Approximate Γ-calibration of Discrete Classification Tasks

Jessica Finocchiaro
Computer Science, Boston College
finocch@bc.edu

Victor Ganson
Computer Science, Boston College
gansonv@bc.edu

Drona Khurana
Computer Science, University of Colorado Boulder
drona.khurana@colorado.edu

Abstract

One prominent method of evaluating machine learning model trustworthiness is the notion of *calibration*. In the binary outcome setting, a probabilistic predictor is calibrated if outcomes are realized according to a model's distributional prediction, conditioned on this prediction. Straightforward extensions of binary calibration definitions to probabilistic multiclass classifiers suffer from an exponential complexity blowup as the space of predictions grows exponentially in the number of classes n. As a remedy, Noarov and Roth [35] propose multiclass calibration with predictions that are *properties* of the outcome distribution, reducing complexity from growing in the number of classes n to the *dimension* d of the property, called its elicitation complexity. Previous work on approximate property calibration is generally limited to continuous scalar properties, despite many relevant properties of interest being discrete, like the mode or rankings. We characterize the approximate property calibration of discrete properties which are strongly orderable by using Lipschitz continuous scalar properties as an intermediary. This work is the first to our knowledge to provide approximate calibration results for discrete properties. Along the way, we construct algorithms for designing these Lipschitz properties, which we prove can be post-processed to obtain the original discrete property.

1 Introduction

One prominent paradigm of evaluating machine learning models examines *calibration*, where a model's probabilistic predictions are roughly realized. That is, instead of conditioning on a feature instance $X = x$ (as desired in evaluating Bayes predictors), we condition on a model's prediction $f(x) = p$, and want to observe outcomes roughly aligned with p. Intuitively, this is typically formalized as $\mathbb{E}[Y|f(X) = p] \approx p$ for all $p \in \mathbf{im}(f)$. Empirically, continuous predictors must have their predictions *binned* to obtain multiple instances of the same conditioning event. The sample complexity of calibration is known to grow polynomially in the number of bins on the space [10, 24, 47].

This poses a prohibitive challenge for *multiclass* calibration: when bins represent approximate probabilistic predictions, the number of bins naïvely grows exponentially in the number of classes, posing substantial computational and statistical challenges. To circumvent this, many works on multiclass calibration reduce the granularity of the binned space [4, 23, 24, 28, 30, 47]. These approaches generally combine full distributional predictions with some polynomial binning approach, and often lack decision-theoretic guarantees for the task implicitly guiding the polynomial binning. Derr et al. [11] outline a few possible approaches to defining multiclass calibration, consolidating

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

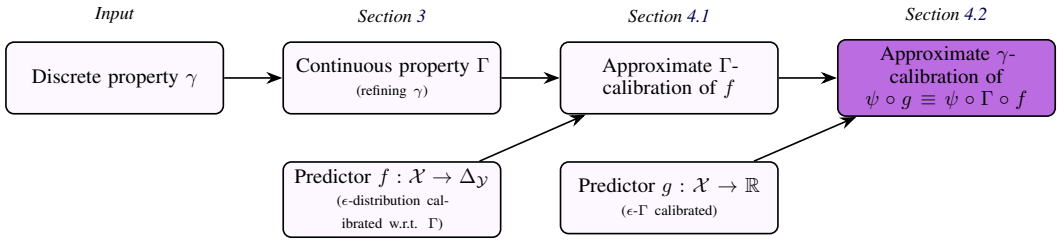


Figure 1: Flowchart of results. While we start with a discrete property γ, we use continuous properties Γ as intermediaries for approximate calibration. We discuss the construction of these properties in Section 3, implications for approximate Γ-calibration in Section 4.1, and further implications for approximate γ-calibration in Section 4.2, where post-processing is formalized through ψ.

many of these approaches to what they call *distribution calibration with respect to* γ, where $\gamma : \Delta_{\mathcal{Y}} \to \mathcal{R}$ maps distributions to reports is called a *property* guiding the binning approach. In contrast, they show distribution calibration with respect to γ is a stronger notion than an alternative called Γ-calibration— initially introduced by Gneiting and Resin [22], Noarov and Roth [35], in which a model estimates the property γ itself— rather than a distribution— while still providing decision-theoretic guarantees.

Critically, results surrounding approximate γ-calibration assume the prediction target γ is *continuous*. These *continuous* (calibrated) predictions are often used to make *discrete* decisions. This continuous-to-discrete jump often poses challenges for obtaining decision-theoretic guarantees, as demonstrated by Noarov and Roth [35]'s assumption of continuous properties and Derr et al. [11] only giving exact calibration results for discrete properties. From a technical perspective, it is difficult to obtain *approximate* calibration guarantees for discrete prediction tasks since discrete predictions do not provide the opportunity to communicate a model's uncertainty, which is critical to understand if a model is "close to" calibration. More precisely, this tension emerges from the need for a calibration error metric that is continuous in the prediction space, which is impossible if predictions themselves are not continuous.

To bridge this gap, we propose using *Lipschitz property elicitation* to obtain a continuous predictor of an intermediate continuous property Γ which can be post-processed to obtain the desired discrete decision property γ. We can obtain approximate calibration guarantees with this smoothed property, while bounding the probability of miscalibration in the discrete space. The dimension of this smoothed property is an important object of study in its own right. If the prediction dimension d is significantly lower than predicting a full distribution over n labels, this improves the computational complexity of gradient-based optimization algorithms. Moreover, the number of bins evaluated in empirical estimation of miscalibration grows exponentially in the prediction dimension d [10]. Since sample complexity grows in the number of bins, a prediction dimension $d < n$ can significantly improve computational and statistical complexity.

Figure 1 outlines the organization of the paper. In Section 3, we first characterize the discrete properties that can be obtained by post-processing a continuous property $\Gamma : \Delta_{\mathcal{Y}} \to \mathbb{R}$ in 1 dimension. Our characterization is constructive, providing two algorithms yielding Lipschitz properties. With this characterization, we proceed to obtain bounds on the approximate Γ-calibration of probabilistic predictors in Section 4.1, then bound the approximate discrete γ-calibration of either probabilistic or Γ-predictors in Section 4.2 by using the smoothed property Γ as an intermediary.

2 Background and related work

Consider a supervised learning task in a multiclass setting where features take values in $\mathcal{X}$, labels take values in a finite set $\mathcal{Y}$ with $3 \leq |\mathcal{Y}| := n < \infty$, and data is sampled i.i.d. from a measurable distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$. We often contrast two predictors: a *distributional* predictor $f : \mathcal{X} \to \Delta_{\mathcal{Y}}$ yielding distributions over labels in $\mathcal{Y}$, and a *surrogate* predictor $g : \mathcal{X} \to \mathbb{R}^d$ for some surrogate dimension d. In particular, we often focus on scalar predictors $g : \mathcal{X} \to \mathbb{R}$. Let $D_{Y|\{\hat{x}:f(\hat{x})=u\}} \equiv D_{Y|f,u}$ denote the marginal distribution on labels conditioned on observing the prediction u, where we use the latter for shorthand. For a distribution $p \in \Delta_{\mathcal{Y}}$, we denote $p_y := \Pr[Y = y]$.

We consider loss functions measuring the error of *reports* in $\mathcal{R}$ against *ground truth labels* in $\mathcal{Y}$, denoting $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ as a surrogate loss like hinge or cross-entropy, and $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ with $|\mathcal{R}|$ finite as a discrete loss like 0-1 loss or a Borda score. We often represent a discrete loss ℓ as a cost matrix, where $\ell(r, y) = \ell_{r,y}$. Similarly, we will discuss surrogate properties $\Gamma : \Delta_\mathcal{Y} \to \mathbb{R}^d$, and discrete properties $\gamma : \Delta_\mathcal{Y} \to 2^\mathcal{R} \setminus \{\emptyset\}$ for a finite set $\mathcal{R}$, which may be set-valued, but never empty. We let $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ be shorthand for $\gamma : \Delta_\mathcal{Y} \to 2^\mathcal{R} \setminus \{\emptyset\}$.

2.1 Properties, identifiability, embeddings

The elicitation of discrete properties was first characterized by Lambert and Shoham [31], Lambert et al. [33]. In essence, a property $\Gamma : \Delta_\mathcal{Y} \to \mathbb{R}^d$ or $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ maps distributions over labels to "reports" in some set. The primary question driving the property elicitation research is which properties can be *elicited* by specific loss functions, as elicitation tends to be a simpler tool for analysis than *consistency* [7, 46], which is necessary for establishing PAC learning bounds through empirical risk minimization.[†] Lambert establishes a tight connection in the geometry of discrete properties and their elicitability through power diagrams [2]; however, their characterization yields discrete losses, which are usually not tractable to directly optimize. To circumvent this, the field has progressed in two diverging directions: restricting to the study of continuous properties [17, 19, 35, 38, 44], or through the design of *surrogate* loss functions [6, 15, 29, 40, 41, 45] whose surrogate predictions in $\mathbb{R}^d$ can later be *linked* back to discrete decisions by a *link* function $\psi : \mathbb{R}^d \to \mathcal{R}$. As one prominent example, the categorical cross-entropy loss elicits the identity property $\Gamma(p) = p$, which is continuous. This continuous property is then often "linked" to the discrete property of the mode $\psi(u) = \arg\max_y u_y$.

Definition 1 ((Direct, indirect) elicitation, refinement)**.** *A loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$* (directly) elicits *a property $\Gamma : \Delta_\mathcal{Y} \to \mathbb{R}^d$ if, for all $p \in \Delta_\mathcal{Y}$, we have*

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

Moreover, we say L indirectly elicits *a property $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ if it elicits Γ and there exists a link $\psi : \mathbb{R}^d \to \mathcal{R}$ such that $r \in \Gamma(p) \implies \psi(r) \in \gamma(p)$ for all $p \in \Delta_\mathcal{Y}$. If such a link exists, we say that Γ* refines *γ.*

Often, we refer to the *level set* or set-valued inverse of a property $\Gamma^{-1}(u) = \{p \in \Delta_\mathcal{Y} \mid u \in \Gamma(p)\}$. Note that for discrete properties $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ elicited by discrete losses $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$, the definition of elicitation follows in the same manner, replacing $\mathbb{R}^d$ with $\mathcal{R}$. However, in this case, the property might be set-valued on the boundary between two level sets r_i and r_{i+1}, which is set of measure 0. In this case, we sometimes study the object $\gamma^{-1}(r_i) \cap \gamma^{-1}(r_{i+1})$ as the boundary.

Fissler and Ziegel [17], Frongillo and Kash [18] and later Frongillo and Kash [20] formalize the notion of *elicitation complexity* for classes of loss functions or properties, and Nueve et al. [36] examines tradeoffs between elicitation complexity and distributional assumptions on guarantees. This notion of complexity is parameterized by the dimension d of inputs to the loss function. A lower dimension d can improve computational costs for gradient-based optimization methods.

Definition 2 (Elicitation complexity [20])**.** *For $d \in \mathbb{N} \cup \{\infty\}$, let $\mathcal{E}_d$ denote the class of all elicitable properties $\Gamma : \Delta_\mathcal{Y} \to \mathbb{R}^d$. Let $\mathcal{C}$ be a class of properties and $d \in \mathbb{N} \cup \{\infty\}$. A property $\Gamma : \Delta_\mathcal{Y} \to \mathcal{R}$ is d-elicitable with respect to $\mathcal{C}$ if there exists an intermediate property $\hat{\Gamma} \in \mathcal{C} \cap \mathcal{E}_d$ and a map ψ such that $\Gamma = \psi \circ \hat{\Gamma}$. The elicitation complexity of Γ is $\mathbf{elic}_\mathcal{C}(\Gamma) = \min\{d \mid \Gamma$ is d-elicitable with respect to $\mathcal{C}\}$.*

For example, while categorical cross-entropy takes $d = n$-dimensional inputs, the highest sensible dimension possible, the identity property it elicits $\Gamma(p) = p \in \mathbb{R}^n$ can be post-processed into *any* discrete property $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$, even those which are not elicitable themselves, providing a trivial upper bound on elicitation complexity of n for all properties γ for all classes $\mathcal{C}$ containing the identity.

We are particularly interested in the *Lipschitz continuous* elicitation complexity of discrete properties γ, where a property $\Gamma : \Delta_\mathcal{Y} \to \mathbb{R}^d$ is *Lipschitz*. Let $\mathcal{C}_{K-\mathrm{Lip}}$ denote the class of K-Lipschitz properties, and $\mathcal{C}_\mathrm{Lip}$ denote the set of properties which are K-Lipschitz for some $K \geq 0$. Without loss of generality, we assume K-Lipschitz property values are bounded in the relevant $[0, 1]^d$ hypercube.

[†] See works like [1, 5–7, 14, 15, 29, 40, 46] for a more nuanced discussion on the relationship between property elicitation, consistency, and classification calibration.

Characterizations of these notions of complexity are most complete in 1 dimension, and are established in the context of *orderable* properties Finocchiaro et al. [13], Khurana et al. [29]. We pose a notion of *strong orderability* necessary for obtaining Lipschitz properties. With these properties, we construct differentiable surrogates $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ indirectly eliciting strongly orderable properties; the intermediate property directly elicited by L is in $\mathcal{C}_{\text{Lip}}$, which we use in Section 4 for approximate calibration guarantees.

Definition 3 ((Strongly) Orderable property [13])**.** *A finite property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ is* orderable *if there is an enumeration of $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$ such that for all $i \leq |\mathcal{R}| - 1$, we have $\gamma^{-1}(r_i) \cap \gamma^{-1}(r_{i+1})$ is a hyperplane intersected with $\Delta_{\mathcal{Y}}$. Moreover, we say γ is* strongly orderable *if it is orderable and, for all $i \in \{2, \ldots, |\mathcal{R}| - 1\}$, we have $\inf_{p \in \gamma^{-1}(r_{i-1}), q \in \gamma^{-1}(r_{i+1})} \|p - q\|$ is bounded away from 0.*

For intuition, a property is strongly orderable if the level sets denoted by black boundaries in Figure 2c are bounded away from each other.

2.2 Calibration in multiclass settings

While calibration of predictors has been well-studied in binary outcome settings, definitions in multiclass settings are not universally agreed upon, and results are generally definition-specific. Recently, Derr et al. [11] categorize three primary approaches to defining calibration in multiclass settings, though we focus on two: *distribution calibration with respect to γ* and *Γ-calibration*. For intuition, consider calibration error which measures (a) conditioned on a piece of information, such as a model's prediction, the (b) distance between some prediction against a baseline. The computational and statistical challenge posed by naïve extensions of multiclass calibration is that the granularity of provided conditioning information in (a) grows exponentially in the number of classes n.

Derr et al. characterize the first set of multiclass calibration definitions as *distribution calibration with respect to γ*, where $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ is some given property representing the conditioning event (a), with granularity $|\mathcal{R}| = \mathbf{poly}(n)$ [4, 24, 47].

Definition 4 (Approximate distribution calibration with respect to γ)**.** *A distributional predictor $f : \mathcal{X} \to \Delta_{\mathcal{Y}}$ is* ϵ-distribution calibrated with respect to γ *if*

$$\mathbb{E}_X \| f(X) - D_{Y|\gamma \circ f, \gamma(f(X))} \| \leq \epsilon \, .$$

Of course, the choice of norm affects the meaning of a calibration bound; this is demonstrated by considering the diameter of the simplex for different p-norms: with $p = 1$, the simplex has diameter 2; with $p = 2$, the diameter is $\sqrt{2}$, and with $p = \infty$, the diameter is 1. Generally speaking, we use an unspecified L_p norm for $p \geq 1$ following the precedent of Garg et al. [21]. Garg et al. [21, Definition 2.3] discusses the most common metric choices being L_1, L_2, and L_∞ norms, and proceed to give results relating approximate calibration coefficients with different choices of L_p-norms. Relationships with other calibration error metrics are discussed in Section B.

Moving one step beyond distribution calibration with respect to γ, Γ-calibration focuses on the distance of predictions and baseline in (b), measuring distance in *prediction space* $\mathbf{im}(\Gamma)$ rather than distribution space $\Delta_{\mathcal{Y}}$. For discrete predictions and properties, this distance is not continuous in function space, and approximate calibration is not empirically stable. To sidestep this issue, Noarov and Roth [35] and Gneiting and Resin [22] assume properties Γ are continuous, Derr et al. [11] only give *exact* Γ-calibration results for discrete properties, Hu et al. [28] focus on binary outcome settings (emphasizing connections to calibration error metrics), and Collina et al. [10] study the sample complexity of multicalibration for continuous scalar properties.

Derr et al. [11] discuss approximate Γ-calibration in terms of an arbitrary metric, but do not give significant approximate results for discrete properties. With this insight, we use continuous properties so that we can use L_p-norms as a distance metric, and bound the approximate calibration of discrete properties through 0-1 error.

Definition 5 (Approximate Γ-calibration for continuous properties)**.** *Consider a continuous property $\Gamma : \Delta_{\mathcal{Y}} \to \mathbb{R}^d$ and Γ-predictor $g : \mathcal{X} \to \mathbb{R}^d$. g is ϵ-approximately Γ-calibrated if, $\mathbb{E}_X \|\Gamma(D_{Y|g,g(X)}) - g(X)\| \leq \epsilon$.*

When defining a miscalibration metric for discrete properties, we use the 0-1 metric $m(s, r) = \mathbb{1}(s \neq r)$. While this is not the only possible miscalibration metric for discrete properties, it does serve as an

upper bound on other decision-theoretic metrics, which are generally 0 if correct decisions are made. See Section B for further discussion.

The discrete predictor in the following definition $h : \mathcal{X} \to \mathcal{R}$ is often not directly optimzed, but formed as a post-processing of some continuous predictor. That is, we often have $h = \psi \circ g$ for a link function $\psi : \mathbb{R}^d \to \mathcal{R}$.

Definition 6 (Approximate γ-calibration for discrete properties)**.** *Consider a discrete property $\gamma : \Delta_\mathcal{Y} \to \mathcal{R}$ for a finite set $\mathcal{R}$ and γ-predictor $h : \mathcal{X} \to \mathcal{R}$. h is ϵ-approximately γ-calibrated if* $\Pr_X[\gamma(D_{Y|h,h(X)}) \neq h(X)] \leq \epsilon$.

With preliminaries established, we now restate our goal: we aim to understand the relationship between the approximate distribution calibration of a predictor $f : \mathcal{X} \to \Delta_\mathcal{Y}$ and the approximate γ-calibration of its post-processing $\psi \circ \Gamma \circ f$, where Γ is a continuous property refining γ. We now proceed by constructing Γ for strongly orderable γ.

3 Characterizing 1-dimensional Lipschitz elicitable properties

In this section, we establish that orderable properties are refined by 1-dimensional Lipschitz properties $\Gamma : \Delta_\mathcal{Y} \to \mathbb{R}$, and as a consequence, we conclude that for any strongly orderable property γ, $\mathbf{elic}_{\mathcal{C}_{\text{Lip}}}(\gamma) = 1$. We later apply this characterization to establish approximate calibration guarantees for such discrete orderable properties. Notably, our characterization is constructive, as Algorithm 1 and Algorithm 2 yield Lipschitz properties Γ refining γ. Algorithm 1 modifies the (piecewise linear, convex, but not differentiable) surrogate construction of Finocchiaro et al. [13, Theorem 11] and interpolates to "smooth" their surrogate; such an interpolation is not straightforward to ensure property refinement. This approach is also very similar to the approach of Khurana et al. [29, Appendix H], though they do not establish Lipschitzness. We conjecture this approach might be generalized to higher dimensions, but leave this to future work.

In contrast, Algorithm 2 uses the geometry of the discrete property γ directly, constructing a property Γ which is a piecewise ratio of expectations, where the functions forming the numerator and denominator are defined by the hyperplanes defining an orderable property. Crucially, it is not clear if it possible to generalize this approach for $d > 1$. With the Lipschitz property Γ in both algorithms, we apply the characterizations of Steinwart et al. [44] and Finocchiaro and Frongillo [12] to construct a surrogate loss indirectly eliciting γ.

Both Algorithm 1 and Algorithm 2 return surrogate losses eliciting Lipschitz continuous properties, but these properties are not necessarily bounded in $[0, 1]^d$. They are however, generally bounded, and therefore can be scaled by $L^*(u, y) := L(\frac{u-\Gamma_{\min}}{\Gamma_{\max}-\Gamma_{\min}}, y)$. For simplicity, we just return L.

Motivating example Consider the loss matrix $\ell^\dagger$ and elicited discrete property γ as follows, with outcomes labeled $\mathcal{Y} = \{1, 2, 3\}$.

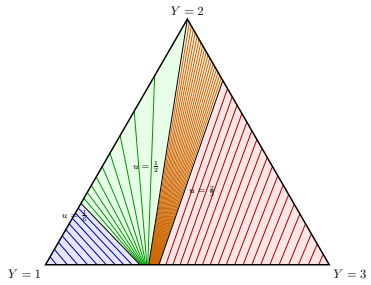
$$\ell = \begin{bmatrix} 0 & 3 & 5 \\ 1 & 0 & 3 \\ 3 & 1 & 0 \end{bmatrix}, \qquad \gamma(p) = \begin{cases} 1 & -3p_1 + p_2 \leq -2 \\ 2 & -3p_1 + p_2 \geq -2 \wedge 5p_1 + 4p_2 \geq 3 \\ 3 & 5p_1 + 4p_2 \leq 3 \end{cases} \tag{1}$$

While ℓ and γ are discrete, we seek a continuous relaxation Γ of γ in order to show $\mathbf{elic}_{\mathcal{C}_{\text{Lip}}}(\gamma) = 1$. Figure 2c depicts the level sets $\Gamma^{-1}(u)$ of the continuous properties yielded from Algorithm 1 and Algorithm 2, respectively. Applications of our algorithms to this example are in Section C.

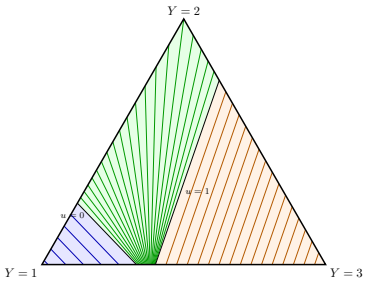
3.1 Algorithm from embeddings

Algorithm 1 uses the *embeddings* approach of Finocchiaro et al. [15]. Given a discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$, Finocchiaro et al. provide a construction of a piecewise linear and convex surrogate $L : \mathbb{R}^{n-1} \times \mathcal{Y} \to \mathbb{R}_+$ by "embedding" discrete reports into a continuous space for optimization through the function $\varphi : \mathcal{R} \to \mathbb{R}^{n-1}$. When properties are *orderable*, Finocchiaro et al. [13, Theorem 11] produces a 1-dimensional surrogate loss $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ and embedding $\varphi : \mathcal{R} \to \mathbb{R}$ so that L indirectly elicits γ elicited by ℓ. However, since their embeddings are piecewise linear, they do

[†]Recall $\ell_{r,y} = \ell(r, y)$ is the discrete loss with report r and outcome y.



(a) Level sets $\Gamma^{-1}(u)$ of the property Γ constructed in Algorithm 1.



(b) Level sets of the property constructed in Algorithm 2.

(c) Comparison of the continuous properties produced from Algorithm 1 and Algorithm 2, respectively, applied to Equation (1) The black contours depict the property γ in (1), or a slight refinement of it in the case of the embedding.

not elicit a Lipschitz property. We must take care to "smooth" these embeddings while maintaining property refinement guarantees.

Algorithm 1 is very similar to the construction given in Khurana et al. [29, Appendix H]. However, their construction linearly interpolates between loss values between integers in $[k]$. By adding midpoints to the interpolation set, we are able to verify refinement in Algorithm 1 as level sets of Γ coincide with boundaries of γ at these midpoints. We let $\delta^+ L(u, \cdot)$ denote the right directional derivative with respect to the first argument at u, and likewise for δ^- as the left derivative.

Algorithm 1 Algorithm for generating continuous property from discrete embedding

procedure SMOOTH EMBEDDING(Orderable γ)

Construct embedding $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ and embedding function $\varphi : \mathcal{R} \to \mathbb{R}$.

Take interpolation set $U = \varphi(\mathcal{R}) \cup \{\frac{\varphi(r_i)+\varphi(r_{i+1})}{2} : i \in [|\mathcal{R}| - 1]\}$

Take function $V(u, y) = \begin{cases} \frac{d}{du} L(u, y) & L \text{ differentiable} \\ 0 & \text{sign}(\delta^+(L(u, y))) \neq \text{sign}(\delta^-(L(u, y))) \\ \frac{\delta^+ L(u,y)+\delta^- L(u,y)}{2} & \text{otherwise} \end{cases}$

Construct $\bar{V}$ by linearly interpolating between $\{V(u, \cdot) : u \in U\}$ on $\text{conv}(U)$, and matching V outside $\text{conv}(U)$.

Integrate $\bar{V}(u, y)$ to get $\bar{L}(u, y) = \int_0^u \bar{V}(z, y)dz$

Construct $\psi : \mathbb{R} \to \mathcal{R}$ as a rounding based on the interpolation set U.

return $\bar{L}, \psi$.

The loss function constructed in Algorithm 1 is piecewise quadratic as the integration of piecewise linear functions. This construction is not unique, as exemplified by the scoring rule characterization of Steinwart et al. [44] and Finocchiaro and Frongillo [12], but one benefit of this form is that the derived property Γ is Lipschitz continuous with constant $\max_{u,y} |\bar{V}(u, y)|$; since $\bar{V}$ is bounded, this constant is bounded.

Proposition 7. *Given a strongly orderable property γ, Algorithm 1 returns a surrogate loss $\bar{L}$ and link ψ such that $\bar{L}$ elicits a Lipschitz continuous property Γ refining γ through the link ψ. Therefore,* $\mathbf{elic}_{\mathcal{C}_{\text{Lip}}}(\gamma) = 1$.

See proof in Section A.

3.2 Algorithm from normals

We now present another construction for constructing Lipschitz continuous refinements of orderable properties that directly leverages the geometric structure of the property itself, instead of "smoothing" a non-smooth surrogate. In essence, Algorithm 2 constructs a surrogate property Γ as a piecewise ratio of expectations of two linear functions, which are inner products of vectors normal to the boundaries of level sets. Strong orderability of γ is necessary here since it enables us to construct precisely one normal vector per boundary. We say a set of vectors $o_1, \ldots, o_{k-1}$ defining level sets $\gamma^{-1}(r_1), \ldots, \gamma^{-1}(r_k)$ is *oriented* if $p \in \gamma^{-1}(r_i) \iff \langle o_{i+1}, p\rangle \leq 0 \leq \langle o_i, p\rangle$ for all $i \in 1, \ldots, k$, omitting the undefined inequalities for $i \in \{1, k\}$.

Algorithm 2 Algorithm for generating Piecewise ROE from normals, where $\chi(x, I)$ is the clipping of x to interval I.

procedure GENERATE PIECEWISE ROE(Orderable γ)
  for all $i \in 1, \dots, |\mathcal{R}| - 1$ **do**
    Find $n-1$ distributions $p_{ij} \in \gamma^{-1}(r_i) \cap \gamma^{-1}(r_{i+1}) \cap \text{relint}(\Delta_{\mathcal{Y}})$ to form $P_i \in \mathbb{R}^{n-1 \times n}$, and define a normal $o_i \in \ker(P_i)$ using SVD.
    Ensure o_i is oriented
  end for
  Construct identification function
  $V(u, y) = u - \chi(u, [0, |\mathcal{R}| - 2]) - o_{1,y} - \sum_{j=1}^{|\mathcal{R}|-1} \chi(u - (j-1), [0, 1])(o_{j+1,y} - o_{j,y})$
  $L(u, y) = \int_0^u V(x, y)dx$
  $\psi : x \mapsto \chi(\lceil x \rceil, [0, |\mathcal{R}| - 1]) + 1$
  return L, ψ

Proposition 8. *Let γ be a strongly orderable finite property. Then Algorithm 2 yields a surrogate loss L and link ψ such that L elicits a Lipschitz continuous property $\Gamma : \Delta_{\mathcal{Y}} \to \mathbb{R}$ which refines γ through the link ψ.*

See proof in Section A.

3.3 Comparison of algorithms and discussion of Lipschitz constants

While both Algorithm 1 and Algorithm 2 yield Lipschitz properties Γ, they might yield *different* properties, as constructed; in fact, this is demonstrated in Figure 2c, where the additional interpolation point at the middle embedding point in Algorithm 1 adds an additional constraint at the cost of the Lipschitz constant increasing. Algorithm 1 can be modified to only interpolate between embedding midpoints, in which case, it would yield the same surrogate L up to scaling as Algorithm 2. However, we intentionally leave Algorithm 1 as written, as we expect generalizations to higher dimensions $d > 1$ will need to interpolate between these embedding points as well. See Section 5 for more discussion.

To understand if this Lipschitz constant is *optimal*, the primary remaining design modification one might consider is the choice of interpolation. While linear interpolation yields properties which are piecewise ratios of expectations and surrogates which are piecewise quadratic, one might plausibly improve the Lipschitz constant of Γ with a different interpolation. Observe that, in our setting, since properties are defined as the roots of identification functions, we can use the implicit function theorem to write the gradient $\nabla_p \Gamma(p)$ as the ratio $-\frac{[V(\Gamma(p),y)]_y}{\mathbb{E}_{Y \sim p} V'(\Gamma(p),Y)}$. With a linear interpolation (and strong orderability guaranteeing non-zero slope piecewise), we can guarantee Γ is Lipschitz. In contrast, for any interpolation with gradient approaching 0, we lose Lipschitzness; general constructions with non-linear interpolations which always yield Lipschitz constants are difficult to construct, and establishing an *improved* general Lipschitz constant is an even loftier goal. However, modifications in fixed instances which improve Lipschitz constants might be possible.

4 Application to Γ-calibration

The characterization of Lipschitz elicitation complexity, particularly with the construction of Lipschitz refining properties in Algorithms 1 and 2, enables us to establish bounds on the approximate Γ-calibration of both distributional predictors post-processed by direct computation of Γ (Section 4.1), and from Γ-predictors to discrete γ-predictors (Section 4.2). The results of these sections can be composed to obtain bounds on the probability of miscalibration with respect to a discrete property when given a distributional predictor $f : \mathcal{X} \to \Delta_{\mathcal{Y}}$.

4.1 Relationship between Distribution calibration and Γ-calibration

Theorem 9 (Γ-calibration bounds by post-processing distributional predictor)**.** *Let $\Gamma : \Delta_{\mathcal{Y}} \to \mathbb{R}$ be K-Lipschitz. Then a predictor f that is ϵ-distribution calibrated with respect to Γ yields predictor $\Gamma \circ f$ which is $K\epsilon$-approximately Γ-calibrated.*

Proof.

$$\begin{aligned}\mathbb{E}_{X,Y}\|\Gamma(f(X)) - \Gamma(D_{Y|\Gamma\circ f,\Gamma(f(X))})\| &\leq \mathbb{E}_{X,Y}[K\|f(X) - D_{Y|\Gamma\circ f,\Gamma(f(X))}\|] \\ &= K\mathbb{E}_{X,Y}\|f(X) - D_{Y|\Gamma\circ f,\Gamma(f(X))}\| \\ &\leq K\epsilon\end{aligned}$$

Observe the first inequality follows in expectation since the Lipschitz inequality holds pointwise for all $p \in \Delta_{\mathcal{Y}}$. □

This result is notably a tighter bound than that of Derr et al. [11, Proposition 7], which gives a bound of $K\epsilon|\mathcal{Y}|$ because they allow the metric measuring distance to be arbitrary. They establish their bound using total variation distance rather than pulling the Lipschitz constant out using properties of the expectation function.

As one important corollary, we observe that if the smoothed property is a contraction mapping (it is Lipschitz with constant $K < 1$), then the calibration error bound is unchanged.

Corollary 10. *Let $\Gamma : \Delta_{\mathcal{Y}} \to \mathbb{R}$ be a contraction mapping. Then a predictor f that is ϵ-distribution calibrated with respect to Γ yields predictor $\Gamma \circ f$ which is ϵ-Γ-calibrated.*

For properties Γ that are not contraction mappings, small distribution calibration with respect to Γ might be deceiving, as small calibration error can lead to very bad decisions with respect to Γ. As a lower bound, we construct a counterexample of a distributional predictor $f : \mathcal{X} \to \Delta_{\mathcal{Y}}$ where f is ϵ-distribution calibrated, but *not* ϵ-approximately Γ calibrated, despite $\mathbf{im}(\Gamma)$ being lower-dimensional than $\mathbf{im}(f)$.

Proposition 11. *Let $\Gamma : \Delta_{\mathcal{Y}} \to \mathbb{R}^d$ be a Lipschitz property with optimal constant $K = \inf\{k \geq 0 : \|\Gamma(p) - \Gamma(q)\| \leq k\|p - q\| \forall p, q \in \Delta_{\mathcal{Y}}\}$. Then for every $C < K$, there exists a data distribution over $\Delta(\mathcal{X} \times \mathcal{Y})$ and distributional predictor $f : \mathcal{X} \to \Delta_{\mathcal{Y}}$ such that $\mathbb{E}_{X,Y}\|f(X) - D_{Y|\Gamma\circ f,\Gamma(f(X))}\| = \epsilon$, yet f fails $C\epsilon$-approximate Γ calibration. That is, $\mathbb{E}_{X,Y}\|\Gamma(f(X)) - \Gamma(D_{Y|\Gamma\circ f,\Gamma(f(X))})\| > C\epsilon$.*

Proof. We show this by constructing a counterexample. For simplicity, we consider $\mathcal{X} = \{x\}$. Fix $C < K$. There must be some p, q such that $\|\Gamma(p) - \Gamma(q)\| > C\|p - q\|$, otherwise K would not be the optimal Lipschitz constant. Consider predictor $f(x) = p$ and distribution D such that $D_{Y|\Gamma(f(x))} = D_Y = q$. Now consider $\epsilon := \|p - q\| = \mathbb{E}_Y\|f(x) - D_Y\| = \mathbb{E}_{X,Y}\|f(X) - D_{Y|\Gamma\circ f,\Gamma(f(X))}\|$. This strategic choice of ϵ ensures f is ϵ-distribution calibrated with respect to Γ by construction. Therefore, we have $\|\Gamma(p) - \Gamma(q)\| > C\epsilon$, and the result follows. □

While Proposition 11 is a negative result in the sense that one can have high Γ-miscalibration with low distribution calibration, this result as emphasizes how distribution calibration might be misleading, and cannot preserve decision-theoretic guarantees for highly variable properties Γ. See Figure 3 in Section D for visual intuition about this bound.

4.2 Γ-calibration for discrete decisions γ

We now turn our focus from the Γ-calibration of *distributional* predictors $f : \mathcal{X} \to \Delta_{\mathcal{Y}}$ to the γ-calibration of Γ-predictors $g : \mathcal{X} \to \mathbb{R}$. For intuition, one can think of a predictor $g : \mathcal{X} \to \mathbb{R}$ either as an intermediate post-processing of a distributional predictor $g = \Gamma \circ f$, or as a directly-learned artifact of empirical when minimizing the surrogate loss L eliciting Γ.

Observe that some increase in calibration error is inevitable with the introduction of discretization. Hu and Wu [27] introduce *calibration decision loss* (CDL) to this effect, which defines calibration error as the worst case error induced by discretization over all possible thresholdings in the binary setting.

In what follows, we need a few geometric objects. If $g : \mathcal{X} \to \mathbb{R}$ is a scalar predictor, we let $\partial(u)$ denote the closest Γ-level set boundary aligning with a γ-level set boundary, and let $\delta(u)$ denote the distance (in Γ-space) to that boundary, with $\delta_{\min}$ the minimum $\delta(u)$ over a finite $\mathbf{im}(g)$. Visual intuition is provided in Figure 3.

Formally, we define the set S of surrogate reports "aligned" with discrete boundaries $S := \{u \in \mathbb{R} : \exists i \in [|\mathcal{R}| - 1] \text{ s.t. } \Gamma^{-1}(u) = \gamma^{-1}(r_i) \cap \gamma^{-1}(r_{i+1})\}$. This enables us to define $\partial(u) :=$

$\arg\min_{u' \in S} \|u - u'\|$, and $\delta(u) = \|u - \partial(u)\|$. In particular, if $\mathbf{im}(g)$ is finite, and $\delta_{\min} := \min_{u \in \mathbf{im}(g)} \delta(u)$.

Before proceeding to bound discrete calibration error, we note that without distributional assumptions, our bound might be unavoidably vacuous as a result of discretization. In Theorem 12, we assume $D_{Y|g,u}$ is Lipschitz in u; without such an assumption, the additive term in our bound is $2K$. Similar assumptions are common in the literature [8, 26, 43]. As K is typically greater than $\frac{1}{2}$, the second term in the bound would be greater than 1. To gain some intuition for why this bound can be vacuous, consider the property $\gamma(p) = \mathbb{1}(\mathbb{E}_{Y \sim p} Y \geq \frac{3}{2})$ for outcomes in $\mathcal{Y} = \{1, 2, 3\}$ and smooth property $\Gamma(p) = \mathbb{E}_{Y \sim p} Y$. If there is one feature $\mathcal{X} = \{x\}$ and a scalar predictor $g(x) = \frac{3}{2} - \frac{\epsilon}{2}$, while we have $\Gamma(D_{Y|g, \frac{3}{2} - \frac{\epsilon}{2}}) = \frac{3}{2} + \frac{\epsilon}{2}$, then we have $\|\Gamma(D_{Y|g, \frac{3}{2} - \frac{\epsilon}{2}}) - g(x)\| = \epsilon$. This counterexample is an artifact of the $\delta_{\min}$ denominator (smallest distance from a prediction to discrete boundary) being small (here, $\frac{\epsilon}{2}$) a value of $\delta_{\min} \leq \epsilon$ makes the bound vacuous, and further guarantees on calibration are impossible.

Theorem 12 (Calibration bounds induced by discretization)**.** *Consider a K-Lipschitz continuous property $\Gamma : \Delta\mathcal{Y} \to \mathbb{R}$ refining $\gamma : \Delta\mathcal{Y} \to \mathcal{R}$ by the link $\psi : \mathbb{R} \to \mathcal{R}$, and predictor $g : \mathcal{X} \to \mathbb{R}$ that is ϵ-approximately Γ calibrated with $\mathbf{im}(g)$ finite. Moreover, assume $D_{Y|g,u}$ is C-Lipschitz continuous in u, and $\mathbf{diam}(\gamma, \psi) := \max_{r \in \mathbf{im}(\gamma)} \max_{u,u' : \psi(u) = r} |u - u'|$. Then the post-processed predictor $\psi \circ g$ yields $\Pr[\psi(g(X)) \neq \gamma(D_{Y|\psi \circ g, \psi(g(X))})] \leq \frac{\epsilon + KC\mathbf{diam}(\gamma, \psi)}{\delta_{\min}}$).*

Proof.

$$
\begin{aligned}
&\mathbb{E}_X \|g(X) - \Gamma(D_{Y|\psi \circ g, \psi(g(X))})\| \\
&\leq \mathbb{E}_X \|g(X) - \Gamma(D_{Y|g,g(X)\}})\| + \mathbb{E}_X \|\Gamma(D_{Y|\psi \circ g, \psi(g(X))}) - \Gamma(D_{Y|g,g(X)})\| \\
&\leq \epsilon + \mathbb{E}_X \|\Gamma(D_{Y|\psi \circ g, \psi(g(X))}) - \Gamma(D_{Y|g,g(X)})\| \\
&\leq \epsilon + K\mathbb{E}_X \|D_{Y|\psi \circ g, \psi(g(X))\}} - D_{Y|g,g(X)}\| \\
&\leq \epsilon + KC\mathbf{diam}(\gamma, \psi)
\end{aligned}
$$

Note the last inequality follows from C-Lipschitz continuity of the marginal distribution in prediction space, and the diameter is the maximal distance between any predictions mapping to the same report $r \in \mathbf{im}(\gamma)$. With this bound on distance from $g(X)$ to $\Gamma(D_{Y|\psi \circ g, \psi(g(X))})$, we can apply Lemma 13 to observe

$$
\begin{aligned}
&\Pr[\gamma(D_{Y|\psi \circ g, \psi(g(X))}) \neq \psi(g(X))] \\
&\leq \Pr[\|\Gamma(D_{Y|\psi \circ g, \psi(g(X))}) - g(X)\| \geq \delta(g(X))] \\
&\leq \Pr[\delta(g(X)) < t] + \Pr[\|\Gamma(D_{Y|\psi \circ g, \psi(g(X))}) - g(X)\| \geq t] \qquad \forall t > 0 \\
&\leq \Pr[\delta(g(X)) < t] + \frac{\epsilon + KC\mathbf{diam}(\gamma, \psi)}{t} \,.
\end{aligned}
$$

The last part of the theorem statement holds by setting $t = \delta_{\min}$ and observing $\Pr[\delta(g(X)) < \delta_{\min}] = 0$. □

Importantly, as $\mathbf{diam}(\gamma, \psi)$ shrinks, one could intuitively think of γ as "approaching" Γ. In this case, the second term in the numerator disappears. In Section E, we also present an alternative notion of approximate γ-calibration which uses the expected distance to a calibration set as the error metric instead of 0-1 error.

This result can be interpreted in two primary ways: first, the bounds from Section 4.1 and Section 4.2 can be composed to conclude that a ϵ-distribution calibrated predictor $f : \mathcal{X} \to \Delta\mathcal{Y}$ can be post-processed to conclude that $\psi \circ \Gamma \circ f$ is $\frac{(K\epsilon + CK\mathbf{diam}(\gamma, \psi))}{\delta_{\min}}$-approximately γ calibrated. In contrast, we can apply Theorem 12 directly to a scalar predictor $g : \mathcal{X} \to \mathbb{R}$. As discussed after Proposition 11, the Lipschitz constant of the smoothed property Γ might impose a *cost* to calibration by predicting property values directly, so without making stringent distributional assumptions, it is hard to generally compare the Γ-miscalibration of two predictors $f : \mathcal{X} \to \Delta\mathcal{Y}$ and $g : \mathcal{X} \to \Delta\mathcal{Y}$. That is, we cannot say anything about the ordinal relationship between calibration error for a distributional predictor attained from ERM on a n-dimensional score-based loss ϵ^f and calibration error for a scalar predictor attained by ERM on a Γ-specific loss ϵ^g. Recall, however, that gradient-based optimization methods

on $g : \mathcal{X} \to \mathbb{R}^d$ are generally more efficient than the same methods applied to $f : \mathcal{X} \to \Delta_{\mathcal{Y}}$ for $d \ll n$. Moreover, straightforward extensions of the results from Collina et al. [10] imply a sample complexity of $\epsilon^{-(d+2)}$, exponentially improving in prediction dimension d.

5 Discussion

We have established bounds on the approximate calibration of discrete properties, particularly for *strongly orderable* discrete properties. This work is the first to our knowledge to establish such approximate calibration bounds for discrete properties using norm-based calibration error metrics, in contrast to the worst-case miscalibration error metrics in CDL [27] and CCE [43]. In order to establish this approximate bound, we "smoothed" discrete properties into Lipschitz continuous properties whose level sets can be mapped back to the discrete property, and established approximate calibration of the smoothed properties.

Future work Many directions of future work remain: first, our characterization of Lipschitz elicitation complexity is limited to one dimension, as is Theorem 12, relying on strong orderability of a property γ, seen in the presence of $\mathbf{diam}(\gamma, \psi)$ in the bound in Theorem 12. An interesting immediate direction of future work is to extend our characterization to general d-dimensional bounds.

Another line of future work emerges from the observation that one advantage of distribution calibration (with respect to γ) emerges from strong connections between (multi)calibration and to omniprediction, where a model can be post-processed to yield decision-theoretic guarantees for a wide suite of properties γ. We conjecture this characterization might explain some of the results of the efficiency of the `TreeCal` algorithm of Fishelson et al. [16] also used by Peng [37].

Finally, one might consider modifications of our proposed algorithms which give the *smoothest* possible properties, possibly by modifying interpolation points to not be midpoints in Algorithm 1 or uniformly spaced in Algorithm 2, tightening the bounds derived in Section 4.

Broader impacts While our work is theoretical in nature, our results make salient the deception of "low calibration error" made possible by simply picking a property Γ that is quite smooth. We view this as not unique to our work, but do advise proceeding with caution when drawing conclusions based on low calibration error.

Acknowledgments and Disclosure of Funding

We would like to thank Rabanus Derr, Sanket Shah, Rafael Frongillo, Bo Waggoner, Georgy Noarov, Natalie Collina, Sarah Fleming, Patrick Lanza, and anonymous reviewers for helpful discussions and feedback. Victor Ganson was funded by a Undergraduate Research Fellowship from Boston College.

References

[1] A. Agarwal and S. Agarwal. On Consistent Surrogate Risk Minimization and Property Elicitation. In *Proceedings of The 28th Conference on Learning Theory*, pages 4–22. PMLR, June 2015. URL https://proceedings.mlr.press/v40/Agarwal15.html. ISSN: 1938-7228.

[2] F. Aurenhammer. Power Diagrams: Properties, Algorithms and Applications. *SIAM Journal on Computing*, 16(1):78–96, Feb. 1987. ISSN 0097-5397, 1095-7111. doi: 10.1137/0216006. URL http://epubs.siam.org/doi/10.1137/0216006.

[3] J. Bailie and R. Derr. Property elicitation on imprecise probabilities, 2025. URL https://arxiv.org/abs/2507.05857.

[4] K. Bairaktari and H. L. Nguyen. Sample-efficient multiclass calibration under ℓ_p error. *arXiv preprint arXiv:2509.23000*, 2025.

[5] H. Bao. Proper Losses, Moduli of Convexity, and Surrogate Regret Bounds. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 525–547. PMLR, July 2023. URL https://proceedings.mlr.press/v195/bao23a.html. ISSN: 2640-3498.

[6] H. Bao, C. Scott, and M. Sugiyama. Calibrated Surrogate Losses for Adversarially Robust Classification, May 2021. URL http://arxiv.org/abs/2005.13748. arXiv:2005.13748 [stat].

[7] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156, Mar. 2006. ISSN 0162-1459. doi: 10.1198/016214505000000907. URL https://doi.org/10.1198/016214505000000907. Publisher: ASA Website _eprint: https://doi.org/10.1198/016214505000000907.

[8] J. Błasiok and P. Nakkiran. Smooth ece: Principled reliability diagrams via kernel smoothing. In *International Conference on Learning Representations*, 2024.

[9] J. Błasiok, P. Gopalan, L. Hu, and V. Guruswami. A Unifying Theory of Distance from Calibration. *Symposium on the Theory of Computing*, 2023. doi: 10.1145/3564246.3585182.

[10] N. Collina, J. Lu, G. Noarov, and A. Roth. The sample complexity of multicalibration, 2026. URL https://arxiv.org/abs/2604.21923.

[11] R. Derr, J. Finocchiaro, and R. C. Williamson. Three types of calibration with properties and their semantic and formal relationships. *Journal of Machine Learning Research*, 2026.

[12] J. Finocchiaro and R. Frongillo. Convex Elicitation of Continuous Properties. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/e9510081ac30ffa83f10b68cde1cac07-Abstract.html.

[13] J. Finocchiaro, R. Frongillo, and B. Waggoner. Embedding Dimension of Polyhedral Losses. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 1558–1585. PMLR, July 2020. URL https://proceedings.mlr.press/v125/finocchiaro20a.html. ISSN: 2640-3498.

[14] J. Finocchiaro, R. Frongillo, and B. Waggoner. Unifying lower bounds on prediction dimension of convex surrogates. In *Advances in Neural Information Processing Systems*, volume 34, pages 22046–22057. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b91a76b0b2fa7ce160212f53f3d2edba-Abstract.html.

[15] J. Finocchiaro, R. M. Frongillo, and B. Waggoner. An Embedding Framework for the Design and Analysis of Consistent Polyhedral Surrogates. *Journal of Machine Learning Research*, 25 (63):1–60, 2024. URL https://www.jmlr.org/papers/v25/22-0743.html.

[16] M. Fishelson, N. Golowich, M. Mohri, and J. Schneider. High-dimensional calibration from swap regret. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[17] T. Fissler and J. Ziegel. Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4):1680–1707, 2016. doi: 10.1214.

[18] R. Frongillo and I. A. Kash. Vector-Valued Property Elicitation. In *Proceedings of The 28th Conference on Learning Theory*, pages 710–727. PMLR, June 2015. URL https://proceedings.mlr.press/v40/Frongillo15.html. ISSN: 1938-7228.

[19] R. Frongillo and I. A. Kash. Elicitation complexity of statistical properties. *Biometrika*, 108(4):857–879, Dec. 2021. URL https://academic.oup.com/biomet/article-abstract/108/4/857/5955754?redirectedFrom=fulltext.

[20] R. M. Frongillo and I. A. Kash. Elicitation complexity of statistical properties. *Biometrika*, 108 (4):857–879, 2021.

[21] S. Garg, C. Jung, O. Reingold, and A. Roth. Oracle Efficient Online Multicalibration and Omniprediction, July 2023. URL http://arxiv.org/abs/2307.08999. arXiv:2307.08999 [cs, stat].

[22] T. Gneiting and J. Resin. Regression diagnostics meets forecast evaluation: conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17(2):3226–3286, Jan. 2023. ISSN 1935-7524, 1935-7524. doi: 10.1214/23-EJS2180. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-17/issue-2/Regression-diagnostics-meets-forecast-evaluation--conditional-calibration-reliab 10.1214/23-EJS2180.full. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

[23] P. Gopalan, M. P. Kim, M. A. Singhal, and S. Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, pages 3193–3234. PMLR, 2022.

[24] P. Gopalan, L. Hu, and G. Rothblum. On Computationally Efficient Multi-Class Calibration. *arXiv.org*, 2024. doi: 10.48550/arxiv.2402.07821.

[25] N. Haghtalab, M. Qiao, K. Yang, and E. Zhao. Truthfulness of calibration measures. *Advances in Neural Information Processing Systems*, 37:117237–117290, 2024.

[26] J. Hartline, Y. Wu, and Y. Yang. Smooth calibration and decision making. In *6th Symposium on Foundations of Responsible Computing (FORC 2025)*, pages 16–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2025.

[27] L. Hu and Y. Wu. Predict to Minimize Swap Regret for All Payoff-Bounded Tasks, Apr. 2024. URL http://arxiv.org/abs/2404.13503. arXiv:2404.13503 [cs, stat].

[28] L. Hu, H. Luo, S. Senapati, and V. Sharan. Efficient swap multicalibration of elicitable properties, 2025. URL https://arxiv.org/abs/2511.04907.

[29] D. Khurana, A. Thilagar, D. Kimpara, and R. Frongillo. Consistency conditions for differentiable surrogate losses. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[30] M. Kull, M. P. Nieto, M. Kängsepp, T. M. S. Filho, H. Song, and P. A. Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *Neural Information Processing Systems*, 2019.

[31] N. Lambert and Y. Shoham. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM conference on Electronic commerce*, EC '09, pages 109–118, New York, NY, USA, July 2009. Association for Computing Machinery. ISBN 978-1-60558-458-4. doi: 10.1145/1566374.1566391. URL https://dl.acm.org/doi/10.1145/1566374.1566391.

[32] N. S. Lambert. Elicitation and Evaluation of Statistical Forecasts. 2019.

[33] N. S. Lambert, D. M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM conference on Electronic commerce*, EC '08, pages 129–138, New York, NY, USA, July 2008. Association for Computing Machinery. ISBN 978-1-60558-169-9. doi: 10.1145/1386790.1386813. URL https://dl.acm.org/doi/10.1145/1386790.1386813.

[34] M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[35] G. Noarov and A. Roth. The Statistical Scope of Multicalibration. In *Proceedings of the 40th International Conference on Machine Learning*, pages 26283–26310. PMLR, July 2023. URL https://proceedings.mlr.press/v202/noarov23a.html. ISSN: 2640-3498.

[36] E. Nueve, B. Waggoner, D. Kimpara, and J. Finocchiaro. Trading off consistency and dimensionality of convex surrogates for multiclass classification. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 103005–103030. Curran Associates, Inc., 2024. doi: 10.52202/079017-3272. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ba9e6976157ee25a10c2e9d95102590f-Paper-Conference.pdf.

[37] B. Peng. High dimensional online calibration in polynomial time. *arXiv preprint arXiv:2504.09096*, 2025.

[38] M. Pęski and C. Stewart. Nondistortionary belief elicitation, 2026. URL https://arxiv.org/abs/2506.12167.

[39] M. Qiao and E. Zhao. Truthfulness of decision-theoretic calibration measures. In N. Haghtalab and A. Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4686–4739. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/qiao25a.html.

[40] H. G. Ramaswamy and S. Agarwal. Convex Calibration Dimension for Multiclass Loss Matrices. *Journal of Machine Learning Research*, 17(14):1–45, 2016. ISSN 1533-7928. URL http://jmlr.org/papers/v17/14-316.html.

[41] H. G. Ramaswamy, A. Tewari, and S. Agarwal. Consistent algorithms for multiclass classification with an abstain option, 2016. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-12/issue-1/Consistent-algorithms-for-multiclass-classification-with-an-abstain-option/10.1214/17-EJS1388.full.

[42] R. T. Rockafellar and R. J. Wets. *Variational analysis*. Springer, 1998.

[43] R. Rossellini, J. A. Soloff, R. F. Barber, Z. Ren, and R. Willett. Can a calibration metric be both testable and actionable? In *The Thirty Eighth Annual Conference on Learning Theory*, pages 4937–4972. PMLR, 2025.

[44] I. Steinwart, C. Pasin, R. C. Williamson, and S. Zhang. Elicitation and Identification of Properties. *Conference on Learning Theory*, 2014.

[45] Y. Wang and C. Scott. Weston-watkins hinge loss and ordered partitions. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19873–19883. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e5e6851e7f7ffd3530e7389e183aa468-Paper.pdf.

[46] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, Feb. 2004. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1079120130. URL https://projecteuclid.org/journals/annals-of-statistics/volume-32/issue-1/Statistical-behavior-and-consistency-of-classification-methods-based-on-convex/10.1214/aos/1079120130.full. Publisher: Institute of Mathematical Statistics.

[47] S. Zhao, M. P. Kim, R. Sahoo, T. Ma, and S. Ermon. Calibrating Predictions to Decisions: A Novel Approach to Multi-Class Calibration. *Neural Information Processing Systems*, 2021.

A Omitted proofs

Proposition 7. *Given a strongly orderable property γ, Algorithm 1 returns a surrogate loss $\bar{L}$ and link ψ such that $\bar{L}$ elicits a Lipschitz continuous property Γ refining γ through the link ψ. Therefore,* $\mathbf{elic}_{\mathcal{C}_{\mathrm{Lip}}}(\gamma) = 1$.

Proof of Proposition 7. We show (a) a closed form of Γ elicited by $\bar{L}$ constructed in Algorithm 1, (b) Γ is Lipschitz, and (c) Γ refines γ.

(a) The spirit of Algorithm 1 is to "smooth" a non-smooth loss into a smooth one by interpolating a step function V into a piecewise linear function $\bar{V}$. Steinwart et al. [44, Theorem 5] establishes that if a continuous function $\bar{V}(u, y)$ yields loss $\bar{L}(u, y) := \int_0^u \bar{V}(z, y)dz$, then $\bar{L}$ elicits the property Γ, Γ is the root of $\mathbb{E}_{Y\sim p}\bar{V}(\cdot, Y)$. That is, $\mathbb{E}_{Y\sim p}\bar{V}(u, Y) = 0 \iff u = \Gamma(p)$. Therefore, we can derive $\Gamma(p)$ by examining the root of $\bar{V}$. Since $\bar{V}$ is piecewise linear, we examine each piece j independently for its root:

$$\mathbb{E}_{Y\sim p}\bar{V}(u, Y) = 0 \iff \sum_y p_y(a_y^{(j)}u + b_y^{(j)}) = 0 \iff u = -\frac{\langle p, b^{(j)}\rangle}{\langle p, a^{(j)}\rangle}\,.$$

Writing $\bar{V}(u, y)$ as a linear interpolation, we obtain $\bar{V}(u, y) = V(s_{i-1}, y) + (u - s_{i-1})\left(\frac{V(s_i,y)-V(s_{i-1},y)}{s_i - s_{i-1}}\right)$ for some $s_{i-1}, s_i \in U$, which enables us to re-write

$$\Gamma(p) = -\frac{\mathbb{E}_{Y\sim p}V(s_{j(p)-1}, Y) - s_{j(p)-1}\frac{\mathbb{E}_{Y\sim p}V(s_{j(p)},Y)-\mathbb{E}_{Y\sim p}V(s_{j(p)-1},Y)}{s_{j(p)}-s_{j(p)-1}}}{\frac{\mathbb{E}_{Y\sim p}V(s_{j(p)},Y)-\mathbb{E}_{Y\sim p}V(s_{j(p)-1},Y)}{s_{j(p)}-s_{j(p)-1}}}\,,$$

where $j(p) = \min\{l : \mathbb{E}_{Y\sim p}V(U_l, Y) > 0\}$, where U_l is the lth ordered element of U (and $\min(\emptyset) := 0$). Crucially, this implies that the level sets defining cases of $\Gamma(p)$ are contained in U, as $j(p)$ increments.

(b) Notably, Γ is Lipschitz continuous as long as $\langle p, a^{(j)}\rangle$ is bounded away from 0 for all pieces j. If this value did approach 0 for some p, then the identification function becomes flat on an interval of nonzero (Lebesgue) measure, meaning the hyperplanes separating discrete level sets are not bounded away from each other, which contradicts strong orderability of γ.

(c) Γ is a piecewise ratio of expectations, where pieces correspond to the interpolation set. We claim that, for each $i \in \{1, 2, \ldots, k-1\}$, there is an element $s \in U$ of the interpolation set such that $\Gamma^{-1}(s) = \gamma^{-1}(r_i) \cap \gamma^{-1}(r_{i+1})$. To construct the link function, it helps to decompose $U = \varphi(\mathcal{R}) \cup S$, where $S = \{s_1, \ldots, s_{k-1}\}$ is the constructed set of midpoints. Consider the link function $\psi(u) = r_k$, with $k = \#\{j : s_j < u\} + 1$.

Since $\psi(u)$ only increments at elements of S, $j(p)$ only increments at elements of U, and elements of U are ordered alternating embedding points and midpoints (in S), $\psi \circ \Gamma$ can be rewritten $p \mapsto \lceil \frac{j(p)}{2} \rceil$, which aggregates level sets at the interpolation points in S. Therefore, Γ refines γ.

□

Proposition 8. *Let γ be a strongly orderable finite property. Then Algorithm 2 yields a surrogate loss L and link ψ such that L elicits a Lipschitz continuous property $\Gamma : \Delta_{\mathcal{Y}} \to \mathbb{R}$ which refines γ through the link ψ.*

Proof of Proposition 8. Without loss of generality, define $\gamma(p) = \min\{i \in [k-1] : \langle o_i, p\rangle \geq 0\}$. If this set is empty, define $\gamma(p) = 0$.

We aim to (a) derive a closed form of the property Γ elicited by the loss returned in Algorithm 2, (b) Γ is continuous in u for all $y \in \mathcal{Y}$, and (c) there is a link function $\psi : \mathbb{R} \to \{1, \ldots, k\}$ such that $\psi(\Gamma(p)) = \gamma(p)$. If all of these statements hold, then Γ refines γ by the link function ψ.

(a) Closed form of Γ Consider 3 cases for the construction of V:

$$V(u,y) = \begin{cases} u - o_{1,y} & u \le 0 \\ -o_{i,y} - (u-(i-1))(o_{i+1,y} - o_{i,y}) & u \in [0, k-1] \\ u - (k-1) - o_{k,y} & u \ge k-1 \end{cases},$$

where the intermediate case is a simplification arising from a telescoping sum. For each of these cases, we use the scoring rule characterization of Steinwart et al. [44] and set $\mathbb{E}_{Y\sim p} V(u,Y) = 0 \iff u = \Gamma(p)$ to obtain

$$\Gamma(p) = \begin{cases} \langle o_1, p\rangle & \langle o_1, p\rangle \le 0 \\ \frac{\langle o_i, p\rangle}{\langle o_i - o_{i+1}, p\rangle} + (i-1) & \exists i \in [k-1] : \langle o_{i+1}, p\rangle \le 0 \le \langle o_i, p\rangle \\ \langle o_k, p\rangle + (k-1) & \langle o_k, p\rangle \ge 0 \end{cases}.$$

Since the normals $\vec{o}$ are oriented, we conclude that every $p \in \Delta_{\mathcal{Y}}$ satisfies one of these criteria. Moreover, on the boundaries, we can see that Γ is continuous the boundary criterion of $\langle o_i, p\rangle = 0$ implies $\Gamma(p) = (i-1)$, while the boundary criterion of $\langle o_{i+1}, p\rangle = 0$ implies $\Gamma(p) = 1 + (i-1) = i$.

(b) Γ is Lipschitz Observe that a ratio of linear functions is Lipschitz continuous as long as the numerator and denominator are bounded, and the denominator is bounded away from 0. Since the domain of γ is the simplex, and is therefore bounded, so are the numerator and denominator. Moreover, the denominator is bounded away from 0 if all pairs of hyperplanes forming boundaries of γ are bounded away from each other, which follows from strong orderability of γ.

(c) Finally, consider the link $\psi : x \mapsto \chi(\lceil x \rceil, [0, |\mathcal{R}| - 1]) + 1$. If $\Gamma(p) < 0$, then $\psi(\Gamma(p)) = 1$, for $\Gamma(p) \in [1, k]$, we have $\psi(\Gamma(p)) = \lceil \Gamma(p) \rceil$, and for $\Gamma(p) \ge k = |\mathcal{R}| - 1$, we have $\psi(\Gamma(p)) = |\mathcal{R}|$. Therefore, for all $p \in \Delta_{\mathcal{Y}}$, $\Gamma(p) < 0 \implies \langle o_1, p\rangle < 0 \implies \gamma(p) = 1 = \chi(\lceil (\Gamma(p)) \rceil, [0, k-1]) + 1$. Similarly, $\Gamma(p) \in [i-1, i) \implies \langle o_i, p\rangle \ge 0 \ge \langle o_{i+1}, p\rangle \implies \gamma(p) = i + 1 = \lceil \Gamma(p) \rceil = \chi(\lceil \Gamma(p) \rceil, [0, k-1]) + 1$ for any $i \in [k]$. Finally, $\Gamma(p) \ge k \implies \langle o_k, p\rangle \ge 0 \implies \gamma(p) = k = \chi(\lceil (\Gamma(p)), [0, k-1]) + 1$. Therefore, Γ refines γ by the link ψ. □

Lemma 13. *Suppose a discrete property $\gamma : \Delta_{\mathcal{Y}} \to \mathcal{R}$ is refined by Lipschitz continuous property $\Gamma : \Delta_{\mathcal{Y}} \to \mathbb{R}$ through the link $\psi : \mathbb{R} \to \mathcal{R}$. Let $g : \mathcal{X} \to \mathbb{R}$ be a scalar predictor. Then for any $p \in \Delta_{\mathcal{Y}}$, we have $\Pr_{X,Y}[\gamma(p) \ne \psi(g(X))] \le \Pr[\delta(g(X)) < t] + \Pr[\|\Gamma(p) - g(X)\| \ge t]$ for all $t > 0$.*

Proof of Lemma 13. Observe that for any $u \in \mathbb{R}$ and $p \in \Delta_{\mathcal{Y}}$, we have $\|\Gamma(p) - u\| \le \delta(u) \implies \gamma(p) = \psi(u)$.

Therefore

$$\begin{aligned} \Pr[\gamma(p) \ne \psi(g(X))] &\le \Pr[\|\Gamma(p) - g(X)\| \ge \delta(g(X))] \\ &\le \Pr[\delta(g(X)) < t] + \Pr[\|\Gamma(p) - g(X)\| \ge t] \qquad \forall t > 0 \end{aligned}$$

The second inequality follows from decomposing the first statement into two events. □

B Omitted discussion of calibration error metrics

Calibration error metrics It is worth discussing the emergence of *calibration error metrics* as a unit of study in their own right. Rossellini et al. [43] pose axioms of calibration metrics of *testability* — the ability to estimate a metric with finitely many samples — and *actionability* — obtaining decision-theoretic guarantees from an approximately "calibrated" predictor. Haghtalab et al. [25] discusses *truthfulness* of calibration metrics, lated extended by Qiao and Zhao [39], who apply their results to decision-theoretic (*actionable*) calibration metrics. While expected calibration error [34] is the "canonical" calibration error metric, it is not empirically testable with finite samples because of a binning requirement. In contrast, Błasiok et al. [9] propose Distance to Calibration as a testable calibration error metric, but it is notably not actionable. Other decision-theoretic calibration metrics have emerged in binary settings, like Calibration Decision Loss [27] and Cutoff Calibration Error [43], but for the purposes of this paper, we generally consider calibration error measured by an expected

norm. We leave it as a further line of inquiry to understand formally which calibration error metrics are suitable for models which estimate properties $\Gamma(p)$ instead of distributions p. Notably Bailie and Derr [3] recently examine the elicitation of "imprecise probabilities," which may be helpful for this problem.

Other possible notions of γ-miscalibration in light of our work Two natural other formalizations of approximate γ-calibration arise from our work. The first is to simply use Γ-calibration error as an approximation of γ-calibration error. If this was desired, one could simply apply the bounds established in Section 4.1, and not proceed to Section 4.2. However, this metric might be subject to Lipschitz-constant-hacking, so we note this approach warrants care.

Another potential formalization to approximate γ-calibration is to use the expected (or maximum) distance to exact γ-calibration. Formally, if we let $d(u, S) := \inf_{u' \in S} d(u, u')$, then we can write this distance as $\mathbb{E}_X d(D_{Y|\psi\circ g,\psi(g(X))}, \gamma^{-1}(\psi(g(X))))$. Notably, if, for a report $r \in \mathcal{R}$, if $D_{Y|\psi\circ g,r} \in \gamma^{-1}(r)$, then the distance is 0. Therefore, this expected distance is positive if and only if there is some report $r \in \mathcal{R}$ (with P_X-measure > 0) such that $D_{Y|\psi\circ g,r} \notin \gamma^{-1}(r)$. However, it is worth noting that the practical auditability of the variant $\hat{\delta}_{\min}$ presented there requires being able to estimate the marginal distribution $D_{Y|\psi\circ g,r}$ for all $r \in \mathcal{R}$. See Section E for more details.

C Applications of Algorithms to Equation (1)

C.1 Algorithm 1 applied to Equation (1)

To demonstrate how Algorithm 1 works, we revisit Equation (1), where $L(u, y) := \begin{cases} \max(-3u, u, 3u - 6) & y = 1 \\ \max(3 - 3u, \frac{1}{2}u - \frac{1}{2}, 3u - 8) & y = 2 \\ \max(-3u + 5, 5 - 2u, \frac{9}{2} - \frac{3}{2}u, 3u - 9) & y = 3 \end{cases}$ as constructed from Finocchiaro et al. [13, Theorem 11] embeds ℓ, with nondifferentiable embedding points $\varphi(\mathcal{R}) = \{0, 1, 3\}$.
We construct the interpolation set with these embedding points, and the (ordered) midpoints to obtain $U = \{0, \frac{1}{2}, 1, 2, 3\}$. Moreover, we take the "pseudo-identification" function V and interpolate to obtain $\bar{V}$ for outcome $y = 1$

$$V(u,1) = \begin{cases} -3 & u < 0 \\ 0 & u = 0 \\ 1 & u \in (0,3) \\ 2 & u = 3 \\ 3 & u > 3 \end{cases} \qquad \bar{V}(u,1) = \begin{cases} u & u < 0 \\ 2u & u \in [0, \frac{1}{2}] \\ 1 & u \in (\frac{1}{2}, 2] \\ u - 1 & u > 2 \end{cases}, \tag{2}$$

and likewise for $y = 2, 3$.

Integrating $\bar{V}(u, y)$ yields a loss $\bar{L}(u, y) = \int_0^u \bar{V}(z, y)dz \implies$

$$\bar{L}(u,1) = \begin{cases} \frac{u^2}{2} & u < 0 \\ u^2 & u \in [0, \frac{1}{2}] \\ u - \frac{1}{4} & u \in (\frac{1}{2}, 2] \\ \frac{u^2}{2} - u + \frac{7}{4} & u > 2 \end{cases} \quad \text{eliciting} \quad \Gamma(p) = \begin{cases} \frac{6p_2+5p_3}{(4p_1+2p_3)} & 4p_1 + 3p_2 \le 1 \\ \frac{-4p_1+24p_2+9p_3}{(24p_2+2p_3)} & 4p_2 + 3p_3 \ge 1 \wedge p_2 + \frac{11}{4}p_3 \le 1 \\ \frac{-4p_1+2p_2+8p_3}{(2p_2+p_3)} & p_2 + \frac{11}{4}p_3 \ge 1 \wedge p_2 + 5p_3 \le 2 \\ \frac{4p_1+8p_2+18p_3}{(4p_1+5p_2+6p_3)} & p_2 + 5p_3 \ge 2 \end{cases}$$

Importantly, since $\bar{V}(u, y)$ is continuous in its first argument, we can apply the scoring rule characterization of Steinwart et al. [44, Theorem 5] to claim that $\bar{V}$ is an *identification function*, and $\bar{L}(u, y) = \int_0^u \bar{V}(z, y)dz$ elicits the property which is the root of $\mathbb{E}_{Y\sim p}\bar{V}(\cdot, Y)$. By design, the interpolation points $\frac{1}{2}$ and 2 correspond to the reports where the set-valued inverse $\Gamma^{-1}(\frac{1}{2}) = \{p \in \Delta_{\mathcal{Y}} : -3p_1 + p_2 = -2\}$ and $\Gamma^{-1}(2) = \{p \in \Delta_{\mathcal{Y}} : 5p_1 + 4p_2 = 3\}$.

Therefore, we construct the link ψ such that $\psi \circ \Gamma(p) \in \gamma(p)$ (with equality up to the boundaries of set-valued inverses) will have $\psi(r) = \begin{cases} 1 & r \le \frac{1}{2} \\ 2 & r \in (\frac{1}{2}, 2] \\ 3 & r > 2 \end{cases}$.

C.2 Algorithm 2 applied to Equation (1)

To demonstrate how Algorithm 2 works, we revisit Equation (1). With the property

$$\gamma(p) = \begin{cases} 1 & p_2 \leq -2 + 3p_1 \\ 2 & -2 + 3p_1 < p_2 \leq \frac{3-5p_1}{4} \\ 3 & p_2 \geq \frac{3-5p_1}{4} \end{cases},$$

we can take the distributions $p_{11} = [0.7, 0.1, 0.2]$ and $p_{12} = [0.68, 0.04, 0.28]$ and use SVD to obtain $o_1 = \frac{1}{\sqrt{14}}[-1, 3, 2]$ and with $p_{21} = [0.5, 0.125, 0.375]$ and $p_{22} = [.25, .4375, .3125]$ to obtain $o_2 = \frac{1}{\sqrt{14}}[-2, -1, 3]$.

We can verify these are both oriented, and derive the identification function

$$V(r, y) = \begin{cases} r - o_{1,y} & r \leq 0 \\ -o_{1,y} + (o_{1,y} - o_{2,y})r & r \in [0, 1] \\ r - 1 - o_{2,y} & r \geq 1 \end{cases} \quad \Gamma(p) = \begin{cases} \langle o_1, p \rangle & p_2 \leq -2 + 3p_1 \\ \frac{\langle o_1, p \rangle}{\langle p, o_1 - o_2 \rangle} & -2 + 3p_1 \leq p_2 \leq \frac{3-5p_1}{4} \\ \langle o_2, p \rangle + 1 & p_2 \geq \frac{3-5p_1}{4} \end{cases} \tag{3}$$

First, we can verify that when $p_2 = 3p_1 - 2$, by simplex constraints, we have $p_3 = 3 - 4p_1$, and we can parameterize $\gamma^{-1}(1) \cap \gamma^{-1}(2) = \{(p_1, 3p_1 - 2, 3 - 4p_1) : p_1 \in [\frac{2}{3}, \frac{3}{4}]\}$ with one variable. With this parameterization, we have $\langle o_1, p \rangle = 0 = \frac{\langle o_1, p \rangle}{\langle p, o_1 - o_2 \rangle}$ for all $p \in \gamma^{-1}(1) \cap \gamma^{-1}(2)$, and similarly for the boundary $p_2 = \frac{3-5p_1}{4}$.

D Visual intuition for Γ-calibration lower bound

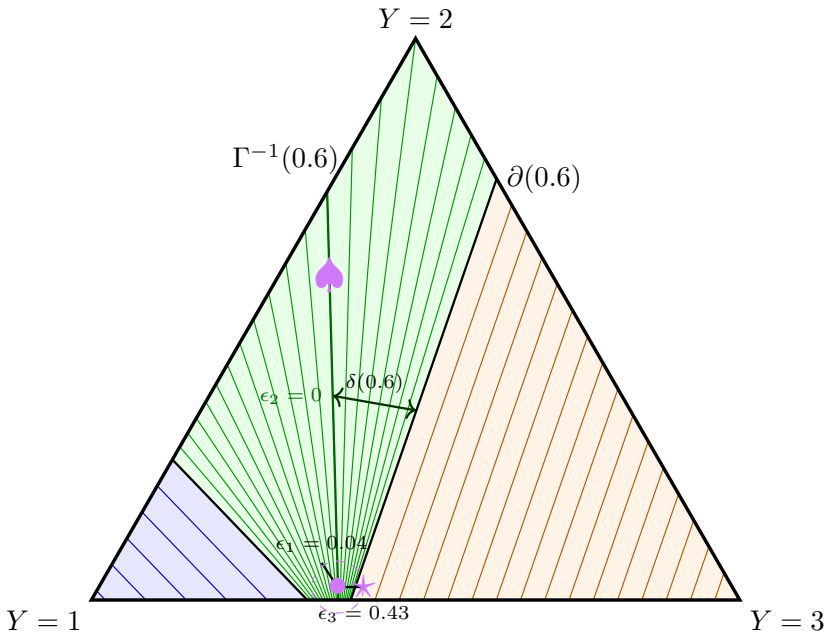


Figure 3: Let $\bullet$ denote a prediction $f(x) \in \Delta_{\mathcal{Y}}$, and consider two alternate marginal distributions $D_{Y|\{\hat{x}:\Gamma(f(\hat{x}))=\Gamma(\bullet)\}}$, $\star$ and $\spadesuit$. We also consider the level set $\Gamma^{-1}(0.6)$, in bold green, its nearest discrete boundary $\partial(0.6)$, and the distance to that boundary $\delta(0.6)$. $\delta_{\min}$ is then defined as the minimum distance over all $\delta(u)$.

Consider a predictor $f(x) = \bullet$ as in Figure 3, where $\mathcal{X} = \{x\}$ for simplicity. Moreover, we compare two alternate marginal distributions $D_{Y|\Gamma\ circf,\Gamma(\bullet)\}}$, $\star$ and $\spadesuit$, where Γ is the smooth property returned from Algorithm 2 in Section C.

If $D_{Y|\{\hat{x}:\Gamma(f(\hat{x}))=\Gamma(\bullet)\}} = \star$, then f is $\epsilon_1 = 0.04$ distribution calibrated with respect to Γ, but it is $\epsilon_3 \approx 0.43$-approximately Γ-calibrated. This jump in the calibration error emerges as $\bullet$ sits on the "most non-smooth" region of Γ. In contrast, if $D_{Y|\{\hat{x}:\Gamma(f(\hat{x}))=\Gamma(\bullet)\}} = \spadesuit$, then $\Gamma \circ f$ is approximately 0.48-distribution calibrated with respect to Γ, yet perfectly Γ-calibrated.

E Bounding expected distance as γ-miscalibration

As discussed in Section B, 0-1 miscalibration is in some sense the most pessimistic possible definition of γ-miscalibration. We now consider the expected distance to γ-calibration as a metric of γ-miscalibration.

For notational simplicity, in this section, we define a norm-induced distance metric $d(\cdot, \cdot)$ instead of writing $\| \cdot - \cdot \|$, and define the distance from a distribution to set $d(p, S) := \inf_{p' \in S} d(p, p')$ as the infimal distance to the set.

In this section, we also denote the shorthard $U := g(X) \in \mathbb{R}$, and $R := \psi(U) \in \mathcal{R}$. For $u \in \mathbb{R}$ or $r \in \mathcal{R}$, we let $p_u := \Pr[Y \mid \{\hat{x} : g(\hat{x}) = u\}]$, and likewise $p_r := \Pr[Y \mid \{\hat{x} : \psi(g(\hat{x})) = r\}]$, with p_U and p_R the corresponding random elements of $\Delta_\mathcal{Y}$ obtained by plugging in $U = g(X)$ and $R = \psi(U)$. Finally, we denote the level sets $\Gamma^{-1}(u) \equiv \Gamma_u$, and likewise for γ for a more concise notation in this section.

Observe that properties, as $\arg\min$s of expected losses are *correspondences*, or set-valued maps, in their most general form.

Assumption 14 (Metric regularity of Γ^{-1} at realized conditionals)**.** *There is a constant $\kappa < \infty$ such that*

$$d(p_u, \Gamma_u) \;\le\; \kappa\, d(u, \Gamma(p_u)) \qquad \textit{for every } u \in \operatorname{supp}(U). \tag{MR}$$

Observe that this condition is slightly weaker than the typical notion of metric regularity [42], which must hold for all (u, p) pairs, rather than (u, p_u), which is only parameterized by u.

Theorem 15. *Suppose Γ refines the elicitable property γ via ψ, $g : \mathcal{X} \to \mathbb{R}^d$ is ϵ-Γ-calibrated, and that Assumption 14 holds. Then*

$$\mathbb{E}\big[d(p_R, \gamma_R)\big] \;\le\; \kappa\epsilon. \tag{$*$}$$

Proof. The proof proceeds in four steps: first, we establish $d(p_U, \gamma_R) \le \kappa d(U, \Gamma(p_U))$ almost surely, show $p_r = \mathbb{E}_U[p_U \mid \psi(U) = r]$ for all $r \in \mathcal{R}$, apply Jensen's inequality, and apply the law of iterated expectations.

First, to show $d(p_U, \gamma_R) \le \kappa d(U, \Gamma(p_U))$ almost surely, fix $u \in \mathbb{R}$ and suppose $p \in \Gamma_u$. Since Γ refines γ via ψ, $\psi(u) \in \psi(\Gamma(p)) \subseteq \gamma(p)$, i.e. $p \in \gamma_{\psi(u)}$. Hence $\Gamma_u \subseteq \gamma_{\psi(u)}$ for all $u \in \mathbb{R}$. Consequently, for every $p \in \Delta_\mathcal{Y}$, we have $d(p, \gamma_{\psi(u)}) \le d(p, \Gamma_u)$. Specifically using $p = p_u$ (defined for $u \in \operatorname{supp}(U)$) and using Assumption 14,

$$d(p_u, \gamma_{\psi(u)}) \le d(p_u, \Gamma_u) \le \kappa\, d(u, \Gamma(p_u)) \qquad \text{for every } u \in \operatorname{supp}(U). \tag{4}$$

Since $U \in \operatorname{supp}(U)$ almost surely, evaluating (4) at the random variable U and using $R = \psi(U)$ gives

$$d(p_U, \gamma_R) \;\le\; \kappa\, d(U, \Gamma(p_U)) \qquad \text{a.s.} \tag{5}$$

Now, to see that $p_r = \mathbb{E}_U[p_U \mid \psi(U) = r]$, we apply the tower property

$$\Pr[Y = y \mid R = r] = \mathbb{E}\big[\mathbf{1}_{\{Y=y\}} \mid R = r\big] = \mathbb{E}\Big[\mathbb{E}\big[\mathbf{1}_{\{Y=y\}} \mid U\big] \;\Big|\; R = r\Big] = \mathbb{E}\big[p_U(y) \mid R = r\big]\,. \tag{6}$$

Therefore, as elements of $\Delta_\mathcal{Y}$, we have $p_r = \mathbb{E}\big[p_U \mid R = r\big]$.

Now since γ is elicitable, the level set $\gamma_r \subseteq \Delta_\mathcal{Y}$ is convex [32], and since d is induced by a norm, $\eta \mapsto d(\eta, \gamma_r)$ is a convex and continuous function on $\Delta_\mathcal{Y}$. Since $p_r = \mathbb{E}\big[p_U \mid R = r\big]$, we apply Jensen's inequality,

$$d(p_r, \gamma_r) = d\big(\mathbb{E}[p_U \mid R = r],\, \gamma_r\big) \;\le\; \mathbb{E}\big[d(p_U, \gamma_r) \mid R = r\big]\,. \tag{7}$$

Considering the event $\{R = r\}$, Equation (5) gives $d(p_U, \gamma_r) \le \kappa\, d(U, \Gamma(p_U))$, so taking conditional expectations,

$$\mathbb{E}\big[d(p_U, \gamma_r) \mid R = r\big] \;\le\; \kappa\, \mathbb{E}\big[d(U, \Gamma(p_U)) \mid R = r\big]. \tag{8}$$

Combining Equation (7) and Equation (8),

$$d(p_r, \gamma_r) \;\le\; \kappa\, \mathbb{E}\big[d(U, \Gamma(p_U)) \mid R = r\big] \qquad \text{for every } r \in \mathcal{R}. \tag{9}$$

Finally, we multiply Equation (9) by $\Pr[R = r]$ and sum over r to obtain

$$\sum_r (R = r)\, d(p_r, \gamma_r) \le \kappa \sum_r (R = r)\, \mathbb{E}\big[d(U, \Gamma(p_U)) \mid R = r\big]. \tag{10}$$

The left side of Equation (10) is $\mathbb{E}[d(p_R, \gamma_R)]$ and, by the law of iterated expectations, the right side is $\kappa\, \mathbb{E}[d(U, \Gamma(p_U))]$. Hence

$$\mathbb{E}\big[d(p_R, \gamma_R)\big] \le \kappa\, \mathbb{E}\big[d(U, \Gamma(p_U))\big] \;\le\; \kappa\epsilon, \tag{11}$$

which is Equation ($*$).

□

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our theorems include statements of assumptions, like whether items of study are discrete or continuous.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: Discussion of limitations are included not only in the conclusion, but also in relevant section. For example, we include a discussion before our main theorem statement (providing bounds on miscalibration) about how our bounds might be vacuous in certain settings.

   Guidelines:

   - The answer [N/A] 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: We take care to ensure assumptions are carefully and clearly stated. For example, strong orderability of discrete properties is explicitly stated in theorem statements, and not simply assumed.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The paper contains no experimental results.

Guidelines:

- The answer [N/A] 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: [N/A]

   Justification: The paper does not include code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: No experiments.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: No experiments.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: No experiments.

   Guidelines:

   - The answer [N/A] 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: We have read and reviewed the NeurIPS code of ethics, and our work conforms to these standards.

   Guidelines:

   - The answer [N/A] 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: While our societal impacts are limited as a theory paper, we do briefly discuss implications of acting on misleadingly low calibration errors.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Theory paper without these types of risks.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: No existing assets used.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: No new assets released.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: Crowdsourcing not conducted.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: No human subjects data collected.

    Guidelines:

    - The answer [N/A] 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.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs only used to generate TikZ figures and for an "advanced search engine" in writing proofs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.