{R}^d \to \mathcal{R}$ are consistent for the target loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$. If L is polyhedral, then (L, ψ) guarantee a linear regret transfer for ℓ, i.e. there exists $\alpha \geq 0$ such that, for all $\mathcal{D}$ and all measurable $h : \mathcal{X} \to \mathbb{R}^d$,*

$$R_\ell(\psi \circ h; \mathcal{D}) \leq \alpha R_L(h; \mathcal{D}).$$

Proof. We first recall that by Fact 1, consistency implies that (L, ψ) are calibrated for ℓ and that (L, ψ) indirectly elicit ℓ. Next, by Observation 1, it suffices to show a linear *conditional* regret transfer, i.e. for all $p \in \Delta_{\mathcal{Y}}$ and $u \in \mathbb{R}^d$, we show $R_\ell(\psi(u), p) \leq \alpha R_L(u, p)$.

By Lemma 4, the polyhedral loss L has a finite set $U \subset \mathbb{R}^d$ of predictions such that (a) for each $u \in U$, the level set Γ_u is a polytope, and (b) $\cup_{u \in U} \Gamma_u = \Delta_{\mathcal{Y}}$. Let $\mathcal{Q}_u \subset \Delta_{\mathcal{Y}}$ be the finite set of vertices of the polytope Γ_u, and define the finite set $\mathcal{Q} = \cup_{u \in U} \mathcal{Q}_u$.

By Lemma 1, for each $q \in \mathcal{Q}$, there exists $\alpha_q \geq 0$ such that $R_\ell(\psi(u), q) \leq \alpha_q R_L(u, q)$ for all u. We choose

$$\alpha = \max_{q \in \mathcal{Q}} \alpha_q.$$

To prove the conditional regret transfer, consider any $p \in \Delta_{\mathcal{Y}}$ and any $u \in \mathbb{R}^d$. There exists $u \in U$ such that $p \in \Gamma_u$, a polytope. So we can write p as a convex combination of its vertices, i.e.

$$p = \sum_{q \in \mathcal{Q}_u} \beta(q) q$$

for some probability distribution β. Recall that $\mathcal{Q}_u \subseteq \Gamma_u$ and R_L and R_ℓ are linear in p on Γ_u by Lemma 3. So, for any u':

$$\begin{aligned}
R_\ell(\psi(u'), p) &= R_\ell\left(\psi(u')\ , \sum_{q \in \mathcal{Q}_u} \beta(q) q\right) \\
&= \sum_{q \in \mathcal{Q}_u} \beta(q) R_\ell(\psi(u'), q) \\
&\leq \sum_{q \in \mathcal{Q}_u} \beta(q) \alpha_q R_L(u', q) \\
&\leq \alpha \sum_{q \in \mathcal{Q}_u} \beta(q) R_L(u', q) \\
&= \alpha R_L(u', p).
\end{aligned}$$

□

B Omitted Proofs: Lower Bound

This section contains omitted proofs from Section 4.

Theorem (Theorem 4)**.** *Suppose the surrogate loss L and link ψ satisfy a regret transfer of ζ for a target loss ℓ. If L, ψ, and ℓ satisfy Assumption 1, then there exists $c > 0$ such that, for some $\epsilon^* > 0$, for all $0 \leq \epsilon < \epsilon^*$, $\zeta(\epsilon) \geq c\sqrt{\epsilon}$.*

Proof outline: By assumption we have a boundary report u_0 which is L-optimal for a distribution p_0. We have some r, r' which are both optimal for p_0, and $\psi(u_0) = r'$. First, we will choose a p_1 where r is uniquely optimal, hence u_0 is a strictly suboptimal choice. We then consider a sequence of distributions $p_\lambda = (1-\lambda)p_0 + \lambda p_1$, approaching p_0 as $\lambda \to 0$. For all such p_λ, it will happen that r is optimal while u_0 and $r' = \psi(u_0)$ are strictly suboptimal. We show that $R_\ell(r', p_\lambda) = c_\ell \lambda$ for some constant c_ℓ and all small enough λ. Meanwhile, we will show that $R_L(u_0, p_\lambda) \leq O(\lambda^2)$, proving the result. The last fact will use the properties of strong smoothness and strong convexity in a neighborhood of u_0.

Proof. Obtain α, u_0, p_0, r, r', and an open neighborhood of u_0 from Assumption 1 and the definition of boundary report. Assume without loss of generality that $\psi(u_0) = r'$; otherwise, swap the roles of r and r'.

Linearity of $R_\ell(r', p_\lambda)$. As ℓ is non-redundant by assumption, there exists some $p_1 \in \mathring{\gamma}_r$, the relative interior of the full-dimensional level set γ_r. We therefore have $R_\ell(r', p_1) = \langle p_1, \ell(r') - \ell(r)\rangle =: c_\ell > 0$, and $R_\ell(r', p_0) = 0$. Let $p_\lambda := (1-\lambda)p_0 + \lambda p_1$. By convexity of γ_r, we have $p_\lambda \in \gamma_r$ for all $\lambda \in [0,1]$, which gives $R_\ell(r', p_\lambda) = \lambda c_\ell$.

Obtaining the global minimizer u_λ of L_λ. Let $L_\lambda : \mathbb{R}^d \to \mathbb{R}_+$ be given by $L_\lambda(u) = \langle p_\lambda, L(u)\rangle = (1-\lambda)\langle p_0, L(u)\rangle + \lambda\langle p_1, L(u)\rangle$. Let $\delta > 0$ such that the above open neighborhood of u_0 contains the Euclidean ball $B_\delta(u_0)$ of radius δ around u_0. Let $u_1 \in \Gamma(p_1)$. We argue that for all small enough λ, $L_\lambda(u)$ is uniquely minimized by some $u_\lambda \in B_\delta(u_0)$. For any $u \notin B_\delta(u_0)$, we have, using local strong convexity and the optimality of u_1,

$$\begin{aligned} L_\lambda(u) - L_\lambda(u_0) &= (1-\lambda)\left(L_0(u) - L_0(u_0)\right) + \lambda\left(L_1(u) - L_1(u_0)\right) \\ &\geq (1-\lambda)\left(\frac{\alpha}{2}\delta^2\right) + \lambda\left(L_1(u_1) - L_1(u_0)\right) \\ &> 0 \end{aligned}$$

if $\lambda < \lambda^* := \alpha\delta^2/(2\alpha\delta^2 + 4L_1(u_0) - 4L_1(u_1))$. For the remainder of the proof, let $\lambda < \lambda^*$. Then any $u \notin B_\delta(u_0)$ has $L_\lambda(u) > L_\lambda(u_0)$, hence is suboptimal. By α-strong convexity of L_0 on $B_\delta(u_0)$, L_λ is strictly convex on $B_\delta(u_0)$. So it has a unique minimizer u_λ, and by the above argument this is the global minimizer of L_λ. Then $\underline{L}(p_\lambda) = L_\lambda(u_\lambda)$, and thus $R_L(u_0, p_\lambda) = L_\lambda(u_0) - L_\lambda(u_\lambda)$. We also observe here that $R_L(u_0, p_\lambda)$ is continuous in λ, e.g. because the Bayes risk of L is continuous in p as is $\langle p, L(u_0)\rangle$. It is also zero when $\lambda = 0$.

Showing R_L is quadratic in λ. By assumption, the gradient of L_y is locally Lipschitz for all $y \in \mathcal{Y}$. We will apply this fact to the compact set $\mathcal{C} = \{u \in \mathbb{R}^d : \|u - u_1\| \leq \|u_0 - u_1\| + \delta\}$. By compactness, we have a finite subcover of open neighborhoods; let β be the minimum Lipschitz constant over this finite set of neighborhoods. We thus have that L_y is β-strongly smooth on $\mathcal{C}$, and hence so is L_λ for any $\lambda \in [0,1]$.

We now upper bound $\|u_\lambda - u_0\|_2$, and then apply strong smoothness to upper bound $R_L(u_0, p_\lambda) = L_\lambda(u_0) - L_\lambda(u_\lambda)$. Consider the first-order optimality condition of L_λ:

$$\begin{aligned} 0 = \nabla L_\lambda(u_\lambda) &= (1-\lambda)\nabla L_0(u_\lambda) + \lambda \nabla L_1(u_\lambda) \\ \implies (1-\lambda)\|\nabla L_0(u_\lambda)\|_2 &= \lambda\|\nabla L_1(u_\lambda)\|_2 \ . \end{aligned}$$

By optimality of u_0 and u_1, strong convexity of L_0 and strong smoothness of L_1, and the triangle inequality, we have

$$\begin{aligned} \|\nabla L_0(u_\lambda)\|_2 &= \|\nabla L_0(u_\lambda) - \nabla L_0(u_0)\|_2 \geq \alpha\|u_\lambda - u_0\|_2 \ , \\ \|\nabla L_1(u_\lambda)\|_2 &= \|\nabla L_1(u_\lambda) - \nabla L_1(u_1)\|_2 \leq \beta\|u_\lambda - u_1\|_2 \\ &\leq \beta\left(\|u_\lambda - u_0\|_2 + \|u_0 - u_1\|_2\right) \ . \end{aligned}$$

Combining,

$$\begin{aligned} (1-\lambda)\alpha\|u_\lambda - u_0\|_2 &\leq (1-\lambda)\|\nabla L_0(u_\lambda)\|_2 \\ &= \lambda\|\nabla L_1(u_\lambda)\|_2 \\ &\leq \lambda\beta\left(\|u_\lambda - u_0\|_2 + \|u_0 - u_1\|_2\right) \ . \end{aligned}$$

Now rearranging and taking $\lambda \leq \frac{1}{2}\frac{\alpha}{\alpha+\beta}$, we have

$$\|u_\lambda - u_0\|_2 \leq \frac{\lambda\beta}{(1-\lambda)\alpha - \lambda\beta}\|u_0 - u_1\|_2 \leq \lambda\frac{2\beta}{\alpha}\|u_0 - u_1\|_2 \ .$$

Finally, from strong smoothness of L_λ and optimality of u_λ,

$$L_\lambda(u_0) - L_\lambda(u_\lambda) \leq \frac{\beta}{2}\|u_0 - u_\lambda\|_2^2 \leq \frac{\beta}{2}\left(\lambda\frac{2\beta}{\alpha}\|u_0 - u_1\|_2\right)^2 = c_L\lambda^2 \ ,$$

where $c_L = \frac{2\beta^3}{\alpha^2}\|u_0 - u_1\|_2^2 > 0$.

To conclude: we have found a $\lambda^* > 0$ and shown that for all $0 \leq \lambda < \lambda^*$, $R_\ell(r', p_\lambda) = c_\ell\lambda$ and $R_L(u_0, p_\lambda) \leq c_L\lambda^2$. In particular, let $\epsilon^* = \sup_{0\leq\lambda<\lambda^*} R_L(u_0, p_\lambda)$. Then for all $0 \leq \epsilon < \epsilon^*$, by continuity, we can choose $\lambda < \lambda^*$ such that $R_L(u_0, p_\lambda) = \epsilon \leq c_L\lambda^2$. Meanwhile, $R_\ell(\psi(u_0), p_\lambda) = c_\ell\lambda \geq \frac{c_\ell}{\sqrt{c_L}}\sqrt{\epsilon}$. Recalling that $\zeta(R_L(u_0, p_\lambda)) \geq R_\ell(\psi(u_0), p_\lambda)$ by definition, this implies $\zeta(\epsilon) \geq c\sqrt{\epsilon}$ for all $\epsilon < \epsilon^*$, with $c = \frac{c_\ell}{\sqrt{c_L}}$. □

C Omitted Proofs: Constant Derivation

This section contains omitted proofs from Section 5.

C.1 Hoffman constants

First we appeal to a known fact, the existence of Hoffman constants for systems of linear inequalities. See Zalinescu [35] for a modern treatment.

Theorem 7 (Hoffman constant [14])**.** *Given a matrix $A \in \mathbb{R}^{m\times n}$, there exists some smallest $H(A) \geq 0$, called the* Hoffman constant *(with respect to $\|\cdot\|_\infty$), such that for all $b \in \mathbb{R}^m$ and all $x \in \mathbb{R}^n$,*

$$d_\infty(x, S(A,b)) \leq H(A)\|(Ax - b)_+\|_\infty \ , \tag{5}$$

where $S(A,b) = \{x \in \mathbb{R}^n \mid Ax \leq b\}$ and $(u)_+ \doteq \max(u, 0)$ component-wise.

Lemma (Lemma 5)**.** *Let $L : \mathbb{R}^d \to \mathbb{R}_+^{\mathcal{Y}}$ be a polyhedral loss with $\Gamma = \mathrm{prop}[L]$. Then for any fixed p, there exists some smallest constant $H_{L,p} \geq 0$ such that $d_\infty(u, \Gamma(p)) \leq H_{L,p}R_L(u,p)$ for all $u \in \mathbb{R}^d$.*

Proof. Since L is polyhedral, there exist $a_1, \ldots, a_m \in \mathbb{R}^d$ and $c \in \mathbb{R}^m$ such that we may write $\langle p, L(u)\rangle = \max_{1\leq j\leq m} a_j \cdot u + c_j$. Let $A \in \mathbb{R}^{m\times d}$ be the matrix with rows a_j, and let $b = \underline{L}(p)\mathbb{1} - c$, where $\mathbb{1} \in \mathbb{R}^m$ is the all-ones vector. Then we have

$$\begin{aligned}
S(A,b) &\doteq \{u \in \mathbb{R}^d \mid Au \leq b\} \\
&= \{u \in \mathbb{R}^d \mid Au + c \leq \underline{L}(p)\mathbb{1}\} \\
&= \{u \in \mathbb{R}^d \mid \forall i\, (Au + c)_i \leq \underline{L}(p)\} \\
&= \{u \in \mathbb{R}^d \mid \max_i\, (Au + c)_i \leq \underline{L}(p)\} \\
&= \{u \in \mathbb{R}^d \mid \langle p, L(u)\rangle \leq \underline{L}(p)\} \\
&= \Gamma(p) \ .
\end{aligned}$$

Similarly, we have $\max_i\, (Au - b)_i = \langle p, L(u)\rangle - \underline{L}(p) = R_L(u,p) \geq 0$. Thus,

$$\begin{aligned}
\|(Au - b)_+\|_\infty &= \max_i\, ((Au - b)_+)_i \\
&= \max((Au - b)_1, \ldots, (Au - b)_m, 0) \\
&= \max(\max_i\, (Au - b)_i,\, 0) \\
&= \max_i\, (Au - b)_i \\
&= R_L(u,p) \ .
\end{aligned}$$

Now applying Theorem 7, we have

$$\begin{aligned} d_\infty(u, \Gamma(p)) &= d_\infty(u, S(A,b)) \\ &\leq H(A)\|(Au-b)_+\|_\infty \\ &= H(A)R_L(u,p)\ . \end{aligned}$$

□

C.2 Separated links

Lemma (Lemma 6)**.** *Let polyhedral surrogate $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$, discrete loss $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$, and link $\psi : \mathbb{R}^d \to \mathcal{R}$ be given such that (L, ψ) is calibrated with respect to ℓ. Then there exists $\epsilon > 0$ such that ψ is ϵ-separated with respect to $\Gamma \doteq \mathrm{prop}[L]$ and $\gamma \doteq \mathrm{prop}[\ell]$.*

Proof. Suppose that ψ is not ϵ-separated for any $\epsilon > 0$. Then letting $\epsilon_i \doteq 1/i$ we have sequences $\{p_i\}_i \subset \Delta_{\mathcal{Y}}$ and $\{u_i\}_i \subset \mathbb{R}^d$ such that for all $i \in \mathbb{N}$ we have both $\psi(u_i) \notin \gamma(p_i)$ and $d_\infty(u_i, \Gamma(p_i)) \leq \epsilon_i$. First, observe that there are only finitely many values for $\gamma(p_i)$ and $\Gamma(p_i)$, as $\mathcal{R}$ is finite and L is polyhedral. Thus, there must be some $p \in \Delta_{\mathcal{Y}}$ and some infinite subsequence indexed by $j \in J \subseteq \mathbb{N}$ where for all $j \in J$, we have $\psi(u_j) \notin \gamma(p)$ and $\Gamma(p_j) = \Gamma(p)$.

Next, observe that, as L is polyhedral, the expected loss $\langle p, L(u)\rangle$ is β-Lipschitz in $\|\cdot\|_\infty$ for some $\beta > 0$. Thus, for all $j \in J$, we have

$$\begin{aligned} d_\infty(u_i, \Gamma(p)) \leq \epsilon_j &\implies \exists u^* \in \Gamma(p) \|u_j - u^*\|_\infty \leq \epsilon_j \\ &\implies |\langle p, L(u_j)\rangle - \langle p, L(u^*)\rangle| \leq \beta\epsilon_j \\ &\implies |\langle p, L(u_j)\rangle - \underline{L}(p)| \leq \beta\epsilon_j\ . \end{aligned}$$

Finally, for this p, we have

$$\inf_{u:\psi(u)\notin\gamma(p)} \langle p, L(u)\rangle \leq \inf_{j\in J} \langle p, L(u_j)\rangle = \underline{L}(p)\ ,$$

contradicting the calibration of ψ. □

C.3 Combining the loss and link

Lemma 7. *Let $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ be a discrete target loss, $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral surrogate loss, and $\psi : \mathbb{R}^d \to \mathcal{R}$ a link function. If (L, ψ) indirectly elicit ℓ and ψ is ϵ-separated, then for all u and p,*

$$R_\ell(\psi(u), p) \leq \frac{C_\ell H_{L,p}}{\epsilon} R_L(u,p).$$

Proof. If $\psi(u) \in \gamma(p)$, then $R_\ell(u,p) = 0$ and we are done. Otherwise, applying the definition of ϵ-separated and Lemma 5,

$$\begin{aligned} \epsilon &< d_\infty(u, \Gamma(p)) \\ &\leq H_{L,p} R_L(u,p). \end{aligned}$$

So $R_\ell(\psi(u), p) \leq C_\ell \leq \frac{C_\ell H_{L,p}}{\epsilon} R_L(u,p)$. □

Theorem (Constructive linear transfer, Theorem 5)**.** *Let $\ell : \mathcal{R} \to \mathbb{R}^{\mathcal{Y}}_+$ be a discrete target loss, $L : \mathbb{R}^d \to \mathbb{R}^{\mathcal{Y}}_+$ be a polyhedral surrogate loss, and $\psi : \mathbb{R}^d \to \mathcal{R}$ a link function. If (L, ψ) are consistent for ℓ, then*

$$(\forall h, \mathcal{D}) \quad R_\ell(\psi \circ h; \mathcal{D}) \leq \frac{C_\ell H_L}{\epsilon_\psi} R_L(h; \mathcal{D})\ .$$

The proof closely mirrors the proof of the nonconstructive upper bound, Theorem 1.

Proof. By Lemma 6, ψ is separated and ϵ_ψ well-defined. By Lemma 7, for each $p \in \mathcal{Q}$, $R_\ell(\psi(u), p) \leq \frac{C_\ell H_L}{\epsilon_\psi} R_L(u,p)$ for all u. Now consider a general p, which is in some full-dimensional

polytope level set Γ_u. Write $p = \sum_{q \in \mathcal{Q}_u} \beta(q) q$ for some probability distribution β, where $\mathcal{Q}_u$ is the set of vertices of Γ_u. By Lemma 3, R_L and R_ℓ are linear in p on Γ_u, so for any u',

$$\begin{aligned} R_\ell(\psi(u'), p) &= \sum_{q \in \mathcal{Q}_u} \beta(q) R_\ell(\psi(u'), q) \\ &\leq \sum_{q \in \mathcal{Q}_u} \beta(q) \frac{C_\ell H_{L,p}}{\epsilon_\psi} R_L(u', q) \\ &\leq \frac{C_\ell H_L}{\epsilon_\psi} \sum_{q \in \mathcal{Q}_u} \beta(q) R_L(u', q) \\ &\leq \frac{C_\ell H_L}{\epsilon_\psi} R_L(u', p). \end{aligned}$$

By Observation 1, this conditional regret transfer implies a full regret transfer, with the same constant. □