Unifying Lower Bounds on Prediction Dimension of Consistent Convex Surrogates

Jessie Finocchiaro
University of Colorado Boulder
jessica.finocchiaro@colorado.edu

Rafael Frongillo
University of Colorado Boulder
raf@colorado.edu

Bo Waggoner
University of Colorado Boulder
bwag@colorado.edu

Abstract

The convex consistency dimension of a supervised learning task is the lowest prediction dimension d such that there exists a convex surrogate $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ that is consistent for the given task. We present a new tool based on property elicitation, d-flats, for lower-bounding convex consistency dimension. This tool unifies approaches from a variety of domains, including continuous and discrete prediction problems. We use d-flats to obtain a new lower bound on the convex consistency dimension of risk measures, resolving an open question due to Frongillo and Kash (NeurIPS 2015). In discrete prediction settings, we show that the d-flats approach recovers and even tightens previous lower bounds using feasible subspace dimension.

1 Introduction

A loss function is called a *surrogate* when it is used to solve a related, but not identical, "target" problem of interest. Selecting a hypothesis by minimizing surrogate risk is one of the most widespread techniques in supervised machine learning. There are two main reasons why a surrogate loss is necessary: (I) the target problem is to minimize a loss, the *target loss*, that does not satisfy some desiderata such as continuity or convexity; or (II) the target problem is to estimate some *target statistic* and some associated surrogate loss is required to do so, as in many continuous estimation problems. In both settings, a key criteria for choosing a surrogate loss is *consistency*, a precursor to excess risk bounds and convergence rates. Roughly speaking, consistency means that minimizing surrogate risk corresponds to solving the target problem of interest, i.e. in (I) the target risk is also minimized, or in (II) the continuous prediction approaches the true conditional statistic.

Despite the ubiquity of surrogate losses, we lack general frameworks to design and analyze consistent surrogates. This state of affairs is especially dire when one seeks low *prediction dimension*, the dimension of the surrogate prediction domain. For example, in multiclass classification with n labels, the prediction domain might be $\mathbb{R}^n$. In many type (I) settings, such as structured prediction and extreme classification, the prediction dimension of any convex and consistent surrogate often becomes intractably large, forcing one to sacrifice consistency for computational efficiency. To understand whether this sacrifice is necessary, recent work developed tools like the feasible subspace dimension to lower bound the prediction dimension of any consistent convex surrogate [33]. Challenges of type (II) include estimating risk measures such as conditional value at risk (CVaR), with applications in financial regulation, robust engineering design, and algorithmic fairness [1, 14, 35, 42]. Risk measures

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

are not elicitable, meaning they cannot be specified via a target loss, and thus we seek a surrogate loss of low (or at least finite) prediction dimension. Recent work [15, 19, 20] gives prediction dimension bounds for some of these risk measures, but without the requirement that the surrogate be convex; bounds for convex surrogates are left as a major open question.

We present a new tool, *d-flats*, which unifies existing techniques to bound the convex consistency dimension in both settings above. Using this tool, we resolve the above open question for type (II), giving the first prediction dimension bounds for risk measures with respect to convex surrogates. We also resolve a similar open question for the mode and modal interval, posed by Dearborn and Frongillo [10]. In settings of type (I), d-flats recover and tighten the feasible subspace dimension result of Ramaswamy and Agarwal [33]. Our framework rests on *property elicitation*, a weaker and simpler condition than calibration, as a way to understand consistency across a wide variety of domains.

The "four quadrants" of problem types. Above, we discuss a significant divergence in previous frameworks: constructing a surrogate given a *target loss* versus a *target statistic*. In addition to the two possible targets, we may have one of two domains: a *discrete* (i.e. finite) target prediction space, like a classification problem, or a *continuous* one, like a regression or point estimation problem. We informally refer to the four resulting cases—target loss vs. target statistic, and discrete vs. continuous predictions—as the "four quadrants" of supervised learning problems, shown in Table 1. In the context of these quadrants, Figure 1 gives a roadmap of our main results.

Literature on consistency and calibration. We focus on surrogate losses $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ that are consistent, roughly meaning that minimizing L-loss corresponds to solving the target problem of interest.

We give informal definitions of consistency in § 2.2, with formal definitions in § A.

When given a *target loss* ℓ, we roughly define L to be consistent if minimizing L, and applying a link function, minimizes ℓ [33, 39, 41, 44]. When given instead a target statistic such as the conditional quantile or variance, we introduce a notion of consistency in line

	Target loss	*Target statistic*
Discrete prediction	**Q1**, e.g. classification	**Q2**, e.g. hierarchical classification
Continuous estimation	**Q3**, e.g. least-squares regression	**Q4**, e.g. variance estimation

Table 1: The four quadrants of problem types, with an example for each as discussed in § 3.1.

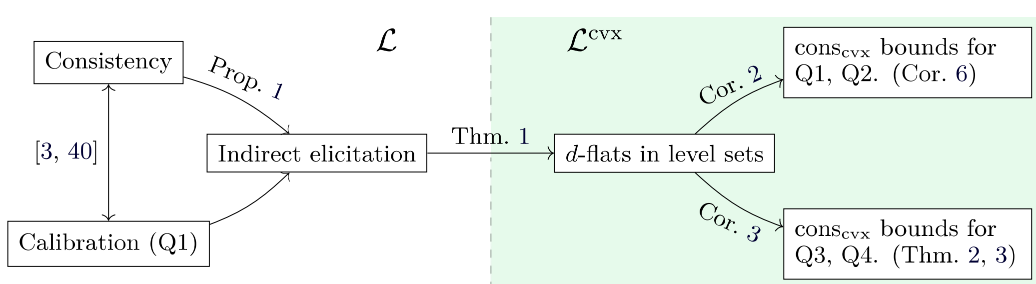

Figure 1: Flow and implications of our results. Compared to calibration, we suggest indirect elicitation as a simpler but almost-as-powerful necessary condition for consistency. In particular, we obtain a testable condition (Theorem 1), based on d-flats, for the existence of a d-dimensional consistent convex surrogate. This condition recovers and strengthens existing calibration-based results for Q1, while simultaneously applying to other quadrants. We illustrate the breadth and power of d-flats by resolving two open questions for Q3 and Q4 in § 4.

with classical statistics [11, 21, 36]. Here we define L to be consistent if minimizing L and applying a link function yields predictions converging to the correct statistic value. The key observation which underpins our approach is that consistency for target losses is a special case of consistency for target statistics (Lemma 1). Therefore, property elicitation—which studies the exact minimizers of loss functions—allows us to give general lower bounds on prediction dimension of any convex surrogates corresponding to a target task; these bounds apply across all four quadrants. See § 2.3 for other prior work on notions of prediction dimension.

As definitions of consistency are difficult to apply directly, the literature often focuses on a weaker condition called *calibration*, which only applies when given a target loss, e.g. Quadrants 1 and 3. Particularly, several authors [3, 28, 33, 41, 44] show the equivalence of consistency and calibration in Quadrant 1. We discuss the additional relationship of elicitation and calibration in § C, and re-derive Proposition 1 via calibration.

2 Setting

In supervised learning, data is drawn from a distribution D over the space $\mathcal{X} \times \mathcal{Y}$ and the goal is to produce a hypothesis $f : \mathcal{X} \to \mathcal{R}$. Here $\mathcal{X}$ is the *feature space*, $\mathcal{Y}$ the *label space*, and $\mathcal{R}$ the *report* or *prediction space*, possibly different from $\mathcal{Y}$. For example, in ranking problems, $\mathcal{R}$ may be all $|\mathcal{Y}|!$ permutations over the $|\mathcal{Y}|$ labels forming $\mathcal{Y}$. We focus on surrogate losses, target problems, and their relationships to conditional distributions $p := D_x = \Pr[Y|X = x]$ over $\mathcal{Y}$ given some $x \in X$. We can often abstract away x, working directly with a set of (conditional) distributions over outcomes $\mathcal{P} \subseteq \Delta_\mathcal{Y}$, where $\Delta_\mathcal{Y} := \{p \in \mathbb{R}^{|\mathcal{Y}|}_+ \mid \|p\| = 1\}$ is the probability simplex over labels. We then write e.g. $\mathbb{E}_p \ell(r, Y)$ to mean the expected loss of prediction $r \in \mathcal{R}$ when $Y \sim p$.

If given, we use $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ to denote a *target loss*, with predictions $r \in \mathcal{R}$. Similarly, $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ will typically denote a *surrogate* loss, with surrogate predictions $u \in \mathbb{R}^d$. In this case, d is the prediction dimension of d. We write $\mathcal{L}_d$ for the set of (Borel) $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{Y}$-measurable and lower semi-continuous surrogates $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ such that $\mathbb{E}_{Y\sim p} L(u, Y) < \infty$ for all $u \in \mathbb{R}^d, p \in \mathcal{P}$, that are minimizable in that $\arg\min_u \mathbb{E}_p L(u, Y)$ is nonempty for all $p \in \mathcal{P}$. (See § F.1 for a discussion of this assumption.) Moreover, $\mathcal{L}^{\text{cvx}}_d \subseteq \mathcal{L}_d$ is the set of convex (in $\mathbb{R}^d$ for every $y \in \mathcal{Y}$) losses in $\mathcal{L}_d$. Set $\mathcal{L} = \cup_{d\in\mathbb{N}} \mathcal{L}_d$, and $\mathcal{L}^{\text{cvx}} = \cup_{d\in\mathbb{N}} \mathcal{L}^{\text{cvx}}_d$. A loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ is *discrete* if $\mathcal{R}$ is a finite set.

2.1 Property elicitation

Arising from the statistics and economics literature, property elicitation is similar to calibration, but only characterizes exact minimizers of a surrogate [17, 18, 25–27, 30, 37]. Specifically, given a statistic or *property* Γ of interest, which maps a distribution $p \in \mathcal{P} \subseteq \Delta_\mathcal{Y}$ to the set of desired or correct predictions, the minimizers of L should precisely coincide with Γ. For example, squared loss $L(r, y) = (r - y)^2$ elicits the mean $\Gamma(p) = \{\mathbb{E}_p Y\}$.

Definition 1 (Property, elicits)**.** *A* property *is a set-valued function* $\Gamma : \mathcal{P} \to 2^\mathcal{R} \setminus \{\emptyset\}$*, which we denote* $\Gamma : \mathcal{P} \rightrightarrows \mathcal{R}$*. A loss* $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ elicits *the property* Γ *if*

$$\forall p \in \mathcal{P}, \;\; \Gamma(p) = \arg\min_{u\in\mathcal{R}} \mathbb{E}_p L(u, Y) \; . \tag{1}$$

The *level set* of Γ at value $r \in \mathcal{R}$ is $\Gamma_r := \{p \in \mathcal{P} : r \in \Gamma(p)\}$. We call a property $\Gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ *discrete* if $\mathcal{R}$ is a finite set, as in Quadrants 1 and 2. A property is *single-valued* if $|\Gamma(p)| = 1$ for all $p \in \mathcal{P}$, in which case we may write $\Gamma : \mathcal{P} \to \mathcal{R}$ and $\Gamma(p) \in \mathcal{R}$. As an example, the mean is single-valued. We define the *range* of a property by $\operatorname{range} \Gamma = \bigcup_{p\in\mathcal{P}} \Gamma(p) \subseteq \mathcal{R}$. When $L \in \mathcal{L}$, we use $\Gamma := \text{prop}_\mathcal{P}[L]$ to denote the unique property elicited by L (for distributions in $\mathcal{P}$) from eq. (1). Typically, we denote the target property by γ, and the surrogate by Γ.

2.2 Consistency and indirect elicitation

As discussed above, notions of consistency have appeared in the literature with respect to target losses, and to target statistics or properties. We give informal definitions of both notions here, with formal versions deferred to § A.

Definition 2 (Consistent: loss (informal))**.** *A loss $L \in \mathcal{L}$ and link (L, ψ) are* consistent *with respect to a target loss ℓ if, for all distributions D over $\mathcal{X} \times \mathcal{Y}$ and all sequences of measurable hypothesis functions $\{f_m : \mathcal{X} \to \mathcal{R}\}$,*

$$\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y) \implies \mathbb{E}_D \ell((\psi \circ f_m)(X), Y) \to \inf_f \mathbb{E}_D \ell((\psi \circ f)(X), Y) \ .$$

Consistency with respect to a property follows similarly, but instead of converging to the optimal target loss, one should approach the optimal (conditional) property value.

Definition 3 (Consistent: property (informal))**.** *Suppose we are given a loss $L \in \mathcal{L}$, link function $\psi : \mathbb{R}^d \to \mathcal{R}$, and property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$. Moreover, let $\mu : \mathcal{R} \times \mathcal{P} \to \mathbb{R}_+$ be any function satisfying $\mu(r, p) = 0 \iff r \in \gamma(p)$. We say (L, ψ) is* consistent *with respect to γ if, there exists a μ such that, for all D over $\mathcal{X} \times \mathcal{Y}$ and sequences of measurable functions $\{f_m : \mathcal{X} \to \mathcal{R}\}$,*

$$\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y) \implies \mathbb{E}_X \mu(\psi \circ f_m(X), D_X) \to 0 \ . \tag{2}$$

Lemma 1 in § A shows that, in fact, one can capture consistency with respect to a target loss as a special case of consistency with respect to a target property. Specifically, given a target loss ℓ, one can take $\gamma = \mathrm{prop}_{\mathcal{P}}[\ell]$ and define $\mu(r, p) := \mathbb{E}_p \ell(r, Y) - \min_{r'} \mathbb{E}_p \ell(r', Y)$ to be the ℓ-regret of the report r. This observation allows us to translate consistency from Quadrant 1 to Quadrant 2, and from Quadrant 3 to Quadrant 4; in particular, it will allow us to prove bounds for all four quadrants simultaneously.

As observed in the literature, e.g. [2, 40], both notions of consistency imply in particular that the link function must map exactly optimal surrogate reports to exactly optimal target reports. In property elicitation, this condition is known as *indirect elicitation*: for single-valued properties, Γ and ψ indirectly elicit γ if $\gamma = \psi \circ \Gamma$. The definition below covers the general set-valued case as well.

Definition 4 (Indirect Elicitation)**.** *A surrogate loss and link (L, ψ)* indirectly elicit *a property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ if L elicits a property $\Gamma : \mathcal{P} \rightrightarrows \mathbb{R}^d$ such that for all $u \in \mathbb{R}^d$, we have $\Gamma_u \subseteq \gamma_{\psi(u)}$. We say L* indirectly elicits *γ if such a link ψ exists.*

Proposition 1. *For a surrogate $L \in \mathcal{L}$, if the pair (L, ψ) is consistent with respect to a property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ or a loss ℓ eliciting γ, then (L, ψ) indirectly elicits γ.*

In other words, indirect elicitation is a necessary condition for consistency. In light of Lemma 1, we can use this fact to build prediction dimension lower bounds across all four quadrants.

Implicit in the above elicitation definitions is that L is minimizable: since $\Gamma = \mathrm{prop}_{\mathcal{P}}[L]$ is nonempty everywhere, the expected loss $\mathbb{E}_p L(\cdot, Y)$ always achieves a minimum. This restriction is also implicit in previous work, e.g., [2]. See § F.1 for further discussion.

2.3 Convex consistency dimension and elicitation complexity

Various works have studied the minimum prediction dimension d needed in order to construct a consistent surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$, typically through proxies such as calibration [2, 33, 40] and property elicitation [15, 18, 20]. Motivated by the importance of convex surrogates in machine learning, Ramaswamy and Agarwal [33] introduce the following definition for Quadrant 1; we generalize it to all quadrants.

Definition 5 (Convex Consistency Dimension)**.** *Given target loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ or property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$, its* convex consistency dimension $\mathrm{cons}_{\mathrm{cvx}}(\cdot)$ *is the minimum dimension d such that $\exists L \in \mathcal{L}_d^{\mathrm{cvx}}$ and link ψ such that (L, ψ) is consistent with respect to ℓ or γ.*

In the case of a target property γ, Lambert et al. [27] similarly introduce the notion of *elicitation complexity*. Later generalized by Frongillo and Kash [20], elicitation complexity is the lowest prediction dimension of an elicitable property, from some class of properties, from which one can compute γ. We give here the definition for convex-elicitable properties.

Definition 6 (Convex Elicitation Complexity)**.** *Given a target property γ, the* convex elicitation complexity $\mathrm{elic}_{\mathrm{cvx}}(\gamma)$ *is the minimum dimension d such that there is a $L \in \mathcal{L}_d^{\mathrm{cvx}}$ indirectly eliciting γ.*

As consistency implies indirect elicitation, we have the following.

Corollary 1. *Given a property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ or loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ eliciting γ, we have* $\mathrm{elic}_{\mathrm{cvx}}(\gamma) \leq \mathrm{cons}_{\mathrm{cvx}}(\gamma) = \mathrm{cons}_{\mathrm{cvx}}(\ell)$*.*

Finally, related to our work is the *embedding dimension* of Finocchiaro et al. [12], which is a lower bound on both convex elicitation complexity of discrete properties and convex consistency dimension of discrete losses and finite statistics.

3 Lower bounding convex consistency dimension via d-flats

We now turn to the question of bounding the convex consistency dimension for a given task. From Proposition 1, given a target property γ or loss ℓ with $\gamma = \mathrm{prop}_{\mathcal{P}}[\ell]$, this task reduces to lower bounding the convex consistency dimension of γ. Theorem 1, crystallized from the proofs of Ramaswamy and Agarwal [33, Theorem 16] and Agarwal and Agarwal [2, Theorem 9], considers a particular distribution p and surrogate prediction $u \in \mathbb{R}^d$ which is optimal for p. Theorem 1 will show that if d is small, then the level set $\{p \in \mathcal{P} : u \in \arg\min_{u'} \mathbb{E}_p L(u', Y)\}$ must be large; in fact, it must roughly contain a high-dimensional *flat* (of codimension d). By definition of indirect elicitation, there is some level set γ_r (where u is linked to r) containing this flat as well. We can then leverage the contrapositive of this result: if γ has a level set intricate enough not to contain any high-dimensional flats, then γ cannot have a low-dimensional consistent convex surrogate.

Definition 7 (d-flat)**.** *For $d \in \mathbb{N}$, a d-*flat*, or simply* flat*, is a nonempty set $F = \ker_{\mathcal{P}} W := \{q \in \mathcal{P} : \mathbb{E}_q W = \vec{0}\}$ for some measurable $W : \mathcal{Y} \to \mathbb{R}^d$.*

The following lemma yields consistency bounds when combined with Proposition 1. A similar result is found in Agarwal and Agarwal [2, Theorem 9], which bounds the dimension of level sets of a single-valued $\mathrm{prop}_{\mathcal{P}}[L]$. Theorem 1 instead bounds the dimension of flats contained in the level sets, an additional power which we leverage in our examples.

Theorem 1. *Let $\Gamma : \mathcal{P} \rightrightarrows \mathbb{R}^d$ be (directly) elicited by $L \in \mathcal{L}_d^{\mathrm{cvx}}$ for some $d \in \mathbb{N}$. Let $\mathcal{Y}$ be either a finite set, or $\mathcal{Y} = \mathbb{R}$, in which case we assume each $p \in \mathcal{P}$ admits a Lebesgue density supported on the same set for all $p \in \mathcal{P}$.[1] For all $u \in \mathrm{range}\,\Gamma$ and $p \in \Gamma_u$, there is some d-flat F such that $p \in F \subseteq \Gamma_u$.*

Proof (finite case). We will prove the result for the case of finite $\mathcal{Y}$, and defer the $\mathcal{Y} = \mathbb{R}$ case to § B. As L is convex and elicits Γ, we have $u \in \Gamma(p) \iff \vec{0} \in \partial \mathbb{E}_p L(u, Y)$. With $\mathcal{Y}$ finite, this is additionally equivalent to $\vec{0} \in \oplus_y p_y \partial L(u, y)$, where $\oplus$ denotes the Minkowski sum [23, Theorem 4.1.1].[2] Expanding, we have $\oplus_y p_y \partial L(u, y) = \{\sum_{y \in \mathcal{Y}} p_y x_y \mid x_y \in \partial L(u, y)\ \forall y \in \mathcal{Y}\}$, and thus there is a W such that $Wp = \sum_y p_y x_y = \vec{0}$ where $W = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$; cf. [33, $\mathbf{A}^m$ in Theorem 16]. Let $V_{u,p} : \mathcal{Y} \to \mathbb{R}^d, y \mapsto W_y$ be the function encoding the columns of W. Observe that $\mathbb{E}_p V_{u,p} = \vec{0}$. We take the flat $F := \ker_{\mathcal{P}} V_{u,p}$, and have $p \in F$ by construction. To see $F \subseteq \Gamma_u$, from the chain of equivalences above, we have for any $q \in \mathcal{P}$ that $q \in \ker_{\mathcal{P}} V_{u,p} \implies \vec{0} \in \partial \mathbb{E}_q L(u, Y) \implies u \in \Gamma(q) \implies q \in \Gamma_u$. □

[1] This assumption is largely for technical convenience, to ensure that $\mathcal{V}_{u,p}$ does not depend on p. Any such assumption would suffice, and we suspect even that condition can be relaxed.

[2] ∂ represents the subdifferential $\partial f(x) = \{z : f(x') - f(x) \geq \langle z, x' - x \rangle\ \forall x'\}$.

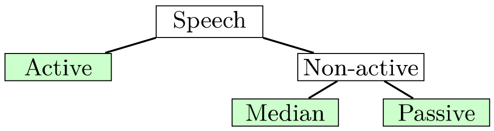

Figure 2: Hierarchical prediction example with labeling tone of speech. We take $\mathcal{Y} = \mathcal{R}$ to be the leaves of this tree, shown in blue.

Theorem 1 now allows us to derive bounds on convex consistency dimension by considering distributions and property values that are either single-valued (Corollary 2) or on the relative interior of the simplex with finite $\mathcal{Y}$ (Corollary 3). Proofs are deferred to § B.

Corollary 2. *Let target property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ and $d \in \mathbb{N}$ be given. Let $\mathcal{Y}$ be either a finite set, or $\mathcal{Y} = \mathbb{R}$, in which case we assume each $p \in \mathcal{P}$ admits a Lebesgue density supported on the same set for all $p \in \mathcal{P}$. Let $p \in \mathcal{P}$ with $|\gamma(p)| = 1$, and take $\gamma(p) = \{r\}$. If there is no d-flat F with $p \in F \subseteq \gamma_r$, then* $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$.

Corollary 3. *Let an elicitable target property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ be given, where $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ is defined over a finite set of outcomes $\mathcal{Y}$, and let $d \in \mathbb{N}$. Let $p \in \mathrm{relint}(\mathcal{P})$. If there is no d-flat F with $p \in F \subseteq \gamma_r$, then* $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$.

3.1 Illustrating the condition in all four quadrants

We now illustrate how to apply Theorem 1 to construct lower bounds on convex consistency dimension for targets across all four quadrants of Table 1. Throughout the examples, we will have $|\mathcal{Y}| = 3$ so that the probability simplex can be visualized in two dimensions (Figure 3). For each, we take $d = 1$, and thus ask whether any 1-flat (a line in the figures) passes through the point p while staying within the corresponding level set.

Q1: Classification with an abstain option. The abstain target loss is a well-studied variation of 0-1 loss that allows for an "abstain" report that gives a lesser punishment 1/2 for abstaining, $r = \perp$ [7, 8, 29, 33, 34]. Formally, the target loss is $\ell^{1/2}(r, y) := \mathbf{I}\{r \notin \{y, \perp\}\} + (1/2)\mathbf{I}\{r = \perp\}$. Since we are given a discrete target loss, this problem fits nicely into Quadrant 1.

To apply Theorem 1, we first consider the abstain property γ elicited by $\ell^{1/2}$, where one predicts the most likely outcome y if $Pr[Y = y] \geq 1/2$ and otherwise "abstains" by predicting $\perp$. For the depicted distribution $p \in \mathrm{relint}(\gamma_\perp)$, we cannot fit a 1-flat (line) fully contained in $\gamma_\perp$ that passes through p. By Corollary 3, we can conclude $\mathrm{cons}_{\mathrm{cvx}}(\gamma^{1/2}) \geq 2$ when $|\mathcal{Y}| = 3$, meaning there is no consistent convex surrogate in 1 dimension. This lower bound matches the upper bound from the convex surrogate of Ramaswamy and Agarwal [33].

Q2: Variation of hierarchical classification. Ramaswamy et al. [32] study hierarchical classification tasks, in which labels are arranged in a tree and one wishes to predict the deepest node in a tree that is "likely enough" [5, 43]. Consider the variation of this task where one can only predict leaves of this tree. For example, Figure 2 depicts a speech classification task where speech is either active or non-active, and non-active is further subdivided into median and passive. It is natural to predict active if that label is more likely than both non-active labels combined, and otherwise to predict the most likely of median and passive:

$$\gamma(p) = \begin{cases} \text{active} & p_{\text{active}} \geq 1/2 \\ \text{median} & p_{\text{active}} \leq 1/2 \wedge p_{\text{median}} \geq p_{\text{passive}} \\ \text{passive} & p_{\text{active}} \leq 1/2 \wedge p_{\text{passive}} \geq p_{\text{median}} \end{cases} .$$

This "T-shaped" property, depicted in Figure 3 (Q2), falls under Quadrant 2, as it is not elicited by any target loss.[3] Like abstain, we cannot fit a 1-flat (line) entirely contained in the level set γ_{passive} through the depicted p, so Corollary 3 gives $\mathrm{cons}_{\mathrm{cvx}}(\gamma) = 2$.

Q3: Least-squares regression Squared loss is commonly used in machine learning and statistics for continuous estimation, making it the canonical choice for Quadrant 3.

[3]The cells of finite elicitable properties form power diagrams, a generalization of Voronoi diagrams, which disallow this "T-shaped" configuration [17, 26].

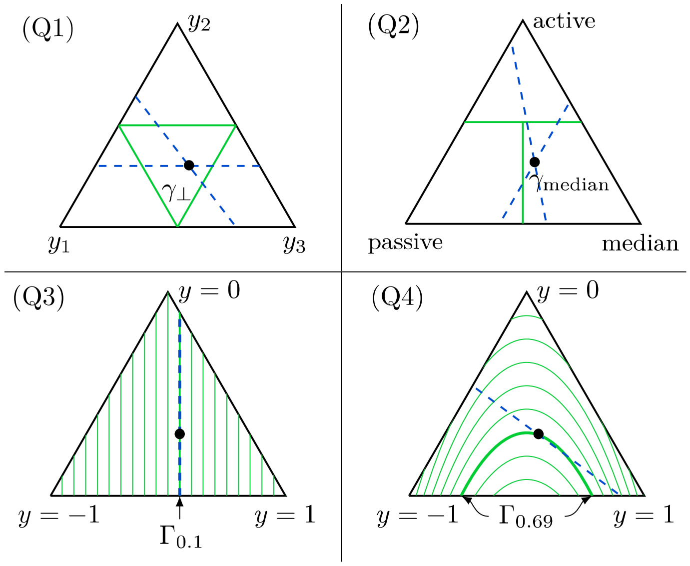

Figure 3: Example properties for each quadrant. Throughout, we take $\bullet$ to be the distribution $p = (0.3, 0.3, 0.4)$ according to the left, top, and right outcomes respectively. (Q1,Q2) We cannot fit a 1-flat (line) through p without leaving the level sets $\gamma_\perp$ and γ_{median}, respectively; Theorem 1 implies that there is no 1-dimensional consistent convex surrogate for either problem. (Q3) Squared error is a 1-dimensional convex loss, and indeed it elicits the mean of Y, whose level sets are all 1-flats. (Q4) The level sets of the variance are curved and cannot fit a 1-flat; from Theorem 1 there is no 1-dimensional convex surrogate consistent for the variance.

Squared loss is a 1-dimensional convex loss which elicits the mean $\Gamma(p) = \mathbb{E}_p[Y]$. Theorem 1 therefore states that we can fit a 1-flat through any distribution p while staying within the corresponding level set. In fact, the level sets of the mean are all exactly 1-flats, as demonstrated in Figure 3 (Q3).

Q4: Variance Consider the task of estimating the variance $\text{Var}(p) = \mathbb{E}_p[Y^2] - \mathbb{E}_p[Y]^2$. The variance is not (directly) elicitable as its level sets are not convex [27, 31], meaning this task falls under Quadrant 4. Interestingly, the fact that the variance is not elicitable does not yield a lower bound on elicitation complexity of 2, as it does not rule out the variance being a link of a real-valued convex-elicitable property; cf. Frongillo and Kash [20, Remark 1]. In § F.2, we show $\text{elic}_{\text{cvx}}(\text{Var}) = 2$, meaning the lowest dimension of a convex loss to estimate conditional variance is 2. This lower bound will follow from Theorem 2 in § 4 using the fact that variance is the Bayes risk of squared loss. While perhaps intuitively obvious, even this simple result is novel.

3.2 Relation to feasible subspace dimension

In Quadrant 1, Ramaswamy and Agarwal [33] give a lower bound on convex consistency dimension roughly by the co-dimension of the *subspace of feasible directions* $\mathcal{S}_{\mathcal{C}}(p)$ of a convex set $\mathcal{C}$ at a given distribution p such that $p \in \mathcal{C}$, which is loosely the "most full" subspace of $\mathcal{C}$ containing a neighborhood around p.

$$\mathcal{S}_{\mathcal{C}}(p) = \{v \in \mathbb{R}^n \mid \exists \epsilon_0 > 0 \text{ such that } p + \epsilon v \in \mathcal{C} \forall \epsilon \in (-\epsilon_0, \epsilon_0)\}$$

Theorem 1 subsumes the bounds given by Ramaswamy and Agarwal [33] by showing that, if there is a d-flat through p fully contained in a level set γ_r (so we can apply Theorem 1) then the subspace of feasible directions at the same $p \in \mathcal{C} := \gamma_r$ has co-dimension at most d, discussed in detail in § D.1.

Proposition 2. *Suppose we are given a discrete loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ eliciting the property $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$. Fix $p \in \text{relint}(\Delta_{\mathcal{Y}})$ and take $r \in \mathcal{R}$ such that $p \in \gamma_r$. If $\text{cons}_{\text{cvx}}(\ell) = d$, then there exists a d-flat $F \subseteq \gamma_r$ through p. Moreover, F is a subspace of feasible directions over the set γ_r intersected with the simplex. Therefore, $\text{codim}(\mathcal{S}_{\gamma_r}(p)) \leq d$, and in turn, this implies $\text{cc}\dim(\ell) \geq d \geq \text{codim}(\mathcal{S}_{\gamma_r}(p))$.*

In other words, any d-flat through p is a subspace of feasible directions of co-dimension at most d, so Theorem 1 provides a weakly tighter lower bound on convex consistency dimension than Ramaswamy and Agarwal [33, Theorem 16]. In fact, the d-flats bound can be strictly tighter; in § D we show that the abstain example from Figure 3 (Q1) yields a d-flats lower bound of 2 and a feasible subspace dimension lower bound of 1. This gap stems from the fact that feasible subspace dimension uses only local information of the property to construct lower bounds, while d-flats in Theorem 1 allow us to additionally use global information. See Figure 4 in § D for an illustration.

4 Application: Risk Measures, Mode, and Modal Interval

We now turn to two main applications of Theorem 1: new lower bounds on the convex consistency dimension of risk measures (§ 4.1) and the mode and modal interval (§ 4.2). In both cases, we build on previous results due to Frongillo and Kash [19, 20] and Dearborn and Frongillo [10] which showed lower bounds with respect to *identifiable* properties; a property is d-identifiable if its level sets are all d-flats, as in Figure 3 (Q3). In contrast, properties elicited by convex losses are generally not identifiable, particularly when the loss is non-smooth. For example, the properties elicited by hinge loss and the abstain surrogate are not identifiable, as their level sets are not flats; see Figure 3 (Q1). It therefore might appear that entirely new ideas are needed. Indeed, both papers above pose developing similar bounds with respect to convex-elicitable properties as a major open question.

Using our d-flats framework, we resolve both open questions with new lower bounds in both settings. Our framework clarifies the relationship between d-identifiable properties and properties elicited by d-dimensional convex losses: the level sets of the former are d-flats by definition, while the level sets of the latter are *unions* of d-flats by Theorem 1. A careful examination of the arguments of Frongillo and Kash [19, 20] and Dearborn and Frongillo [10] reveals that they largely rely on the containment of d-flats in level sets, rather than the full structure of identifiable properties. As such, although quite subtle in the case of risk measures, the general structure of these previous proofs go through for convex-elicitable properties: since no d-flat could be contained in a particular level set, no union of d-flats could be either. Our lower bounds therefore match both of these papers, though we conjecture that our convex consistency bounds could be tightened in some cases.

4.1 Risk measures (Q4)

The problem of estimating a risk or uncertainty measure of Y is of central importance in financial regulation [1, 6, 14] and robust engineering design [4, 35, 38]. Risk measures include the upper confidence bound $\mathbb{E}[Y] + \lambda\sqrt{\mathrm{Var}[Y]}$, or the conditional value at risk (CVaR) defined below in eq. (3), in either conditional or unconditional contexts. Uncertainty measures include the variance, entropy, or norm of the distribution of Y. Risk and uncertainty measures are typically not elicitable, so this problem falls under Quadrant 4. Frongillo and Kash [19, 20] give prediction dimension lower bounds for a broad class of risk and uncertainty measures, namely Bayes risks. As stated above, these bounds are with respect to identifiable properties, and bounds for convex surrogates are left as a major open question.

We resolve this open question using our d-flats framework, giving a matching result for convex-elicitable properties (Theorem 2). First we recall the definition of the Bayes risk.

Definition 8. *Given loss function $L : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ for some report set $\mathcal{R}$, the* Bayes risk *of L is defined as $\underline{L}(p) := \inf_{r\in\mathcal{R}} \mathbb{E}_p L(r, Y)$.*

Condition 1. *For some $r \in \operatorname{range}\Gamma$, the level set $\Gamma_r = \ker_{\mathcal{P}} V$ is a d-flat presented by some $V : \mathcal{Y} \to \mathbb{R}^d$ such that $0 \in \operatorname{int}\{\mathbb{E}_p V : p \in \mathcal{P}\}$.*

Theorem 2. *Let $\mathcal{P}$ be a convex set of Lebesgue densities supported on the same set for all $p \in \mathcal{P}$. Let $\Gamma : \mathcal{P} \to \mathbb{R}^d$ satisfy Condition 1 for some $r \in \mathbb{R}^d$. Let $L \in \mathcal{L}^{\mathrm{cvx}}$ elicit Γ such that $\underline{L}$ is non-constant on Γ_r. Then $\mathrm{cons}_{\mathrm{cvx}}(\underline{L}) \geq \mathrm{elic}_{\mathrm{cvx}}(\underline{L}) \geq d + 1$.*

To illustrate the theorem, we briefly apply it to one of the most prominent financial risk measures, the conditional value at risk (CVaR). Several other applications from Frongillo and Kash [19, 20], such as other risk measures, entropy, and norms, follow similarly. The authors observe that CVaR can be expressed as a Bayes risk; for $0 < \alpha < 1$, we may define

$$\mathrm{CVaR}_\alpha(p) = \inf_{r\in\mathbb{R}} \mathbb{E}_p \left\{ \tfrac{1}{\alpha}(r - Y)\mathbb{1}_{r\geq Y} - r \right\} \ , \tag{3}$$

which is the Bayes risk of the transformed pinball loss $L_\alpha(r,y) = \frac{1}{\alpha}(r-y)\mathbb{1}_{r\geq y} - r$. In turn, L_α elicits the α-quantile, the quantity $q_\alpha(p)$ such that $\Pr_p[Y \geq q_\alpha(p)] = \alpha$. Following Frongillo and Kash [20], we will restrict to the set $\mathcal{P}_q$ of probability measures over $\mathbb{R}$ with connected support and whose CDFs are strictly increasing on their support, so that q_α is single-valued. Under mild assumptions, we find that there is no consistent real-valued convex surrogate for CVaR_α.

Corollary 4. *Let $\mathcal{P}$ be a convex set of continuous Lebesgue densities on $\mathcal{Y} = \mathbb{R}$ with all $p \in \mathcal{P}$ having support on the same interval. If we have $p_1, p_2, p_3, p_2' \in \mathcal{P}$ with $q_\alpha(p_1) < q_\alpha(p_2) < q_\alpha(p_3)$ and $\mathrm{CVaR}_\alpha(p_2) \neq \mathrm{CVaR}_\alpha(p_2')$, then $\mathrm{cons}_{\mathrm{cvx}}(\mathrm{CVaR}_\alpha) \geq \mathrm{elic}_{\mathrm{cvx}}(\mathrm{CVaR}_\alpha) \geq 2$.*

As first shown by Fissler et al. [15], the pair $(\mathrm{CVaR}_\alpha, q_\alpha)$ is jointly identifiable and elicitable, but not by any convex loss [13, Prop. 4.2.31].We conjecture the stronger statement $\mathrm{elic}_{\mathrm{cvx}}(\mathrm{CVaR}_\alpha) \geq 3$, which if true would constitute an interesting gap between elicitation complexity for identifiable and convex-elicitable properties.

4.2 Mode and modal interval (Q4, Q3)

For finite $|\mathcal{Y}|$, the mode $\gamma_{\mathrm{mode}}(p) = \arg\max_{y\in\mathcal{Y}} p(y)$ is elicited by 0-1 loss. By contrast, for $\mathcal{Y} = \mathbb{R}$, the mode is not elicitable [22], landing it in Quadrant 4. Defining the mode is subtle for general distributions; here let us assume p has a smooth and bounded Lesbegue density f_p, and define the mode the same way, $\gamma_{\mathrm{mode}}(p) = \arg\max_{y\in\mathcal{Y}} f_p(y)$. Dearborn and Frongillo [10] recently showed a strong impossibility result, that the mode has countably infinite elicitation complexity with respect to identifiable properties. In other words, it is as hard to elicit the mode as the full distribution p itself. Complexity with respect to convex-elicitable properties is left as an important open question.

We resolve this question, with a matching infinite lower bound for convex-elicitable properties. In light of our d-flats framework, the result is nearly immediate, as the proof in Dearborn and Frongillo [10] already showed that the level sets of the mode cannot contain any d-flats.

Theorem 3. *The mode has $\mathrm{cons}_{\mathrm{cvx}}(\gamma_{\mathrm{mode}}) = \mathrm{elic}_{\mathrm{cvx}}(\gamma_{\mathrm{mode}}) = \infty$ (countably infinite) with respect to $\mathcal{P}$, the class of probability measures on $\mathcal{Y} = \mathbb{R}$ with a smooth and bounded density and such that γ_{mode} is single-valued.*

Proof. The proof of Dearborn and Frongillo [10, Theorem 1] gives a distribution $p \in \mathcal{P}$ with $\gamma_{\mathrm{mode}}(p) = 0 =: u$. It then introduces an arbitrary identification function $V : \hat{\mathcal{R}} \times \mathcal{Y} \to \mathbb{R}^k$, $k \in \mathbb{N}$, and value $r \in \hat{\mathcal{R}}$ such that $p \in \ker_{\mathcal{P}} V(r,\cdot)$. Letting $F = \ker_{\mathcal{P}} V(r,\cdot)$, we therefore have an arbitrary k-flat containing p. The proof then proceeds to construct some $p' \in F$ with $\gamma_{\mathrm{mode}}(p') \neq u$. Corollary 3 now gives $\mathrm{cons}_{\mathrm{cvx}}(\gamma_{\mathrm{mode}}) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma_{\mathrm{mode}}) \geq k+1$. As k was arbitrary, the result follows. □

A closely related property for any $\beta > 0$ is the (midpoint of the) modal interval of width 2β, given by $\gamma_\beta(p) = \arg\max_{x\in\mathbb{R}} p([x-\beta, x+\beta])$. Interestingly, unlike the mode for $\mathcal{Y} = \mathbb{R}$, the modal interval is elicitable, by the target loss $\ell_\beta(r,y) = \mathbb{1}\{|r-y| > \beta\}$. The problem of estimating the modal interval therefore could be thought of as falling under Quadrant 3.

As observed in Dearborn and Frongillo [10, Corollary 1], the properties γ_{mode} and γ_β coincide with the family of distributions needed in their Theorem 1, meaning the conclusion of Theorem 3 transfers to the modal interval as well.

Corollary 5. *For any $\beta > 0$, the modal interval $\gamma_\beta : \mathcal{P}_\beta \to \mathbb{R}$ has $\mathrm{cons}_{\mathrm{cvx}}(\gamma_\beta) = \mathrm{elic}_{\mathrm{cvx}}(\gamma_\beta) = \infty$ (countably infinite) with respect to $\mathcal{P}_\beta$, the class probability measures on $\mathcal{Y} = \mathbb{R}$ with a smooth and bounded density, and such that γ_{mode} and γ_β are single-valued.*

Thus, while γ_β is elicitable, it does not have any consistent finite-dimensional convex surrogate. While this statement may seem counter-intuitive, recall that the mode for finite $|\mathcal{Y}|$ has $\mathrm{cons}_{\mathrm{cvx}}(\gamma_{\mathrm{mode}}) = |\mathcal{Y}| - 1$. Taking the limit as $|\mathcal{Y}| \to \infty$, one may therefore expect an infinite convex consistency dimension for both the mode and modal interval.

5 Conclusions and future work

In this work, we introduce a new tool to generate lower bounds on the convex consistency dimension of general prediction tasks. This tool is simultaneously broader, stronger, and easier to understand than previous results. Its breadth is demonstrated by applying to multiple problem types simultaneously (§ 3), while its strength is demonstrated by proving new bounds on convex consistency dimension (§ 4), and ease is apparent when observing that indirect elicitation is a strictly weaker notion than calibration – the most common proxy for consistency. We then apply our framework to yield new bounds on convex consistency dimension for entropy, risk measures, the mode, and modal intervals.

Several important questions remain open. Particularly for the discrete settings, we would like to know whether one can lift the restriction that surrogates always achieve a minimum; we conjecture positively (see § F.1). The observation that our bounds are as tight as calibration-based bounds, yet we use the weaker condition of indirect elicitation, motivates the study of how much weaker indirect elicitation is than calibration. More broadly, we would like to characterize $\mathrm{cons}_{\mathrm{cvx}}$ and $\mathrm{elic}_{\mathrm{cvx}}$ and develop a general framework for constructing surrogates achieving the best possible prediction dimension.

Acknowledgments and Disclosure of Funding

We thank Adam Bloniarz and Nishant Mehta for several key suggestions and insights. This material is based upon work supported by the National Science Foundation under Grant No. IIS-2045347 and Graduate Research Fellowship No. DGE-1650115.

References

[1] Carlo Acerbi and Balazs Szekely. Back-testing expected shortfall. *Risk*, November 2014.

[2] Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *JMLR Workshop and Conference Proceedings*, volume 40, pages 1–19, 2015. URL `http://www.jmlr.org/proceedings/papers/v40/Agarwal15.pdf`.

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

[4] Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33):3190–3218, 2007. URL `http://www.sciencedirect.com/science/article/pii/S0045782507001259`.

[5] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[6] Sean D. Campbell. *A review of backtesting and backtesting procedures*. Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, 2005. URL `http://www.federalreserve.gov/pubs/FEDS/2005/200521/200521pap.pdf`.

[7] Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957.

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

[9] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.

[10] Krisztina Dearborn and Rafael Frongillo. On the indirect elicitability of the mode and modal interval. *Annals of the Institute of Statistical Mathematics*, 72(5):1095–1108, 2020.

[11] Jianoing Fan and Qiwei Yao. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660, 09 1998. ISSN 0006-3444. doi: 10.1093/biomet/85.3.645. URL `https://doi.org/10.1093/biomet/85.3.645`.

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

[13] Tobias Fissler. *On higher order elicitability and some limit theorems on the Poisson and Wiener space*. PhD thesis, 2017.

[14] Tobias Fissler, Johanna Fasciati-Ziegel, and Tilmann Gneiting. Expected Shortfall is jointly elicitable with Value at Risk-Implications for backtesting. *Risk Magazine*, 2016.

[15] Tobias Fissler, Johanna F Ziegel, and others. Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016.

[16] Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

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

[18] Rafael Frongillo and Ian Kash. Vector-Valued Property Elicitation. In *Proceedings of the 28th Conference on Learning Theory*, pages 1–18, 2015.

[19] Rafael Frongillo and Ian A. Kash. On Elicitation Complexity. In *Advances in Neural Information Processing Systems 29*, 2015.

[20] Rafael Frongillo and Ian A Kash. Elicitation Complexity of Statistical Properties. *Biometrika*, 11 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa093. URL https://doi.org/10.1093/biomet/asaa093.

[21] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[22] C. Heinrich. The mode functional is not elicitable. *Biometrika*, page ast048, 2013. URL http://biomet.oxfordjournals.org/content/early/2013/11/27/biomet.ast048.short.

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

[24] Aleksandr Davidovich Ioffe and Vladimir Mikhailovich Tikhomirov. On minimization of integral functionals. *Functional Analysis and Its Applications*, 3(3):218–227, 1969.

[25] Nicolas S. Lambert. Elicitation and evaluation of statistical forecasts. 2018. URL https://web.stanford.edu/~nlambert/papers/elicitability.pdf.

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

[27] Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

[28] Yi Lin. A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1):73–82, 2004.

[29] David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6150–6160, 2018.

[30] Kent Osband and Stefan Reichelstein. Information-eliciting compensation schemes. *Journal of Public Economics*, 27(1):107–115, June 1985. ISSN 0047-2727. doi: 10.1016/0047-2727(85)90031-3. URL http://www.sciencedirect.com/science/article/pii/0047272785900313.

[31] Kent Harold Osband. *Providing Incentives for Better Cost Forecasting*. University of California, Berkeley, 1985.

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

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

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

[35] R. Tyrrell Rockafellar and Johannes O. Royset. On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety*, 95(5):499–510, 2010. URL http://www.sciencedirect.com/science/article/pii/S0951832010000177.

[36] David Ruppert, M. P. Wand, Ulla Holst, and Ola Hösjer. Local polynomial variance-function estimation. *Technometrics*, 39(3):262–273, 1997. doi: 10.1080/00401706.1997.10485117. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1997.10485117.

[37] L.J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, pages 783–801, 1971.

[38] Muhammad Shahbaz, Zhong-Hua Han, W. P. Song, and M. Nadeem Aizud. Surrogate-based robust design optimization of airfoil using inexpensive Monte Carlo method. In *Applied Sciences and Technology (IBCAST), 2016 13th International Bhurban Conference on*, pages 497–504. IEEE, 2016. URL http://ieeexplore.ieee.org/abstract/document/7429924/.

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

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

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

[42] Robert Williamson and Aditya Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797. PMLR, 2019.

[43] Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, and Liming Chen. Hierarchical classification of emotional speech. *IEEE Transactions on Multimedia*, 37, 2007.

[44] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

A Consistency Implies Indirect Elicitation

In this section, we connect consistency of any surrogate to an indirect elicitation requirement. This will allow us to show indirect elicitation gives state-of-the-art lower bounds on convex consistency dimension.

We start by formalizing consistency in two ways that generalize across our four quadrants. First, given a target loss ℓ, we say L is consistent if optimizing L and applying a link ψ optimizes ℓ (Definition 9). Second, given a target property γ, such as the α-quantile, we say L is consistent if optimizing L implies approaching, in some sense, the correct statistic $\gamma(D_x)$ of the conditional distributions $D_x = \Pr[Y|X = x]$ (Definition 10). We then observe that Definition 9 is subsumed by Definition 10, and use this to show consistency implies L indirectly elicits $\mathrm{prop}_{\mathcal{P}}[\ell]$ or γ respectively.

Condition 2 (Covers)**.** *A set $\mathcal{D} \subseteq \Delta(\mathcal{X} \times \mathcal{Y})$* covers *a convex set $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ if, for all $p \in \mathcal{P}$, there exists $D \in \mathcal{D}$ and $x \in \mathcal{X}$ such that D has a point mass on x and $p = D_x$.*

Definition 9 (Consistent: loss)**.** *A loss $L \in \mathcal{L}$ and link (L, ψ) are $\mathcal{D}$-*consistent *for a set $\mathcal{D}$ of distributions over $\mathcal{X} \times \mathcal{Y}$ with respect to a target loss ℓ if, for all $D \in \mathcal{D}$ and all sequences of measurable hypothesis functions $\{f_m : \mathcal{X} \to \mathcal{R}\}$,*

$$\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y) \implies \mathbb{E}_D \ell((\psi \circ f_m)(X), Y) \to \inf_f \mathbb{E}_D \ell((\psi \circ f)(X), Y) \ .$$

For a given convex set $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$, we simply say (L, ψ) is consistent *if it is $\mathcal{D}$-consistent for some $\mathcal{D}$ covering $\mathcal{P}$.*

Instead of a target loss ℓ, one may want to learn a target property, i.e. a conditional statistic such as the expected value, variance, or entropy. In this case, following the tradition in the statistics literature on conditional estimation [11, 21, 36], we formalize consistency as converging to the correct conditional estimates of the property. Convergence is measured by functions $\mu(r, p)$ that formalize how close r is to "correct" for conditional distribution p. In particular we should have $\mu(r, p) = 0 \iff r \in \gamma(p)$.

Definition 10 (Consistent: property)**.** *Suppose we are given a loss $L \in \mathcal{L}$, link function $\psi : \mathbb{R}^d \to \mathcal{R}$, and property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$. Moreover, let $\mu : \mathcal{R} \times \mathcal{P} \to \mathbb{R}_+$ be any function satisfying $\mu(r, p) = 0 \iff r \in \gamma(p)$. We say (L, ψ) is $(\mu, \mathcal{D})$-*consistent with respect to γ *if, for all $D \in \mathcal{D}$ and sequences of measurable functions $\{f_m : \mathcal{X} \to \mathcal{R}\}$,*

$$\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y) \implies \mathbb{E}_X \mu(\psi \circ f_m(X), D_X) \to 0 \ . \tag{4}$$

*We simply say (L, ψ) is μ-*consistent *if it is $(\mu, \mathcal{D})$-consistent for some $\mathcal{D}$ covering $\mathcal{P}$. Additionally, we say (L, ψ) is* consistent *if there is a μ such that (L, ψ) is μ-consistent.*

Typical definitions of consistency require $\mathcal{D}$ to be the set of all distributions over $\mathcal{X} \times \mathcal{Y}$, while our conditions are much weaker. As the main focus of this paper is lower bounds on the prediction dimension, i.e., showing that surrogates of a certain prediction dimension cannot exist, these weaker conditions translate to stronger impossibility statements.

Given a target loss ℓ, we can define a statistic γ, the property it elicits. Intuitively, consistency of a surrogate L with respect to ℓ and γ are equivalent, i.e. in both cases estimates should converge to values that minimize ℓ-loss. We formalize this by letting μ be the ℓ-regret, $R_\ell := \mathbb{E}_p \ell(r, Y) - \min_{r'} \mathbb{E}_p \ell(r', Y)$, yielding Lemma 1.

Lemma 1. *Let a convex $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ be given. Given a surrogate loss $L \in \mathcal{L}$, link ψ, and target loss ℓ, set $\mu(r, p) := \mathbb{E}_p \ell(r, Y) - \min_{r'} \mathbb{E}_p \ell(r', Y)$ as the excess risk of ℓ, R_ℓ. Then there is a $\mathcal{D}$ covering $\mathcal{P}$ such that (L, ψ) is $\mathcal{D}$-consistent with respect to ℓ if and only if (L, ψ) is $(\mu, \mathcal{D})$-consistent with respect to $\gamma := \mathrm{prop}_{\mathcal{P}}[\ell]$.*

Proof. First, observe that $\mu(r, p) = 0 \iff \mathbb{E}_p \ell(r, Y) = \inf_{r' \in \mathcal{R}} \mathbb{E}_p \ell(r', Y) \iff r \in \gamma(p)$. Now suppose (L, ψ) are consistent with respect to ℓ, and take any sequence $\{f_m\}$ of measurable hypotheses. Rewriting the right-hand side of Definition 9,

$$\begin{aligned} &\mathbb{E}_D \ell(\psi \circ f_m(X), Y) \to \inf\nolimits_f \mathbb{E}_D \ell(\psi \circ f(X), Y) && (5)\\ \iff\ &\mathbb{E}_X R_\ell(\psi \circ f_m(X), D_X) \to 0 \\ \iff\ &\mathbb{E}_X \mu(\psi \circ f_m(X), D_X) \to 0 \ . && (6) \end{aligned}$$

Therefore, $\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y)$ implies (5) if and only if it implies (6). Observe that the assumptions on $\mathcal{L}$ allow us to apply the Fubini-Tonelli Theorem [16, Theorem 2.37], which yields the equivalence of eq. 5 to the next line. $\square$

Because each target loss in $\mathcal{L}$ elicits some property, but not all target properties can be elicited by a loss (e.g. the variance), consistency with respect to a property is the strictly broader notion. In a loose sense, Proposition 1 lets us translate problems about target losses to be about the properties these losses elicit. This points to indirect elicitation as a natural necessary condition for consistency, as formalized in Proposition 1.

Proposition 1. *For a surrogate $L \in \mathcal{L}$, if the pair (L, ψ) is consistent with respect to a property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ or a loss ℓ eliciting γ, then (L, ψ) indirectly elicits γ.*

Proof. By Lemma 1, it suffices to show the result for consistency with respect to a property γ, setting $\gamma := \text{prop}_{\mathcal{P}}[\ell]$ if ℓ is given instead. We show the contrapositive; suppose (L, ψ) does not indirectly elicit γ, meaning we have some $p \in \mathcal{P}$ so that $u \in \Gamma(p)$ but $\psi(u) \notin \gamma(p)$, where $\Gamma := \text{prop}_{\mathcal{P}}[L]$. Observe that we use the fact $\Gamma(p) \neq \emptyset$. By definition, if we had consistency, there must be some distribution D on $\mathcal{X} \times \mathcal{Y}$ with a point mass on some $x \in \mathcal{X}$ and $D_x = p$. Consider a constant sequence $\{f_m\}$ with $f_m = f'$ such that $f'(x) = u$, so that $\mathbb{E}_D L(f_m(X), Y) = \mathbb{E}_{D_x} L(f_m(x), Y) = \mathbb{E}_p L(u, Y)$. Since $u \in \Gamma(p)$, we have $\mathbb{E}_p L(u, Y) = \inf_f \mathbb{E}_{D_x} L(f(x), Y) = \inf_f \mathbb{E}_D L(f(X), Y)$. In particular, we have $\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y)$. However, we have $\mathbb{E}_X \mu(\psi \circ f_m(X), D_X) = \mu(f_m(x), p) = \mu(\psi(u), p) \neq 0$, since $\psi(u) \notin \gamma(p)$. Therefore (L, ψ) is not consistent with respect to γ (Definition 10). $\square$

This result allows us to state elicitation complexity as a lower bound for convex consistency dimension.

Corollary 1. *Given a property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ or loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ eliciting γ, we have $\text{elic}_{\text{cvx}}(\gamma) \leq \text{cons}_{\text{cvx}}(\gamma) = \text{cons}_{\text{cvx}}(\ell)$.*

B Implications of Convex Indirect Elicitation Bounds

We start by fully proving Theorem 1.

Theorem 1. *Let $\Gamma : \mathcal{P} \rightrightarrows \mathbb{R}^d$ be (directly) elicited by $L \in \mathcal{L}_d^{\text{cvx}}$ for some $d \in \mathbb{N}$. Let $\mathcal{Y}$ be either a finite set, or $\mathcal{Y} = \mathbb{R}$, in which case we assume each $p \in \mathcal{P}$ admits a Lebesgue density supported on the same set for all $p \in \mathcal{P}$.[4] For all $u \in \text{range}\,\Gamma$ and $p \in \Gamma_u$, there is some d-flat F such that $p \in F \subseteq \Gamma_u$.*

Proof. As L is convex and elicits Γ, we have $u \in \Gamma(p) \iff \vec{0} \in \partial \mathbb{E}_p L(u, Y)$. We proceed in two cases, depending on $|\mathcal{Y}|$.

Finite $\mathcal{Y}$: If $\mathcal{Y}$ is finite, this is additionally equivalent to $\vec{0} \in \oplus_y p_y \partial L(u, y)$, where $\oplus$ denotes the Minkowski sum [23, Theorem 4.1.1].[5] Expanding, we have $\oplus_y p_y \partial L(u, y) = \{\sum_{y \in \mathcal{Y}} p_y x_y \mid x_y \in \partial L(u, y)\ \forall y \in \mathcal{Y}\}$, and thus $Wp = \sum_y p_y x_y = \vec{0}$ where $W = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$; cf. [33, $\mathbf{A}^m$ in Theorem 16]. Let $V_{u,p} : \mathcal{Y} \to \mathbb{R}^d, y \mapsto W_y$ be the function encoding the columns of W. Observe that $\mathbb{E}_p V_{u,p} = \vec{0}$.

$\mathcal{Y} = \mathbb{R}$: Any $L \in \mathcal{L}_d^{\text{cvx}}$ satisfies the assumptions of [24], so we may interchange subdifferentiation and expectation. Specifically, letting $\mathcal{V}_{u,p} = \{V : \mathcal{Y} \to \mathbb{R}^d \mid V \text{ measurable}, V(y) \in \partial L(u, y)\ p\text{-a.s.}\}$, we have $\partial \mathbb{E}_p L(u, Y) = \{\int V(y) dp(y) \mid V \in \mathcal{V}_{u,p}\}$. As $\vec{0} \in \partial \mathbb{E}_p L(u, Y)$, in particular, there is some $V_{u,p} \in \mathcal{V}_{u,p}$ such that $\mathbb{E}_p V_{u,p} = 0$. For any $q \in \mathcal{P}$, as by assumption q is supported on the same set as p, we have $V_{u,p}(y) \in \partial L(u, y)$ q-a.s., so that $V_{u,p} \in \mathcal{V}_{u,q}$. Thus, $\mathbb{E}_q V_{u,p} = 0$ implies $0 \in \partial \mathbb{E}_q L(u, Y)$ by the above.

[4] This assumption is largely for technical convenience, to ensure that $\mathcal{V}_{u,p}$ does not depend on p. Any such assumption would suffice, and we suspect even that condition can be relaxed.

[5] ∂ represents the subdifferential $\partial f(x) = \{z : f(x') - f(x) \geq \langle z, x' - x \rangle\ \forall x'\}$.

In both cases, we take the flat $F := \ker_{\mathcal{P}} V_{u,p}$, and have $p \in F$ by construction. To see $F \subseteq \Gamma_u$, from the chain of equivalences above, we have for any $q \in \mathcal{P}$ that $q \in \ker_{\mathcal{P}} V_{u,p} \implies \vec{0} \in \partial \mathbb{E}_q L(u, Y) \implies u \in \Gamma(q) \implies q \in \Gamma_u$. $\square$

In order to apply Theorem 1 to various properties, we need the following lemmas about separating hyperplanes.

A hyperplane weakly separates two sets if its two closed halfspaces respectively contain the two sets.

Lemma 2. *If $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ is an elicitable property, then for any pair of predictions $r, r' \in \mathcal{R}$ where $\gamma_r \neq \gamma_{r'}$, there is a hyperplane $H = \{x \in \mathbb{R}^{\mathcal{Y}} : v \cdot x = 0\}$, for some $v \in \mathbb{R}^{\mathcal{Y}}$, that weakly separates γ_r and $\gamma_{r'}$ and has $\gamma_r \cap H = \gamma_{r'} \cap H = \gamma_r \cap \gamma_{r'}$.*

Proof. Let ℓ elicit γ. Let $v = \ell(r, \cdot) - \ell(r', \cdot)$, interpreted as a nonzero vector in $\mathbb{R}^{\mathcal{Y}}$. Let $H = \{q : v \cdot q = 0\}$. If $v \cdot q < 0$, then r' cannot be optimal, so $q \notin \gamma_{r'}$. So $\gamma_{r'} \subseteq \{q : v \cdot q \geq 0\}$. Symmetrically, $\gamma_r \subseteq \{q : v \cdot q \leq 0\}$. This is weak separation, and it immediately implies that $\gamma_r \cap \gamma_{r'} \subseteq H$. Finally, if and only if $v \cdot q = 0$, i.e. $q \in H$, by definition the expected losses of both reports are the same. So $q \in \gamma_r \cap H \iff q \in \gamma_{r'} \cap H$. This gives $\gamma_r \cap H = \gamma_{r'} \cap H = \gamma_r \cap \gamma_{r'} \cap H = \gamma_r \cap \gamma_{r'}$. $\square$

Lemma 3. *Suppose we are given an elicitable property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$, where $\mathcal{Y}$ is finite, and distribution $p \in \mathrm{relint}(\mathcal{P})$ such that $p \in \gamma_r \cap \gamma_{r'}$ for $r, r' \in \mathcal{R}$. Then for any flat F containing p, $F \subseteq \gamma_r \iff F \subseteq \gamma_{r'}$.*

Proof. If $\gamma_r = \gamma_{r'}$, we are done. Otherwise, Lemma 2 gives a hyperplane $H = \{x \in \mathbb{R}^{\mathcal{Y}} : v \cdot x = 0\}$ and a guarantee that $\gamma_r \subseteq \{q \in \Delta_{\mathcal{Y}} : v \cdot q \leq 0\}$, while $\gamma_{r'} \subseteq \{q \in \Delta_{\mathcal{Y}} : v \cdot q \geq 0\}$, and finally $\gamma_r \cap \gamma_{r'} \subseteq H$.

Suppose $F \subseteq \gamma_r$; we wish to show $F \subseteq \gamma_{r'}$. Let $q \in F$. By Lemma 7(i), we have $p \in \mathrm{relint}(F)$, so there exists $\epsilon > 0$ so that $q' = p - \epsilon(q - p) \in F$.

Now, suppose for contradiction that $q \notin \gamma_{r'}$. Then $v \cdot q < 0$: containment in γ_r gives $v \cdot q \leq 0$, and if $v \cdot q = 0$ then $q \in \gamma_r \cap H \implies q \in \gamma_{r'}$, a contradiction. But, noting that $p \in H$, we have $v \cdot q' = -\epsilon(v \cdot q) > 0$, so q' is not in γ_r. This contradicts the assumption $F \subseteq \gamma_r$. Therefore, we must have $q \in \gamma_{r'}$, so we have shown $F \subseteq \gamma_{r'}$. Because r and r' were completely symmetric, this completes the proof. $\square$

Now we can understand the application of Theorem 1.

Corollary 2. *Let target property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ and $d \in \mathbb{N}$ be given. Let $\mathcal{Y}$ be either a finite set, or $\mathcal{Y} = \mathbb{R}$, in which case we assume each $p \in \mathcal{P}$ admits a Lebesgue density supported on the same set for all $p \in \mathcal{P}$. Let $p \in \mathcal{P}$ with $|\gamma(p)| = 1$, and take $\gamma(p) = \{r\}$. If there is no d-flat F with $p \in F \subseteq \gamma_r$, then $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$.*

Proof. Let (L, ψ) indirectly elicit γ, where $L \in \mathcal{L}_d^{\mathrm{cvx}}$, and let $\Gamma = \mathrm{prop}_{\mathcal{P}}[L]$. As Γ is non-empty, there is some $u \in \Gamma(p)$. Since γ is single-valued at p, we have $r = \psi(u)$; by Theorem 1, we know there is a d-flat $F = \ker_{\mathcal{P}} V_{u,p}$ so that $p \in F \subseteq \Gamma_u$. By definition of indirect elicitation, we additionally have $\Gamma_u \subseteq \gamma_r$. Thus, we have $p \in F \subseteq \gamma_r$. If no flat F satisfies the above conditions, then no $L \in \mathcal{L}_d^{\mathrm{cvx}}$ indirectly elicits γ, so $\mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$, and recall $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma)$ by Corollary 1. $\square$

Corollary 3. *Let an elicitable target property $\gamma : \mathcal{P} \rightrightarrows \mathcal{R}$ be given, where $\mathcal{P} \subseteq \Delta_{\mathcal{Y}}$ is defined over a finite set of outcomes $\mathcal{Y}$, and let $d \in \mathbb{N}$. Let $p \in \mathrm{relint}(\mathcal{P})$. If there is no d-flat F with $p \in F \subseteq \gamma_r$, then $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$.*

Proof. Let (L, ψ) indirectly elicit γ and the convex function L and elicit Γ. As Γ is non-empty, there is some $u \in \Gamma(p)$, and suppose $r' = \psi(u)$. Take $F \subseteq \Gamma_u$ to be the flat that exists by Theorem 1. If $r = r'$, then $p \in F \subseteq \Gamma_u \subseteq \gamma_r$ by indirect elicitation. Otherwise, by Lemma 3, for elicitable properties with $p \in \gamma_r \cap \gamma_{r'}$, we observe $p \in F \subseteq \gamma_r \iff p \in F \subseteq \gamma_{r'}$.

As above, if no flat F satisfies the above conditions, then no $L \in \mathcal{L}_d^{\mathrm{cvx}}$ indirectly elicits γ, so $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \mathrm{elic}_{\mathrm{cvx}}(\gamma) \geq d + 1$, recalling Corollary 1 for the first inequality. $\square$

C Definitions of Calibration

When given a discrete target loss, such as for classification-like problems, direct empirical risk minimization is typically NP-hard, forcing one to find a more tractable surrogate. To ensure consistency, the literature has embraced the notion of *calibration* from Steinwart and Christmann [40, Chapter 3], which aligns with the definition in Tewari and Bartlett [41] for multiclass classification, and its generalizations to arbitrary discrete target losses [2, 33]. Calibration is more tractable and weaker than consistency, yet the two are equivalent under suitable assumptions [33, 41], notably in Quadrant 1. Intuitively, calibration says one cannot achieve the optimal surrogate loss while linking to a suboptimal target prediction.

Definition 11 (Calibrated: Quadrant 1)**.** *Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ be a discrete target loss. A surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ and link $\psi : \mathbb{R}^d \to \mathcal{R}$ pair (L, ψ) is* $\mathcal{P}$-calibrated with respect to ℓ *if*

$$\forall p \in \mathcal{P} : \inf_{u \in \mathbb{R}^d : \psi(u) \notin \arg\min_r \mathbb{E}_p \ell(r,Y)} \mathbb{E}_p L(u, Y) > \inf_{u \in \mathbb{R}^d} \mathbb{E}_p L(u, Y) \ . \quad (7)$$

We simply say L is calibrated if $\mathcal{P} = \Delta_{\mathcal{Y}}$.

Many works characterize calibrated surrogates for specific discrete target losses [3, 28, 41, 44], including the canonical 0-1 loss for binary and multiclass classification. We give another definition of calibration which is a special case of calibration via Steinwart and Christmann [40], and show it is equivalent to Definition 11 in discrete prediction settings, but can be applied in continuous estimation settings as well. We use this more general definition of calibration when proving statements about the relationship between consistency, calibration, and indirect elicitation.

The close connection between indirect elicitation and consistency was first explored by Agarwal and Agarwal [2]. In particular, calibration of $L \in \mathcal{L}$ with respect to ℓ implies indirect elicitation quite directly: take $u \in \mathbb{R}^d$ and $p \in \Gamma_u$, implying $u \in \Gamma(p)$. From eq. (1), $\mathbb{E}_p L(u, Y) = \inf_{u' \in \mathbb{R}^d} \mathbb{E}_p L(u', Y)$, so we must have $\psi(u) \in \gamma(p)$ from eq. (7), as desired.

For a given $p \in \mathcal{P}$, the (conditional) *regret*, or excess risk, of a loss L is given by $R_L(u, p) := \mathbb{E}_p L(u, Y) - \inf_{u^*} \mathbb{E}_p L(u^*, Y)$.

Definition 12 (Calibrated: Quadrants 1 and 3)**.** *A loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ is* $\mathcal{P}$-calibrated *with respect to a loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ if there is a link $\psi : \mathbb{R}^d \to \mathcal{R}$ such that, for all distributions $p \in \mathcal{P}$, there exists a function $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$ with ζ continuous at 0^+ and $\zeta(0) = 0$ such that for all $u \in \mathbb{R}^d$, we have*

$$\ell(\psi(u); p) - \underline{\ell}(p) \leq \zeta \left(\mathbb{E}_p L(u, Y) - \underline{L}(p) \right) \ . \quad (8)$$

If $\mathcal{P} = \Delta_{\mathcal{Y}}$, we simply say (L, ψ) is calibrated.

Consider the following four conditions: Suppose we are given $\zeta : \mathbb{R}_+ \to \mathbb{R}_+$.

- A ζ satisfies $\zeta : 0 \mapsto 0$ and is continuous at 0.
- B $\epsilon_m \to 0 \implies \zeta(\epsilon_m) \to 0$.
- C Given $\zeta : \mathbb{R} \to \mathbb{R}_+$, for all $u \in \mathbb{R}^d$, $R_\ell(\psi(u); p) \leq \zeta(R_L(u; p))$.
- D For all $p \in \mathcal{P}$ and sequences $\{u_m\}$ so that $R_L(u_m; p) \to 0$, we have $R_\ell(\psi(u_m); p) \to 0$.

The existence of a function ζ so that $(A \wedge C)$ defines calibration as in Definition 12, and we show $A \iff B$ in Lemma 4. Lemma 5 shows calibration if and only if D, which yields a condition equivalent to calibration without dependence the function ζ.

Proposition 3. *When $\mathcal{R}$ and $\mathcal{Y}$ are finite, a continuous loss and link (L, ψ) are $\mathcal{P}$-calibrated with respect to a target loss ℓ via Definition 12 if and only if they are $\mathcal{P}$-calibrated via Definition 11.*

Proof. $\implies$ We prove the contrapositive; if (L, ψ) is not calibrated with respect to ℓ by Definition 11, then it is not calibrated via Definition 12 either. If (L, ψ) are not calibrated with respect to ℓ by Definition 11, then there is a $p \in \mathcal{P}$ so that $\inf_{u : \psi(u) \not\in \gamma(p)} \mathbb{E}_p L(u, Y) = \inf_u \mathbb{E}_p L(u, Y)$. Thus there is a sequence $\{u_m\}$ so that $\lim_{m \to \infty} \psi(u_m) \not\in \gamma(p)$ and

$\mathbb{E}_p L(u_m, Y) \to \underline{L}(p)$. Now we have $R_L(u_m; p) \to 0$ but $R_\ell(\psi(u_m); p) \not\to 0$, so by Lemma 5, we contradict calibration by Definition 12.

$\Longleftarrow$ Suppose there was a function ζ satisfying the bound in eq. (8) for a fixed distribution $p \in \mathcal{P}$. Observe the bound in eq. (7) can be written as $R_L(u, p) > 0$ for all $p \in \Delta_\mathcal{Y}$ and u such that $\psi(u) \neq \gamma(p)$. By eq. (8), for any sequence $\{u_m\}$ so that $\psi(u_m) \not\to \gamma(p)$, we have must have $\zeta(R_\ell(\psi(u_m), p)) \not\to 0$ as we would otherwise contradict the bound in eq. (8) since $R_\ell(\psi(u), p) \not\to 0$. Therefore $R_L(u_m, p) \not\to 0$; thus, the strict inequality holds. $\square$

The following Lemma shows that conditions A and B are equivalent, so that we can using condition B in lieu of condition A in the proof of Lemma 5

Lemma 4. *A function $\zeta : \mathbb{R} \to \mathbb{R}$ is continuous at 0 and $\zeta(0) = 0$ if and only if the sequence $\{u_m\} \to 0 \implies \zeta(u_m) \to 0$.*

Proof. $\implies$ Suppose we have a sequence $\{u_m\} \to 0$. By continuity, we have $\lim_{u_m \to 0} \zeta(u_m) = \zeta(0) = 0$, so $\zeta(u_m) \to 0$.

$\Longleftarrow$ Suppose $\zeta(0) \neq 0$ but ζ was continuous at 0. The constant sequence $\{u_m\} = 0$ then converges to 0, but as ζ is continuous at 0, we must have $\lim_{m\to\infty} \zeta(u_m) = \zeta(0) \neq 0$, so $\zeta(u_m) \not\to 0$.

Now suppose $\zeta(0) = 0$ but ζ was not continuous at 0. There must be a sequence $\{u_m\} \to 0$ so that $\lim_{m\to\infty} \zeta(u_m) \neq \zeta(0) = 0$, so $\zeta(u_m) \not\to 0$. $\square$

Lemma 5 now gives a condition equivalent to calibration without requiring one to already have a function ζ in mind.

Lemma 5. *A continuous surrogate and link (L, ψ) are $\mathcal{P}$-calibrated (via definition 12) with respect to ℓ if and only if, for all $p \in \mathcal{P}$ and sequences $\{u_m\}$ so that $R_L(u_m; p) \to 0$, we have $R_\ell(\psi(u_m); p) \to 0$.*

Proof. $\implies$ Take a sequence $\{u_m\}$ so that $R_L(u_m; p) \to 0$. Since $\zeta(0) = 0$ and ζ is continuous at 0, we have $\zeta(R_L(u_m; p)) \to 0$. As the bound from Equation (8) is satisfied for all $u \in \mathbb{R}^d$ by assumption, we observe

$$\begin{aligned}
&\forall m,\, 0 \leq R_\ell(\psi(u_m); p) \leq \zeta(R_L(u_m; p)) \\
&\implies 0 \leq \lim_{m\to\infty} R_\ell(\psi(u_m); p) \leq \lim_{m\to\infty} \zeta(R_L(u_m; p)) = 0 \\
&\implies 0 = \lim_{m\to\infty} R_\ell(\psi(u_m); p) \ .
\end{aligned}$$

$\Longleftarrow$ Fix $p \in \mathcal{P}$, and consider $\zeta(c) := \sup_{u : R_L(u,p) \leq c} R_\ell(\psi(u); p)$. We will show $R_L(u_m; p) \to 0 \implies R_\ell(\psi(u_m); p) \to 0$ gives calibration via the function ζ constructed above. With ζ as constructed, we observe that the bound in equation (8) is satisfied for all $u \in \mathbb{R}^d$ and apply Lemma 4 to observe that if there is a sequence $\{\epsilon_m\} \to 0$ so that $\zeta(\epsilon_m) \not\to 0$, it is because $R_L(u_m, p) \not\to 0 \not\Longleftarrow R_\ell(\psi(u_m), p) \to 0$.

Now, we observe that the bound in Equation (8) is satisfied for all $u \in \mathbb{R}^d$ by construction of ζ. Let $S(v) := \{u' \in \mathbb{R}^d : R_L(u'; p) \leq R_L(v, p)\}$. Showing $R_\ell(\psi(u); p) \leq \sup_{u' \in S(u)} R_\ell(\psi(u'); p)$ for all $u \in \mathbb{R}^d$ gives the condition C. As u is in the space over which the supremum is being taken (as $R_L(u; p) \leq R_L(u; p)$), we then have calibration by definition of the supremum.

Now suppose there exists a sequence $\{\epsilon_m\} \to 0$ so that $\zeta(\epsilon_m) \not\to 0$. Consider $S(\epsilon) = \{u \in \mathbb{R}^d : R_L(u, p) \leq \epsilon\}$.

$$\begin{aligned}
\epsilon_1 \leq \epsilon_2 &\implies S(\epsilon_1) \subseteq S(\epsilon_2) \\
&\implies \zeta(\epsilon_1) \leq \zeta(\epsilon_2) \ .
\end{aligned}$$

Now suppose there exists a sequence $\{u_m\}$ so that $R_L(u_m, p) \to 0$. Then for all $\epsilon > 0$, there exists a $m' \in \mathbb{N}$ so that $R_L(u_m, p) < \epsilon$ for all $m \geq m'$. Since this is true for all ϵ, we have $S(\epsilon)$ nonempty for all $\epsilon > 0$, and therefore $\zeta(c)$ is discrete for all $c > 0$. Now if $\zeta(\epsilon_m) \not\to 0$, it must be because $R_\ell(\psi(u_m), p) \not\to 0$ for some sequence converging to zero surrogate regret, and therefore we contradict the statement $R_L(u_m, p) \to 0 \implies R_\ell(\psi(u_m), p) \to 0$.

Moreover, we argue that such a sequence of $\{u_m\}$ with converging surrogate regret always exists by continuity and boundedness from below of the surrogate loss, since we can take the constant sequence at the (attained) infimum. $\square$

C.1 Relating calibration, consistency, and indirect elicitation.

Even with the more general notion of calibration that extends beyond discrete predictions, we still have consistency implying calibration.

Proposition 4. *If a loss and link (L, ψ) are consistent with respect to a loss ℓ, then they are calibrated with respect to ℓ.*

Proof. We show the contrapositive. If (L, ψ) are not calibrated with respect to ℓ, then there is a sequence $\{u_m\}$ such that $R_L(u_m; p) \to 0$ but $R_\ell(\psi(u_m); p) \not\to 0$ via Lemma 5. Suppose $D \sim \mathcal{X} \times \mathcal{Y}$ has only one $x \in \mathcal{X}$ with $Pr_D(X = x) > 0$ so that $p := D_x$ and $\mathbb{E}_D f(X, Y) = \mathbb{E}_p f(x, Y)$. Consider any sequence of functions $\{f_m\} \to f$ with $f_m(x) = u_m$ for all f_m. Now we have $\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_D L(f(X), Y)$, but $\mathbb{E}_D \ell(\psi \circ f(X), Y) \not\to \inf_f \mathbb{E}_D \ell(\psi \circ f(X), Y)$, and therefore (L, ψ) is not consistent with respect to ℓ. $\square$

Moreover, we have calibration implying indirect elicitation.

Lemma 6. *If a surrogate and link (L, ψ) with $L \in \mathcal{L}$ are calibrated with respect to a loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$, then L indirectly elicits the property $\gamma := \mathrm{prop}_{\mathcal{P}}[\ell]$.*

Proof. Let Γ be the unique property directly elicited by L, and fix $p \in \Delta_{\mathcal{Y}}$ with u such that $p \in \Gamma_u$. We know such a u exists since $\Gamma(p) \neq \emptyset$. As $p \in \Gamma_u$, then $\zeta(\mathbb{E}_p L(u, Y) - \underline{L}(p)) = \zeta(0) = 0$, we observe the bound $\ell(\psi(u); p) \leq \underline{\ell}(p)$. We also have $\ell(\psi(u); p) \geq \underline{\ell}(p)$ by definition of $\underline{\ell}$, so we must have $\ell(\psi(u); p) = \underline{\ell}(p) = \ell(\gamma(p); p)$, and therefore, $p \in \gamma_{\psi(u)}$. Thus, we have $\Gamma_u \subseteq \gamma_{\psi(u)}$, so L indirectly elicits γ. $\square$

Combining the two results, we can observe the result of Proposition 1 another way: *through calibration.*

D Quadrant 1: Previous Lower Bounds and Comparisons

The main known technique for lower bounds on surrogate dimensions is given by Ramaswamy and Agarwal [33] for the Quadrant 1 (target loss and discrete predictions). The proof heavily builds around the "limits of sequences" in the definition of calibration. By restricting slightly to the broad class of minimizable losses $\mathcal{L}^{\mathrm{cvx}}$, we show their bound follows relatively directly from Corollary 3. (We conjecture that the minimizability restriction to $\mathcal{L}^{\mathrm{cvx}}$ can be lifted; see § 5.) Ramaswamy and Agarwal [33] construct what they call the subspace of feasible dimensions and give bounds in terms of its dimension.

Definition 13 (Subspace of feasible directions)**.** *The* subspace of feasible directions $\mathcal{S}_{\mathcal{C}}(p)$ *of a convex set* $\mathcal{C} \subseteq \mathbb{R}^n$ *at* $p \in \mathcal{C}$ *is* $\mathcal{S}_{\mathcal{C}}(p) = \{v \in \mathbb{R}^n : \exists \epsilon_0 > 0$ *such that* $p + \epsilon v \in \mathcal{C}\ \forall \epsilon \in (-\epsilon_0, \epsilon_0)\}$.

Ramaswamy and Agarwal [33] gives a lower bound on the dimensionality of all consistent convex surrogates, i.e. $\mathrm{cons}_{\mathrm{cvx}}(\ell) \geq \|p\|_0 - \dim(\mathcal{S}_{\gamma_r}(p)) - 1$ for all p and $r \in \gamma(p)$, particularly in the setting where one is given a discrete prediction problem and target loss over finite outcomes. It turns out that the subspace of feasible directions is essentially a special case of a flat described by Theorem 1. So, by making a slight restriction to the class of minimizable convex surrogates $\mathcal{L}^{\mathrm{cvx}}$, we can derive this lower bound from our general technique in a way that we find shorter and simpler.

Corollary 6 ([33] Theorem 18)**.** *Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ be a discrete loss eliciting $\gamma : \Delta_{\mathcal{Y}} \rightrightarrows \mathcal{R}$ with $\mathcal{Y}$ finite. Then for all $p \in \Delta_{\mathcal{Y}}$ and $r \in \gamma(p)$,*

$$\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \|p\|_0 - \dim(\mathcal{S}_{\gamma_r}(p)) - 1 \ . \tag{9}$$

Sketch. If $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \leq d$, then there is a $L \in \mathcal{L}_d^{\mathrm{cvx}}$ so that L is consistent with respect to γ, and in turn, indirectly elicits γ. Theorem 1 says that there is some d-flat $F = \ker_{\mathcal{P}} V$ such that $p \in F \subseteq \gamma_r$. In particular, if $p \in \mathrm{relint}(\Delta_{\mathcal{Y}})$, we can see $\dim(F) = \dim(\mathcal{S}_{\gamma_r}(p))$. Since $\mathrm{affhull}(\Delta_{\mathcal{Y}})$ has dimension $|\mathcal{Y}| - 1 = \|p\|_0 - 1$, by rank-nullity and $\mathrm{rank}(V) \leq d$ (more precisely, the corresponding linear map $q \mapsto \mathbb{E}_q V$) we have $d \geq \|p\|_0 - 1 - \dim(\mathcal{S}_{\gamma_r}(p))$.

When $p \notin \mathrm{relint}(\Delta_{\mathcal{Y}})$, we can project down to the subsimplex on the support of p, again of dimension $\|p\|_0 - 1$, and modify L and ℓ accordingly. Now p is in the relative interior of this subsimplex, so the above gives $\mathrm{cons}_{\mathrm{cvx}}(\gamma) \geq \|p\|_0 - 1 - \dim(\mathcal{S}_{\gamma_r}(p))$, where now $\mathcal{S}$ is relative to $\mathbb{R}^{\mathrm{supp}(p)}$. Finally, the feasible subspace dimension in the projected space is the same as in the original space because of p's location on a face of $\Delta_{\mathcal{Y}}$. $\square$

There are some cases where the bound provided by Corollaries 2 and 3 is strictly tighter than the bound provided by feasible subspace dimension in Corollary 6. For an example of how Corollary 2 applies to a discrete property for which there is no target loss – a non-elicitable property, i.e. Quadrant 2, which is not considered by Ramaswamy et al. [34] – we refer the reader to Figure 3.

Example: Abstain Recall the abstain target loss $\ell^{abs}(r, y) := \mathbf{I}\{r \notin \{y, \bot\}\} + (1/2)\mathbf{I}\{r = \bot\}$, we can consider the *abstain property* it elicits, where one predicts the most likely outcome y if $Pr[Y = y|x] \geq 1/2$ and "abstain" by predicting $\bot$ otherwise. Ramaswamy and Agarwal [33] present a convex surrogate for the abstain loss that takes as input a prediction whose dimension is logarithmic in the number of outcomes, yielding new upper bounds on $\mathrm{cons}_{\mathrm{cvx}}(\ell^{abs})$ which are an exponential improvement over previous results, e.g., [9].

To lower bound the dimension of convex surrogates, we can consider two different distributions; in the first, our bound yields a strict gap over the feasible subspace dimension bound, and in the second, the bounds are equal. First, we choose $p = \bullet$ to be the uniform distribution (see Figure 4). In this case, the bound by feasible subspace dimension yields $\mathrm{cons}_{\mathrm{cvx}}(\ell^{abs}) \geq 3 - 2 - 1 = 0$, as the feasible subspace dimension is 2 since we are on the relative interior of the level set and simplex, as shown in Figure 4 (L).

However, consider any 1-flat containing $\bullet$. When intersected with the simplex, one can see that any line (a 1-flat, since $\bullet \in \mathrm{relint}(\Delta_{\mathcal{Y}})$) in the simplex through $\bullet$ also leaves the cell $\gamma_\bot$, which contains p. See Figure 4 (R) for intuition; a 1-flat through $p \in \mathrm{relint}(\Delta_{\mathcal{Y}})$ would be a line in such a figure. Therefore, we have no 1-flat containing p staying in $\gamma_\bot$, so we obtain a better lower bound, $\mathrm{cons}_{\mathrm{cvx}}(\ell^{abs}) \geq 2$. Combining this with the upper bounds given by [34], we observe the bound $\mathrm{cons}_{\mathrm{cvx}}(\ell^{abs}) = 2$ is tight in this case with $|\mathcal{Y}| = 3$.

Our bounds sometimes match those of [33]; consider the distribution $\star = (1/4, 1/4, 1/2)$, shown in Figure 4. The feasible subspace dimension of both $\gamma_\bot$ and γ_3 at $\star$ is 1, since one only moves toward the distributions $(0, 1/2, 1/2)$ and $(1/2, 0, 1/2)$ without leaving the level sets, and the three points are collinear in $\mathrm{affhull}(\Delta_{\mathcal{Y}})$, suggesting $\mathcal{S}_{\gamma_\bot}(q) = 1$. This yields $\mathrm{cons}_{\mathrm{cvx}}(\ell^{abs}) \geq 3 - 1 - 1 = 1$. The same line segment defines a flat contained in both $\gamma_\bot$ and γ_3, so we have $\mathrm{cons}_{\mathrm{cvx}}(\ell^{abs}) \geq 1$ by Corollary 3, matching the feasible subspace dimension bound.

Bounds using d-flats appear to work well at distributions where previous bounds via feasible subspace dimension would have been vacuous. In essence, flats allow us a "global" view of the property we are eliciting, while the feasible subspace method only permits a "local" look at the property, so we find our method works better for distributions in $\mathrm{relint}(\Delta_{\mathcal{Y}})$.

D.1 Reconstructing Ramaswamy and Agarwal [33, Thm. 16]

Lemma 7. *Let the d-flat $F \subseteq \mathcal{P}$ (defined over finite $\mathcal{Y}$) contain some $p \in \mathrm{relint}(\mathcal{P})$. Then*

(i) $p \in \mathrm{relint}(F)$;

(ii) $\dim(\mathcal{S}_F(p)) \geq \dim(\mathrm{affhull}(\mathcal{P})) - d$.

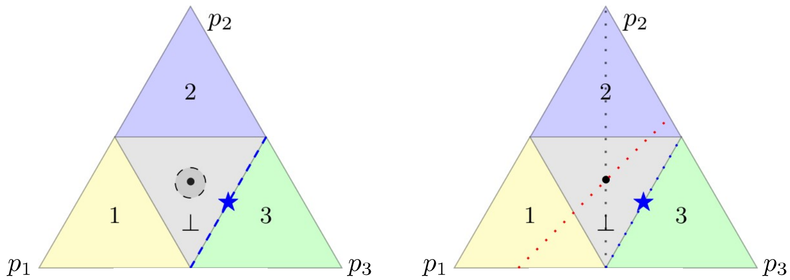

Figure 4: (Left) Feasible subspace dimension $\mathcal{S}_{\gamma_\perp}(\bullet) = 2$ and $\mathcal{S}_{\gamma_\perp}(\star) = 1$, giving the bound $\text{cons}_{\text{cvx}}(\ell^{abs}) \geq 3 - 1 - 1 = 1$. (Right) No 1-flat through $\bullet$ (a line since $\bullet \in \text{relint}(\Delta_\mathcal{Y})$) stays fully contained in $\gamma_\perp$, so $\text{cons}_{\text{cvx}}(\ell^{abs}) \geq 2$.

Proof. As F is a d-flat, we have some $W : \mathcal{Y} \to \mathbb{R}^d$ such that $F = \ker_\mathcal{P} W$. Throughout, given a point (typically a distribution) p and convex set P, we define $P_p := P - \{p\}$. Define $T_W : \text{span}(\mathcal{P}_p) \to \mathbb{R}^d, v \mapsto \mathbb{E}_v W$.

(i) Since $p \in \text{relint}(\mathcal{P})$, for all $q \in \mathcal{P}$, there is some small enough $\epsilon > 0$ so that for all $\alpha \in (-\epsilon, \epsilon)$, the point $q_\alpha := p - \alpha(q - p)$ is still in $\mathcal{P}$. In particular, for $q \in F$, we claim $q_\alpha \in F$. As $p, q \in F$, we have $\mathbb{E}_p W = \mathbb{E}_q W = \vec{0}$. By linearity of expectation, we then have $\mathbb{E}_{q_\alpha} W = \vec{0}$. This implies $q_\alpha \in F$, and therefore $p \in \text{relint}(F)$.

(ii) We first show $\text{span}(F_p) = \mathcal{S}_F(p)$. First, take $v \in \mathcal{S}_F(p)$, and take ϵ_0 as in the definition. For $\epsilon = \epsilon_0/2$, we then have $p + \epsilon v \in F \implies \epsilon v \in F_p$, and therefore, $v \in \text{span}(F_p)$. Now take $v \in \text{span}(F_p)$. Since $p \in \text{relint}(F)$ (i), we have $\vec{0} \in \text{relint}(F_p)$. Therefore there is an $\epsilon_0 > 0$ so that $\epsilon v \in F_p$ for all $\epsilon \in (-\epsilon_0, \epsilon_0)$ by convexity of F. Therefore, $v \in \mathcal{S}_F(p)$, and we observe $\mathcal{S}_F(p) = \text{span}(F_p)$.

We now show $\mathcal{S}_F(p) = \ker(T_W)$. Observe that $\mathcal{S}_F(p) \subseteq \ker(T_W)$ follows trivially from the definitions of the two functions. Now let $v \in \ker(T_W)$, and $v' \in F_p$. This means $\mathbb{E}_v W = \vec{0}$, so it suffices to show $v = cv' \in F_p$, thus showing $v \in \mathcal{S}_F(p)$. Since $p \in \text{relint}(\mathcal{P})$, we must have $\vec{0} \in \text{relint}(F_p)$, so we know there is some small enough $\epsilon > 0$ so that $-\alpha v' \in F_p$ for $\alpha \in (-\epsilon, \epsilon)$. Take $c = -\alpha$, and we conclude $v \in \mathcal{S}_F(p)$. Therefore, $\ker(T_W) = \mathcal{S}_F(p)$.

We finally want to show $\dim(\text{affhull}(\mathcal{P})) = \dim(\text{span}(\mathcal{P}_p))$. Consider that any $q \in \text{span}(\mathcal{P}_p)$ can be written as a scalar multiple of an element of $\mathcal{P}_p$, which can be written as a convex combination of elements of the minimal basis $\mathcal{P}_p$. In particular, since $\vec{0} \in \mathcal{P}_p$, it can be written as an affine combination of elements of the basis, so $\dim(\text{affhull}(\mathcal{P})) \geq \dim(\text{span}(\mathcal{P}_p))$. We also have $\text{affhull}(\mathcal{P}) - \{p\} \subseteq \text{span}(\mathcal{P}_p)$, so $\dim(\text{affhull}(\mathcal{P})) = \dim(\text{affhull}(\mathcal{P}) - \{p\}) \leq \text{span}(\mathcal{P}_p)$. Therefore, $\dim(\text{affhull}(\mathcal{P})) = \dim(\text{span}(\mathcal{P}_p))$.

As $\mathcal{Y}$ is a finite set, $\text{span}(\mathcal{P}_p)$ is a finite-dimensional vector space. The rank-nullity theorem states $\dim(\text{im}(T_W)) + \dim(\ker(T_W)) = \dim(\text{span}(\mathcal{P}_p)) = \dim(\text{affhull}(\mathcal{P}))$. As $\dim(\text{im}(T_W)) \leq d$, and we have shown above that $\mathcal{S}_F(p) = \text{span}(F_p) = \ker(T_W)$, the conclusion follows. □

Corollary 6 ([33] Theorem 18). *Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ be a discrete loss eliciting $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ with $\mathcal{Y}$ finite. Then for all $p \in \Delta_\mathcal{Y}$ and $r \in \gamma(p)$,*

$$\text{cons}_{\text{cvx}}(\gamma) \geq \|p\|_0 - \dim(\mathcal{S}_{\gamma_r}(p)) - 1 \ . \tag{9}$$

Proof. Let $L \in \mathcal{L}_d^{\text{cvx}}$ be a calibrated surrogate for ℓ, and let $\Gamma := \text{prop}_{\Delta_\mathcal{Y}}[L]$. Consider $\mathcal{Y}' := \{y \in \mathcal{Y} : p_y > 0\}$ and $p' = (p_y)_{y \in \mathcal{Y}'} \in \Delta_{\mathcal{Y}'}$. Take $L' := L|_{\mathcal{Y}'}$ and $\ell' := \ell|_{\mathcal{Y}'}$. Define $h : \mathbb{R}^{\mathcal{Y}'} \to \mathbb{R}_\mathcal{Y}$ such that $h(q') = q$ such that $q_y = q'_y$ for $y \in \mathcal{Y}'$ and $q_y = 0$ otherwise. Take $\Gamma' = \Gamma \circ h$, $\gamma' = \gamma \circ h$.

We wish to first show L' indirectly elicits γ'. Since L indirectly elicits γ, we have a link ψ such that for all $u \in \mathbb{R}^d$, $\Gamma_u \subseteq \gamma_{\psi(u)}$. As $\Gamma'(q) = \Gamma(h(q))$ and $\gamma'(q) = \gamma(h(q))$, we have

$q \in \Gamma'_u \iff h(q) \in \Gamma_u \implies h(q) \in \gamma_{\psi(u)} \iff (q_y)_{y\in\mathcal{Y}'} \in \gamma'_{\psi(u)}$, and therefore, L' indirectly elicits γ' via the link $\psi \circ \mathrm{proj}(\mathcal{Y}')$, where $\mathrm{proj}(\mathcal{Y}') : q \mapsto (q_y)_{y\in\mathcal{Y}'}$.

We aim to show $\dim(\mathcal{S}_{\gamma_r}(p)) \geq \dim(\mathcal{S}_{\gamma'_r}(p'))$. We do this by showing that $h(\mathcal{S}_{\gamma'_r}(p')) \subseteq \mathcal{S}_{\gamma_r}(p)$, and the result holds as h is linear and injective. Suppose $v \in h(\mathcal{S}_{\gamma'_r}(p'))$, then there exists a v' so that $v = h(v')$ and an $\epsilon_0 > 0$ such that $\epsilon v' + p' \in \gamma'_r$ for all $\epsilon \in (-\epsilon_0, \epsilon_0)$. Since h is linear and recall $h(\gamma'_r) \subseteq \gamma_r$, this implies $\epsilon v + p \in \gamma_r$ for all $\epsilon \in (-\epsilon_0, \epsilon_0)$. Therefore $v \in \mathcal{S}_{\gamma_r}(p)$, and the result follows.

As L' indirectly elicits γ', by Corollary 3, we know there exists a d-flat F with $p' \in F \subseteq \gamma'_r$. Taking $\mathcal{P} = \Delta_{\mathcal{Y}'}$, we know $p' \in \mathrm{relint}(\Delta_{\mathcal{Y}'})$ by construction, so we can apply Lemma 7(ii), which gives $\dim(\mathcal{S}_F(p')) \geq \dim(\mathrm{affhull}(\Delta_{\mathcal{Y}'})) - d = \|p\|_0 - 1 - d$.[6] Additionally, $\mathcal{S}_F(p') \subseteq \mathcal{S}_{\gamma'_r}(p')$ by subset inclusion of the sets themselves. Chaining these results, we obtain

$$\dim(\mathcal{S}_{\gamma_r}(p)) \geq \dim(\mathcal{S}_{\gamma'_r}(p')) \geq \dim(\mathcal{S}_F(p')) \geq \|p\|_0 - 1 - d \ .$$

□

E Proof of Theorem 2

Throughout this section, we will assume $\mathcal{P}$ is convex. See Frongillo and Kash [20, §E.5] for a discussion of how to relax this assumption.

E.1 General setting of elicitation complexity

We briefly introduce the general notion of elicitation complexity, of which Definition 6 is a special case, as some statements are more naturally made in this general setting.

Definition 14. Γ' refines Γ *if for all* $r' \in \mathrm{range}\,\Gamma'$ *there exists* $r \in \mathrm{range}\,\Gamma$ *with* $\Gamma'_{r'} \subseteq \Gamma_r$.

Equivalently, Γ' refines Γ if there is a link function $\psi : \mathrm{range}\,\Gamma' \to \mathrm{range}\,\Gamma$ such that $\Gamma'_{r'} \subseteq \Gamma_{\psi(r')}$ for all $r' \in \mathrm{range}\,\Gamma'$.

Definition 15. *For* $k \in \mathbb{N} \cup \{\infty\}$, *let* $\mathcal{E}_k(\mathcal{P})$ *denote the class of all elicitable properties* $\Gamma : \mathcal{P} \to \mathbb{R}^k$, *and* $\mathcal{E}(\mathcal{P}) := \bigcup_{k\in\mathbb{N}\cup\{\infty\}} \mathcal{E}_k(\mathcal{P})$. *When* $\mathcal{P}$ *is implicit we simply write* $\mathcal{E}$.

Definition 16. *Let* $\mathcal{C}$ *be a class of properties. The* elicitation complexity *of a property* Γ *with respect to* $\mathcal{C}$, *denoted* $\mathrm{elic}_{\mathcal{C}}(\Gamma)$, *is the minimum value of* $k \in \mathbb{N} \cup \{\infty\}$ *such that there exists* $\hat{\Gamma} \in \mathcal{C} \cap \mathcal{E}_k(\mathcal{P})$ *that refines* Γ.

E.2 Supporting statements

Proposition 5 (Osband [31])**.** *Let* Γ *be elicitable. Then* Γ_r *is convex for all* $r \in \mathrm{range}\,\Gamma$.

Lemma 8 (Set-valued extension of Frongillo and Kash [20, Lemma 4])**.** *If* Γ' *refines* Γ *then* $\mathrm{elic}_{\mathcal{C}}(\Gamma') \geq \mathrm{elic}_{\mathcal{C}}(\Gamma)$.

Proof. As Γ' refines Γ, we have some $\psi : \mathrm{range}\,\Gamma' \to \mathrm{range}\,\Gamma$ such that for all $r' \in \mathrm{range}\,\Gamma'$ we have $\Gamma'_{r'} \subseteq \Gamma_{\psi(r')}$. Suppose we have $\hat{\Gamma} \in \mathcal{C}$ and $\varphi : \mathrm{range}\,\hat{\Gamma} \to \mathrm{range}\,\Gamma'$ such that for all $u \in \mathrm{range}\,\hat{\Gamma}$ we have $\hat{\Gamma}_u \subseteq \Gamma'_{\varphi(u)}$. Then for all $u \in \mathrm{range}\,\hat{\Gamma}$ we have $\hat{\Gamma}_u \subseteq \Gamma'_{\varphi(u)} \subseteq \Gamma_{(\psi\circ\varphi)(u)}$. In particular, if $\mathrm{elic}_{\mathcal{C}}(\Gamma') = m$, then we have such a $\hat{\Gamma} : \mathcal{P} \rightrightarrows \mathbb{R}^m$, and hence $\mathrm{elic}_{\mathcal{C}}(\Gamma) \leq m$. □

Lemma 9 (Frongillo and Kash [20, Lemma 8])**.** *Suppose* $L \in \mathcal{L}$ *elicits* $\Gamma : \mathcal{P} \to \mathcal{R}$ *and has Bayes risk* $\underline{L}$. *Then for any* $p, p' \in \mathcal{P}$ *with* $\Gamma(p) \neq \Gamma(p')$, *we have* $\underline{L}(\lambda p + (1-\lambda)p') > \lambda\underline{L}(p) + (1-\lambda)\underline{L}(p')$ *for all* $\lambda \in (0, 1)$.

Lemma 10 (Adapted from Frongillo and Kash [20, Theorem 4])**.** *If* L *elicits a single-valued* Γ, *and* $\hat{\Gamma}$ *refines* $\underline{L}$, *then* $\hat{\Gamma}$ *refines* Γ.

[6] To reason about $\dim(\mathrm{affhull}(\Delta_{\mathcal{Y}'})) = \|p\|_0 - 1$, observe that the uniform distribution on $\Delta_{\mathcal{Y}'}$ has full support and therefore requires $\|p\|_0 - 1$ elements in its basis.

Proof. Suppose for a contradiction that $\hat{\Gamma}$ does not refine Γ. Then we have some $u \in \operatorname{range}\hat{\Gamma}$ such that for all $r \in \operatorname{range}\Gamma$ we have $\hat{\Gamma}_u \not\subseteq \Gamma_r$. In particular, recalling that Γ is single-valued, we must have $p, p' \in \hat{\Gamma}_u$ such that $\Gamma(p) \neq \Gamma(p')$. Moreover, as $\hat{\Gamma}$ refines $\underline{L}$, we also have $\underline{L}(p) = \underline{L}(p')$. From Lemma 9 and $\lambda = 1/2$ we have $\underline{L}(q) > \frac{1}{2}\underline{L}(p) + \frac{1}{2}\underline{L}(p') = \underline{L}(p)$, where $q = \frac{1}{2}p + \frac{1}{2}p'$. As the level set $\hat{\Gamma}_u$ is convex by Proposition 5, we also have $q \in \hat{\Gamma}_u$, and hence $\underline{L}(q) = \underline{L}(p)$, a contradiction. □

Lemma 11 (Minor modifications from Frongillo and Kash [20])**.** *Let $\mathcal{V}$ be a real vector space. Let $f : \mathcal{V} \to \mathbb{R}^k$ be linear and $C \subseteq \mathcal{V}$ convex with* $\operatorname{span} C = \mathcal{V}$*, and let $m \in \mathbb{N}$. Suppose that $0 \in \operatorname{int} f(C)$, and for all $v \in S := C \cap \ker f$, there exists a linear $\hat{f}_v : \mathcal{V} \to \mathbb{R}^m$ with $v \in C \cap \ker \hat{f}_v \subseteq S$. Then $m \geq k$. If $m = k$, we additionally have $0 \in \operatorname{int} \hat{f}_v(C)$ for some $v \in S$.*

Proof. The condition $0 \in \operatorname{int} f(C)$ is equivalent to the existence of some $v_1, \ldots v_{k+1} \in C$ such that $0 \in \operatorname{int}\operatorname{conv}\{f(v_i) : i \in \{1, \ldots, k+1\}\}$. Let $\alpha_1, \ldots, \alpha_{k+1} > 0$, $\sum_{i=1}^{k+1} \alpha_i = 1$, such that $\sum_{i=1}^{k+1} \alpha_i f(v_i) = 0$. As these are barycentric coordinates, this choice of α_i is unique, a fact which will be important later. We will take $v = \sum_{i=1}^{k+1} \alpha_i v_i$, an element of C by convexity, and thus an element of S as $f(v) = 0$.

Let $\hat{f}_v : \mathcal{V} \to \mathbb{R}^m$ be linear with $v \in \hat{S} := C \cap \ker \hat{f}_v \subseteq S$. Let $\beta_1, \ldots, \beta_{k+1} \in \mathbb{R}$, $\sum_{i=1}^{k+1} \beta_i = 0$, such that $\sum_{i=1}^{k+1} \beta_i \hat{f}_v(v_i) = 0$. We will show that the β_i must be identically zero, i.e. that $\{\hat{f}_v(v_i) : i \in \{1, \ldots, k+1\}\}$ are affinely independent. By construction, $v' := \sum_{i=1}^{k+1} \beta_i v_i \in \ker \hat{f}_v$, and as $v \in \ker \hat{f}_v$, for all $\lambda > 0$ we have $v_\lambda := v + \lambda v' = \sum_{i=1}^{k+1}(\alpha_i + \lambda\beta_i)v_i \in \ker \hat{f}_v$. Taking λ sufficiently small, we have $\gamma_i := \alpha_i + \lambda\beta_i > 0$ for all i, and $\sum_{i=1}^{k+1} \gamma_i = \sum_{i=1}^{k+1} \alpha_i + \lambda \sum_{i=1}^{k+1} \beta_i = 1$. By convexity of C, we have $v_\lambda \in C$. Now $v_\lambda \in C \cap \ker \hat{f}_v \subseteq S = C \cap \ker f$, and in particular $v_\lambda \in \ker f$. Thus, $f(v_\lambda) = \sum_{i=1}^{k+1} \gamma_i f(v_i) = 0$. By the uniqueness of barycentric coordinates, for all $i \in \{1, \ldots, k+1\}$, we must have $\gamma_i = \alpha_i$ and thus $\beta_i = 0$, as desired.

As $\hat{f}_v(C)$ contains $k+1$ affinely independent points, we have $m \geq \dim \operatorname{im} \hat{f}_v \geq k$. When $m = k$, by affine independence, the set $\operatorname{conv}\{\hat{f}_v(v_i) : i \in \{1, \ldots, k+1\}\}$ has dimension k in $\mathbb{R}^k$. As $0 = \hat{f}_v(v) = \sum_{i=1}^{k+1} \alpha_i \hat{f}_v(v_i)$, and $\alpha_i > 0$ for all i, we conclude $0 \in \operatorname{int}\operatorname{conv}\{\hat{f}_v(v_i) : i \in \{1, \ldots, k+1\}\} \subseteq \operatorname{int} \hat{f}_v(C)$. □

Lemma 12 (Frongillo and Kash [20, Lemma 14])**.** *Let $\mathcal{V}$ be a real vector space. Let $f : \mathcal{V} \to \mathbb{R}^k$ be linear, $C \subseteq \mathcal{V}$ convex with* $\operatorname{span} C = \mathcal{V}$*, and let $S = C \cap \ker f$. If $0 \in \operatorname{int} f(C)$ then* $\operatorname{span} S = \ker f$*.*

E.3 Proving the lower bound for Bayes risks

Let $\mathcal{C}_d^*$ be the class of properties Γ which are elicited by a convex loss $L \in \mathcal{L}_d^{\text{cvx}}$ for some $d \in \mathbb{N}$, and let $\mathcal{C}^* := \bigcup_{d \in \mathbb{N}} \mathcal{C}_d^*$. Then for all properties γ, if $\operatorname{elic}_{\mathcal{C}^*}(\gamma) < \infty$, we have $\operatorname{elic}_{\mathcal{C}^*}(\gamma) = \operatorname{elic}_{\text{cvx}}(\gamma)$, a fact we use tacitly in the proof.

Theorem 2. *Let $\mathcal{P}$ be a convex set of Lebesgue densities supported on the same set for all $p \in \mathcal{P}$. Let $\Gamma : \mathcal{P} \to \mathbb{R}^d$ satisfy Condition 1 for some $r \in \mathbb{R}^d$. Let $L \in \mathcal{L}^{\text{cvx}}$ elicit Γ such that $\underline{L}$ is non-constant on Γ_r. Then* $\operatorname{cons}_{\text{cvx}}(\underline{L}) \geq \operatorname{elic}_{\text{cvx}}(\underline{L}) \geq d + 1$*.*

Proof. Let $V : \mathcal{Y} \to \mathbb{R}^d$ and r be given by the statement of the theorem and from Condition 1. Let $m = \operatorname{elic}_{\mathcal{C}^*}(\underline{L})$, so that we have $\hat{\Gamma} \in \mathcal{C}_m^*$ which refines $\underline{L}$. By Lemma 10 we have $\hat{\Gamma}$ refines Γ.

We now establish the conditions of Lemma 11 for $C = \mathcal{P}$. Let $f : \operatorname{span}\mathcal{P} \to \mathbb{R}^d$, $p \mapsto \mathbb{E}_p V$. From Condition 1, we have $0 \in \operatorname{int} f(\mathcal{P})$ and $\ker f \cap \mathcal{P} = \ker_{\mathcal{P}} V = \Gamma_r$. Now let $p \in \Gamma_r$ be arbitrary, and take any $u \in \hat{\Gamma}(p)$. As Γ is single-valued, $r \in \operatorname{range}\Gamma$ is the unique value with $p \in \Gamma_r$. As $\hat{\Gamma}$ refines Γ, there exists $r' \in \operatorname{range}\Gamma$ with $\hat{\Gamma}_u \subseteq \Gamma_{r'}$, and since $p \in \hat{\Gamma}_u$, we conclude $r' = r$ from the above. From Theorem 1, we have some $\hat{V}_{u,p}$ with

$p \in \ker_{\mathcal{P}} \hat{V}_{u,p} \subseteq \hat{\Gamma}_u \subseteq \Gamma_r = \ker_{\mathcal{P}} V$. Letting $\hat{f}_p : \mathrm{span}\mathcal{P} \to \mathbb{R}^d$, $p \mapsto \mathbb{E}_p \hat{V}_{u,p}$, we have now satisfied the conditions of Lemma 11. We conclude $m \geq d$, and moreover, if $m = d$, then there exists some $q \in \Gamma_r$ such that $0 \in \mathrm{int}\, \hat{f}_q(\mathcal{P})$.

Now suppose $m = d$ for a contradiction. Let $\hat{S} := \ker f_q \cap \mathcal{P}$. Applying Lemma 12 to the functions f and $\hat{f}_q$ we have $\mathrm{span}\ker f = \mathrm{span}\Gamma_r$ and $\mathrm{span}\ker \hat{f}_q = \mathrm{span}\hat{S}$. As $\hat{S} \subseteq \Gamma_r$, we have $\ker \hat{f}_q = \mathrm{span}\hat{S} \subseteq \mathrm{span}\Gamma_r = \ker f$. By the first isomorphism theorem, we also have $\mathrm{codim}\ker \hat{f}_q = \mathrm{codim}\ker f = d$, as the images of these linear maps span all of $\mathbb{R}^d$. By the third isomorphism theorem we conclude $\Gamma_r = \hat{S}$. Moreover, as $\hat{S} \subseteq \hat{\Gamma}_u \subseteq \Gamma_r$, we have $\hat{S} = \hat{\Gamma}_u = \Gamma_r$.

We now see that $\underline{L}$ is constant on Γ_r since there is some link function $\psi : \mathbb{R}^m \to \mathbb{R}$ such that $\Gamma_r = \hat{\Gamma}_u \subseteq \underline{L}_{\psi(u)}$, meaning $\underline{L}(p) = \psi(u)$ for all $p \in \Gamma_r$. This statement contradicts the assumption that $\underline{L}$ is non-constant on Γ_r. $\square$

F Omitted Discussion and Examples

F.1 Note on restricting minimizable assumption

While some popular surrogates such as logistic and exponential loss are not minimizable, these losses are still covered in Corollary 3 and Theorem 2 as $\Gamma(p) \neq \emptyset$ when $p \in \mathcal{P} := \mathrm{relint}(\Delta_{\mathcal{Y}})$; moreover, by thresholding $L'(u, y) = \max(L(u, y), \epsilon)$ for sufficiently small $\epsilon > 0$ we can achieve $L' \in \mathcal{L}$ for both. We expect that a generalization of property elicitation which allows for "infinite" predictions (e.g., along a prescribed ray) would allow us to assume minimizability without loss of generality, as convex losses would always admit this more general minimizer.

F.2 Lower-bounding the convex consistency dimension of the variance

Corollary 7. *Let $\mathcal{P}$ be a set of continuous Lebesgue densities on $\mathcal{Y} = \mathbb{R}$ with all $p \in \mathcal{P}$ having the same support. If there exist $p, q, q' \in \mathcal{P}$ with $\mathbb{E}_p Y = \mathbb{E}_q Y \neq \mathbb{E}_{q'} Y$ and $\mathrm{Var}(p) \neq \mathrm{Var}(q)$, then $\mathrm{cons}_{\mathrm{cvx}}(\mathrm{Var}) = \mathrm{elic}_{\mathrm{cvx}}(\mathrm{Var}) = 2$.*

Proof. For the upper bound, we may elicit the first two moments via the convex loss $L(r, y) = (r_1 - y)^2 + (r_2 - y^2)^2$, and recover the variance via $\psi(r) = r_2 - r_1^2$, giving $\mathrm{elic}_{\mathrm{cvx}}(\mathrm{Var}) \leq 2$. Now for the lower bound. Without loss of generality, $\mathbb{E}_q Y < \mathbb{E}_{q'} Y$. Let $r = \frac{1}{2}\mathbb{E}_q Y + \frac{1}{2}\mathbb{E}_{q'} Y$, and define $V : \mathcal{Y} \to \mathbb{R}, y \mapsto y - r$. Then $\ker_{\mathcal{P}} V = \{p' \in \mathcal{P} \mid \mathbb{E}_{p'} Y = r\} = \Gamma_r$ where $\Gamma : p' \mapsto \mathbb{E}_{p'} Y$ is the mean. As $\mathbb{E}_q Y < r < \mathbb{E}_{q'} Y$, we conclude $\mathbb{E}_q V < 0 < \mathbb{E}_{q'} V$. We have now satisfied Condition 1 for $d = 1$. To apply Theorem 2, it remains to show that Var is non-constant on Γ_r. By our assumptions and the definition of Var, we have $\mathbb{E}_p Y^2 \neq \mathbb{E}_q Y^2$. Letting $p_1 = \frac{1}{2} q + \frac{1}{2} q'$, $p_2 = \frac{1}{2} p + \frac{1}{2} q'$, we have $\mathbb{E}_{p_i} Y = r$ for $i \in \{1, 2\}$, but $\mathbb{E}_{p_1} Y^2 = \frac{1}{2}\mathbb{E}_q Y^2 + \frac{1}{2}\mathbb{E}_{q'} Y^2 \neq \frac{1}{2}\mathbb{E}_p Y^2 + \frac{1}{2}\mathbb{E}_{q'} Y^2 = \mathbb{E}_{p_2} Y^2$. As p_1, p_2 have the same mean but different second moments, we conclude $\mathrm{Var}(p_1) \neq \mathrm{Var}(p_2)$. $\square$

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] **See discussion in the Introduction; it is hard to forecast impacts of theoretical work, but we try to explain some impacts.**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] **Sometimes details are deferred to Appendix, but we make an effort to explicitly state when that happens.**
   (b) Did you include complete proofs of all theoretical results? [Yes] **Sometimes proofs are deferred to Appendix, but we make an effort to explicitly state when that happens.**
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]