Trading off Consistency and Dimensionality of Convex Surrogates for Multiclass Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

In multiclass classification over n outcomes, we typically optimize some *surrogate loss* $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ assigning real-valued error to predictions in $\mathbb{R}^d$. In this paradigm, outcomes must be embedded into the reals with dimension $d \approx n$ in order to design a *consistent* surrogate loss. Consistent losses are well-motivated theoretically, yet for large n, such as in information retrieval and structured prediction tasks, their optimization may be computationally infeasible. In practice, outcomes are typically embedded into some $\mathbb{R}^d$ for $d \ll n$, with little known about their suitability for multiclass classification. We investigate two approaches for trading off consistency and dimensionality in multiclass classification while using a convex surrogate loss. We first formalize *partial consistency* when the optimized surrogate has dimension $d \ll n$. We then check if partial consistency holds under a given embedding and low-noise assumption, providing insight into when to use a particular embedding into $\mathbb{R}^d$. Finally, we present a new method to construct (fully) consistent losses with $d \ll n$ out of multiple problem instances. Our practical approach leverages parallelism to sidestep lower bounds on d.

1 Introduction

Multiclass classification, due to its combinatorial and discontinuous nature, is intractable to optimize directly, which drives machine learners to optimize some nicer *surrogate loss*. To ensure these surrogates properly "correspond" to the discrete classification task, we seek to design *consistent* surrogates. If one uses a consistent surrogate loss, in the limit of infinite data and model expressivity, one ends up with the same classifications as if one had solved the original intractable problem directly with probability 1.

Surrogate losses form the backbone of gradient-based optimization for classification tasks. Optimizing a surrogate is easier than direct optimization, but a large dimension d of the surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ can make gradient-based optimization intractable. Therefore, previous literature has operated under the premise that the prediction dimension d should be as low as possible, subject to consistency for the classification task [Ramaswamy and Agarwal, 2016, Finocchiaro et al., 2024, 2020]. For multi-class classification, the lower bound on d is $n - 1$ [Ramaswamy and Agarwal, 2016].

These previous works implicitly focus on a binary approach to consistency: a surrogate is either consistent for every possible label distribution, or it is not consistent. But there is a way out: lower bounds on the surrogate dimension d rely on edge-cases that rarely show up in reality [Ramaswamy and Agarwal, 2016]. As a result, practitioners are often willing to trade-off the guarantee of consistency in order to improve the computational tractability of optimization. However, we currently lack rigorous analysis tools to analyze many of the partially-consistent surrogates commonly used in practice. Thus, *unlike previous works, our work focuses on this more realistic paradigm of partial*

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

consistency. We apply our unique approach to rigorously analyze a popular surrogate construction that encompasses methods such as one-hot and binary encoding. Our approach allows for fine-grained control of the trade-off between consistency and dimension.

Prior works have informally brushed upon the proposed partial-consistency paradigm, without rigorous study. For example, Agarwal and Agarwal [2015] impose a low-noise assumption to construct a surrogate for classification with $d = \log(n)$. However, their work does not provide any way to control the consistency-dimension trade-off. Similarly, Struminsky et al. [2018] characterize the excess risk bounds of inconsistent surrogates, which teaches us about the learning rates for inconsistent surrogates, but not *under which distributional assumptions* we can recover consistency guarantees.

Using different techniques than both of these approaches, we seek to understand the tradeoffs of consistency, surrogate prediction dimension, and number of problem instances through the use of polytope embeddings which are common in the literature [Wainwright et al., 2008, Blondel et al., 2020]. When embedding outcomes into $d \ll n$ dimensions, we first show there always exists a set of distributions where *hallucinations* occur: where the report minimizing the surrogate leads to a prediction $\hat{y}$ such that the underlying true distribution has no weight on the prediction; that is, $Pr[Y = \hat{y}] = 0$ (Theorem 3). Following this, we show that every polytope embedding is partially consistent under strong enough low-noise assumptions (Theorem 5). Finally, we demonstrate through leveraging the embedding structure and multiple problem instances that the mode (in particular, a full rank ordering) over n outcomes embedded into a $\frac{n}{2}$ dimensional surrogate space is elicitable over all distributions via $O(n^2)$ problem instances (Theorem 10). This alternative approach to recovering consistency is parallelizable, detangling the complexity of gradient computation of one high-dimensional surrogate.

2 Background and Notation

Let $\mathcal{Y}$ be a finite label space, and throughout let $n = |\mathcal{Y}|$. Define $\mathbb{R}^{\mathcal{Y}}_{+}$ to be the nonnegative orthant. Let $\Delta_{\mathcal{Y}} = \{p \in \mathbb{R}^{\mathcal{Y}}_{+} \mid \|p\|_1 = 1\}$ be the set of probability distributions on $\mathcal{Y}$, represented as vectors. We denote the point mass distribution of an outcome $y \in \mathcal{Y}$ by $\delta_y \in \Delta_{\mathcal{Y}}$. Let $[d] := \{1, \dots, d\}$. In general, we denote a discrete loss by $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ with outcomes denoted by $y \in \mathcal{Y}$ and a surrogate loss by $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ with surrogate reports $u \in \mathbb{R}^d$ and outcomes $y \in \mathcal{Y}$. The surrogate must be accompanied by a link $\psi : \mathbb{R}^d \to \mathcal{Y}$ mapping the convex surrogate model's predictions back into the discrete target space, and we discuss consistency of a *pair* (L, ψ) with respect to the target ℓ.

For $\epsilon > 0$, we define an epsilon ball via $B_\epsilon(u) = \{u \in \mathbb{R}^d \mid \|u - x\|_2 < \epsilon\}$ and $B_\epsilon := B_\epsilon(\vec{0})$. Given a closed convex set $\mathcal{C} \subset \mathbb{R}^d$, we define a projection operation onto $\mathcal{C}$ via $\text{Proj}_{\mathcal{C}}(u) := \arg\min_{x \in \mathcal{C}} \|u - x\|_2$. Full tables of notation are found in Appendix A.

2.1 Property Elicitation, Consistency, and Prediction Dimension

Discrete label prediction requires optimization of a target loss function, ℓ, e.g. multi-class classification and 0-1 loss. When designing surrogate losses, consistency is the key notion of correspondence between surrogate and target loss. Intuitively, consistency implies that minimizing surrogate risk corresponds to solving the target problem. Finocchiaro et al. [2021] show that surrogate loss consistency is a necessary precursor to excess risk bounds and convergence rates.

Consistency is generally a difficult condition to work with directly. Hence, we will use the notion of *calibration*, which is equivalent to consistency in our setting with finite outcomes. Our approach follows from the property elicitation literature, which allows us to abstract away from the feature space $\mathcal{X}$ and focus on the conditional distributions over the labels, $p = \Pr[Y \mid X = x] \in \Delta_{\mathcal{Y}}$ [Bartlett et al., 2006, Zhang, 2004, Ramaswamy and Agarwal, 2016, Steinwart, 2007]. In this approach, the central object of study is a *property* which maps label distributions to reports that minimize the loss.

Definition 1 (Property, Elicits, Level Set)**.** *Let $\mathcal{R}$ be an arbitrary report set. For $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$, a property is a set-valued function $\Gamma : \mathcal{P} \to 2^{\mathcal{R}} \setminus \{\varnothing\}$, which we denote $\Gamma : \mathcal{P} \rightrightarrows \mathcal{Y}$. A loss $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ elicits the property Γ on $\mathcal{P}$ if*

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

If L elicits a property, it is unique and we denote it prop$[L]$. The level set of Γ for report r is the set $\Gamma_r := \{p \in \mathcal{P} \mid r = \Gamma(p)\}$. If prop$[L] = \Gamma$ and $|\Gamma(p)| = 1$ for all $p \in \mathcal{P}$, we say that L is strictly proper for Γ.

Once a model is optimized wrt. a surrogate L, it predicts reports in the surrogate space, $\mathbb{R}^d$. Then, to map surrogate reports to discrete labels, the surrogate loss must be paired with a link, $\psi : \mathbb{R}^d \to \mathcal{Y}$. Intuitively, a surrogate and link pair (L, ψ) are calibrated with respect to a target loss ℓ, if the optimal expected surrogate loss when making the *incorrect classification* (by ψ) is strictly greater than the optimal surrogate loss.

Definition 2 (ℓ-Calibrated Loss)**.** *Given discrete loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$, and link function $\psi : \mathbb{R}^d \to \mathcal{Y}$. We say that (L, ψ) is ℓ-calibrated over $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ if, for all $p \in \mathcal{P}$,*

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

If $\mathcal{P}$ is not specified, then we are discussing calibration over $\Delta_{\mathcal{Y}}$.

Our analysis crucially relies on the ability to specify $\mathcal{P}$ when invoking the definition of calibration. This is because the surrogates we analyze break the $d = n - 1$ lower bound on the dimension of any consistent surrogate loss. So the surrogates will not be calibrated over the whole simplex $\Delta_{\mathcal{Y}}$. To aid in our analysis, we use a condition that shows that converging to a property value implies calibration for the target loss itself [Agarwal and Agarwal, 2015].

Definition 3 (ℓ-Calibrated Property)**.** *Let $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$, $\Gamma : \mathcal{P} \rightrightarrows \mathbb{R}^d$, discrete loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, and $\psi : \mathbb{R}^d \to \mathcal{Y}$. We will say (Γ, ψ) is ℓ-calibrated for all $p \in \mathcal{P}$ and all sequences in $\{u_m\}$ in $\mathbb{R}^d$ if,*

$$u_m \to \Gamma(p) \Rightarrow \mathbb{E}_{Y \sim p}[\ell(\psi(u_m), Y)] \to \min_{r \in \mathcal{Y}} \mathbb{E}_{Y \sim p}[\ell(r, Y)] \; .$$

Theorem 1 ([Agarwal and Agarwal, 2015, Theorem 3])**.** *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ and $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$. Let $\Gamma : \mathcal{P} \rightrightarrows \mathbb{R}^d$ and $\psi : \mathbb{R}^d \to \mathcal{Y}$ be such that Γ is elicitable and (Γ, ψ) is an ℓ-calibrated property over $\mathcal{P}$. Let $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ be a convex function for all $y \in \mathcal{Y}$ and strictly proper for Γ i.e. prop$[L] = \Gamma$ and $|\Gamma(p)| = 1$ for all $p \in \mathcal{P}$. Then, (L, ψ) is ℓ-calibrated over $\mathcal{P}$.*

Finally, we present the 0-1 loss that we analyze, which is the target loss for multiclass classification.

Definition 4 (0-1 Loss)**.** *We denote the 0-1 loss by $\ell_{0-1} : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$ such that $\ell_{0-1}(y, \hat{y}) := \mathbb{1}_{y \neq \hat{y}}$. Observe $\gamma^{mode}(p) :=$ prop$[\ell_{0-1}](p) = \{y \in \mathcal{Y} | y \in \arg\max_y p_y\}$.*

3 Polytope Embedding and Existence of Calibrated Regions

Often, discrete outcomes are embedded in continuous space onto the vertices of the simplex via one-hot encoding, or the vertices of the unit cube via binary encoding [Seger, 2018]. Generalizing, we introduce an approach to surrogate construction inspired by Wainwright et al. [2008] and Blondel et al. [2020] that encompasses the aforementioned embedding methods. This construction utilizes embeddings onto arbitrary low-dimensional polytopes $\varphi : \mathcal{Y} \to \mathbb{R}^d$. Then, an embedding scheme naturally induces a large class of loss functions L^G_φ defined by the embedding, any G-Bregman Divergence, and a link function ψ^φ.

Our analysis begins by defining a condition stronger than inconsistency that arises when embedding into $d < n - 1$ dimensions for multiclass classification. To this end, we introduce the notion of *hallucination* as a means to characterize the "worst case" behavior of a surrogate pair (§ 3.2). In a positive manner, we characterize the *calibration regions* of various embeddings (§ 3.3), which are sets $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ such that our surrogate and link pair $(L^G_\varphi, \psi^\varphi)$ are ℓ-calibrated over $\mathcal{P}$. We refer the reader to the Appendix B for omitted full proofs.

3.1 Polytope Embedding Construction

A Convex Polytope $P \subset \mathbb{R}^d$, or simply a polytope, is the convex hull of a finite number of points $u_1, \ldots, u_n \in \mathbb{R}^d$. An extreme point of a convex set A, is a point $u \in A$ such that if $u = \lambda y + (1 - \lambda) z$ with $y, z \in A$ and $\lambda \in [0, 1]$, then $y = u$ and/or $z = u$. We shall denote by vert(P) a polytope's set of extreme points. A polytope can be expressed by the convex hull of its extreme points, i.e. $P = \text{conv}\,(\text{vert}(P))$ [Brondsted, 2012, Theorem 7.2]. Additional definitions pertaining to polytopes are used for proofs that are omitted to the appendix, we refer the reader to (§ B.1) for said definitions.

We propose the following embedding procedure that allows one to construct surrogate losses with almost *any* polytope, and *any* Bregman divergence.

Construction 1 (Polytope Embedding). *Given $\mathcal{Y}$ outcomes, $|\mathcal{Y}| = n$, choose a polytope $P \subset \mathbb{R}^d$ such that $|vert(P)| = n$. Choose a bijection between $\mathcal{Y}$ and $vert(P)$. According to this bijection, assign each vertex a unique outcome so that $\{v_y | y \in \mathcal{Y}\} = vert(P)$. Then the polytope embedding $\varphi : \Delta_{\mathcal{Y}} \to P$ is $\varphi(p) := \sum_{y \in \mathcal{Y}} p_y v_y$, which is the sum of p-scaled vectors*

Following the work of Blondel [2019] and their proposed Projection-based losses, we use the extremely general class of Bregman divergences (Definition 5) and a polytope embedding φ to define an induced loss L^G_φ (Definition 6).

Definition 5 (Bregman Divergence). *Given a strictly convex function $G : \mathbb{R}^d \to \mathbb{R}$, $D_G(u, v) := G(v) - [G(u) + \langle dG_v, u - v\rangle]$ is a Bregman divergence where dG_v denotes a subgradient of G at v. For this work, we shall always assume that* $\mathrm{dom}(G) = \mathbb{R}^d$.

Definition 6 ((D_G, φ) Induced Loss). *Given a Bregman divergence D_G and a polytope embedding φ, we say (D_G, φ) induces a loss $L^G_\varphi : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ defined as $L^G_\varphi(u, y) := D_G(u, v_y) = G(v_y) - [G(u) + \langle dG_{v_y}, u - v_y\rangle]$.*

We show that for any $p \in \Delta_{\mathcal{Y}}$, the report that uniquely minimizes the expectation of the loss L^G_φ is $\varphi(p)$, the embedding point of p. Furthermore, the polytope P contains all of, and only the minimizing reports in expectation under L^G_φ.

Proposition 2. *For a given induced loss L^G_φ, the unique report which minimizes the expected loss is $u^* := \arg\min_{u \in \mathbb{R}^d} \mathbb{E}_{Y\sim p}[L^G_\varphi(u, Y)] = \varphi(p)$ such that $u^* \in P$. Furthermore, every $\hat{u} \in P$ is a minimizer of $\mathbb{E}_{Y\sim\hat{p}}[L^G_\varphi(u, Y)]$ for some $\hat{p} \in \Delta_{\mathcal{Y}}$.*

We now define the maximum a posteriori (MAP) link, which will be used in conjunction with an induced loss L^G_φ to form a surrogate pair for the 0-1 loss. The MAP link projects surrogate predictions onto the polytope P, then links to the nearest vertex of P, and is commonly used in the literature [Tsochantaridis et al., 2005, Blondel, 2019, Xue et al., 2016].

Definition 7 (MAP Link). *Let φ be a polytope embedding. The MAP link $\psi^\varphi : \mathbb{R}^d \to \mathcal{Y}$ is defined as $\psi^\varphi(u) = \arg\min_{y \in \mathcal{Y}} ||Proj_P(u) - v_y||_2$ The level set of the link for y is $\psi^\varphi_y = \{u \in \mathbb{R}^d | y = \psi^\varphi(u)\}$. We break ties arbitrarily but deterministically.*

3.2 Hallucination Regions

Since our polytope embedding violates surrogate dimension bounds, calibration for 0-1 loss will not hold for all distributions. In particular, we show there always exists some distribution p such that $p_y = 0$ yet $\mathbb{E}_{Y\sim p} L^G_\varphi(u, Y)$ is minimized at some u such that $\psi^P(u) = y$. This implies a "worst case" inconsistency where the reported outcome could never actually occur with respect to our embedding of n events via φ into vert(P).

Definition 8 (Hallucination). *Given (L, ψ) such that $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$, $|\mathcal{Y}| = n$, $d < n$, and $\psi : \mathbb{R}^d \to \mathcal{Y}$, we say that a hallucination occurs at a surrogate report $u \in \mathbb{R}^d$ if, for some p, $u \in \arg\min_{\hat{u} \in \mathbb{R}^d} \mathbb{E}_{\mathcal{Y}\sim p}[L(\hat{u}, Y)]$ and $\psi(u) := y$ but $p_y = 0$. We denote by $\mathcal{H} \subseteq P \subset \mathbb{R}^d$ as the hallucination region as the elements of P at which hallucinations can occur.*

We express the subspace of the surrogate space where hallucinations can occur as the hallucination region denoted by $\mathcal{H}$. In Theorem 3, we characterize the hallucination region for any polytope embedding while using the surrogate pair $(L^G_\varphi, \psi^\varphi)$ and show that $\mathcal{H}$ is never empty.

Theorem 3. *For any given pair $(L^G_\varphi, \psi^\varphi)$ and ℓ_{0-1} with embedding dimension $d < n - 1$; it holds that $\mathcal{H} = \cup_{y \in \mathcal{Y}} \mathrm{conv}\,(vert(P) \setminus \{v_y\}) \cap \psi^\varphi_y$ and furthermore $\mathcal{H} \neq \varnothing$.*

Sketch. Fix $y \in \mathcal{Y}$. We abuse notation and write vert$(P_{-y}) :=$ vert$(P) \setminus \{v_y\}$. Observe $\mathrm{conv}\,(\text{vert}(P_{-y})) \cap \psi^\varphi_y \subseteq \mathcal{H}$ since any point in this set can be expressed as a convex combination without needing vertex v_y implying there is a distribution embedded by φ to said point which has no weight on y. To show that $\mathcal{H} \subseteq \cup_{y \in \mathcal{Y}} \mathrm{conv}\,(\text{vert}(P_{-y})) \cap \psi^\varphi_y$. Assume there exists a point $u \notin \mathrm{conv}\,(\text{vert}(P) \setminus v_y) \cap \psi^\varphi_y$ such that there exists some $p \in \Delta_{\mathcal{Y}}$ where $\varphi(p) = u$, $p_y = 0$, and $\psi^\varphi(u) = y$. Since $\psi^\varphi(u) = y$ and $u \notin \mathrm{conv}\,(\text{vert}(P_{-y})) \cap \psi^\varphi_y$, it must be the case that

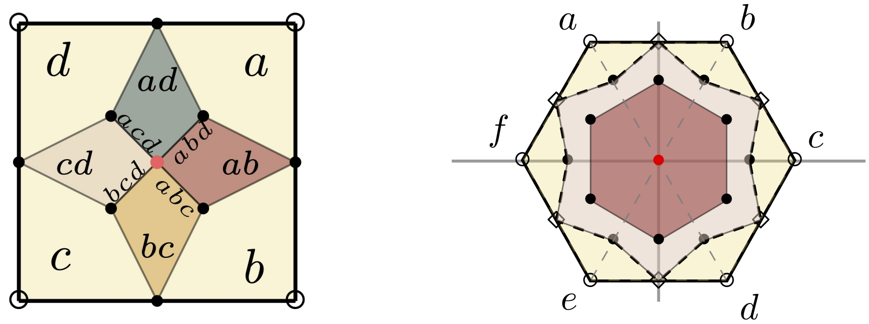

Figure 1: (Left) Mode level sets of $\Delta_{\mathcal{Y}}$ where $\mathcal{Y} = \{a, b, c, d\}$ embedded into a two dimensional unit cube. The center red point denotes the origin $(0, 0)$ which is the hallucination region. (Right) An embedding of $\Delta_{\mathcal{Y}}$ where $\mathcal{Y} = \{a, b, c, d, e, f\}$ into a three-dimensional permutahedron: the beige region expresses strict calibration regions, the light pink regions expresses regions with inconsistency, and the auburn region expresses regions with hallucinations. For example, consider the report $u = \vec{0}$. Since losses are convex, if $p = (0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0)$, then $\mathrm{conv}(\{b, e\})$ (dashed grey) is optimal, which includes u. However, $\vec{0}$ is also contained in $\mathrm{conv}(\{a, d\})$ which is optimal for the distribution $p' = (\frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0)$. Therefore, we cannot distinguish the optimal reports for a hallucination at $\vec{0}$.

$u \notin \mathrm{conv}(\mathrm{vert}(P_{-y}))$. However, that implies that u is strictly in the vertex figure and thus must have weight on the coefficient for y. Thus, forming a contradiction that $p_y = 0$ which implies that $\mathcal{H} \subseteq \cup_{y \in \mathcal{Y}} \mathrm{conv}(\mathrm{vert}(P_{-y})) \cap \psi_y^{\varphi}$. Finally, using Helly's Theorem [Rockafellar, 1997, Corollary 21.3.2], we are able to show the non-emptiness of $\mathcal{H}$. $\square$

Theorem 3 suggests that using machine learning in high-risk settings such as medical and legal applications while violating the known $n-1$ dimensional bound for surrogate losses in multiclass classification is inherently ill-advised without human intervention given the possibility for hallucinations. Furthermore, hallucinations may be forced by the target loss, as in the case of Hamming loss (see Appendix C). In these cases practitioners should carefully consider the choice of target loss. We conjecture that hallucinations are common for many structured prediction losses. However this is not a concern in our primary loss of study of multi-class classification.

3.3 Calibration Regions

Ideally, we would like calibration to hold over the entire simplex since that would imply minimizing surrogate risk would always correspond to solving the target problem regardless of the true underlying distribution. We observe that the mode's embedded level sets in the polytope overlap (see Figure 1L), which is unsurprising given that we are violating the lower bounds on surrogate prediction for the mode and hence calibration does not hold over the entire simplex. Since $|2^{\mathcal{Y}} \setminus \{\varnothing\}|$ is a finite set, we know that the number of unique mode level sets is finite. Although every point in the polytope is a minimizing report for some distribution, if multiple distributions with non-intersecting mode sets are embedded to the same point, there is no way to define a link function that is correct in all cases. However, if the union of mode sets for the p's mapped to any $u \in P$ is a singleton, regardless of the underlying distribution[*], a link ψ would be calibrated over the union if it mapped u to the mentioned singleton. Given (L, ψ), φ, and a target loss ℓ, we define strict calibrated regions as the points for which calibration holds regardless of the actual distribution realized, which are possible at said points.

Definition 9 (Strict Calibrated Region)**.** *Suppose we are given (L, ψ), φ, and a target loss ℓ. We say $R \subseteq P$ is a* strict calibrated region *via (L, ψ) with respect to ℓ if (L, ψ) is ℓ-calibrated for all $p \in \varphi^{-1}(R) := \{p : \varphi(p) \in R\}$.*

For any $y \in \mathcal{Y}$, we define $R_y := R \cap \psi_y$. We let $R_{\mathcal{Y}} := \cup_{y \in \mathcal{Y}} R_y$.

By violating lower bounds, we are in a partially consistent paradigm where surrogate reports do not necessarily correspond to a unique distribution p. However, strict calibration regions allow us to

[*]We leave the more general case of linking u when $\bigcap_{p \in \varphi^{-1}(u)} \gamma(p) \neq \varnothing$ to future work.

check whether or not the loss is calibrated for the distribution p generating the data — even without explicit access to p. One simply has to check whether the report u is in $R_{\mathcal{Y}}$.

In Theorem 4, regardless of one's chosen P, we show that there always exists a non-zero Lebesgue measurable strict calibration region and that $(L_{\varphi}^{G}, \psi^{\varphi})$ is calibrated for the 0-1 loss overall distributions embedded into the strict calibration region. This result shows that our surrogate and link construction for *any* d, always yields calibration regions in a robust sense — lending support to the practical use and study of these surrogates.

Theorem 4. *Let D_G be a Bregman divergence, φ be any polytope embedding, ψ^{φ} be the MAP link, and L_{φ}^{G} be the loss induced by (D_G, φ). There exists a $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ with non-zero Lebesgue measure and $\varphi(\mathcal{P}) \subseteq R_{\mathcal{Y}}$ via $(L_{\varphi}^{G}, \psi^{\varphi})$ with respect to ℓ_{0-1}.*

Although strict calibration regions R_y exist for each outcome $y \in \mathcal{Y}$ via the polytope embedding, tightly characterizing strict calibration regions is non-trivial. Since the level sets of elicitable properties are convex within the underlying simplex, characterizing the strict calibration regions becomes a collision detection problem, which is often computationally hard.

4 Restoring Inconsistent Surrogates via Low-Noise Assumptions

Looking towards application, we refine our results on the existence of strict calibration regions by examining a low-noise assumption, which provides an interpretable calibration region (§ 4.1). We show which low-noise assumptions imply calibration when embedding 2^d outcomes into d dimensions and $d!$ outcomes into d dimensions (§ 4.2). We refer the reader to Appendix B for omitted proofs.

4.1 Calibration via Low Noise Assumptions

We demonstrate that every polytope embedding leads to calibration under some low-noise assumption. Our results enable practictioners to choose the dimension d, unlike in previous works. Following previous work [Agarwal and Agarwal, 2015], we define a low noise assumption to be a subset of the probability simplex with low noise on the label distribution parameterized by $\hat{\alpha}$: $\Theta_{\hat{\alpha}} = \{p \in \Delta_{\mathcal{Y}} \mid \max_{y\in\mathcal{Y}} p_y \geq 1 - \hat{\alpha}\}$ where $\hat{\alpha} \in [0, 1]$. Given $\alpha \in (0, 1]$ and $y \in \mathcal{Y}$, we define the set $\Psi_{\alpha}^{y} = \{(1 - \alpha)\delta_y + \alpha\delta_{\hat{y}} \mid \hat{y} \in \mathcal{Y}\}$. With an embedding φ onto P, we define the set $P_{\alpha}^{y} := \varphi(\text{conv}(\Psi_{\alpha}^{y}))$, a scaled version of P anchored at v_y, that moves vertices $(1 - \alpha)$ towards y, (Figure 2R).

Theorem 5. *Let D_G be a Bregman divergence, φ be any polytope embedding, and L_{φ}^{G} be the loss induced by (D_G, φ). There exists an $\alpha \in [0, .5)$ such that for the link $\psi_{\alpha}^{\varphi}(u) = \arg\min_{y\in\mathcal{Y}} \|u - P_{\alpha}^{y}\|_2$, $(L_{\varphi}^{G}, \psi_{\alpha}^{\varphi})$ is ℓ_{0-1}-calibrated over the distributions $\Theta_{\alpha} := \{p \in \Delta_{\mathcal{Y}} \mid \max_{y\in\mathcal{Y}} p_y \geq 1 - \alpha\}$.*

Proof. **Part 1 (Choosing $\alpha \in [0, .5)$)**: By Theorem 4, there exists an $\epsilon > 0$ such that $B_{\epsilon}(v_y) \cap P \subseteq R_y$ for all $y \in \mathcal{Y}$. Given that vert(P) are unique points, there exists a sufficiently small $\epsilon' > 0$ such that $B_{\epsilon'}(v) \cap B_{\epsilon'}(\hat{v}) = \varnothing$ for all $v, \hat{v} \in \text{vert}(P)$ where $v \neq \hat{v}$. Let $\epsilon'' = \min(\epsilon, \epsilon')$. For any $y \in \mathcal{Y}$, observe the set $\text{conv}(\Psi_{\alpha}^{y})$, defined using any $\alpha \in [0, .5)$, is a scaled-down translated unit simplex and that for all $p \in \text{conv}(\Psi_{\alpha}^{y}) \subset \Delta_{\mathcal{Y}}$ it holds that $y = \text{mode}(p)$.

We shall show that for some sufficiently small $\alpha \in [0, .5)$, P_{α}^{y} is a scaled down version of P positioned at the respective vertex v_y. Furthermore, we shall show that $P_{\alpha}^{y} \subset B_{\epsilon''}(v_y) \cap P \subseteq R_y$ for all $y \in \mathcal{Y}$. Observe that by linearity of φ,

$$P_{\alpha}^{y} := \varphi(\text{conv}(\Psi_{\alpha}^{y})) = \text{conv}(\varphi(\{(1-\alpha)\delta_y + \alpha\delta_{\hat{y}} | \hat{y} \in \mathcal{Y}\})) = \text{conv}(\{(1-\alpha)v_y + \alpha v_{\hat{y}} | \hat{y} \in \mathcal{Y}\})$$

and hence, P_{α}^{y} is a scaled version of P positioned at v_y. Hence for some sufficiently small α, $(1 - \alpha)v_y + \alpha v_{\hat{y}} \in B_{\epsilon''}(v_y)$ for all $\hat{y}$ and hence $P_{\alpha}^{y} \subseteq B_{\epsilon''}(v_y) \subseteq R_y$. With said sufficiently small α, define ψ_{α}^{P} and the respective sets $\text{conv}(\Psi_{\alpha}^{y})$ for each $y \in \mathcal{Y}$. Using the previous α, define the set Θ_{α} as well.

Part 2 (Showing Calibration): Recall, by Proposition 2, for any $p \in \Delta_{\mathcal{Y}}$, $u = \varphi(p)$ minimizes the expected surrogate loss $\mathbb{E}_{\mathcal{Y}\sim p}[L_{\varphi}^{G}(u, Y)]$. For any fixed $y \in \mathcal{Y}$, observe that $\text{conv}\{(1-\alpha)\delta_y + \alpha\delta_{\hat{y}} \mid$

$\hat{y} \in \mathcal{Y}\} = \{p : p_y \geq 1 - \alpha\} \subset \Delta_{\mathcal{Y}}$ and hence, by Proposition 2, $\cup_{y \in \mathcal{Y}} P^y_\alpha$ contains all of the minimizing surrogate reports with respect to Θ_α. By our choice of α and the construction of ψ^P_α, every $u \in \cup_{y \in \mathcal{Y}} P^y_\alpha$ is linked to the proper unique mode outcome since $\cup_{y \in \mathcal{Y}} P^y_\alpha \subseteq R_{\mathcal{Y}}$. Assuming a low-noise condition where $p \in \Theta_\alpha$, any $u \notin \cup_{y \in \mathcal{Y}} P^y_\alpha$ is never optimal for any low-noise distribution. In such cases, we project the point to the nearest P^y_α as a matter of convention. Given that calibration is a result pertaining to minimizing reports, this design choice is non-influential. Finally, since every $\cup_{y \in \mathcal{Y}} P^y_\alpha \subseteq R_{\mathcal{Y}}$, by the definition of strict calibration region, it holds that $(L^G_\varphi, \psi^\varphi_\alpha)$ is ℓ_{0-1}-calibrated for Θ_α. □

4.2 Embedding into the Unit Cube and Permutahedron under Low-Noise

In this section, we demonstrate embedding onto the unit cube and the permutahedron [Blondel et al., 2020, Seger, 2018]. We show that by embedding 2^d outcomes into a d dimensional unit cube $P^\square$, $(L^G_\varphi, \psi^{P^\square}_\alpha)$ is calibrated over Θ_α for all $\alpha \in [0, \frac{1}{2})$. Furthermore, we found that by embedding $d!$ outcomes into a d dimensional permutahedron P^w, $(L^G_\varphi, \psi^{P^w}_\alpha)$ is calibrated for Θ_α for $\alpha \in (0, \frac{1}{d})$. Theorem 6 enables us to simultaneously study the aforementioned embeddings.

Theorem 6. *Let D_G be a Bregman divergence, φ be any polytope embedding, and L^G_φ be the loss induced by (D_G, φ). Fix $\alpha \in [0, .5)$ and with it define Θ_α. If for all $y, \hat{y} \in \mathcal{Y}$ such that $y \neq \hat{y}$ it holds that $P^y_\alpha \cap P^{\hat{y}}_\alpha = \varnothing$, then $(L^G_\varphi, \psi^\varphi_\alpha)$ is ℓ_{0-1}-calibrated for Θ_α where $\psi^\varphi_\alpha(u) = \arg\min_{y \in \mathcal{Y}} \|u - P^y_\alpha\|_2$.*

Proof. Pick an α such that for all $y, \hat{y} \in \mathcal{Y}$, $P^y_\alpha \cap P^{\hat{y}}_\alpha = \varnothing$. Define Θ_α and ψ^P_α accordingly. For $p \in \Theta_\alpha$ and some $y \in \mathcal{Y}$, say a sequence $\{u_m\}$ converges to prop$[L^G_\varphi](p) = \varphi(p) \in P^y_\alpha$, where the equality follows from Proposition 2. Given that each P^y_α is closed and pairwise disjoint, there exists some $\hat{\epsilon} > 0$ such that for all $y, \hat{y} \in \mathcal{Y}$ where $y \neq \hat{y}$, it also holds that $(P^y_\alpha + B_{\hat{\epsilon}}) \cap (P^{\hat{y}}_\alpha + B_{\hat{\epsilon}}) = \varnothing$ where $+$ denotes the Minkowski sum. Since $\{u_m\}$ converges to $\varphi(p)$, there exists some $N \in \mathbb{N}$ such that for all $n \geq N$, $\|u_n - \varphi(p)\|_2 < \hat{\epsilon}$. By the definition of ψ^φ_α, any u_n where $n \geq N$ will be mapped to y, the correct unique report given that prop$[L^G_\varphi](p) \in P^y_\alpha$. Hence, (prop$[L^G_\varphi], \psi^\varphi_\alpha)$ is ℓ_{0-1}-calibrated property with respect to Θ_α. Finally, since L^G_φ is strictly proper for prop$[L^G_\varphi]$, by Theorem 1, we have that $(L^G_\varphi, \psi^\varphi_\alpha)$ is ℓ_{0-1}-calibrated for Θ_α. □

Unit Cube Define a unit cube in d-dimensions by $P^\square := \text{conv}\,(\{-1, 1\}^d)$. Binary encoding outcomes into the elements of $\{-1, 1\}^d$ (the vertices of a unit cube) is a commonly used method in practice (e.g., [Seger, 2018, Yu and Blaschko, 2018]). We show that calibration holds under a low noise assumption of Θ_α when $\alpha < .5$.

Corollary 7. *Let φ be an embedding from 2^d outcomes into the vertices of $P^\square$ in d-dimensions and define an induced loss L^G_φ. Fix $\alpha \in [0, .5)$ and define Θ_α. $(L^G_\varphi, \psi^{P^\square}_\alpha)$ is ℓ_{0-1}-calibrated for Θ_α.*

Corollary 7 suggests that binary encoding is an appropriate methodology when one has a prior over the data that the mode of the label distribution $\Pr[Y \mid X = x]$ is greater than half for all $x \in \mathcal{X}$. Interestingly, the bound of α is not dependent on the dimension of d. We now present a result for embedding outcomes into a factorially lower dimension via the permutahedron. Intuitively, ranking can be recast as a multiclass classification problem, in which case the outcomes are orderings of the d possible labels.

Permutahedron Let $\mathcal{S}_d$ express the set of permutations on $[d]$. The permutahedron associated with a vector $w \in \mathbb{R}^d$ is defined to be the convex hull of the permutations of the indices of w, i.e., $P^w := \text{conv}\,\{\pi(w) \mid \pi \in \mathcal{S}_d\} \subset \mathbb{R}^d$. The permutahedron may serve as an embedding from $d!$ outcomes into d-dimensions; it is a natural choice for embedding full rankings over d items.

Corollary 8. *Let φ be an embedding from $d!$ outcomes into the vertices of P^w in d dimensions such that $w = (0, \frac{1}{\beta d}, \frac{2}{\beta d}, \ldots, \frac{d-1}{\beta d}) \in \mathbb{R}^d$ where $\beta = \frac{d-1}{2}$. Fix $\alpha \in (0, \frac{1}{d})$. Then $(L^G_\varphi, \psi^{P^w}_\alpha)$ is ℓ_{0-1}-calibrated over Θ_α.*

The calibration region in Corollary 8 show that consistency in Θ_α shrinks exponentially in d. Unless one has a prior that the data follows some form of a power distribution, Corollary 8 suggests not to factorially embed outcomes.

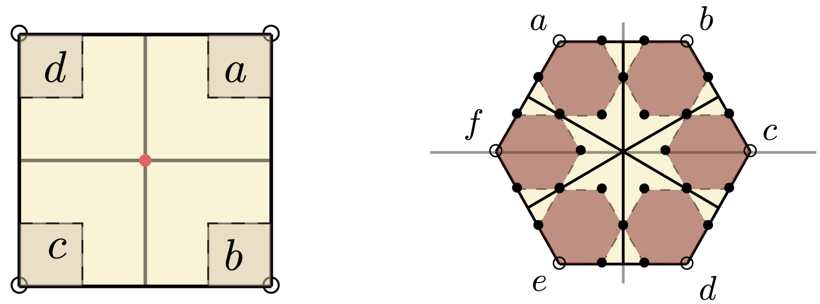

Figure 2: (Left) Corners represent the strict calibration regions for Θ_α where $\mathcal{Y} = \{a, b, c, d\}$ is embedded into a two dimensional unit cube such that $\alpha = .25$. (Right) Auburn regions show that strict calibration holds for Θ_α where $\mathcal{Y} = \{a, b, c, d, e, f\}$ is embedded into a three-dimensional permutahedron such that $\alpha = \frac{1}{3} - \epsilon$.

5 Elicitation in Low Dimensions with Multiple Problem Instances

The tools developed in previous sections now enable us to address the setting in which we require full consistency, $\mathcal{P} = \Delta_{\mathcal{Y}}$, but also desire surrogate prediction dimension $d \ll n-1$. We side-step the $n-1$ lower bound by utilizing multiple problem instances and aggregation of the outputs. Although cumulatively we have a larger surrogate prediction dimension than $n-1$, each individual problem instance has a less than $n-1$ surrogate prediction dimension. This approach is well-motivated since it allows for distributed computing of separate, smaller models which leads to faster convergence overall since in general optimization is at least $poly(d)$.

Definition 10. *Extending Definition 1, we say a loss and link pair (L, ψ), where $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ and $\psi : \mathbb{R}^d \to \mathcal{Y}$, elicits a property $\Gamma : \mathcal{P} \rightrightarrows \mathcal{Y}$ on $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ if $\forall\ p \in \mathcal{P}$, $\Gamma(p) = \psi(\arg\min_{u \in \mathbb{R}^d} \mathbb{E}_{Y \sim p}[L(u, Y)])$.*

Definition 11 ((n, d, m)-Polytope Elicitable)**.** *Suppose we are given a property $\gamma : \mathcal{P} \rightrightarrows \mathcal{Y}$ such that $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ and $|\mathcal{Y}| = n$ finite outcomes. Say we have m unique polytope embeddings $\{\varphi_j : \Delta_{\mathcal{Y}} \to \mathbb{R}^d\}_{j=1}^m$ where $d < n-1$, and a set of induced losses $\{L^G_{\varphi_j}\}_{j=1}^m$ and links $\psi_j : \mathbb{R}^d \to \mathcal{B}_j$ defined wrt. φ_j, where $\mathcal{B}_j$ is an arbitrary report set. For each $j \in [m]$, assume the pair $(L^G_{\varphi_j}, \psi_j)$ elicits the property $\Gamma_j : \mathcal{P} \rightrightarrows \mathcal{B}_j$. If there exists a function $\Upsilon : \mathcal{B}_1 \times \cdots \times \mathcal{B}_m \rightrightarrows \mathcal{Y}$ such that for any $p \in \Delta_{\mathcal{Y}}$ it holds that $\Upsilon(\Gamma_1(p), \ldots, \Gamma_m(p)) = \gamma(p)$, we say that γ is (n, d, m)-Polytope Elicitable over $\mathcal{P}$.*

Equivalently, we will also say that the pair $(\{(L^G_{\varphi_j}, \psi_j)\}_{j=1}^m, \Upsilon)$ (n, d, m)-Polytope elicits the property γ with respect to $\mathcal{P}$.

We shall express a d-cross polytope by $P^{\oplus} := \operatorname{conv}(\{\pi((\pm 1, 0, \ldots, 0)) \mid \pi \in \mathcal{S}_d\})$ where $(\pm 1, 0, \ldots, 0) \in \mathbb{R}^d$. Observe that a d-cross polytope has $2d$ vertices. For any vertex of a d-cross polytope $v \in \text{vert}(P^{\oplus})$, we shall say that $(v, -v)$ forms a diagonal vertex pair.

Lemma 9. *Say we are given a cross-polytope embedding $\varphi : \Delta_{2d} \to P^{\oplus}$ and induced loss L^G_{φ}. Let (v_{a_i}, v_{b_i}), be the i^{th} diagonal pair (i.e. $\varphi(\delta_{a_i}) = v_{a_i}$). Define the property $\Gamma^{\varphi} : \Delta_{2d} \to \mathcal{B}$ element-wise by*

$$\Gamma^{\varphi}(p)_i := \begin{cases} (<, a_i, b_i) & \text{if } p_{a_i} < p_{b_i} \\ (>, a_i, b_i) & \text{if } p_{a_i} > p_{b_i} \\ (=, a_i, b_i) & \text{if } p_{a_i} = p_{b_i}. \end{cases}$$

Furthermore define the link $\psi^{P^{\oplus}} : \mathbb{R}^d \to \mathcal{B}$ with respect to each diagonal pair as

$$\psi(u; v_{a_i}, v_{b_i})_i^{P^{\oplus}} := \begin{cases} (<, a_i, b_i) & \text{if } ||u - v_{a_i}||_2 > ||u - v_{b_i}||_2 \\ (>, a_i, b_i) & \text{if } ||u - v_{a_i}||_2 < ||u - v_{b_i}||_2 \\ (=, a_i, b_i) & \text{o.w.} \end{cases}$$

Then $(L^G_{\varphi}, \psi^{P^{\oplus}})$ elicits Γ^{φ}.

The following theorem states that by using multiple problem instances, based on Lemma 9, we can Polytope-elicit the mode. Algorithm 1 outlines how to aggregate the individual solutions to infer the mode. We defer the proof to Appendix B.

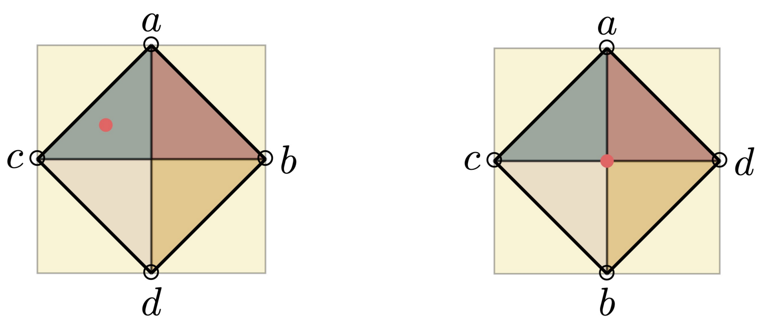

Figure 3: Four outcomes embedded in $\mathbb{R}^2$ in two different ways, with the minimizing reports • for a distribution p." (Left) Configuration φ_1 with • at $(-.5, .3)$ implying $p_a > p_d$ and $p_b > p_c$. (Right) Configuration φ_2 with • at $(0, 0)$ implying $p_a = p_b$ and $p_c = p_d$. This implies the true distribution is $p = (0.4, 0.4, 0.1, 0.1)$."

Theorem 10. *Let $d \geq 2$. The mode is $(2d, d, m)$-Polytope Elicitable for some $m \in [2d-1, d(2d-1)]$.*

Algorithm 1 Elicit mode via comparisons and the d-Cross Polytopes

---

Require: $M = \{(L^G_{\varphi_j}, \psi_j^{P^\oplus})\}_{j=1}^m$

Learn a model $h_j : \mathcal{X} \to \mathbb{R}^d$ for each instance $(L^G_{\varphi_j}, \psi_j^{P^\oplus}) \in M$

For some fixed $x \in \mathcal{X}$, collect all $B_j \leftarrow \psi_j^{P^\oplus}(h_j(x))$ where $B_j \in \mathcal{B}_j$

Report $R \leftarrow \text{FindMaxes}^\dagger(B_1, \ldots, B_m)$

Although Theorem 10 states that the mode is $(2d, d, m)$-Polytope Elicitable for some $m \in [2d - 1, d(2d-1)]$, it does not state how we select said $\{(L^G_{\varphi_j}, \psi_j^{P^\oplus})\}_{j=1}^m$ problem instances in an optimal manner. Unfortunately, selecting the min number of problem instances reduces to a a minimum set cover problem which is computationally hard. Even so, through a greedy approach, one can choose problem instances that are log approximate optimal relative to the true best configuration. In practice using real data, given that these are asymptotic results, we may have conflicting logic for the provided individual reports. In Appendix D, we discuss an approach of how to address this in practice.

6 Discussion and Conclusion

This work examines various tradeoffs between surrogate loss dimension, restricting the region of consistency in the simplex when using the 0-1 loss, and number of problem instances. Since our analysis is based on an embedding approach commonly used in practice, our work provides theoretical guidance for practitioners choosing an embedding. We see several possible future directions. The first is a deeper investigation into hallucinations. Future work could investigate the size of the hallucination region in theory, and the frequency of reports in the hallucination region in practice. Another direction would be to construct a method that efficiently identifies the strict calibration regions and the distributions embedded into them. This would provide better guidance on whether or not a particular polytope embedding aligns with one's prior over the data. Finally, another direction is to identify other properties that can be elicited via multiple problem instances while also reducing the dimension of any one instance.

Broader Impacts: Our work broadly informs the selection of loss functions for machine learning. Thus our work may influence practitioners' choice of loss function. Of course, such loss functions can be used for ethical or unethical purposes. We do not know of particular risks of negative impacts of this work beyond risks of machine learning in general.

[†]Given all comparisons, a sorting algorithm can be used to compute the set of $r \in \mathcal{Y}$ such that p_r is maximum.

References

Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *Conference on Learning Theory*, pages 4–22. PMLR, 2015.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005.

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.

Mathieu Blondel. Structured prediction with projection oracles. *Advances in neural information processing systems*, 32, 2019.

Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning with fenchel-young losses. *The Journal of Machine Learning Research*, 21(1):1314–1382, 2020.

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

Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. Embedding dimension of polyhedral losses. In *Conference on Learning Theory*, pages 1558–1585. PMLR, 2020.

Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. Unifying lower bounds on prediction dimension of consistent convex surrogates. *arXiv preprint arXiv:2102.08218*, 2021.

Jessie Finocchiaro, Rafael M Frongillo, and Bo Waggoner. An embedding framework for the design and analysis of consistent polyhedral surrogates. *Journal of Machine Learning Research*, 25(63): 1–60, 2024.

Peter M Gruber. *Convex and discrete geometry*, volume 336. Springer, 2007.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

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

R Tyrrell Rockafellar. *Convex analysis*, volume 11. Princeton university press, 1997.

Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Kirill Struminsky, Simon Lacoste-Julien, and Anton Osokin. Quantifying learning guarantees for convex but inconsistent surrogates. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Yexiang Xue, Zhiyuan Li, Stefano Ermon, Carla P Gomes, and Bart Selman. Solving marginal map problems with np oracles and parity constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

Jiaqian Yu and Matthew B Blaschko. The lovász hinge: A novel convex surrogate for submodular losses. *IEEE transactions on pattern analysis and machine intelligence*, 42(3):735–748, 2018.

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

A Notation tables

Notation	Explanation
$r \in \mathcal{Y}$	Prediction space
$y \in \mathcal{Y}$	Label space
$\Delta_{\mathcal{Y}}$	Simplex over $\mathcal{Y}$
$[d] := \{1, \ldots, d\}$	Index set
$\mathbb{1}_S \in \{0,1\}^d$ s.t. $(\mathbb{1}_S)_i = 1 \Leftrightarrow i \in S$	0-1 Indicator on set $S \subseteq [d]$
$\mathcal{C} \subset \mathbb{R}^d$	Closed convex set
$u \in \mathbb{R}^d$	Surrogate prediction space
$\text{Proj}_{\mathcal{C}}(u) := \arg\min_{x \in \mathcal{C}} \lVert u - x \rVert_2$	Projection onto closed convex set
$\pi \in \mathcal{S}_d$	Permutations of $[d]$
$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$	Discrete loss
$L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$	Surrogate loss
$\mathbb{E}_{Y \sim p}[\ell(r, Y)]$	Expected discrete loss
$\mathbb{E}_{Y \sim p}[L(u, Y)]$	Expected surrogate loss

Table 1: Table of general notation

Notation	Explanation
$P \subset \mathbb{R}^d$	Polytope
$P^{\square} := \text{conv}\left(\{-1, 1\}^d\right)$	Unit cube
$P^{w} := \text{conv}\left\{\pi \cdot w \mid \pi \in \mathcal{S}_d\right\} \subset \mathbb{R}^d$ s.t. $w \in \mathbb{R}^d$	Permutahedron
$P^{\oplus} := \text{conv}\left(\{\pi((\pm 1, 0, \ldots, 0)) \mid \pi \in \mathcal{S}_d\}\right)$	Cross polytope

Table 2: Table of polytope notation

B Polytopes, Omitted Proofs, and Results

B.1 Polytopes

A Convex Polytope $P \subset \mathbb{R}^d$, or simply a polytope, is the convex hull of a finite number of points $u_1, \ldots, u_n \in \mathbb{R}^d$. An extreme point of a convex set A, is a point $u \in A$ such that if $u = \lambda y + (1-\lambda)z$ with $y, z \in A$ and $\lambda \in [0,1]$, then $y = u$ and/or $z = u$. We shall denote by $\text{vert}(P)$ a polytope's set of extreme points. A polytope can be expressed by the convex hull of its extreme points, i.e. $P = \text{conv}\,(\text{vert}(P))$ [Brondsted, 2012, Theorem 7.2].

We define the dimension of P via $\dim(P) := \dim(\text{affhull}(P))$ where $\text{affhull}(P)$ denotes the smallest affine set containing P. A set $F \subseteq P$ is a face of P is there exists a hyperplane $H(y,\alpha) := \{u \in \mathbb{R}^d \mid \langle u, y\rangle = \alpha\}$ such that $F = P \cap H$ and $P \subseteq H^+$ such that $H^+(y,\alpha) := \{u \in \mathbb{R}^d \mid \langle u, y\rangle \leq \alpha\}$. Let $F_i(P)$ where $i \in [d-1]$ denote set of faces of dim i of a polytope P. A face of dimension zero is called a vertex and a face of dimension one is called an edge. We define the edge set of a polytope P by $E(P) := \{\text{conv}\,((v_i, v_j)) \mid (v_i, v_j) \subseteq \binom{\text{vert}(P)}{2}, \text{conv}\,((v_i, v_j)) \in F_1(P)\}$. We define the neighbors of a vertex v by $\text{ne}(v; P) := \{\hat{v} \in \text{vert}(P) \mid \ \text{conv}\,((v, \hat{v})) \in E(P)\}$. We will denote $\text{conv}\,((v,\hat{v})) \in E(P)$ by as $e_{v,\hat{v}}$ and $\text{ne}(v;P)$ by $\text{ne}(v)$ when clear from context.

B.2 Omitted Proofs from § 3

Proposition 11. *For a given induced loss L^G_φ, the unique report which minimizes the expected loss is $u^* := \arg\min_{u \in \mathbb{R}^d} \mathbb{E}_{Y\sim p}[L^G_\varphi(u, Y)] = \varphi(p)$ such that $u^* \in P$. Furthermore, every $\hat{u} \in P$ is a minimizer of $\mathbb{E}_{Y\sim\hat{p}}[L^G_\varphi(u,Y)]$ for some $\hat{p} \in \Delta_{\mathcal{Y}}$.*

Proof. By [Banerjee et al., 2005, Theorem 1], the minimizer of $\mathbb{E}_{Y\sim p}[L^G_\varphi(u,Y)]$ is $\sum_{y\in\mathcal{Y}} p_y v_y = \varphi(p)$. Thus, by the construction of the polytope embedding, it holds that $u^* = \varphi(p)$. Since Bregman divergences are defined with respect to strictly convex functions, u^* uniquely minimizes $\mathbb{E}_{Y\sim p}[L^G_\varphi(u,Y)]$.

Conversely, every $\hat{u} \in P$ is expressible as a convex combination of vertices; hence, by the definition of φ, for some distribution, say $\hat{p} \in \Delta_{\mathcal{Y}}$, it holds $\hat{u} = \varphi(\hat{p})$. Therefore, it holds that $\hat{u}$ minimizes $\mathbb{E}_{Y\sim\hat{p}}[L^G_\varphi(u,Y)]$. □

Theorem 12. *For any given pair $(L^G_\varphi, \psi^\varphi)$ and ℓ_{0-1} with embedding dimension $d < n-1$; it holds that $\mathcal{H} = \cup_{y\in\mathcal{Y}}\text{conv}\,(\text{vert}(P) \setminus \{v_y\}) \cap \psi^\varphi_y$ and furthermore $\mathcal{H} \neq \varnothing$.*

Proof. Choose a $y \in \mathcal{Y}$. We abuse notation and write $\text{vert}(P) \setminus v_y := \text{vert}(P) \setminus \{v_y\}$. Observe all $u \in \text{conv}\,(\text{vert}(P) \setminus v_y) \cap \psi^\varphi_y$ can be expressed as a convex combination of vertices without needing vertex v_y. The coefficients of said convex combination express a $p \in \Delta_{\mathcal{Y}}$ that is embedded to the point $u \in P$ where $p_y = 0$. Yet, by Proposition 2, said u is an expected minimizer of L^G_φ with respect to p. Given the intersection with ψ^φ_y and by Definition 8, it holds that $\cup_{y\in\mathcal{Y}}\text{conv}\,(\text{vert}(P)\setminus v_y) \cap \psi^\varphi_y \subseteq \mathcal{H}$.

We now shall show that $\mathcal{H} \subseteq \cup_{y\in\mathcal{Y}}\text{conv}\,(\text{vert}(P)\setminus v_y) \cap \psi^\varphi_y$. Fix $y \in \mathcal{Y}$. Assume there exists a point $u \notin \text{conv}\,(\text{vert}(P)\setminus v_y) \cap \psi^\varphi_y$ such that there exists some $p \in \Delta_{\mathcal{Y}}$ where $\varphi(p) = u$, $p_y = 0$, and $\psi^\varphi(u) = y$. Since $\psi^\varphi(u) = y$ and $u \notin \text{conv}\,(\text{vert}(P)\setminus v_y)) \cap \psi^\varphi_y$, it must be the case that $u \notin \text{conv}\,(\text{vert}(P)\setminus v_y)$. However, that implies that u is strictly in the vertex figure and thus must have weight on the coefficient for y. Thus, forming a contradiction that $p_y = 0$ which implies that $\mathcal{H} = \cup_{y\in\mathcal{Y}}\text{conv}\,(\text{vert}(P)\setminus v_y) \cap \psi^\varphi_y$.

To show non-emptiness of $\mathcal{H}$, we shall use Helly's Theorem (Rockafellar [1997], Corollary 21.3.2). W.l.o.g, assign an index such that $\mathcal{Y} = \{y_1, \ldots, y_d, y_{d+1}, \ldots, y_n\}$. Observe the elements of the set $\{\mathcal{Y} \setminus y_i\}_{i=1}^n$ each differ by one element. W.l.o.g, pick the first $d+1$ elements of the previous set. Observe $|\cap_{i=1}^{d+1} \mathcal{Y}\setminus y_i| = |\mathcal{Y} \setminus \{y_1, \ldots, y_d, y_{d+1}\}| = n - (d+1) > 0$. Hence, by Helly's theorem and uniqueness of y_i's, $\cap_{y\in\mathcal{Y}}\text{conv}\,(\text{vert}(P)\setminus v_y) \neq \varnothing$.

Pick a point $u' \in \cap_{y\in\mathcal{Y}}\text{conv}\,(\text{vert}(P)\setminus v_y)$. Since ψ^φ is well-defined, u' will be linked to some outcome $y' \in \mathcal{Y}$ and thus $u' \in \text{conv}\,(\text{vert}(P)\setminus v_{y'}) \cap \psi^\varphi_{y'} \subset \mathcal{H}$. Yet, u' can be expressed as a convex combination which does not use $v_{y'}$ since it lies in $\cap_{y \in \mathcal{Y}} \mathrm{conv}\,(\mathrm{vert}(P) \setminus v_y)$. Thus, by using Proposition 2 and by the definition of Hallucination (Def. 8), we have that $\mathcal{H} \neq \varnothing$. $\square$

Lemma 1 (Proposition 1.2.4)**.** *[Hiriart-Urruty and Lemaréchal, 2004] If φ is an affine transformation of $\mathbb{R}^n$ and $A \subset \mathbb{R}^n$ is convex, then then the image $\varphi(A)$ is also convex. In particular, if the set A is a convex polytope, the image is also a convex polytope.*

Lemma 2. *Let D_G be a Bregman divergence, φ be any polytope embedding, ψ be the MAP link, and L_φ^G be the loss induced by (D_G, φ). Assume the target loss is ℓ_{0-1}. If a point is in a strict calibrated region such that $u \in R_y$ for some $y \in \mathcal{Y}$, it is necessary that $u \in \mathrm{conv}\,(\{v_y\} \cup \mathrm{ne}(v_y)) \setminus \mathrm{conv}\,(\mathrm{ne}(v_y))$.*

Proof. If $u \in R_y$ and $u \in P \setminus \big(\mathrm{conv}\,(\{v_y\} \cup \mathrm{ne}(v_y)) \setminus \mathrm{conv}\,(\mathrm{ne}(v_y))\big)$, then u can be expressed as a convex combination which has no weight on the coefficient for v_y. Hence, there exists a distribution embedded into u where y would not be the mode, thus violating the initial claim that $u \in R_y$. $\square$

Lemma 3. *Let D_G be a Bregman divergence, φ be any polytope embedding, ψ be the MAP link, and L_φ^G be the loss induced by (D_G, φ). For any $u \in e_{(v_i, v_j)} \in E(P)$, it holds that $|\varphi^{-1}(u)| = 1$.*

Proof. Observe, the two vertices of an edge define the convex hull making up the edge and hence, by (Gruber [2007] ,Theorem 2.3) the two vertices are affinely independent. Therefore, all elements of the edge have a unique convex combination which are expressed by the convex combinations of the edge's vertices. Given the relation of the embedding φ and convex combinations of vertices expressing distributions, it holds that $|\varphi^{-1}(u)| = 1$. $\square$

Lemma 4. *Let D_G be a bregman divergence, φ be a polytope embedding, and L_φ^G be the induced loss by (D_G, φ). For all $y \in \mathcal{Y}$, it holds that $\dim(\varphi(\mathit{mode}_y)) = \dim(P) \geq 2$.*

Proof. By the construction of φ, we know that $\dim(P) \geq 2$. Fix $y \in \mathcal{Y}$. By Lemma 3, we know that any edge connected from v_y and $\hat{v} \in \mathrm{ne}(v_y)$, the distributions embedded into the half of the line segment closer to v_y, y is in the mode. By Lemma 1, we know that $\varphi(\gamma_y^{\text{mode}})$ is a convex set. Thus, the convex hull of the half line segments is part of $\varphi(\gamma_y^{\text{mode}})$. Since each vertex has at least $\dim(P)$ neighbors, it holds that $\dim(\varphi(\gamma_y^{\text{mode}})) = \dim(P)$. $\square$

Theorem 13. *Let D_G be a Bregman divergence, φ be any polytope embedding, ψ^φ be the MAP link, and L_φ^G be the loss induced by (D_G, φ). There exists a $\mathcal{P} \subseteq \Delta_\mathcal{Y}$ with non-zero Lebesgue measure and $\varphi(\mathcal{P}) \subseteq R_\mathcal{Y}$ via $(L_\varphi^G, \psi^\varphi)$ with respect to ℓ_{0-1}.*

Proof. Recall that $\gamma^{\text{mode}}(p) := \text{prop}[\ell_{0-1}](p) = \text{mode}(p)$. Fix $y \in \mathcal{Y}$. For contradiction, assume for any $\hat{y} \in \mathcal{Y}$ where $y \neq \hat{y}$, it holds that $B_\epsilon(v_y) \cap \varphi(\gamma_{\hat{y}}^{\text{mode}}) \neq \varnothing$ for all $\epsilon > 0$. By Lemma 3, it holds that $\mathrm{conv}\,(\{v_y\} \cup m_{v_y,\alpha}) \subseteq \varphi(\gamma_y^{\text{mode}})$ where $m_{v_y,\alpha} := \{(1-\alpha)v_y + \alpha\overline{v} \mid \overline{v} \in \mathrm{ne}(v_y)\}$ defined by any $\alpha \in (0, .5)$. Furthermore, the elements of $\cup_{m \in m_{v_y,\alpha}} \mathrm{conv}\,(\{v_y\} \cup \{m\})$ have one distribution embedded onto it where y is the only valid mode thus, we know that $\varphi(\text{mode}_{\hat{y}}) \cap \cup_{m \in m_{v_y,\alpha}} \mathrm{conv}\,(\{v_y\} \cup \{m\}) = \varnothing$. Since $\varphi(\gamma_{\hat{y}}^{\text{mode}}) \subset P$ is closed and convex, there must exist some non-negative min distance between $\varphi(\gamma_{\hat{y}}^{\text{mode}})$ and v_y which we shall denote by d_v. For any $\epsilon \in (0, d_{v_y})$, we can define $B_\epsilon(v_y)$ such that $B_\epsilon(v_y) \cap \varphi(\gamma_{\hat{y}}^{\text{mode}}) = \varnothing$, forming a contradiction.

For each $v_y \in \text{vert}(P)$ define a d_{v_y} and let $\epsilon' \in \cap_{v_y \in \text{vert}(P)}(0, d_{v_y})$. By the construction of P and the definition of ψ^φ, there exists a $\epsilon'' > 0$ such that for all $u \in B_{\epsilon''}(v_y)$ it holds that $\psi(u) = y$ and $B_{\epsilon''}(v_y) \subset \psi_y^\varphi$. For any $y \in \mathcal{Y}$, we know that $B_{\min\{\epsilon', \epsilon''\}}(v_y)) \cap P \subseteq R_y$ by the construction of our epsilon ball. We claim $\varphi^{-1}(B_{\min\{\epsilon', \epsilon''\}}(v_y) \cap P)$ is a set of distributions for which calibration holds.

For $p \in \Delta_\mathcal{Y}$ such that $\varphi(p) \in B_{\min\{\epsilon', \epsilon''\}}(v_y) \cap P$ for some $v_y \in \text{vert}(P)$, suppose a sequence $\{u_m\}$ converges to $\text{prop}[L_\varphi^G](p) = \varphi(p)$ (equality by Proposition 2). By construction of $B_{\min\{\epsilon', \epsilon''\}}(v_y) \cap P$, $\psi^\varphi(\varphi(p)) = y \in \text{mode}(p)$ and hence, a minimizing report for $\ell_{0-1}(y; p)$. Furthermore, since $B_{\min\{\epsilon', \epsilon''\}}(v_y) \subset \psi_{\varphi^{-1}(v_y)}^\varphi$, all elements within $B_{\min\{\epsilon', \epsilon''\}}(v_y)$ link to y. Since $\{u_m\}$ converges to $\text{prop}[L_\varphi^G](p)$, there exists some $N \in \mathbb{N}$ and $n \geq N$, such that $\|u_n - \varphi(p)\|_2 < \min\{\epsilon', \epsilon''\}$, meaning that $\mathbb{E}_{\mathcal{Y}\sim p}[\ell_{0-1}(\psi^{\varphi}(u_m), Y)] \to \min_{y\in\mathcal{Y}} \mathbb{E}_{\mathcal{Y}\sim p}[\ell_{0-1}(y, Y)]$. Hence, for any $v_y \in \text{vert}(P)$, $(\text{prop}[L^G_\varphi], \psi^\varphi)$ is ℓ_{0-1}-calibrated property with respect to $\varphi^{-1}(B_{\min\{\epsilon',\epsilon''\}}(v_y) \cap P)$. Furthermore, by the construction of $B_{\min \epsilon',\epsilon''}(v_y)$ for each $v_y \in \text{vert}(P)$, we have that L^G_φ is strictly for $\text{prop}[L^G_\varphi]$. Thus, by Theorem 1, $(L^G_\varphi, \psi^\varphi)$ is ℓ_{0-1}-calibrated for at least the distributions $\mathcal{P} = \cup_{v_y\in\text{vert}(P)}\varphi^{-1}(B_{\min\{\epsilon',\epsilon''\}}(v_y) \cap P)$ as well as $\varphi(\mathcal{P}) \subseteq R_{\mathcal{Y}}$. Furthermore, since $B_{\min\{\epsilon',\epsilon''\}}$ for each $v_y \in \text{vert}(P)$ is non-empty, we have that $\mathcal{P} \neq \varnothing$. $\square$

B.3 Omitted Proofs from § 4

Corollary 14. *Let φ be an embedding from 2^d outcomes into the vertices of $P^\square$ in d-dimensions and define an induced loss L^G_φ. Fix $\alpha \in [0, .5)$ and define Θ_α. $(L^G_\varphi, \psi^{P^\square}_\alpha)$ is ℓ_{0-1}-calibrated for Θ_α.*

Proof. W.l.o.g, say the outcome $y_1 \in \mathcal{Y}$ is embedded into $\mathbb{1}_{[d]} \in \text{vert}(P^\square)$. Say $\alpha = .5$. Observe that

$$\Psi^{y_1}_\alpha = \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1-\alpha \\ \alpha \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1-\alpha \\ 0 \\ \alpha \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 1-\alpha \\ 0 \\ \vdots \\ \alpha \\ 0 \end{pmatrix}, \begin{pmatrix} 1-\alpha \\ 0 \\ \vdots \\ 0 \\ \alpha \end{pmatrix} \right\}$$

and that $1 \geq (1-\alpha) \pm \alpha \geq 0$ for any $\alpha \in (0, .5)$. Hence, for any $\alpha \in (0, .5)$ it holds that $P^{y_1}_{0.5} = \text{conv}\left(\{0,1\}^d\right)$ and furthermore $P^{y_1}_\alpha \subset P^{y_1}_{0.5} \subset \mathbb{R}^d_{>0}$. By symmetry of $P^\square$ and the linearity of φ, for any $\alpha \in (0, .5)$ and $y \in \mathcal{Y}$, we have that P^y_α is a strict subset of the orthant that contains v_y. Hence, for all $y, \hat{y} \in \mathcal{Y}$ such that $y \neq \hat{y}$, it holds that $P^y_\alpha \cap P^{\hat{y}}_\alpha = \varnothing$. Thus by Theorem 6, $(L^G_\varphi, \psi^{P^\square}_\alpha)$ is ℓ_{0-1}-calibrated for Θ_α where $\alpha \in (0, .5)$. $\square$

Corollary 15. *Let φ be an embedding from $d!$ outcomes into the vertices of P^w in d dimensions such that $w = (0, \frac{1}{\beta d}, \frac{2}{\beta d}, \ldots, \frac{d-1}{\beta d}) \in \mathbb{R}^d$ where $\beta = \frac{d-1}{2}$. Fix $\alpha \in (0, \frac{1}{d})$. Then $(L^G_\varphi, \psi^{P^w}_\alpha)$ is ℓ_{0-1}-calibrated over Θ_α.*

Proof. Let $\Delta_d := \text{conv}\left(\{\mathbb{1}_i \in \mathbb{R}^d \mid i \in [d]\}\right)$ and observe $P^w \subset \Delta_d$ since for all π, $\|\pi \cdot w\|_1 = \|w\|_1 = 1$. Observe that P^w can be symmetrically partitioned into $d!$ regions with disjoint interiors, one for each permutation $\pi \in \mathcal{S}_d$ via $\Delta^\pi_d := \{u \in \Delta_d \mid u_1 \leq \cdots \leq u_d\}$. Fix $\pi \in \mathcal{S}_d$ and w.l.o.g assume π is associated with the constraints $\Delta^\pi_w := \{u \in \Delta_w \mid u_1 \leq \cdots \leq u_d\}$ implying that $\pi(w) = (\frac{0}{\beta d}, \frac{1}{\beta d}, \ldots, \frac{d-1}{\beta d})$. Let $\alpha = \frac{1}{d}$ and define Θ_α. With respect to Θ_α, let $y := \varphi^{-1}(\pi(w)) \in \mathcal{Y}$ and $\hat{y} := \varphi^{-1}(\hat{\pi}(w)) \in \mathcal{Y}$ such that $\hat{\pi} \in \mathcal{S}_d$. Thus the set $\Psi^y_\alpha := \{(1-\frac{1}{d})\delta_y + (\frac{1}{d})\delta_{\hat{y}} \mid y, \hat{y} \in \mathcal{Y}\}$ is mapped via φ to the following points

$$\varphi(\Psi^y_\alpha) = \{(1-\frac{1}{d})(\pi(w)) + (\frac{1}{d})(\hat{\pi}(w)) \mid \hat{\pi} \in \mathcal{S}_d\}$$

within the permutahedron.

We shall show that $P^y_\alpha \subseteq \Delta^\pi_d$. If this were not true, there would exists an element of $w^{\pi,\hat{\pi}} \in \varphi(\Psi^y_\alpha)$ such such that for some pair of adjacent indices, say $i, i+1 \in [d-1]$, $w^{\pi,\hat{\pi}}_i > w^{\pi,\hat{\pi}}_{i+1}$. For sake of contradiction, fix $i \in [d-1]$ and assume there exists a $\hat{\pi} \in \mathcal{S}_d$ such that $w^{\pi,\hat{\pi}}_i > w^{\pi,\hat{\pi}}_{i+1}$. Observe that any element of $\hat{\pi}(w)$ can be expressed by $\frac{j}{\beta d}$ using some $j \in \{0, 1, \ldots, d-1\}$. Thus,

$$\begin{aligned}
& w_i^{\pi,\hat{\pi}} > w_{i+1}^{\pi,\hat{\pi}} \\
& \Leftrightarrow (1-\frac{1}{d})(\frac{i-1}{\beta d}) + (\frac{1}{d})(\hat{\pi}(w))_j > (1-\frac{1}{d})(\frac{i}{\beta d}) + (\frac{1}{d})(\hat{\pi}(w))_{\hat{j}} \\
& \Rightarrow \quad (1-\frac{1}{d})(\frac{i-1}{\beta d}) + (\frac{1}{d})(\frac{j}{\beta d}) > (1-\frac{1}{d})(\frac{i}{\beta d}) + (\frac{1}{d})(\frac{\hat{j}}{\beta d}) \qquad \text{Multiply by } \beta d \\
& \Rightarrow \quad (i-1)(1-\frac{1}{d}) + j(\frac{1}{d}) > i(1-\frac{1}{d}) + \hat{j}(\frac{1}{d}) \\
& \Rightarrow \quad 1-d > \hat{j} - j
\end{aligned}$$

for some $j, \hat{j} \in \{0, 1, \ldots, d-1\}$ where $j \neq \hat{j}$.

Case 1: $(j < \hat{j})$: The smallest value possible for $\hat{j} - j$ is $0 - (d-1)$ however, $1 - d \not> 1 - d$.

Case 2:$(j > \hat{j})$: The smallest value possible for $\hat{j} - j$ is 1 however, $1 - d \not> 1$.

Hence, $P_\alpha^y \subseteq \Delta_d^\pi$ and specifically, there can exists an extreme point of P_α^y that lies on the boundary of Δ_d^π as shown in **Case 1**. However, if $\alpha \in (0, \frac{1}{d})$, every extreme point of P_α^y moves closer to $\pi(w)$ (besides the extreme point itself already on $\pi(w)$) and therefore P_α^y lies strictly within Δ_d^π. By symmetry of P^w and the linearity of φ, this would imply that for all $y', y'' \in \mathcal{Y}$ such that $y' \neq y''$ it holds that $P_\alpha^{y'} \cap P_\alpha^{y''} = \varnothing$. Thus by Corollary 6, $(L_\varphi^G, \psi_\alpha^{P^w})$ is ℓ_{0-1}-calibrated for Θ_α where $\alpha \in (0, \frac{1}{d})$. □

B.4 Omitted Proofs from § 5

Lemma 16. *Say we are given a cross-polytope embedding $\varphi : \Delta_{2d} \to P^\oplus$ and induced loss L_φ^G. Let (v_{a_i}, v_{b_i}), be the i^{th} diagonal pair (i.e. $\varphi(\delta_{a_i}) = v_{a_i}$). Define the property $\Gamma^\varphi : \Delta_{2d} \to \mathcal{B}$ element-wise by*

$$\Gamma^\varphi(p)_i := \begin{cases} (<, a_i, b_i) & \text{if } p_{a_i} < p_{b_i} \\ (>, a_i, b_i) & \text{if } p_{a_i} > p_{b_i} \\ (=, a_i, b_i) & \text{if } p_{a_i} = p_{b_i}. \end{cases}$$

Furthermore define the link $\psi^{P^\oplus} : \mathbb{R}^d \to \mathcal{B}$ with respect to each diagonal pair as

$$\psi(u; v_{a_i}, v_{b_i})_i^{P^\oplus} := \begin{cases} (<, a_i, b_i) & \text{if } ||u - v_{a_i}||_2 > ||u - v_{b_i}||_2 \\ (>, a_i, b_i) & \text{if } ||u - v_{a_i}||_2 < ||u - v_{b_i}||_2 \\ (=, a_i, b_i) & \text{o.w.} \end{cases}$$

Then $(L_\varphi^G, \psi^{P^\oplus})$ elicits Γ^φ.

Proof. W.l.o.g, fix a diagonal pair (v_a, v_b) and let $v_a := \mathbb{1}_1$ and $v_b := -\mathbb{1}_1$. Define the embedding φ accordingly. We will show that the following is true for all distributions mapped via φ to $u \in P^\oplus$.

$$\begin{aligned}
&||u - v_a||_2 > ||u - v_b||_2 \iff p_a < p_b \\
\text{OR } &||u - v_a||_2 < ||u - v_b||_2 \iff p_a > p_b \\
\text{OR } &||u - v_a||_2 = ||u - v_b||_2 \iff p_a = p_b.
\end{aligned}$$

First, fix $p \in \Delta_{2d}$. Recall, by Proposition 2, the minimizing report for L_φ^G in expectation is $u = \varphi(p) \in P \subset \mathbb{R}^d$. We will prove the forward direction of the first and second lines. Then the reverse directions follow from the contrapositives.

Case 1, $\implies$: Assume for contradiction that $p_a < p_b$ and $||\varphi(p) - v_a||_2 < ||\varphi(p) - v_b||_2$. Then

$$\begin{aligned}\langle \varphi(p) - \mathbb{1}_1, \varphi(p) - \mathbb{1}_1 \rangle <& \langle \varphi(p) + \mathbb{1}_1, \varphi(p) + \mathbb{1}_1 \rangle \\ (u_1 - 1)^2 + \sum_{i=1} u_i^2 <& (u_1 + 1)^2 + \sum_{i=1} u_i^2 \\ -u_1 <& u_1 \ .\end{aligned}$$

By the definition of a d-cross polytope $P^{\oplus} := \mathrm{conv}\left(\{\pi((\pm 1, 0, \ldots, 0)) \mid \pi \in \mathcal{S}_d\}\right)$ and the orthogonal relation between vertices, to express a $u \in P^{\oplus}$ as a convex combination of vertices, each diagonal pair of vertices coefficients solely influence the position along a single unit basis vector. Hence, due to the definition of φ, we have $u_1 = \mathbb{1}_1 \cdot p_a - \mathbb{1}_1 \cdot p_b < 0$ since we have assumed that $p_a < p_b$. Hence $-u_1 < u_1 < 0$, a contradiction.

Case 2, $\implies$: Assume $p_a > p_b$ and $||\varphi(p) - v_a||_2 < ||\varphi(p) - v_b||_2$. By symmetry with case 1, all the inequalities are reversed, leading to the contradiction that $-u_1 > u_1 > 0$.

Case 3: ($p_a = p_b$): Follows from the if and only ifs of cases 1 and 2.

Hence $(L^G_{\varphi}, \psi_{\varphi})$ elicits Γ^{φ}.

□

Theorem 10. *Let $d \geq 2$. The mode is $(2d, d, m)$-Polytope Elicitable for some $m \in [2d-1, d(2d-1)]$.*

Proof. We will elicit the mode via the intermediate properties, Γ^{φ_j}, defined in Lemma 9. First we construct a set of embeddings so that we guarantee that all the φ_j's allow comparison between any pair of outcome probabilities. For example, for each unique pair $(a, b)_j \in \binom{\mathcal{Y}}{2}$ define an embedding: $\varphi_j(\delta_a) = \mathbb{1}_1$ and $\varphi_j(\delta_b) = -\mathbb{1}_1$, and embed every other remaining report $r \in \mathcal{Y} \setminus \{a, b\}$ arbitrarily. Since $(L^G_{\varphi}, \psi^{P^{\oplus}})$ elicits Γ^{φ}, minimizing each $L^G_{\varphi_j}$ with a separate model yields us comparisons via the link $\psi^{P^{\oplus}}$. To find the set $r \in \mathcal{Y}$ such that p_r is maximum, we use a sorting algorithm that uses pairwise comparisons, such as bubble sort. Hence with Υ as Algorithm 1, we have that $\Upsilon(\{L^G_{\varphi_j}, \psi^{P^{\oplus}}\}) = \mathrm{mode}(p)$.

Assuming there exist φ_js such that there is no redundancy in comparison pairs between each Γ^{φ_j}, we would need only $\frac{d(2d-1)}{d} = 2d - 1$ problem instances. Hence, we establish our lower bound on the needed number of problem instances. □

C Hamming Loss Hallucination Example

Hamming loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is defined by $\ell(y, \hat{y}) = \sum_{i=1}^{d} \mathbb{1}_{y_i \neq \hat{y}_i}$ where $\mathcal{Y} = \{-1, 1, \}^d$. Suppose $d = 3$ and we have the following indexing over outcomes

$$\begin{aligned}\mathcal{Y} := \{&y_1 \equiv (1,1,1), y_2 \equiv (1,1,-1), y_3 \equiv (1,-1,1), y_4 \equiv (-1,1,1), \\ &y_5 \equiv (-1,-1,1), y_6 \equiv (1,-1,-1), y_7 \equiv (-1,1,-1), y_8 \equiv (-1,-1,-1)\} \ .\end{aligned}$$

Let us define the following distribution

$$p_\epsilon = (0, \frac{1}{3} - \epsilon, \frac{1}{3} - \epsilon, \frac{1}{3} - \epsilon, 0, 0, 0, 3\epsilon) \in \Delta_{\mathcal{Y}}$$

such that $\epsilon > 0$.

- $\mathbb{E}_{Y \sim p_\epsilon}[\ell(y_1, Y)] = 1 + 6\epsilon$
- $\mathbb{E}_{Y \sim p_\epsilon}[\ell(y_2, Y)] = \mathbb{E}_{Y \sim p_\epsilon}[\ell(y_3, Y)] = \mathbb{E}_{Y \sim p_\epsilon}[\ell(y_4, Y)] = \frac{4}{3} + 2\epsilon$
- $\mathbb{E}_{Y \sim p_\epsilon}[\ell(y_5, Y)] = \mathbb{E}_{Y \sim p_\epsilon}[\ell(y_6, Y)] = \mathbb{E}_{Y \sim p_\epsilon}[\ell(y_7, Y)] = \frac{7}{3} - 4\epsilon$
- $\mathbb{E}_{Y \sim p_\epsilon}[\ell(y_8, Y)] = 2 - 6\epsilon$

For all $\epsilon \in [0, \frac{1}{12})$, the minimizing report in expectation is $y_1 = (1, 1, 1)$. However, $p_{\epsilon,1} = 0$ and thus, a hallucination would occur under a calibrated surrogate and link pair.

D Linking under Multiple Problem Instances

As stated in § 5, when using real data, given that these are asymptotic results, we may have conflicting logic for the provided individual reports. In this section, we provide an approach such that the algorithm still reports information in the aforementioned scenario and will reduce to Algorithm 1 asymptotically. We build a binary relation table $M \in \{0,1\}^{n \times n}$ with the provided reports. Based on M, we select a largest subset of $S \subseteq \mathcal{Y}$ such that when M is restricted to rows and columns corresponding to the elements of S, denoted by M_S, we have that M_S is reflexive, antisymmetric, transitive, and strongly connected implying M_S has a total-order relation defined over its elements. Having a total-order relation infers the mode can be found via comparisons. The algorithm returns (R, S), where R is the mode set with respect to the elements of S.

Algorithm 2 Elicit mode via comparisons and the d-Cross Polytopes over well-defined partial orderings

Require: $M = \{(L^G_{\varphi_j}, \psi_j^{P^\oplus})\}_{j=1}^m$

Learn a model $h_j : \mathcal{X} \to \mathbb{R}^d$ for each instance $(L^G_{\varphi_j}, \psi_j^{P^\oplus}) \in M$

For some fixed $x \in \mathcal{X}$, collect all $B_j \leftarrow \psi_j^{P^\oplus}(h_j(x))$ where $B_j \in \mathcal{B}_j$

Build $M \in \{0,1\}^{n \times n}$ binary relation table with provided $\{B_j\}_{j=1}^m$ as such

- Label rows top to bottom by $y_1, \ldots, y_n$ and columns left to right by $y_1, \ldots, y_n$.
- For all $(\cdot, p_{y_i}, p_{y_k}) \in B_j$, if $p_{y_i} \leq p_{y_k}$ set $M[i,k] = 1$ and 0 otherwise.

Select largest subset $S \subseteq \mathcal{Y}$ such that M_S is reflexive, antisymmetric, transitive, and strongly connected.

Report $(R, S) \leftarrow$ FindMaxElements-of-$S(M; S)$

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Any claimed result in the abstract is proved within this work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, our paper discuss how these results are asymptotic.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Yes, we thoroughly introduce every necessary definition and past result necessary to understand the assumptions that hold under our results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The results of this work have rigorous proofs presented next to the results or referenced clearly in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification:This paper does not include experiments requiring code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: None of our conducted work for this paper violates the code of ethics presented.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the body of our paper, we provide a broader impace section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The answer NA means that the paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.